

# Effective Data Analysis

Hard and soft skills

Mona Khalil

MEAP

 MANNING



# Effective Data Analysis

Hard and soft skills

Mona Khalil

MEAP



MANNING



# Effective Data Analysis

1. [welcome](#)
2. [1 What does an analyst do?](#)
3. [2 From Question to Deliverable](#)
4. [3 Testing and Evaluating Hypotheses](#)
5. [4 The Statistics You \(Probably\) Learned: T-Tests, ANOVAs, and Correlations](#)
6. [5 The Statistics You \(Probably\) Didn't Learn: Non-Parametric Tests, Chi-Square Tests, and Responsible Interpretation](#)
7. [6 Are you measuring what you think you're measuring?](#)
8. [7 The Art of Metrics: Tracking Performance for Organizational Success](#)
9. [8 Navigating Sensitive and Protected Data](#)
10. [9 The World of Statistical Modeling](#)

# welcome

Thank you for purchasing the MEAP for *Effective Data Analysis*. I hope this book will be of immediate use to you in your work as an analyst. With your help, the final book will be a tool for you to accelerate your career in data analytics, data science, and more.

Early in my career in research and analytics, I discovered a large gap between the technical skills I was taught (statistics, R, SPSS, SQL, etc.) and the delivery of a final product that provides tangible, actionable recommendations to my stakeholders. Like many junior analysts, I learned through trial and error, with many failed deliverables I recreated until they were understood by the team who requested them.

With some amazing mentors, I grew in my capacity as a data scientist and a data analyst, eventually growing into a leadership role. Along the way, I sought to support and mentor others who were early in their career, discovering many of them shared the same struggles that I once did. This book is my intention to put that mentorship to paper and create a definitive set of resources for you to maximize your contribution and value in analytics while growing your career.

This book is written assuming you have most or all of the following foundational skills of analytics:

- Knowledge of relational databases and how to query them with SQL
- Knowledge of univariate parametric statistical tests (e.g., t-tests, ANOVAs, linear regression)
- Knowledge of Python (pandas, matplotlib, seaborn, numpy)

Throughout this book, we will cover a wide range of skills designed to support you in your day to day work, giving you the skills necessary to build a rich set of experience in your domain of expertise. By the end of this book, you will have learned:

- How to ask the right questions of your data, including hypothesis development, operationalizing challenging concepts, and choosing data sources and data collection methods that best answer your question
- How to use statistical tests effectively, including appropriate selection of tests based on the characteristics of your data, using non-parametric tests, and interpreting the results responsibly
- Developing effective measurements and metrics to guide the success of your business or organization
- Building a toolkit of resources, including flexibility in synthesizing data for your tests and models, strategically choosing an approach to modeling, and automating repeatable analytics processes to optimize your time
- Building a data-informed culture with your stakeholders and organization

Please leave comments in the [liveBook Discussion forum](#) and let me know what you think about this book so far. My intention is to put together a resource that I wish existed for my own career and those of many people I've supported, and your feedback will support me in achieving that goal.

Thank you again for your interest in this book and for purchasing the MEAP!

#### **In this book**

[welcome](#) [1 What does an analyst do?](#) [2 From Question to Deliverable](#) [3 Testing and Evaluating Hypotheses](#) [4 The Statistics You \(Probably\) Learned: T-Tests, ANOVAs, and Correlations](#) [5 The Statistics You \(Probably\) Didn't Learn: Non-Parametric Tests, Chi-Square Tests, and Responsible Interpretation](#) [6 Are you measuring what you think you're measuring?](#) [7 The Art of Metrics: Tracking Performance for Organizational Success](#) [8 Navigating Sensitive and Protected Data](#) [9 The World of Statistical Modeling](#)

# 1 What does an analyst do?

## This chapter covers

- Introducing analytics
- A review of common analytic domains
- Using a data analyst's toolkit
- Preparing for your first role

So you're a newly minted data analyst—congratulations! Perhaps you just finished school and are looking for your first role, or maybe you just started your first job in this field. It's possible you planned this career path, but it's also possible you landed here without being as prepared as you would have liked. Maybe you're part of a large team, or maybe you're the first and only analyst at your organization. All of that's okay! There are *so many* paths into the world of data, and each one brings its unique challenges. If you're looking to do the best work you can, become a data professional, and an expert in making data-informed decisions, then this book was written for you.

Analytics is *everywhere*. The role of a data analyst has been seen in nearly every type of organization for *decades*, and mature organizations will almost always have multiple teams dedicated to the effective use of data. These dedicated functions have familiar names like business analytics, business intelligence, product analytics, and data science, and are dedicated to providing the organization with information needed to make strategic decisions. These days, organizations have more data than ever, making it especially critical to understand your users, customers, and stakeholders in your work. Being data-driven is more important—and doable—than ever.

A lot of attention and hype is focused on working with data. Much of that is tied to the work of a data scientist or machine learning practitioner, training models to generate predictions that inform or make decisions. The varied applications of machine learning and data science methodology can elevate the value generated within a business. However, much of that value benefits from a strong foundation in analytics.

Across many titles in a data practice, being an effective analyst is necessary to derive value from your stakeholders. Throughout this book, we will cover various topics foundational to being a skilled analyst capable of producing deliverables that continue delivering value for your organization. We will cover a range of soft and technical skills covered less often in a data analyst or data scientist curriculum and strategies to set yourself up for success.

## 1.1 What is analytics?

*Analytics* is an all-encompassing term for a broad domain with many definitions. For this book, we will define analytics as the *practice of leveraging data to discover and communicate patterns, trends, and insights that inform decisions*. An analyst leverages a range of methods to describe and infer information about a data set. These can include descriptive statistics, inferential statistics, statistical models, financial models, and more. The specific methods used vary by field, with a set of core approaches and best practices that tend to be used by the majority of analysts.

Analytics within an organization is frequently organized into one or more of the following domains and departments:

### 1.1.1 Business Intelligence

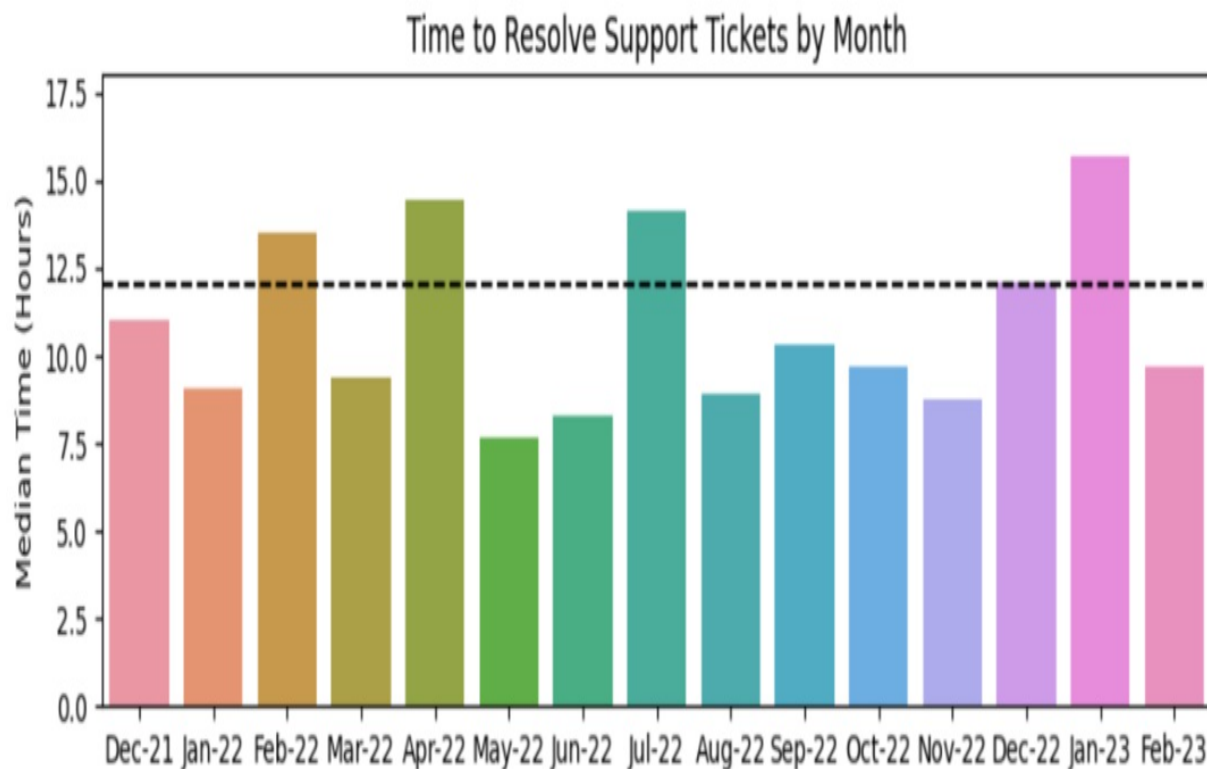
A business intelligence or business analytics team enables tracking and analyzing data about an organization's performance and makes informed and strategic operational decisions across various functions. This type of team can employ a wide variety of methods of synthesizing data and communicating results but typically aims to present results in a clear and readable format for stakeholders less familiar with the interpretation of statistics and mathematics.

The specific tasks and workflow owned by a business analytics team vary by the domain and size of an organization but will typically involve the following.

#### Developing Metrics and KPIs

Setting and tracking core **metrics** (standardized quantitative data tracked over time) and **key performance indicators** (KPIs; the most important indicators of performance) is foundational to the success of a data-informed organization. Many business intelligence teams will track a combination of *standard metrics* (used across industry or field) and *custom metrics* (unique to the organization) to provide a comprehensive picture of performance. These metrics are distilled into tools such as **dashboards** for ease of consumption, understanding, and decision-making.

**Figure 1.1** Line graph of a support team KPI with a threshold for the goal of resolving tickets in less than 12 hours. Metrics and KPIs generate value when tracked over time.



## Developing Reports to Generate Business Insights

In addition to developing and tracking metrics, a business intelligence team will often dedicate its time to generating novel insights about the function and operation of the business. They may identify areas of inefficiency, revenue-generating opportunities, answer questions from stakeholders to enable them to make increasingly strategic decisions. These results are often

shared as **reports** or **presentations**.

## **Developing Dashboards for Ease of Information Consumption**

Nearly every type of analytics team produces **dashboards** as a deliverable. Dashboards are highly curated visual representations of data, typically containing interactive charts and graphs that provide insight into a specific area of the organization. Dashboard also preferably use data sources that are automatically refreshed when new records are added, minimizing the amount of time the team spends supporting routine updates for stakeholders.

Business intelligence teams will typically use a business intelligence tool (BI tool) that is either purchased as a software (e.g., Tableau, PowerBI) or built and maintained by the organization. The team will often create dashboards to track metrics, key performance indicators, and trends that are monitored regularly by stakeholders. These tools are powerful assistants to the team, enabling much of the organization to become data-driven in their decision-making without needing direct support from the business intelligence analysts.

**Figure 1.2 A dashboard typically contains summary information and the highest-value visualizations for quick interpretation.**

Win Rate

36% -

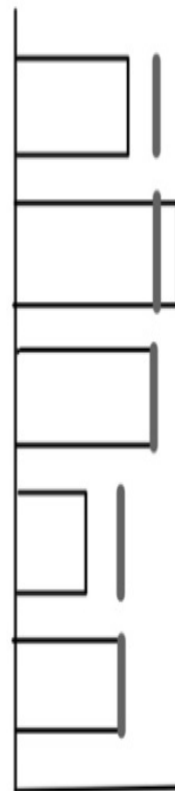
Revenue this Year

\$1.5m ↑

Total Customers

3,000 ↑

Weekly Sales



## **Distilling and Communicating Results to Business Stakeholders**

A business intelligence team is highly flexible in their delivery of insights to the stakeholders they support. Depending on the purpose of the analysis and the data literacy of their stakeholders, they will need to tailor their use of statistical and mathematical language, the depth of their analysis, and the formatting of the deliverable in order to maximize value. Deliverables may include dashboards, reports, summarized insights, or presentations.

### **A Note on Terminology**

It's important to quickly note that *business intelligence* and *business analytics* are not entirely interchangeable. Gartner defines business analytics as the specific application of analysis and statistical methods to inform a business. Some sources describe business intelligence as a more encompassing function that can include skills and tasks such as data mining, machine learning, data engineering, data governance, and more. In practice, the use of these terms may be interchangeable and continually evolving with the needs of an organization.

Further, depending on the size and structure of an organization, a business intelligence function can include additional specializations such as marketing analytics, financial analytics, and human resources analytics. However, the primary distinguishing characteristic of *business intelligence* is that it supports the internal operational need for data within an organization.

### **1.1.2 Marketing Analytics**

*Marketing analytics* finds patterns in data related to an organization's marketing efforts. Evaluating and optimizing email campaigns, advertisements, conversion rates, and customer/prospective customer engagement are all common areas of focus within marketing analytics.

A marketing analytics team will often perform similar tasks to a business intelligence team. For example, a marketing analyst may track metrics and KPIs for a marketing team, create a dashboard, and develop an ad-hoc attribution model to understand where visitors are converting to users in the

pipeline.

## Experimentation

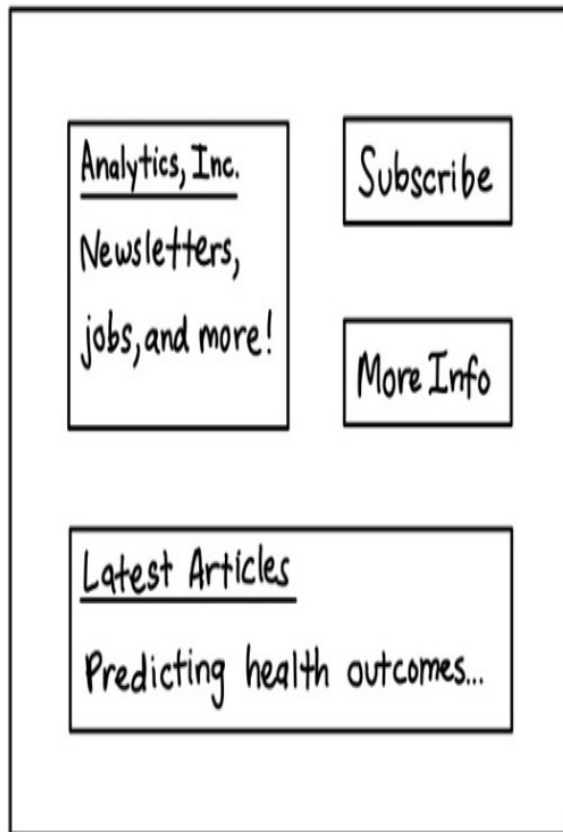
**Experimentation** refers to the process of testing a hypothesis in controlled conditions to discover cause and effect relationships. Many organizations use experimental procedures in order to guide the design and improvement of websites, applications, and products; without these approaches, teams may find themselves guessing which approaches would create a desired outcome for their users.

A/B testing is a common experimental procedure used by marketing analytics teams to understand how small iterations impact the engagement of prospective customers or users. These tests split a subset of users into one or more *experiment* and *control* groups, showing variations of an advertisement that invites them to ask for a product demonstration or convert to paid subscribers. The experiment group with the highest conversion rate is considered the winner, and is then shown to the entire prospect or user base.

By splitting your users, customers, or prospects into separate groups and testing variations of text, colors, images, calls to action, etc., you can comprehensively understand their wants, needs, interests, and behavioral trends over time. Further, many A/B tests are conducted using the same statistical tests that you may have covered in a college statistics course! We'll discuss these far more in depth in chapter 3.

Experiments are generally delivered to stakeholders as a **report** summarizing findings, impact, and recommended next steps.

**Figure 1.3 Example of two conditions in an A/B test. Small, iterative variations like this can significantly improve user engagement and revenue.**



Variation A



Variation B

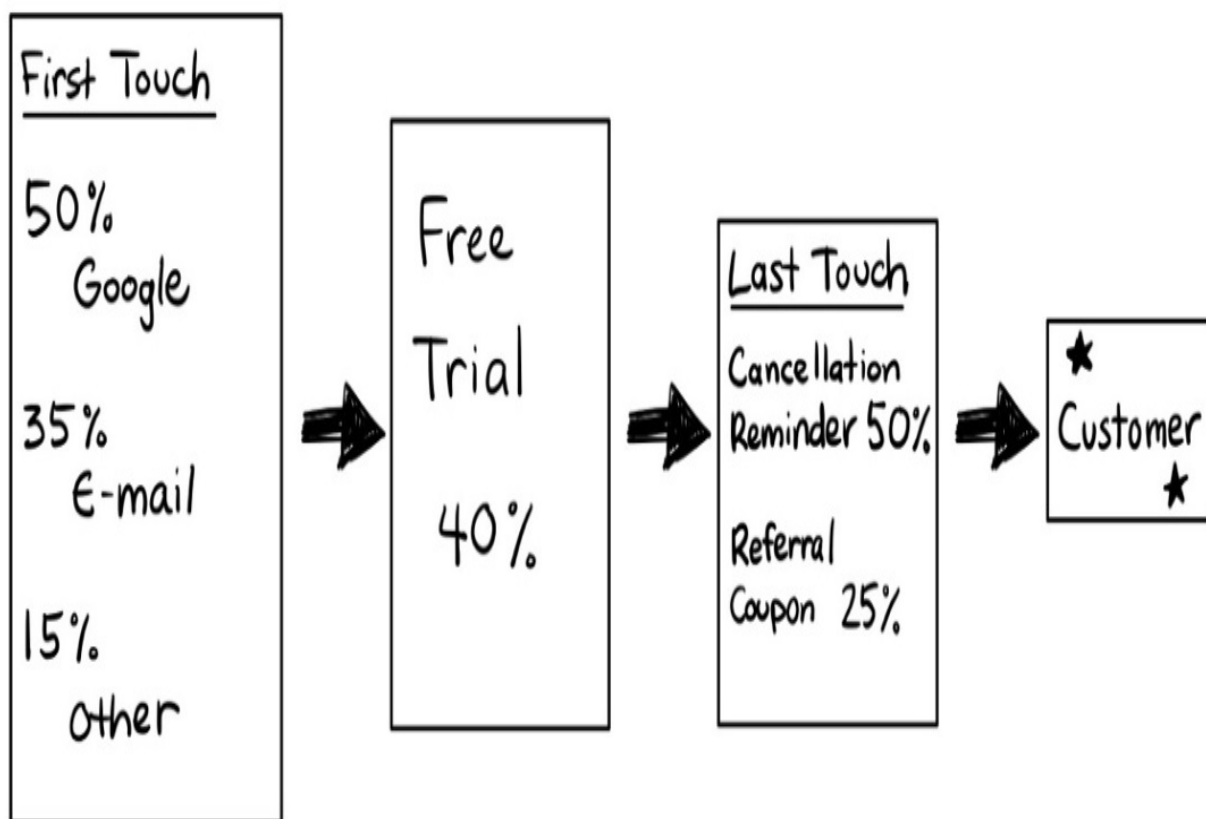
## Attribution Modeling

**Attribution modeling** is the analysis of each *touchpoint*, or step, prior to a purchase or subscription. For example, a prospective user may follow these steps in order: click on an advertisement, visit the marketing page, start a free trial, and subscribe. A percentage of users at each step will complete the following step, with a very small number making it to the end of the funnel.

The task of the marketing analytics team is to determine what proportion of the success (subscription) can be *attributed* to each touchpoint and understand which are the most valuable in the customer acquisition process.

Some simple methods include first-touch attribution (attributing all credit to the first touchpoint) and last-touch attribution (attributing all credit to the final touchpoint). More complex approaches include multi-touch attribution and algorithmic techniques using statistical models. Each of the above involves delivering an analysis breaking down the top sources of traffic or subscriptions at the selected touchpoint.

**Figure 1.4 Attribution model showing first/last touch and example intermediary steps. Each model breaks down the sources at the touchpoint to understand which is most successful at generating new customers.**



## Competitive Analysis

**Competitive analysis** involves various approaches to researching and obtaining publicly available data on competitor performance and business practices. This type of analysis helps an organization determine its market fit, ideal user profiles and understand specific areas where its competitors tend to

win or lose. A marketing analytics team may be involved directly in the research and compiling of information for the competitive analysis, as well as any comparisons of quantitative data discovered in the research process. This function is often performed collaboratively with a finance or financial analytics team.

### **1.1.3 Financial Analytics**

Financial analytics teams leverage payment and financial data about an organization to understand trends in its performance over time. Generating financial insights encompass a range of similar tools and methods to business analysis and may involve cross-functional overlap with marketing analytics or other functions where the health of the business is concerned.

Depending on the business, a financial analytics team may include functions that require specialized coursework or skill sets (e.g., risk analysis). An investment firm will need a different set of deliverables from a financial analysis team than a software company, and jobs at these types of companies will have correspondingly different requirements. The following section highlights financial analytic team approaches common to many types of businesses.

#### **Financial Metrics**

Financial analytics teams will monitor and report on a comprehensive set of standard metrics such as revenue, profitability, and customer lifetime value tracked by nearly every organization where those metrics are applicable. Each metric is monitored within organizations to understand the growth trajectory and the impact of various team functions on that growth potential. These metrics often serve as *outcome measures* for other teams seeking to understand the impact of more specific actions on organizational performance.

#### **Risk Analysis**

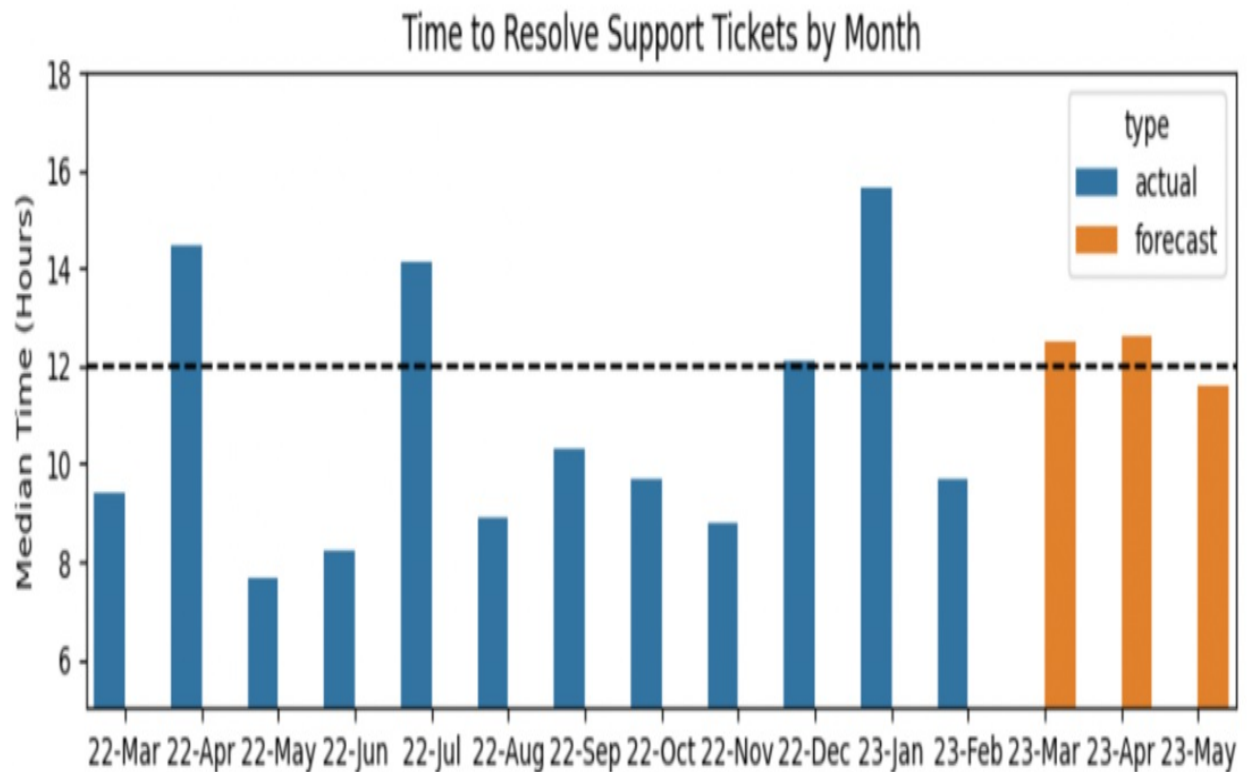
**Risk analysis** assesses the likelihood of different types of risk to an

organization, such as a reduction in revenue, an increase in customer churn, or an increase in operational costs. Financial analysts perform simulations and develop forecasting models and other approaches to quantify a business's numerous potential risks. The mathematical models a risk analysis includes can be complex but are ultimately limited by the number of factors that can be accounted for in a model.

## **Business Forecasting**

**Forecasting** models use historical data to provide insight into the expected financial performance of an organization. These can include projected growth based on seasonal and most recent trends, augmented by organizational factors and broader economic indicators. A range of statistical methods are used for this type of analysis, and organizations hiring to meet this need will often specify a requirement for skills in standard forecasting methods.

**Figure 1.5** Figure 1.1, with an additional 3-month rolling average provided as a forecast. Forecasting methods range from simple calculations such as this one to more complex time series modeling approaches. A simple forecast like the above will typically have less variability than the actual data.



### 1.1.4 Product analytics

Product analytics is the analysis of product usage and users to understand and continually improve their experience with a product. Product analysis typically resides within a research & development (R&D) department, supporting a product team in understanding users' needs, the value of investments, and more. This function is quite common at software companies. Product analysis can be performed in a *decentralized* capacity, where product managers, software engineers, and other team members work together to answer questions leveraging data; in a *centralized* capacity, product analysts answer questions using data to support the department's ability to make data-informed decisions.

### Opportunity sizing

An essential component of product development involves appropriately quantifying the **opportunity cost** of pursuing a specific line of work. A product analytics team will try to answer questions about the expected impact

on subscriptions, user engagement, productivity, or any metric of interest based on the range of available data related to the opportunity. For example, a product team is considering redesigning parts of the website dedicated to a specific segment of users. The team discovers from available data that this segment of users has proportionally low engagement (website visits per week), tends to generate more support tickets than other segments, and tends to cancel their subscriptions more frequently. This new design addresses the most common sources of confusion mentioned in support tickets.

The picture provided by this range of data sources allows the product team to develop a hypothesis around the expected outcomes associated with redesigning parts of the site, compare expected labor costs to projected loss in revenue or engagement associated with *not* pursuing the opportunity, and more. This type of analysis is typically followed up by using the initial data points as success metrics and outcomes of interest to evaluate after the project is complete.

## **Experimentation**

The experimental procedures that marketing analytics teams use are also frequently owned and performed by product analytics or *growth* teams. In addition to simple iterations on layouts, buttons, text, etc., product analytics teams will use a broad range of methods to design and evaluate more sophisticated experiments. These may include longer-running A/B tests on complex workflows with multiple outcome metrics and a more comprehensive range of statistical tests for evaluation.

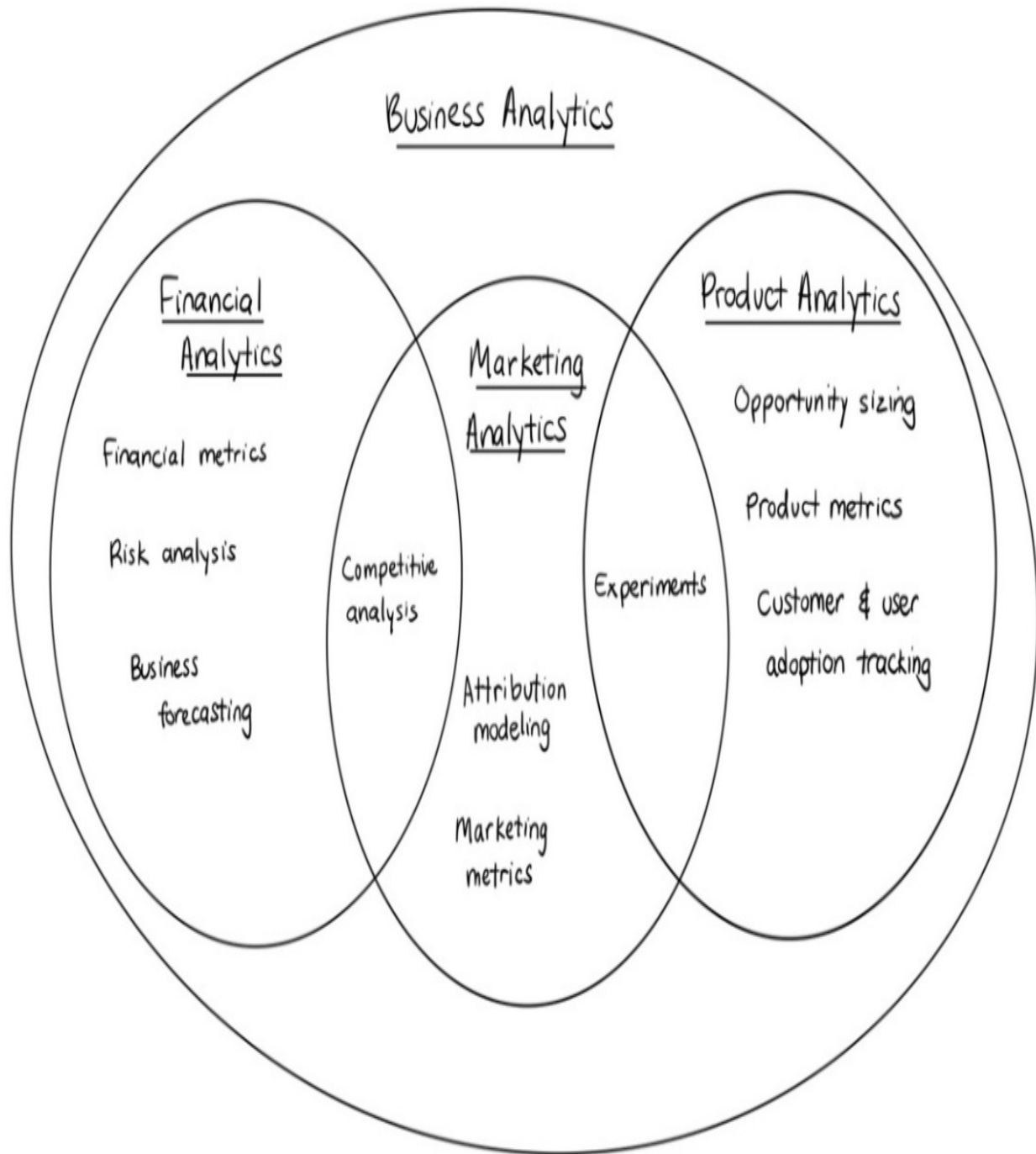
In addition to between-group evaluations, product analytics teams will use pre/post comparisons to measure impact, quasi-experimental designs for when a true experiment is impossible, and others. The appropriate use of these methods and statistical tests to evaluate them will be covered in chapters 3 and 4.

### **1.1.5 How distinct are these fields?**

Analytics functions and teams have a noticeable overlap in methods, tasks, approaches, and stakeholders. The line between teams and functions may blur

within an organization or for an individual role. The shape of an analytics practice within an organization constantly evolves, and you will readily discover opportunities for increased collaboration and division of labor. This is especially true earlier in your career, when you may have a similar education and skillset as other analysts you meet. Over time, you will build a profile of specialized skills unique to the analytics function you work with and greater exposure to the needs and problems of that type of team.

**Figure 1.6 Concentric circles showing common areas of overlap according to categories of deliverables provided by different analytics teams.**



## 1.2 The Data Analyst's Toolkit

A data analyst who has completed an education, training program, or coursework in this field will generally be exposed to various tools and languages necessary to complete their work. The availability of the tools

varies considerably by company. In your first role, you may find access to a range of sophisticated proprietary tooling maintained by an engineering team, or you may only have access to free versions of software you learned to use in a classroom.

Regardless of your organization's previous investment in data tooling, you will benefit from accessing the following categories of tools for your work.

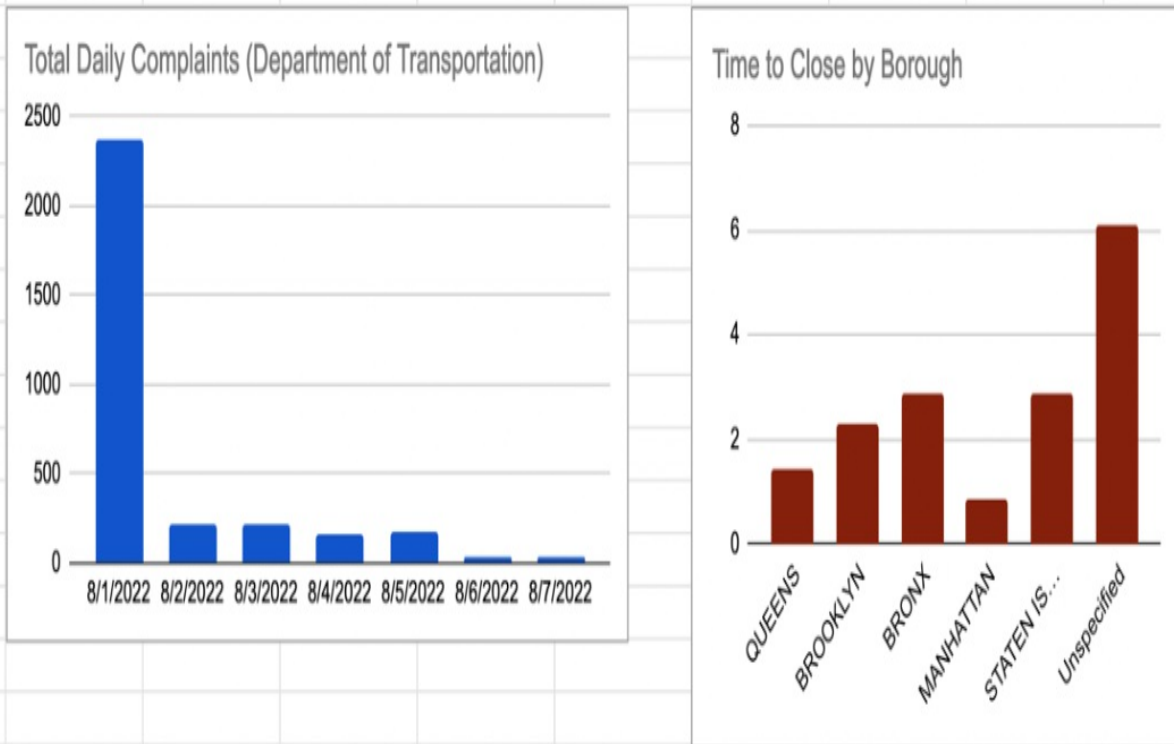
### **1.2.1 Spreadsheet Tools**

In *all* titles and seniority levels, a data practitioner needs a readily available spreadsheet tool to directly manipulate, shape, present, and interact with data. Spreadsheets are often considered the *least common denominator* of the data world. Appropriately using a programming language and development environment can mitigate the frustration of interacting with large, slow spreadsheets. As the most widely used data manipulation and analysis software on the market, there's no avoiding the periodic need for a spreadsheet.

If you cannot access a proprietary spreadsheet tool such as Microsoft Excel, the freely available G-Suite and Google Sheets will meet most of your needs to manipulate data and add charts, formulas, and pivot tables. G-Suite enables you to collaborate with teammates on projects and quickly support stakeholders with their analyses. When a spreadsheet no longer meets your analytic needs, you can directly connect to and import data from an appropriately formatted sheet in a development environment of your choice using R or Python.

**Figure 1.7 Don't underestimate the value of a spreadsheet for easy analysis and sharing!**

## 311 Complaint Summary



### 1.2.2 Querying Language

The majority of organizations store data in *tabular format* (stored in rows and columns, like you would see in a spreadsheet) across multiple sources. When working at organizations that collect and store large amounts of data from a website, application, or business process, you will typically have access to a **data warehouse**. Data warehouses are large storage systems that synthesize the data an organization collects from multiple sources, allowing teams to curate the structure of data into the following objects:

- **Tables**, which are sets of logically organized tabular data. Each row in a table is a record, and each column contains information about that record.
- **Views**, which are curated results of a query containing logically organized information that may come from multiple underlying tables.

These are often constructed by data engineering or analytics engineering teams to better enable analysts to generate insights.

- **Schemas**, which are logically organized sets of tables.

In this situations, analysts are usually expected to draw from these data sources using an appropriate *querying language*. This is almost always a dialect of **SQL** (structured query language) [2] similar to those taught in classrooms, bootcamps, and online tutorials. Even if you are new to an organization and have never worked with their specific type of data warehouse, being familiar with SQL will give you the ability to quickly access, discover, and manipulate data in that warehouse. If the “dialect” of SQL differs from what you are used to, each data warehouse will typically have documentation on the functions that differ from one type to another (e.g., functions to manipulate dates often differ between warehouses).

Without a well-maintained data warehouse, a data analyst will still benefit from the knowledge and use of SQL. Manipulating data in programming languages such as R or Python involves the use of functions and methods with similar syntax to SQL statements. For example, the following is an example of how the total population by state is calculated in SQL, R, and Python, respectively (note that this code will not run, as we don’t have this table!):

```
SELECT      #A
    state,
    SUM(population) AS total_population
FROM city_populations
GROUP BY state
```

```
city_populations %>%      #B
  group_by(state) %>%
  summarise(total_population = sum(population))
```

```
city_populations.groupby('state')['population'].sum()      #C
```

The syntax is quite similar across each language—using a `sum`, and `group by` function after selecting each column. In addition, both R and Python have functions that allow you to use SQL queries to manipulate data if you are more comfortable with that syntax.

If you have little to no opportunity to interact with a data warehouse, it may be a good idea to proactively identify opportunities to incorporate SQL in your data manipulation workflows. For example, you can write SQL or Python scripts that process data for you, saving time and reducing errors. Eventually, you will likely be required to use these tools more actively in your career, and keeping this skill fresh in your mind will benefit your long-term growth and opportunities.

### 1.2.3 Statistical Computing/Programming Language

A **statistical software** or **modeling language** is essential for any analytics job where you expect to evaluate data using descriptive or inferential statistics. Like a data warehouse, the preferred software depends on the team and organization. SAS was a popular statistical software suite for decades and continues to be used in government agencies and some large corporations. Many smaller organizations, marketing agencies, and non-profits use SPSS for statistical analysis, especially when they primarily hire researchers and analysts with degrees in the social sciences (statistics courses in these programs frequently use SPSS).

If a team prefers proprietary software, it may still be beneficial to incorporate the use of a language such as R or Python into your workflow. In R, you can access, interact with, and save SPSS, SAS, and STATA files using the haven library or the upload tool available in RStudio user interface. All of the same can be accomplished in Python using the pandas library.

#### Using R

R is a popular programming language for statistical computing in data analytics, data science, and research space [3]. Its use compared to Python (discussed more below) varies by industry, team area of expertise, seniority level, and type of project. R tends to be more widely used in the biological sciences, social sciences, and statistics. If you work with an organization or academic institution in these areas, you may be more likely to encounter R as the technology of choice in your work or coursework.

If you're experienced in using spreadsheets or proprietary statistical software

such as SPSS, SAS, or STATA, the R community has a range of resources designed to ease the transition to your first programming language. [4][5] If you anticipate needing to develop explanatory statistical models as part of your work (see Chapter 8), R has easy-to-use native modeling capabilities and a wide ecosystem of packages for less commonly used statistical tests. It also has a well-structured collection of packages augmenting its base capabilities called the Tidyverse [6].

## **Using Python**

Python quickly became the most popular programming language in the data world and is one of the top languages of choice for developers in general [7]. It tends to be most popular among data science teams, especially those working with larger data sources and developing machine learning models as part of their workflow. Those with a math, engineering, or physics background may have been exposed to Python during their education.

If you expect your work as an analyst will grow to include predictive modeling (see Chapter 8) or are interested in developing a career in data science, machine learning, or data engineering, Python may be an ideal choice of language for your work. There is a wide range of tutorials, online courses, books, and other resources to support learning Python.

## **Choosing a Language**

There is a long-standing debate about the benefits of R or Python for data practitioners. As you grow your career, I recommend learning to read and interface with both languages to enable you to work with a broader range of stakeholders and peers in an organization.

If your team has a preferred language and an established set of code, resources, and best practices in that language, it's most effective to adopt and contribute to the existing framework.

### **1.2.4 Data Visualization Tool**

Your deliverables as an analyst will almost always include data visualization

to aid stakeholder interpretation of your work. A dedicated data visualization and dashboard creation tool will support your productivity as an analyst.

## **Static Visualizations**

Reports and presentations that include charts, graphs, and other visuals require, at minimum, the ability to generate static (non-interactive) visualizations using your spreadsheet tool or programming language of choice. A written or oral presentation usually needs visuals for stakeholders to interpret your work appropriately.

As with other elements of the data analyst toolkit, the choice of tool for creating data visualizations depends on the needs and practices of your stakeholders and team. If you expect teammates to collaborate with and interact with data in spreadsheets, using the charting capabilities in that tool will allow for the greatest level of interactivity and ease of ability to update with new data. If you're generating reports or deliverables using R or Python, both have robust libraries allowing you to create sophisticated visualizations.

## **Dynamic Visualizations and Dashboards**

Unless your deliverables are in the form of presentations or static reports, your work as an analyst will benefit from creating reproducible tools for your stakeholders. A dynamic and interactive dashboard is the most common reproducible tool that allows others to explore insights without you needing to refresh and update documents.

There are a range of open source and proprietary dashboard and business intelligence solutions available on the market. For the moment, I recommend you consider making use of a dashboard tool (e.g., Tableau, PowerBI) in instances where you expect your stakeholders will require any of the following:

- Reviewing the same analysis repeatedly over time with new data
- Drilling down into an analysis to view subsets or subgroups of data in customizable ways beyond what fits into a report or presentation
- Having predictable questions beyond what your team can support in an

ad-hoc capacity

## **1.2.5 Adding to your Toolkit**

Over time, augmenting your toolkit will enable you to continually increase the value of your work and reduce time spent repeating routine work. You can employ various strategies to incrementally save time and effort, freeing up the capacity for further improvement. Regardless of the size or seniority of your team, if you're a solo analyst at an organization, or if the overall investment in the data practice at your organization is low, we will discuss strategic investments you can recommend to the organization to elevate the visibility and value of your efforts. We will discuss the amazing range of available tools and strategies for investing in a data toolkit in chapter 11.

## **1.3 Preparing for Your Role**

While this book assumes you are in the early stages of your career, each chapter aims to prepare you to solve common problems successfully that an analyst faces in their work. The success of an analyst in solving each of these problems is highly dependent on access to mentorship, guidance, and skills not taught in common technical resources. The ability to choose the most appropriate statistical test, justify a hypothesis, or build a high-value dataset is as crucial as the ability to write performant SQL queries and efficient Python code. The former, however, is more challenging to prepare for and can slow down performance and career growth.

### **1.3.1 What to Expect as an Analyst**

An analyst's career can branch into numerous directions based on your interests, opportunities in an organization, and skillset (technical and non-technical skills). Knowing how to avoid common challenges and pitfalls will set the foundation for your career in analytics, data science, or other data practices and better enable you to excel as a professional.

### **Career Trajectories**

Entering an analytics career offers opportunities to grow and mature within the function and branch out into other adjacent practices. Some examples include:

- Data Science
- Data Engineering
- Research Science
- Technical Communication

Analytics in organizations have been around for decades, and their core functions will continue to exist as newer fields like data science mature and differentiate into specialized roles. Analytics is a valuable foundation for all data practitioners, as well as being an inherently valuable field in itself. It's an excellent skill set that can enable you to grow your career into another domain, develop your expertise, and increase your leadership capabilities in and outside the data world.

## **Demonstrating Value**

Take a look at the following scenario:

### **Managing Stakeholder Requests**

Clara is a data analyst at a startup in the education technology space. Her team of 3 analysts supports stakeholders in their marketing, fundraising, programs, and human resources decisions. They maintain a backlog and schedule of work deliverables and support ad-hoc requests from team leads and executives. These ad-hoc requests often have strict and limited turnaround times (2 business days or less). In the past year, the team is finding it more challenging to meet the deadlines of routine requests due to the increase in requests from the growing leadership team.

Clara's team lead has requested additional headcount to support the influx of requests. As part of the request process, the company has asked for a summary of the expected return on investment (ROI) and value for the business associated with the increased headcount. The executive team reviewing the request has responded with questions about why their requests

have a long turnaround and are causing disruption since they are considered relatively straightforward.

If the scenario is familiar, you're not alone. Being an analyst requires more than a formulaic approach to processing datasets and generating findings. It includes managing stakeholder expectations, strong communication about expected timelines and processes associated with fulfilling requests, and more. It's easy for stakeholders to fall into an *analytics fallacy*, where the simplicity of the deliverable (e.g., a summary statistic, table, or chart) is perceived as indicative of the level of simplicity in producing that deliverable. This can contribute to misalignment in communication, investment decisions in the data practice, and rapid turnaround times for deliverables.

Quantifying the return on investment (ROI) in data is not usually accomplished using straightforward calculations or metrics. It takes collaboration, qualitative insights, and a mature relationship with the people whose decisions you support. Throughout this book, we will review strategies for aligning with your organization on the value of an investment in analytics and minimizing miscommunications on deliverables.

### **1.3.2 What you will Learn in this Book**

Analytics coursework, books, and other curricula teach comprehensive *direct technical skills*, such as the programming languages, software, and statistical tests you will use daily in your role. You may have spent time practicing SQL for retrieving data from relational databases, Python or R for processing and evaluating the data, a business intelligence tool for visualizing the data, and more. These topics are well covered in a range of great resources that you can access in ways that best fit your learning style.

This book is intended to serve as a resource for excelling at the *soft and indirect technical skills* associated with building expertise in analytics. Skills such as developing strong communication styles with non-technical stakeholders, understanding the limitations of measurement, and effectively managing a project from stakeholder question to deliverable are taught on the job, by a mentor, or learned by trial and error in your work. A comprehensive

guide on these topics can save you months or years in your career progression with the accomplishments you can make and mistakes you can avoid. Managing these skills will enable you to be effective, regardless of the degree of support available in your current work environment.

## 1.4 Summary

- There is a wide range of analytics domains (e.g., marketing analytics, product analytics). Each has a standard set of workflows and deliverables and unique methods to solve problems within the function.
- Analysts typically use spreadsheet tools, querying languages, programming languages, and data visualization tools to complete their work and develop deliverables for stakeholders.
- Being successful as an analyst involves more than producing the output assigned to you; it involves strategic stakeholder communication and alignment to create value over time.

## 1.5 References

[1] “Definition of Data Warehouse - Gartner Information Technology Glossary,” *Gartner*. <http://www.gartner.com/en/information-technology/glossary/data-warehouse>

[2] B. Kelechava, “The SQL Standard - ISO/IEC 9075:2016 (ANSI X3.135),” *The ANSI Blog*, Oct. 05, 2018. <http://blog.ansi.org/2018/10/sql-standard-iso-iec-9075-2016-ansi-x3-135/>

[3] R Core Team, “R: The R Project for Statistical Computing,” *R-project.org*, 2022. <https://www.r-project.org/>

[4] D. Johnson, “Spreadsheet workflows in R,” *education.rstudio.com*, Aug. 17, 2020. <https://education.rstudio.com/blog/2020/08/spreadsheets-using-r/>

[5] J. L. & A. Horst, *R for Excel Users*. 2020. Accessed: Mar. 05, 2023. [Online]. Available: <https://jules32.github.io/r-for-excel-users/>

[6] “Tidyverse,” *www.tidyverse.org*. <https://www.tidyverse.org/>

[7] “Stack Overflow Developer Survey 2022,” *Stack Overflow*, 2022.  
<https://survey.stackoverflow.co/2022/#technology-most-popular-technologies>

# 2 From Question to Deliverable

## This chapter covers

- Preparing an end-to-end analytics project
- Setting expectations with stakeholders
- Managing the interpretation of results
- Identifying opportunities to create resources for reproducibility

All analytics projects begin with a question. From tracking organizational finances to understanding product users to test a marketing campaign, questions guide data analysis, statistical methods, and visualizations to communicate insights. The answer you provide to the question will ideally provide strategic information and direction to your stakeholders and their work.

Most analytics teams have a range of responsibilities beyond statistical analysis and presenting results. Analysts often consult on the appropriate interpretation and usage of findings, and guide stakeholders through asking valuable questions. Much of the success of a project depends on involvement in the entire *project lifecycle*, from the initial inquiry to the follow-up and recommended actions taken based on findings.

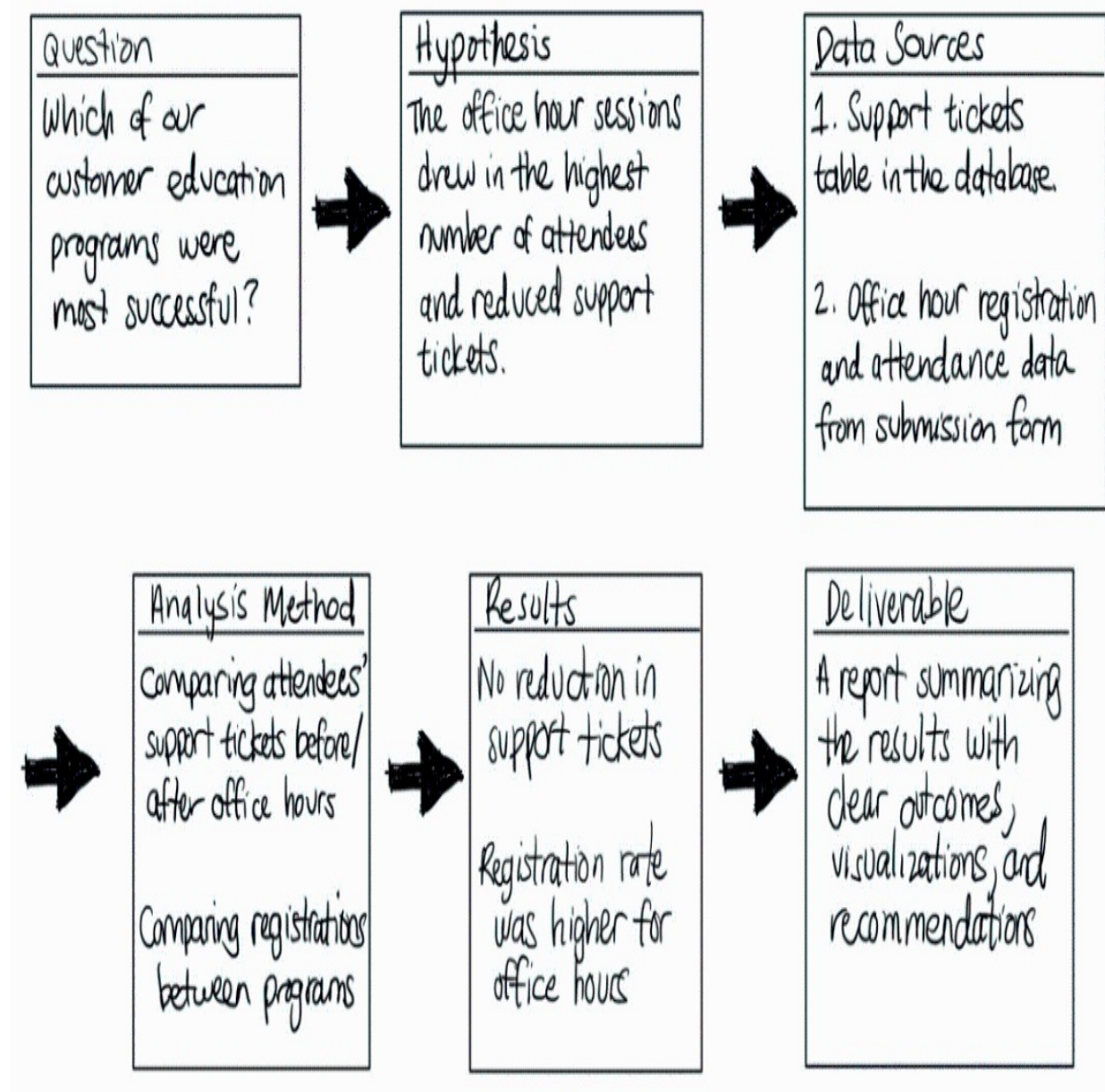
## 2.1 The Lifecycle of an Analytics Project

With a question in hand, your responsibility as an analyst is to distill the organizational process, research idea, or curiosity into something you can define, measure, and report on. While some routine questions and analyses have well-structured metrics and data sources, most novel questions your team addresses will not have an available data source, metric, or statistical analysis method to guide your approach.

Many businesses draw heavily from scientific methods in their approaches to deriving insights. The step-by-step process recommended here will give you

the appropriate tools to make confident decisions, align and clarify areas of ambiguity with your stakeholders, and make concrete recommendations. For each stage of the project lifecycle, you will be provided with a checklist to proactively identify the best path forward into the next stage.

**Figure 2.1 Flowchart of an analytics lifecycle looking at the impact of a customer education and outreach program. Each step distills the question into measurable items that can be analyzed and used to produce actionable recommendations.**



## 2.1.1 Questions and Hypotheses

### How do you measure that?

Questions you receive as an analyst are informed by your stakeholders' domain expertise, previous experience, and heuristics they're familiar with in their teams. These same heuristics rarely translate to a singular data source, metric, or method of operationalization in analysis and can lead to confusion around definitions. The first step of effectively working with a stakeholder is to agree on how to define their question.

### Operationalizing the question

**Operationalization** is the process of translating an abstract concept into a process you can measure. The term is commonly used in social sciences research methods and statistics courses to define the process of distilling complex behavioral and social phenomena. Operationalizing concepts is valuable as an analyst since the business and organizational processes you interact with are complex and typically involve a dimension of human behavior and processes that don't exist in a vacuum. Many behavioral and cognitive processes can't be measured directly, so additional steps and diligence are necessary to develop assessments agreed upon within an academic discipline.

Operationalizing a process involves aligning on precise definitions of the concepts in your stakeholders' questions. We'll demonstrate this with a hypothetical product analytics team throughout this chapter:

#### Operationalizing Customer Behavior

Jane is a Product Manager at a software company. She wants to learn whether customers found it easier to use the website's Help section after updates to their search functionality. Sam is a Product Analyst working with Jane's team and notices that *easier* is a heuristic that can have multiple definitions with the available data at their company:

Do customers spend less time on the Help section? Do they call the customer

support center less frequently? Do they respond positively to the feedback question that pops up on the Help Center screen?

Sam responded to Jane's request by proposing alternative, more specific definitions of her identified outcome – an *easier* customer experience. Sam's proposed definitions are not the only methods of defining *easier* in this context – dozens of possible measures likely indicate an easier customer experience. The questions Sam identified for the analysis in this context are based on the practical availability of data at the company. The list of questions being evaluated may expand or be revised as operationalizing an easy customer experience becomes better understood at the company.

In an academic setting, operationalizing behavioral and cognitive processes involves rigorous peer review, survey and measurement development, and psychometric testing to ensure the reliability and validity of the developed metrics. In a business or organizational setting, the rigorous peer review process is rarely feasible for analytics teams to apply in daily work. I recommend the following items to consider as you operationalize concepts in partnership with your stakeholders:

- What terms in the question are vague or could have multiple meanings?
- What specific, operationalized versions of the question are most straightforward or practical to measure with the available data at your organization (to be discussed in more detail in the next section)?
- What specific, operationalized versions of the question can have multiple viable competing definitions (e.g., are there arguments that both “more time” and “less time” can be considered desirable, positive outcomes)?
- Are there industry standard, peer-reviewed, or otherwise widely agreed upon measurable definitions of the concepts in the original question?

### **2.1.2 Data Sources**

The data you can access is often the primary driver of how questions are ultimately operationalized. Data is typically available to analytics teams in various *tabular* formats (e.g., CSVs, relational databases) as well as unstructured formats that require additional processing time. Each type of

source has its strengths and limitations in how you approach your work; we will discuss these in greater depth in chapter 10.

We'll discuss three types of data sources that Sam's product analytics team would use to answer their questions. These are great examples of data available at many organizations, but are by no means an exhaustive list. We will discuss *a great many* data sources in chapters 5 and 10 as well as the strengths and limitations of each.

## Clickstream and Behavioral Data

**Clickstream** and **behavioral data** are high-value sources of information for analytics teams whose organizations primarily operate in the digital world. *Clickstream data* refers to the record of a user's activity and navigation on a website or application. It allows you to track the sequence of *clicks* a user makes, and often includes the following information:

- The order in which the events happened within a single session on the website/application, allowing you to generate the full sequence of steps performed.
- The time spent on each page, which can be used to understand large-scale patterns usage patterns on your website/application.
- Individual elements or links clicked on a page, showing granular information on how users interact with your website/application.

Clickstream data is frequently used by product and marketing analytics teams in A/B testing, attribution modeling, website analytics, and more. It's often one of the *largest* datasets collected by an organization.

*Behavioral data* is a broader term that encompasses data related to a user or customer's actions, behaviors, and decisions. Clickstream data is a *type* of behavioral data, which can also include a wide variety of other data sources (e.g., purchases, email opens, survey responses). These data sources are integral to the purpose of product and marketing analytics: understand the behavior of users, customers, and prospects to create a better product or service.

Here is an example of a table that captures data on page views of the website:

**Figure 2.2** A sample of data from a table on page view events.

event_time	visit_id	session_id	user_id	page
2022-10-24 22:18:23.751292	11680601	1446638	915441	Settings
2022-10-24 23:22:23.751292	10088479	1088166	18840	What's New
2022-10-24 23:56:23.751292	11935535	1791443	736562	What's New
2022-10-25 01:57:23.751292	6617370	1571759	254287	What's New
2022-10-25 02:15:23.751292	3944598	1876849	182038	Help Center

The types of metadata in each column seen here are common to this type of data source, allowing you to understand which pages are most frequently visited, revisited and the length of time spent on the page. Many comprehensive sources of page view data will also enable you to track users' journeys across a website.

Taking a look at the characteristics of this table, we can see the following:

- A unique visitor ID (*user\_id*) allows Sam to track customer page views over time.
- A unique session ID (*session\_id*) for each time a customer visits the website, tracking all pages they visit during that time.
- The *Help Center* is available as a page, with no further detail on where a customer navigated (e.g., articles read, searches performed).
- The amount of time spent on the page is *unavailable*, limiting Sam's ability to assess trends or changes in time spent at the Help Center.

While sources of clickstream and behavioral data can be some of the most valuable for an organization to understand its users, it does have its limitations worth noting. This is typically collected using a third-party software, which can be expensive. It also requires manual effort by technical teams to proactively define the events they want to track (e.g., clicks on the “Add to Cart” button). Data can also be missing if users have ad-blockers installed, which can prevent browser tracking from third-party sources.

## **Customer Support Records**

Organizations with a Support team or function will often keep records of calls, chats, and other customer communications for analysis. Business Analytics teams will use this data to track metrics on the volume of communications, time to resolve customer issues and customer satisfaction as indicators of the team’s performance.

This type of data can also be used as a measure for analytics projects looking to impact the customer’s experience. In the case of our example, Jane’s product team was looking to implement changes that would improve the experience of customers and their ease of use of the website. If these changes are successful, it’s reasonable to develop a hypothesis around changes in the volume of communications the Support team receives.

Below is a sample of the dataset on chat support available at Sam’s organization:

**Figure 2.3 A sample of data from a table on chat support requests.**

user_id	chat_id	support_rep_id	chat_start	visited_help_ctr
915441	10550	139	2022-10-24 22:18:23.751292	True
18840	5180	451	2022-10-24 23:22:23.751292	True
736562	3270	357	2022-10-24 23:56:23.751292	False
254287	7431	58	2022-10-25 01:57:23.751292	True
182038	13558	258	2022-10-25 02:15:23.751292	True

The dataset contains the following characteristics we can consider in the analysis:

- The visitor ID (user\_id) is not in the same format as the user\_id field on the page views table, which means Sam may not be able to connect users between data sources.
- The column visited\_help\_ctr is a Boolean value (True/False) indicating whether or not the customer visited the Help Center before starting the chat support conversation. This may be a helpful *proxy* metric for Help Center visitor volume.
- The dataset contains columns allowing for multiple methods of measuring chat volume: (1) total chat requests, (2) chat requests per support rep, and (3) queue time.

Data from support calls, chats, and tickets are a great source of information on the areas in which your customers and users struggle with the product or service provided. However, the software and systems used to manage customer support can vary widely in the data they capture, how they're set up, and how easily they can be incorporated into a data warehouse.

## Survey Responses

The vast majority of organizations conduct surveys of their users, customers, population of interest, and employees. This data is collected from questionnaires or interviews, and can be administered in a variety of ways. Survey data is used to discover insights, opinions, perceptions, and self-reported behaviors on a set of topics flexibly defined by the organization.

Simple, one-question surveys are often capable of being built into websites and applications to gather large volumes of data on very specific topics. However, when those questions are asked using a third-party software, it may require additional work to collect, download, and connect that data to other information about your users and customers. For example, Sam's team discovered some potential survey data to use for their project:

### Leveraging Survey Questions

The product and marketing teams recently added a pop-up on the bottom of Help Center articles asking customers, *Did this article answer your question?* Customers can select “Yes” or “No” in response to the question. Unfortunately, Sam's team has discovered that this dataset is unavailable in their data warehouse for analysis. The team has access to the following information via a vendor's website:

- An aggregate *Customer Satisfaction* score is computed in a dashboard as a percentage of customers that responded “Yes” over time.
- No information about the users or which Help Center articles they visited is provided.
- No information is provided on the response rate to the question.

Sam's team has requested that the data engineering team ingest the data from the help center survey in the data warehouse. They anticipate the effort will take 4-6 weeks, at which point Sam will conduct a follow-up analysis.

While it's easily available and widely used, survey data does have significant limitations in terms of quality and accuracy. We'll discuss this in depth in

chapter 6, where we'll learn how to design effective measures (including surveys).

## Identifying Characteristics Of your Dataset

What does that timestamp mean? You may be surprised to learn that a specific piece data doesn't always mean what you think it does. Analysts will benefit from an initial effort to challenge their assumptions about the characteristics of a dataset *before* conducting an analysis in order to reduce the likelihood of inaccurate conclusions and misinterpretations of results.

As we've covered throughout the previous three sections, many data sources have limitations that are key determinants of their usability:

- Clickstream data can contain missing or inaccurate data from users with ad-blockers, making it difficult to use when precise measurements are necessary. For example, if you need a list of *every* user who visited a page, you may find that the list is incomplete.
- Data collected about business processes using third-party software (e.g., customer support data) can vary widely in its ease of access and available data about the support interaction.
- Data collected from surveys can be low in quality and accuracy when best practices in the science of measurement are not used.

As you build domain expertise at your organization, you will develop an understanding of the types of questions you need to ask to make sure you're able to ask the questions you want of your data. The following are examples of questions you can ask:

- Is the data raw, *event-level* information (e.g., one row per page view, click, call)? Or is it partially aggregated at some level (e.g., per user per day)?
- Are you able to connect different datasets to the same user or customer using a shared primary key?
- What timestamps are available on each row? Do you have the ability to capture time between actions, time to complete an action, or volume of actions over time?

- What percentage of rows have missing data in a field you're considering using for your analysis?
- What fields are *missing* from your dataset that might be valuable if you had them? How are your questions or hypotheses impacted by not having those fields available?

After answering these questions and determining the scope of analysis possible with available data, you can move on to the appropriate measurement and method selection.

#### **What's in the dataset?**

Sam's team has determined that some of the fields of interest are not available in the data sources he's been looking at. There is no measure of the time spent on a page, and the survey question on the support page asking whether an article was helpful is only available in aggregate form. Thus, the original research questions need to be revised.

With an understanding of the characteristics and limitations of each dataset, Sam's team can narrow down and revise the precise research questions agreed upon with Jane's product team:

- Do customers spend less time on the Help section?
- Do they call the customer support center less frequently?
- Do they respond positively to the feedback question on the Help Center screen?

The first two questions can be answered without the ability to join users between the two datasets. The third question can only be partially answered using the vendor's aggregate summary and is limited in its ability to understand the nuances of *why* a customer may respond "Yes," "No," or choose not to respond at all.

Sam has aligned with Jane on the data available to their team, which questions are possible to answer, and at what depth. Jane agrees that the analysis results will be valuable to the team, even with limited granular data.

## 2.1.3 Measures and Methods of Analysis

Research methods, study designs, and the statistical tests to evaluate them are essential tools in your analyst's toolkit. There are countless ways to ask and answer questions; study designs are intended to provide structure and guidance to your work. In this section, we'll primarily focus on the role, value, and procedures used in the process of designing a study or research project, diving into the data, choosing a statistical test, and preparing your results. We will cover the types of methods available in *lots* of depth in chapters 3, 4, 5, 6, and 8.

### Exploring Dataset Characteristics

Analysts will almost always engage in the crucial process of **exploratory data analysis** (EDA). This is a series of steps in which the main focus is on understanding the characteristics and patterns within the data, often beyond the hypotheses being investigated. Nearly every analytics project should include a detailed EDA to understand overall trends, patterns, and limitations of the data you are working with. It can help inform your

EDA often includes steps such as the following:

- **Exploring the shape of distributions:** when the data is plotted on a histogram or boxplot, what does the shape look like?
- **Investigating potential outliers:** are there records with extremely high or low values for a given measure? Is there a legitimate reason for those values? Should they be kept or removed from your dataset, and why?
- **Investigating missing and duplicate values:** what number or percentage of records have missing data? Is this missing for legitimate reasons, or is there an issue? How should these missing values be handled? Should the entire row be removed from the dataset, should an appropriate value be imputed into the row (e.g., the overall average), or should the missing value be left alone? Conversely, are there duplicate rows for information that's supposed to be unique?
- **Investigating summary statistics:** what is the mean, median, standard deviation, minimum, maximum, number of values, and number of unique values?

This is also *not* an exhaustive list; we will be discussing EDA throughout this book as it applies to different types of projects. We will also cover each of the statistical concepts and summary statistics mentioned in this section in chapters 4 and 5.

## Choosing Methods and Statistical Tests

If you take care to operationalize and structure your questions and hypotheses, choosing a research method and statistical tests will be a clear and easy decision. There are dozens of research and study designs available, though the majority of questions can be answered with only a handful of approaches. We will discuss this process in depth in chapter 3.

After selecting an approach, it's valuable to touch base with your stakeholders to set expectations on the type of information you will show them: will there be charts comparing average values between groups? Will you show a series of boxplots for each of the groups included in your statistical test? Setting expectations early saves you time in the development of your report; we will discuss strategies for tailoring the results of your methods and analysis to the data proficiency level of your stakeholders.

### Developing a Plan of Analysis

Sam communicates the data discovered and the proposed analysis plan to the rest of the team:

- Descriptive statistics showing (1) total Help Center visits and chat support requests over time, (2) average Help Center and chat support requests *per unique user* over time, and (3) average customer satisfaction scores over time.
- A *between-subjects study design* (discussed in chapter 3) comparing daily Help Center visits and chat support requests in the 90 days *before* the search functionality change was deployed and 90 days *after*. Additional follow-up comparisons will be completed 90 days after the first comparison.
- An *independent samples t-test* (discussed in chapter 4) to assess the

statistical significance of any differences in daily Help Center visits and chat support requests before and after the changes.

Sam receives positive feedback from the team on the analysis plan and recommends an additional *non-parametric* statistical test to add to the final step. With their support, Sam can begin preparing the data for analysis.

## Applying Best Practices

I recommend the following considerations when preparing your dataset for analysis and statistical comparison:

- **Budget time appropriately:** Running statistical tests usually takes the *least* time compared to every other step discussed in this chapter. You can expect planning, data preparation, EDA, and interpretation to require a far more significant time investment. Make sure to consider this when communicating expected deliverable deadlines to stakeholders.
- **Lead with your question:** The questions you ask should guide the methods and statistical tests you choose—*not* the other way around. If you try to fit a question into a specific type of statistical model, you risk confusing stakeholders and misinterpreting trends in your data.
- **Simpler is often better:** It can be tempting to start your analysis with complex statistical modeling to grow your skillset and derive better insights—I highly recommend you exercise caution with this! Start with a more straightforward test where possible, and look for examples using the same tests to ensure you’re using the right approach, that your data is in the right shape, and that you thoroughly understand the results.

### 2.1.4 Interpreting Results

Interpreting and distilling the results of statistical tests for stakeholder communication is the final component of an analytics project—and arguably the most essential step. Tailoring results communication for the intended audience is crucial to creating value with your work.

Sam has finished all steps in the analytical plan: the descriptive statistics have been calculated, the statistical tests show significant decreases in daily chat

support volume after the search functionality changes, and the team is excited to share their findings. What exactly should they communicate to Jane and the product team?

## **Assess the Statistical Knowledge of your Stakeholders**

When scoping a project, aligning with your stakeholders on their experience and understanding of basic statistical concepts is often valuable: Do they understand correlations? Statistical significance? Means comparisons (e.g., t-tests)? Each piece of information tells you how much detail into the statistical results to focus on in your final deliverable.

If you learned statistics and research methods in an academic setting, you likely learned to share your findings in a standardized *Results* section of a paper. These are often written as rote recitations of statistical test coefficients and values, making for easy reproducibility by other academic professionals. Unless you are preparing a publication for peer review, this format is *not* ideal for communication outside of academia and the classroom. Instead, I recommend aiding interpretation with clear summary statements and visuals.

Sam's final report includes a summary of findings with statements describing the methods used and the significance of the statistical tests for Jane's team. Sam confirmed with Jane that she's familiar with statistical significance and would find knowing the coefficient values returned by the statistical tests used helpful. Below is an excerpt from the results section of the report, which also includes bar graphs comparing the values for each measure in the 90 days before and after the search bar changes:

Daily Help Center page visits and daily volume of chat support requests were compared in the 90 days before and after deploying search functionality changes. The daily chat support requests decreased significantly ( $p < .01$ ) in the 90 days after changes were made. Daily Help Center visits did not change significantly ( $p = .42$ ) in this time period. Additionally, customer satisfaction scores increased by 5%.

The choice, application, and interpretation of statistical tests can frequently confuse your stakeholders (especially those outside technical roles). Statistics

education in most undergraduate and graduate curricula is pretty limited, and opportunities to advance data literacy in day-to-day work are highly asymmetrical across functions, domains, and organizations. While Jane understood the summary above, it's feasible that other stakeholders in the same organization will gloss over the details of the statistical tests and limitations if they are unfamiliar with their meaning. Choose your level of detail carefully!

## 2.1.5 Exercises

You are part of a Business Analytics team at a high-end fitness company. The marketing team has reached out to your team with a request for help answering a question: *What impact has the recent promotion at the gym (one month free for new members) had on the business?*

1. What *operational concepts* and *heuristics* are referenced in the above question? How might you translate the heuristics into measurable concepts?
2. Which datasets may be valuable to investigate when answering the stakeholder question? What columns or fields might you look for in each?
  - a. Customer gym check-ins
  - b. Customer payment records
  - c. Company payroll records
  - d. Customer experience survey feedback
3. What methods or statistical tests might you use to measure the impact of the promotion? (If you're unfamiliar with the appropriate choices, you can return to this question after reading chapters 3 and 4.)
4. The Director of Marketing has informed you that most team members are unfamiliar with statistics. How might you tailor your presentation to their experience?

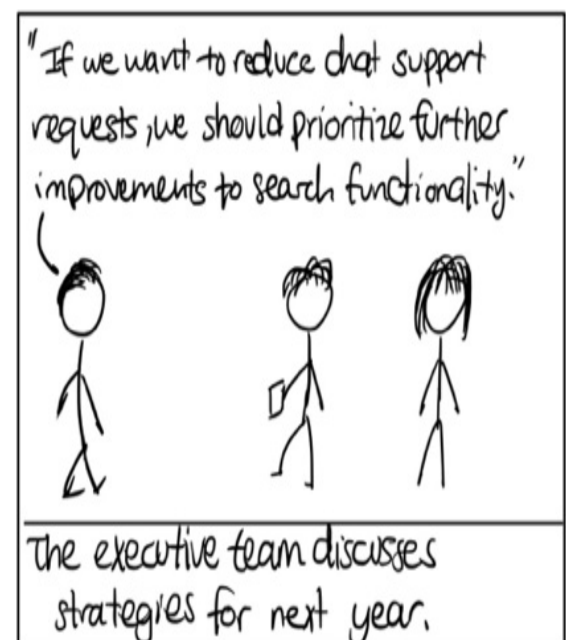
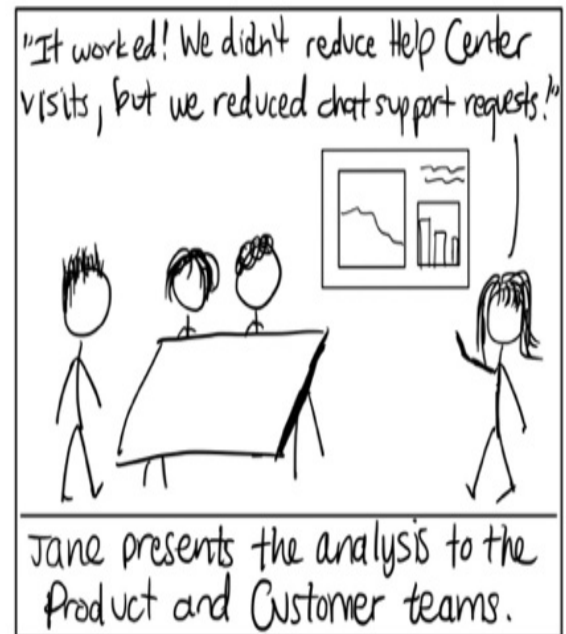
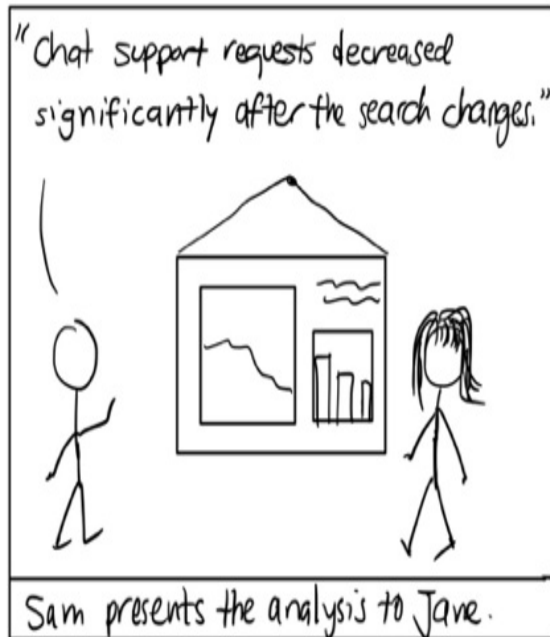
Use the checklists, recommendations, and example scenarios in each section as a guide for operationalizing each question, suggesting appropriate datasets and metrics, choosing statistical tests, and identifying appropriate levels of detail for the stakeholders in marketing.

## 2.2 Communicating with Stakeholders

You have aligned with stakeholders, operationalized their questions, identified appropriate data sources, performed an analysis, and written a well-structured report with proper visuals and summary statements aligned with the expertise of your stakeholders. Is your job done?

Not quite – an analyst's role includes creating resources to aid stakeholder interpretations and next steps and communicating results to *their* stakeholders to ensure all parties receive the appropriate message about your work. Let's take a look at this scenario:

**Figure 2.4 Analytics Telephone can diffuse the quality of your insights.**



*Analytics telephone* is a situation that occurs when results from peer-reviewed articles, internal analyses, statistical modeling, or any other type of synthesis of quantitative or qualitative information are distilled and summarized from one source to another, ultimately losing their meaning.

This is commonly seen in the communication of scientific findings in news media. For example, an article is published detailing a study showing a positive association between a behavior and a health outcome in a small sample of adults. A press release summary is produced, excluding the details about the sample and limits of the association. A local news channel reviews the press release and reports that engaging in the behavior causes the health outcome without mentioning the limitations. The public then assumes that engaging in the behavior will cause the health outcome, and when it does not, it can ultimately lead to distrust in future findings presented to them.

As an analyst, it's valuable to be mindful that the findings and results from your project can generate excitement among your teammates, who are eager to read and share with others! In that process, it's easy for your findings to be diluted into colloquial language that lacks the precise wording used by professionals in the data world. Analysts should be cautious when telling stakeholders their findings indicate *causation* or *proof*. It's challenging to walk back from those claims at a later date if new data surfaces with contradictory findings, and this can result in a lack of trust in the analytics function of the organization.

### **2.2.1 Guiding the Interpretation of Results**

Language and words matter in analytics. The terms used to describe findings carry tremendous weight in how consumers of your work interpret—or misinterpret—the results. Words such as *cause*, *prove*, *associate*, *predict*, *suggest*, *difference*, and *findings* may have distinct definitions in research and analytics, but they are often conflated in conversational speech.

#### **The Scope of Interpretation**

Analysts benefit from the strategic communication of two concepts in their findings:

- *The scope of interpretation* is the acceptable degree to which your results can be generalized beyond the specific findings communicated. This includes generalizability to a broader population beyond what was included in your work and the interpretation of null or alternative

hypotheses (e.g., does a non-significant result mean there is no relationship between two variables?).

- *Precision of language* is the responsible and intentional use of keywords that aid stakeholders in interpreting correlation, causation, statistical significance, and other concepts. This strategy minimizes conflation with colloquial terminology and provides a roadmap to your stakeholders on the appropriate interpretation of your research design, methodology, and findings.

Careful consideration of these concepts in your deliverables will guide stakeholders in interpretations you *can* make and those you *cannot*, given the work completed. It can also guide appropriate follow-up questions and subsequent research steps to continue a strong partnership with your team.

Over time, both strategies will be valuable for managing expectations, taking effective action, and building a data-informed and data-literate culture within your organization.

**Figure 2.5 A slide such as this example that delineates the scope of interpretation helps ensure the long-term success of your work.**

## Summary

- Chat support requests decreased in the 90 days after the Help Center changes.
  - This indicates initial evidence of the impact of these changes.
  - We recommend re-evaluating this measure after an additional 90 days to assess the scope of the impact.
- Daily Help Center visits did not change in the 90 days after the changes.
  - This likely indicates that Help Center visits were *not* the correct behavior to measure for this outcome.
- While customer satisfaction scores increased slightly in the same time period, we do not recommend using this metric as an indication of success.

---

Let's return to Sam and Jane's customer Help Center visits and chat volume analysis. The scenario in Figure 2.5 above shows what's possible when *analytics telephone* occurs—a situation where your results are shared between teams and the eventual conclusions drawn are beyond the scope of your work. In the course of Sam's colleagues sharing the team's results, the scope of interpretation became diluted until the executive team was making broad inferences about the value of future efforts to optimize search functionality on the Help Center and suggesting that those changes would *cause* further reductions in chat support volume. Suppose these inferences become recommendations for additional optimization work before due diligence on this inference is completed. In that case, multiple teams' time and effort can be dedicated to work whose justification is based on a faulty premise.

At the recommendation of the team, Sam made the following changes to the report and presentation:

- Added the following details on the methodology:
  - The number of users visiting the Help Center and contacting chat support in the 90 days before/after the search changes
  - The number of users doing each of the above in the last year as a benchmark
- Added the following details on the results:
  - Updated the language on the lack of association between search functionality changes and Help Center visits to indicate that no relationship was *detected* and there was insufficient information to show whether there was an impact on Help Center experience.
- Added clear hypotheses for each of the operationalized questions, indicating the expected *association* between search functionality changes and outcomes rather than an expected *causal* relationship.

## Enumerate Limitations and Next Steps

Mitigating the likelihood of analytics telephone scenarios is done with a few intentional steps and information communicated as part of your lifecycle report. I recommend considering the following steps, especially when you expect your results will be shared with a wide audience.

- **Include a limitations section in your report or presentation:** this is a standard section in peer-reviewed papers that is valuable to communicate in your reports as a slide or page for stakeholders to read. Include a list of bullet points of data unavailable for in-depth analysis, the scope of interpretation, and any interpretations you *cannot* make with your findings.
- **Include a section with suggestions for further research:** this is also a standard section in peer-reviewed papers, helping provide a strategic lens into future research in a topic area. Enumerating recommendations for further research and evaluation steps is an easy way to provide a roadmap for stakeholders looking for the strategic investment of time and resources.
- **Create a guide to statistical interpretation to share with stakeholders:** if you don't already have this as a resource, find or develop an appropriate guide to understanding correlation, causation, statistical significance, and generalizing findings. We will discuss

creating this resource in depth in Chapter 11.

Sam's team was informed of the executive team's discussion about recommending continued work on optimizing the search functionality of the Help Center. In line with the above steps, Sam's team can augment the report and presentation initially delivered to Jane's team with some simple information that clarifies the project and its scope.

A slide incorporating the first two bullet points can be added to the presentation:

**Figure 2.6 Addendum slide to Sam's original presentation.**

## Limitations

- The full dataset for customer satisfaction scores was not available.
- Chat support request volume should be re-evaluated at +90 and +180 days to assess continued reduction.
- Chat support reduction was assessed after this specific search change. Hypotheses on further search optimization should be evaluated in follow-up tests before generalizing.
- The original goal of reducing Help Center page views should be examined in greater depth before drawing conclusions.

This addendum provides direct clarification to the leadership teams planning strategic efforts. In response to this information, the executive team recommends further research into the efficacy of two modular changes to the search functionality before proposing a much more extensive overhaul to the

feature. Thus, the additional information provided Sam's team with two new clear deliverables that add information to the decisions made by the product and customer teams.

## **2.2.2 Results that Don't Support Hypotheses**

As an analyst, you will *regularly and frequently* discover findings that do not support your hypotheses or those of your stakeholders. This happens to every analyst and is *not* an inherent reflection of your capability of working with your stakeholders. If you went for long periods in your career without findings that contradict hypotheses, I *would* be concerned with the accuracy of your results and methodological approaches to your work. I repeat: this is a part of the job.

Hypotheses aren't developed in a vacuum; they're usually tied to strongly held beliefs, domain knowledge, and heuristics about an organizational process or behavior. You can expect to experience resistance, skepticism, or pushback on these findings at multiple points throughout your career. Nonetheless, the frequency with which findings don't support hypotheses does not make it easy to communicate this to stakeholders.

### **Findings Misaligned with Hypotheses**

Even when a hypothesis is not enumerated as part of the analytics lifecycle, stakeholders will frequently have expectations about the analysis outcome based on preconceived notions of patterns or behaviors in the domain area. They may plan to take actions aligned with one or more of these expectations, and results that don't match those findings can create frustration and delay work if started ahead of the analysis.

When this misalignment occurs frequently, it often indicates a more significant culture shift necessary within an organization to derive value from quantitative insights. However, even in organizations with a mature approach to analysis, this can *still* happen. High-quality research takes time and questions free of bias, and not everyone you work with will have the time or ability to approach your work the way you expect.

I recommend handling each of the following misalignments in structured ways:

- **Results that *oppose* the hypothesis:** When you have statistically significant results that directly contradict stakeholders' hypotheses, it's beneficial to discuss the contextual background that informed the hypothesis in the first place. What guided them to develop the hypothesis in the first place? Can you break down the hypothesis into granular behaviors that can be measured and examined in more depth?
- **Results that show no significant relationship:** The lack of statistically significant results can be interpreted by stakeholders as the absence of a relationship or conflated with an *opposite* relationship. These results can also be interpreted as their work being ineffective toward achieving a desired outcome. In this case, discussing the behaviors and outcomes chosen for comparison in detail is beneficial. Were they the proper measures to assess the behavior or outcome of interest? Are there more appropriate measures that can be considered for future analyses?
- **Non-significant differences *supporting* the hypothesis:** Statistical significance can be challenging to explain to stakeholders unfamiliar with the concept and its application. When reporting this type of result to stakeholders, it can be valuable to communicate those initial findings are promising and that additional time, users, or data is necessary to report on the findings confidently.

## **Findings Misaligned with Communal Knowledge**

Over time, organizations build up collective knowledge from various sources: customer interviews, free-text surveys, competitive research, peer-reviewed articles, product feedback, and more. As time goes by, this knowledge guides the strategy and direction of the organization. However, that knowledge can become outdated and misaligned with current customers or stakeholders.

There are two common types of misalignments in this area:

- **Quantitative findings contradict qualitative findings:** organizations commonly augment quantitative findings with qualitative data such as free-text survey comments, product feedback, customer interviews, and

focus groups. Many smaller organizations with fewer customers, clients, or external stakeholders will lean heavily on the latter instead of investing in an analytics function.

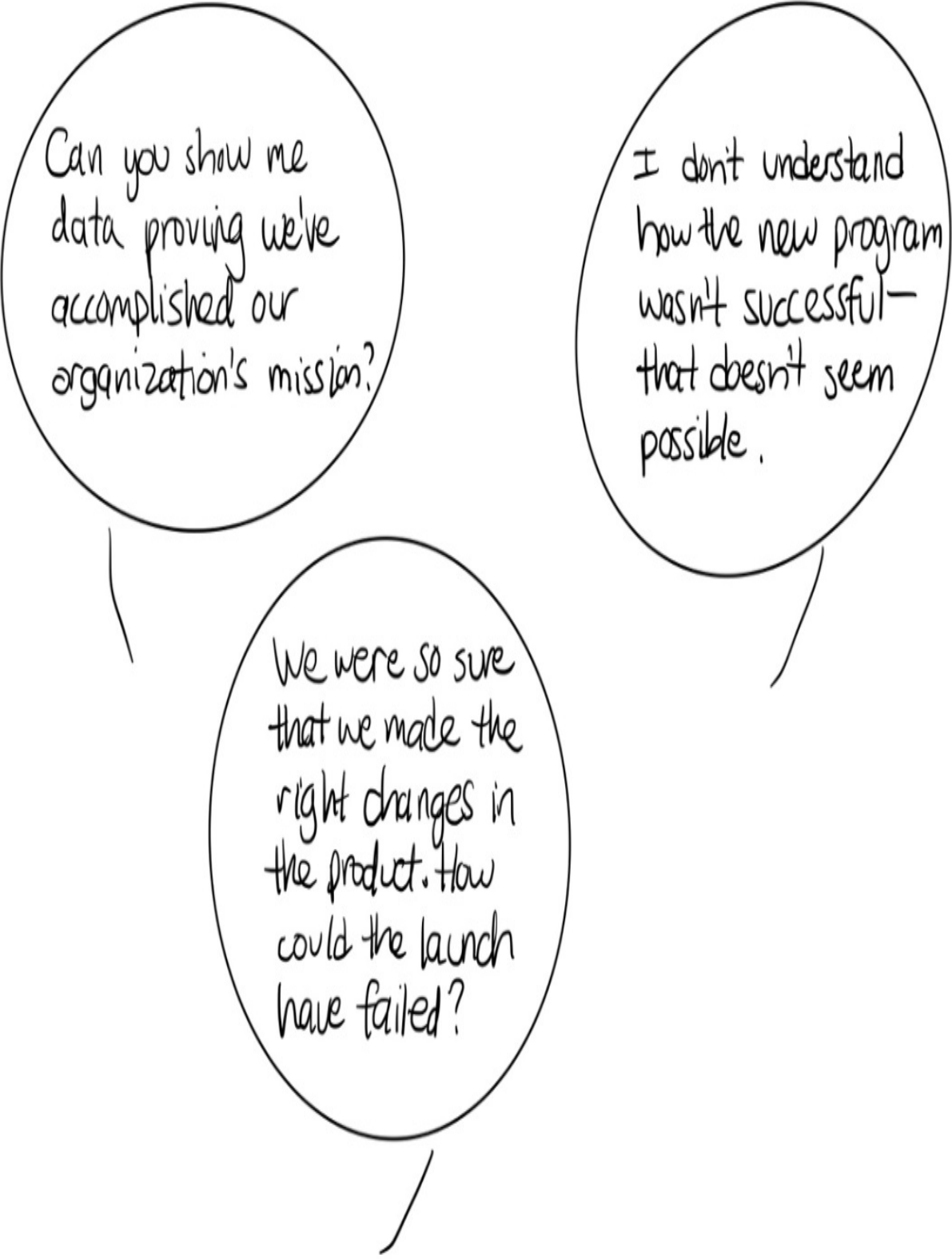
**Figure 2.7 Example quotes referencing qualitative findings.**



When recent quantitative findings do not align with qualitative findings, I recommend the following:

- Check for overlap in users or customers between quantitative and qualitative findings. You will likely find that qualitative insights were generated using a small subset of highly-engaged people not representative of the broader base of users.
- Check the recency of references to qualitative insights compared to quantitative insights. In most cases, it's easier to refresh data from a statistical analysis than to redo a set of interviews, focus groups, or another qualitative method. Raise questions about the applicability of findings that may be outdated and not representative of the current base of users.
- **Quantitative findings contradict common organizational beliefs:** good ideas and rapid feedback loops help small organizations get off the ground and scale rapidly. The information gained in the early days of developing a product or service is necessary to understand the need being met and the likelihood of success. These early insights become foundational to the mission and goals of the organization. They can be difficult to challenge as the organization matures, and quantitative insights demonstrate a different picture from what was previously believed.

**Figure 2.8 Example quotes referencing common organizational beliefs.**



Can you show me data proving we've accomplished our organization's mission?

I don't understand how the new program wasn't successful—that doesn't seem possible.

We were so sure that we made the right changes in the product. How could the launch have failed?

When quantitative findings do not align with strongly held beliefs, I

recommend the following items as a long-term strategy for your team. This is *not* quickly resolved within the scope of a single project.

- Work with core stakeholders to understand the source of the organizational belief – was there early research done to inform these beliefs? Can the research be updated with a larger, currently-representative group of customers? Can you develop a strategy to highlight where there are gaps in the sources of information being used?
- Scope out a project roadmap to understand how and where the user base has grown and changed. This will help mature commonly held beliefs at the organization and demonstrate where customers' profiles, needs, or behaviors have changed from previous years.

Let's return to Sam's analytics team. During the presentation of the proposed project follow-up plan, a marketing team leader expressed concern with the product team's goal to reduce chat support volume.

*“I thought customers who engage with chat support are less likely to cancel their subscriptions with us. Why would we want to reduce it?”*

The marketing team leader referenced an analysis performed several years prior, showing that customers who contacted the support team at least once were less likely to cancel subscriptions after one year. Sam's team shared an updated version of the analysis to answer the question posed during the presentation, showing that the conclusions from 5 years ago were no longer accurate for the current customer base.

## **Final Note on Results**

Similar to the results you evaluate, aligning with stakeholders is an expected part of the job of an analyst. The data literacy of your stakeholders will vary widely based on their domain expertise, previous experience, and the expectations of your current organization. We don't yet have widespread data literacy or competency education in schools, nor is it necessary to be effective in many roles.

It's generally helpful for your team to understand the degree of data *accessibility* of each stakeholder you work with to tailor messages to them and their teams. The comprehensive knowledge of stakeholder needs will allow you to tailor resources to their level and enable increased comprehension of your work over time.

Until a comprehensive data literacy curriculum is part of an early education curriculum, the role of an analyst will include communication and *data translation* at multiple levels. We will continue incorporating communication strategies throughout this book to build your confidence in this fundamental skill.

### 2.2.3 Exercises

Let's return to your analysis plan for the Business Analytics team of the high-end fitness company. You have developed a report and presentation for the marketing team detailing the impact on the number of new paying customers, gym check-ins, and customer satisfaction.

1. The number of new paying customers increased significantly in the 30 days after the promotion launched compared to the previous month. Write a 1-2 sentence summary detailing these findings, guiding stakeholders through the interpretation.
2. The marketing team is excited to hear the results you shared and recommends a strategy of providing one month free to all new members at the next executive team meeting. Can you identify or explore any limitations with this strategy with the available data?
3. The customer satisfaction score decreased slightly in the 30 days after the promotion launched. Does the promotion cause this? How can you communicate the *scope of interpretation* for this finding?
4. The number of check-ins at the gym did not change in the 30 days before and after the promotion launched. A marketing team member informs you that they had previously learned from new members that they tend to check in at much higher rates in the first 90 days of their gym membership. How can you reconcile your findings with the qualitative results cited?

As with the previous exercises, it's important to note that there is no *single* correct answer to each question. It's valuable to document and be prepared to explain your rationale for any interpretation of the results.

## 2.3 Reproducibility

The technical steps of an analytics project are designed to be repeated: identifying, retrieving, cleaning, processing, and analyzing data should ideally be possible for others to *reproduce* using the report you publish as a source of *documentation* on the steps taken. Additionally, many of these steps are repetitive should you wish to redo or duplicate the analysis later, making an eye toward *reproducibility* beneficial to your peers and a significant time-saver for you.

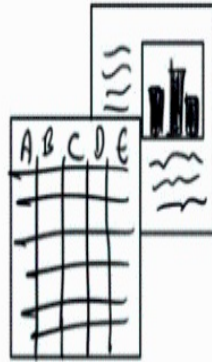
**Reproducibility** is the capacity for a scientific study, analysis, or project to be replicated by your peers. A project is considered *reproducible* if the steps to recreate the methods, datasets, measures, and statistical tests are documented with the necessary detail for others to understand the steps taken and redo the same project. In academic sciences, *reproducibility* is usually an essential condition for the publication of findings in peer-reviewed journals. Outside of work in a research institution, providing sufficient documentation for an analytics project is highly dependent on the tools available within the organization and the capacity of the team to detail their work.

**Figure 2.9 Recommended steps involved in reproducing a project.**



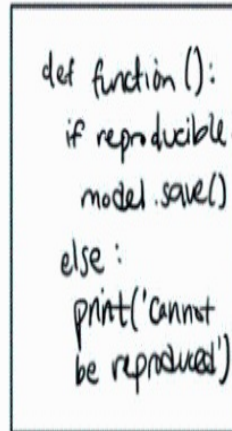
1

Document the research that informed your ideas.



2

Document, save, and share your data.



3

Add your code to a repository. Where possible, use accessible open source tools.



4

Share your results and how you reached your conclusions.

Regardless of the implementation of *reproducible practices* in the broader analytics team, it's beneficial to keep this detailed documentation for your work wherever possible. Ensuring that others can re-test your findings with new or similar datasets or *build upon* your findings and create additional insights based on your work can have widespread benefits on the data-driven capacity of your team and organization.

### 2.3.1 Documenting Work

In analytics projects, the *documentation* of your work is defined as a record

of relevant detail so that it can be *reproduced* or augmented by your peers at a later date. These records are rarely surfaced in detail as part of your stakeholder deliverables. Still, they are crucial should your stakeholders request detailed follow-ups or want to dig deeper into your work.

Depending on available tools and software, analytics teams may keep a separate internal record system or have reporting software that enables more granular view and edit permissions for the underlying queries and code. Regardless of the system or level of diligence of your specific team, there are some goals worth striving for in your work that will improve its accuracy and save you time in the long run:

- If you leave the team or organization for another role, the rest of the team can understand the steps you took in a project and replicate them.
- At a high level, your stakeholders can understand the purpose of the steps you took in your projects.
- You can redo a project without having to rewrite your queries or code.
- You can draw from previous projects' queries and code where applicable to save time and improve consistency across your work.
- You are able to revisit a project two years later and understand the rationale and context for the work based on the record kept.

With these principles in mind, let's discuss strategies for keeping high-quality records for each step in your analytics with minimal additional effort.

## **Questions and Hypotheses**

The questions and hypotheses developed as part of a project are presented in most final deliverables. Beyond restating them for your audience, it's valuable to share a summary of the rationale and context and any background research motivating the project. Questions asked of you by stakeholders do not exist in a vacuum, nor are they developed at random. Enumerating the sources of information that guided the question helps get everyone on the same page. It allows others to build the knowledge base you and your stakeholders have developed as part of the project.

If you're familiar with the format of peer-reviewed papers, you know that a

lengthy introduction section on all relevant background research precedes the statement of hypotheses and methods. Ideally, the reader is guided to the question and hypothesis based on the information provided.

A complete detailed introduction is rarely necessary outside of academia. Instead, the principles of structuring this section can be applied to the documentation you create for your questions and hypotheses. The detail included in each of the following can be tailored to the deliverable (e.g., a report or presentation).

- Begin the section with the most *general* background information. This may include the organization's motivation for solving a problem, a component of a company's values, or other broad goals.
- Summarize any research conducted or previous work informing decisions on the organizational problem over time. Each example discussed should be increasingly narrower in scope, bringing increased focus to the current questions and problems the team is attempting to solve.
- By the end of the introductory section, it should be relatively clear to your audience why the question is being asked and why resources are dedicated to answering it instead of other questions.

Structuring your introduction or background section in this manner will proactively answer many questions you can expect to receive from readers who aren't familiar with the conversations, meetings, and organizational history motivating a project. It helps new team members familiarize themselves with the knowledge base built by more tenured people within the organization. It enables you to reach a resolution when findings are misaligned with collective knowledge.

Sam's team writes a *Background* section for the detailed report deliverable:

### **Background**

Creating a seamless customer experience is one of the company's core values. For years, we have sought to provide insights into the customer experience and understand what behaviors indicate a positive experience or may instead indicate friction when using the product.

Visits to the Help Center, conversations with our Support Team, and customer satisfaction are key progress indicators measured across the company. Our research shows that high customer satisfaction consistently predicts customers renewing their subscriptions. Previous research (4 years ago) had identified increased visits to the Help Center and outreach to the Support Team as predictors of customer renewal. However, these trends have shifted with the growth of our customer base. The more customers visit the Help Center and contacts Support, the *less* likely they are to renew their subscription.

In the past two years, we've seen a substantial increase in the percentage of customers who visit the Help Center and contact Support. This has placed a strain on the Support staff and created concern, as our renewal rates have decreased in that period. We're also aware that in any month, at least 50% of customers contacting Support had first visited the Help Center and could not find the resources they needed. To that end, the Product Team aims to improve the experience and functionality of the Help Center to reduce the volume of requests to Support and mitigate customer cancellation risk.

Each statistic referenced in the Background section above includes a link to another report or a reference for further information. The full report containing the summary is referenced in the appendix of the stakeholder slideshow presentation and was shared across the organization as a full record of the work completed. An abbreviated *Background* section was developed for the slideshow presentation:

**Figure 2.10 Background slide in Sam's presentation summarizing information in the long-form report.**

## Background

- Recent research shows that higher frequency of visiting the Help Center and contacting Support is associated with decreased likelihood of customer subscription renewal.
- The percent of customers visiting the Help Center and contacting Support has been increasing for 2 years.
- Over 50% of customers contacting support first visited the Help Center and were unable to find the resources they needed.

---

As you add relevant background information to your deliverables, ask stakeholders for feedback on the value of the *level of detail* and *type* of the information supplied. Iterating on your approach to background information will increase the value you produce for the organization over time.

## Data Sources

Keeping a record of data sources used, and the methods for retrieving them is crucial to reproducible work. This record is the *most* important to keep in developing your work. A complete history of all queries run to retrieve data from your database, datasets from third parties, or other information is necessary to rerun your analysis later and debug or correct any issues you identify as part of your work.

Imagine a stakeholder identifies an error in your summary results, but you do not recall the exact steps you took to generate that summary in the first place!

I recommend avoiding this scenario by keeping a *comprehensive history* of how you retrieved, shaped, and processed your data.

The record of data sources is usually quite simple compared to writing a background summary and is easily done as you perform the analysis initially. I recommend the following steps in this process:

- Keep a record of all queries used in the final analysis. Depending on your organization's reporting or business intelligence software, this may be a built-in capability that requires no additional work.
- Keep an additional record of queries in exploratory steps that you chose *not* to include in the final report. These are useful if your stakeholders ask why you chose or chose not to take your work in a specific direction.
- Keep a record of all links to third-party data (e.g., a dataset from a government database) and code used to retrieve that data.
- Add comments to your queries and code to document their specific purpose and how to use them.
- When communicating with stakeholders in the final report and presentation, share the high-level data sources rather than the specific queries and code. Link to them or provide information on how to access them in your report.

Retaining a well-documented record of queries, code (including appropriate docstrings for your code), and data sources ensures you can edit or update your findings at a later date. This record can also be a starting place for future analyses, metrics, or data warehouse tables, saving your team bandwidth over time.

Sam's team uses reporting software that allows readers to view the report's underlying queries powering charts and summaries. Thus, a single summary is written for both the report and slideshow:

#### **Datasets**

The following data sources were used in this analysis: (1) Page View events, which include visits to the Help Center, (2) Support ticket data, which includes chat support requests, and (3) aggregate customer satisfaction scores available in our *Customer Experience Pro* account. Average scores were

compared between July to September and October to December, representing 90 days before and after changes were made.

In addition to the summary, Sam included comments in each query underlying the report indicating its purpose, the date range for which data is intended to be retrieved, and a brief rationale for any records filtered out of the final dataset. Since all datasets will likely be included in follow-up projects making changes to the Help Center, evaluating those changes will take a fraction of the time.

## Measures and Methods

Nearly all consumers of your work will require sufficient context to understand how you choose to summarize and present quantitative results. This includes aggregating, tracking, and summarizing data, sharing the statistical techniques used, and the steps taken to evaluate your results. Though we use summary tables, charts, and graphs to aid in the visual interpretation of the data, it's not always immediately apparent *why* you choose to measure and display something in a specific manner.

Documenting methods involves creating a section on the methodological steps you take and relevant context interspersed throughout the presentation of results. This information is geared toward answering *why* you chose the steps you did to make sense of the data. This documentation strategy includes the following:

- Write a *Methods* section in all forms of deliverables (reports, presentations, etc.). This should include a list of **exploratory data analysis** steps taken to produce charts and summary tables, a list of statistical tests used, and any special considerations in how the data was evaluated.
- Include clear titles, labels, and brief descriptions of all charts and summary tables in your deliverables.
- Include a brief explanation of why data is summarized in a specific way – especially when it differs from methods stakeholders are used to viewing (e.g., weekly totals instead of monthly).

Sam's team included the following summary in their long-form report, which is rewritten and abbreviated in a bullet-point format for the presentation:

### **Methods**

Descriptive statistics were shown for the 90 days before and 90 days after the changes to the Help Center. Each measure where granular data was available (Help Center page views, chat support requests) had a daily total calculated, and a mean/median was calculated based on those totals. This aligns with the team's established daily volume metric for both measures. The overall average customer satisfaction score for the 90 days before/after the change was included. No other customer satisfaction aggregations were shown for this analysis due to the lack of availability of granular data.

A repeated measures t-test was used to compare the mean (average) daily volume of Help Center views and chat support requests in the 90 days before and after the changes were made. The results were evaluated with a 95% confidence interval.

In addition, Sam included the following notes in the presentation, where aggregate information on each metric was shown in a chart:

- A description of why the weekly volume was shown in the chart instead of daily volume (aggregating by week accounted for a drop in page views over the weekend, making it easier to see the increase over time).
- A reminder that the bar graph showing before/after customer satisfaction scores did not include row-level granular data.

Each strategy supports Sam's team in being proactive about expected questions from stakeholders and consumers of the report. This documentation saves time and effort for the team, builds stronger relationships with stakeholders, and aids stakeholders in developing a skillset in analytical methods.

### **2.3.2 Exercises**

Now that we have a comprehensive example of the documentation included in a report for reproducibility let's add relevant documentation to your

analysis plan for the Business Analytics team of the high-end fitness company.

1. The company has been routinely running new member promotions and incentives for joining since it opened its first location 7 years ago. It's seen a 20-fold increase in business since then and has 18 locations across three cities. Write a *Background* summary for your stakeholder presentation slideshow that highlights relevant information motivating the analysis of the current promotion.
2. Provide a summary slide of the data sources used to analyze the promotion. In this summary, share the datasets that were *not* included and why.
3. Each data source is available in the company's data warehouse with the granular detail of each record. What type of documentation might you include about the queries used to retrieve that data?
4. Provide a summary slide of the methods used to analyze the data.

Keep returning to this example project as we review specific technical skills throughout this book. You'll have an opportunity to review and evaluate the appropriate level of detail for different stakeholders with each new topic discussed.

## 2.4 Summary

- The lifecycle of an analytics project starts with a **question**. Previous knowledge motivating the question guides the development of hypotheses, which informs the datasets and methods used to evaluate the question.
- **Operationalization** involves translating a concept into something *measurable*. This usually involves working with your stakeholders to take a heuristic (e.g., customer satisfaction) and creating a technical definition that can be directly assessed using data.
- Organizations have many different **data sources** that can be leveraged for analysis. Some examples include clickstream data (tracking clicks, page views, and other individual events on a website or application), customer support data, and survey data. Each data source has strengths and limitations that you will need to consider when determining if the

information you have can be used to answer your question.

- **The scope of interpretation** is the degree to which you can reasonably generalize your findings beyond the people included in your analysis. Your assessment of the appropriate scope of your findings will help guide your strategic recommendations to stakeholders.
- **Communicating results** to stakeholders involves tailoring final deliverables to their understanding of analytics, statistics, and previous information about the topic. There are many areas where you can expect follow-up questions and strategies you can apply for responding to common types of follow-ups.
- Documenting your background research, context, datasets, measures, and methods ensures that your work is appropriately **reproducible**, saves time, improves accuracy, and optimizes your team's ability to take on new projects.
- Managing the above effectively requires developing an improved understanding of your stakeholders and their needs over time. However, you can apply many great strategies in your work today to better set you up for success.

# 3 Testing and Evaluating Hypotheses

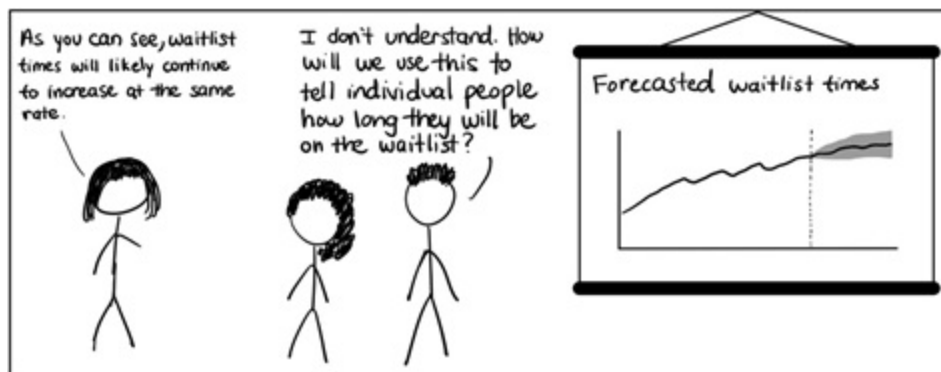
## This chapter covers

- Conducting appropriate research to inform your hypothesis
- Choosing and implementing methods for gathering information
- Choosing and implementing a research design for your analysis
- Using testing and evaluation methods in different research programs

Now that you know what to expect at each step of an end-to-end analytics project, let's zoom in on the process of operationalizing questions, developing hypotheses, and choosing an appropriate method for evaluating your hypothesis.

This may seem like *a lot* of focus for what's ultimately a few sentences in your final deliverable, but the skills you learn here will support you in synthesizing results and presenting the right analytic approach for your final deliverable.

**Figure 3.1** Many frustrating situations can be avoided by aligning on the question, hypothesis, and expected deliverable, as has happened many times to myself and most analysts I've worked with.



The overarching topics in this chapter are usually covered in undergraduate

and graduate courses in the sciences, with titles such as *Research Methods* or *Experimental Design*. The methods we will cover primarily draw from these curricula; however, we will take a less traditional approach to cover each topic by focusing on *probabilistic* methods of thinking about our hypotheses and how to evaluate them. As in previous chapters, we will focus on a range of applied examples in a business or organization outside of those typically covered in academic coursework.

At this point, you may wonder why this book contains lengthy instructions for what amounts to a small portion of your deliverable—especially since we aren’t covering statistical tests used in the evaluation process until chapters 4 and 5. You’re right to be skeptical—and hear me out! Regardless of whether you’ve practiced these skills in a formal capacity, I argue this is *the most crucial chapter* to follow in depth. Here’s why:

- **Your hypothesis is your foundation.** How you structure your hypothesis helps guide the audience through the methods of analysis you are using. Documenting your background evidence and informed guesses sets a clear standard for your organization and how they should do the same in their work.
- **Investigating a question and hypothesis is the core of data analytics.** An analyst asks and answers questions, choosing from various methods to evaluate data. Whether comparing groups, tracking a metric, or training a machine learning model, you are drawing on this skill set to decide how to reach your goal.
- **Mastering this skill set can determine the success of your career.** This applies to most professionals in the world of data (data analyst, data scientist, etc.). If you can demonstrate rigor in asking and answering questions, you will find it easier to succeed in your work and career.

## 3.1 Informing a Hypothesis

I will start us off with a personal anecdote:

### Learning to Lead with a Question

During the first semester of my Ph.D. program, my advisor sent me a public

dataset to evaluate for potential analysis and publication. The dataset contained survey responses on adolescent behaviors and opinions across the United States. I was instructed to explore the available measures for interesting research questions and return with a hypothesis. After weeks of poring through the data catalog, published papers about the dataset, and some theoretical frameworks in our field, I came up empty. No amount of research got me closer to the right question, hypothesis, and methods.

It took discontinuing the degree and working as an analyst for several years to understand what led to that project's failure. As a new student and junior researcher, I approached the project with a naïve understanding of the analytic constraints I was operating within. I did not understand the limitations of available data, how to navigate those gaps, and where I should exercise my agency as a researcher and make a decision with the best information available. I was under the impression that if I consumed as much information as possible, a straight-forward question and hypothesis would emerge, representing the next sensible direction for the field.

Chapter 2 emphasized that research is not conducted in a vacuum. We seek information to inform our hypothesis and make an informed decision about how to structure a project. Conversely, there is rarely a complete and ideal set of information to guide your processes. You will often use your best judgment and acknowledge the information you have and don't have and the rationale for your decisions so others can contribute over time.

For most of your projects, you will synthesize information from a handful of sources to set the context for your stakeholders. We'll discuss strategies you can use to gather information, even when you may lack sufficient context and information.

### **3.1.1 Collecting Background Information**

Starting a new project can be daunting. *Where do you begin to understand what's already been studied or researched? How do you know you're on the right track? Will your analysis and findings make sense to stakeholders who know the domain better than you?*

If you've ever asked yourself the above questions, you're far from alone. Many analysts are brought into projects because of their experience working with *data* but not the domain area. In these cases, accumulating background knowledge is necessary to understand the context of the project.

Let's introduce at our case study for the chapter:

### **Starting an Analytics Project**

Jay is an analyst on the Insights team at a non-profit raising money for cancer research. His typical tasks involve analyzing the success of fundraising campaigns, donor and volunteer engagement efforts, and generating reports for the board of directors. The team received a request for a new effort to bring in adolescent volunteers. The organization hopes to understand what factors contribute to adolescent interest in volunteering and what positive outcomes they can expect.

Jay has been designated the project lead for the effort to conduct research. Since this is a new area of focus for the Insights team, they will have to synthesize a large body of work to understand the topic better.

I recommend a strategic approach to collecting information about a topic: *research, interviews, and exploration.*

### **Research**

**Research** is defined as (1) the investigation and synthesis of available information to establish a baseline understanding of a topic and (2) the process of investigating a topic as a study or experiment to gain *new* information. The first definition enables the effective execution of the second definition of research; in an ideal situation, a feedback loop can develop, leading to continued information gain on a topic.

Outside of an academic setting and specialized roles, few analysts are involved in publishing peer-reviewed papers. However, the principles of information synthesis remain the same: you compile an understanding of existing research and leverage it in your decisions about how to approach

your investigations.

When researching a topic, I recommend the following steps:

1. **Identify academic domains that research your topic.** You can often draw a wealth of knowledge from academic and public sources. Does your project require an understanding of human behavior? Take a look at domains of study within psychology and sociology. Economic or job trends? Look for topics within macro and labor economics.
2. **Search for peer-reviewed papers about your topic.** If you're unfamiliar with your research topic, some trial and error may be necessary to identify the terminology used in a specific field. Once you've identified search terms that return appropriate results, look for 3 to 10 papers to inform your topic, prioritizing recent papers published by different authors.
3. **Evaluate the papers you've selected.** Recent papers from top journals in a field can provide a lens into cutting-edge questions about a domain area and an existing synthesis of the field in the introduction. In addition to high-profile and recent research, look for papers with methods and samples closest to the population you will be working with (e.g., people in similar demographic groups and regions).
4. **Search for synthesized information outside of peer review.** Many resources draw from peer review that exists outside the academic system. Government agencies publish datasets and reports regularly, providing insights into a topic over time. Many fields have journalists or industry experts that publish work on a topic and can be followed for a layperson's evaluation of cutting-edge findings.

Let's return to the Insights team at the non-profit to understand how they can strategically synthesize research:

### **Synthesizing Background Information**

Jay's first step is to understand where the organization has gaps in knowledge. The team understands the success of adult volunteering efforts but has no experience working with *adolescents*. They know adolescents may have different motivations, availability, and financial resources and need guardian permission to participate. The team starts their search for papers on

adolescent volunteering behaviors. They guess that these are likely covered in the field of *developmental psychology*.

Next, Jay searches Google Scholar and PLoS One for papers on *adolescent volunteering*. He narrows the search to papers published in the last 20 years, saving eight papers he believes are most relevant.

Jay's third step is to read the papers in depth to understand their methods and findings. One paper looks at outcomes associated with adolescent volunteering, finding that adolescents who volunteer are more like to *continue* volunteering in adulthood. The study took place in Australia, which he notes may have economic and social conditions different from the United States, where he is located. However, the sample is quite large ( $n > 2000$ ), and the study controls for socioeconomic status, indicating that the impact of volunteering was seen regardless of income. He saves this paper for the team to review.

Finally, Jay searches for information on the benefits of adolescent volunteering outside of peer-reviewed papers. He discovers informational pages on several non-profit websites summarizing the benefits of volunteering. They cite additional studies that were not discovered in his initial search. They also give his team a concrete example of how their organization can communicate the benefits of volunteering for adolescents.

In applying the above steps, Jay identifies information to guide questions and hypotheses while providing summary information that can be shared with other teams to drive their decisions.

## **Stakeholder Interviews**

Your stakeholders can be a wealth of information on the domain-specific context and rationale of the work they request from you. Domain experts likely have access to information about resources in their field (e.g., publications, conferences, industry experts) and lessons learned from their hands-on experience. As you build relationships with those you support, you will find opportunities for a *bidirectional flow of information* that enables you and your teams to better make decisions in your roles.

If you're unsure where to start with appropriate questions, I recommend building and iterating on a list of standard information you find helpful in your work. This will change over time, become more comprehensive, and better reflect the needs of your projects.

Here are some examples to get you started:

- **General context:** This is especially important when working with a new team or one that has shifted focus. Why is the project important, and why now?
- **Background information:** Ask your stakeholders what sources of information (publications, podcasts, talks, etc.) informed the decision to pursue a project.
- **Expected outcomes:** As discussed in chapter 2, ask your stakeholders what they expect (or **hypothesize**) to happen due to pursuing the project or initiative. What is the value of the project succeeding or failing?

Let's learn how Jay asks for additional context from a partner team:

Jay sets up a meeting with the project lead for the volunteer initiative on the Program Management team. He has questions prepared for the project lead, Emma, to fill gaps in his knowledge and share the results of his research.

#### Clarifying Open Questions with Stakeholders

Jay: Can you tell me about the motivation for reaching out to adolescents as potential volunteers? This is a new direction for us, and I want to understand why this is valuable.

Emma: The number of volunteer registrations is far lower than last year. My team manager spoke to the Program Management team at another non-profit, and they've successfully increased registration by engaging adolescents.

Jay: What helped you decide this was the right initiative to pursue?

Emma: We did some research to see if other volunteer programs existed for adolescents at high schools or local community centers. We saw a few events at community centers, but none had a consistent presence or advertisements

discussing the benefits of volunteering.

Jay: We found a good amount of information about that from peer-reviewed literature. In addition, the websites of many non-profits have well-designed pages summarizing the benefits to volunteers, the community, and more.

Emma: Thanks for the information. A page on our website summarizing what we've learned about the benefits of volunteering may be a great addition to this project plan.

Jay: What are the specific outcomes you're hoping for by reaching out to potential adolescent volunteers?

Emma: We hope to increase the volunteer registration rate and tenure—the time a person continues to volunteer with us. It's also important to know what benefits might exist for adolescents who volunteer since that's likely different from adults. We hope they do better in school, have fewer disciplinary issues, and improve their well-being. We want to report on each benefit to the board and in future grant applications.

A simple conversation asking your stakeholders the proper contextual questions can go a long way in a new project. You'll likely learn the information you need to close the gap between your deliverable and their *expectations* for the deliverable.

## **Exploring Available Data**

The third step is to explore the available data at your organization. This is done before the *exploratory data analysis* of data collected as part of your project and serves to help operationalize your concepts (as discussed in chapter 2) and determine the *size of the opportunity* your organization is pursuing.

*Opportunity sizing* is a term commonly used in Product Management, which refers to the process associated with quantifying the scope of the potential impact of a project or course of action. The process is done through the synthesis of external research and context, with additional exploration and evidence gathered from data at the organization (sound familiar?). The

outcome of an opportunity sizing effort is typically an estimated range of users, customers, or behaviors expected to be impacted by the project or action. When done for multiple potential efforts, it can be an excellent tool for prioritizing work within an organization.

Let's look at how Jay gathers information to estimate the size of the opportunity to engage adolescent volunteers:

### **Opportunity Sizing**

Jay searches the organization's shared drive for information that can help estimate the potential impact of an adolescent volunteering initiative. The drive contains a historical record of presentations, whitepapers, program evaluations, and submitted grants for multiple efforts across five years. He also searches their city's database of grants to determine how many opportunities may be available to their organization by pursuing this effort.

Jay discovers the most recent performance report, which shows that the number of active volunteers has decreased for six months and is 20% lower than last year. New registrations are down, and volunteer tenure has slightly decreased in the same period.

In the donor database, Jay finds several previous donors have asked when the organization will directly engage younger volunteers. He also finds several available grants for organizations engaging youth in their community, totaling an opportunity of more than \$200,000.

### **Recap**

Since each project is different, you will likely draw more heavily from different approaches based on available information. Due to bandwidth or informational constraints, you will also likely take on projects without a thorough background investigation. When faced with these limitations, I recommend including information from *at least* two of the above steps and, in your deliverable, clearly highlighting areas where you have gaps in knowledge.

### 3.1.2 Constructing Your Hypothesis

With background research complete, you're ready to construct an informed hypothesis. In a research methods class, students learn a formalized method of stating a *null hypothesis* ( $H_0$ ) and *alternative hypothesis* ( $H_1$ ).

A study is conducted to determine if sufficient evidence exists to *reject* the null hypothesis and *accept* the alternative hypothesis. Here's an example of a hypothesis represented in this standardized format:

#### Stating your Hypotheses

H0: The test scores in the treatment group (individual tutoring) are equal to or less than in the control group (no tutoring).

H1: There are significantly higher test scores in the treatment group (individual tutoring) compared to the control group (no tutoring).

This demonstrates a hypothesis with a **directional prediction**, indicating a desired *direction* of differences for the test group (higher scores). Specifying a direction in an alternative hypothesis is not required, though most studies have an inherent "desired" direction for the outcome. If the results support the statistical criteria you set for your evaluation, you *reject* the null hypothesis. You fail to reject the null hypothesis if your results do *not* meet the directional and statistical criteria.

A strong hypothesis adheres to the following criteria:

- It identifies the **independent variables** (predictors) and **dependent variables** (outcomes).
- It is a declarative statement about the expected outcomes.
- It is a clear, concise statement easily interpretable by your stakeholders and audience.

You may notice that the criteria for an alternative hypothesis include a specified *direction*, not a specific, *quantifiable estimate* of the expected change. In most studies and analyses, this is acceptable and well-understood by your audiences. Where possible, I recommend taking additional steps to

estimate the *quantifiable difference* expected as part of your study, program, or experiment. This can be well-received by your stakeholders and provides you with additional numerical criteria to evaluate against your expectations.

## Quantifying a Hypothesis

The methods of defining a hypothesis that we've discussed so far are most closely aligned with the school of **frequentist hypothesis testing**.

Frequentist testing is a specific interpretation of probability that aims to make inferences about the broader population based on samples of data from that population. As part of that school of research and statistics, hypotheses are typically structured as an educated guess about the presence of group differences, and sometimes the direction of those differences. However, you may notice that we have stopped short of making specific, numerical estimates of the magnitude of those differences.

Quantifying a hypothesis is a process more aligned with the **Bayesian** approach to hypothesis testing, which interprets probabilities as a number representing the degree of belief about an event, based on the evidence available. This is not often taught in introductory statistics courses, but can be a valuable approach to your work that encourages you to think strategically about the actual differences and changes that you hypothesize will occur.

Quantifying a hypothesis is simple – estimate the size of the difference between groups and state it as part of your hypothesis. Now, I'm not suggesting you make something up! As part of this process, you will have an opportunity to learn more about existing information on your topic of interest either through peer-reviewed research, existing data at your organization, or domain knowledge from your stakeholders.

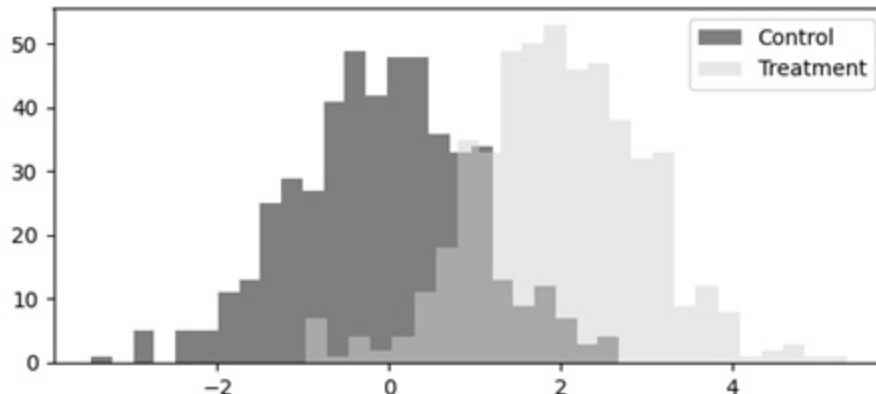
Beyond quantifying a hypothesis, we won't go into depth on probability and Bayesian statistics – there are already fantastic resources out there to build that knowledge. However, this is a great first step toward *thinking probabilistically* as an analyst.

Let's look at an example using Python. We will generate two overlapping distributions to simulate a hypothetical *treatment* and *control* group for the

hypothesis in the previous section. The distribution for the treatment group is shifted two standard deviations to the right to demonstrate what a highly effective treatment that has a statistically significant difference will look like.

```
import numpy as np # A
import matplotlib.pyplot as plt
c = np.random.normal(0, 1, size=500) # B
t = np.random.normal(2, 1, size=500)
plt.hist(c, alpha=0.5, bins=25, color="black") # C
plt.hist(t, alpha=0.5, bins=25, color="lightgray")
plt.legend(["Control", "Treatment"])
```

**Figure 3.2 Distributions showing a hypothetical group difference between the treatment and control group.**



This is a simplification of the process of estimating the underlying distributions representing a hypothesized change. It assumes you have access to a distribution or parameters about the population or a larger portion of the sample. From there, we hypothesize that the treatment group distribution will *shift* two standard deviations to the right, indicating a significant difference from the control group.

If we were actually conducting a study on test scores, we could estimate the distribution of test scores by looking at all students' grades in the school district. You can start quantifying your hypothesis by understanding the shape of existing data about the population you work with and if you have sample or population data from whom you will draw your control group. From there, we can make the following assumptions:

- The control group will have a distribution roughly identical to the

student population

- The treatment group will have a distribution representing a positive shift in the number of standard deviations (or a specific point increase in raw test scores) in line with those found in previous studies.

Where comprehensive research is not available to estimate change, we can quantify expected change using the best available information from qualitative sources and the goals of the program or study.

Let's look at how Jay's team defines and quantifies their hypotheses:

### **Quantifying Hypothesized Changes**

Jay constructs the following hypotheses for his report on the upcoming evaluation of the new volunteer engagement:

*H0: There is no significant difference in the registration rate between adult and adolescent volunteering events.*

*H1: There is a significant difference in the registration rate between adult and adolescent volunteering events.*

In addition to a null and alternative hypothesis with an expected direction, Jay attempts to estimate the quantifiable change in registrations for the program team. He recalls that one peer-reviewed paper found that more than 50% of adolescents participate in volunteering activities, compared to less than 30% of adults. This aligns with the organization's knowledge of volunteer registration rates at engagement activities for adults: about 25% of event attendees will register to volunteer. The organization estimates 5000 adolescents from local middle and high schools will attend the volunteer fairs at schools and community centers. If 50% of them register across the six months that the events are scheduled, the number of weekly volunteer registrations for the organization will increase by 60% overall. Jay *quantifies* his hypotheses with this estimation:

*H0: There is no significant difference in the registration rate between adult and adolescent volunteering events.*

*H1: There is registration rate for adolescent volunteering events will be 50%, compared to 25% for adult volunteering events.*

Jay can comprehensively evaluate the program outcomes with a quantified hypothesis. He can compare the actual vs. hypothesized changes and monitor how closely their performance aligns with reported trends on volunteering behavior and non-profit success.

### **3.1.3 Exercises**

You are part of a Product Analytics team at an e-commerce company that designs A/B test experiments to increase subscriptions and improve users' experience with the software. You are designing a series of experiments to answer the question: *What page layouts, tooltips, and recommendations decrease the rate of abandoned shopping carts without a purchase?*

1. What sources of information can you leverage to collect background information on the expected outcomes of the software changes?
  - a. What questions will you ask your stakeholders to gain the appropriate context for the experiment?
2. You decide to design an experiment that compares 3 experiment groups with different layouts and one control group. Write a null and alternative hypothesis based on the research question.
3. Update your null and alternative hypothesis to *quantify* the expected outcome for the experimental and control groups.

When completing this exercise, you can define the expected outcomes (hypothesized group differences, direction, and quantified values) using an appropriate example. You can also suggest data sources internal to an organization that would be valuable to have access to (e.g., database tables, reports).

## **3.2 Methods of Gathering Evidence**

With a defined hypothesis, the next step is to collect and report on data to test and evaluate the hypothesis. But what shape should the data be in? How exactly is everything structured? There are *a lot* of methods to choose from,

guided by your question. We'll cover three methods under which most research can be classified: descriptions, associations, and causal relationships. The usage differs by the discipline of study and type of data usually collected; however, these approaches are common to work in Product Analytics, Marketing Analytics, Business Analytics, and more.

### 3.2.1 Descriptions

The simplest data analysis is a presentation of *descriptive information*. As the name suggests, this method *describes* a phenomenon without manipulating a test variable or condition. This approach to analysis is ideal for understanding new data sources, developing metrics, and opportunity sizing.

Descriptive methods can involve existing data analysis or active data collection and can be performed on quantitative and qualitative information. Data is often presented using measures of central tendency, trends over time, or group differences.

#### Descriptive Statistics

**Descriptive statistics** refers to methods used to summarize insights from a quantitative dataset. An analyst determines which descriptive measures are most appropriate for the dataset and what information they want stakeholders to glean from the data. These include measures of central tendency (mean, median, mode) and measures of the distribution (standard deviation) for continuous data, counts, or proportions for categorical data.

Reporting on descriptive statistics can be part of an inferential statistical workflow or exist as a standalone deliverable if it meets the stakeholder's needs. Many reports and dashboards rely entirely on descriptive statistics to deliver value. While the actual analysis is straight-forward, descriptive statistics can be the most useful routine insights used within an organization.

When creating deliverables that *only* rely on descriptive statistics, I recommend considering the following challenges:

- **Selecting statistics:** Choosing appropriate descriptive statistics requires

you thoroughly explore the dataset, the shape of its distribution, and understand what you **cannot** interpret if you exclude a statistic from your final deliverable. A graph with a mean, median, sum, or percentage of the total will lead your readers to very different conclusions about the same information.

- **Guiding interpretations:** Deliverables relying on descriptive statistics are often designed to be straight-forward self-service tools (e.g., a dashboard). However, stakeholders have a broad range of analytic proficiency and may draw different conclusions from the same information set. Include strong guidance for what you can and cannot conclude from a specific tool.
- **Conflating explanation with the cause:** Descriptive data points tracked over time are common within organizations. If their presentation is not paired with a strong understanding of what **impacts** the tracked data points, stakeholders may be left with poor estimations of what causes changes and trends.

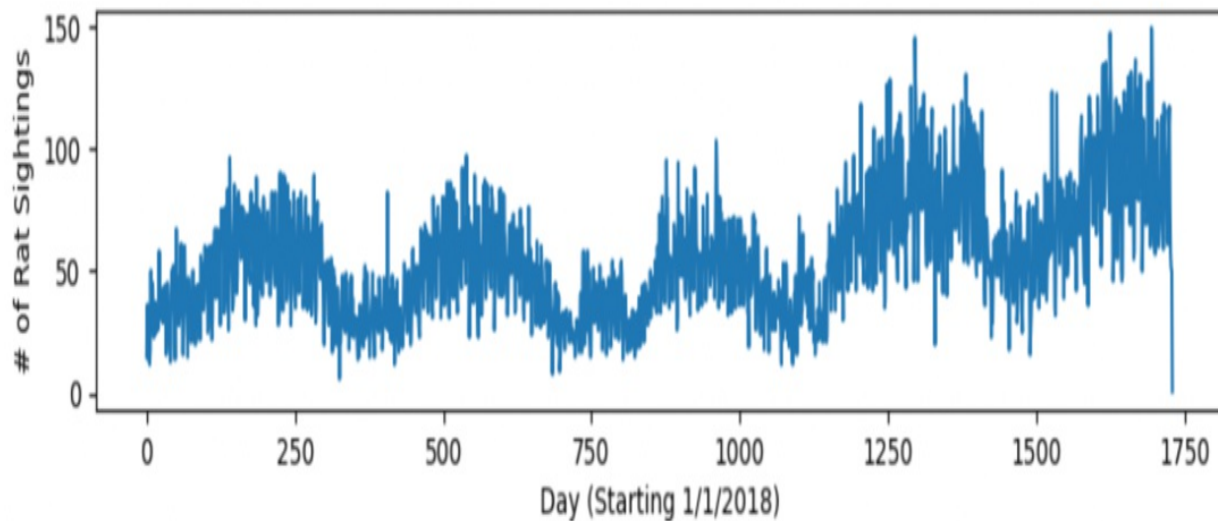
Let's see an example of how different descriptive statistic presentations can impact interpretation. We'll use a dataset called `rat_sightings.csv`, a subset of the NYC Open Data 311 dataset. Each row is the number of calls to the 311 hotline about public rat sightings per day between January 1, 2018, and June 30, 2022 (the entire 311 hotline dataset contains billions of rows about thousands of call types).

How might we answer the question, *have the number of rat sightings changed over time?* Our independent variable is the *number of rat sightings*, and our dependent variable is *time*; no group differences or pre/post comparisons are necessary for this question.

We can import the dataset in Python and generate a line plot as follows:

```
import pandas as pd #A
import matplotlib.pyplot as plt
rats = pd.read_csv("rat_sightings.csv", index_col=0) #B
plt.plot(rats["rat_sightings"]) #C
plt.xlabel("Day (Starting 1/1/2018)")
plt.ylabel("# of Rat Sightings")
```

**Figure 3.3** Time series plot of daily reports of rat sightings in NYC.



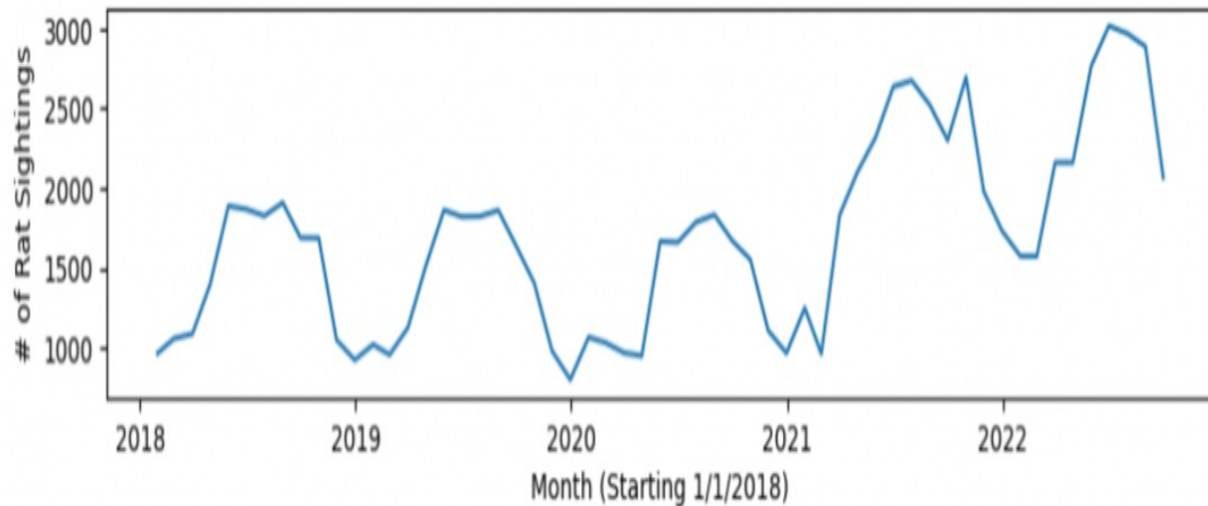
From this graph, we can see the following:

- The number of daily rat sightings has been trending upward in recent years
- There are consistent weekly and monthly seasonal trends in the number of rat sightings

The daily plot above makes it challenging to estimate the true *volume* of rat sightings in larger time periods (e.g., a week, a month). If this were a final deliverable, you could expect to receive follow-up questions to create views at different granularities. This can be achieved through aggregations such as a mean, median, or sum:

```
rats = rats.reset_index()
rats["day"] = pd.to_datetime(rats["day"]) #A
rats_group = rats.groupby(pd.Grouper(key="day", axis=0, freq="M"))
    ["sum", "mean", "median"]
) #B
rats_group.columns = rats_group.columns.get_level_values(1) #C
plt.plot(rats_group["sum"]) #D
plt.xlabel("Month (Starting 1/1/2018)")
plt.ylabel("# of Rat Sightings")
```

**Figure 3.4** Time series plot of total monthly rat sightings in NYC.

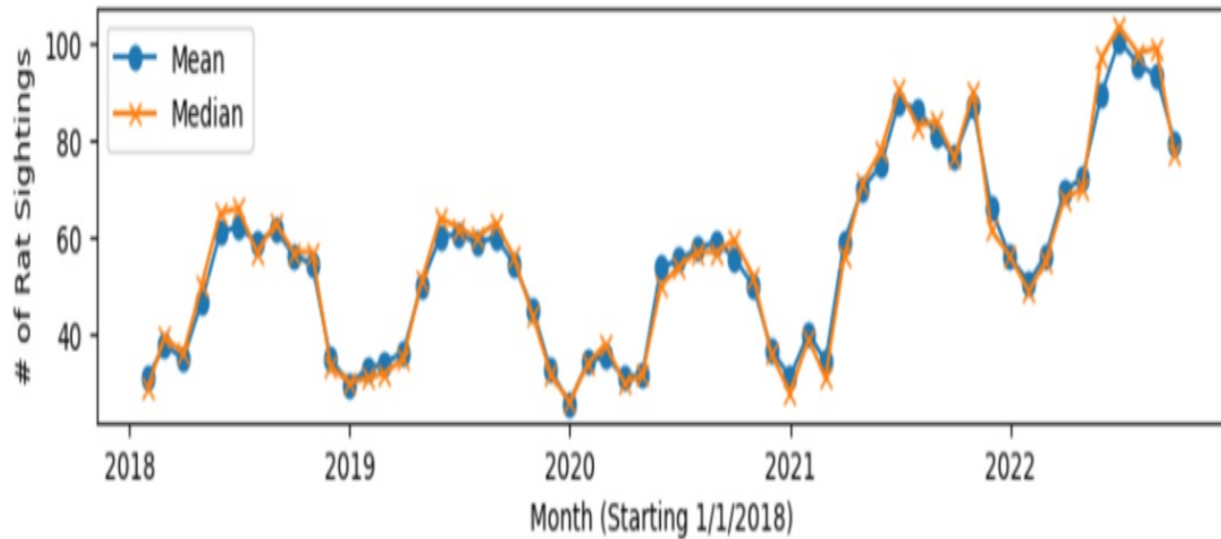


This visual makes it easier for stakeholders to understand the *volume* of rat sightings over time. The weekly seasonality causing the dense spikes has been removed, highlighting the monthly seasonality associated with colder months (approx. November through March) and warmer months (April through October) in New York City.

Finally, we will compare the mean and median values. This will help us determine if there are properties of the underlying distribution we should highlight in a deliverable.

```
plt.plot(rats_group["mean"], marker="o") #A
plt.plot(rats_group["median"], marker="x") #B
plt.legend(["Mean", "Median"]) #C
```

**Figure 3.5 Time series plot of mean and median monthly rat sightings in NYC.**



There's minimal difference between the mean and median values over time, indicating that there is likely a consistent normal distribution underlying the dataset. We can also see that the shape of the data over time is identical to the graph showing the *sum* of rat sightings. Given that this second graph adds little new information, we can leave it out of our final deliverable. Instead, we can include a table or sentence describing the mean/median values and how they have changed over time.

Let's return to Jay's Insights team and review the statistics he includes in his report:

Three months have passed, and 20 of the planned 40 volunteering events have concluded.

For those 20 events, the registration rate was an average of 36%. This is lower than the anticipated rate (50%); however, this is higher than the rate for the 22 adult volunteering events (21%), and the overall number of registrations is the highest in the organization's history. The average volunteer tenure has not changed; however, we do not expect to determine if there are changes for at least another 3 to 6 months.

The descriptive information above highlights the percentage change (a

*relative* value) and a rolling average value of weekly registrations (an *absolute* value). Including both together is more valuable than each measure on its own and sets the context for statistical tests performed in the evaluation.

## Qualitative Research

*Qualitative analysis* involves the synthesis, analysis, and interpretation of non-numerical data. This often requires deriving insights from unstructured or free-form language data recorded as text. As an analyst, you may be asked to leverage methods used in humanities and social sciences (e.g., 1:1 interviews, focus groups) or to derive insights from much larger samples of text data using *natural language processing*.

We will discuss each of these approaches in more detail throughout this book. For this chapter, I recommend the following takeaways when deriving insights from qualitative data:

- Quotes and themes from interviews, focus groups, or free text survey responses are excellent **aids** to bring quantitative insights to life. A slide or section with anecdotes that support your interpretation can help ground your analysis in the experience of the people you collect data from.
- Small sample qualitative methods (e.g., 1:1 interviews, focus groups) can be an appropriate starting point for research but struggle with **generalizability** when performed independently. You will likely **need** to complement qualitative with quantitative ones.
- Natural language processing approaches (e.g., sentiment analysis, topic modeling) can support generating insights for larger samples of text data but can be confusing for someone unfamiliar with the methods. Set strong expectations with stakeholders on what the deliverable will look like and how it is derived.

## Words of Warning: Description vs. Inference

Descriptive methods are intended to summarize the characteristics of *a specific set of data*. But how do you know any group differences are

statistically meaningful, or if you can *infer* that your measures exist in the broader population?

If you are presenting descriptive information in a deliverable, consider the following limitations and communicate them to your intended audience:

- A trend, mean, or median value is **not** sufficient to infer that your findings exist beyond the dataset you are working with.
- A mean or median value between groups is **not** sufficient to determine if differences are large enough to be meaningful.
- A current trend is not guaranteed to **continue**—especially if you don’t yet understand the factors influencing the trend.

### 3.2.2 Correlations

One of the most common approaches to comparing continuous data is to look for **associations** between variables. These associations usually take the form of a measure of **covariance** (an unstandardized measure of how two variables *vary together*) or *correlation*, which is a standardized covariance measure. The term **correlation** is well understood outside of data practitioners and can be explained to your stakeholders using plain language terminology and mathematical concepts from high school algebra.

*Pearson’s correlation* is the most well-known method of identifying linear associations between variables. It can be used with any continuous data, does not require you to standardize your units, and the direction of the relationship is easy to interpret and explain (e.g., a negative correlation coefficient indicates a negative relationship). You can also expect that most stakeholders and partners outside a data team have encountered Pearson’s correlations in their careers.

Let’s build on our example `rat_sightings.csv` data. We saw in the previous section that there is a seasonality to the number of rat sightings in New York City, with more reported during warmer months and fewer during winter months. We can explore whether there are associations between rat sightings and weather parameters (temperature, humidity, wind speed, or precipitation) on a given day by combining these two data sources. A new file,

weather.csv, contains daily weather parameters from January 1, 2018, to December 31, 2020 (3 out of the five years included in the rats dataset). We join the dataset and generate a matrix of Pearson's correlations as follows:

```
weather = pd.read_csv("weather.csv", index_col=0) #A
rats_weather = weather.join(rats, how="left").fillna(0) #B
corrs = rats_weather.corr() #C
corrs.style.background_gradient(cmap="RdBu", vmin=-1)
```

**Figure 3.6** Pearson's Correlations between the number of daily rat sightings and weather parameters.

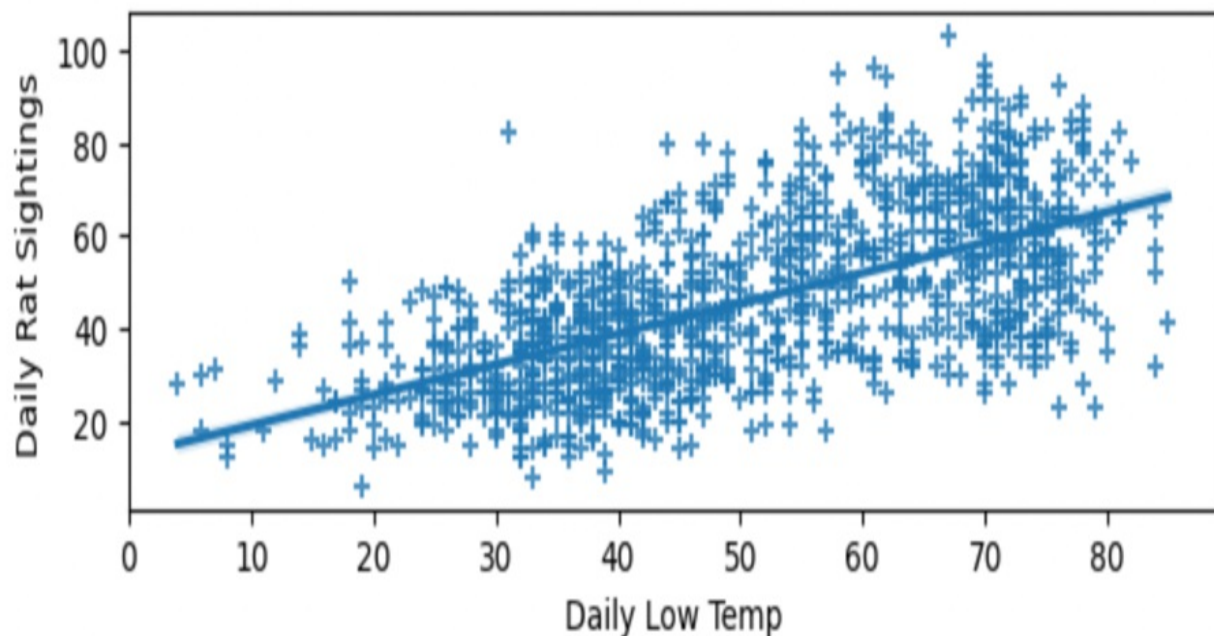
	rat_sightings	high_temp	low_temp	humidity	wind_speed	precip
rat_sightings	1.000000	0.600707	0.615463	0.153749	-0.242205	-0.029722
high_temp	0.600707	1.000000	0.962917	0.151777	-0.231311	-0.036839
low_temp	0.615463	0.962917	1.000000	0.177102	-0.260153	-0.026765
humidity	0.153749	0.151777	0.177102	1.000000	0.029353	0.233285
wind_speed	-0.242205	-0.231311	-0.260153	0.029353	1.000000	0.212698
precip	-0.029722	-0.036839	-0.026765	0.233285	0.212698	1.000000

The daily sightings of rats have a strong positive linear correlation with high and low temperatures on the same day. We can see that the high and low temperatures are correlated at nearly  $r = 1$ , indicating they are likely not independent, and we should select *one* of the variables to highlight the relationship. We also see a weak to moderate *negative* correlation with wind speed and little to no association with humidity or precipitation.

An association with a strong Pearson's correlation coefficient is often easy to visualize and share as a scatterplot in reports or dashboards. If the trend is unclear, add a regression line to the plot to demonstrate the linear relationship better.

```
import seaborn as sns #A
sns.regplot( #B
    x="low_temp",
    y="rat_sightings",
    data=rats_weather,
    marker="+",
)
plt.xlabel("Daily Low Temp") #C
plt.ylabel("Daily Rat Sightings")
```

**Figure 3.7** Scatterplot of daily high temperatures vs. daily rat sightings.



As we saw in the correlation matrix, there is a clear positive correlation between the daily low temperature and the number of rat sightings. The warmer the temperature, the more rat sightings people report to the city's 311 hotline.

## **Deliverables Using Correlations**

The correlational relationship above is an example of an association that is valuable to deliver in advance of a complete statistical analysis with predictors of change. A simple deliverable communicating the expected increase in rat sightings associated with warmer months or a heat wave will allow interested parties (e.g., a government agency or a restaurant) to make preparations based on the information. It won't be sufficient information to comprehensively *reduce* rat populations—that requires more sophisticated methods we will cover in later sections. However, this is an example of knowledge that generates value by sharing in advance of more complex analyses.

When sharing deliverables based on correlations, I recommend the following steps to ensure the accuracy of the results you share:

- If using Pearson's correlation, explore and visualize all correlations for the **linearity** of the trend. The scatterplot above shows that the trend is approximately linear; in many cases, the relationship may be better fit by a **curvilinear** trend line. We'll discuss methods for achieving this in chapter 4.
- Ensure that the default Pearson's correlation is appropriate for your analysis when generating correlations. If you are generating correlations between ordinal data points or are more interested in the **relative** scores of your variables, Spearman's correlation is a more appropriate choice for your analysis.
- As demonstrated above, correlations can provide a great starting point for planning and decision-making. However, they are rarely sufficient if the goal is to move the needle on one of the measures. Work with your stakeholders to determine if the correlations you report on are appropriate for their needs or if an experimental method is necessary.

Let's return to Jay and the Insights team at the non-profit to see the associations they discovered in their evaluation and how they report on them.

Jay has shared the initial descriptive summary with Emma, the Program Manager in charge of the youth volunteering events. The initial findings seem promising; however, she notes that the number and percentage of event

attendees registering to volunteer varies widely. She asks Jay if he can identify some factors correlated with event registration rates.

Since only 20 events have occurred by this point, Jay decides to retrieve data from the last 100 adult events and select information about them: the number of staff, the number of event attendees, the amount of money spent on food and catering, and the number of registrations. From these data points, he derives the ratio of attendees to registrations and the ratio of staff to attendees.

Jay discovers the strongest correlation ( $r = 0.65$ ) is between the ratio of attendees to registrations and the ratio of staff to attendees. When he creates a scatterplot of these two variables, he sees a clear linear trend between the variables; when he generates separate trend lines for adult vs. youth events, they appear to be nearly identical.

Jay incorporates his findings into the draft report he prepares for the program team. He recognizes that communicating the insight to the team *now* can potentially benefit the volunteer events scheduled in the coming weeks. He informs Emma of the relationship he discovers and recommends increasing the total number of staff scheduled to support larger events. He stresses that the relationship he discovered is only an association and will follow up with a more in-depth analysis after the scheduled volunteer events have passed.

Correlations can be inherently valuable insights to share as part of your deliverables. Sometimes, they may even derive value if shared before the final deliverable. When doing so, manage stakeholder expectations about the validity and non-causal nature of the relationship you are communicating.

### **Words of Warning: Correlation vs. Causation**

You may have heard the phrase: *correlation does not equal causation*. This is emphasized in statistics and research methods curricula. It's a phrase associated with analytics humor—for example, tylervigen.com has a website dedicated to spurious correlations.

Conflating correlation with causation can pose challenges for an analytics

team. In addition to stakeholder misalignment, conflating correlational with causal relationships can detract from efforts to effect change on an outcome, leading to poor quality recommendations and inefficient resource use. The examples of strong correlational relationships we discussed are inherently valuable to share on their own, with an emphasis that no requisite work was done to establish cause and effect.

In the case of associating rat sightings with warmer temperatures, we can look at the following information to dissect the limitations of this relationship:

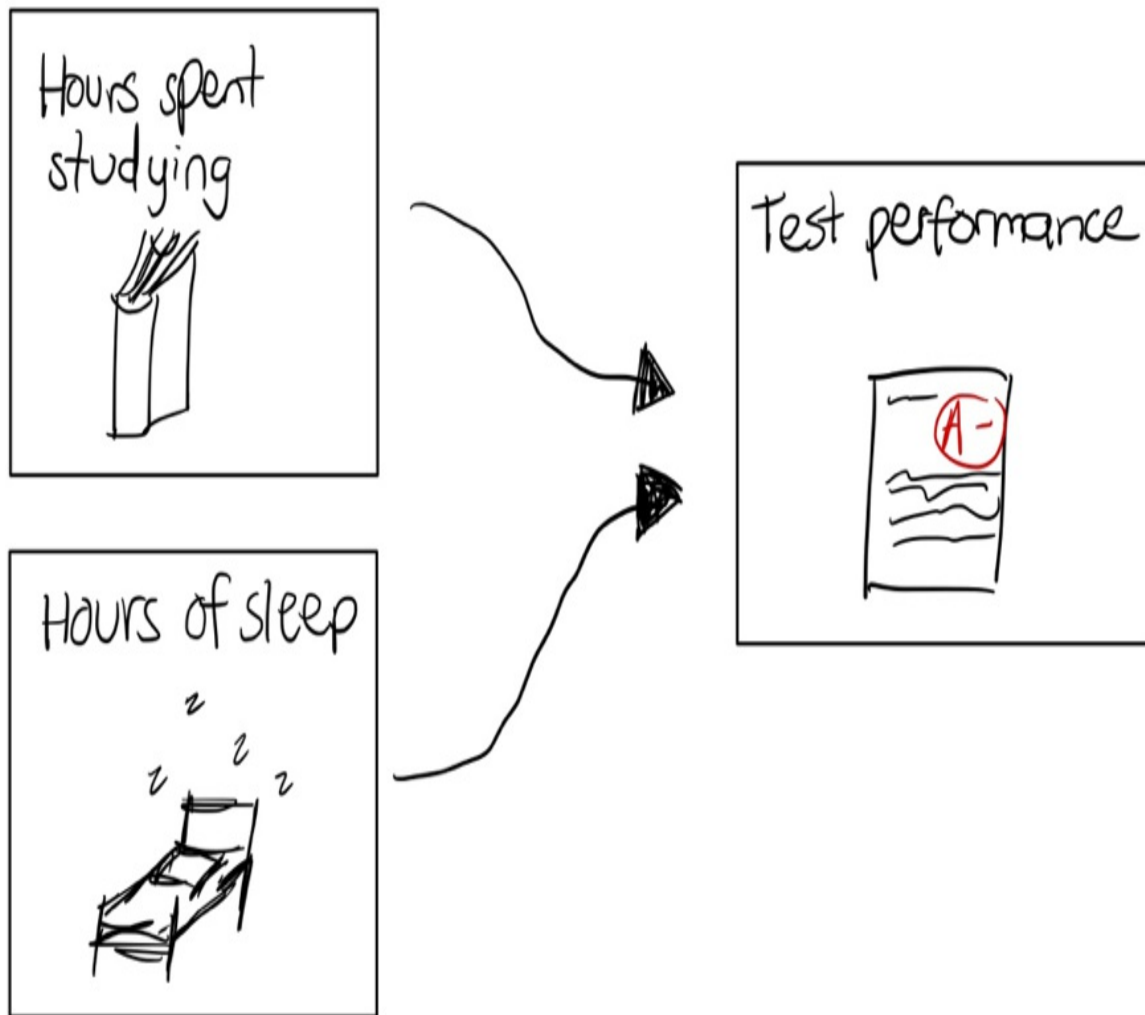
- Does warmer weather **cause** more rat sightings? If we attribute a causal relationship to this question, can we manipulate our independent variable (temperature) to change the dependent variable (the number of rat sightings)?
- Are rat sightings representative of the rat **population** or the **visibility** of the rat population?
- Is the goal to reduce the rat **population** or the **visibility** of the rat population? (We will expand more on this in chapters 6 and 7.)

Investigating these questions will guide your messaging to stakeholders, help you focus on what you can and cannot manipulate in your evaluations, and set you up for an *experimental design* to more appropriately attribute cause and effect to a phenomenon you are analyzing.

### 3.2.3 Experiments

An *experiment* is an investigation where an independent variable is directly manipulated, and the dependent variable is observed and measured. A researcher will seek to control as many conditions of the experiment as possible so that changes in the dependent variable can be confidently attributed to the manipulation of the independent variable. Simply put, the goal is to determine to the best of one's ability whether A *causes* B.

**Figure 3.8 A hypothesized causal relationship between two independent and one dependent variable.**



In an ideal situation, an experiment meets the following conditions:

- **Random selection:** Participants or subjects in an experiment should be randomly selected from the population. Selection can be a purely random or **stratified** sample, where participants are chosen proportionately to relevant subgroups in the population.
- **Random group assignment:** Selected participants should be randomly assigned to one of the groups in the experiment (e.g., treatment vs. control). Just as with random selection, assignment can be purely random or stratified.
- **Controlled environment:** The experiment should occur in highly controlled conditions, where as many variables as possible are managed or removed from the environment to better attribute cause to the

independent variable. For example, an A/B test **only** changed the color of a button on a website and tracked differences in newsletter subscription rates between groups. Since only the color differed between the three groups (red, blue, green), the team can confidently say that the red button **caused** or **influenced** more people to subscribe to the newsletter.

- **Manipulating independent variables:** The independent variable should be a condition that you can directly **manipulate** and change to confidently attribute cause to the changes you control for as the researcher.

In an academic or clinical setting, you may be familiar with a *randomized controlled trial* as the ideal standard of experiment used to attribute cause. These include trials for new medication, medical treatment, psychological studies, and more. Experimentation is also commonly used in other industries, absent the laboratory settings associated with the practice. Experiments are used in non-profits to evaluate the effectiveness of programs. Businesses can use experiments to assess the efficacy of iterative changes on a website or product.

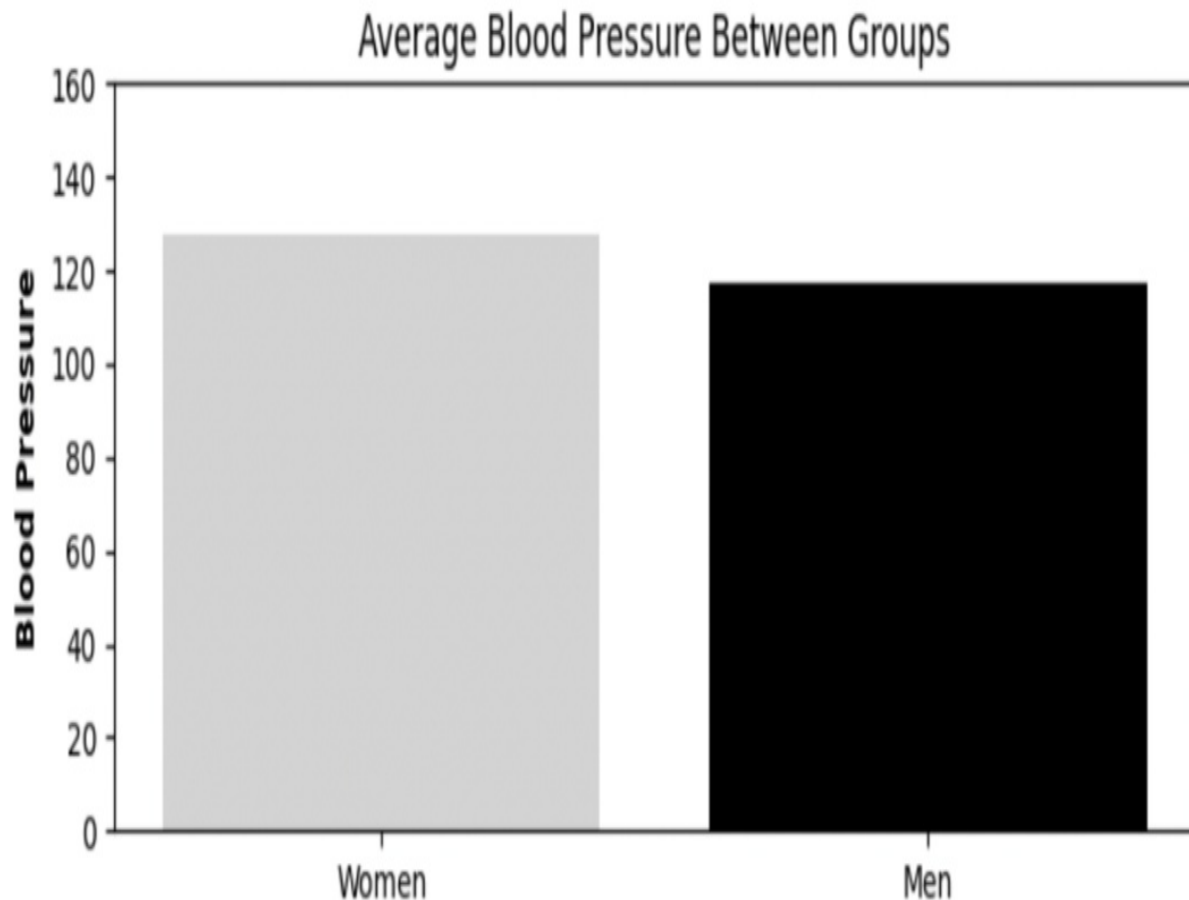
## Quasi-Experiments

In many cases, you will want to design experiments where the independent variable is a characteristic inherent to your participants and not something you can directly manipulate and assign. Statistically significant differences will often exist between participant demographic groups or inherent characteristics; depending on the questions you are answering, you will likely either be looking to control for these variables or measure them inherently as part of your core evaluation. The latter study design is known as a *quasi-experiment*.

For example, a study comparing the efficacy of a blood pressure medication between men and women has a valid body of research suggesting there will be differences in blood pressure decreases between those groups. However, the researcher cannot randomize participants into the Male/Female categories – they can only work with the characteristics of the participants selected for the study. Figure 3.8 shows a simple comparison of blood pressure records

between participants after receiving medication for four weeks:

**Figure 3.9** Quasi-experiments have a similar design to controlled experiments without random group assignment.



Quasi-experiments comparing participant demographics are common in academic research, clinical trials, and non-profits, where differences between groups are often expected and meaningful phenomena to report on. In a business setting, participant demographic groups are frequently used as the basis for *segmenting* users into cohorts based on observed or expected behaviors.

Here are some example quasi-experimental research questions across industries:

- Do male/female and elementary/middle school students see improved

- math scores when participating in an individualized tutoring program?
- Do customers renew their subscriptions for longer in rural, urban, or suburban areas?
  - How do youth from different family income brackets benefit from extra tutoring?
  - How often do users at different career stages visit and engage with our website and product?

Quasi-experimental studies can be incorporated into a randomized control trial or exist as a standalone evaluation. In both cases, researchers will typically compare multiple participant characteristic groups to ensure appropriate documentation of all sub-group trends. A combination of sub-groups is often compared for *interaction effects*. This view of participants is more prone to Type I false-positive errors but can also lead to more granular insights.

### **Words of Warning: Evidence vs. Proof**

Using the word “proof” is a personal faux pas of mine as a data professional. I have stakeholders who jokingly share that they know not to use that word around me, as if it were a generally inappropriate term to use in the workplace. While I may be considered on the lookout for attempts to “prove” something with data, this is evidence of a strong relationship with the teams I collaborate with. My colleagues will catch and correct themselves mid-sentence as we discuss projects. Over time, I have found that they are far more prepared for the times we find evidence contradicting previous knowledge within the organization.

Evidence and proof are not the same thing. *Evidence* is information supporting a hypothesis or theory; *proof* is a claim treated as a rule and not designed to be refuted. Even if someone references data or evidence supporting a “proven” statement, that does not make it data-informed. A “proven” statement or belief is remarkably impervious to change or new information that contradicts previous information. As analysts and researchers, we collect *evidence* with the understanding that future

information can counter previous information and be accepted if the methods used to collect and analyze it are sound.

This may seem like splitting hairs over two words. Still, I emphasize this based on my experience of how the language used within an organization is reflective of its data-informed culture. Organizations and teams that frequently assume information is *proof* of a phenomenon are often resistant to change and new information, even when ignoring that information can have a negative impact on the organization. In fact, many such organizations tend to approach data in a backward capacity – looking for information that *proves* a strongly-held belief.

Managing the misuse and misinterpretation of findings can be challenging for an analyst; you may functionally take on an entire organization's culture without sufficient resources. At a *minimum*, I recommend sticking to a script around the interpretation of findings in your deliverables and communications:

- **Use your language carefully:** Phrases like “we discovered evidence in support of the hypothesis” can go a long way when re-emphasized across deliverables.
- **Guide stakeholder interpretations:** As we discussed in chapter 2, a slide or section with recommended statements of interpretation can be **very** helpful to your stakeholders if they are less experienced in leveraging data and evidence in their work.
- **Provide context for why findings change:** If you present findings that contradict previous findings or strongly held beliefs, include context for why the findings or trends may have changed. Has your user base transformed since the analysis was last performed? Was your study conducted on a different sample than usual? Did you approach your work in a novel way?

Standardizing your communication on the above can go a *long way* toward building a data-informed relationship with your stakeholders, even when you lack the resources to create an organizational culture shift around using data.

### 3.2.4 Types of Study Design

When researchers choose to conduct an experiment, they have to select an appropriate **study design** in order to identify how to best evaluate their participants. The *study design* refers to the method of assigning participants into *comparison groups*, which serve as the independent variable of the study. The study design informs what statistical tests should be used to determine if group differences are statistically meaningful.

When evaluating multiple hypotheses, a study can feasibly include more than one study design. However, a single hypothesis is usually evaluated with one study design, and one type of statistical test.

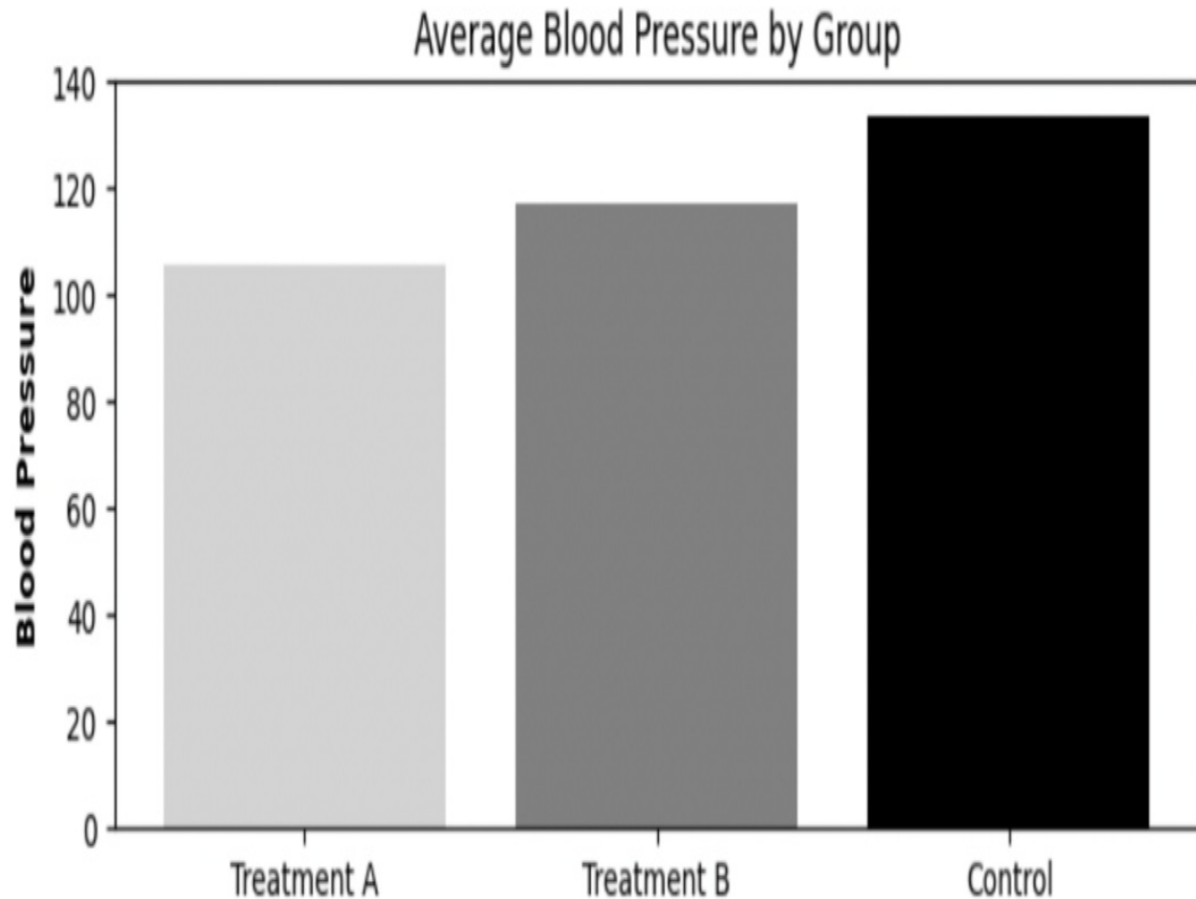
We'll be discussing some of the most common study designs used in academia, businesses, and other types of organizations. This list is *far* from exhaustive; there are dozens of study designs used to answer specialized questions in specific domain areas beyond the scope of this book. However, the study designs we cover here will likely apply to 90% or more of the use cases you will encounter in the first few years of your analytics career.

## Between-Subjects

A **between-subjects design** compares your dependent variable between *separate groups of participants*, which is your *independent variable* (e.g., total purchases made on a website compared between geographic locations). This method is used in the majority of experiment and study examples we've discussed so far.

In this design, participants belong to *mutually exclusive groups* either based on an inherent characteristic (e.g., age) or **random assignment** to a group (e.g., treatment vs. control). Statistical tests for *independent samples* (referring to different, independent groups, which will be covered in chapter 4) are then used to compare values on the dependent variable between each group in the study.

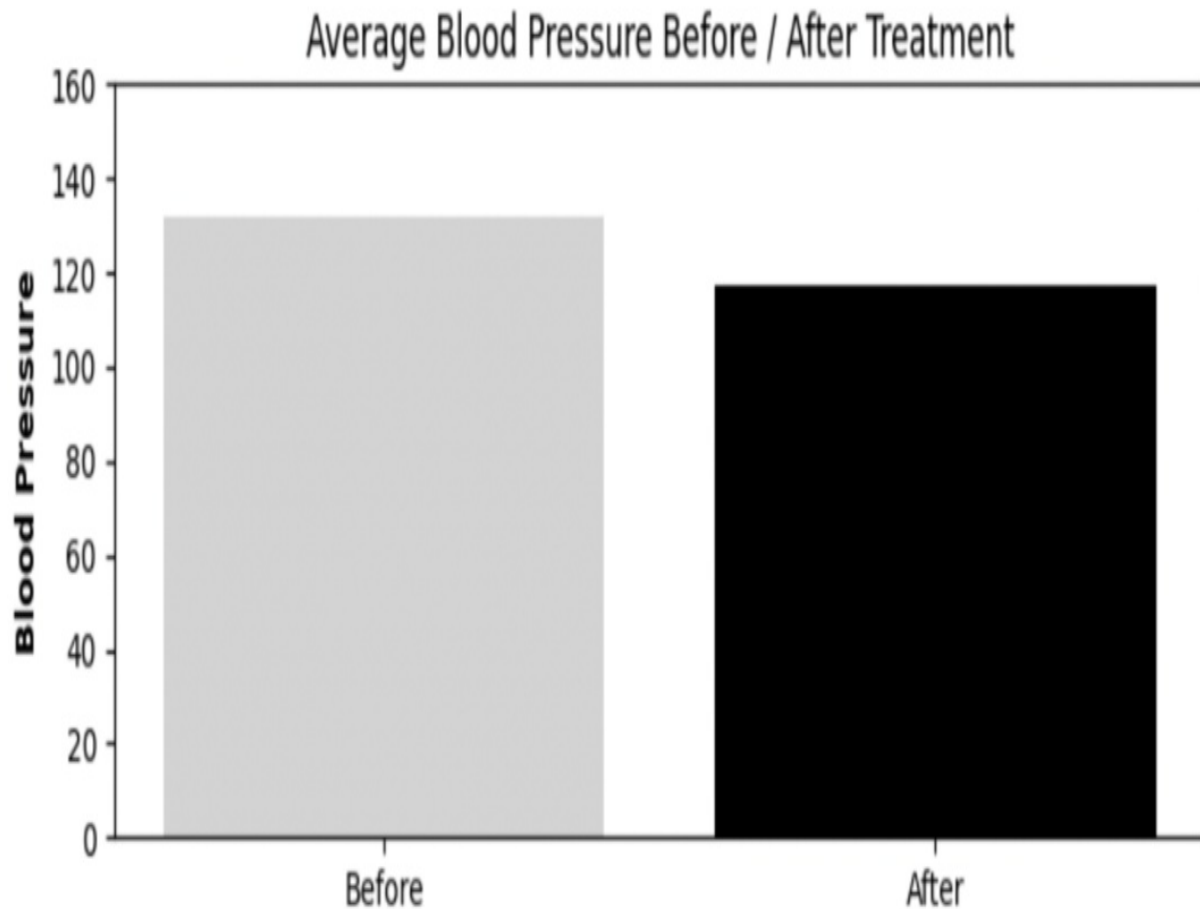
**Figure 3.10 Between-subjects comparisons evaluate differences between mutually exclusive characteristics or assignments.**



### Within subjects

A **within-subjects design** (also known as a *repeated measures design*) involves exposing *every participant* to *every independent* variable condition in an experiment. Participants are repeatedly measured on the dependent variable before and after exposure to the independent variable condition. The study design may include a two-group pre/post comparison or repeated assessment across multiple time points. Experiments with a repeated measures design will use different variations of statistical tests than those used to evaluate a between-subjects design.

**Figure 3.11 Within-subjects comparisons evaluate differences before and after treatment.**

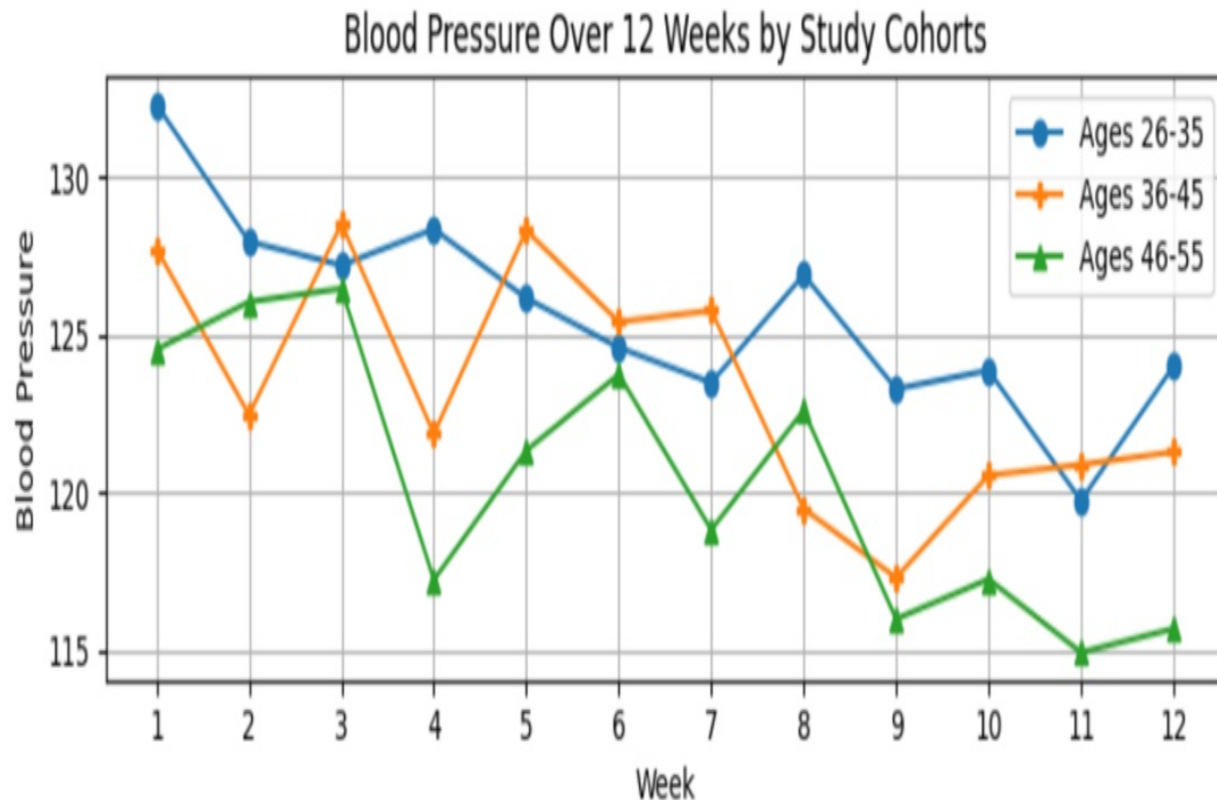


## Cohort Comparisons

A **cohort study** is a type of study that groups participants into meaningful cohorts to evaluate over extended time intervals. Participants can be assigned to cohorts based on characteristics that change over time (e.g., age or the year they subscribed to a service), static characteristics (e.g., school attended, subscription tier).

Cohort study designs combine within-subjects and between-subjects approaches in order to track meaningful changes over time period that would otherwise be missed with a simple before vs. after comparison. These studies also exclude random assignment and instead seek to identify whether measurable differences in trends occur over time between mutually exclusive groups.

**Figure 3.12 Cohort comparisons evaluate *meaningful* cohort groups over time.**



## Longitudinal Comparisons

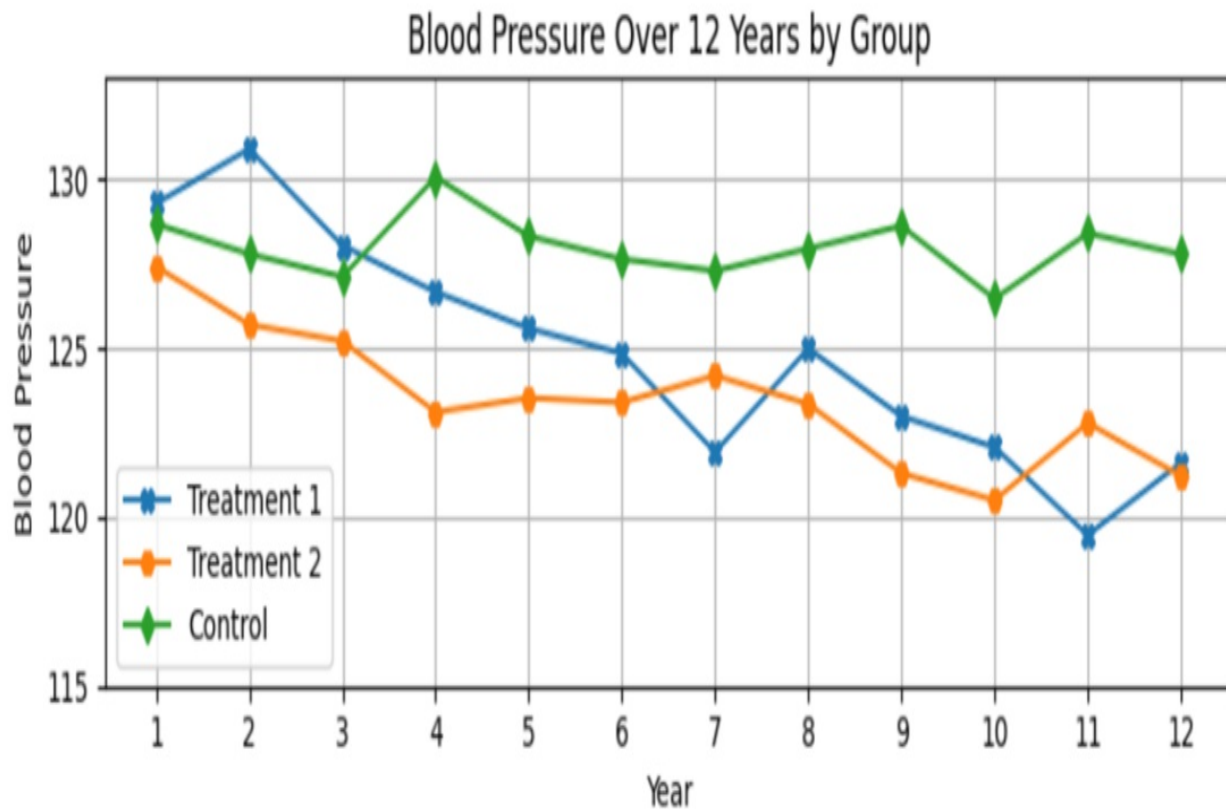
Both within-subject and cohort designs are types of **longitudinal comparisons**. A longitudinal study design assesses participants *repeatedly* at *multiple points over time*, many collecting data repeatedly over weeks, months, or even years. These study designs look for discernable changes and trends in a dependent variable over the desired time period.

Longitudinal studies can also incorporate both between-subjects and within-subjects comparisons in order to assess differences between randomly assigned groups or cohorts over time. Participants may be assigned to one of multiple treatment groups for a multi-year pharmaceutical trial, or different geographical cohorts might be monitored to understand their spending patterns over a calendar year.

Longitudinal tests with a limited number of time periods are often assessed

using repeated measures univariate tests (e.g., t-tests, ANOVAs). Those who follow participants for longer periods of time may require more specialized statistical methods tailored to the question (e.g., survival analysis, growth curve analysis).

**Figure 3.13 Longitudinal comparisons evaluate participants over months, years, or even lifetimes.**



Our analyst Jay is ready for the final step of his evaluation:

#### **Evaluating the Efficacy of a Program**

The program team has completed all of its 40 youth volunteering events. Jay has retrieved and prepared the data for analysis, re-calculated his descriptive statistics, and updated correlations on the relationship between staff-to-attendee ratios and the registration rate (percentage of attendees who register to volunteer).

He also performs an *independent samples* t-test to assess his *between-subjects*

*design* comparing the registration rate of the 40 youth volunteering events to the 38 adult volunteering events held in the previous six months—the same duration of time in which the youth events were held. He finds that the youth events had significantly higher registration rates than the adult volunteering events. There was *no* difference in the staff-to-attendee ratio that may account for this finding, so Jay concludes that the youth volunteering events were more effective at registering new volunteers than adult events.

In 3, 6, and 9 months, he intends to follow up on his analysis to compare the *tenure* of the new adolescent and adult cohorts that registered and the percentage that are still active. He plans to perform a repeated measures ANOVA to assess the differences in tenure and a survival regression analysis to assess trends in continued volunteering activity.

Jay's complete analysis plan included a range of valuable descriptive methods (weekly trends), associations (correlations between staff-to-attendee ratio and registration rates), and experiments using a between-subjects and within-subjects comparison. Each approach provides the program team with recommendations on appropriately registering new volunteers and staff events and continuing to engage them over time.

### **3.2.5 Exercises**

Your Product Analytics team at the e-commerce company has completed the A/B test comparing three different website layouts implemented with the goal of *decreasing the rate of abandoned shopping carts without a purchase*.

1. What type of experiment and study design was used?
2. As part of your analysis, you can access the following demographic information about users on your website: geographic region, web browser type, time of visit, and number of previous visits. Which comparisons or cohorts might you include as part of your final deliverable?
3. One of the three website layouts performed better than the other two, which performed *worse* than the original website layout. What conclusions might you draw about this finding, and what action would you recommend?

4. The Product Management team informs you that many website layouts perform better or worse for short periods of time before reaching a true value when customers get used to them. How might you recommend adjusting your experiment to account for this phenomenon?

You can build on the decisions made in your previous exercises, writing recommended study designs and follow-ups that align with your original examples.

## 3.3 Types of Research Programs

We've covered a comprehensive range of methods for gathering evidence (descriptive, associative, experimental) and types of study designs (between-subjects, within-subjects, cohorts, longitudinal). Combining these two, you can appropriately design a study to answer almost any measurable research question.

As an analyst, you will likely evaluate successive or concurrent studies requiring more in-depth strategies to ensure success. A single team or organization will typically specialize in one or two types of *research programs*. We will discuss some of the most common programs in academic institutions and organizations: basic/applied research, A/B testing, and program evaluation.

### 3.3.1 Basic and Applied Research

**Basic research** refers to a program whose primary goal is to contribute to the overall available knowledge base in a specialized field. Studies are usually designed and conducted in succession, accumulating and advancing knowledge on the research area over a long period of time. This is a common approach in academic and other laboratory research, where a team sometimes dedicates their career to the topic of interest.

The value of a basic research program is usually measured by the impact of the accumulation of findings over time rather than an individual study. After a series of studies are published, the research team will use them to form the basis of a larger theory (e.g., the theory of evolution). Research teams will

contribute findings that support, refute, and augment the theory. Over time, a more sophisticated view of the research topic is developed and disseminated for a broader audience.

An *applied research program* seeks to collect and analyze data about a specific, targeted population to build direct knowledge about that population and influence how practitioners engage with them. Studies are designed to generate direct, actionable insights that translate to programs, products, and services tailored to the population from which the sample was drawn. This method is standard in community and psychological research, organizational settings, and non-profits.

While applied research programs don't have a primary goal of contributing to basic theories within the field of study, they will usually affect an accrual of knowledge about distinct subsets of a population (e.g., children in a geographic area, second-generation immigrant youth, persons with a specific disability).

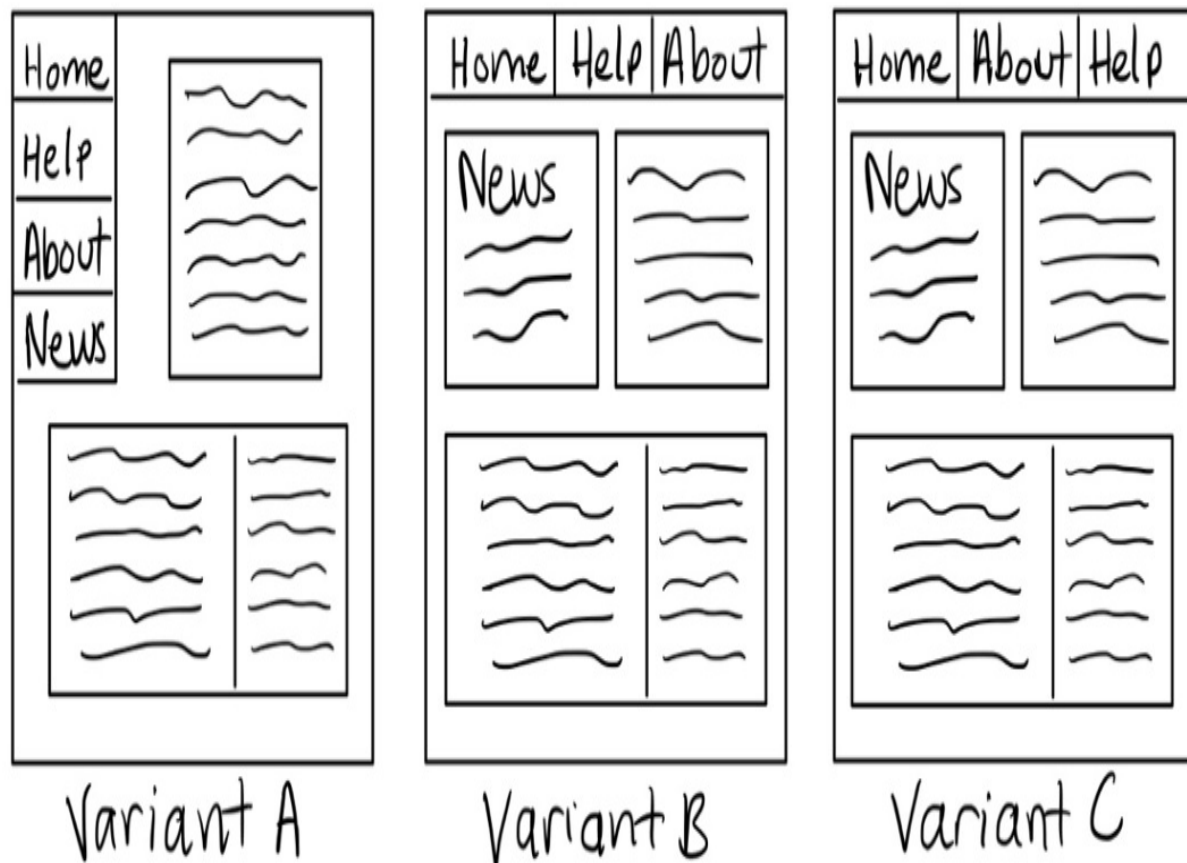
### 3.3.2 A/B Testing

In business settings, experiments are often designed as **A/B tests**. An A/B test is a type of experiment comparing two variations of a variable against each other to determine which performs better on key business metrics. Users are randomly assigned to a variant (e.g., group A, B, and C) for a duration. At the conclusion of the experiment, one or more statistical tests are used to compare the performance of the groups.

A successful A/B test typically examines the impact of small, modular changes to a website on conversions, visits, subscriptions, or time to complete a workflow. Each test is expected to have a limited scope of impact—a small increase in the critical business metrics or information gained about what types of changes *don't* have an effect. The actual value is in building and scaling an *A/B testing experimentation program* within an organization. When this type of program is mature, dozens to hundreds of experiments are run concurrently or in rapid succession to accumulate knowledge about how users interact with your product or service (basic research) while having a direct, measurable impact on how they use the product with each change

(applied research).

**Figure 3.14** Example of an A/B/C group comparison with minor changes to a website.



A/B testing leverages the same principles we have discussed up to this point in the chapter and the statistical tests and metrics we will discuss in depth later in the book. The laboratory of an A/B testing program is essentially a website or application, and the population of the testing program is the base of current and potential users.

### 3.3.3 Program Evaluation

**Program evaluation** is the most common strategy to assess the efficacy of non-profit, government, and many academic programs that liaise directly with institutions. These institutions design *programs* intending to meet a need

or provide a service within a specific population. Participants are assessed on outcome measures before, throughout, and several points after the program. Over time, data collected can enable an organization to systematically enhance its programs and their impact on the target population.

The goals of an evaluation are typically narrower than the previous two types of research programs discussed. A basic/applied research or A/B testing program will continually expand its areas of study as it generates findings about a topic. Program evaluation will often retain a specific focus over time (e.g., reducing the prevalence of a disease) as it improves its ability to achieve a goal cost-efficiently.

#### **Incorporating the Evaluation into the Broader Research Program**

The research Jay conducted to evaluate the efficacy of adolescent volunteering events is part of a larger series of *programs* at the non-profit. The organization's overarching goal is to raise money for cancer research and increase public awareness about new scientific discoveries and challenges associated with different forms of cancer. Multiple initiatives are run within the organization—volunteer engagement, fundraising, awareness campaigns, and more. The new adolescent program provides an additional component to evaluate, report on, and continually improve the efficacy of volunteer engagement.

In addition, the new adolescent volunteering initiative is beginning a new *program* for the organization—engaging adolescents in volunteering with the organization at fundraisers, charity events, walk/run events, and more. As part of this program, the organization will conduct evaluations of volunteers 3, 6, 12, and 24 months after their first volunteer engagement. The evaluations will include school performance, well-being, peer connections, family relationships, and more. This information can be leveraged for grant opportunities, peer-reviewed research, and more. The organization hypothesizes that adolescents volunteering with the organization will see improvement in school performance, well-being measures, and increased peer connections with other volunteers.

### **3.4 Summary**

- Developing a data-informed and quantifiable hypothesis involves synthesizing peer-reviewed and public research, gathering stakeholder information, and conducting foundational analyses of available data at your organization.
- **Research** the process of investigating a topic as a study or experiment to gain *new* information about that topic. The goal is to accumulate information over time, enabling you to build expertise and better inform decisions, practice, and policy.
- Data can be evaluated as one of the following:
  - **Descriptive information**, which seeks to describe *only* existing information in a dataset without making inferences, predictions, or quantifying the relationships between variables.
  - **Correlations**, which measure a non-causal linear association between two variables. Correlations range from -1 to 1, with values further from 0 indicating a stronger relationship.
  - **Causal/predictive analyses** aims to determine if one phenomenon (your independent variable) causes something else to happen (your dependent variable). This is usually tested as an **experiment** under controlled settings to be able to confidently isolate an independent variable as the cause of the dependent variable.
- **Experimental designs** comparing differences between groups are often set up as **between-subjects comparisons** (e.g., random assignment between an experiment and control group) or **within-subjects designs** (e.g., a single experiment group compared before and after treatment). These standard designs allow for comparison across vast arrays of research.
- **Cohort** comparisons (e.g., participants grouped by age and tracked for the duration of an experiment) and **longitudinal studies** (tracking participants over long periods) are less common due but highly valuable in many domains of study, research, and work.
- Individual study designs usually become part of more extensive research programs within an organization.
  - **Basic and applied research programs** seek to contribute to the overall knowledge base about a specific topic. These are common in academic settings.
  - **A/B testing programs** seek to continually improve on a product and/or business outcome through a continuous program of

concurrent or successive experiments run on large numbers of users for short periods of time. These programs are common in marketing and product analytics, and are primarily run in business settings.

- **Program evaluations** are typical in non-profit and government environments. These seek to evaluate the efficacy of new and ongoing programs for the populations they serve, continually improving their ability to achieve a specific goal (e.g., reducing rates of cancer).

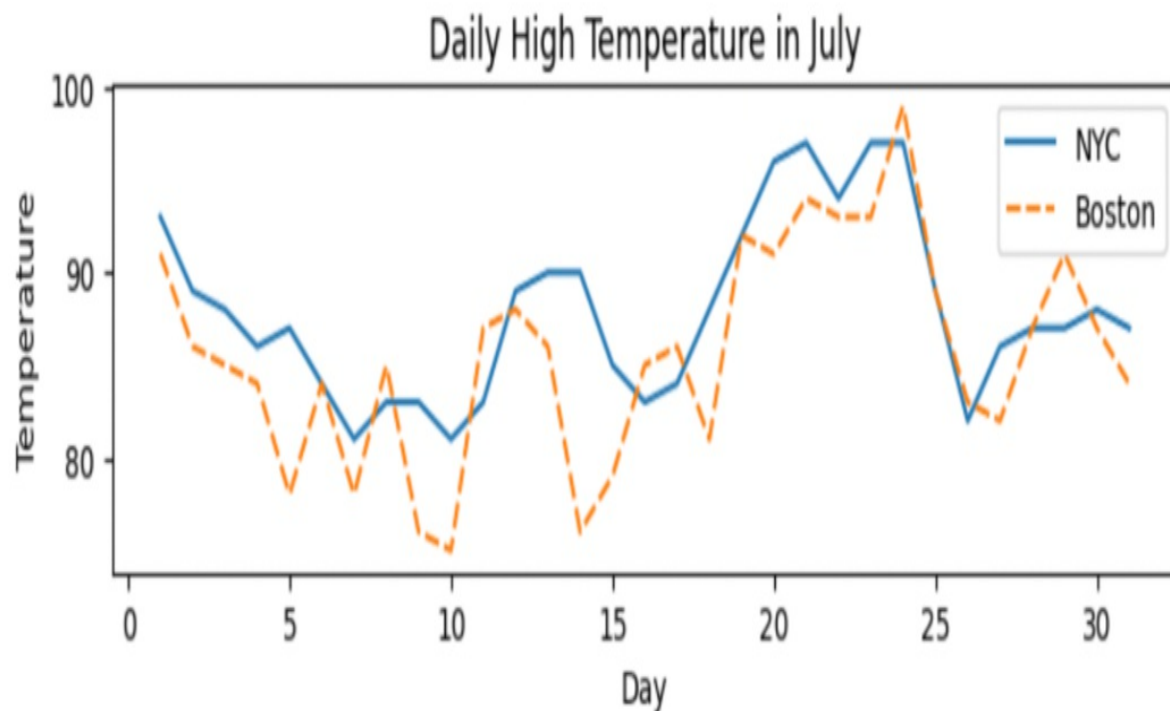
# 4 The Statistics You (Probably) Learned: T-Tests, ANOVAs, and Correlations

**This chapter covers**

- Breaking down summary statistics and their underlying logic
- Using parametric statistical tests appropriately
- Understanding and managing the limitations of parametric statistical tests

Take a look at the bar graph below comparing the daily high temperature over a month between New York City and Boston:

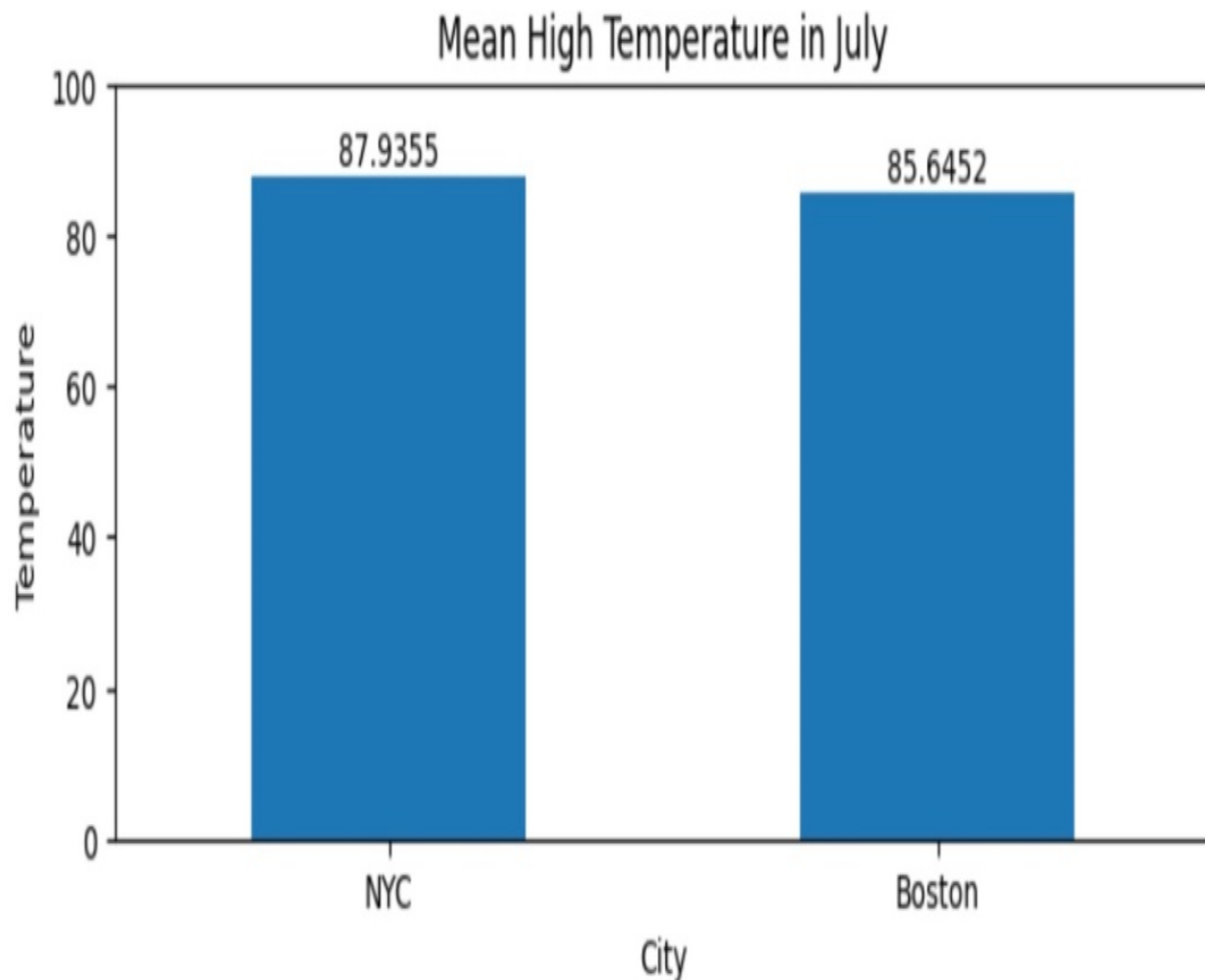
**Figure 4.1 Comparison of temperatures in July between New York City and Boston**



Can you determine which city is warmer in July? You can see that there's likely a relationship between the weather patterns of each city, which is a sensible hypothesis given the geographical proximity of New York City and Boston. However, there are clear day-to-day deviations in how the daily temperatures fluctuate, making it challenging to visually discern if one city has a higher temperature.

Take a look at an alternate view of the same data, which takes the mean of each daily high temperature per city and plots it on a bar graph:

**Figure 4.2 Comparison of the mean daily temperature between New York City and Boston**



You can see that the average temperature for New York City is slightly higher than in Boston, but is the difference in temperatures meaningful? How

do you know? How much of a difference indicates that one city is *meaningfully* warmer than the other? In all likelihood, if these questions were asked of multiple people, you would get a range of answers. This indicates that there is no agreed-upon threshold by which the *numerical* difference becomes meaningful.

Statistical tests can create rigor and alignment in the interpretation of numerical differences. There are common sets of methods used by most statisticians, social scientists, and analysts. Across a wide variety of domains of study and types of questions, practitioners use similar criteria to evaluate the coefficients of statistical tests that allow them to conclude whether or not they achieve statistical significance.

Despite these benefits, there are assumptions and limitations associated with common statistical tests and a troublesome history associated with their development and widespread use. We will cover the context and development of the most common statistical tests, coefficients, and evaluation criteria and break down the mathematical logic behind each approach. These skills will enable you to share highly accurate and actionable results with your stakeholders.

## 4.1 The Logic of Summary Statistics

You are likely aware that statistical tests are a *toolkit* for evaluating the characteristics of large quantities of data. Your dataset often represents only a subset (sample) of a broader population whose characteristics you want to *infer* in your work.

Before we decompose the logic of inferential statistics (e.g., t-tests, ANOVAs), we will review the core logic of measures of central tendency, the mathematical components in their equations, and the tradeoffs associated with reporting each measure. Let's continue with the example of daily high temperatures in New York City and Boston. We will begin by importing the raw dataset and displaying the list of daily high temperatures for New York City:

```
import pandas as pd      #A
weather = pd.read_csv("nyc_boston_weather.csv", index_col=0)    #B
```

```
print(list(weather.nyc))    #C
```

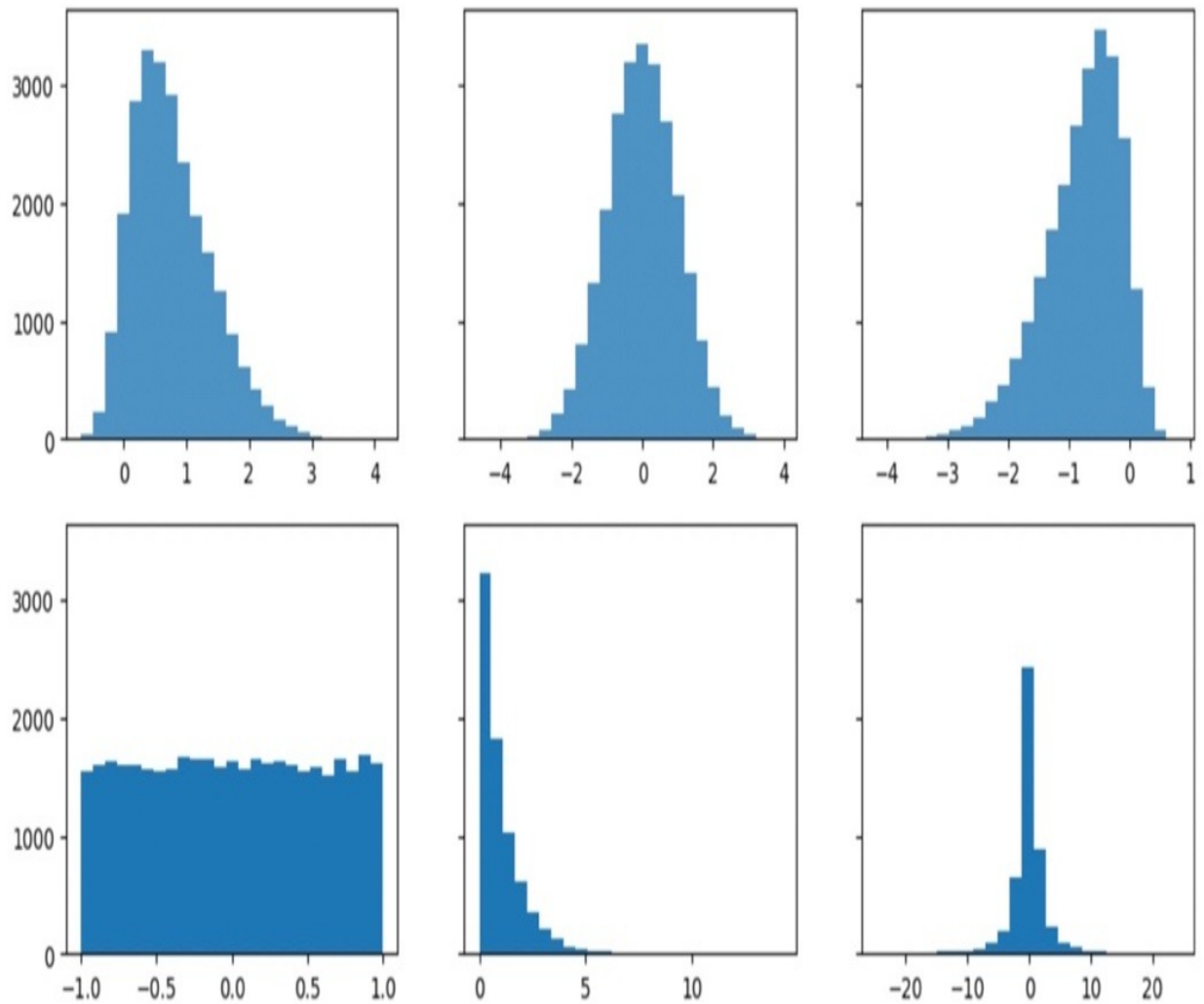
If you had only seen Figure 4.2 and not Figure 4.1, you would not know about the range, fluctuations, and heat waves depicted in the raw dataset. By visually inspecting the list of 31 values, you can see that an average of 87.9 degrees provides a limited view of the dataset. The temperature ranges between 81 degrees on the coolest day and 97 degrees on the hottest, and there are five sequential days where the temperature is in the high 90-degree range.

This is true for any method of summarizing a dataset: some dataset characteristics are highlighted, and some are lost.

### **4.1.1 Summarizing Properties of Your Data**

*Summary statistics* are single-value measures that describe a property of the *distribution* of a dataset. The mean, median, and mode are often referenced in introductory statistics courses but are by no means the only measure of value in your work.

**Figure 4.3 Summary statistics describe characteristics of the shape of a distribution.**



I recommend breaking down summary statistics into three categories in your evaluation and reporting:

- Measures of *central tendency* (e.g., the mean, median, and mode)
- Measures of *variability* (e.g., standard deviation)
- Measures of normality of a distribution (e.g., skewness, kurtosis)

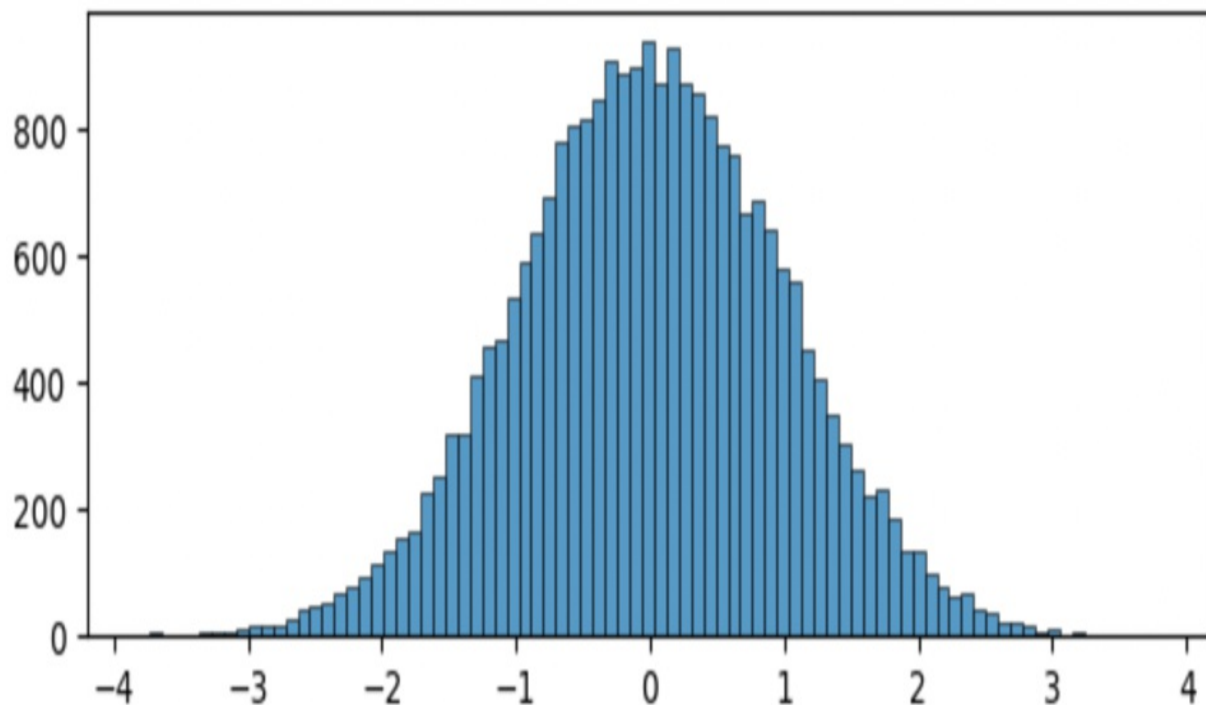
For this chapter, we will focus on the first two categories of statistics from the perspective of best use, logic, and limitations. Each of these is arguably necessary to evaluate before using inferential statistical tests. We will discuss the pros and cons of using measures as part of the *metrics* you report in chapter 6.

## Assumptions

Before we break down summary statistics, let's discuss the assumptions each category of summary statistics makes about the shape of your data. When your data does not meet the assumptions, your measures may not provide an accurate picture to your stakeholders. Some of these assumptions include:

- **Normality:** the assumption that your data roughly fits the shape of a bell curve (normal distribution).
- **Centrality:** where your data has a meaningful midpoint representing a “typical” data point.
- **Symmetry:** where your distribution has a similar number of data points to the left and right of the mean.

Figure 4.4 A standard normal distribution with a mean of 0 and standard deviation of 1



*Centrality* and *symmetry* are included in the *normality* assumption and exist as standalone assumptions for different measures. We will discuss the benefits and limitations of using each measure based on the distribution of your data.

## Measures of Central Tendency

Measures of central tendency are single-point measures of the “typical” records in your dataset. As the name suggests, these measures assume your data is clustered at a meaningful *center*. We will focus on the appropriate use of the most widely used measures: the mean, median, and mode.

The *arithmetic mean* is the most common and widely used statistic for summarizing numerical data. Analysts will often start their work by taking means of their data. Using this measure has a *lot* of benefits:

- The majority of stakeholders you collaborate with will be familiar with the *mean*.
- The mean calculation is relatively easy to explain to stakeholders unfamiliar with the metric.
- The use of the mean is widespread, so you will likely have benchmark comparisons available at your organization, in peer-reviewed literature, and in public data sources.

The arithmetic mean also has key assumptions and limitations:

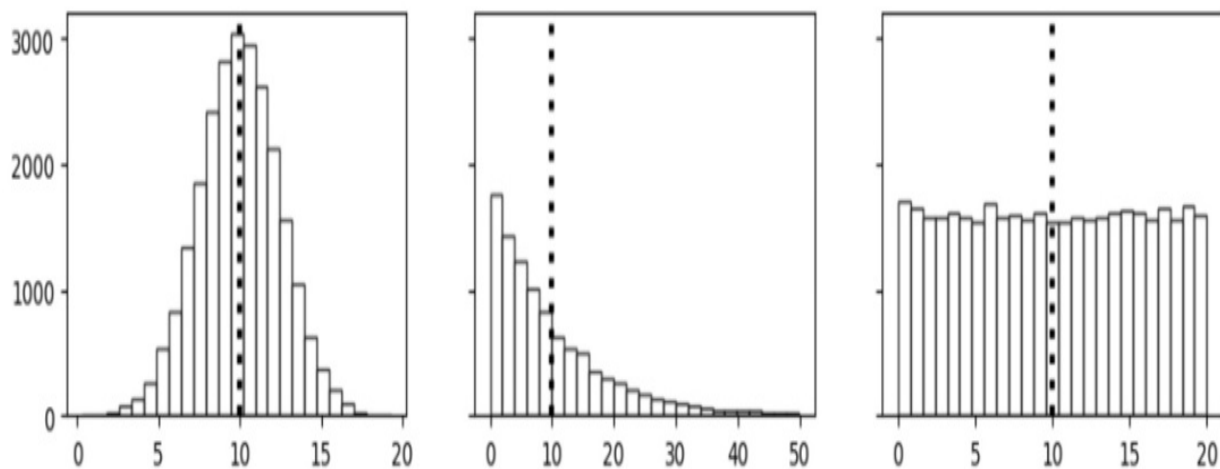
- Outliers and the skew of your distribution heavily impact the mean calculation.

**Figure 4.5** A mean calculation is highly sensitive to skewed data and outliers. The mean noticeably decreases when the highest outlier value of 97 is replaced with a value closer to the rest of the set.

$$\begin{array}{l} \{83, 97, 78, 81, 77\} = \frac{416}{5} = 83.2 \\ \updownarrow \\ \{83, 82, 78, 81, 77\} = \frac{401}{5} = 80.2 \end{array} \quad \begin{array}{l} \nwarrow \\ \nwarrow \end{array} \text{Mean}$$

- An appropriately representative mean calculation assumes that your data has a meaningful midpoint or *center*. The mean can mask differences in the shape of your distribution and interpretation of the *center*.
- In practice, the mean is often interpreted as your dataset's “typical” value. As an analyst, you will benefit from including interpretations of the shape of your distribution for your stakeholders to understand the summary statistic best.

**Figure 4.6 Three distinct distributions with an identical mean of 10. The interpretation of the mean or “average” is very different for each distribution.**



The *median* is simply the *midpoint* of a sorted series of data points. The median has several advantages over the mean:

- By definition, it represents the *midpoint* rather than a weighted calculation. It may be more appropriate to report for distributions without a meaningful center or symmetry (e.g., the second distribution in Figure 4.6).
- The median is also relatively well understood by many of your stakeholders.
- The median is more robust to skew and outliers than the mean. It can be a more appropriate representation of a “typical” record when a distribution is not symmetrical.

**Figure 4.7 The median is robust to skewed data and outliers. When the highest outlier value of 97 is replaced with a value closer to the rest of the set, the median remains the same.**

$$\begin{array}{l}
 \{83, 97, 78, 81, 77\} \rightarrow \{77, 78, \underline{81}, 83, 97\} = 81 \\
 \updownarrow \\
 \{83, 82, 78, 81, 77\} \rightarrow \{77, 78, \underline{81}, 82, 83\} = 81
 \end{array}$$

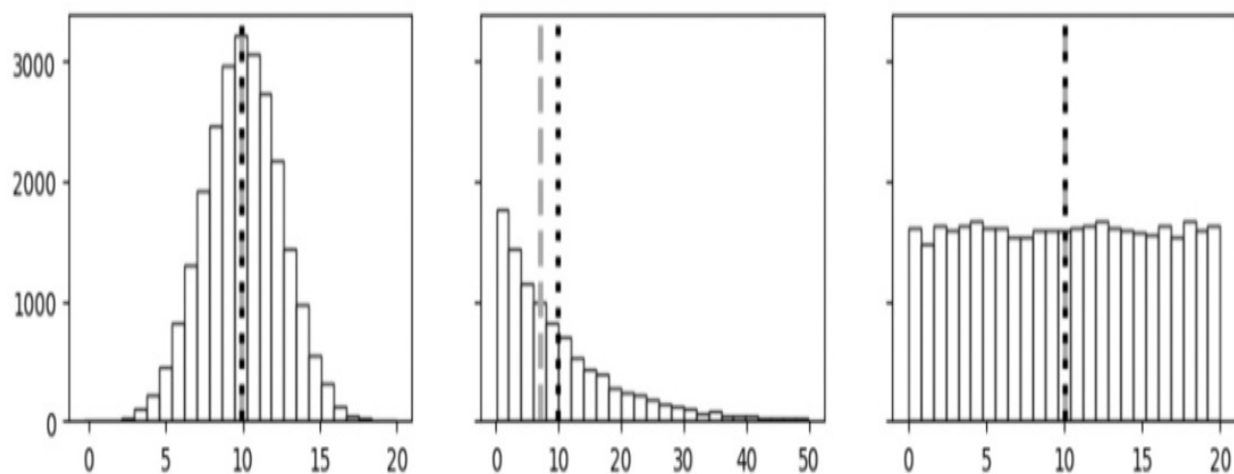
Median

As with the mean, there are key limitations to note about the median:

- The median can be more robust to change than the mean. If you compare changes in a median over time or between groups, you may be less likely to detect differences.
- Reporting both median and mean values can create confusion for your stakeholders. You may need to provide context and a justification for reporting each measure.

When preparing a report, you often choose between the mean and median to share with stakeholders based on the measure that provides the greatest clarity and value. Outside of direct stakeholder reporting, observing the differences between the mean and median indicates that your dataset is likely skewed or otherwise non-normal. You may want to note or correct the non-normality as part of your statistical analysis (see section 4.3).

**Figure 4.8** The mean (black dotted line) and median (gray dashed line) are approximately identical in the first and third distributions. The measures noticeably deviate in the second skewed distribution.



The *mode* is the most frequent value that occurs in your dataset. In practice, it is leveraged far less often than the mean and median and requires more context to appropriately explain its importance in reports.

**Figure 4.9** The mode is the most frequently occurring value in a dataset.

$$\{ 83, 74, 61, \underline{84}, 75, \underline{84}, 86, \underline{84}, 72 \} = 84$$

Mode

$$\{ \underline{75}, 64, 61, 76, \underline{72}, \underline{75}, 68, \underline{72}, 66, 71 \} = 72, 75$$

You may sometimes need to round or bin values to derive a meaningful mode. When testing a rounding calculation on a series with a large range or floating-point continuous data, your choice of bins or decimal point to round to can drastically change your outcome.

**Figure 4.10** Rounding floating-point values can highlight different mode values.

$$\{7.58, 6.44, 3.01, 8.19, 6.41, 5.22, 5.32\} = \text{N/A}$$

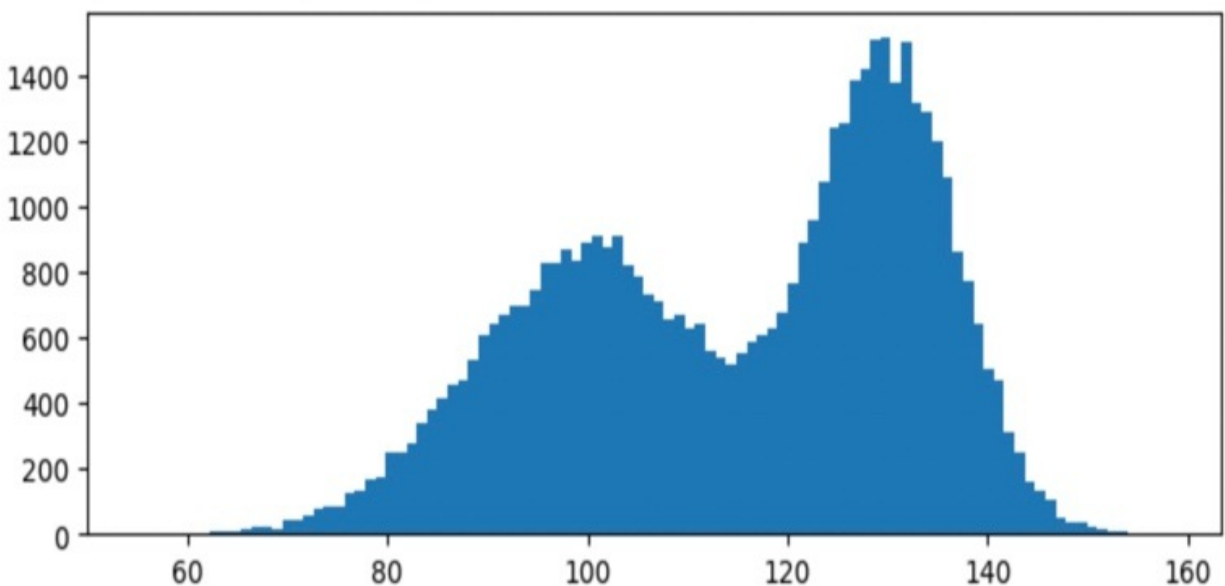
↓

$$\{7.6, \underline{6.4}, 3.0, 8.2, \underline{6.4}, 5.2, 5.3\} = 6.4$$

Mode ↗

Additionally, the mode is often helpful as a relative calculation to describe the shape of a distribution. A dataset may have many relative modes best discovered by observing the distribution. Taking counts to find the most frequent value gives you the *absolute* mode.

**Figure 4.11 A bimodal distribution has two modes best discovered by visual observation.**



You may go through years in your career without ever reporting a mode to your stakeholders. Though it's rarely used, I recommend considering the following conditions for where a mode is valuable to highlight:

- If a single value represents a vast proportion of the dataset
- If rounding continuous or floating-point values yields a meaningful set of bins for representing the data, or a mode with a substantial frequency
- If a distribution has multiple peaks with relative modes (*multimodal distribution*). These are often best discovered through visual observation of the distribution.

It's worth noting that when summarizing categorical data (e.g., counts of users in each city reported as a bar graph), you are reporting the dataset's *mode* (most frequent category). Representing this data type as a percentage/relative proportion of the categories instead of a count by group will be far more effective for your stakeholders to understand. We will expand more on representing this type of data in chapter 7.

**Let's introduce our case study for the chapter:**

Naomi is a research scientist at a pharmaceutical company. Her job includes data collection, analysis, and reporting for clinical trials of new experimental medications. The company regularly publishes its findings to government agencies, in public reports, and peer-reviewed papers in collaboration with academic teams.

Naomi is tasked with preparing an analysis to evaluate the efficacy of a new drug for treating insomnia in a randomized control trial that compared the new drug to a placebo. Participants were brought into the lab to monitor their sleep quality on three separate occasions throughout the trial. In total, 473 participants were in the experimental group (received the experimental drug), and 455 were in the control group (received the placebo). Participants did not know which group they were assigned to. The participants were monitored for their total sleep hours each night and the number of sleep interruptions.

For the first part of her analysis, Naomi will evaluate whether there are statistically significant differences on the *final* day of the sleep quality evaluation. She begins by calculating measures of central tendency for the dataset and generates histogram plots of each outcome measure broken out by the study group. She first creates the following summary table for participants in the experimental group:

### Hours of Sleep Interruptions

Mean	6.33	2.4
Median	7	2
Mode	7	2

The mean hours of sleep is lower than the median, whereas the reverse is true for the number of sleep interruptions. When Naomi creates a chart showing the distribution of both metrics, she discovers that Hours of Sleep is *negatively* skewed, with most participants reporting approximately 7-9 hours of sleep. She also finds that the number of sleep interruptions only ranges from 0 to 7, with most participants (52%) reporting one interruption.

Naomi begins the summary of her descriptive statistics for a paper to be submitted for peer review with the mean and median for both measures and the mode for the number of sleep interruptions.

### Measures of Variability

*Variability* is the degree to which your data diverges from the mean or median value. Measures of variability give you an estimate of the width of your dataset and insight into the representativeness of the mean or median. We will focus on measures in increasing order of complexity: the range, interquartile range, standard deviation, and standard error.

The *range* is the difference between a dataset's highest and lowest values. It's reported as the enumerated difference between the two values or a single value subtracting those values. In practice, it's often valuable to report both values together.

**Figure 4.12 The range depicts the entire width of the dataset.**

$$\{83, 92, 78, 81, 67\} \rightarrow \{67, 78, 81, 83, 92\}$$

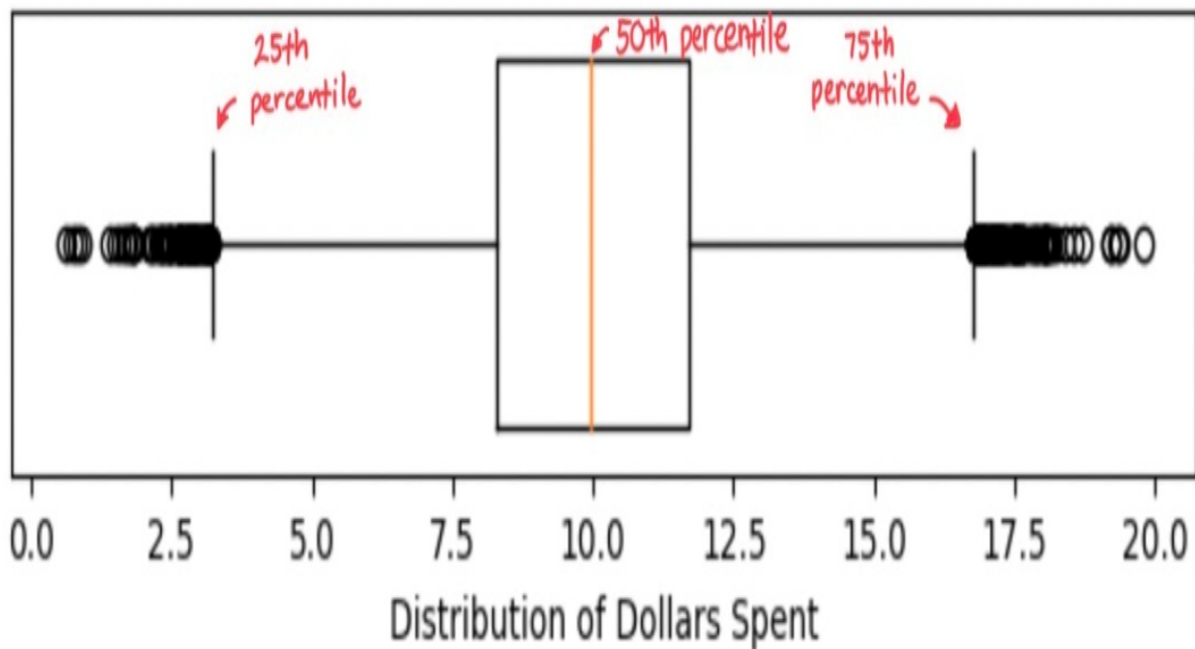
$$67 \text{ to } 92 \quad (\text{OR}) \quad 92 - 67 = 25$$

Temperatures ranged from 67 to 92,  
a difference of 25 degrees.

In addition to the full range, the interquartile range (IQR) shows the spread of the middle 50% of your data points from the 25<sup>th</sup> to the 75<sup>th</sup> percentile. This can be compared to the overall range to better describe the spread of your dataset between percentiles. With the median, range, and interquartile range, you can calculate the distance between any set of quartiles in your distribution.

In most cases, you will visually observe these ranges rather than just calculate and interpret the values. This is often done using a *boxplot* or *box and whisker plot*.

**Figure 4.13** A boxplot shows the median and interquartile ranges in the box and the 5<sup>th</sup>/95<sup>th</sup> percentiles in the whiskers by default. Values outside of the whiskers are typically treated as outliers.



Effectively communicating the results of a range calculation requires appropriate context to clarify its importance to your stakeholders in addition/in place of other measures. If you decide it's valuable to include in your findings, you can consider contextualizing it with statements such as the following:

- *The middle 50% of participants finished the 10k race between 43 and 67 minutes.*
- *Test scores ranged between 42% and 93%, with the median student receiving a 74%.*
- *50% of website visitors stay on the home page between 8 and 17 seconds.*

The second measure of variation we will discuss is the *standard deviation*. This measures the dispersal of your data from the mean, often defined as the *average distance from the mean*. The standard deviation is derived from the *variance* of a dataset by taking the square root. These two calculations serve a similar purpose for reporting purposes and will therefore be discussed together.

**Figure 4.14** The standard deviation essentially takes an average of the differences from the mean.

$$\sigma = \sqrt{\frac{\sum (x - \mu)^2}{N}} \rightarrow$$

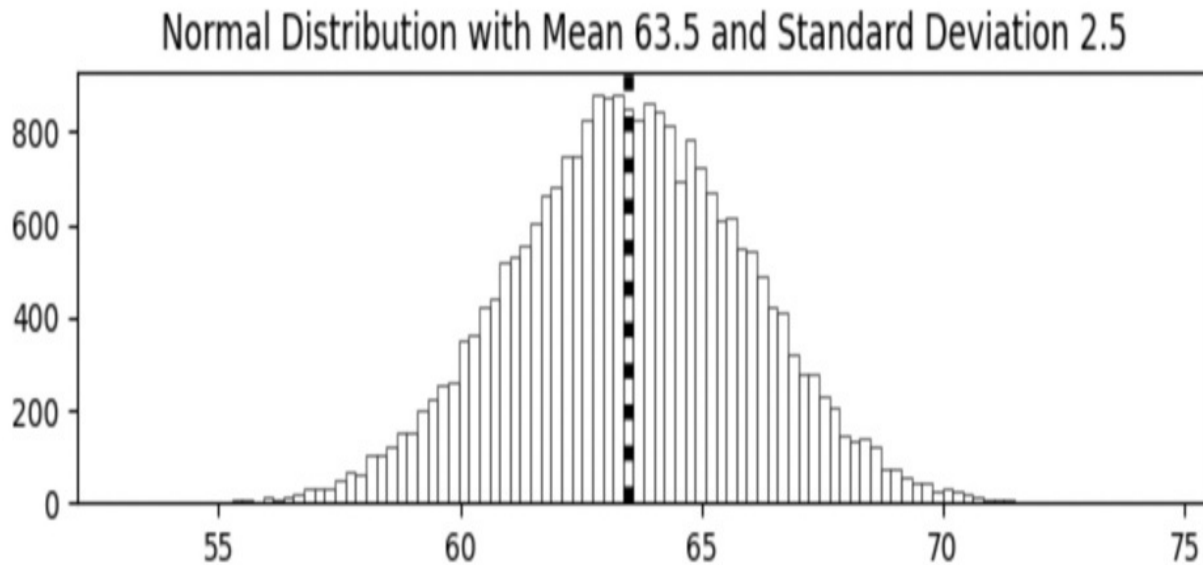
Data	Mean	Difference	Squared
83	83.2	-0.2	.04
97	↓	13.8	190.44
78		-5.2	27.04
81		-2.2	4.84
77		-6.2	38.44

sum = 260.8  $\rightarrow \sqrt{\frac{260.8}{5}} = 7.22$

If you have a mean and standard deviation and assume your data is normally distributed, you can easily approximate the shape of the dataset. Similar to the range and IQR, the standard deviation can be used as a coordinate system to estimate the proportions of data points between two values. To demonstrate, let's generate a *normal curve* representing the approximate distribution of heights (in inches) of men in the United States.

```
import numpy as np      #A
import matplotlib.pyplot as plt
import seaborn as sns
m, sd = 63.5, 2.5      #B
dist = np.random.normal(loc=m, scale=sd, size=25000)
sns.histplot(dist, bins = 100, color = "white")      #C
plt.axvline(np.mean(dist), color = "black", linestyle = "dotted")
plt.title(f"Normal Distribution with Mean {m} and Standard Deviat
```

**Figure 4.15** A normal distribution of heights (in inches) for men in the United States is easily generated if the mean and standard deviation are known.



In peer-reviewed papers and technical reports, the standard deviation and the mean are almost always included in the summary statistics. If you include the standard deviation in your reporting to less technical stakeholders, you will likely need to provide a layperson's explanation to minimize confusion.

Throughout my career, I have found the following explanations valuable in teaching statistics to undergraduate students and communicating with stakeholders:

- The standard deviation shows how much, on *average*, participants differ from the mean.
- The standard deviation estimates the most common range of data points you can expect to encounter above and below the mean.
- The standard deviation can be a reference point for how close the majority of data is to your mean: approximately 68% of data points are within one standard deviation from the mean, and 95% are within two standard deviations.

The final variance measure in this section is the *standard error of the mean (SEM or SE)*. The standard error estimates the distance between the sample mean, and the overall population mean. It's calculated by dividing the standard deviation by the square root of the total sample size. In this way, it differs from the previous measures by *inferring* a property about the broader

population rather than just describing the sample.

**Figure 4.16 Deriving the standard deviation and standard error from the initial variance calculation.**

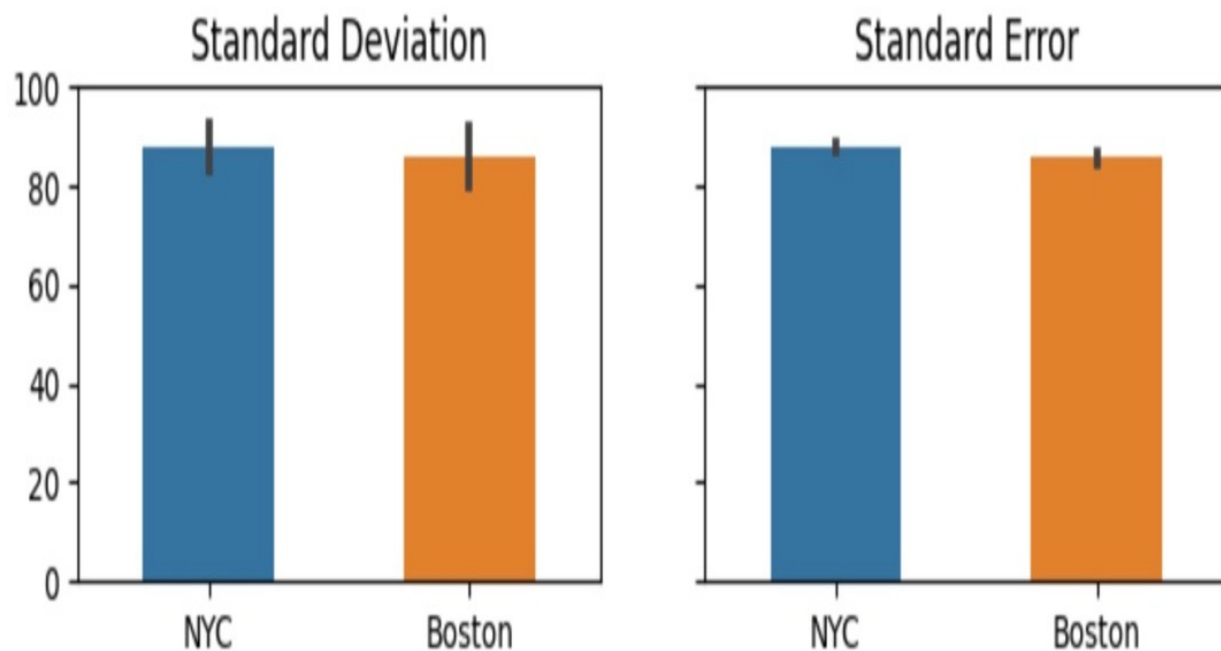
The diagram shows a sequence of three mathematical expressions connected by arrows. The first expression is  $\sigma^2 = \frac{\sum (x - \mu)^2}{N}$ , with a handwritten arrow pointing to it from the word "Variance". The second expression is  $\sigma = \sqrt{\sigma^2}$ , with a handwritten arrow pointing to it from the words "Standard deviation". The third expression is  $SE = \frac{\sigma}{\sqrt{n}}$ , with a handwritten arrow pointing to it from the words "Standard error".

$$\sigma^2 = \frac{\sum (x - \mu)^2}{N} \rightarrow \sigma = \sqrt{\sigma^2} \rightarrow SE = \frac{\sigma}{\sqrt{n}}$$

The standard error is a common choice for augmenting visualizations such as bar graphs to add context on variability within/between groups. It's an option in the seaborn barplot in Python and an easy addition in data visualization tools like Tableau.

You can often assume your audience will readily identify the error bars as a measure of variability. However, they may not be familiar with the underlying measures generating the error bars. You may benefit from clarifying the differences between a standard deviation, standard error, and confidence interval and your reason for choosing the specific measure in your deliverable.

**Figure 4.17 Bar graphs with error bars are very common visualizations but run the risk of misrepresenting the underlying data and creating confusion with stakeholders around the type and purpose of the error bars.**



I *strongly* caution against using this common type of visualization without first ensuring you meet the following assumptions and conditions:

- Your dataset is a *sample* from a larger population with a roughly *symmetrical* distribution. A bar graph with error bars will *not* depict a skewed distribution and may ultimately misrepresent the underlying shape of your data.
- The population your dataset is drawing from is not measurable or measured in its entirety for your analysis (e.g., the population of interest is all adults in the United States).
- The representativeness of your sample mean to the theoretical population mean is of value to your stakeholders to understand the deliverables you are creating.

If the above conditions are satisfied, the standard error of the mean is a great first indication of potentially detectable *statistically significant differences* between groups using an appropriate inferential statistical test.

### 4.1.2 Recap

If we synthesize the measures we have covered, we can answer questions

about the characteristics of our dataset, such as the following:

- What does the most typical data point look like (mean, median, mode)?
- How close to that “typical” data point are most records in the dataset (variance, standard deviation, interquartile range)?
- How wide is the entire or majority of the dataset (range, interquartile range)?
- How close to the true population mean is your sample mean (standard error)?

Each descriptive statistic you report has a tradeoff: some dataset properties are prioritized, and others are masked. Many descriptive (and inferential) statistics also have underlying assumptions about the shape and properties of your dataset that *must be checked and met before reporting on their values!*

I emphasize this as an analyst who understands that many of us don’t have the structures to enable us to apply statistical rigor to our work. So, I will leave you with some key takeaways about when to report on each summary statistic and when:

- Use the median to report on skewed or asymmetrical distributions. This measure will mask the impact of extreme outliers by prioritizing the *relative position* of data points.
- Use the mean or median with symmetrical data with a meaningful center.
- Use the mode to report on a distribution with a high concentration of values within a bin that the mode can represent. Include another measure of central tendency, such as the mean or median, in this reporting.
- Use the standard deviation when a dataset is relatively symmetrical.
- Suppose a dataset has no meaningful center (e.g., the third graph in Figure 4.8). In that case, you may want to describe the range, median, and interquartile range *and* include a visualization of the distribution for your stakeholders.
- Use the standard error if you assess your sample mean’s approximation of the true population mean or if you want to demonstrate statistically meaningful differences between groups (we will elaborate more on this in the next section).
- Check and report on all of the above before running statistical tests.

### 4.1.3 Activity

Run the following code in the Python environment of your choice (terminal, Jupyter Notebook, etc.). You will need to have numpy, matplotlib, and pandas installed.

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
dist = pd.Series(np.sqrt(np.random.exponential(1,75000)))
plt.hist(dist, bins = 100)
```

1. How would you describe the shape of this distribution?
2. What is the mean and median of the distribution? Which of these measures would you use to share with a stakeholder?
3. What is the mode of the distribution? How does it change when you round values to different numbers of decimal points? Is there a meaningful value you would consider reporting to stakeholders?
4. What is the standard deviation of the distribution? What does it tell you about how much it deviates from the mean? Can you determine if the distribution is symmetrical from this value?
5. Write a summary of the statistics values you have discovered so far. Based on the examples provided, you will refine the summary in the following sections.

## 4.2 Making Inferences: Group Comparisons

Until now, we have discussed the logic, usage, and assumptions of statistics used to describe a dataset and infer basic information about a sample's relationship to the population mean. In many cases, your work as an analyst will include drawing conclusions about the *significance* of relationships between variables or differences between groups using inferential statistics.

Most introductory statistics courses teach the same univariate, parametric methods of comparisons and options for testing significance. Many practitioners stop at this set of methods and repeatedly apply them to an incredible breadth of questions and fields of study.

**Figure 4.18** Statistical comparisons like correlations, t-tests, ANOVAs, and others are used *everywhere*. The example above is from a program evaluation I delivered to a non-profit in 2015.

Table 20

*Correlation matrix for demographic questions.*

	1	2	3	4	5	6	7	8	9	10
1	1									
2	-0.08	1								
3	0.62**	-0.21	1							
4	-0.46**	0.11	-0.72**	1						
5	0.19	-0.04	0.45**	-0.41**	1					
6	0.15	0.25	-0.02	-0.08	-0.04	1				
7	0.32*	0.02	0.21	-0.18	-0.01	0.28*	1			
8	0.58**	-0.38*	0.54**	-0.52**	0.39*	0.17	0.24	1		
9	0.14	0.28*	-0.18	0.11	-0.11	0.38**	0.06	-0.10	1	
10	0.11	0.16	0.12	0.04	-0.09	0.18	0.13	0.11	0.30*	1

\* p<.05

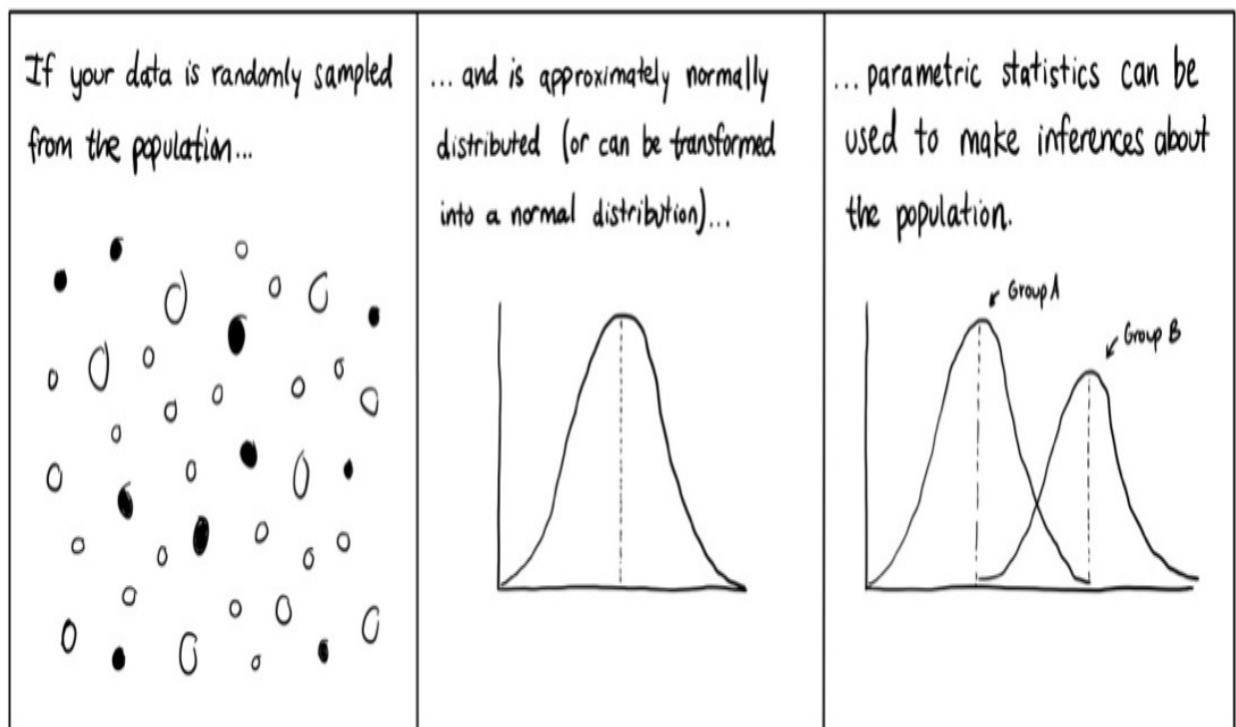
\*\* p<.01

While these tests aim to have a broad application, data professionals will frequently apply them without exploring alternative (non-parametric) statistical tests that may be a better fit for the data they work with. This section will discuss parametric tests, their limitations, and how to maximize the value of your inferences and conclusions. I recommend reading carefully if you're unsure what this section refers to. **The improper usage of parametric statistics can lead to patently wrong conclusions (e.g., identifying a group difference where there is none), spending countless hours and resources at an organization, and risking the reputation of the analytics function.**

### 4.2.1 Parametric Tests

The term *parametric* refers to inferring a value about a parameter (measurable value) of the population. From that definition stems *parametric statistics*, the branch of statistics inferring fixed parameters about a population. In other words, these statistical tests assume that true population data fit the specific shape of a probability distribution, can be modeled as such, and can be estimated based on a *representative sample* of data from that population.

**Figure 4.19** Parametric statistics assume that the population data follows a specific probability distribution and that you can make inferences about the parameters of that distribution based on your sample data.



Parametric statistical tests are ever-present in analytics. If you took an introductory statistics course in an undergraduate or graduate program, you likely covered a range of *univariate* approaches designed to evaluate one dependent variable per test. Many of the following tests may be familiar to you:

- The *t-test* is used to identify differences between the means of two groups. Comparisons can be between groups (independent samples) or

within groups, typically comparing values before and after a change or intervention (paired samples).

- The ANOVA (analysis of variance) is used to identify differences between the means of two or more groups. Unlike a t-test, an ANOVA can include multiple groups per independent variable *and* multiple factors (e.g., a two-way ANOVA has two factors).
- Pearson's correlation is used to identify linear relationships between two continuous variables. Unlike the previous methods, the coefficient (*r-value*) is standardized and can be directly interpreted for the strength and direction of the relationship.
- Linear and logistic regression are predictive models used to measure the relationship between a dependent variable (continuous and categorical, respectively) and one or more independent variables.

We will elaborate more on correlation and regression methods in a later section. However, this section's assumptions and interpretation of parametric statistics apply to these methods and should be considered foundational to the next topic.

## Assumptions

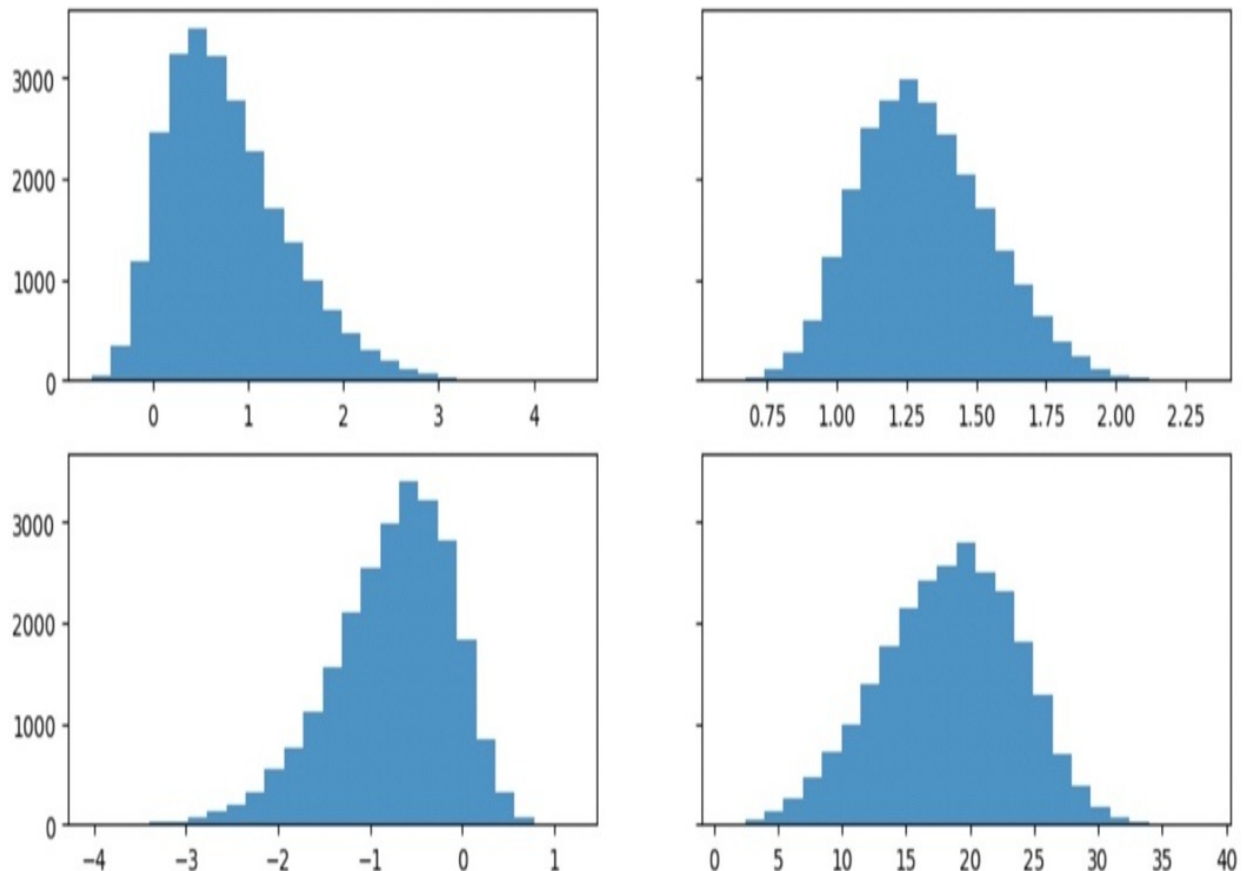
In addition to the assumptions of the measures of central tendency and variation discussed in the previous section, parametric statistics have strict assumptions about the shape of the data within and between groups. Meeting these assumptions is *necessary* for making accurate inferences about your data.

The first assumption of parametric statistics is that the data is shaped according to a distribution the underlying population is believed to follow. In the majority of tests that we'll cover in this chapter, the underlying population is believed to follow a *normal distribution* (the assumption of *normality*). To meet this assumption, your data either needs to be *approximately normally distributed* or *capable of being transformed into a normal distribution*. This process is also called *normalizing* your data. It can be done via a number of mathematical transformations to the entire dataset, resulting in a reshaping of the distribution. For example, we can transform a positively skewed distribution by taking the square root of all of the data

series' values and transform a negatively skewed distribution by squaring the data.

```
from scipy.stats import skewnorm      #A
import matplotlib.pyplot as plt
positive_skew = skewnorm.rvs(4, size = 25000)      #B
negative_skew = skewnorm.rvs(-3, size = 25000)
fig, ax = plt.subplots(2, 2, sharey = True)      #C
ax[0][0].hist(positive_skew, bins = 25)
ax[0][1].hist(np.sqrt(positive_skew+1), bins = 25)
ax[1][0].hist(negative_skew, bins = 25)
ax[1][1].hist((negative_skew+5)**2, bins = 25)
```

**Figure 4.20** Skewed data can often be transformed into an approximately normal distribution.



Not every distribution can be effectively normalized for analysis with parametric statistical tests. Many *data types*, such as categorical and discrete count data, are not appropriate for numerical transformation. Some distributions won't transform into the desired shape if manipulated, even

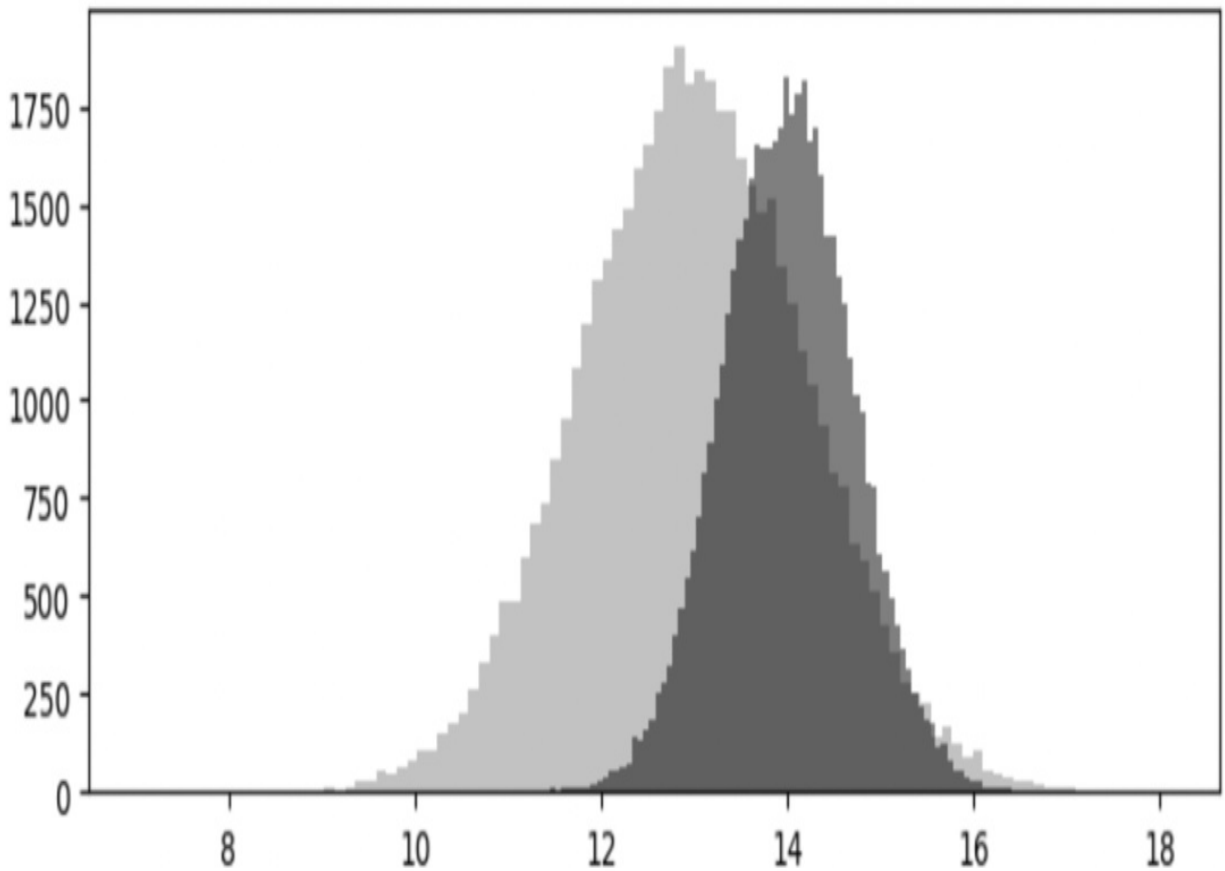
when suitable data types are used. When your data is uniformly distributed, extremely skewed with significant outliers, or is multimodal (has more than one mode), you will likely need to use a non-parametric statistical test to evaluate it.

The second assumption is the *independence* of data points in your dataset. Unless otherwise indicated with the statistical test you use (e.g., a repeated measures t-test or ANOVA), the probability of events in your dataset are assumed *not* to impact the probability of other events. In practice, a lack of independence of data points might look like one of the following situations:

- Participants in a laboratory changed their answers on a survey after learning how their peers answered the same questions.
- Participants in an A/B test are randomized, but users within the same company compare and notice their user interfaces look different.
- Participants in the control group of an intervention notice that experimental group participants are experiencing more positive outcomes.

The third related assumption is the *equality of variances between groups* (also known as homoscedasticity). Parametric tests assume that the population(s) your samples are drawn from vary equally on your outcome measure of interest. Tests such as the t-test and ANOVA include the standard deviation (square root of the variance) in the denominator of the calculation; if one of the groups has a much higher variance, the calculation will be skewed, and results will be unreliable.

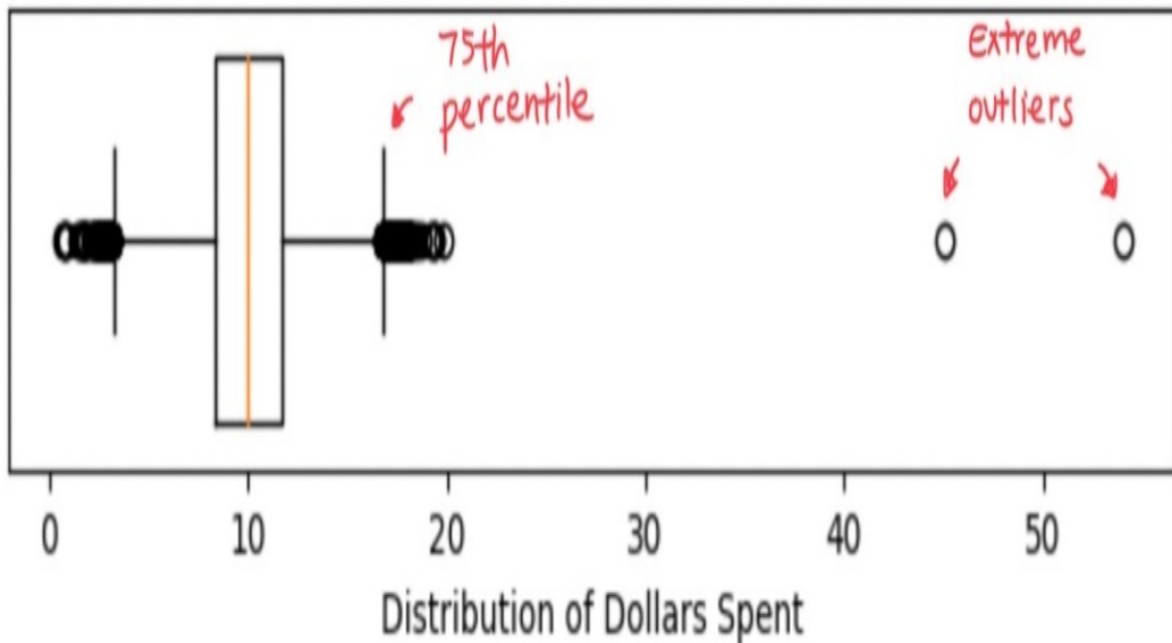
**Figure 4.21 Parametric statistics generally assume that your samples have equal variances. In this example, the unequal variance leads to greater overlap between the two distributions.**



If you determine that your samples have unequal variances, you can use adjusted versions of t-tests and ANOVAs (Welch's tests) that are more robust to violations of this assumption. Non-parametric tests for group comparisons may also be better choices for your work.

The fourth explicit assumption is the absence of *numerical outliers*. Parametric tests assume that your dataset lacks extreme outliers, and failing to correct them can significantly impact the accuracy of your results. In most cases, numerical outliers can be easily identified through visual dataset observation.

**Figure 4.22** Extreme outliers are often easily detected by generating a boxplot or a histogram of your data.



It's recommended to take one of the following steps to handle outliers in your dataset:

- Systematically removing the values: this can be accomplished by taking only a limited range of data around the median (e.g., the interquartile range). It is *not* recommended to drop an individual value – **that can quickly turn into p-hacking, which we will discuss in section 4.3!**
- Transforming your data: if your dataset is skewed and contains outliers, you can attempt one of the transformation methods shown in Figure 4.20 to correct for the extreme values.

To put all of these assumptions together, parametric statistics require the use of sample data with distributions that have the following characteristics:

- Distributed according to the parameters of the underlying distribution (e.g., normal distribution) or can be transformed into the distribution assumed by the test
- Have measures where events/participants do not impact each other's results
- Are of the same or similar width and shape
- Do not have individual or small clusters of data points that have an

extreme numerical deviation from the mean and median

If you have taken a statistics course in an undergraduate or graduate curriculum, you likely covered these topics as part of your education. So why are we spending so much time covering things you may already know?

In my analytics career, I've seen that these steps are often neglected in the application of parametric statistical tests. It's common to apply a t-test or ANOVA to your data without first making the necessary checks and quickly drawing conclusions about the significance/non-significance of the results. In practice, we are often limited in time and capacity and have stakeholders who don't have the statistical knowledge to inspect our work in detail.

For the sake of the accuracy of your results and the long-term accrual of accurate information at your organization, please, do not neglect these steps. You run a genuine risk of your results and conclusions being completely wrong. If the time and diligence to appropriately apply parametric statistics is not feasible in your workflow, I *strongly* recommend using non-parametric statistics instead.

## **Coefficients and Statistical Significance**

Statistical test calculations provide a *coefficient* or a numerical value for interpreting the strength and direction of the relationship between your groups or variables. Coefficients differ based on the statistical test used but are generally standardized values that can be used to evaluate your results against each other and a contingency table.

Let's use the t-value from a t-test as an example. The t-value represents the difference between the means of two samples (independent or repeated measures) or between a sample mean and hypothesized value (one-sample t-test). A larger t-value indicates a larger difference between groups.

In most cases, a coefficient's numerical value is insufficient to determine if your results support your hypothesis. Coefficients can be compared *against* each other within the same test (e.g., multiple t-values from different t-tests). Still, they cannot be compared against other coefficients (e.g., a t-value vs. an

F-value in an ANOVA) and, on their own, provide limited information about whether the differences between your groups are statistically meaningful.

Appropriate interpretation of coefficients requires two additional pieces of information:

- The *degrees of freedom* and p-value threshold, which is one less than your sample size (e.g., if you have 200 data points, your degrees of freedom is 199).
- An appropriate *p-value* as a critical threshold. This is also known as the alpha level.

With this information, you can evaluate whether your results are statistically significant. Likely, you are already familiar with this process if you are an analyst – the t-test evaluation is covered fairly early in undergraduate statistics coursework, and degrees of freedom and the p-value are ubiquitous in our work. However, there are clear limitations with these approaches and situations where the validity of parametric tests falls apart.

Yes – even if you check and meet all of your assumptions for using a parametric test, you can *still* generate superfluous results if your sample size is inappropriate for the test being used. Let's demonstrate these limitations with a t-distribution table for a two-tailed t-test:

**Table 4.1 Abbreviated t-distribution table showing that increasing the degrees of freedom has diminishing returns on the t-critical values at each alpha level/p-value threshold.**

alpha level	0.1	0.05	0.01	0.005	0.001
degrees of freedom					
10	1.81	2.23	3.17	3.58	4.59
20	1.72	2.09	2.85	3.15	3.85

30	1.70	2.04	2.75	3.03	3.65
40	1.68	2.02	2.70	2.97	3.55
50	1.68	2.01	2.68	2.94	3.50
60	1.67	2.00	2.66	2.91	3.46
70	1.67	1.99	2.65	2.90	3.44
80	1.66	1.99	2.64	2.89	3.42
90	1.66	1.99	2.63	2.88	3.40
100	1.66	1.98	2.63	2.87	3.39
150	1.66	1.98	2.61	2.85	3.36
200	1.65	1.97	2.60	2.84	3.34

The degrees of freedom listed in the first column represent the group sample sizes ( $n_1 + n_2 - 2$ ). As the sample size increases, the critical t-value for statistical significance at each p-value (listed in the columns) decreases. You'll notice that the decrease is exponential, reaching a point of diminishing returns after an n of around 50 to 100. However, the t-value formula does *not*

have a similar point of diminishing returns and will continue to increase with your sample size in accordance with its formula.

**Figure 4.23 Formula for an independent samples t-test**

$$t = \frac{X_1 - X_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

sample means

sample standard deviations, squared

sample sizes

When the sample sizes  $n_1$  and  $n_2$  increase, the size of the overall t-value increases with no other changes to the mean or standard deviations. Let's take two samples with the following summary information:

**Table 4.2 Sample test score data for two groups of students.**

	Group 1	Group 2
Mean	80.4	79.9
Standard Deviation	4	3.8

Sample Size	45	44
-------------	----	----

If you calculate the t-value for these two groups, your t-value is far below the critical threshold at the current sample size.

**Figure 4.24 The two groups have a non-significant difference.**

$$t = \frac{80.4 - 79.9}{\sqrt{\frac{4^2}{45} + \frac{3.8^2}{44}}} = 0.6047$$

X not significant

If you double the sample size for each group to 90 and 88, respectively, you get the following result:

**Figure 4.25 The two groups still have a non-significant difference, but the t-value is larger.**

$$t = \frac{80.4 - 79.9}{\sqrt{\frac{4^2}{90} + \frac{3.8^2}{88}}} = 0.8551$$

X not significant

If you increase the sample size again by ten times the original number of participants, the t-value increases considerably and far exceeds the critical t-

value.

**Figure 4.26 Increasing the sample size by a factor of 10 yields a statistically significant result.**

$$t = \frac{80.4 - 79.9}{\sqrt{\frac{4^2}{900} + \frac{3.8^2}{880}}} = 2.7042$$

✓ statistically significant

As analysts in the age of big data, we frequently work with datasets substantially larger than in previous decades. Collecting data from participants in academic settings is time-consuming and costly, which leads the majority of researchers to moderately constrain their sample sizes (e.g., 100-200 participants). Data is often highly available and extremely cheap to capture in fields such as marketing or product analytics. It's increasingly common to access large data samples over extended periods and compute statistics on thousands or millions of records. When running parametric statistical tests, such large sample sizes can yield significant differences even when the group means being compared are nearly identical. The recommendations made from these results are unlikely to be valuable or actionable.

There are some steps you can take to correct for issues with datasets whose magnitudes exceed a few hundred records:

- Increase your significance threshold from .05 to .01 or .001.
- Use effect size measures such as Cohen's d to measure the magnitude of differences between your group means. These calculations are not impacted by sample size.
- Set a minimum threshold of difference between group means that is meaningful based on your domain knowledge (e.g., student test scores with an average difference of 0.5% is likely, not meaningful) and the implications of the differences (e.g., how much revenue does a 0.2%

increase in conversion rate mean for your organization). The easiest way to do this is to use the confidence interval to compare the true difference between means to the desired value.

**Let's return to our case study for the chapter:**

Naomi is preparing her results for analysis. She has the following summary information about her primary measure of interest in the drug trial:

Hours of Sleep	Experimental	Control
Mean	6.54	6.11
Std. Deviation	1.7	1.8
Sample Size	473	455

The distribution of *hours of sleep* is normally distributed, with no extreme outliers. The two groups also have approximately equal variances. Since this dataset meets all of the assumptions of parametric statistical tests, Naomi elects to use an independent samples t-test to determine whether the differences between the two groups are statistically meaningful. She sets an alpha-level threshold of .001 because of her large sample size. Her criteria for statistical significance must be appropriate due to the implications of reporting inaccurate results on a trial for a new medication.

The results yield a t-value of 3.738. With 926 degrees of freedom, she concludes that her results are statistically significant.

In this book, we've discussed the concept of statistical significance, alpha levels, and p-values at length. The p-value is a universal tool in applying inferential statistics, and you're likely familiar with interpreting p-values of your statistical tests. However, providing a layperson's explanation of the value and its application can be challenging. It's not particularly intuitive. The first introduction to the p-value and its meaning in nearly every undergraduate statistics course I taught led to a classroom of confused faces.

The p-value is a value between 0 and 1 representing the probability of your

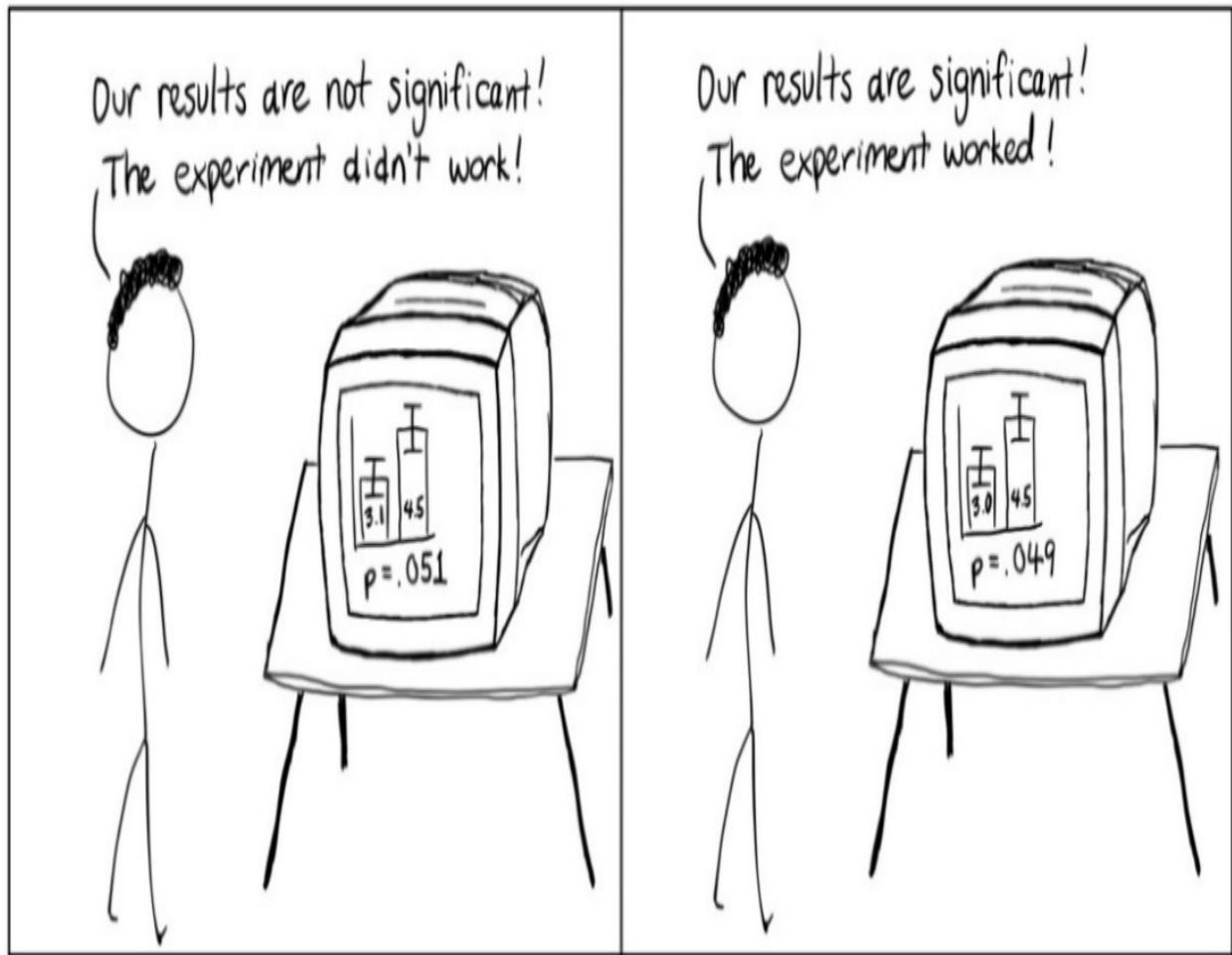
test coefficient (e.g., t-value, F-value) being the same or larger than your test yielded, assuming that your null hypothesis is true. The smaller the p-value, the more substantial the evidence that your null hypothesis is false. In less technical jargon, assuming there is *no true difference between the experiment and control group in the population*, a p-value of .05 indicates a 5% probability you would see the observed magnitude of differences between your sample means. The smaller the p-value, the less your null hypothesis will likely be true.

The p-value is **NOT** defined as any of the following, though you may frequently encounter these interpretations:

- The probability that your results occurred due to chance
- The probability that your alternative hypothesis is true
- A static, universal threshold where all values above .05 are not significant and all values below .05 are significant

Like some statistical tests and concepts discussed in this section, the p-value's development and widespread application have a rocky history. Pearson developed the concept in the early 20<sup>th</sup> century to mitigate the need to manually compare your test statistic to a critical value (see Table 4.1). The test was popularized in the 1950s by Fisher with the recommended .05 threshold commonly used today.

**Figure 4.27 The p-value is often treated as a magical boundary that unlocks findings considered worthy of peer-reviewed publication.**



Using the p-value as an immutable threshold constrains the quality of an analyst's work. There's rarely a meaningful difference between a p-value slightly above or below your chosen alpha level. It's often easier to use a rigid interpretation of findings with a less restricted alpha level and present potentially erroneous results. Regardless, it can sometimes be challenging to present findings in some contexts with a flexible interpretation of the p-value (e.g., peer-reviewed articles, program evaluations) and have them perceived as legitimate.

If you find yourself in a situation where you are expected to use the broadly accepted interpretation of a p-value, I recommend the following steps to maximize the quality of your deliverable:

- Set your alpha level intentionally at the start of any experiment alongside your hypothesis generation based on the following:

- Your field of study or work (an experiment on user behavior on a website will usually have less restrictive criteria than a medical trial)
  - The number of groups and interaction effects in your study design (more groups and interactions produce a higher likelihood of false positive results)
  - The implications of getting it wrong and reporting false positive results (recommending a website design vs. recommending a new type of therapy or educational intervention)
- The degree of control over your experiment (a highly controlled laboratory setting can potentially limit the number of confounding effects, allowing you to set more conservative thresholds than studies in real-world settings)
- Check or re-check all of the assumptions of your test. If you are unclear whether certain assumptions are met, consider running tests (e.g., Welch's test for equality of variance) to validate your visual observations.
- Determine the appropriate *minimum* sample size to detect an effect using an *a priori power analysis*. Many free sample size calculators are available online, and it's also possible to do so in most statistical software. With the limitations of sample-size sensitive parametric tests in mind, set a goal of collecting more than the minimum. For example, the following code determines the minimum sample size necessary to detect a small effect size of 0.3 at 80% power (the most commonly used threshold), an alpha level of .05, and with four groups being compared.

```
from statsmodels.stats.power import FTestAnovaPower    #A
pwr = FTestAnovaPower()
sample = pwr.solve_power(effect_size = 0.3,
    power = 0.8,
    alpha = 0.05,
    k_groups = 4)    #B
print(sample)
```

- If your p-value is slightly above the alpha level, consider collecting additional data with a **fixed sample size** to determine if the gap between your test coefficient and the critical value can be reduced or eliminated. Do *not* just collect data until you reach your desired threshold. That's one method of p-hacking, which we will discuss later in this chapter.
- Leverage and report on effect size measures such as Cohen's *d* alongside

your measure of statistical significance to provide a robust picture of the magnitude of your results.

In general, marketing and product analytics units in business have opportunities to be flexible with their interpretations of statistical significance. If you can set a margin of error and apply qualitative judgment to results, I recommend many of the same steps: set your margin of error intentionally alongside your alpha level, collect an appropriately-sized sample, and report on effect sizes.

## 4.2.2 Activity

The following code performs an *a priori power analysis* to determine the minimum sample size necessary to detect a medium-sized effect (effect\_size = 0.5) in a t-test at 80% power (power = 0.8). These two parameters are common defaults in an a priori test.

Run the code in the Python environment of your choice (terminal, Jupyter Notebook, etc.). You will need statsmodels installed for this step and numpy and scipy for the rest of this activity.

```
from statsmodels.stats.power import TTestIndPower    #A

pwr = TTestIndPower()    #B
sample = pwr.solve_power(effect_size = 0.5,
    power = 0.8,
    alpha = 0.05)
print(sample)
```

1. What is the minimum recommended sample size for a t-test? How does the value change when you adjust the alpha level to .01? .001?
2. Run an independent samples t-test using the two normally distributed samples of data generated with the following code. Replace the value of 0 for n with the recommended sample size you just calculated for alpha = 0.05 (divide the value by 2, as the test recommends a *total* sample size). Are the results statistically significant at the p=.05 threshold?

```
import numpy as np
from scipy import stats as st    #A
```

```

n = 0
mu, sigma = 75.5, 6.2
mu2, sigma2 = 77.9, 6.5
X1 = np.random.normal(mu, sigma, n)
X2 = np.random.normal(mu2, sigma2, n)    #B

result = st.ttest_ind(X1, X2)
print(result)    #C

```

3. Replace the value of  $n$  with the recommended sample size at  $\alpha = 0.01$  (don't forget to divide the value by 2). Is the result statistically significant at the  $p=0.01$  threshold?
4. Summarize the changes you saw between each t-test conducted with different sample sizes. Why did the t-value and p-value change the way they did?
5. Note how the t-test results change with each alpha and sample size adjustment.

## 4.3 Making Inferences: Correlation and Regression

A correlation is a measure of the relationship between two variables. It's often one of the first steps taken to identify patterns in a dataset and establish an association between variables later examined for potential causal relationships in a regression model. A thorough understanding of correlation and regression is foundational to advanced statistics, predictive modeling, and machine learning.

### 4.3.1 Correlation Coefficients

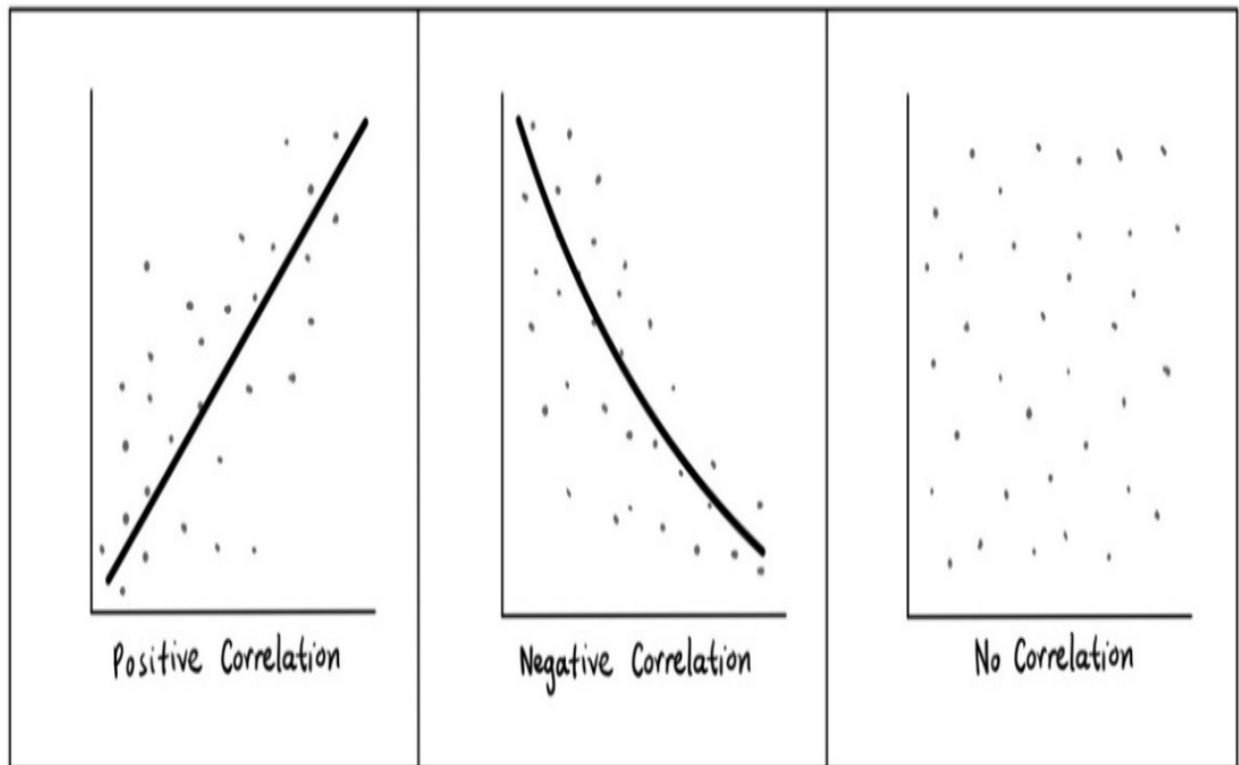
There are several types of correlation coefficients you can use to evaluate relationships between two variables:

- **Pearson's correlation** measures linear relationships between two continuous variables. It's the most commonly used among correlational methods. To effectively leverage this coefficient, your data must meet the assumptions of other parametric statistics and represent a *linear* trend. If data is *not* checked for linearity, your coefficient can indicate a far weaker relationship than actually exists.

- **Spearman's correlation** is a non-parametric statistic that compares the *ranked position* of each data point between two variables. It's often used for ordinal data and variables with non-linear relationships. We will discuss this method in Chapter 5.
- **Kendall's rank correlation** or Kendall's tau is a measure of ordinal association between data points calculated by measuring the number of pairs with identical and disparate ranks. It's used less often than Spearman's correlation but can better identify some ordinal relationships. We will discuss this method in Chapter 5.
- **Point-biserial correlation** is a special type of Pearson's correlation used to measure associations between one binary variable and one continuous variable. It's calculated by measuring the difference between the two group means for the continuous variable. It is one of several available coefficients for measuring associations between a binary and continuous variable.

All of these coefficients benefit from using the same standardized scale; values range from -1 to 1, with values closer to 1 or -1 indicating a *stronger* relationship, the +/- sign indicating the *direction* of the relationship, and values closer to 0 indicating a weak to no relationship.

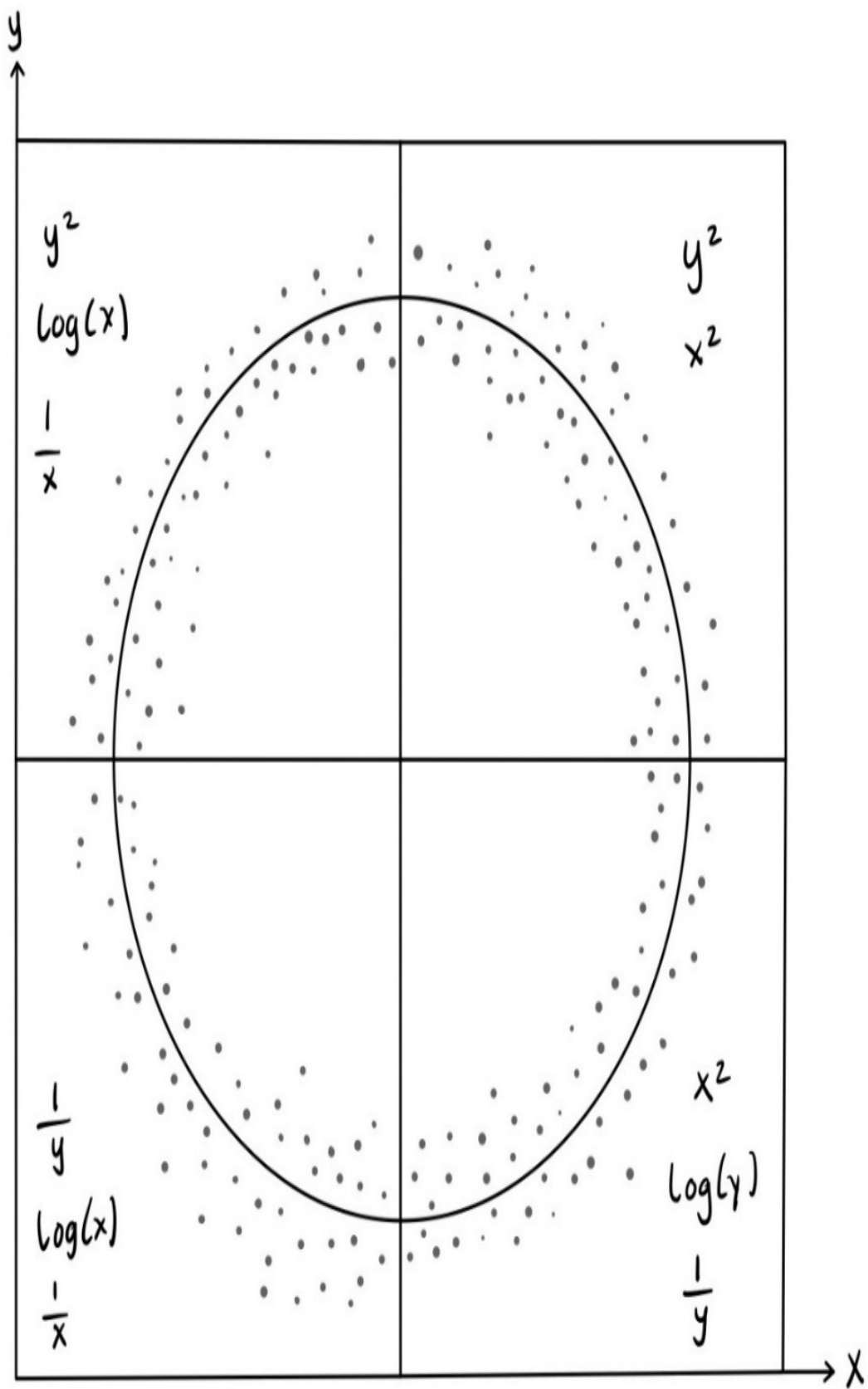
**Figure 4.28** Linear and some non-linear correlations can be easily visualized.



Choosing a correlation coefficient is often dictated by the type of data you are working with (e.g., when relationships are not linear or one variable is not continuous). When measuring associations between two continuous variables, you will generally benefit from visually observing a scatterplot of the relationship and determining if it can be *transformed* into a linear relationship.

For example, the negative correlation shown in Figure 4.29 depicts two variables with a *curvilinear* relationship. The best-fit curve is easy to visualize, but one or more variables will need to be transformed to create a linear variable for Pearson's  $r$  coefficient to represent the strength of the relationship accurately. The *Circle of Transformations* is a common diagnostic tool for identifying appropriate transformations to your variables.

**Figure 4.29** The circle of transformation recommends possible transformations to test based on the shape of the two variables you are comparing shown in a scatterplot.



Often, you will benefit from testing more than one of the transformations to determine if one method yields a higher correlation coefficient that better fits the data.

### 4.3.2 Regression Modeling

Like correlation, regression is a method for investigating the strength and direction of a relationship between two variables. Rather than providing a single coefficient to describe the relationship, a regression is used to model the relationship between a *dependent* variable and one or more *independent* variables. Regression modeling is used extensively in *predictive* and *causal* modeling, which we will discuss at length in Chapter 9.

A *linear regression* models a line of best fit to describe the relationship between the dependent variable and one or more independent variables. The equation for a simple linear regression (one independent variable) is provided in one of the following forms:

#### Linear Regression Formula with one Independent Variable

$$y = mx + b$$

This is recognizable as the formula for the *slope of a line*, where  $b$  is the y-intercept (x-value where  $y = 0$ ) and  $m$  is the slope (the change in  $y$  for a 1-unit change in  $x$ ). A *multiple linear regression* equation (more than one independent variable) will often be presented in the following format:

#### Alternative Linear Regression Formula with two Independent Variables

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

In this version of the formula,  $\beta_0$  is the y-intercept, and  $\beta_1/\beta_2$  are the respective slopes for each predictor. Both methods of representing a regression equation are appropriate for simple and multiple regression models. However, the latter is sometimes more prevalent in academic settings and for models with multiple predictors.

Linear regression is a *parametric* statistic that makes similar assumptions to the previous tests we've discussed. It assumes that your data represents a set of independent events and that a *linear relationship* exists between your dependent variable and its independent variables. Any data not meeting these assumptions should be appropriately transformed (see Figures 4.20 and 4.29). Linear regressions also make the following assumptions:

- The variables in your dataset are *multivariate normal*. This means that across the variables in your model, their *combined* distribution follows what's known as a multivariate normal distribution. This is often assessed by generating a Q-Q plot to compare the *quantiles* of each variable to those of a normal distribution.
- The independent variables are *not* highly correlated with each other, which is typically referred to as *multicollinearity*. This is generally evaluated by evaluating correlation values between the independent variables and selecting between variables when there are strong correlations.
- The spread of errors (residuals) is consistent for all values of the independent variables, known as *homoscedasticity*. This is typically evaluated by plotting residuals against predicted values (a residual plot). When this assumption is violated, it's recommended to use a Weighted Least-Squares regression that weights observations based on the size of their errors or to transform the dependent variable using a square root or logarithm similar to how you might in the case of non-linear relationships.

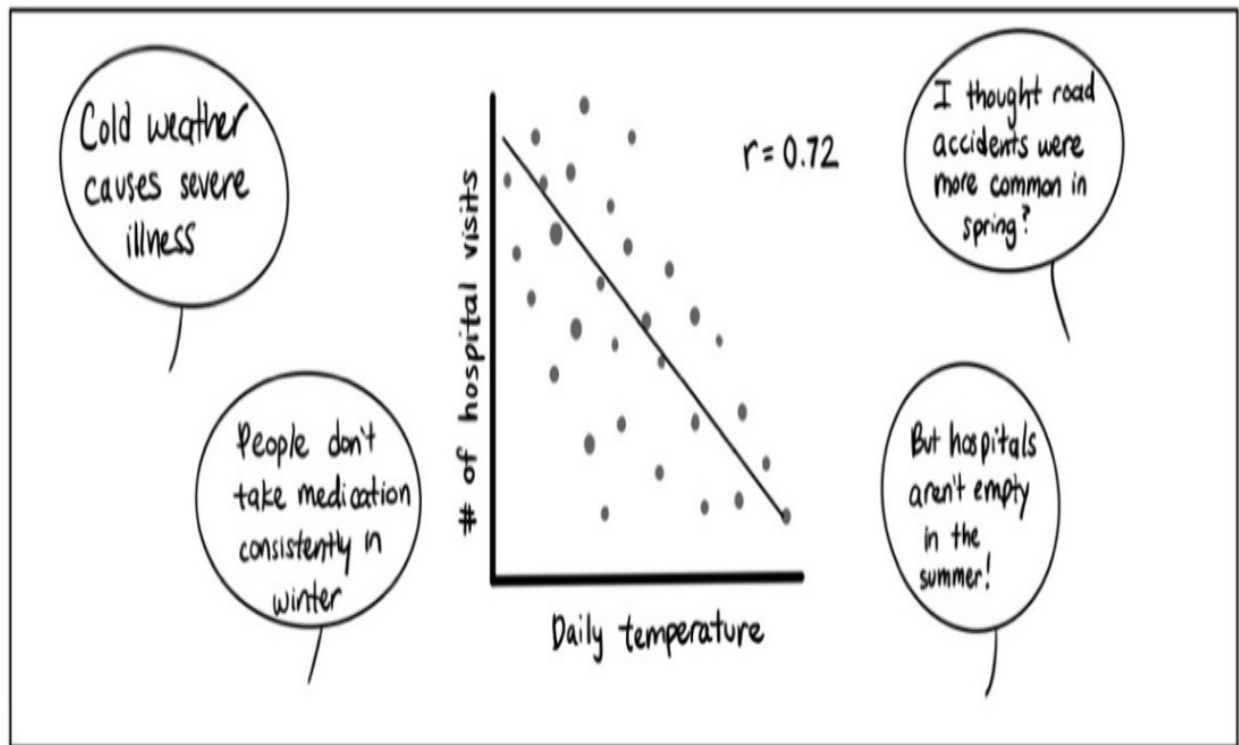
We will discuss the statistical modeling in depth in chapter 9, including Python implementations of regression models for different types of deliverables.

### 4.3.3 Reporting on Correlations and Regressions

Correlations are one of the most widely known and understood statistical concepts. Many stakeholders can quickly gain value from visualizations, coefficients, and summaries with minimal additional context. By extension, many of the interpretations of correlations can be applied to regression modeling in the presentation of your final deliverable.

In practice, how your stakeholders interpret correlation results can be an early diagnostic for the general comfort level with data and statistics across your organizations.

**Figure 4.30** Interpretations of correlational results can provide insight into the misconceptions about their purpose and limitations.



In one of my roles in data science, our team discovered this issue when reporting on correlations to various stakeholders. We identified some patterns of misinterpretation:

- Attributing a direct causal relationship between the two variables
- Adding interpretation based on previously held beliefs
- Disputing the relationship based on partial information or previously held beliefs

To better assist interpretations, we developed a guide to evaluating correlations (see Figure 4.31) and specific recommendations for interpreting results. The recommendations were delivered in presentations to large audiences that were recorded, disseminated, and archived for a large portion

of the organization to refer to over time.

**Figure 4.31** Example of a slide created to guide statistical interpretations of correlations.

## Interpretation of Results

✓ There is a negative correlation between the variables – the lower the temperature, the more hospital visits tend to occur.

✓ We have several hypotheses about the underlying causes of this relationship that require investigation.

✗ A lower daily temperature causes more hospital visits.

✗ We think that certain diseases spread more often in winter, causing the surge in hospital visits.

When reporting regression results, it may be necessary to distinguish between *predictive* relationships and *causal* relationships for your stakeholders (these are not the same, and we will discuss this at length in Chapter 9). The predictive nature of a regression model is implied in its selection of

independent and dependent variables, and its results are even more easily interpreted as causation.

In your deliverables and presentations, you may want to consider the following strategies for mitigating misinterpretation:

- Isolate and present the strongest independent variable relationships with your dependent variables. These may be best communicated as univariate correlations with scatterplots.
- Include clear, consistent language on what conclusions your stakeholders *can* draw and limitations highlighting what they *cannot* reasonably conclude.

## 4.4 Activity

We haven't yet answered the first question of this chapter – is Boston or New York City warmer in July?

1. Import the `nyc_boston_weather.csv` dataset associated with this book. Generate distributions to visualize the data.
2. Check all of the assumptions of the t-test. Make any necessary transformations to normalize the data.
3. Determine if you have a sufficient sample size by running an *a priori* power analysis with an alpha level of .05, a medium effect of 0.5, and 80% power.
4. Run an independent samples t-test to determine if there is a significant difference between Boston and New York City's weather in July of 2022. Which city is warmer, if any?
5. Prepare a summary of your findings for a stakeholder who does not have direct experience with inferential statistics. Include statements on how you *can* and *cannot* interpret the results.

## 4.5 Summary

- **Measures of central tendency** such as the mean, median, and mode are used to quickly assess the characteristics of a dataset. Each can be used

in reporting to stakeholders; however, valuable information about outliers, skew, and shape can be lost if only one measure is reported.

- **Measures of variability** tell you how much your dataset deviates from the mean or median. These measures give you an estimate of the spread of your dataset and a first point of comparison between two or more distributions.
- **Parametric statistical tests** are widespread across nearly every domain of analytics. These tests make explicit *assumptions* about the parameters and characteristics of the underlying population distribution.
- Many parametric tests assume that your population is **normally distributed**. These tests require that your data can be represented as a normal distribution through trimming, transformation, or other appropriate steps.
- The majority of statistical tests leverage the **p-value** in the interpretation of the test coefficient. This value estimates the probability that you would observe the magnitude of group differences if there were no actual differences in the population. This value is often used as a threshold to determine *statistical significance*.
- Each statistical test has a **minimum recommended sample size** to detect an effect between groups or variables. Many tests (e.g., t-tests) also have a theoretical upper limit on your sample size before you risk generating false-positive results.
- Making inferences using regression modeling requires that you meet many of the same assumptions as tests comparing two or more groups (e.g., t-tests, ANOVAs). In addition, Pearson's Correlation and linear regression require that your variables have a linear relationship or can be transformed into a linear relationship.
- Reporting the results of inferential statistical tests to non-technical stakeholders requires precise language to guide teams through the appropriate interpretation and the limitations of your findings.

# 5 The Statistics You (Probably) Didn't Learn: Non-Parametric Tests, Chi-Square Tests, and Responsible Interpretation

## This chapter covers

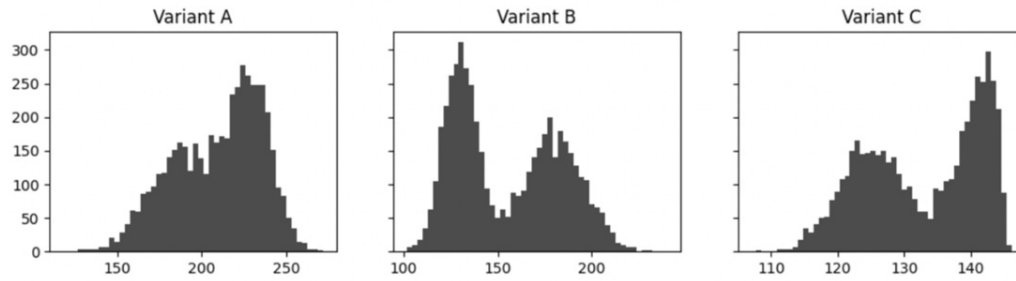
- The history and original purpose of common statistical tests
- Evaluating and using non-parametric alternatives to common parametric tests
- Leveraging the chi-square test for categorical comparisons
- Mitigating the likelihood of false positive and false negative results
- Using statistics responsibly to ensure the accuracy of your findings

“The number of ways you can misunderstand statistics is infinite. The number of ways you can understand it is finite.”

– *Dr. Lawrence Tatum*

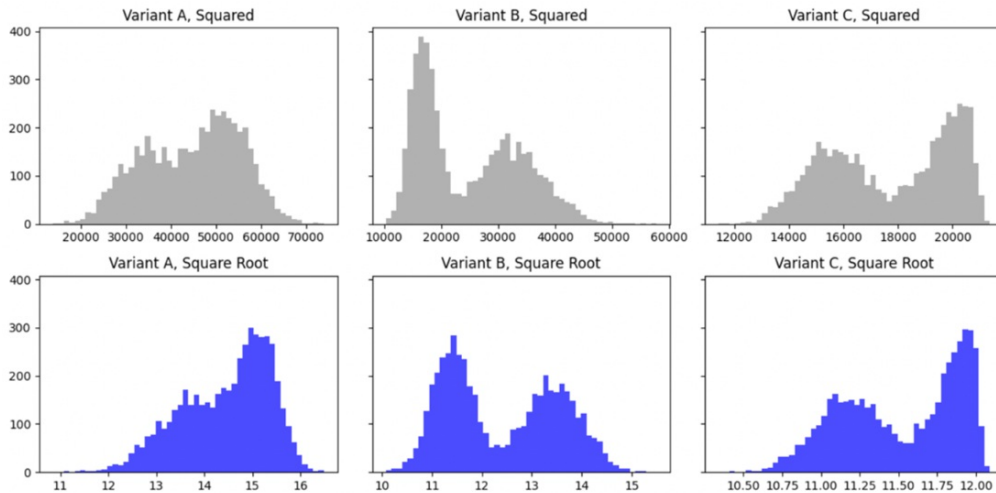
You are an analyst on a product analytics team at a software company. The product team is evaluating whether one of the new page versions on the app leads to customers completing a workflow faster. You were asked to conduct a between-subjects ANOVA to assess the results of an A/B/C test. In your diligence as an analyst, you start by exploring the distributions of data to check the assumption of normality. The distributions look like this:

**Figure 5.1 The distribution of each group is bimodal.**



You try transforming each distribution with a few of the recommended approaches discussed in chapter 4 but cannot change the shape of the bimodal distributions.

**Figure 5.2** Square and square root transformations of the original bimodal distributions.



What do you do? Do you give up and *not* conduct your analysis? Based on the assumptions discussed in the previous chapter, you are unlikely to yield accurate results with a one-way ANOVA. You risk providing recommendations to the product team, leading to less optimal app user experiences.

Parametric tests cannot be used for every situation. However, the tests we covered in chapter 4 are *not* the only options available to you as an analyst. When you cannot expect to produce reliable results with a t-test, ANOVA, or any other method we have covered so far, you have a wide range of *non-parametric* tests available to evaluate your data and use for making inferences about the broader population.

To provide an appropriate context for the underlying logic of parametric statistical tests, we will first cover a brief history of the development of these tests. Equipped with an understanding of their intended purpose, you will be prepared to answer challenging stakeholder questions, communicate the limitations of a test, and think critically about an extensive range of questions you may answer in your organization and interviews as you seek to grow your career.

## 5.1 The Landscape of Statistics: Past and Present

In some form, statistics have been actively leveraged for centuries. Probability theory dates back to the 17<sup>th</sup> century when it was used to predict uncertain events (e.g., the number of annual births and deaths in a town). The theory was expanded over the following centuries, with methods and approaches (e.g., Bayes' theorem) still widely used today.

Most *parametric* statistical tests we're familiar with in analytics were developed in the last 100 to 120 years. Since their development, they have grown in influence and often dominate the methodological choices of fields in the social sciences, humanities, and others. But their rise to prominence does not necessarily reflect their efficacy across possible research questions.

### 5.1.1 The Evolution of Statistical Methods

The critical historical development of parametric statistics is usually left out of statistics education. If you sat through courses as I did, you might have been taught formulas for each test as a rule of law or formulas you need to memorize the same way you do in a calculus class. Statistics, however, is *not* the same as mathematics, and none of the formulas you learned were discovered—they were developed with a purpose in mind.

#### Tests Developed for a Purpose

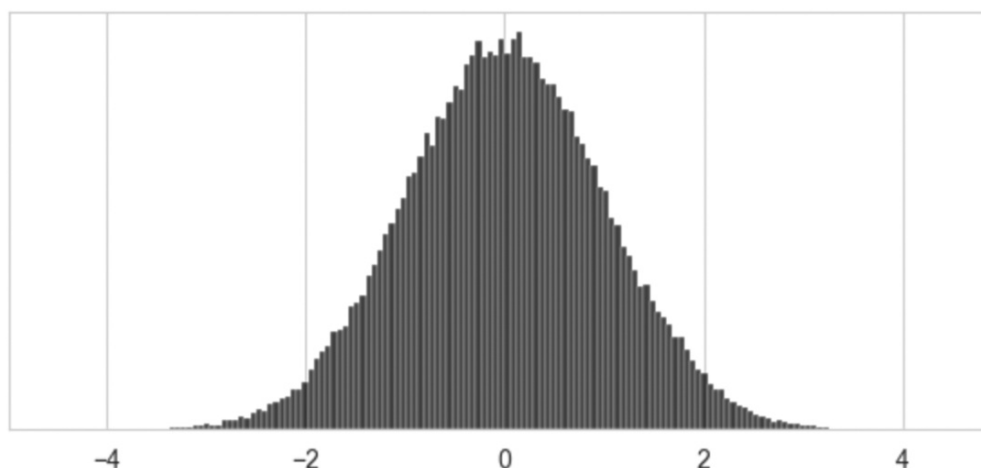
The story of how parametric statistics rose to prominence lies with the eugenics movement. From the turn of the 20<sup>th</sup> century to the 1930s, many statistical tests we know were developed by well-known eugenicists (Francis

Galton, Karl Pearson, Ronald Fisher) as part of their efforts to operationalize the concept of *intelligence*.

In the view of this movement, intelligence was treated as a universal concept in which someone's intellectual capabilities were considered biologically determined. The IQ test was developed as a measure by which people can be "ranked" and differentiated by race or ethnicity. The ideas and philosophies of eugenics and its focus on differentiating groups of people as superior or inferior were cited by Nazi Germany as a key guiding factor in its atrocities.

While this type of explicit thought is less prevalent in today's academic world, many of the methods and tests developed as part of efforts to define and measure intelligence are still used today. The ideas and concepts of eugenics guided the development of statistical significance thresholds, the bell curve (the scaled shape of the distribution of intelligence scores), and correlations that still dominate academic and statistical thinking. We're taught to frame questions as looking for *differences* between groups or *correlations* between variables. We're taught that the normal distribution is ever-present in measures of behavioral and cognitive phenomena. We *still* use the IQ test to measure intellectual capabilities and discuss a general intelligence capable of being reduced to a single number. We *still* teach these topics in social sciences as if they're truth rather than hypotheses.

**Figure 5.3** The bell curve/normal curve is ever present in statistics and *does* occur naturally for a handful of phenomena, such as the distribution of human heights.



## The Development of Non-Parametric Statistical Tests

In the decades after developing parametric methods, statisticians responded to the limitations of parametric methods and their assumptions about the underlying distribution of the population data. These assumptions (e.g., normality) were often unrealistic to meet and limited the practical applications of their tests and the ability to approach research questions as group differences or correlation/regression problems.

Statisticians developed non-parametric methods in the mid-20th century, which do not make assumptions about the distribution or shape of the data. A range of tests were created, and the approach grew in popularity over the decades. We continue to see advancements in new model development and model refinement today, with many new classes of non-parametric methods developed with machine learning advancements.

### Further Reading

Regardless of your research questions, belief on this topic, or the type of analytics work you do, it's essential to understand the context of the tools we use as analysts. In aggregate, the methods we regularly used were developed to compare, differentiate, and rank people. The way of thinking about the questions we ask is still a part of statistics and social sciences education a century later.

We won't discuss a comprehensive history of statistics and its development in this book; this topic can easily span multiple books and has been well-documented by amazing authors in the past decades. If you're interested in learning more, I recommend reading the following books:

- *The Theory that Would Not Die*, by Sharon McGrayne: this book is centered around the history and application of Bayesian statistical methods, which the purveyors of frequentist statistics heavily criticized.
- *The Mismeasure of Man*, by Stephen J. Gould: this book comprehensively criticizes *The Bell Curve* (promoting the above ideas). It is an excellent source for explaining statistical methods such as factor analyses.

## 5.1.2 Choosing Your Approaches Responsibly

To succeed as analysts, we still need to leverage parametric statistics as appropriate. Despite their limitations and history, I recommend the following takeaways when choosing a research method and statistical test:

- We're trained as researchers and analysts to frame our questions as looking for group differences, correlations, or causal relationships. This is the primary focus of this book. Previous chapters covered research questions and hypotheses with this structure due to the prevalence and high availability of training material on these topics. However, looking for differences and correlations is *not* the only method you can use. You can leverage dozens of quantitative, qualitative, and mixed-methods approaches. For example:
  - **Observational research** aims to observe and record data about behavior and events without manipulation or intervention. The overall goal of this method is to *describe* (see Chapter 2 for more detail on descriptive methods) and understand a phenomenon. This is often the *best method to use* in analytics when first evaluating a new dataset.
  - **Observation-oriented modeling** focuses on relationships and patterns in the data and how they relate to observed phenomena being studied. This approach uses graphical representations, statistical models, and machine learning algorithms to build models based on observations. You can read more about this method in the referenced paper, *Observation Oriented Modeling* [1].

**Figure 5.4 Visual Representation of a Causal Phenomenon with Observation-Oriented Modeling**

- ① Observation: Users struggle to set up and use your software
- ② Action: Do they visit your support site?
- ③ Question: Are users who visit your support site more likely to complete the setup process?



- For each research method we covered in chapter 4 (group differences, correlations, and predictive/causal relationships), you can use numerous non-parametric and semi-parametric statistical tests to test your hypotheses. Many are available in SAS or R packages in specific research fields.
- The near-ubiquitous parametric statistical tests taught in an introductory statistics course (e.g., t-test, ANOVA, Pearson's correlation) are not necessarily *better* than tests used less frequently. There were aggressive efforts within academic statistics to popularize parametric methods and discredit probabilistic methods (*The Theory that Would Not Die* discusses this extensively).
- Statistical tests were designed assuming you leverage a finite sample representing a relatively small proportion of the broader population. For example, many psychological studies and clinical trials recruit participants in the hundreds or thousands. In today's world, an analyst can easily work with datasets containing millions of records. If your statistical test assumes a smaller sample size, you will *not* yield meaningful or accurate results (shown in chapter 4).

## 5.2 Non-Parametric Statistics

We spent Chapter 4 discussing tests that make rigid assumptions about the shape of your data. In many cases, those assumptions cannot be met, or the underlying distribution of the data is unknown. *Non-parametric statistics* is a class of methods that doesn't make assumptions about the underlying

distribution of the data. They're commonly used instead of parametric statistics when assumptions cannot be met. Many tests offer additional flexibility on data types, enabling you to compare categorical and ordinal data.

As an analyst, you can apply a non-parametric alternative for each parametric test we covered. Each of the methods we will cover is available in R or Python, and most are also easily applied in SPSS, SAS, and STATA.

### 5.2.1 Comparisons Between Groups on Continuous or Ordinal Data

The most prevalent parametric statistical tests assume that you use *continuous data* captured about a phenomenon of interest between or within groups. When your data is *not* normal, you have several possible choices to evaluate your results.

#### Comparisons Between Two Groups

The first test we will cover is the Mann-Whitney  $U$  test (also known as the Wilcoxon rank-sum test). This compares the *medians* of two *independent samples* and is performed by taking the *sum of ranks* between groups and calculating a  $U$ -statistic from the group sums. The sum of ranks is calculated by adding the ranked values for each group; each sum is used to calculate a  $U$ -value for each group, and the lowest among them is selected as the  $U$ -statistic. The  $U$ -statistic is compared to a critical value using a  $U$ -table, just as we've done with parametric tests in Chapter 4. If the  $U$ -statistic is **lower** (*not* higher) than the critical threshold, it's considered statistically significant.

**Figure 5.5** Steps to calculating the  $U$ -statistic

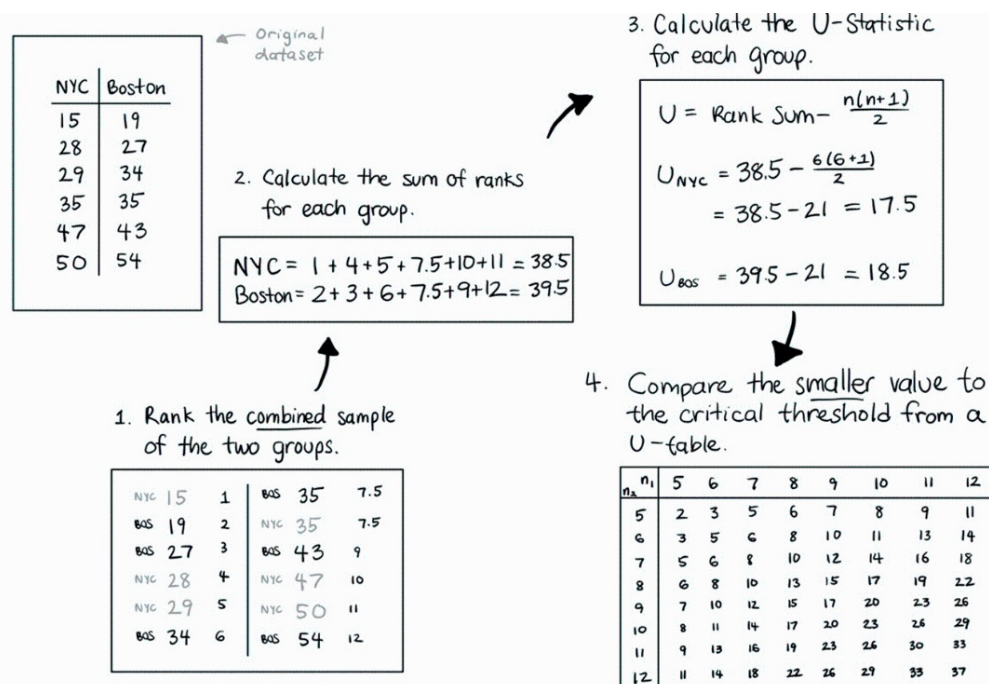


Figure 5.5 shows an example of a hand-calculated  $U$ -statistic comparing daily weather data between New York City and Boston. The  $U$  statistic is calculated using six values per city for a total sample size of 12. The  $U$  statistic for the city with the *lower* value (New York City) is compared to the appropriate critical value in a  $U$  table. In the same manner as the tests we discussed in chapter 4, we can conclude that the temperature difference between New York City and Boston is highly significant.

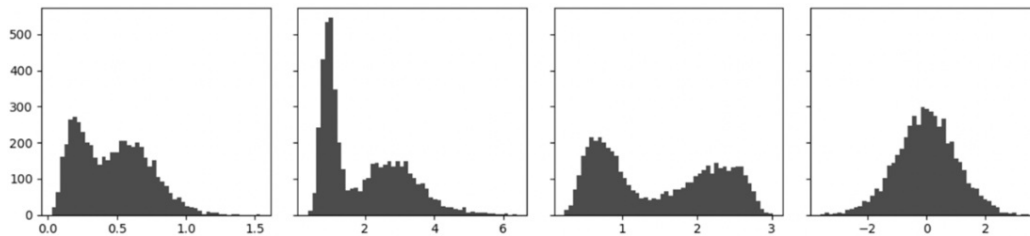
The Mann-Whitney  $U$  test is robust to violations of the assumptions of a  $t$ -test, which means that you can leverage this test in most cases where you cannot rely on a parametric test. The test compares the *relative position* of each data point to the rest of the dataset rather than being impacted by the numerical value of that data point. Your data can be non-normal or have unequal variances with no impact on the validity of your results.

To summarize the value and advantages of this approach:

- It's agnostic to the shape of your distribution, producing reliable results when parametric test assumptions are violated.
- It's more robust than a  $t$ -test to moderate differences in the sample sizes of each group.
- In addition to continuous data, the Mann-Whitney  $U$  test (and all non-

parametric tests we will cover in this section) can be conducted with *ordinal* data, making statistical inference possible on far more data types than parametric tests alone.

**Figure 5.6** The Mann-Whitney  $U$  test can be leveraged for any distribution shapes displayed, whereas the  $t$ -test is only appropriate for the final (normal) distribution shape.



As with every statistical test we've covered thus far, the Mann-Whitney  $U$  test is not a silver bullet free of limitations. As an analyst, you will need to evaluate the properties of your data to determine if this non-parametric test is appropriate and take additional steps to interpret the results:

- When your sample is too small, you will likely generate results with false-positive errors. You can see this in our example above, where a minimal temperature difference was highly statistically significant with a sample size of 12 across groups.
- The test is also highly sensitive to *large* sample sizes, where you are more likely to produce false-negative results.
- The Mann-Whitney  $U$  test performs poorly when there are many tied ranks (e.g., one tie in this dataset). Where possible, including additional digits in your floating-point data can minimize the impact and frequency of tied ranks.
- There's a lack of consensus among statisticians and analysts on many aspects of the Mann-Whitney  $U$  test and its usage. For example, there are different recommended approaches for handling tied ranks and the appropriate minimum sample size necessary to draw practical conclusions. Most statistical software and packages choose between approaches and require you to adhere to their choice unless you can write a custom module.
- If you search for the application of the Mann-Whitney  $U$  test, you'll also discover the lack of agreement between otherwise reputable resources on basic information about the test. My research found discrepancies in

the formula, assumptions, data types, minimum sample sizes, and other information. Be prepared to dig into peer-reviewed research on this test to ensure the information you seek is accurate.

- Unlike the  $t$ -statistic, the  $U$ -statistic does not include negative values indicating the direction of the relationship. You must compare the  $U$ -values or medians between groups to determine which has *significantly higher ranks*.

**Figure 5.7**  $U$ -values or rank sums agree directionally with the *median* of a dataset but don't always align with the interpretation of the mean. Be careful in conflating your interpretations!

NYC 16	1	BOS 35	7
BOS 19	2	NYC 35	8
BOS 27	3	BOS 39	9
NYC 28	4	NYC 49	10
NYC 29	5	NYC 50	11
BOS 34	6	BOS 51	12

	NYC	Boston
Rank Sum	36	37
Median	32	34.5
Mean	34.5	34.17

means and rank sums do not need to agree

When reporting results of a Mann-Whitney  $U$  test (or any non-parametric test that uses a rank-sum method), I strongly recommend presenting your stakeholders with *median* values between groups rather than *mean* values (discussed extensively in Chapter 4). Rank sum values, by definition, will align with which group has a higher median but will *not* reliably agree with the interpretation of mean values. Including both or only the mean may create confusion or inaccurate interpretations of your results.

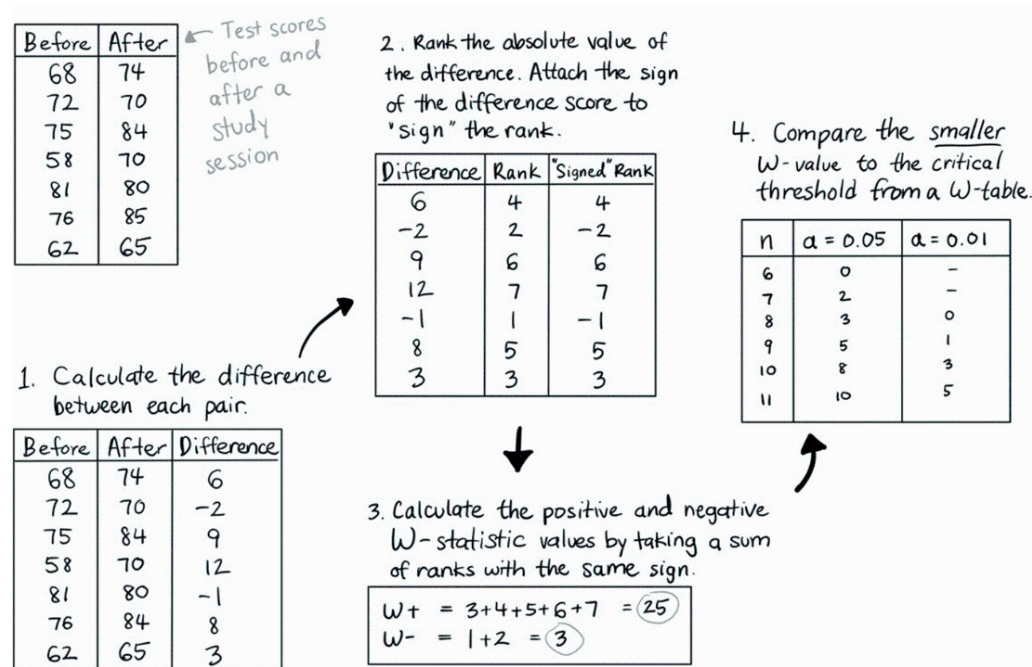
## Comparisons Within Two Groups

The Wilcoxon signed-rank test is a non-parametric test for *repeated measures comparisons* between two groups. It's designed to be used in place of a within-subjects  $t$ -test when the assumption of normality cannot be met.

Similar to the Mann-Whitney  $U$  test, the Wilcoxon signed-rank test leverages the rank of each data point to calculate its test statistic. Since this is a repeated-measures test, it ranks the difference in values between each *pair* of data (e.g., a participant score before and after treatment). The  $W$ -statistic is generated by taking the rank of the *absolute value of differences* between pairs and then attaching the sign of the original difference value to the rank.

The resulting set of rank values combines negative and positive values, grouped by sign and added together for the *sum of ranks* calculation. The lower W-value between the sum of positive ranks (difference values in the *positive* direction) and negative ranks (difference values in the *negative* direction) is selected as the test statistic and compared to a critical value.

**Figure 5.8 Steps to calculating the W-statistic**



The Wilcoxon signed-rank test shares many advantages and limitations with the Mann-Whitney *U* test. The test can be used with continuous and ordinal data and does not require you to have normally distributed data. However, it *does* require that two assumptions be met to ensure the accuracy of results:

- The differences between pairs of observations should be *symmetrical* (e.g., not skewed) around the median. This is tested by visually observing the distribution of differences and checking the skewness and kurtosis values.
- Ties in ranks should be eliminated by introducing minor, decimal-value differences at random. A large number of ties in your difference values will reduce the statistical power of your test and increase your likelihood of false-positive results.

In addition to reporting statistical test results, it's likely valuable to report summary information about medians and proportions to your stakeholders for clarity. A synopsis of your findings might include the following:

The median score was 8 points higher (76%) after the study session, which was determined to be highly statistically significant. 70% of students saw an increase in scores after the study session

As with any non-parametric test leveraging rank sums and relative positioning, it's important to stress that your statistical tests did *not* evaluate mean scores. In my experience, this is easier to convey if you use clear statements on why you chose your summary metrics and depict visualizations such as boxplots. This can help to focus your readers on the relative position of your data points in their distribution rather than looking for an average score.

Let's revisit our case study from Chapter 4. As a recap, Naomi is a research scientist at a pharmaceutical company analyzing the results of a randomized control trial on a new medication to treat insomnia. She used an independent samples t-test to evaluate differences in the experiment and control groups on the primary measure of interest, *hours of sleep*. Additional measures captured violated the assumptions of the t-test and ANOVA; thus, Naomi evaluated them using non-parametric approaches as shown below:

#### **Analyzing Non-Normal and Ordinal Data**

In addition to the continuous, normally distributed data that Naomi collected for the sleep study, she has several variables in her dataset that were unsuitable for parametric statistical tests. Each of the following was collected or derived as part of the study:

1. In addition to the time spent awake during the sleep cycle, the study captured the *number of sleep interruptions* as a separate measure to identify the number of times a participant woke up. This is captured as discrete count data ranging from 0 to 7, making it an *ordinal* dataset.
2. The difference in the number of hours of sleep before the beginning and end of the sleep trial is captured separately for the experiment and control groups and is highly skewed for both.

3. Participants' age, which has a bimodal distribution.

Naomi decides to conduct two statistical tests. She chooses a Mann-Whitney *U* test to compare the number of sleep interruptions between the experiment and control group at the end of the study. She also decides that a separate Wilcoxon signed-rank test for each participant group (experiment and control) is appropriate to compare the number of sleep hours at the trial's beginning and end.

She first calculates the *U*-statistic, yielding a highly significant value of 79,635 and a p-value of less than .001. She notes that the experiment group has a median of 2 sleep interruptions, and the control group has a median of 3. She is aware that the *U* test is highly sensitive to larger sample sizes and includes a note in her report that she may need further exploration to answer this research question appropriately.

For her second question, Naomi is interested in comparing the hours of sleep before and after the trial for the experimental and control groups. She calculates the *W*-statistic for each of the two tests being conducted separately. She obtains a *W*-statistic for the experiment group of 8,794 and a p-value less than .001. She notes that the experiment group had a median of 7.75 hours of sleep during the final sleep study, compared to 5.86 hours during the first sleep study before receiving the experimental medication.

By comparison, the test comparing hours of sleep for the control group had a *W*-statistic of 54,674 and a p-value of 0.64. The group had a median of 5.93 hours of sleep during the final sleep study, compared to 5.83 during the sleep study at the start of the experiment.

She concludes that the experimental drug was highly effective at increasing the median hours of sleep and reducing the number of sleep interruptions compared to the placebo.

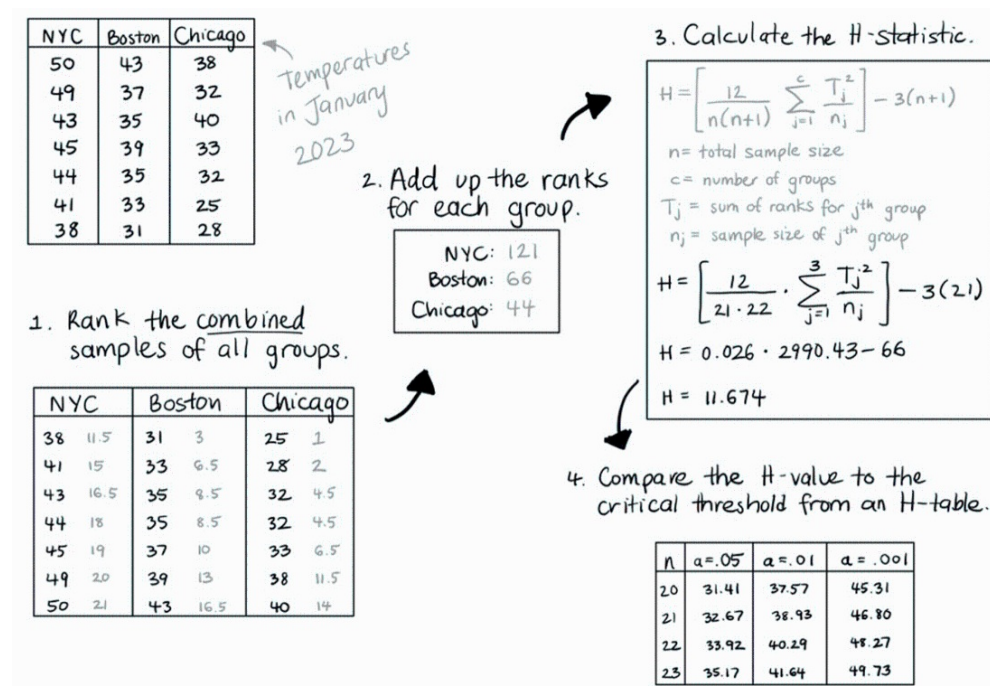
### **Comparisons Between Two or More Groups**

There are two common alternatives to the one-way ANOVA used to compare three or more groups. The Kruskal-Wallis test compares the medians of two

or more independent samples. Similar to the Mann-Whitney  $U$  test, it calculates the sum of ranks between each of the groups and applies a weight to each rank based on the sample size of the group. This test can also compare main and interaction effects with two independent variables, as you would expect to do with a two-way ANOVA.

The Kruskal-Wallis test also does not make assumptions about the underlying shape of your data. The  $H$ -statistic of this test provides information on the magnitude of the difference between groups but does not indicate the direction of that difference. The Kruskal-Wallis test can compare as few as two groups in conjunction with or in place of the Mann-Whitney  $U$  test.

**Figure 5.9 Steps to calculating the  $H$ -statistic**



The Kruskal-Wallis test also shares the same applications (non-normal or ordinal data) and limitations (lower power compared to a one-way ANOVA) as the tests we've discussed. There are some considerations worth noting in your usage of this test:

- The Kruskal-Wallis test requires a post hoc comparison to determine *which* groups have significant differences when comparing three or more groups. Dunn's test is the most common, applying the same type

of Bonferroni correction used in a one-way ANOVA. (Like with a one-way ANOVA, you risk false-positive errors when making multiple pairwise group comparisons.)

- Despite the  $H$ -statistic's more complex formula than the Mann-Whitney  $U$  test, their performance is identical when conducted between two groups. The test statistic results will differ, but their distributions are calibrated, so the resulting p-values will nearly match between tests. We can demonstrate this in Python using two simulated datasets:

```
import numpy as np      #A
from scipy import stats as st

group_a = st.skewnorm.rvs(a=9, scale=2.2, size=99) + 4.5      #B
group_b = st.skewnorm.rvs(a=11, scale=1, size=99) + 4.6

H = st.kruskal(group_a, group_b)      #C
U = st.mannwhitneyu(group_a, group_b)

print(f"Kruskal-Wallis Test, H={H[0]}, p={H[1]}").      #D
print(f"Mann-Whitney U-test, U={U[0]}, p={U[1]}")

Kruskal-Wallis Test, H=9.869, p=0.002
Mann-Whitney U-test, U=6167.0, p=0.002
```

- If you expect to run more than one test to compare a variable number of groups (e.g., comparing student test scores between grades 6, 7, and 8 in two schools using two one-way tests), using a Kruskal-Wallis for both will make for an easier comparison of test statistics than using a Mann-Whitney for the second comparison (even though you can do so without jeopardizing the accuracy of your results).

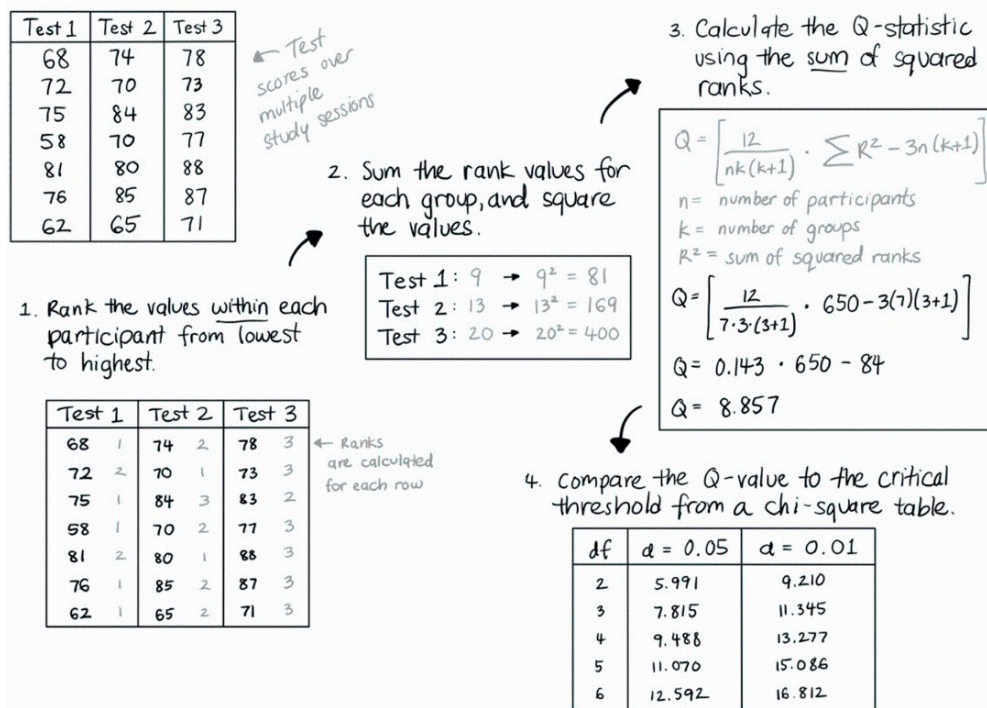
In practice, I have seen both the Mann-Whitney and Kruskal-Wallis test used for non-parametric comparisons. The leading case in which I've seen a clear preference for a Kruskal-Wallis test is when colleagues expected to run group comparisons programmatically over time. We anticipated variation in the number of groups being compared (between 2 and 3), so running the Kruskal-Wallis test allowed for the automation of the periodic calculations.

## Comparisons Within Three or More Groups

Finally, let's discuss the Friedman test as an alternative to the one-way repeated measures ANOVA. This test compares the ranks of three or more related samples and can be used instead of an ANOVA when the assumptions of normality or sphericity are violated.

The Q-statistic (also referred to as the  $X^2$ /chi-square statistic) is calculated as the *sum of squared ranks*, ranked for each participant and summed/squared for each group. If you reject the null hypothesis, the Wilcoxon rank-sum test can be used as a post hoc test with a Bonferroni correction to identify where sample differences are present.

**Figure 5.10 Steps to calculating the Q-statistic**



The Friedman test leverages the same sum of ranks approach as all of the non-parametric tests we've covered so far, with some notable exceptions:

- Unlike the Kruskal-Wallis test, the Friedman test *must* compare at least three groups. You can't use it interchangeably with the Wilcoxon rank-sum the way you can with an independent samples comparison.
- The previous tests in this section take an opposite approach to their parametric counterparts in interpreting the test statistic—the test value

needs to be *below* the critical threshold to be statistically significant. The Friedman test does **not** take this same approach; instead, the Q-statistic takes the same approach as the univariate parametric tests and must be higher than the  $X^2$  critical value to achieve statistical significance.

When your analysis yields a significant result, you will usually need to conduct post hoc tests to identify the significant group differences, as expected with repeated measures ANOVA. The most common approach is the manual application of the Wilcoxon rank-sum test for each pair of groups and applying a Bonferroni correction to the p-values to minimize the likelihood of false positive results.

### **Wait! Why Not Use Non-Parametric Tests for Everything?**

The non-parametric tests we've covered have clear advantages over parametric tests in many cases—they can be used with more data types, they're largely agnostic to the underlying shape of your data, and their calculations are relatively straightforward and available in most statistical software. That's great! So why don't we use them everywhere and toss parametric statistics to the wind?

Non-parametric methods leveraging rank-sum comparisons have some key limitations that can limit their applicability:

- Parametric and non-parametric tests *cannot* be compared 1:1 since parametric tests typically compare means/variances, and non-parametric tests compare medians/ranks between groups.
- Due to their usage of median and rank comparisons, you may have to modify your hypothesis to account for the different types of comparison (the median and rank instead of the mean).
- When your data *does* meet a parametric test's assumptions, a non-parametric test will typically have lower statistical power. This is especially true when working with large or small sample sizes. While there are upper limits to the sample size you can use with a parametric test, you're more likely to generate false positive results with a large  $n$  when using a non-parametric approach.
- While the non-parametric tests we covered are available in Python

packages such as `scipy` and `statsmodels`, many of their post-hoc tests are not. You may have to leverage additional packages such as `scikit-posthocs` or do some manual work to conduct multiple comparisons depending on your comparison of choice. An example post hoc comparison is shown below:

```
import numpy as np #A
from scipy import stats as st
import scikit_posthocs as sp

group_a = st.skewnorm.rvs(a=9, scale=2.2, size=99) + 4.6 #B
group_b = st.skewnorm.rvs(a=11, scale=1.5, size=99) + 4.6
group_c = st.skewnorm.rvs(a=9.1, scale=2.0, size=99) + 4.6

data = [group_a, group_b, group_c]
H = st.kruskal(group_a, group_b) #C
post_hoc = sp.posthoc_dunn(data, p_adjust="bonferroni")

print(f"Kruskal-Wallis Test, H={H[0]}, p={H[1]}") #D
print(post_hoc)

Kruskal-Wallis Test, W=9.866, p=0.007

      1      2      3
1 1.000  0.122  0.896
2 0.122  1.000  0.006
3 0.896  0.006  1.000
```

Where possible, applying *both* a parametric and non-parametric statistical test is an excellent strategy for validating your results and counterbalancing the limitations of each. Having two statistical tests with convergent results provides strong evidence supporting your hypothesis. When your tests *diverge*, you have an opportunity to tease apart whether your group *means* or *medians/ranks* differ. You are also provided with diagnostic information to help you identify risks of false-positive errors or violations of assumptions.

**Table 5.1 Divergent test results aren't necessarily bad – they can teach a lot about your data.**

	Suggested Interpretations	Stakeholder Communications
Parametric and non-parametric	Results are highly likely to reflect true positive differences between the actual	“Students in Classroom A had higher average and higher ranked test scores than

tests are significant	and relative position of values between groups.	Classroom B.”
Parametric tests are significant, only	Sample size may be too small to detect group differences with a non-parametric test. Parametric test results may reflect false positive errors. Assumptions of the parametric test may not be met.	“Students in Classroom A had significantly higher average scores than Classroom B. However, the two classes appear similar in rank, indicating that the average differences can be attributed to a small number of high performers.”
Non-parametric tests are significant, only	Assumptions of the parametric test may not be met. The non-parametric test results may reflect false positive errors. The dataset may contain extreme outliers.	“Classroom A had higher scores relative to Classroom B.” “Average scores did not differ between classrooms; however, Classroom A ranked higher than Classroom B.”

The robust and widely known parametric methods we covered are far from the only non-parametric methods available for you to analyze data. Statisticians continually propose new methods and publish them in statistical journals. Many offer advantages when working with specific data types or answering questions that don’t quite fit into the paradigm we’ve covered. If you frequently answer atypical questions with quantitative data, I strongly recommend keeping yourself up to date with this type of statistical research.

## Summary

Your ability to apply parametric and non-parametric tests as appropriate will significantly bolster your career and capacity to deliver high-quality and accurate results. While they’re not often taught in introductory statistics coursework (and, truthfully, often left out of intermediate and advanced coursework), these tests are proven alternatives to the t-tests and ANOVAs we’re familiar with.

In summary, I recommend the following takeaways on the practical usage of non-parametric tests for group comparisons:

- Non-parametric tests are an appropriate choice when your data doesn't meet the assumptions of normality and equal variances required for parametric group comparisons.
- The tests we covered have been shown in studies [2] to be slightly less sensitive than parametric tests, which means they may have lower statistical power to detect differences between groups.
- Each test can be used with continuous *and* ordinal data, making quantitative analysis possible on more data types than parametric tests.
- Non-parametric tests are *not* a silver bullet. As discussed in chapter 4, you may have difficulty accurately detecting group differences with a small or very large sample size.
- When communicating results to stakeholders, you may need to calibrate expectations about *which* measure of central tendency is being compared (*not* the mean).

## Activity

The following code generates three non-normal distributions. Let's assume that each distribution represents one of the three groups' performances on an assessment, and we are looking to determine which of the three groups has the highest scores.

```
import numpy as np      #A
from scipy import stats as st

np.random.seed(99)     #B

x_a = np.random.normal(loc=47, scale=4, size=55)  #C
x_b = np.random.normal(loc=53, scale=4, size=65)
X1 = np.concatenate([x_a, x_b])
X2 = st.skewnorm.rvs(83, size=120) + 51
X3 = np.random.exponential(scale=10, size=79) + 44
```

Run the code in the Python environment of your choice (terminal, Jupyter Notebook, etc.). You must have `numpy` and `scipy` installed for this step and `matplotlib` for the remainder of the activity.

1. Create a histogram of x1, x2, and x3 to visualize each data series. How would you describe the shape of each distribution?
2. Try transforming x1, x2, and x3 into a normal distribution. Which can and which cannot be successfully transformed?
3. Based on the possible transformations, can you run a between-subjects ANOVA?
4. Run the following code to conduct a Kruskal-Wallis test. The code assumes you have already imported the libraries in question 1. How can you interpret the results?

```
H = st.kruskal(x1, x2, x3)
print(H)      #A
```

5. Double the sample size values in the size parameter for x\_a, x\_b, x2, and x3. How do your results change?
6. Which group has the highest score? What is the best measure to report based on the comparison type in the Kruskal-Wallis?
7. Research the available documentation on the `scikit-posthocs` library. What post hoc tests are available in this library for the Kruskal-Wallis test? Try implementing at least two different tests and compare the output. How do the p-values differ between these tests?

## 5.2.2 Comparing Categorical Data

Based on your research question, there will be situations where your data doesn't allow comparing an ordinal or continuous measure by groups. If your independent and dependent variables are categorical, you will likely need to use a chi-square test for comparisons.

The *chi-square test* is a non-parametric statistical test used to identify differences between categorical variables by comparing frequencies between each category. This is one of the few non-parametric tests included in introductory statistics curricula in undergraduate and graduate courses. In my experience, it's often taught toward the end of the semester with limited focus on the methodology and how it differs from previous tests.

A chi-square test compares *observed to expected frequencies* (how the null hypothesis is conceptualized) between categories assuming no differences

between categories. It can be used as a one-way test (one variable) or a two-way test (a comparison between two variables).

**Figure 5.11 Comparisons of two categorical variables can be presented as a contingency table**

customer_id	sign_up_date	state	subscription_tier
28465	2021-09-09 18:40:27.789150	Florida	Enterprise
31656	2022-01-09 07:42:27.789150	California	Team
33206	2020-10-15 21:01:27.789150	California	Individual
36624	2019-07-08 10:37:27.789150	Florida	Individual
12498	2020-06-10 18:06:27.789150	California	Free
57147	2021-01-10 12:02:27.789150	New York	Enterprise

subscription_tier	Enterprise	Free	Individual	Team	Total
state					
California	301	300	308	303	1212
Colorado	299	335	289	325	1248
Florida	295	318	340	361	1314
New York	296	315	313	302	1226
Total	1191	1268	1250	1291	5000

You can use multiple types of chi-square tests to compare one or two samples:

- The **goodness-of-fit test** compares observed frequencies in a single sample against the expected frequencies.
- The **test of independence** compares observed to expected frequencies between *two* categorical variables. The data compared can be presented as counts of observations or proportions of the dataset.
- The **test for trend** determines if there is a trend or pattern in a categorical variable *over time* or across groups.

Chi-square tests aren't used often in some fields of study and work. Analysts may not apply this test for years after completing statistics coursework. This is not necessarily tied to the applicability of this test—in fact, chi-square tests have many advantages over tests of group comparisons. They can be a helpful tool to derive insights about where there is *disproportionality* in your data compared to expectations (either static values or relative proportions in your dataset).

**Figure 5.12 Breakdown of the equation for the  $X^2$  (chi-squared) test**

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

$\chi^2$  = chi squared  
 $O_i$  = observed value  
 $E_i$  = expected value

$$E_i = \frac{\text{Row Total} \cdot \text{Column Total}}{\text{Total Sample Size}}$$

Unlike each test we've covered, large sample sizes don't negatively impact the chi-square. The test statistic calculation compares each cell's expected value/proportion in a contingency table to the actual value/proportions. As your sample size increases, the test becomes more sensitive and more likely to detect small differences between observed and expected values.

Conversely, smaller sample sizes make it more challenging to detect differences, which can lead to fewer significant results.

## Application of the Chi-Square Test

Suppose you are an analyst on a product analytics team at a SaaS company that offers a subscription service to customers. You recently concluded an A/B/C test on three versions of a new recommendation engine, and you're interested in seeing whether any of the versions significantly impacted user engagement.

To test this hypothesis, you can use a chi-square test of independence to compare the frequencies of users who engaged with each version of the recommendation engine and the control group to see any differences by country. You start by comparing the observed number of users from each of the four countries' user bases included in the experiment (United States, Canada, England, France) broken out by experiment group. This ensures that we have appropriately stratified our sample before analysis.

**Figure 5.13 Summary table of frequencies of users who engaged with the feature broken out by country and experiment group.**

country	Canada	England	France	United States	Total
<b>recommender</b>					
<b>A</b>	415	443	441	434	1733
<b>B</b>	395	385	365	392	1537
<b>C</b>	467	444	421	398	1730
<b>Total</b>	1277	1272	1227	1224	5000

It's unclear whether any country disproportionately represents any of the experimental groups. A chi-square test for *independence* can be conducted to validate that the sample is appropriately stratified:

```
import scipy.stats as st      #A

chi_sq = st.chi2_contingency(assignments)      #B

print(f"Chi-square value: {chi_sq[0].round(3)}")      #C
print(f"p-value: {chi_sq[1].round(3)}")
print(f"Expected Frequencies:\n {chi_sq[3].round(2)}")

Chi-square value: 6.73      #D
p-value: 0.347
Expected Frequencies:
[[442.61 440.88 425.28 424.24]
 [392.55 391.01 377.18 376.26]
 [441.84 440.11 424.54 423.5 ]]
```

The chi-square test shows no significant differences in the number of users assigned to each experiment group. The chi-square value and corresponding p-value leave minimal room for ambiguity, as they are far from significant.

Next, you calculate two tables for your chi-square test: a table showing the number of users who *clicked* the recommendations by group and the number that you *expect* would click by group if there were no differences.

**Figure 5.14** Summary table of the number of recommendation clicks by group.

country	Canada	England	France	United States	Total
recommender					
A	117	131	119	80	447
B	82	76	77	105	340
C	113	120	133	143	509
Total	312	327	329	328	1296

We can see that 1296 out of 5000 users in the experiment clicked on the recommendations. This is a total proportion of 25.96%, which is used to calculate the *expected* click rates per group based on their initial sample size. This resulting summary table will be used as the *expected frequencies* for the chi-square test.

```
expected = assignments * .2596    #A
print(expected)                  #B
```

country	Canada	England	France	United States	Tot
recommender					
A	107.73	115.00	114.48	112.67	449.
B	102.54	99.95	94.75	101.76	399.
C	121.23	115.26	109.29	103.32	449.
Total	331.51	330.21	318.59	317.75	1298.

```
chi_sq = st.chi2_contingency(clicked.iloc[:-1, :-1], expected)
```

```
print(f"Chi-square value: {chi_sq[0].round(3)}")    #B
print(f"p-value: {chi_sq[1].round(3)}")
```

```
Chi-square value: 22.994    #C
p-value: 0.001
```

In addition to the test statistic and p-value, the chi-square test also includes a table of *expected frequencies* if there were no differences in relative group proportions. In most statistical software and packages, you must manually subtract the expected from observed frequencies to interpret the results. If your deliverable includes presenting these differences to stakeholders, you will likely want to color code the results with a heatmap for ease of interpretation.

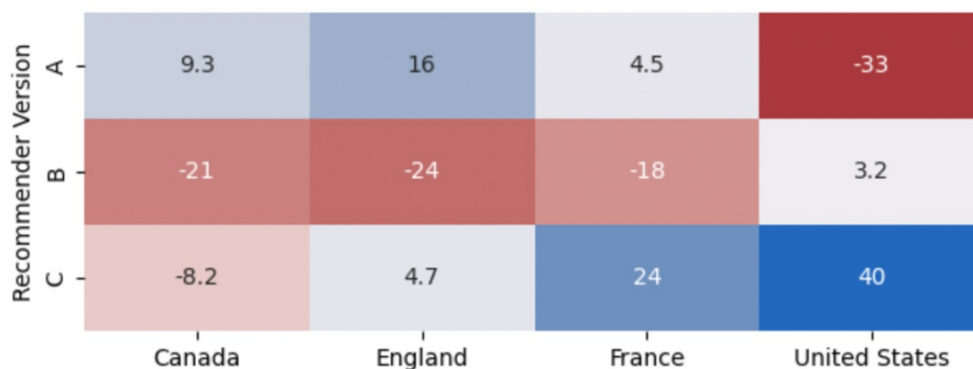
```
import seaborn as sns    #A
from operator import sub
```

```

diffs = list(map(sub, clicked.values, expected.values))    #B
diffs = pd.DataFrame(
    diffs,
    columns=assignments.columns,
    index=assignments.index,
)
sns.heatmap(    #C
    diffs.iloc[:-1, :-1],
    cmap="vlag",
    annot=True,
    cbar=False,
)
plt.xlabel("")
plt.ylabel("Recommender Version")

```

**Figure 5.15 Human-readable heatmap of differences between observed and expected click rates.**



With the same test, you can determine that you have appropriately stratified your sample *and* that each recommender performed differently in different countries. The observed dataset in the first test is then used as the *expected* dataset to compare the proportions of recommender clicks in each country.

## Considerations When Running a Chi-Square Test

If you choose to leverage the chi-square test in your work, there are several limitations to keep in mind when evaluating your methods and results:

- Chi-square tests are highly sensitive to small samples and typically require larger quantities of data than tests comparing continuous and ordinal data.
- Use a chi-square test with a test evaluating continuous or ordinal data.

You must choose an appropriate sample size far more strategically to minimize false positive results.

- Chi-square tests require an expected frequency of at least 5 in each cell to produce accurate results. If your categories have uneven distributions, check your expected frequencies to ensure they meet this minimum.
- The chi-square test doesn't include information about where disproportionality exists in your frequency tables. You will likely need to conduct a post hoc test to identify which pairs of groups are responsible for the significant differences. This application is often manual; for example, pairwise comparisons for a chi-square test can be performed in Python using a loop and re-conducting 2x2 comparisons for each pair. For this example, we'll compare the overall performance of experiment groups.

```
from itertools import combinations #A

pairs = list(combinations(assigned.iloc[:-1, :-1].index, 2)) #B
chisq_values = []
p_values = []

for p in pairs: #C
    c = clicked[(clicked.index == p[0]) | (clicked.index == p[1])]
    chi2, pv, dof, exp = st.chi2_contingency(c, correction=True)
    chisq_values.append(chi2)
    p_values.append(pv)
    print(p, pv.round(3)) #D

('A', 'B') 0.001
('A', 'C') 0.004
('B', 'C') 0.736
```

### 5.2.3 Summary

A lot of value in data exists outside of analyzing continuous or ordinal data. Performing high-value categorical analysis is a skill for which many data professionals lack expertise. This is not for lack of value—much of the untapped value in data exists in less than perfectly structured fields.

Knowing how to compare proportions of data against other variables or benchmarks can highlight where behaviors differ between segments of users and customers, determine whether a factory produces goods that pass quality

inspection at or above a static threshold (a proportion of goods pass), and much more.

## 5.3 Responsible Interpretation

You have just completed a large deliverable for your stakeholders detailing the performance of multiple initiatives. You followed every best practice in designing questions and enumerating hypotheses and diligently chose your statistical tests. The interpretation of your findings should be relatively straightforward at this point, right?

Well, not exactly. Your efforts have made a difference in how your stakeholders leverage your recommendations and how much value they will generate from your findings. Given that your work as an analyst involves *inference* and *predictions*, the accuracy of findings is never guaranteed. It's disturbingly easy to intentionally misconstrue results to support preconceived notions, special interests, and biased messaging. When leveraging statistics or machine learning, it's even easier to miss areas of nuance, previous research, or perspectives outside your own. There are countless ways to use statistics poorly and far fewer ways to use them well.

Instead of continuing this topic (your author can rant eternally about responsible, irresponsible, and malicious uses of statistics), I recommend reading *How to Lie with Statistics* [3]. This book was first published in 1954 and remained a timeless source of advice on applying healthy skepticism to the information you digest.

The topics in this section *are* frequently discussed in statistics curricula (Type 1/Type 2 errors, confounding variables). However, few classes include strategies and considerations for mitigating and avoiding these errors outside of a highly controlled laboratory environment (and even in those settings, mitigating these errors is often highly subjective). We'll discuss real-world strategies to limit these errors and leverage scientific best practices to set your project up for success across areas of study and practice.

### 5.3.1 Errors

Statistical errors refer to discrepancies between the true values of a population compared to the inferences and estimates made based on a sample. The term “error” doesn’t inherently mean that you’re doing something wrong; we’re ultimately using limited information to make educated guesses about events we haven’t measured and that may not have happened yet. There’s no way to eliminate errors from inferential statistical approaches, but understanding the types and sources of errors is crucial for maximizing the accuracy and reliability of your results.

## **Type 1 Errors**

A Type 1 (false positive) error refers to rejecting the null hypothesis when it’s actually true. These errors typically occur when a statistical test identifies a significant difference or effect when, in reality, the difference isn’t meaningful or isn’t present in the broader population.

Some examples include:

- A patient receives a positive result on a COVID-19 test, even though they are not infected with the virus.
- A person who did not commit a crime is taken to trial and found guilty.

In both examples, the false positive result has a drastic real-world impact on the individual. The person found guilty may be fined, imprisoned, and experience long-term disruptions to their economic and social status. The patient may be required to quarantine for an extended time, which stresses their finances, family, or other areas of their life. The consequence of false positive results in analytics is often far more challenging to pinpoint. Even if you can identify when it occurs, estimating the impact is often more theoretical than tangible.

There is no definitive guide for detecting and eliminating Type 1 errors. In an ideal world, you would either be able to compare the results from your sample to the entire population or have the resources to repeatedly test the same phenomenon and see if somebody can replicate your initial findings with a different sample. However, there *are* conditions in which Type 1 errors are more likely to occur:

- **Tests with an alpha level (significance threshold) at or above 0.05** are likelier to return a Type 1 statistically significant result. A p-value of .05 (5% probability of obtaining differences at least as large in future samples) may seem low, but a 5% chance is one out of every 20 based on random variation alone. A p-value of .1 (10%) can occur every 1 in 10 tests.
- **Tests with multiple comparisons** are more likely to produce significant results, even when you use a corrected p-value. When comparing large numbers of categories or interaction effects, your chances of finding a significant result increase by the sheer number of comparisons alone. If you're using tests such as two-factor ANOVAs, limit the number of groups you compare as much as possible. In my work, I limit two-factor ANOVAs to a 2x3 design (one variable with two groups and another with three groups).
- Both **small or very large sample sizes** have a higher chance of returning false positive results. As demonstrated in chapter 2, many test coefficients will eventually cross the statistical significance threshold with a large sample size increase, even if no actual difference is reflected in the population.

Figure 5.16 Your author implores you not to run excessive multiple comparisons.

#### Average Time on Website by Country and Age Cohort

	Age 18-24	Age 25-29	Age 30-34	Age 35-39	Age 40-44	Age 45+
U.S.A.	34.6	46.2	23.9	55.8	32.2	63.9 *
Canada	20.8 *	45.4	18.5	52.3	25.8	41.7
France	59.0	28.1	66.4 **	21.9	39.5	54.7
England	30.8	49.2	24.0	61.4 *	40.9	57.6
Spain	19.0 **	48.2	35.8	64.2 *	22.7	52.8

\*  $p < .05$       \*\*  $p < .01$

Suppose you suspect your statistically significant results may result from a Type 1 error. What do you do? There are some straightforward steps you can take to limit the occurrence of Type 1 errors in your work:

- **Choose appropriate statistical test(s):** Yes! Everything we've discussed in this and the previous chapter—checking assumptions, transforming data, and choosing a test (or multiple tests) based on the characteristics of your data—is a *huge* step toward mitigating Type 1 errors.
- **Set the alpha level (significance threshold) conservatively:** The most commonly used alpha level of 0.05 is *probably* best not to exceed in most cases. While interpreting this threshold flexibly is often beneficial, I *don't* recommend setting higher approximate thresholds (e.g., 0.1). As we've discussed, you may want to set an alpha level below 0.05 (e.g., 0.01) when the stakes of reporting a false-positive result are especially high.
- **Interpret your p-value dynamically:** On the flip side of setting a conservative alpha level, many analyses may benefit from flexible interpretations of the p-value. Suppose you are analyzing data on human behavior (e.g., time a user spent completing a workflow in your app). In that case, you can set a threshold of 0.01 if you believe ensuring that one version of the workflow design is better is vital. However, you probably *won't* want to throw away both versions of the workflow if your analysis yields a p-value of 0.015 if you have reason to believe your results are meaningful (e.g., users who saw version A of the workflow had *consistently* lower times to complete the workflow for the duration of the experiment and across user types). As you grow your expertise in a domain, you will develop confidence in your ability to judge the appropriate thresholds for the phenomena you are analyzing.
- **Replicate your findings:** The most reliable way to validate your findings is to replicate your results with different samples. If you can do so in your work, replication can be performed or estimated in several ways:
  - **Retest your results with a new sample:** If you can, conducting the same assessments and tests with a new sample over time adds significant weight to the validity of your results and recommendations. If you expect to collect new data over time for

your analysis, test the next set separately before incorporating them into the larger sample.

- **Bootstrap your results:** *Bootstrapping* is a technique for estimating the parameters of your population by taking *smaller samples with replacement* from your entire sample. Bootstrapping is flexible and can be applied to most statistical tests. The application of bootstrapping far exceeds resampling for a statistical test. I recommend looking into the many books and resources on bootstrapping approaches and uses. To mitigate Type 1 errors, we'll cover a specific example that can be applied to most tests we have covered. Let's take a sample with  $n=500$  and draw 1000 samples of  $n=50$  with replacement to build a distribution of  $t$  values. The `scipy` package has an implementation of bootstrapping we can use with its other functions.

```
import numpy as np #A
from scipy import stats as st
import matplotlib.pyplot as plt

X1 = np.random.normal(loc=75.5, scale=6.2, size=500) #B
X2 = np.random.normal(loc=76.2, scale=6.5, size=500)

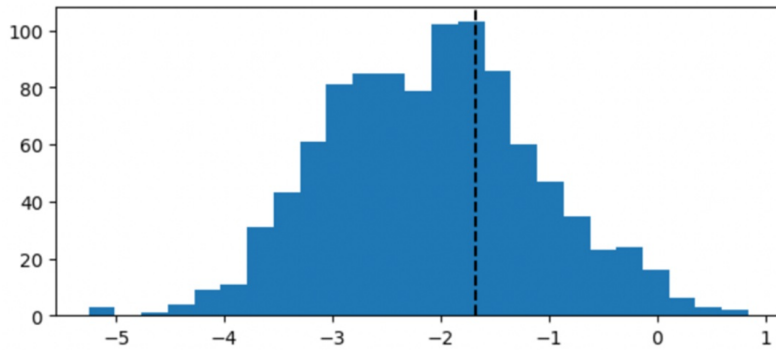
def t_stat(X1, X2): #C
    return st.ttest_ind(X1, X2)[0]

t_values = st.bootstrap( #D
    (X1, X2),
    t_stat,
    n_resamples=1000,
    batch=50,
    method="basic",
    vectorized=False,
    random_state=99,
)

t_crit = -st.t.ppf(q=0.95, df=49) #E

plt.hist(result.bootstrap_distribution, bins=25) #F
plt.axvline(t_crit, color="black", linestyle="dashed")
```

**Figure 5.17** The majority (but not all) of bootstrapped samples have  $t$ -values above the absolute value of the critical threshold



In addition to putting these recommended guardrails in place, mitigating Type 1 errors requires leveraging your domain experience. As you build expertise in an area, you will become familiar with previous research and the group differences and trends reported in those findings. Over time, that experience will provide a comprehensive perspective on variables that may be missing in an analysis, which we will discuss further in the section on confounding variables.

## Type 2 Errors

A Type 2 error occurs when a test result yields a negative result (e.g., a non-significant p-value) when the alternative hypothesis is true in the population. This leads to the conclusion that there are no significant differences between groups or relationships between the variables compared in your analysis.

In practice, Type 1 errors are given more attention than Type 2. However, Type 2 errors can produce just as consequential outcomes in the real world as Type 1. Type 2 errors can look like the following:

- A patient receives a *negative* result on a COVID-19 test, even though they *are* infected with the virus. They do not quarantine or take measures to prevent the spread of the disease and infect several people. One of these people infected becomes critically ill from the disease.
- A person who committed a violent crime is taken to trial and found innocent. The person is released and goes on to commit additional violent crimes.
- A public health study using several million dollars in research grants is conducted over two years. The results come back as negative, even

though the treatment did produce a long-term impact on the participants. The potential benefit to the broader population and the value of the research grants are not realized.

Type 2 errors are more likely to occur under conditions you can control for or detect as an analyst. Some causes and factors contributing to Type 2 errors include the following:

- **Small sample sizes:** a true difference can be challenging to detect when your sample size is too small for the test you are using. If you're unsure how many data points to capture, consider conducting an *a priori power analysis*, as discussed in chapter 4.
- **Inappropriate statistical test:** if you do not meet the necessary assumptions of a statistical test, you're far more likely to generate both Type 1 and Type 2 results. Additionally, if you choose the wrong test for the type of study design (e.g., an independent samples t-test for a repeated measures design), you will likely generate inaccurate results that confuse your stakeholders.
- **High variability:** in addition to the general shape of the distribution, parametric statistical tests assume that your data has a standard variance (measured by the distribution's *kurtosis*). A dataset in which both groups have very high variance will likely return either a false-positive or a false-negative result. In these cases, you may want to use a non-parametric test even if the dataset is normally distributed.

## Confounding Variables

Confounding occurs when the effect of one variable on the outcome of interest is mixed up with the effect of another variable that is often not measured as part of the analysis. In other words, a confounding variable is a *covariate* of your model (a factor that varies *with* your independent and dependent variables) that you have not accounted for.

This can lead to incorrect or misleading results about the relationship between the variables. They're also exceedingly common, given the complexity of most analyses we do relate to human phenomena. Some examples include:

- A researcher finds a relationship between sleep duration and academic performance. However, the study didn't control or account for caffeine consumption related to sleep duration and academic performance.
- An analyst finds a relationship between the geographic location and online product preferences of website users without noting that their median age differs drastically by location.

Controlling for confounding variables is generally accomplished through a combination of strategies in the design of your research and the statistical approaches to your work. Taking these steps early in your analysis process can increase the validity and reliability of your results, saving resources and maximizing value from each analysis you invest time in.

- **Peer-review your initial study design:** this is likely the most impactful step you can take to mitigate confounding variables. Asking others with *domain-specific knowledge* if there's anything you might be missing or didn't account for is an underrated and impactful step you can take to alleviate blind spots in your research. As with other areas we've covered —*not accounting for every possible variable does not mean you failed as an analyst*. Peer review and collaboration are a necessary part of science and analysis.
- **Build a knowledge base of known covarying factors:** where possible, performing and circulating analyses that teach your organization about trends and differences in key covarying characteristics can help better structure your questions. If you, your colleagues, and your organization generally know where key types of users and customers differ on behavior, needs, and outcomes, you will have an easier time accounting for confounding variables. We will discuss this at length in chapter 11.
- **Stratify your randomized samples:** if you are conducting experiments where you randomly select your participants or users (e.g., A/B tests), checking and stratifying the random sample according to *known* confounding variables can help reduce their influence.
- **Leverage a statistical model that accounts for covariates:** when using a t-test, ANOVA, or non-parametric alternative to these tests, you generally are limited to sampling and randomization methods to account for covariates. Techniques such as regression analysis and ANCOVA (analysis of covariance) allow you to include covariates and *control for*

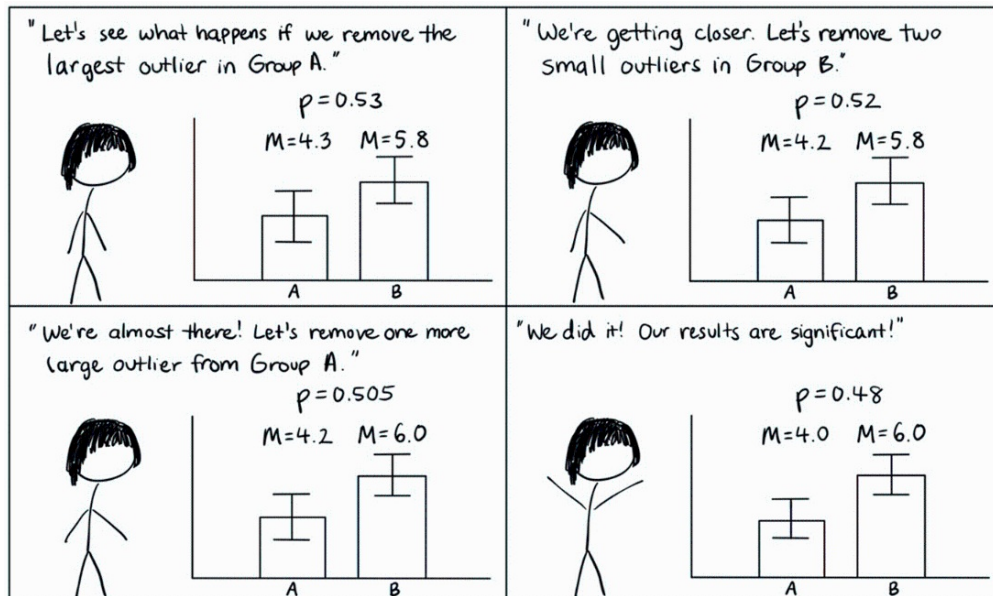
them as part of your model.

### 5.3.2 P-hacking

The previous section discussed issues and situations that can unknowingly impact the accuracy and quality of your results. They're less often due to the intentional action of the researcher but rather the omission or a blind spot concerning study design and variable selection. This section covers the steps that *you* as an analyst can (but should *not*) take that amount to direct manipulation of your results.

P-hacking, also known as data fishing, is an approach to data analysis where researchers and analysts manipulate the data and statistical tests to obtain results that reach statistical significance, often leading to reporting effects that are not reflected in the population. In other words, p-hacking involves continually testing multiple hypotheses, designs, statistical tests, outlier removal methods, or transformations until *something* reaches statistical significance. This is the analyst's version of "*throwing it against the wall until something sticks.*"

**Figure 5.18** P-hacking involves testing a dataset indefinitely until something returns a significant value.



P-hacking can take a wide range of forms. Some of these approaches differ if you're in an academic or industry setting, but any of the following are possible. As analysts, it's our responsibility to be diligent in where the following approaches may be present in our work:

- **Conducting large or indefinite permutations of statistical tests until a desired result is achieved:** this does *not* refer to the steps commonly involved in exploring a dataset to report descriptive findings (e.g., “70% of users visit the app once per week or more”). P-hacking with statistical tests refers to conducting numerous statistical tests to make inferences about the population (e.g., “changing the layout of the app will cause 10% more users to visit the app at least once per week”). At an alpha level of 0.05, at least 5% of the tests you run will reject the null hypothesis and risk committing Type 1 errors. The probability of any test returning a significant result is *additive* (the more tests you run, the more likely you are to find something), as shown in figure 5.16. Thus, running two tests increases the likelihood of rejecting the null to 10%, 15% with three tests, and so on. To mitigate these risks, the following approaches should be avoided:
  - **Changing the variables of the analysis:** during exploratory data analysis, exploring relationships between many variables (e.g., all questions in your survey or study) is a sensible choice. If you're preparing a statistical test, changing your variables and altering your hypotheses and interpretations for those new variables can degrade the quality of your work. This is especially true if you work at an organization with access to many variables at your fingertips—it can be very tempting just to *keep trying new things* until something comes back significant.
  - **Changing the study design:** proposing multiple study designs in your analysis plan at the early stages is an excellent way to approach a problem from numerous angles. You're unlikely to approach your work with the same rigor or accuracy if you're trying every permutation possible to achieve statistical significance.
- **Manual removal of data points:** this refers to the selective removal of individual data points in a dataset (usually outliers) to alter the shape of the distribution and achieve a desired result. This is *never* an appropriate approach to your work. Do not do this.

- **Manual modification of the sample:** this refers to the revision of the existing sample by any of the following:
  - **Non-random or random selection of a subsample that achieves the desired result:** as shown in figure 5.19, bootstrapping of a sample indicates that subsamples will range in their test statistic values and p-values. By random chance alone, an individual subsample can easily fall above or below a critical threshold that does not reflect the difference within the entire sample.
  - **Exclusion of subgroups within the sample:** there is usually value in identifying valuable subgroups and segments among your users, customers, or participants. The time to do that is *not* at random to generate a statistically significant result.

P-hacking is usually performed with the intention of publishing and reporting results. Analysts and researchers commonly experience pressure to generate *something* as far as insights to receive grant funding for a lab or organization, continued employment in academic settings, or the maintenance of relationships with stakeholders.

As analysts and researchers, I understand that we are often under tremendous pressure to prove our worth and value to an organization. Usually, the organization's success relies on the results we are expected to produce.

But p-hacking is *never* worth it. The incorrect and misleading conclusions generated through this approach *will catch up to you and have tremendous negative consequences on others who rely on your findings*.

## 5.4 Summary

- The majority of parametric statistical tests were developed 100 years ago or more for specific types of analysis and have since become the dominant method of choice for most fields of study and work. However, these tests are *not* ubiquitous and are not always the best choice for every analysis.
- Non-parametric statistical tests are alternatives to parametric statistical tests that do not make assumptions about the underlying distribution of the data. They can be used with a broader variety of data types than

parametric statistical tests (you can use either continuous or ordinal data), providing analysts with a wide range of options.

- There is a wide range of non-parametric tests for group comparisons. The most common are (1) the Mann-Whitney *U* test for comparing two independent samples, (2) the Wilcoxon Signed-Rank test for two dependent samples, (3) the Kruskal-Wallis test for two or more independent samples, (4) and the Friedman test for three or more dependent samples. Each of these tests compares the *ranks* and *relative positions* of the data points in each group rather than making calculations based on the actual values of the data points.
- The chi-square test is a common test that you can use to compare one or two categorical variables. The test compares the difference between the observed to expected frequencies for each row and column in a *contingency table*.
- As an analyst, it's essential to evaluate your study design and results for possible type 1 and type 2 errors as well as confounding variables that can reduce the accuracy and validity of your results.
- Part of being a responsible analyst involves ensuring the integrity of your approaches, even when your stakeholders aren't watching. Choose your design ahead of time, and don't be afraid to report non-significant results for the long-term accuracy of your work and the trust of your stakeholders.

## 5.5 References

[1] J. W. Grice, "Observation Oriented Modeling: Preparing Students for Research in the 21st century," *Comprehensive Psychology*, vol. 3, p. 05.08.IT.3.3, Jan. 2014.

[2] G. Y. Kanyongo, G. P. Brook, L. Kyei-Blankson, and G. Gocmen, "Reliability and Statistical Power: How Measurement Fallibility Affects Power and Required Sample Sizes for Several Parametric and Non-parametric Statistics," *Journal of Modern Applied Statistical Methods*, vol. 6, no. 1, pp. 81-90, May 2007, doi: <https://doi.org/10.22237/jmasm/1177992480>

[3] D. Huff, *How to lie with statistics*. New York [U.A.] Penguin Books, 1981.

# 6 Are you measuring what you think you're measuring?

## This chapter covers

- The theoretical underpinnings of effective measurement
- Identifying the strengths and limitations of a measurement
- Reliably measuring information about your concept or process
- Ensuring your measures are valid representations of your concept or process

Have you ever responded to the following question? If not, perhaps you've seen it on a website or a marketing email you received from a service you use.

**Figure 6.1 The Net Promoter Score is used across industries and products.**

On a scale of 0-10, how likely are you to recommend our business to a friend or colleague?										
0	1	2	3	4	5	6	7	8	9	10

The aggregate score calculated with this measure is known as the *Net Promoter Score* (NPS). NPS has been used for 20 years across industries, business sizes, and specializations to gauge *customer satisfaction* and *loyalty* to the business's products and services.

NPS is calculated using the question's 11-point rating scale from 0 to 10, where 0 means "not at all likely" and 10 means "extremely likely." Scores are categorized into three groups:

**Figure 6.2 NPS responses are categorized into three groups: promoters, passives, and detractors**

On a scale of 0-10, how likely are you to recommend our business to a friend or colleague?



- Customers who respond with a 9 or 10 are classified as **promoters**. These are considered enthusiastic customers who are likely to continue using a business's products or services and refer others to the company.
- Customers who respond with a 7 or 8 are classified as **passives**. These are considered neutral by neither actively promoting nor discouraging others from engaging with the business.
- The **detractors** category comprises customers believed to be dissatisfied with the business's products or services and, therefore, likely to discourage others from engaging with the business.

The **NPS** is calculated by subtracting the percentage of detractors from the percentage of promoters, resulting in a standardized score from -100 to 100. The resulting number is used as a *benchmark* against competitors, peers, and unrelated businesses.

How would you rate this approach to measurement? NPS is derived using a straightforward calculation that results in a standardized score. It's simple, easy to understand, and widely adopted. Does that make it a good measure?

Not quite. The NPS is far from perfect, and many businesses and organizations fail to derive value from its measurement. Let's look at some of those limitations:

- NPS is **overly simplistic**, using only one question to infer multiple aspects of customer experience, satisfaction, and loyalty.
- The measure assesses a general likelihood to recommend the business or its services with a **limited ability to understand the reasons for the score provided**. Customers may be dissatisfied with the product, customer service, or a specific feature.
- **NPS interprets its 11-point scale as a universal standard**: People don't treat numerical scales equally, which we will discuss in this

chapter. NPS assumes that all customers likely to recommend the business will respond with a score of 9 or 10.

- The approach to categorization makes the score **susceptible to minor fluctuations**. If a few customers answer with an 8 instead of a 9, they are not included in the score, even though their answers may not represent a material difference.
- NPS focuses only on customers who respond to the survey and **ignores the response rate** and characteristics of customers who refuse to answer.
- Although standardized, **you cannot compare NPS 1:1 across businesses with different customer bases**. Not every business *can* be recommended or referred to a “friend or colleague” in the same way.
- Overemphasis on NPS can **lead to long-term neglect of other customer satisfaction indicators**. This tunnel vision can have detrimental impacts on the performance of the business.

Ideally, NPS should be used as **one possible measure** of customer satisfaction among many; if you cannot do this, it’s better not to use it.

In the following sections, we’ll discuss approaches to assessing a concept that you’ve operationalized, measuring it, and identifying pros and cons with various approaches to measurement.

## 6.1 A Theory of Measurement

Transforming abstract phenomena into something you can observe and quantify is not an easy task. Anyone can develop a survey from scratch and deliver it to an audience of their choosing, but doing so risks capturing inaccurate information or no valuable insights. To measure what you think you’re measuring, it’s important to apply key principles from **measurement theory** to generate high-quality data.

**Measurement theory** is a framework for assigning numerical values to abstract concepts, events, or objects. This framework provides analysts with standardized criteria to develop and evaluate their measures. If you are creating a measure of stress, you have decades of recommendations available on how to operationalize stress, develop and administer a questionnaire,

appropriately quantify degrees of stress, and analyze the results. Stress is inherently subjective and thus can be challenging to quantify. To ensure your results are *interpretable* using quantitative methods, an accurate reflection of people's experiences, and aligned with other definitions of this concept, you can apply best practices from each component of measurement theory:

- **Conceptualize** or define the phenomena of interest. This involves enumerating the components of what you're interested in to **operationalize** the concepts. We cover these steps in chapter 2.
- Develop an appropriate **measurement scale**, identify appropriate measures among **readily available data**, and identify **proxy measures** where direct data capture on your phenomenon is impossible.
- **Review and test your measure's limitations** to ensure they remain consistent, accurate, and stable over time.

Peer-reviewed papers, textbooks, and other resources on measurement theory typically cover it from the perspective of social sciences and psychological research. Researchers assume that you are primarily developing questionnaires, performance evaluations, and other rating scales designed for participants to self-report behavioral and cognitive processes. However, you can apply most steps to develop measures to the various forms of data in most organizations (e.g., data on business processes recorded in a warehouse).

As in previous chapters, we will continue to draw from best practices in the social sciences. If you're tracking user actions on your app, customer sentiment, customer renewals, time to resolve support tickets or nearly anything we've covered up until now, you are measuring a **behavior** or a **cognitive process**. By definition, psychology is the study of those two things. If you've taken a psychology class during your education, this chapter is an excellent place to pause and review the materials from your curriculum and apply the lens of a behavioral scientist in your work.

### 6.1.1 Conceptualizing a Measurement

As an analyst, much of your job at the early stages of a project involves **clearly defining and operationalizing a concept so that you can measure**

**it.** If you're starting with a new question that your organization hasn't yet investigated, your first step is determining if you *can* agree on a definition for measurement. Even with agreement, you may find that existing measures familiar to your organization may be limited in their applicability to a new project. As we saw with NPS, just because something has been measured a certain way for a long time or is widely adopted does not make it the best choice for delivering value.

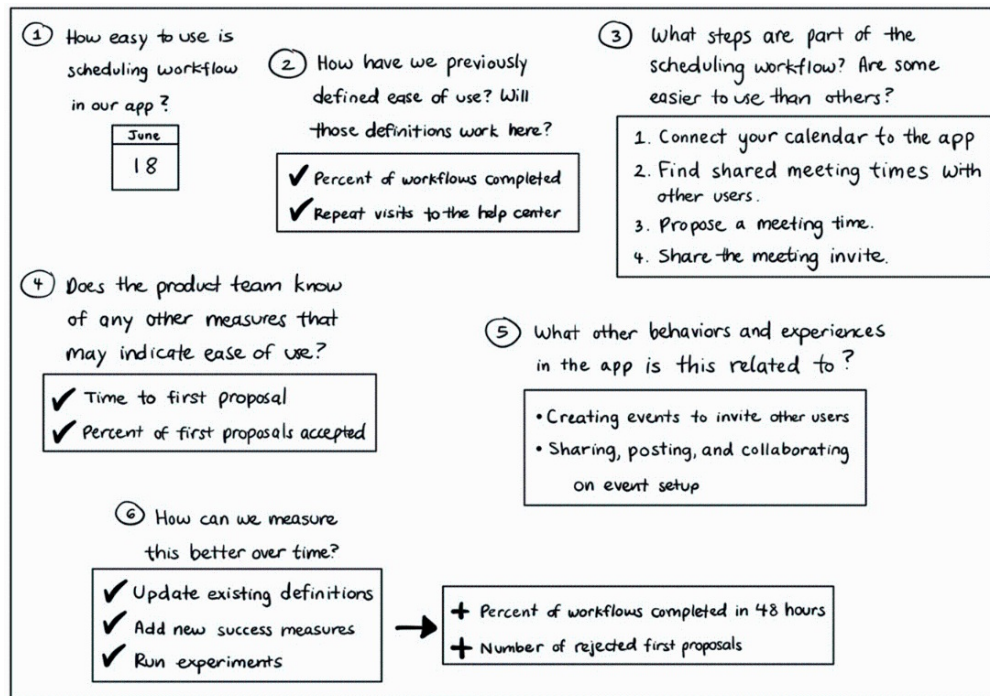
**Conceptualizing** an abstract phenomenon or construct involves providing explicit definitions of the components, dimensions, and characteristics of that construct. This step enables you to create a precise understanding between your team and stakeholders to ensure everyone agrees on what is being measured.

You can break down the process of conceptualizing a measure into the following steps:

- **Identify the phenomenon and the rationale for analyzing it:** What are you or your stakeholders interested in learning more about? Why is it important? Why now?
- **Review existing conceptualizations:** Are there ways this concept has been previously defined? You will likely want to gather appropriate context from peer-reviewed literature, industry standards, peer organization practices, and previous work within your organization.
- **Specify key components of your concept:** Often, measuring a concept involves aggregating multiple behaviors, cognitive processes, sentiments, or data sources. You and your stakeholders will likely need to align on compiling available information into a comprehensive strategy for guiding decisions.
- **Align on the definition of your concept:** Once you have a definition, it's essential to ensure it aligns with how stakeholders see the concept. Doing so at this stage will minimize confusion and save you time in the project lifecycle.
- **Identify related concepts:** How does your concept relate to other similar processes? Do you expect to find any relationships or associations as you measure them?
- **Iterate on the definition:** Your first concept definition will likely not be

the last. As you gain more information about the limitations of your first iteration, you will want to continually improve in alignment with your stakeholders.

**Figure 6.3 Steps to defining and refining the concepts you want to measure**



Clear definitions are necessary to use your organization's data and resources appropriately. However, not everything can be neatly measured. Many concepts lack consistent, agreed-upon definitions, making them difficult to quantify (e.g., intelligence, creativity). Even with abundant resources, you may not develop a measure that all your stakeholders agree on. Measuring subjective concepts is challenging but necessary to understand complex human processes.

**Figure 6.4 Can you measure complex, subjective concepts such as creativity?**



**Let's introduce our case study for the chapter:**

Arthur is an analyst on the Human Resources team at a large supply chain company with over 10,000 employees. The company is interested in understanding their employees' engagement in their work, the company culture, and whether engagement can predict retention and attrition.

Arthur meets with his manager to understand the rationale for the project and its value to the company: Is the company aiming to improve its headcount planning for next year? Is it trying to reduce employee turnover across the company or within specific departments?

Arthur's manager tells him the company has seen increased turnover in some departments. They want to use the insights from this project to get ahead of any potential future attrition.

Arthur searches peer-reviewed literature and finds several measurement scales assessing employee engagement and job satisfaction. He works with his team's manager to align on critical components of employee engagement. Together, they decide to investigate the following topics: (1) satisfaction in the current role, (2) perception of workload, (3) interest in company events, (4) satisfaction with salary and benefits, (4) productivity in current role, and (5) performance in current role.

Arthur hypothesizes that similar concepts, such as motivation in an employee's current role, will be associated with employee engagement. He also hypothesizes that each of the components will correlate with each other.

## **6.2 Choosing a Data Collection Method**

Once you and your stakeholders align on how to break down the concepts, it's time to translate them into measurable indicators for your analysis. Your goal at this stage is to develop clear *observable, measurable* indicators of your concept to use in your work. This process of operationalization (discussed in chapter 2) enables you to decide on the best method for *collecting* the data you need.

There are numerous ways to categorize data collection methods. If you completed statistics or research methods coursework, you might be familiar with the distinction between *direct measures* (e.g., a person's height) and *indirect measures* (e.g., historical archives). This distinction **categorizes data based on the actions of the researcher**, i.e., whether the researcher measured and recorded the data *directly* as opposed to using previously existing information.

## 6.2.1 Types of Measures

Instead of categorizing data based on how the researcher collected it, we will be focusing on how the user or participant *provided* the measure of interest.

- **Self-report measures** such as questionnaires, interviews, and feedback forms provide information that users or participants directly input. Surveys are the most common self-report measure, ranging from a single question (e.g., the NPS question) to lengthy, comprehensive assessments of multiple concepts. Any data recorded *directly* by your users (e.g., in-app feedback form, user preferences, and settings) is included in this category.
- **Behavioral measures** refer to data collected via *direct observation or measurement*. For example, a person's height is directly measured using a standardized instrument (a measuring tape). An increasing volume of primary data is available in many organizations—sales records in a database, fitness data on your phone, and job applications are all direct measures of *behavior*.

In practical scenarios, most organizations possess a blend of self-reported and behavioral data, varying preferences for either type depending on the team involved. For instance, tech companies typically have detailed behavioral data reflecting user interactions with their software. Consequently, their product and analytics teams depend heavily on such data for strategic decision-making. However, the same organizations' marketing and user research teams might lean more toward customer surveys, interviews, and focus groups for insights.

As an analyst, your vantage point provides a more comprehensive

understanding of the data spectrum within an organization compared to stakeholders focused on specific domains. This unique position allows you to act as a consultant, advising stakeholders on the most suitable measurement approach to address a question. As discussed in previous chapters, your approach will be influenced by the question, the organization's available data, and the most efficient measurement method considering your resources.

**Table 6.1 Examples of self-report and behavioral data you may encounter at a software company**

Self-Report Data	Behavioral Data
Employee satisfaction survey	User interaction logs (e.g., clicks, page views)
User feedback survey on software features	Time logs of employees on project management software
Self-assessments of productivity	Reports of software bugs or issues
User surveys on software ease of use	Version control history
User biographies in the software	Helpdesk ticket data

## Self-Report Measures

When you recommend a measurement approach to stakeholders, it's valuable to enter the conversation with an understanding of the pros and cons of each approach. Self-report measures offer several key advantages:

1. **Easy to administer and start from scratch:** self-report measures can be quickly developed and administered to participants. They also make it far easier to gather insights if no data about the phenomenon you are measuring exists.
2. **Ability to assess detailed, personal experiences:** self-report measures allow you to ask participants about their thoughts, feelings, and experiences about a phenomenon that is difficult or impossible to assess using behavioral measures.
3. **Flexibility:** self-report measures are adaptable to the needs of an organization and analytics team. They can be gathered online, in person, and at various milestones relevant to your organization.

Self-report measures also have several *disadvantages* that are important to

share with your stakeholders:

1. **Response bias:** through no fault of their own, participants' responses will vary by a range of social factors that bias the data you collect. A person's cultural background, the setting in which data is collected, and their perception of what response is socially desirable will all influence the quality of the information provided.
2. **Measurement bias:** minor changes to survey measurements can drastically impact the responses you receive. If you ask the same question with slight variations in wording, response scales (e.g., a scale from 1 to 5 vs. 1 to 9) and the descriptors used on the scale (e.g., "Strongly Disagree" to "Strongly Agree") can result in wildly varying response patterns.
3. **Indirect measurement:** self-report measures are limited because you *cannot directly measure the phenomenon of interest*. When surveying people, you are limited to assessing people's *perceptions* of a phenomenon.

The opening of this chapter introduces the Net Promoter Score question as an illustration of a self-report measure. When employing this tool, researchers must thoughtfully weigh the inherent pros and cons tailored to this measure and how it's used.

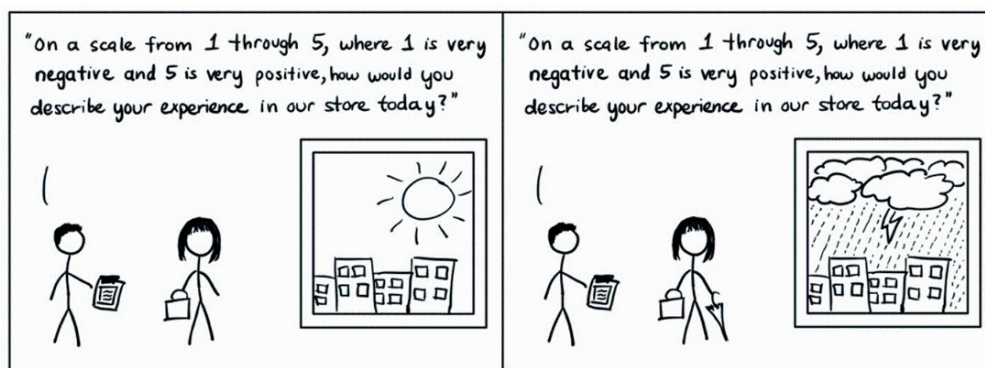
**Table 6.2 Advantages and disadvantages of the self-reported NPS question**

Advantages	Disadvantages
Data is easily stored in a spreadsheet	Answers on the 11-point scale will differ between respondents independent of their sentiment based on immeasurable personal factors
Assesses personal beliefs about the likelihood of recommending a product or service	Changes to the question's wording (e.g., company name, company type, part of a business's offering) will drastically impact results.
Quickly asked online using a survey tool or in person as a quick survey. Often built into marketing and email	Measures respondents' perception of their likelihood to refer, not their actual referral behaviors.

campaign software.

While they offer simplicity, flexibility, and the ability to access personal experiences, self-report measures are susceptible to numerous forms of bias outside your control. When recommending a measurement approach, it's crucial to carefully weigh these factors and consider the specific needs and goals of the stakeholders involved.

**Figure 6.5** The quality of data from self-report measures depends on numerous factors outside of the control of the researcher or analyst.



## Behavioral Measures

By circumventing the need to ask users or participants to report on the phenomenon of interest, **behavioral measures** offer an opportunity to bypass the limitations of self-report measures. Overall, they provide the following advantages:

1. **Direct, objective measurement:** a behavioral measure involves directly recording when a phenomenon or action occurred. The participant or user of interest isn't *asked* about the occurrence, eliminating the bias associated with their self-report.
2. **Precisely timed:** behavioral measures make it easier to associate a precise time with the occurrence of an event. For example, a researcher can directly record a timestamp if they are observing a participant, or a timestamp can be recorded in a database when a user clicks a button.
3. **Comprehensive measurement:** behavioral measures can be combined with other behavioral or self-report data about the same user or

participant. For example, a set of pre-defined behaviors a researcher can observe or multiple user activities can be recorded in a database and synthesized into a picture of their usage of your app.

Each advantage of behavioral measures *enhances the value of the data and insights you can provide*. So why don't we abandon self-report measures entirely? Not every research team or organization can capture behavioral measures for *every* topic they're interested in. Even for those who do, they come with several distinct disadvantages:

1. **Intrusiveness:** many behavioral measures capture information about users and participants through direct observation. A researcher may watch a participant looking for specific behaviors (common in market research observing people in retail stores), or a cookie might be installed on a user's browser to track their clicks and page views. This can often occur without the consent of the person being observed.
2. **Resource-intensive:** most behavioral measures require more resources to capture data than self-report measures. A survey can be developed using a free tool, distributed online, or via an advertisement, and have data captured in a spreadsheet. Participant observation requires the time and labor of a research team. Online tracking measure may require third-party software, compute, and storage resources and the effort of a data engineering team to structure the data for analysis.
3. **Legal and ethical concerns:** many businesses and organization are subject to regulations associated with the collection and retention of data. If you are working with legally confidential data (e.g., Protected Health Information or PHI), your organization may be required to have strict controls around its usage that may limit the types of analyses you can perform. Further, ethical implications may be associated with the data you collect and potential information you can conclude about people via behavioral observation. We will discuss this more in-depth in chapter 8.

To create an equivalent behavioral measure of the NPS question, a researcher must *directly* assess how often users refer friends or colleagues to the business. This quickly becomes complicated – do you count the number of referrals received? How do you count how often they make verbal

recommendations to others that don't lead to new business? How do you tell if they make recommendations to competitors instead?

Short of invasive surveillance of your customers, you're unlikely to get a good read on this phenomenon as a direct measure of behavior. This data type may be more accurate and precise, but it's not always feasible for what you need. An analyst's job involves balancing the choice of measures based on the best use of resources available to their team and organization.

**Let's return to our case study for the chapter:**

Arthur begins the project's next phase by identifying available measures at the organization to assess the components of employee engagement they decided to focus on. He evaluates whether these measures are direct observations of employee actions or employees' subjective evaluations of their experiences.

1. Satisfaction in the current role: Arthur discovers that the company conducts annual employee satisfaction surveys. He reviews the survey items and finds questions related to job satisfaction that can be used as self-report measures for this analysis component.
2. Perception of workload: The company has data on employees' work hours and workload through project management software. Arthur identifies this data as a behavioral measure. Additionally, he finds self-report measures in the employee satisfaction survey that ask about workload perceptions.
3. Interest in company events: Arthur learns that the company maintains attendance records for company-sponsored events. He decides to use this attendance data as a behavioral measure for the *interest in company events* component.
4. Satisfaction with salary and benefits: The employee satisfaction survey includes questions related to satisfaction with compensation and benefits, providing Arthur with self-report measures for this component.
5. Productivity in the current role: Arthur finds that the company has data on employee productivity, such as key performance indicators (KPIs) and project completion rates. These metrics serve as behavioral measures for the productivity component.

6. Performance in the current role: The company's performance management system tracks employee performance through regular evaluations and ratings. Arthur recognizes that while this data is not self-reported by the employee, it's subject to the biases of a self-report measure since the employee's manager reports it.

Arthur determines that each self-report measure associated with the annual employee satisfaction survey is readily available via a survey tool where he can download the dataset as a CSV. Each behavioral measure is housed in a different system and requires additional effort to extract and combine the data with self-report measures.

## 6.2.2 Constructing Self-Report Measures

What's the difference between the following measurement scales when used in a survey?

**Figure 6.6 Varying methods of presenting the same response format**

1	2	3	4	5
Strongly Disagree	Somewhat Disagree	Neutral	Somewhat Agree	Strongly Agree
Strongly Disagree	Somewhat Disagree	Neutral	Somewhat Agree	Strongly Agree
1	2	3	4	5
Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree

Slight variations in wording, response formats, and rating scales can substantially impact the responses you receive on a self-report measurement. In the social sciences, researchers will often directly tweak, manipulate, and test rating scales and response formats to get as close as possible to a read on an abstract concept. These small changes and tests are crucial to minimizing error and deriving value from self-report data.

## Wording the Question

As an analyst, your efforts to *intentionally* craft well-structured, unbiased, and precise questions will enable you to minimize ambiguities, response biases, and other potential issues that may compromise the quality of your data. This attention to detail in self-report measurement design will ultimately lead to more accurate and reliable findings, enabling you to make well-informed decisions and effectively address the research objectives or practical challenges.

Consider the following issues to avoid when wording your question or evaluating an existing measure. Example questions are included for reference:

- Respondents may provide inaccurate responses if a question is **ambiguous or contains complex wording**.
  - **Less effective:** *Do you like our software?*
  - **More effective:** *What aspects of our software do you find most helpful in your day-to-day work?*
- **Leading language** in a question subtly suggests the “desired” answer or contains assumptions about the respondents. Avoid this
  - **Less effective:** *Wouldn't you agree that our software saves time and effort?*
  - **More effective:** *To what extent do you agree that our software saves you time and effort?*
- **Double-barreled statements** are question about two distinct concepts that make it difficult for respondents and users to provide an accurate, comprehensive answer to both parts.
  - **Less effective:** *Do you agree that the software's user interface is intuitive and the customer support team is helpful?*
  - **More effective:** Two distinct questions asking, “*Do you agree that the software's user interface is intuitive?*” and “*Do you agree that the customer support team is helpful?*” with the ability to respond to each separately.
- **Negatively framed questions** are a leading format that can introduce bias in your results. A positively or negatively worded question assessing the same concept will elicit different responses.

- **Less effective:** *How often do you hate using our software?*
- **More effective:** *How often do you experience challenges using our software?*
- In addition to leading and negatively-framed questions, **emotionally charged or judgmental language** will heavily bias responses to your questions. Use neutral and objective language wherever possible.
  - **Less effective:** *How do you feel about the excessive number of notifications you receive in our software?*
  - **More effective:** *What is your opinion on the number of notifications you receive from our software?*
- **Colloquial speech, specialized language, or technical terminology** may lead to confusion and incorrect responses. Use clear, everyday language that you are confident will be easily understood by your target respondents or users.
  - **Less effective:** *To what degree have our software's API integration capabilities impacted your workflow?*
  - **More effective:** *How does the ability to connect our software with other tools impact your workflow?*
- Very often, **the order of your questions** can influence the quality of your answers. If you expect one question to *prime* your respondents in a specific way, carefully consider the order or randomization you use in your measures.
  - **Less effective:** *Do you believe our software's latest feature has improved its usability? How often do you use the new feature in our software?*
  - **More effective:** *How often do you use the new feature in our software? Do you believe our software's latest feature has improved its usability?*

Appropriate wording of questions can take a significant time investment in your work as an analyst. However, you rarely need to start from scratch when measuring new concepts. The field of *psychometrics* has decades of research applying innovations in measurement theory to develop measures for behavioral and cognitive processes. Many of the concepts we analyze in our work are closely related to human behavior and cognition, which enables us to make use of a wide range of available peer-reviewed resources in our work:

1. Countless **tested and validated questionnaires** are available in peer-reviewed papers on Google Scholar or specialized databases. Each was designed to assess a behavioral or cognitive process and can be strategically modified for specific use cases.
2. Academic resources (peer-reviewed papers, books) that include **research on wording questions in self-report measures** can be leveraged to make decisions about creating measures from scratch or modifying existing ones. If a large portion of your work in analytics involves developing self-report measures, I *strongly* recommend taking the time to familiarize yourself with this research.
3. Existing **research on operationalizing your concept of interest and assessing it as a self-report measure** can help guide you, your team, and your stakeholders in developing these measures. This type of research can help build institutional knowledge of your organization's domain, which we will discuss in chapter 12.

**Figure 6.7 Examples of different question wording. How do you think these formats might impact the responses you receive?**

How would you rate the ease of navigation of our software?

Our software is user friendly and easy to navigate, isn't it?

How easy is it to navigate, find help, and complete workflows in our software?

How challenging would you say it is to navigate our software?

How often would you say it's challenging to navigate our software?

## Using the Right Response Scale

Designing self-report measures includes selecting an appropriate format for your response scale. Most questions only use specific response types (e.g., open-ended questions typically use free-text responses). Still, there's a tremendous amount of flexibility in how you set up those formats.

Unsurprisingly, just like with wording your question, small and subtle changes can lead to shifts in the responses you receive.

Let's first break down common types of response formats in self-report

measures. Each of these is available to you in most survey tools, such as Google Forms:

- **Free-text boxes** allow respondents to answer open-ended questions in their own words. These can yield valuable *qualitative* data but may be more challenging to analyze. The text box size can also impact the length of responses you receive.
- **Single-response** questions allow respondents to select only one answer from a dropdown or scale. These typically use Likert scales (e.g., a 5-point scale from “Strongly Disagree” to “Strongly Agree”).
- **Multiple-response** questions allow respondents to select more than one answer from a list. Limited or unlimited choices may be allowed to the respondent from a set of *categorical* options.
- **Ranking questions** ask respondents to rank items in order of preference. These questions focus on *relative* rather than *absolute* preferences or priorities assessed with single-response Likert scales.
- **Sliding scales** and other interactive response formats can allow for more nuance in responses (e.g., a sliding scale from 0 to 100 instead of a 0 to 10-point scale). It’s essential to use these cautiously and only in cases where you are confident that the additional granularity in response is meaningful to participants.

The NPS question is presented to potential respondents with an 11-point *single-response* Likert scale. It’s often followed up with a free-text box asking respondents to provide additional context on why they responded the way they did to the question. These two pieces of information are commonly used in organizations to create and track the **Net Promoter Score** as a metric over time while also breaking down qualitative responses to infer some of the positive and negative aspects of the business or service assessed.

Most surveys use single-response **Likert scales** to gather self-reported information from respondents. Initially developed in the 1930s, the scale consists of a statement or question followed by an *ordinal range of response options* representing the degree to which you agree or feel strongly about the question. Likert scales often contain an *odd number of response options* with a midpoint allowing for a neutral response. The most common Likert scales use 5, 7, or 9 points and contain response statements like those shown in

figure 6.7.

Likert scales are widely used in surveys due to their flexibility in question and statement types, ease of understanding by respondents, and the ability to compare responses across multiple questions and concepts measured. Given their widespread use, benchmarking against other datasets using the same measures and scales is often straightforward. You can see, for example, whether respondents in different states feel the same way about a candidate running for political office or have differing opinions.

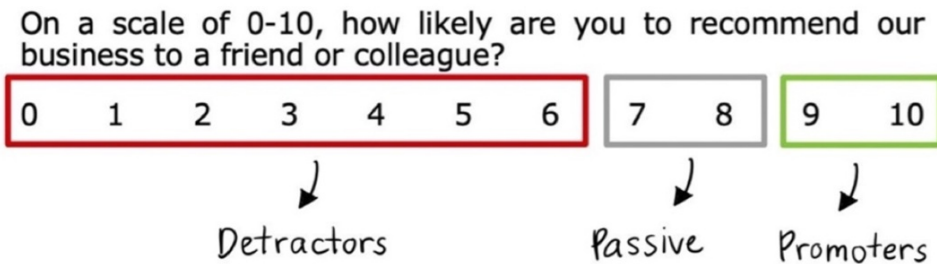
Despite being a universally leveraged response format in surveys, details on how to use them best and why are rarely included in developing these measures (surprise, surprise!). Like with question-wording, extensive research is available on small changes' impact on response formats. Strategically choosing between the available response formats in survey software can help minimize errors in your responses and give you a better estimate of the concepts you're measuring. The following criteria should be applied to *all* Likert scales:

- A **balanced Likert scale** provides an equal number of positive and negative response options. Unbalanced scales that favor positive or negative choices will invariably skew responses in that direction.
- Not everyone has an opinion or feeling on a topic of interest. Likert scales with a **“neutral” midpoint** (e.g., “Neither Agree nor Disagree”) or an **opt-out response option** (e.g., “Unsure/Don’t Know”) are less likely to receive arbitrary responses.
- The **number of response options** impacts the granularity and distribution of responses. Including too few options may limit respondents' abilities to convey their feelings, and too many options can lead to confusion, choice paralysis, and arbitrary selection of values. You can ask most questions with **5 to 7 response items**.

Let's return to the NPS question to see how it performs with these criteria. At face value, the NPS question *does* have a balanced Likert scale. The scale ranges from 0 to 10, with an *implied* range from least to most likely. While the range appears balanced, it's not interpreted in a balanced way for reporting and metrics calculations. Only the top two scale points (9 and 10) are considered *Promoters*, and the next two (7 and 8) are considered *neutral*

points. Most of the scale (0 to 6) are interpreted as negative. The scale is also quite long, with 11 points from least to most likely. Given this granularity, it's challenging to prove whether a person responding with a 9 is more likely to promote your business or organization than someone who responds with an 8 on the scale.

**Figure 6.8 Assessment of the NPS question on the appropriate Likert scale criteria**



- X Interpretation of Likert scale is imbalanced
- X Midpoint available, but evaluated as negative
- X Scale has 11 points, which may create confusion

Once you’ve developed a balanced scale with a strategically chosen number of items and the most suitable option for a midpoint or neutral item, it’s time to fine-tune the scale for your use case. Small changes in formatting and presentation, shown in figure 6.8, can influence how respondents interpret and complete your measures. When developing measures, there are a few formatting decisions to consider:

- **Labeling** your scales involves selecting between the presence of verbal (e.g., “Strongly Agree”) labels and numeric labels (e.g., 1 through 5) for each point on the scale. A scale is typically presented in one of the following ways:
  - **Fully labeled:** a fully labeled scale includes a verbal descriptor that clearly explains the meaning behind each numeric point. This creates clarity on the meaning of each point, which may improve the consistency and accuracy of your responses. They may also enhance the *accessibility* of your survey to respondents. However, a long survey with many fully labeled questions can increase fatigue associated with responding to each question.
  - **Partially labeled:** a partially labeled scale usually includes verbal

descriptors for the end and midpoints. This approach can visually simplify your scale and create less cognitive load than a fully labeled scale. However, participants may interpret the unlabeled scale items differently (e.g., Slightly Agree vs. Agree vs. Somewhat Agree), which may reduce the consistency of your responses.

- **Fully labeled without numerical points:** a fully labeled scale can hide the numerical points associated with each verbal descriptor. This approach allows you to focus on the meaning of each point and remove any preexisting cognitive associations that respondents may have with numerical values.
- **Response order:** the order in which verbal descriptors are displayed (e.g., “Strongly Disagree” to “Strongly Agree” vs. “Strongly Agree” to “Strongly Disagree”) can impact the responses you receive. It’s generally recommended that you keep a *consistent* order of response options for the entirety of a survey. If you see responses that are heavily skewed in one direction, you can consider *counterbalancing* the survey in one of two ways:
  - Present half of the participants with a scale with response options ordered from positive to negative while presenting the other half with a scale from negative to positive.
  - Include multiple questions that assess the same or similar concepts, and strategically vary the wording between positive and negative.

Self-report measurement development is not an *exact* science—there are multiple variations of questions and scales that you can use to collect data about your target population. As an analyst, you can recommend strategically developing questions and scales to increase the quality of your organization’s data over time. Consider selectively testing and iterating on your approaches to asking questions and setting up response scales where possible.

**Let’s return to our case study for the chapter:**

Arthur’s next step is to evaluate the quality of the available self-report measures and measurement scales. He considers the clarity of the questions and the response scales used to determine whether they should be included in his analysis.

He examines each measure in more detail:

1. Satisfaction in the current role: The satisfaction survey uses a 5-point Likert scale with fully-labeled options, ranging from "Very Dissatisfied" to "Very Satisfied." Arthur considers this measure high quality due to its straightforward, easily understandable labels and balanced scale.
2. Perception of workload: The employee satisfaction survey contains questions related to workload perceptions, using a 9-point Likert scale with partially-labeled options from "Light" to "Manageable" to "Overwhelmed". While this scale offers more granularity, the partially-labeled options may lead to some ambiguity in interpretation, potentially reducing the quality of the measure. He also notes that the scale ranges from positive to negative items, which requires him to interpret the numeric scale in the opposite direction of the satisfaction questions.
3. Interest in company events: The company-sponsored event attendance records serve as a behavioral measure. Arthur creates a binary response scale representing attendance (1 = Attended, 0 = Did not attend). This straightforward scale is of high quality and easy to interpret.
4. Satisfaction with salary and benefits: The employee satisfaction survey includes questions about satisfaction with compensation and benefits, using a 4-point Likert scale with numeric-only labels. This scale may be more ambiguous and challenging for some respondents to interpret, potentially reducing the quality of the measure.
5. Productivity in the current role: Arthur notes that the data for this measure is behavioral and skips evaluating any of the available measures as part of this step.
6. Performance in the current role: The company's performance management system uses a rating scale of 1 to 5, with 1 representing poor performance and 5 representing exceptional performance. While this measure is not self-reported, it could be subject to biases similar to self-report measures as the employees' manager reports it. As such, the quality of this measure might be affected by the presence of biases, such as leniency, central tendency, or halo effect.

Arthur decides to include each of the available measures in the analysis, noting which ones are the highest quality as he prepares to compute descriptive statistics for the project. He also notes areas where he believes

each measure can be improved and plans to include these in the limitations and recommendations section of the final report he will create.

### 6.2.3 Interpreting Available Data

As an analyst, you'll often have a combination of self-report data and available data from a source, such as a software application in a data warehouse. Much of this data can be leveraged to track and understand behaviors and combined with self-report measures for a comprehensive picture of your users or participants. While it's readily available for querying and reporting, this type of data can come with unique challenges and limitations that must be addressed to track behavior accurately. We'll discuss some steps to evaluate the quality of available data and its viability for measuring behaviors of interest.

#### Data Collection Practices

The information available in your organization's data warehouse is likely synthesized from various sources. If you work at a software company, you probably have vast amounts of data about the users who access your software. You may also have data from third-party applications used by different teams (e.g., HR software, support ticket software) that gets combined into one place to query, join and manipulate in your analyses and reporting.

When examining the available data in a warehouse, learning how data is collected and surfaced for analytics will help determine if it can be leveraged to measure a specific behavior. Ask yourself and your colleagues the following questions to determine the viability of a data source for your needs:

- **What is the source of the data?** Is the data collected from the primary source of information about your organization (e.g., sales transactions on the company website, records from a software application), a third-party application used by your organization, or both? Often, the data you have querying access to is a highly curated version of the underlying data, so it's valuable to understand what information comes from where.
- **How often is data refreshed?** Is it updated once per hour, per day, or

less frequently? A daily or hourly refresh rate will usually suffice to create self-service analytics tools and insights for your stakeholders. Set expectations with your stakeholders on how often new information is available.

- **Does the data capture objects or actions?** A record in a database can mean many things. Does it correspond to an object (e.g., a show available on an online streaming service) or an action (e.g., a user “clicked” the description of the show)? Objects can sometimes be used as proxies for behavioral data but may have limited insight into the nuanced actions that led to the creation of that record.
- **Is there a way to track changes?** Many times, records in a database are edited in place without the ability to determine what changes were made and when (e.g., a sales transaction record deletes items returned from the purchase history). This can create challenges when tracking information over time.
- **What do the timestamps mean?** Tables in a database will almost always have a timestamp associated with a record. The timestamp can be when an action occurred, an object was created, or the record was uploaded into the data warehouse. Since time is a crucial dimension for most measurements, make sure it means what you think it does.

**Figure 6.9** Each row in this dataset represents a record of a visit to a doctor’s office.

	visit_id	patient_id	doctor_id	date	time	visit_reason
0	1038	2654	5	2023-05-01	10:00	Annual check-up
1	1039	2114	8	2023-05-03	14:30	Sore throat
2	1040	2025	4	2023-05-05	11:45	Allergy consultation
3	1041	2759	6	2023-05-06	09:15	Flu symptoms
4	1042	2281	7	2023-05-08	16:20	Routine vaccination

## Granularity and Aggregation

You can measure behavioral data at different levels of granularity (e.g., individual users, group of users, or an organization). If you work at an organization that serves both business/organizational customers and their

users, this can introduce a level of hierarchy that impacts anything you measure. The individual *and* their organization predict the behaviors you capture. This requires additional strategic decisions about how to create an effective measurement. Overall, asking yourself the following questions will help you determine the level of aggregation (or *grain*) of the data that is appropriate for measurement and metrics (more on that in chapter 7):

- **What grain does this behavior need to be measured at?** The unit of measurement is generally derived from an understanding of what *matters* and what you're trying to change. Is it a web page? A series of webpages creating a flow to checkout your cart and make a purchase? Is it the user's continued visits to your site? How many users at a company tend to visit your app? The grain you choose determines the level of aggregation and any hierarchical relationships you need to consider in your analyses or models (more on that in chapter 9).
- **Is the data pre-aggregated?** You may encounter data in your career that has already been pre-aggregated (e.g., a data warehouse view that's grouped by month, a third-party API that gives you a count of records per hour). You may need to determine what information is lost at the level of aggregation, and whether it's necessary to disaggregate the data to derive value.

**Figure 6.10** Each row in this dataset contains information about visitors and the pages they viewed. What level of aggregation exists in this dataset?

	hour	event_id	user_id	browser	page	total_views
0	2023-04-25 06:00:00	9666349	254311	Chrome	Clothing	2
1	2023-05-01 14:00:00	6701590	710429	Chrome	Checkout Cart	3
2	2023-05-01 23:00:00	5451450	650559	Firefox	Blog	2
3	2023-05-03 00:00:00	20957045	75027	Firefox	Blog	1
4	2023-05-08 00:00:00	20792402	52518	Chrome	Accessories	2

## Data Quality

Organizations invest varying amounts of time and resources to ensure the quality and accuracy of data in a warehouse. The state and structure of your

data warehouse will differ by the industry you are in, the size of your workplace, and how important accurate data is to the performance of the business or organization. Software companies tend to invest heavily in their teams' data stack since there's high potential value associated with data-informed and machine learning features. A small non-profit may simply not have the capacity to develop the infrastructure necessary for a functioning data warehouse, regardless of the value potential of their data.

The questions you should ask about data quality will differ heavily based on the type of business organization you work with. You may not need to ask them for every dataset and project, but you will want to answer these questions **at least once** early in your role and before starting a task. You will avoid many inaccurate findings and misinterpretations if you do!

- **What steps are taken, if any, to handle duplicate, missing, or otherwise inaccurate data?** The ingestion of data into a warehouse is rarely error-proof. You may encounter duplicate records or missing or bad data in your reports if your organization doesn't regularly perform data quality checks. If you don't have visibility into this work, check with the team that maintains the warehouse (e.g., data engineering) to learn more.
- **If you are capturing data using third-party tracking, is there anything that can limit the accuracy and completeness of the dataset?** Many organizations use third-party software to track behaviors online on their website or application (e.g., page views, clicks). Some options for third-party tracking do not work under certain conditions, such as when a user has an ad blocker enabled. This can create issues with missing data, *not at random*, and limit your ability to accurately report information at the grain of the user.
- **What kind of tables are you accessing?** Are you directly querying data ingested into the warehouse, or are there highly curated analytics tables that combine data sources for ease of use? If your organization has an analytics engineering function that develops and maintains analytics models, you will often have automated checks to resolve many typical data quality issues.

**Figure 6.11** A few simple queries can help determine if there are issues with the quality of your data.

Check for Duplicate Records:		Check for Bad Dates	
<pre>SELECT username, COUNT(user_id) AS n_records FROM users GROUP BY username HAVING COUNT(user_id) &gt; 1;</pre>		<pre>SELECT username, signup_date FROM users WHERE     signup_date &gt; CURRENT_DATE;</pre>	
username	n_records	username	signup_date
jd_9705	2	python_user8	2025-10-01
data_analyst01	3	coffee_and_data	2024-12-31

Let's return to our case study for the chapter:

Arthur's next step is to evaluate the quality, granularity, and data collection process of the available behavioral measure: *productivity in the current role*.

Arthur examines the data available in the company's data warehouse to assess productivity. He identifies several potential indicators of productivity:

1. Number of tasks completed
2. Project completion times
3. Adherence to deadlines

Arthur recognizes that these indicators may not fully capture the complexity of productivity across different roles and departments within the company. Some tasks may require more time and effort, making it difficult to directly compare productivity levels based on the number of completed tasks.

Arthur speaks with the IT department and managers responsible for tracking this data to understand the data collection process better. He learns that while most departments have well-established monitoring and reporting productivity processes, some use inconsistent and subjective criteria. He proceeds cautiously using these measures and prepares to exclude them entirely from the analysis if appropriate.

## 6.2.4 Activity

You are a Human Resources Analyst at a company developing a workplace

wellness program to reduce burnout and improve employee retention and productivity. You have been assigned a project to identify or develop self-report and behavioral measures to evaluate the program's impact.

1. Go to [scholar.google.com](https://scholar.google.com) and search for *employee wellbeing questionnaire*. Filter your results to papers published since 2019. Are any papers available about developing or validating a scale on employee wellbeing? If you access the full text of the paper, you will usually find the newly developed questionnaire in the Appendix.
2. Develop or select a question from one of the available surveys you discovered on the following topics. What type of response scale would you use and why? What potential biases or limitations may be associated with different response formats? Suggest ways to minimize their impact on the quality of your results.
  - a. Stress levels related to your work environment
  - b. Perceptions of support they receive from supervisors for their wellbeing
  - c. Overall satisfaction with the workplace
3. Identify potential behavioral measures to track employee engagement and workplace stress. You can assume that multiple data sources are available in a data warehouse on employee hours, project completion times, meeting deadlines, and more. What characteristics would the data need to have for you to accurately use it for measuring engagement?
4. Propose a hypothesis about which behavioral measure(s) might be correlated with the impact of the workplace wellness program.
5. How will you integrate the self-report and behavioral measures? What information would you need to connect these data sources?

## 6.3 Reliability and Validity

Regarding measurement, *reliability* and *validity* are the essential foundation of all conclusions you draw from your data. These two concepts ensure the accuracy and integrity of any analysis you perform. Most strategies for developing reliable and valid measures were designed with self-report measurements in mind, but many of the same steps can be applied when working with behavioral measures. We will cover approaches for working with both types of data.

## 6.3.1 Reliable Measures

*Reliability* refers to a measurement's ability to assess a concept, behavior, or process consistently and repeatably over time and across different respondents or users. A measure must be reliable for stakeholders to trust the numbers they are seeing and the conclusions drawn from its data.

We will discuss several types of reliability, including test-retest reliability, inter-rater reliability, and internal consistency. Each type should be evaluated as appropriate for a measure to be deemed reliable.

### Test-retest Reliability

*Test-retest reliability* refers to how *consistent* a measure is over time. A measure is consistent if the same individuals produce similar results on the measurement when tested and retested over time, assuming your concept of interest has not changed. Some examples of test-retest reliability include:

- A participant in a study with clinical depression will respond with similar answers to a measurement scale assessing depression over time periods, assuming they are still experiencing an episode of depression and are not receiving treatment.
- A software user responds to the NPS question with similar scores over time, assuming that the business or service offered has not substantially changed.

Test-retest reliability is vital for data intended to be measured over time, such as for a business metric or forecasting model. If you're interested in how something has changed and will continue to change, you *must* be confident in the quality of the underlying numbers.

Test-retest reliability can be performed using the following steps:

1. Collect or select the initial dataset on your measure being tested.
2. Choose an appropriate time interval to collect or select your subsequent datasets. This will vary based on what you are measuring and the resources you have available.

3. Collect data for one or more intervals after the initial measurement. To compare measures effectively in most statistical software, you may want to structure your data in a *wide* format, with one row per user or participant and one column per time interval measurement.
4. Compare correlations between scores for each time interval. This is typically done using a Spearman or Pearson's correlation value. The more scores correlate, the higher your test-retest reliability.

Test-retest reliability applies to self-report and behavioral measures about a construct you expect to be stable over time. If the data *consistently* measures a *stable* process about a user over time, these steps are valuable to use before reporting on a measure for metrics or experimentation.

For example, a product analytics team is interested in developing a measure of user engagement on an app. The team starts by measuring the average time spent on the app per user session. To test this measure, the team calculates the weekly average time spent on the app per user. They then select three *non-consecutive weeks* to compare the initial week's data and run Pearson's correlations on each user's time spent on the app. Pearson's correlation values should be pretty high if this is a *stable measure* (e.g., users are consistent in the amount of time they spend on the app). If they weren't as high as expected, the team could explore further potential sources of error, such as changes in user behavior, app functionality, or data aggregation methods. This type of investigation will build *a lot* of internal knowledge of the user base.

**Figure 6.12 Test-retest reliability correlations are expected to be quite high across time periods (e.g., T1, T2, T3), often with an r value above 0.7.**

	Happiness at T1	Stress at T1	Happiness at T2	Stress at T2	Happiness at T3	Stress at T3
Happiness at T1	1.00	-0.60	0.72	-0.65	0.58	-0.52
Stress at T1	-0.60	1.00	-0.45	0.68	-0.50	0.72
Happiness at T2	0.72	-0.45	1.00	-0.72	0.68	-0.63
Stress at T2	-0.65	0.68	-0.72	1.00	-0.75	0.80
Happiness at T3	0.58	-0.50	0.68	-0.75	1.00	-0.72
Stress at T3	-0.52	0.72	-0.63	0.80	-0.72	1.00

## Inter-rater Reliability

*Inter-rater reliability* measures the agreement between two or more people evaluating the same data set. It's typically used when performing analysis tasks that require subjective judgments, ratings, or classifications (e.g., assessing qualitative data or making recommendations based on a large project). High inter-rater reliability means that your raters made *consistent* judgments, indicating the reliability of the criteria used to review the data.

Inter-rater reliability is typically conducted using the following steps:

1. Develop an objective rubric or criteria for raters to follow when evaluating data. This may include categories, a rating scale, or other measures.
2. Have the raters conduct independent evaluations of the data without discussing or sharing input. This will allow them to focus on evaluation without biasing each other's responses.
3. Calculate inter-rater reliability using one or more similarity indicators (e.g., the percentage of cases with the same rating, a Pearson's correlation, or a *Cohen's kappa*).
4. Iterate on the evaluation criteria and processes to improve alignment between raters and the evaluation quality over time.

When done at scale and over long periods of time, measures of inter-rater reliability demonstrate a strong and consistent understanding of the qualitative data you are capturing. If a complex evaluation is part of your role or your organization's success, a system of inter-rater reliability evaluation can be set up to continually improve these processes.

**Figure 6.13 Inter-rater reliability involves systematically evaluating the agreement between individuals where your measurement includes a subjective judgment.**

NPS Comment:

"I have had issues with setting up my account and doing basic tasks. The support team helps when I have a question but it doesn't solve the main problem"

<u>Rater # 1</u>		<u>Rater # 2</u>		<u>Rater # 3</u>	
Product	(negative)	Initial setup	(negative)	Product	(neutral)
Service	(positive)	Service	(neutral)	Initial setup	(negative)
				Service	(negative)

## Internal Consistency

*Internal consistency* evaluates the extent to which a collection of measurements effectively captures data about the same core concept or process. In most analytics projects, you'll likely need to work with multiple measures or develop composite measures to assess complex processes related to your users or participants. For instance, a depression questionnaire typically comprises various questions addressing different symptoms. The participant's responses are then combined to generate a score that reflects the presence of depression. In such cases, it's crucial to take a moment to ensure that the multiple measures you're using are *reliably* measuring the same process before combining them into a single, cohesive measure.

Internal consistency can be evaluated using the following steps:

1. Identify all variables and indicators you hypothesize measure the same underlying concept or process.
2. Calculate correlations between each indicator for each user or participant. You can use any appropriate correlation coefficient for your data (e.g., Pearson's correlation for continuous data, Spearman's rank correlation for ordinal data). Consistent variables should have high correlation coefficients. Variables *unrelated* to the rest should likely be considered a separate process and excluded from any composite measures.
3. Use a statistical method for assessing internal consistency. One of the most common methods is *Cronbach's alpha*, which calculates the average inter-item covariance (unstandardized correlation) between all variables and returns a coefficient from 0 to 1. Cronbach's alpha is not readily available in the Python packages we have covered thus far (e.g., statsmodels, scipy). Instead, it can be imported using the penguin library, which has a range of statistical functions that return comprehensive information for common tests and has functions not available elsewhere.

**Figure 6.14 Steps to calculating Cronbach's alpha**

$$d = \frac{k \cdot \bar{c}}{\bar{v} + (k-1) \cdot \bar{c}}$$

$k$  = number of items

$\bar{c}$  = average covariances between items

$\bar{v}$  = average variance of each item

4. Refine the variables that you include in your composite measure. You may need to remove some or consider adding others related to your concept or process.

To calculate Cronbach's alpha in Python, you can use the pingouin library's `cronbach_alpha` function. We'll use the `weather.csv` data from previous chapters to demonstrate how this calculation is performed. First, we'll create a correlation matrix to view the standardized relationships between each variable.

```
import pandas as pd      #A
import numpy as np

weather = pd.read_csv("weather.csv")    #B
weather_corr = weather.corr()          #C
print(weather_corr)
```

	high_temp	low_temp	humidity	wind_speed	precip
high_temp	1.00	0.96	0.15	-0.23	-0.04
low_temp	0.96	1.00	0.18	-0.26	-0.03
humidity	0.15	0.18	1.00	0.03	0.23
wind_speed	-0.23	-0.26	0.03	1.00	0.21
precip	-0.04	-0.03	0.23	0.21	1.00

We can see that the the temperature variables are strongly correlated with each other and moderately negatively correlated with the wind speed. Next, we will use the pingouin library to calculate the Cronbach's alpha associated with the internal consistency of all items in the dataset.

```
import pingouin as pg    #A
```

```
cr_alpha = pg.cronbach_alpha(data=weather.drop('day', axis=1))
print(cr_alpha)      #C
(0.5446867046720659, array([0.501, 0.586]))
```

The function returns an alpha value of 0.545, which suggests a relatively low internal consistency. In general, peer-reviewed research suggests the following approximate guidelines for how to interpret the strength of an alpha value:

- **0 to 0.69** for low internal consistency.
- **0.7 to 0.79** for moderate internal consistency
- **0.8 to 0.89** for good internal consistency
- **0.9 to 0.95** for excellent internal consistency
- **0.95** or above suggests there may be redundancy among your variables, and you can potentially remove some without reducing the quality of your measure

The guidelines for interpreting the strength of internal consistency are not exact. They were set as approximate guidelines primarily on the strengths of self-report measures collected as quantitative surveys. While these recommended thresholds haven't been directly tested with behavioral measures, they *are* based on correlation coefficients, frequently used with both data types.

## Recap

Reliable measures are central to ensuring that your deliverables as an analyst continue to bring value to your stakeholders long after you complete them. If you can confidently say that your measures are consistent, stable, and have a meaning agreed upon between professionals, you're well equipped to create self-service and reproducible results. This is key to developing effective metrics, which we will discuss in the next chapter.

### 6.3.2 Validity

The standardized and tested measures you are working with enable you to maximize the quality of your data and results. Once these steps are complete,

it's essential to step back and ask yourself the question that this chapter is named after – **are you measuring what you think you're measuring?**

*Validity* is the extent to which a measurement accurately captures the intended concept or process. A valid measure should be a true, comprehensive representation of what you are analyzing. Establishing the validity of your measurement is crucial to deriving meaningful, trustworthy results for your stakeholders to use in their decisions.

Researchers and analysts use several common types of validity as criteria for the quality of their measurements, including face, construct, and criterion. In most cases, assessing measures for these types of validity are straightforward and can be incorporated into your background research and preparation for your analysis.

### **Content / Face Validity**

*Content* or *face validity* refers to the degree to which your measures sufficiently capture all facets of the concept or process you are measuring. It's primarily assessed through a qualitative evaluation of the underlying theoretical concepts, expert judgments, and existing evidence supporting the definition of the concept. When assessing face validity, reviewing **your definitions with subject matter experts is essential**. These are most often your stakeholders or colleagues in your field that work directly with the process you are measuring. Assessing face validity is an *ongoing process* that requires you to keep up to date on new theories and evidence in the domain of study or practice you are working in.

For example, if you want to assess employee job satisfaction, you will first need to *conceptualize* job satisfaction and identify all relevant facets of this concept. You may review peer-reviewed literature in organizational psychology and identify several factors typically included in questionnaires (e.g., satisfaction with work tasks, work environment, manager, colleagues, and compensation for the role). From there, you can use existing peer-reviewed measures or develop your own to assess each of these factors.

### **Construct Validity**

*Construct validity* refers to the extent to which your measurement accurately assesses the theoretical construct it's intended to. It reflects the degree to which the observed patterns in your data align with the relationships you expect based on the underlying theory. Assessing construct validity involves examining two subtypes of this criteria:

- **Convergent validity** refers to the degree to which a measure is related to *other* measures that are theoretically associated with the same construct. This is typically assessed by evaluating the strength of correlations between concepts or behaviors that are theoretically related.
- **Discriminant validity** refers to the degree to which a measure is *not* related to other measures that are theoretically unassociated with the construct. You can also assess this type of validity using correlations; you will generally expect that the strength of the relationship should be *less* than convergent measures.

For example, if you are assessing the construct validity of the NPS question, you will likely want to determine if responses to the question correlate with referral behaviors and if you can track them at your organization. If scores on the question *positively* correlate with referral rates, you can leverage it as a *valid* measure of the likelihood of referring others to your business or organization.

## Criterion Validity

*Criterion validity* refers to the degree to which a measure is associated with an external criterion, such as a well-established measure of the existing concept or an outcome of interest. This form of validity is also typically assessed by looking at correlations between the measure of interest and the *criterion* used to assess it.

Criterion validity can be assessed in two ways, depending on your analytics project:

- **Concurrent validity** assesses the relationship between a measure and its criterion simultaneously. Both measures can be administered at the same time (self-report measures) or evaluated at the same/similar time periods

(behavioral measures).

- **Predictive validity** assesses the relationship between a measure at a time period *before assessing* its criterion. Analysts may collect data on the criterion measure at a subsequent point or select an appropriate time window, after which they expect a predictive relationship to be present in the data.

For example, suppose you want to assess the relationship between a new financial metric and the company's financial performance. In that case, you will likely want to see if it's an appropriate *predictor* of existing financial metrics. You can compare the data on the new financial metric in the current fiscal quarter to a performance metric such as net income or profit in the *following* fiscal quarter. A valid financial metric should correlate strongly with the company's financial performance.

**Let's wrap up our case study for the chapter:**

Arthur's final step before conducting his analysis is to assess the reliability and validity of each measure to ensure the insights on the workplace wellness program are accurate and meaningful. He uses the following steps to evaluate each measure:

1. **Satisfaction in the current role:** Arthur examines the internal consistency of the survey using Cronbach's alpha, yielding a value of 0.72. He also assesses the content validity of the survey by researching existing definitions and measures and determines that the survey used is comprehensive.
2. **Perception of workload:** Arthur yields a Cronbach's alpha of 0.68 for the questions assessing this concept. He assesses content validity by researching existing measures and definitions of perceptions of one's workload.
3. **Interest in company events:** Arthur does not assess the internal consistency of this single-item measure. He primarily focuses on the question's wording to ensure it's a clear and objective measure of employee interest in events.
4. **Satisfaction with salary and benefits:** Arthur yields a Cronbach's alpha of 0.34 with all of the variables used. He reviews existing research

and measures of the concept, concluding that the questions measuring this concept do not all capture the same process.

5. **Productivity in current role:** Since this is a behavioral measure with concerns about reliability, Arthur takes the time to assess the test-retest reliability of the measure across multiple time periods. He discovers that the correlations between time periods are quite high and relatively stable.
6. **Performance in current role:** Arthur assesses inter-rater reliability by comparing performance ratings between managers, supervisors, and directors where available.

By systematically assessing the reliability and validity of each measure, Arthur has ensured that the data collected is of high quality. He will establish confidence in each measure and enable the organization to leverage the findings and each measure as metrics over time.

### 6.3.3 Activity

You have shared the proposed self-report questionnaires and behavioral measures to assess employee wellbeing with your team lead. The surveys have been administered to employees at the company, and you have approximately 600 responses for analysis. You are asked to validate that the measures used were appropriate for measuring each concept of interest in this project.

1. What steps will you take to assess the reliability of each set of measures? As a refresher, we have measures on the following topics:
  - a. Stress levels related to your work environment
  - b. Perceptions of support they receive from supervisors for their wellbeing
  - c. Overall satisfaction with the workplace
2. What forms of validity are most appropriate to assess the above measures?
3. How will you assess the validity of the behavioral measures that you selected? How do those steps differ from the self-report measures?

## 6.4 Summary

- **Self-report measures** are collected directly from individuals, usually through questionnaires or surveys. These measures capture information about perceptions, attitudes, beliefs, or experiences. In analytics, self-report measures provide valuable insights into subjective experiences and allow for assessing factors that may not otherwise be observable.
- Designing accurate self-report measures (e.g., **questionnaires**) requires **careful wording of questions** and assessment items. Leading language, double-barreled statements, and negative wording can all reduce the quality of responses you receive.
- There are many ways to structure the **response format** to questions (e.g. **Likert scales**) that can impact your results. A scale should be **balanced**, and have a moderate number of response items (usually 5 to 7) to maximize the quality of responses.
- **Behavioral measures** are records of behaviors or actions captured by direct observation or recorded in a data warehouse as an interaction within an application or tool. These measures can offer direct insights into user and participant experiences without the bias associated with self-report measures.
- **Reliability** refers to the consistency, stability, and repeatability of a measurement. A reliable measure is expected to produce similar results under consistent conditions. Ensuring reliability in analytics is essential to produce accurate and trustworthy results that can inform decision-making and support evidence-based practices.
- Analysts typically assess three common forms of reliability: (1) **test-retest reliability**, which is an assessment of consistency in measurement responses over time, (2) **inter-rater reliability**, which is a measure of consistency between raters for qualitative evaluations, and (3) **internal consistency**, which is an assessment of the degree to which different variables measure the same underlying construct.
- **Validity** refers to the degree to which a measure accurately assesses the concept or process it's designed to measure. Ensuring that your measures are valid is critical to building confidence in your stakeholders that your results are meaningful, accurate, and reflective of the true phenomenon you are investigating.
- Three forms of validity are essential to assess as part of your measurement: (1) **content/face validity**, which is an assessment of how comprehensively your measure captures the concept or phenomenon of

interest, (2) **construct validity**, which is an assessment of whether you are measuring your intended theoretical construct, and (3) **criterion validity**, which assesses whether your measure is associated with other measures of similar constructs.

# 7 The Art of Metrics: Tracking Performance for Organizational Success

## This chapter covers

- Understanding the value of metrics for the success of an organization
- Identifying measures of success to leverage as metrics
- Designing SMART metrics for effective tracking and decision-making
- Identifying and mitigating common pitfalls and errors when creating metrics
- Powerfully communicating progress and insights to stakeholders

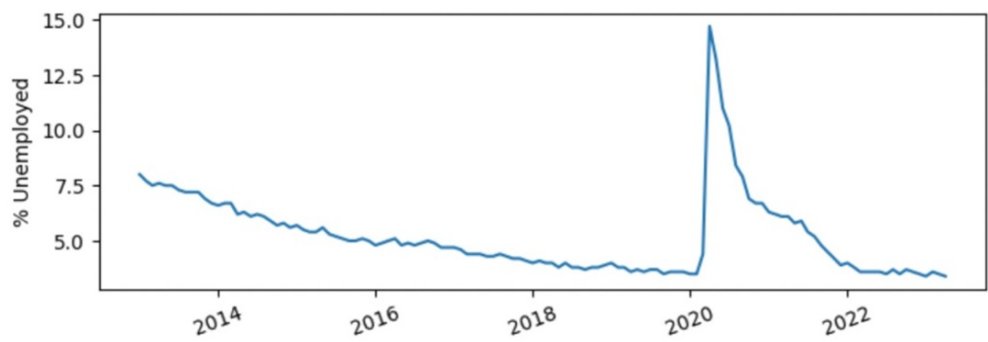
When watching or reading the news, you'll regularly see charts tracking information over time. You're likely familiar with many of them—COVID-19 prevalence, the Air Quality Index (AQI), unemployment rate—and you likely understand what it means when the numbers increase or decrease. These metrics enable you to understand the world and how it changes.

*Metrics* are **standardized quantitative measures** that track processes, outcomes, or activities over time. These invaluable tools allow us to understand immediate and long-term changes that influence nearly every facet of our lives.

Metrics are shown to nearly *everyone*, regardless of their experience with data. Take the following example: the **unemployment rate** is a widely recognized economic indicator across many countries, defined as the *percentage of the labor force not currently employed and actively looking for a job*. The Bureau of Labor Statistics calculates the unemployment rate every month, reporting it to the public as an indicator of the job market's health. People use this metric to make critical life decisions, such as seeking new employment or buying a house. The full concept of *unemployment*, however, is captured as one of multiple indicators providing useful subsets of

information on the state of the job market at any given time.

Figure 7.1 The official unemployment rate reported monthly from 2013 to 2023



To provide a comprehensive picture of unemployment in the U.S., the Bureau of Labor Statistics (BLS) publishes **six metrics** each month. The official unemployment rate is known as U-3, which includes *all* people in the workforce who are unemployed, looking for work, and available for work. The other five measures include subsets of the population not counted in U-3 (e.g., U-4 includes *discouraged* workers not actively looking for work because they believe no jobs are available). Each provides a specific insight, and all six are necessary to understand workforce participation.

Figure 7.2 Each unemployment metric includes different types of unemployed or underemployed workers

	U - 1	U - 2	U - 3	U - 4	U - 5	U - 6
Long-term unemployed	✓		✓	✓	✓	✓
Recently unemployed		✓	✓	✓	✓	✓
Finished temporary employment		✓	✓	✓	✓	✓
Discouraged/no longer searching				✓	✓	✓
Marginally attached/sometimes searching					✓	✓
Underemployed						✓

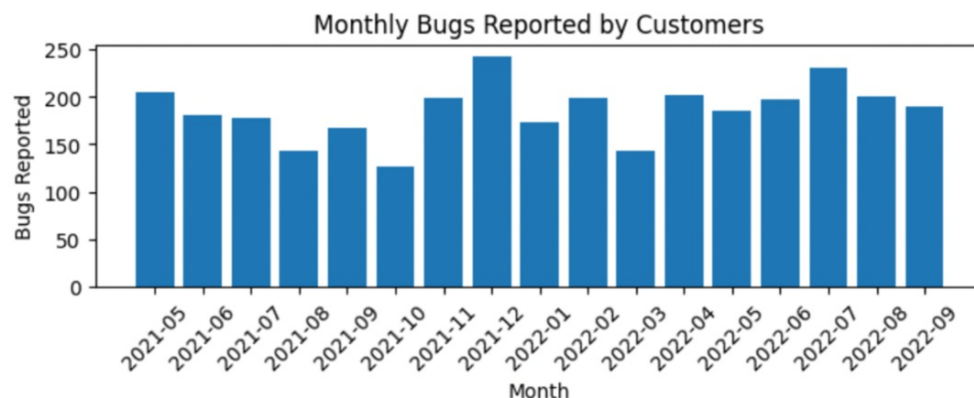
The BLS unemployment metrics have existed in some form for nearly a

century. Decades of research allowed for refining the calculation and data collection into the six indicators shown in figure 7.2. This chapter will cover the skills you need to create similarly clear, actionable, and impactful metrics (without a century of work!). You will learn how to identify valuable processes to monitor, a framework for defining the metric, how to avoid common pitfalls in metric calculations, and how to share them effectively with stakeholders. The metrics you design will enable your organization to understand the impact of their actions, make strategic decisions, and understand the complex processes that influence the organization.

## 7.1 The Role of Metrics in Decision-Making

Metrics provide a concrete way of evaluating performance and trends over time and are essential tools in decision-making, strategic planning, and goal-setting. They're often visualized as a line or bar graph, where the y-axis captures an aggregated measure and the x-axis captures the measure over time.

**Figure 7.3** Metrics are usually depicted as an aggregate measure over time.



Every organization can benefit from metrics to understand the impact of its actions and decisions on performance. When an organization invests time in developing, tracking, and surfacing metrics to stakeholders, they have a competitive advantage in their strategic decision-making. In this section, we will cover three tiers of metrics for organizations to track:

- **Performance metrics** provide critical feedback on the state of the

organization. These are core metrics tracked across many types of organizations and ultimately tell you whether or not the organization is succeeding.

- **Organizational strategy metrics** provide information on the impact of specific actions taken within the organization. These allow teams to dive one level deeper than performance metrics and understand actions that correlate with or predict performance.
- **Accountability metrics** hold individuals and teams responsible for producing results. These often overlap with performance metrics but are designed to make the daily work, processes, and successes of individual teams clear and transparent to the organization.

An organization that understands its performance with metrics is generally positioned for better long-term outcomes. When you understand your users and the quantitative indicators of their needs and experiences, you will spend far less time figuring out what actions to take to meet those needs. An organization that *doesn't* invest in measurement and metric development can easily hit a wall and struggle to achieve its long-term goals.

With the proper steps, organizations at any size or stage of growth can learn to guide their strategy with appropriate metrics. The key is selecting measures that align with their goals and offer actionable insights. These metrics enable the organization to confidently navigate the complex decision-making landscape when well implemented.

### 7.1.1 Tracking Performance

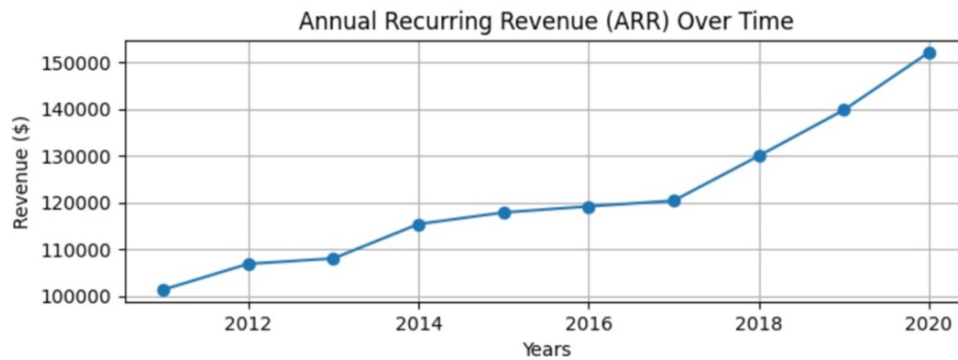
An organization should typically begin using metrics to track measures that clearly indicate its performance. These measures most often correspond directly to the organization's key goals, objectives, and ability to continue succeeding (e.g., revenue, profit). Once these foundational metrics are established, you can delve deeper into secondary indicators of operational success.

At smaller or newer organizations, your role as an analyst may entail developing organizational metrics from the ground up. The majority of organizations will benefit from establishing performance metrics in the

following order:

- **Financial metrics** are critical to understanding the financial health of an organization. Most businesses will track and monitor revenue, profit margins, and customer acquisition costs (CAC). For-profit companies have well-established sets of financial metrics based on the type of business (e.g., business-to-consumer/B2C) and method of collecting revenue (e.g., monthly subscriptions) [1]. Non-profit and government organizations will also track inbound funding and revenue as appropriate. However, they often need to customize these calculations to represent how they receive funds. These metrics are *vital* to investors, board members, and shareholders with a vested interest in the organization's financial performance.
- **Operational metrics** gauge the efficiency of an organization's operational processes. These can include measures such as production volume, downtime on an app, or the acceptance rates of offers extended for employment. Operational metrics are more granular than many financial metrics, typically calculated as *ratios of products/services per employee* or *time to complete an activity*. These metrics are necessary for teams to identify areas of inefficiency and bottleneck, allowing them to improve productivity and reduce costs associated with day-to-day tasks.
- **Customer metrics** focus on customers' experiences and interactions with the organizations. Examples include customer satisfaction scores (CSAT), churn rates, and lifetime value (LTV). These metrics provide insight into how well an organization meets the needs and expectations of its customers. They often guide organizational strategy, which will be discussed in the next section.

**Figure 7.4 Annual recurring revenue (ARR) or monthly recurring revenue (MRR) is a standard financial metric among businesses with subscription-based models.**



## 7.1.2 Informing Organizational Strategy

Once performance metrics are established, organizations can delve deeper into specific areas of operations and factors that correlate with the success of each performance metric. These might include employee productivity, process efficiency, and customer behavior measures. Understanding this next level of metrics and their relationships to organizational performance enables you to create a data-informed organizational strategy.

Performance metrics don't usually tell you the reason for the value you are seeing. If your organization's revenue decreases in a month, you and your stakeholders can hypothesize any potential causes for the change.

**Organizational strategy** metrics enable you to tease apart the factors contributing to performance and take actions to benefit the organization.

Several types of metrics can inform organizational strategy and potentially correlate with performance metrics. In most cases, they can be broken into the following categories:

- **Customer metrics** track the behaviors, experiences, and sentiments of the individuals, businesses, and organizations that leverage your goods and services. Understanding their needs and experiences with your organization provides you with the information you need to retain and grow your customer base. Some examples of customer metrics include engagement/activity rates, referral rates, and customer effort. Each of these metrics can inform how your organization engages with these customers.
- **Product metrics** are used in organizations that offer a product or

service. These measures allow you to understand the behaviors and usage of your product or service and inform your product development strategy. Examples of product metrics include usage rates, feature adoption rates, and time to complete workflows.

- **Market metrics** enable you to understand the market in which your organization operates. These metrics usually comprise data gathered outside your organization, such as the many supply, demand, and price metrics provided by the Bureau of Labor Statistics. Having market data can inform decisions about entering new markets, adjusting pricing strategies, or investing in marketing and advertising.

Identifying appropriate strategy metrics can require a significant time investment to drive success. As an analyst, you must form hypotheses, explore relationships with performance metrics, and recommend which ones to track. It's also important to regularly evaluate and refine organizational metrics, whose relationships with performance metrics may change with its performance, focus, and market conditions. You can expect to iterate and revise strategy metrics more often than performance metrics.

Where possible, establishing a causal link between your strategy and performance metrics can significantly enhance decision-making at your organization. Identifying *causality* means identifying whether changes in strategic initiatives directly lead to changes in performance. Demonstrating causality can be complex—there are entire books on methods of causal inference that you can reference if you are considering building expertise in this topic:

- **Causal Inference: The Mixtape** by Scott Cunningham [2] covers numerous sophisticated approaches to establishing causality when working with complex datasets. The book includes examples in R and leverages economic data in many of its examples.
- **Causality** by Judea Pearl [3] covers a breadth of approaches to causal inference leveraged in academic and non-academic fields. T

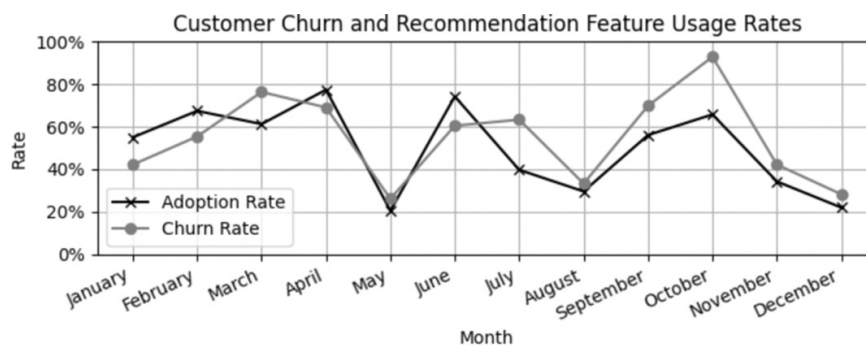
If you are constrained in your ability to leverage sophisticated statistical methods, even simple approaches can provide valuable insights:

- Comparing performance metrics **before and after** an action or change

captured in your strategy metrics (e.g., performing a regression analysis using your strategy metrics as predictors and performance metrics as your outcome).

- Comparing performance metrics **between groups** that were and were not affected by the action or change (e.g., adding group indicators as predictors in your regression model and performing a t-test or ANOVA).

**Figure 7.5 Visually representing strong relationships between strategy and performance metrics can help your stakeholders understand long-term trends and if the relationship strengthens or weakens over time.**



Regression models provide additional rigor to your analysis, but remember that correlation does *not* imply causation. A breadth of factors influences organizational performance; as an analyst, you will continually try to understand the relationships you *haven't* yet measured. Over time, you will build a comprehensive picture of the processes and measures that make up the landscape of your organization.

### 7.1.3 Promoting Accountability

Metrics are often used to measure the performance of individuals and teams in their roles to enable the organization's success. These **accountability metrics**, often called Key Performance Indicators (KPIs), provide a tangible way to set expectations at multiple levels and measure progress toward goals and responsibilities. Organizations often use these metrics to guide performance reviews, promotion decisions, bonus pay, and team development strategies.

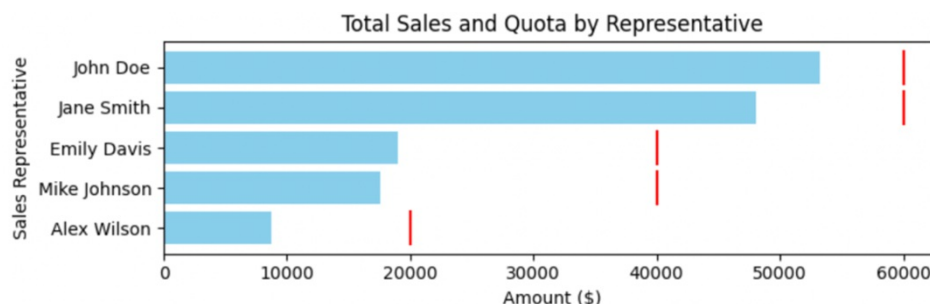
For individuals and teams, these metrics are typically tied to the timely and

accurate completion of tasks and projects. For example, a customer support team may track the *percentage of customer ticket inquiries resolved in under 24 hours*. The team's manager likely looks at the overall metric value for the team, followed by a breakdown by sales representative to understand each person's performance. The information shows which team members meet targets, and which might need additional support, coaching, or reassignment of work.

Accountability metrics at an organization can look like the following, broken down by team:

- **Percentage of sales quota attained** per sales development representative on a sales team, measured quarterly.
- **Average time to fill open roles** per recruiter on a talent acquisition team, measured month over month.
- **Time to resolve bugs** (work representing errors or issues on a site or app) on a software engineering team, measured each quarter.
- **Number of qualified leads** (inbound customers likely to purchase your product or services) generated through marketing campaigns, measured every six months.

**Figure 7.6** A graph showing the sales representative performance compared to their quotas, based on their seniority at the organization.



At organizations, accountability metrics are part of larger strategic goals, such as achieving certain profit margins, reducing overhead costs, or improving customer experience. These metrics align different teams' efforts toward the same purpose and can help ensure the organization remains focused on its strategic objectives.

Similar to organizational strategy metrics, accountability metrics should be revisited regularly to ensure they align appropriately with overall performance metrics. As the organization grows and changes, the metrics teams are held to should evolve to represent their most important work. Ultimately a thoughtful approach to measuring work for accountability can drive performance and encourage a culture of transparency and continuous improvement.

### 7.1.4 Activity

You are an analyst starting a new role supporting a company's marketing department. Your team has some existing dashboards and documentation about the following metrics, which are currently being tracked or were previously considered essential to the company.

Number of website visitors	Average time on page (measured in minutes)
Email open rate (% of emails opened by customers)	Customer lifetime value (LTV)
Customer acquisition cost (CAC)	Number of leads generated
Conversion rate (% of leads that become customers)	Cost per lead
Revenue growth (% increase in revenue)	Spam rate (% of emails marked as spam by the customer)
Social media engagement rate (number of views, shares, and clicks per post)	Return on advertising spend
Number of webinar attendees	On-time launch rate of campaigns

1. From the list of metrics above, identify which might belong to the three metric categories we discussed: **performance**, **operational**, and **accountability** metrics. Metrics can belong to more than one.
2. For metrics most appropriate as operational or accountability metrics, which might need to be measured separately for individual teams instead of the entire marketing organization?
3. Propose a hypothesis for which operational metrics might impact overall performance metrics. How would you test this?

# 7.2 The Key Principles of Metric Design

Designing metrics is a crucial skill set for an analyst. It requires a deep understanding of your organization's goals and objectives and a comprehensive knowledge of the processes and activities you measure. Across nearly any organization, you can apply the foundational principles of metric design to create appropriate and actionable targets to monitor and leverage for success. We will delve into the SMART framework for metric design, ensuring you can connect more granular metrics to organizational goals and set appropriate targets to provide context for evaluating success.

## 7.2.1 Leveraging the SMART Framework

The SMART framework for metric design is a powerful and widely used approach for creating measurable objectives at an organization. SMART (Specific, Measurable, Achievable, Relevant, and Time-bound) is an acronym that encompasses key characteristics necessary for effective performance tracking. By adhering to this framework, an analyst can develop well-defined and highly effective metrics for the organization to drive performance and guide decision-making.

### Specific

A metric should clearly and concisely represent a **specific** aspect of your organization's performance or objective. You should avoid vague and ambiguous language in your definitions to ensure all stakeholders understand what you are measuring. The metric should specify what needs to be accomplished, what success looks like, and the expected results.

**Table 7.1** Many metrics are colloquially discussed as concepts and require stakeholder collaboration to create specific operational definitions.

High-Level Concept	Specific Metric
Team productivity	Percentage of tasks completed on time
Customers supported	Percentage of customer issues and questions resolved on the first contact (First Call

	Resolution/FCR)
Application is easy to use	Median time to complete the setup flow for new users

For example, if your stakeholders ask you to measure "*customer satisfaction*," you will need to work with them to specify the *operational definition* of this concept. The metric you choose should be a *reliable* and *valid* (see chapters 2 and 6) quantitative measure agreed upon by teams who intend to use it for strategic decisions. Using customer self-report data, your specific metric could be the "*percentage of customers reporting that they were 'Satisfied' or 'Highly Satisfied' with their experience.*" This *precise* operational definition will speed up the alignment process with your stakeholders.

## Measurable

A metric needs to be quantifiable and capable of being **measured** objectively. Once you and your stakeholders agree on a quantitative and objective measure, your role as an analyst includes ensuring that your organization *reliably and accurately captures the data over time*. In practice, this requires collaborating with multiple teams to ensure your data can meet the following criteria:

- The data is captured at **regular intervals over time**. For example, if you're using questions from a customer satisfaction survey measured once per quarter, you should expect a similar number of responses to be recorded during each time period.
- The organization is committed to **continued investment in collecting the data**. Metrics are captured over medium to long periods of time (e.g., monthly over 1+ years). Thus, your organization will need to collect data for a minimum agreed-upon time period (e.g., two years) without changes to the underlying data or collection frequency.
- The data has consistent, **defined units of measurement** that can be relied upon to aggregate the data. For example, the formats of the response scales for the customer satisfaction survey should remain consistent. Additionally, the tables in the data warehouse containing responses should *only* change when necessary and with sufficient

warning, allowing you to modify queries and dashboards.

- The data is **high-quality** and **accurate**. You and your stakeholders should be able to trust that the values are correct representations of the information collected. If there is poor-quality data (e.g., wrong date types, duplicate records), appropriate steps should be in place to correct for issues that would impact your metric values.

**Figure 7.7** This dataset contains duplicate primary keys (`ticket_id`) and a bad date value that can degrade the quality of any metrics created.

	ticket_id	customer_id	category	priority	status	created_date
0	1527	cid_901	General	Low	Open	2023-06-10
1	1643	cid_986	General	Low	Closed	2023-05-17
2	3390	cid_569	General	High	Open	1023-01-01
3	4389	cid_754	Billing	Low	Resolved	2023-05-28
4	4389	cid_886	Product Inquiry	High	Open	2023-06-07

## Achievable

Your organization's goals using metrics should be **attainable**, given its available resources. Organizational performance and accountability metrics can be highly motivating *if* teams believe they can be achieved. An overly ambitious or unattainable target can be frustrating and demoralizing, leading to burnout and lower performance for employees and teams. On the contrary, setting too easy goals is unlikely to create meaningful improvement.

Setting achievable and realistic goals requires balance and continuous iteration. You and your stakeholders may easily see that a "100% customer retention rate" metric is unrealistic. However, figuring out an appropriate target requires more understanding of your customers and their needs. If you don't have a concrete understanding of the actions that will contribute to your churn rate, then you are unlikely to influence it.

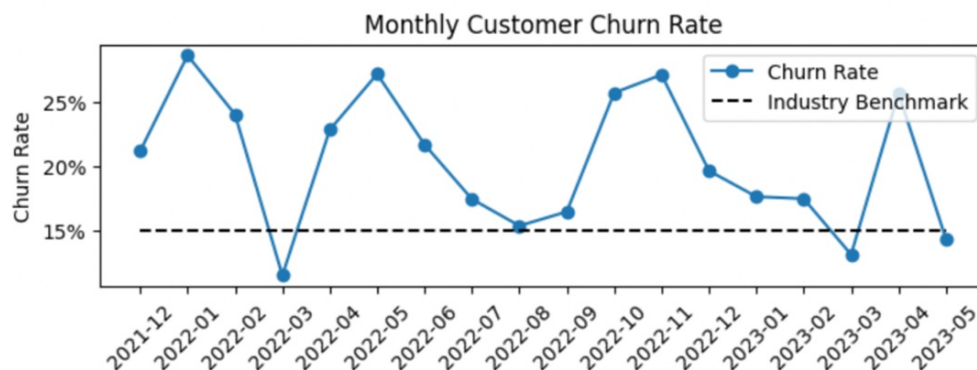
In general, the following information is necessary to set appropriate, achievable targets:

- An understanding of long-term trends in your proposed metric over long

periods of time (ideally, at least two years so you can compare seasonal changes)

- Benchmark comparisons of your proposed metric for organizations with similar characteristics (e.g., the same industry, size, geography, and type of service). These benchmarks are valuable in understanding the range of feasible targets for your metric of interest.
- An understanding of factors that correlate with or predict your proposed metric. If you wish to set targets for improving a metric value, your organization needs to understand what strategic actions will enable them to achieve those targets.
- Information about key factors outside of your control that impact your organization, its key metrics, and targets you set. For example, production targets at a clothing manufacturer will typically be lowered when there are shortages of necessary materials.

**Figure 7.8** This chart shows a company's customer churn rate compared to an industry benchmark. Churn tends to be higher than the benchmark, which suggests there may be room for improvement in the customer retention rate.



We will discuss strategies for establishing baselines and identifying the benchmarks necessary to set achievable goals in section 7.2.2.

## Relevant

Metrics require cross-functional investment to develop, communicate, and continuously monitor progress. For an organization to derive value from this effort, a metric should align with its strategic goals and priorities. It should directly relate to the area measured and provide meaningful insights that

contribute to decision-making and performance for the organization.

Organizations can easily determine the relevance of their *performance metrics*. These metrics are tied to the bottom line (e.g., finances, efficiency, and customers). Identifying the most relevant strategy and accountability metrics requires additional analysis to understand the relationship between these metrics and overall performance outcomes.

For example, an engineering team is interested in understanding its effectiveness in supporting the organization's goals. Leadership proposes looking at the *total lines of code* written by every engineer each quarter to measure productivity. While this is easily measurable, and the team can set targets against it, those targets are unlikely to produce meaningful outcomes for the organization. Lines of code written do *not* directly measure the software's performance, usability, or user value and don't provide actionable insights for product development. Instead, the engineering team will likely benefit from focusing on metrics that capture well-defined indicators of success for their users.

**Table 7.2 Be prepared to investigate the relevance of your metrics and ensure they are tied to the success of your team and organization.**

Domain	Not Relevant	Relevant
Engineering	Number of lines of code	Percentage of Bugs Resolved
Product	Number of features developed	Customer retention rate
Marketing	Number of social media followers	Customer Acquisition Cost (CAC)
Finance	Employee absence rate	Gross profit margin

The examples in this table show that a measure can be specific, measurable, achievable, and still irrelevant to the organization's goals. Often, irrelevant metrics are chosen due to their convenience, availability, and perceived association with the actions taken by individuals and teams. This can lead to months of wasted effort trying to achieve goals that do not create the desired impact on the organization. Be cautious to ensure that a metric meets **all** criteria of the SMART framework we have covered thus far before moving to

the final step!

## Time-bound

A metric should have a **defined period of time** for its measurement and goal-setting. A metric should be linked to a specific timeframe, such as daily, weekly, monthly, or quarterly. This allows teams and organizations to monitor and compare metrics to determine progress. Establishing clear time periods for metric targets (e.g., results reported at the end of the quarter) creates a sense of urgency. It helps teams prioritize work and allocate resources based on progress toward the goal.

Setting time-bound metrics requires understanding the timespan in which the underlying behavior changes. As an analyst, you should expect to account for a range of factors related to the time series data of your metric:

- The **time it takes to complete the behavior or process** impacts the time period you should choose for your metric. If you are creating a metric for the time to complete a workflow that takes an average of 2 months, you will likely need to aggregate the data by quarter instead of weekly or monthly to see meaningful changes.
- The **expected length for initiatives** intended to influence metric targets should be considered when choosing the time period. If you are running an A/B test with an expected duration of two weeks, a quarterly metric may not be at the appropriate granularity to represent any real effects of the test.
- Many processes have **seasonal variation** that creates recurring daily, weekly, or monthly fluctuations. You must consider an aggregation period that removes the seasonality to set actionable targets.
- In addition to seasonal patterns, metrics vary in stability across different periods. If you want to set targets that you can reliably achieve, you will need to find a time period that mitigates instability without eliminating variability. This can be seen clearly in Figure 7.8, where the customer churn rate has a lot of random variation, making it difficult to set goals for an individual month.

While many issues can be mitigated using a larger time period (e.g., monthly

instead of weekly), doing so can obscure the effects of actions taken to impact the metric. An initiative that starts in the third month of a quarter will have minimal impact on that quarter's metric, and it may take the entire *next* quarter to see results. The best time period to choose will depend on the process, your organization's needs, and the shape of the data.

**Table 7.3 Recommended considerations when choosing a time period for metric aggregation.**

Time Period	Visible Seasonality	When to Use	When to Avoid
Daily	Weekly, monthly	Short-term tracking, processes needing a quick response	Long-term tracking
Weekly	Quarterly, yearly	Short to medium-term tracking where daily data contains too much noise	Long-term tracking, short-term tracking where daily fluctuations are meaningful
Monthly	Yearly, multi-year	Medium to long-term tracking & forecasting	Short-term tracking & multi-year tracking (3+ years)
Quarterly	Multi-year	Long-term tracking & forecasting	Short-term tracking

**Let's introduce our case study for the chapter:**

Alex is a sales analyst at a high-growth SaaS (Software as a Service) company. The sales team is looking to expand its base of potential customers but has struggled to determine which marketing efforts are most effective. Alex is tasked with researching and proposing a metric to address this problem and decides to apply the SMART framework to evaluate potential metrics.

**Specific:** Alex recognizes that the sales team needs a metric that reflects the impact of marketing efforts on the sales team's processes, specifically. Thus, he decides to focus on whether sales leads are *qualified* (likely to purchase the software). He proposes a "Percentage of Sales Qualified Leads by

Marketing Source" metric.

**Measurable:** Alex determines that the "Percentage of Sales Qualified Leads by Marketing Source" metric can be quantified by tracking the number of leads originating from each marketing source that the sales team validates as high-quality (Sales Qualified Leads). This is available using the organization's system for recording lead source data.

**Achievable:** Alex proposes implementing the new metric using existing processes for tracking leads. The only change required would be to ensure the sales team tracks the factors in the system that determine whether a lead is *qualified*. He notes that the team can easily be trained, but the data is manually entered and may need to be monitored for quality.

**Relevant:** Alex is confident that the proposed metric directly addresses the sales team's challenge of determining the effectiveness of marketing efforts. It can provide insights into where the best leads are coming from, enabling more targeted and effective marketing.

**Time-bound:** Alex proposes tracking the metric monthly for three months to gather enough data for initial analysis. He knows all of the company's sales metrics are evaluated monthly or quarterly and that too few sales are made every week for a more granular time period to be valuable. After the initial period, he proposes evaluating the metric's effectiveness with the marketing team and making any necessary adjustments.

## Recap

Designing impactful metrics requires strategic consideration of your organization's goals, performance, and the meaning of success. Applying a clear and concise framework like SMART ensures that the metrics you choose aligns with those goals and are quantifiable, attainable, and meaningful in the long term. Overall, the SMART framework offers a clear structure for the research and development of metrics, which can be ambiguous and challenging.

## 7.2.2 Establishing Baselines and Targets

Creating SMART metrics at an organization requires a lot of background information to set goals and take action to achieve them. After identifying a metric, it's necessary to establish *the current baseline state* to understand performance levels before any changes or interventions are made. This serves as the starting point for strategic decision-making. Determining appropriate baseline values for your metrics involves thoroughly analyzing historical data, peer-reviewed research, and industry standards where available.

The analysis in creating metrics baselines creates structure and clarity for your performance improvement efforts. It allows your stakeholders to build a shared understanding of previous trends, user behavior, and external processes related to your organization's goals. From this data-informed perspective, teams can more easily align on what's possible to achieve and with what effort.

## **Analyze Historical Data**

Once you have proposed a *specific* and *measurable* metric, an analyst's role is to explore historical data on the underlying measure comprehensively. This allows you to build subject matter expertise in the behavior being represented and recommend strategic approaches to move the needle on the metric. The resulting insights can help determine if it's *achievable* to set targets, as well as what *time aggregation* is most appropriate, and continually iterate on goals in the long term.

Analyzing historical data for a proposed metric requires examining the data from *many* different perspectives. It involves investigating the shape, trends, anomalies, segments, time-series data, and correlations with other known measures and metrics. No two analyses are identical since your findings will lead you toward more granular investigations. At a minimum, I recommend considering the following questions when examining historical metric data:

1. **What is the shape of the data?** To visualize the distribution, look at the data using a histogram or boxplot. Is the data *skewed*, normally distributed, or neither? How concentrated is the data around the median (kurtosis)? These characteristics will tell you what's *common* and possible regarding metric values.

2. **Are there outliers in the data?** In which direction? Are they extreme enough to skew the mean of your dataset aggregated for a given time period? Atypical records in your distribution tell you about subsets of users or customers that may be falling through the cracks or are super users of your product or service.
3. **Are there differences between segments?** Consider how your organization divides its customers and users (e.g., company size, geographic region, age group). How your organization is already partitioning its user base is a great starting point for investigating differences in new metrics. Try comparing measures of central tendency and overlapping distributions between each segment to understand where meaningful differences exist and whether they support existing hypotheses about those segments.
4. **How does the data change over time?** Metric values are unlikely to stay constant over the long periods in which they grew and evolved. Try partitioning your data into years and other meaningful time periods. Do you notice any new trends?
5. **How do metric values change over a customer or user's tenure with your organization?** What trends do you see if you start calculating data from when a person joins your application, purchases a product or service, or engages with your organization? For example, do new customers have lower scores on your metric than loyal, tenured customers? Are there clear trends in the first 7, 30, 60, or 90 days when someone is a customer with you?
6. **Are metric scores correlated with any of the following?**
  - a. Time of year (month number, week number), indicating seasonality
  - b. Length of time as a customer, indicating growth or decline by tenure with your organization
  - c. User activity metrics (e.g., number of weekly active days), indicating a relationship with one or more other activity and usage behaviors

Let's import a customer activity dataset to show how to examine a metric with these steps. Each dataset row has the number of days out of the month that the customer logged in. We also have information on each customer's region as a potential segment.

```
import pandas as pd
```

```
import matplotlib.pyplot as plt
import seaborn as sns      #A

logins = pd.read_csv("customer_logins.csv")      #B
print(logins.head())      #C
```

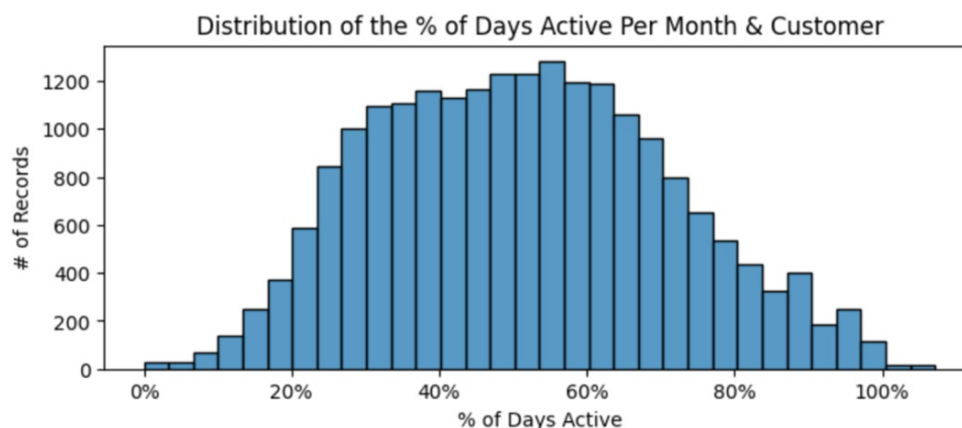
	customer_id	region	month	login_days
0	93	Europe	2020-01-01	21
1	346	Europe	2020-01-01	12
2	404	Asia	2020-01-01	22
3	347	North America	2020-01-01	15
4	403	Asia	2020-01-01	30

We can start exploring patterns in the metric data by generating a new column representing the *percentage* of days in a month that a customer has logged in, standardizing for the slight variation in days per month. We can then generate a histogram of the percentage values recorded *per customer per month*.

```
logins["n_days"] = pd.DatetimeIndex(logins["month"]).days_in_month
logins["login_days_pct"] = logins["login_days"]/logins["n_days"]

sns.histplot(x=logins["login_days_pct"], bins=32)      #C
plt.title("Distribution of the % of Days Active Per Month & Customer")
plt.xlabel("% of Days Active")
plt.ylabel("# of Records")
```

**Figure 7.9** A histogram of the percent of active days in a month shows a possible bimodal distribution



The histogram shows that most customers are active between 25% and 75%

of the days in the month. There are potentially two or more modes in the distribution, which is relatively wide. Very few customers have months when their activity is above 80% or below 20%. If we want to determine if the *same* customers are represented in those outlier months, we can aggregate the dataset to see if any individual customer IDs occur more frequently than others among the outliers.

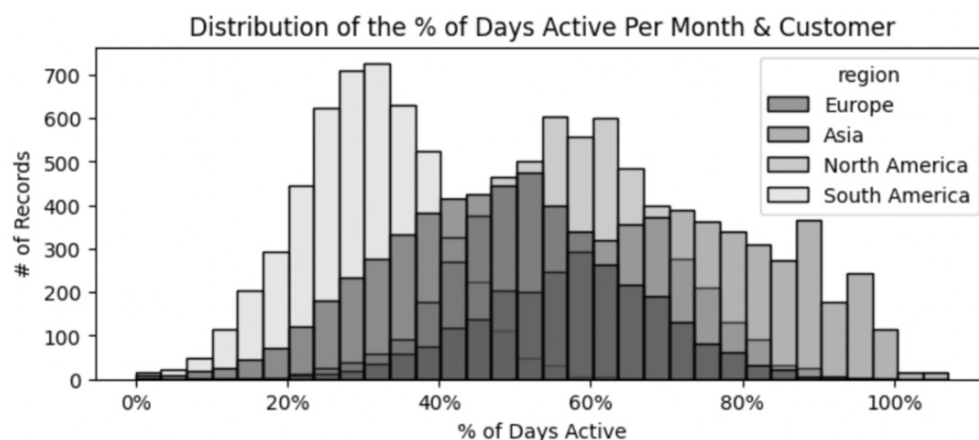
```
low_pct = logins[(logins["login_days_pct"] < .2)]      #A
low_ct = low_pct.groupby("customer_id").size().reset_index(name="
print(low_ct.sort_values(by="ct", ascending=False).head())    #C
```

	customer_id	ct
145	456	10
229	627	9
52	228	9
221	612	9
24	156	9

Let's break out the chart by the region segment provided in the dataset to add more nuance to the original histogram.

```
sns.histplot(x=logins["login_days"], hue=logins["region"], bins=3
```

**Figure 7.10** Histograms of active days broken out by region segment to better understand the data.



We can see that breaking out the metric values by segment adds *a lot* of additional nuance. There are distinct underlying distributions for each region. Customers in Asia seem to have the highest median percentage of active days, followed by North America. Customers in South America tend to be the

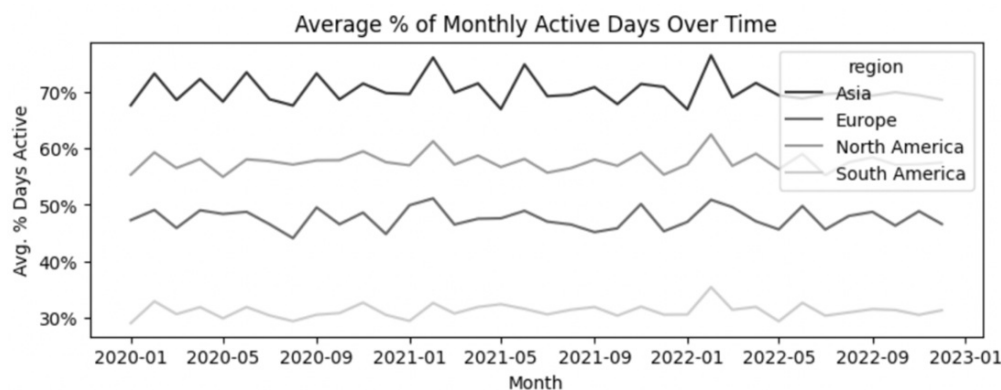
least active. As part of research into establishing baselines, this finding suggests that the metric should have *separate targets and tracking* for each region.

Next, we can explore the time series data broken out by the region segment. We'll start by looking at values over the calendar month across all three years in the dataset.

```
avg_logins = ( #A
    logins.groupby(["month", "region"])["login_days_pct"].mean().
)

sns.lineplot(x="month", #B
    y="login_days_pct ",
    hue= "region",
    data=avg_logins)
plt.xlabel("Month")
plt.ylabel("Avg. % Days Active")
plt.title("Average % of Monthly Active Days Over Time")
```

**Figure 7.11 Average percentage of monthly active days, aggregated by month across three years.**



The data show apparent differences between regions that remain stable over the three years in this dataset. There is some random fluctuation but no discernable trend from one month to another. We cannot say there's evidence of any increase or decrease that has occurred organically until now. Next, let's investigate potential seasonal trends by aggregating the data monthly and yearly. In order to make the chart easier to interpret, we will also convert the month numbers to names.

```
import calendar
```

```
logins['month'] = (      #A
    logins['n_month'].apply(lambda x: calendar.month_name[x])
)

avg_logins_m = (      #B
    logins.groupby(["year",
        "month"])[ "login_days_pct"].mean().reset_index()
)

sns.lineplot(      #C
    x="n_month",
    y="login_days_pct",
    hue= "year",
    data=avg_logins_m
)
plt.xticks(rotation=45)
plt.xlabel("Month Number")
plt.title("Average % of Monthly Active by Month and Year")
```

**Figure 7.12** Monthly active days are shown by year and month number to investigate potential seasonality.

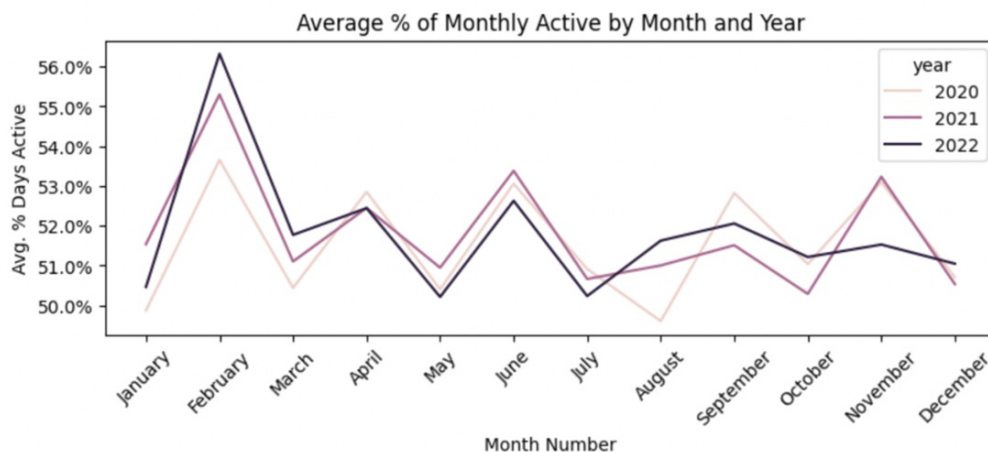


Figure 7.12 shows potential seasonality, where the percentage of active days was slightly higher in February compared to other months. However, the difference is marginal and only noticeable given the small y-axis range (50-58%). The trend should be verified with product, marketing, and customer teams to determine if the increased activity aligns with any known reason (e.g., an annual promotion).

At the organization this dataset is collected from, the next step is to explore potential relationships with existing metrics. Performance metrics such as

contract size, add-on purchase, and contract renewal rates should be compared to customer activity rates. Moderate to strong correlations can be the basis for hypotheses about the downstream impact of efforts to increase the new metric.

Correlational relationships can be used to develop hypotheses about *how* to move the needle on customer activity rate goals and what targets might be reasonable. However, correlations with other processes shouldn't be used to establish the *relevance* of a metric. In fact, a metric with extremely high correlations (e.g., 0.8+) is likely redundant and doesn't add new information to the organization's strategy.

The foundational understanding of historical metrics patterns helps stakeholders leverage and set goals more effectively. With a standard approach to exploration like the one we just covered, you will enable better, more strategic approaches to understanding the metric, user behavior, and how to ensure it meets the SMART criteria.

## **Identify Appropriate Benchmarks**

Recent metric data provide a partial view of the underlying measure and what it looks like to set targets effectively. *Benchmarks* are reference points used to compare against your metric values. *Benchmarking* refers to the process of comparing your metrics to the reference points you identified. Since a benchmark represents a performance target that has been previously achieved, leveraging these data points serves as a highly effective anchor point when setting your organization's own goals.

Many sources of data can be used as benchmarks to contextualize your metric values:

- **Industry benchmark data** provides an organization with the average performance data for its industry. These benchmarks are available for metrics commonly captured within that industry (e.g., win rates for sales deals, salary ranges for a job your organization is hiring for). This data tends to be collected by industry associations or third-party firms that provide or sell the information to individual organizations.

- **Peer-reviewed benchmark data** is valuable for organizations that work with processes and populations with a field of academic study dedicated to publishing novel findings on the topic. This type of data is available in academic journals. For example, a healthcare technology company can compare the accuracy of a new device it's developing to diagnostic timelines and accuracy in a database such as PubMed.
- **Governments, NGOs, and non-profits** collect vast data made available for public consumption. If you find value in benchmarking against social and economic indicators, there's a high likelihood you will have information available in your geographic location. Most of these data sources are collected by surveys designed to be population-level estimates of a phenomenon. For example, the unemployment rate metric we reviewed at the beginning of this chapter is updated monthly, along with dozens of other indicators available on the U.S. Bureau of Labor Statistics website. Measures such as job postings, public health information, population attitudes, public company quarterly earnings, and more are freely available to you and your organization as benchmarks. These high-value data sources can help you understand the broader system in which your organization operates.

**Figure 7.13 A footwear company's annual sales compared to a benchmark captured from the Bureau of Labor Statistics (BLS) Consumer Expenditure Survey**



Your organization will benefit from multiple sources and benchmark data segments for strategic goal-setting. Suppose you can segment and filter benchmark data to represent better your organization's characteristics (e.g., their industry and company size) or their user or customer base (e.g., ages, geographic locations). In that case, your targets will be far more realistic and achievable.

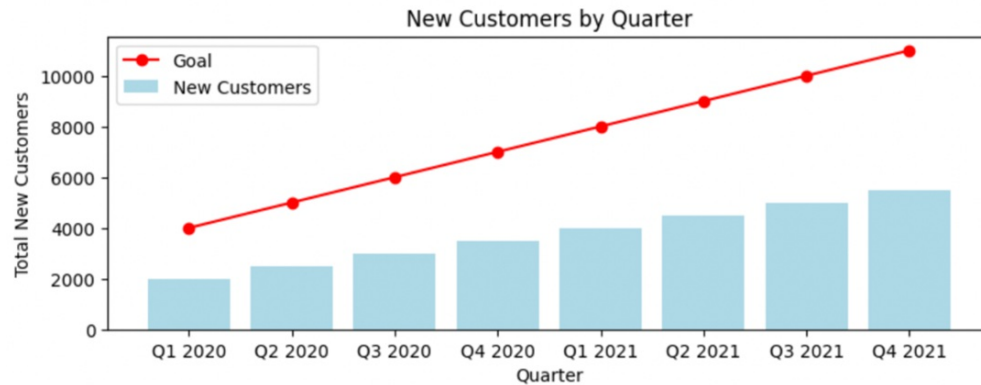
## Set Realistic Targets

Once a baseline is established, the next step is to set targets. A *target* is the desired future state or performance level of a metric you set that represents an improvement in a process or experience. Targets should be ambitious *and* realistic and must align with the organization's strategic goals.

Organizations complete an in-depth analysis of historical data and benchmark metrics to set informed targets and to understand the landscape in which they operate. This helps ensure that targets are *achievable* and *meaningful* in driving growth and improvement. Consider the following questions when proposing targets to help determine their appropriateness for the teams who will be held to account for their success:

- How confident are teams and stakeholders on what actions to take to drive change? Are there known factors that correlate with or predict change in the metric?
- Is there evidence that the metric is movable? Is the metric stable, or does it fluctuate randomly despite efforts to drive growth?
- Do teams and stakeholders have the capacity to drive growth and improvement at the levels they're targeting? Will they be able to improve the *quality* of their operations, or are they simply expected to take on an additional *quantity* of work?
- Will the targets set motivate the teams to achieve them, or will they reduce morale and degrade their performance capacity if they cannot reach their goals?

**Figure 7.14 A target of achieving two times the quarterly new customer registrations per quarter is unlikely to be successful or motivating to the teams responsible for these goals.**



Your analytics team will likely not be setting targets for the *entire* organization, but your expertise on the data will enable you to make informed recommendations on these goals. **If the opportunity exists, it's valuable to advocate for an analytics team member to participate in strategic planning conversations to inform the goals being set for the organization, teams, and individual employees.**

## Iterate

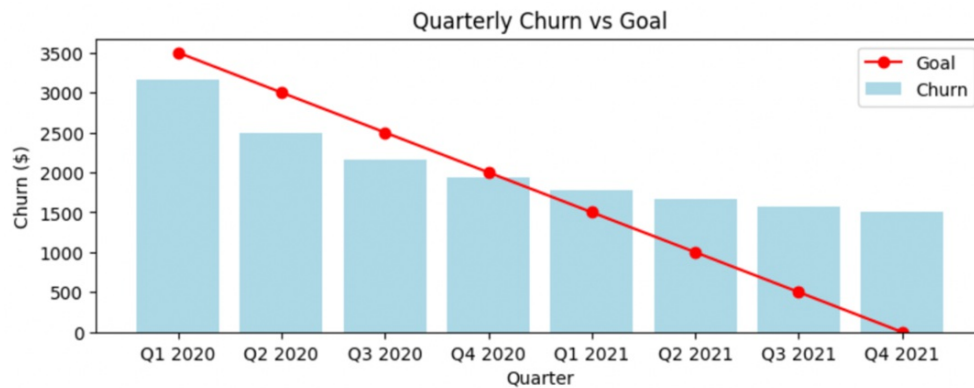
Setting targets is not a one-time task. Organizations and their landscapes constantly evolve, requiring they revisit and adjust metrics and targets accordingly. This iterative approach keeps metrics relevant and ensures they provide valuable insights and strategic direction.

Targets for improvement should rarely stay static from one time period to the next. For example, a goal of decreasing customer churn by five percent each quarter cannot continue until churn is reduced to nothing. Customers will leave, new customers will join, and users will change their behaviors. It's unrealistic to assume that improvement in a single process will continue indefinitely. You and your stakeholders *will* be disappointed, and your organization will spend valuable resources on a task with diminishing returns. As part of iterating, shifting focus between relevant processes over time is beneficial. This balanced approach ensures the organization operates with the latest and most pertinent information.

When setting targets, it's worth reminding stakeholders that growth trajectories cannot continue indefinitely. Determining *when* your improvement rate will slow is challenging, but knowing it will eventually

happen is necessary.

**Figure 7.15 Growth trajectories cannot be linear forever. Within four quarters, a static goal for reducing customer churn quarter over quarter has diminishing returns and eventually becomes unfeasible.**



For example, large tech companies like Facebook and Netflix have millions or billions of users. Their metrics will likely include targets intended to drive the growth of their user base and its engagement. Given the scale of these companies, their total user base is, quite literally, limited by the size of the human population with access to the Internet. Beyond those levels, they are forced to expand into new industries or accept the limits of their business potential.

**Let's return to our case study for the chapter:**

Alex's next step is to conduct an exploratory analysis of historical data for the new proposed Percentage of Sales Qualified Leads metric. This involves reviewing the company's sales data history to understand patterns and trends in the types of leads considered "qualified," how the definition has shifted, and how many leads meet each criterion.

Alex begins by exploring the distribution of the metric using a histogram. He discovers that the metric's distribution has a strong right skew, indicating that very few leads have high scores compared to most leads, whose scores are quite low. When measured over time and grouped by quarter, the percentage of qualified leads fluctuates, but no apparent increase or decrease across three years.

Next, Alex breaks out the distribution into three segments relevant to the company: industry, company size, and geographic region. He discovers that the percentage of qualified leads varies widely by industry, with the highest scores seen among technology companies. This suggests that the company's product is well-suited to the technology industry and would benefit from increasing its focus on this sector.

Finally, Alex explores the strength of correlations with the Percentage of Sales Qualified Leads metric and existing sales metrics, such as the win rate (% of leads won) and the average deal size (dollar value of a contract). He discovers a moderate positive correlation between the new metric, close rate, and average deal size. This suggests that leads with a higher score are more likely to convert to sales and that those contracts will be higher value. This encourages Alex to further recommend the new metric as a potential predictor of sales success.

Alex creates a report with each of the visualizations he created as part of his exploration and an interpretation of each step of the exploration. He presents the findings to the rest of the team. Based on the analysis of historical data and benchmarks for the technology industry performance, the sales team decides to set a goal of increasing the overall sales qualified lead percentage to the same levels seen among technology companies. Achieving this target will likely involve an increased focus on the technology sector when searching for new sales leads and doing a more in-depth analysis of factors in other industries that have them categorized as qualified leads.

### **7.2.3 Activity**

Now that you've identified the marketing team's performance, organizational, and accountability metrics, let's leverage them to develop a SMART metric. We will create and validate one or more metrics to represent the broad objective of measuring "website engagement." Assume that you can access website analytics data, including information on page views, session durations, click event data (e.g., for buttons on a page), and summary data about engagement with social media posts.

1. Refine the objective of measuring "website engagement" to be more

specific. What are some possible **specific** definitions that will indicate engagement with the website? Based on the available data, are there any pros, cons, or conflicting definitions?

2. Propose one or more ways to measure this objective. Is it already available in the list of metrics and datasets? Is there any additional data necessary to collect for your specific definition to be **measurable**?
3. Consider what an improvement would look like for your proposed metric. What would be both meaningful and **achievable**?
4. How can you document where this metric is **relevant** to your organization? How is this important to your team's current goals? What do you hypothesize will be the impact of setting goals for this metric?
5. Propose a specific timeline for achieving this goal. What advantages and disadvantages might exist if your metric is **time-bound** by week, month, or quarter?
6. After assessing each of these criteria, document your **SMART** metric: What is the name of the metric and its goal that you can communicate to your stakeholders and the broader organization?

## 7.3 Avoiding Metric Pitfalls

Not all metrics, or their visual representations, are created equally. While metrics at an organization can guide decision-making and demonstrate progress toward goals, they can also mislead, confuse, and misrepresent reality if not appropriately constructed or displayed. We'll discuss common mistakes in calculating and presenting metrics that can mask underlying trends, distort findings, and confuse you and your stakeholders. These steps are necessary to complete *in addition* to ensuring your metric meets the SMART criteria and will ensure your insights drive informed and accurate decisions.

### 7.3.1 Representation

A SMART metric can be interpreted and understood differently depending on how it's calculated and presented to end users. While some representations (e.g., mean vs. median, bar graph vs. line graph) are suitable given the data's shape, some common ways of displaying metrics should always be avoided.

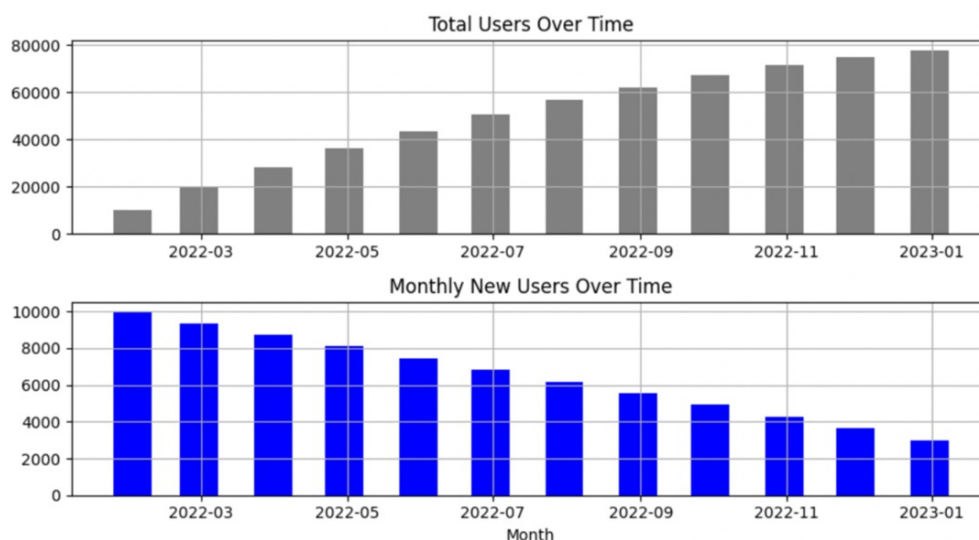
It's all too easy for analysts to fall into the trap of using inappropriate representations that can distort your results and detract from the strategic purpose of the metric.

## Cumulative Values

Metrics that track a count or sum of values over time can be represented as an *incremental* or a *cumulative* calculation. An *incremental metric* includes only new values for a period (e.g., net new monthly sales). This metric type can be directly compared from one time period to another to understand how processes change over time. Incremental metrics are beneficial because they depict fluctuations from one time period to the next, alerting organizations to changes in performance that should inform and motivate actions.

A *cumulative metric* calculates all current and previous values for a given time period. Instead of capturing values for a particular period, a cumulative view of the data shows the total up to that point. While this can be useful for showing total progress over time, cumulative metrics can hide fluctuations or trends that should otherwise motivate strategic choices at the organization.

**Figure 7.16** A cumulative user count shows the overall growth but may disguise a decrease in growth rate.



Cumulative metrics document the state of an organization at various points in

time. For instance, organizations need to adopt a cumulative view of the data to track the growth of their total recurring revenue and customer base over time. This is illustrated in figure 7.16, where the cumulative total of users each month reflects the user base's current state and overall growth trajectory. To fully understand these trends, the cumulative metric needs to be accompanied by an incremental version of the same metric. We can see that although the user base is growing, the growth *rate* is shrinking. This nuance is harder to discern from the cumulative chart alone.

When unaccompanied by an incremental calculation, an otherwise SMART metric calculated as a cumulative value can fall into the trap of being a *vanity metric* for an organization. Vanity metrics are data points that look good on paper, are easy to celebrate, and usually *don't* contribute to actionable insights. A chart with a continually increasing value (e.g., cumulative user base) can be easily celebrated but does *not* inform goals, actions, and strategies.

Ultimately, cumulative metrics are rarely as valuable as analysts and stakeholders believe. They can mistakenly portray areas for concern at an organization in a positive light and often mask subtle trends or shifts in your data. Make sure to accompany cumulative metrics with an *incremental* calculation to ensure your stakeholders know the actual status of goals.

## Granularity

Metric development often includes choosing an appropriate *grain* for your data *before* making aggregate calculations. *Granularity* refers to the level of detail available in each row. Like the tradeoffs you make with measures of central tendency (see chapters 4-5), choosing the grain can highlight or minimize different characteristics of the data you present to stakeholders, impacting your decisions.

Organizations differ *widely* in the granularity considerations based on their customers, business model, and purpose. The following are some common examples you may encounter in a project:

- In B2B (business-to-business), data is usually collected on *customers*

(other businesses) and their *users*, which belong to an individual customer. An analyst must determine whether it's appropriate to calculate a metric at the grain of the user or aggregated at the level of the customer.

- Most organizations collect **geographic information** about their customers and users. Geographic data can be easily aggregated at a variety of grains (e.g., zip code, city, state, country) and types of regions (urban vs. rural).
- **Product categories** at companies tend to have multiple levels of specificity and intersecting classifications for tracking and analysis. For example, the database of inventory at a retail company will classify departments by age groups (e.g., children's clothing vs. adults) by gender (men vs. women) and by type of clothing item (shirt, pants, etc.).

The choice of granularity is ultimately a balancing act. Higher granularity (e.g., zip code instead of state) can offer detailed insights, limiting your ability to identify broader trends. In contrast, lower granularity can obscure trends and hide meaningful subsets of data. As part of your metric explorations, segment your data at various levels of granularity to determine an appropriate *grain* to recommend to your stakeholders.

Let's walk through an example of this exploration and decision process when choosing whether to aggregate data at the user or customer level at a B2B company. The company has several hundred customers across various industries and company sizes. A sample of the customers table in their database looks like the following:

**Figure 7.17 Sample data showing information about customers at a B2B company**

	id	created_date	status	contract_amount	subscription_type	industry	company_size	referral
0	012	2023-02-17	Churned	13614.88	Pro	Tech	Large	No
1	013	2023-04-18	Active	14798.01	Pro	Tech	Large	Yes
2	014	2023-01-24	Inactive	14781.55	Basic	Finance	Small	Yes
3	015	2023-03-18	Churned	11471.28	Basic	Education	Medium	No
4	016	2023-04-07	Churned	18448.20	Pro	Healthcare	Large	Yes

If the company wants to create and monitor a simple *user activity* metric, it can start by creating an aggregate query that results in the following table:

**Figure 7.18 Example of a monthly active user rate aggregated by taking a sum of all users**

	month	total_users	active_users	pct_active_users
0	2022-08	47077	24800	52.679652
1	2022-09	47054	21504	45.700684
2	2022-10	47060	25568	54.330642
3	2022-11	47035	24700	52.514085
4	2022-12	47040	22185	47.161990

Each company in the current customers table ranges between 10 and 1000 employees, shown in the categorical company\_size column. If we take an overall sum for the metric, we are weighting the metric calculation toward the larger customers. Because they have more users, their activity rates will significantly impact the metric more.

To account for the differences in company size, we can *partially aggregate* the data at the company level, generating a user activity score *per company per month*. This intermediary dataset can be aggregated *again* per month overall, effectively removing the weight favoring large companies.

**Figure 7.19 Partially aggregated user activity data by customer and month**

	customer_id	month	n_users	n_active_users	pct_active
0	012	2022-08	697	599	0.859397
1	012	2022-09	692	70	0.101156
2	012	2022-10	695	472	0.679137
3	012	2022-11	695	600	0.863309
4	012	2022-12	695	396	0.569784

In figure 7.19, each customer is treated as an equal entity in the partially aggregated dataset metric. Assuming the company's strategy is to engage the entire customer base, an *average percent of active users by month* will provide more accurate and actionable insights on progress towards goals.

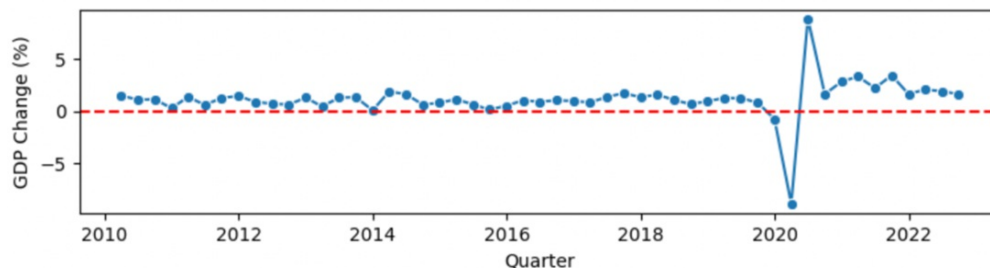
## Composite Metrics

A *composite metric*, or an *index*, combines two or more metrics into a single, standardized score for reporting to users or stakeholders. These are frequently used when measuring complex processes, and reporting on individual metrics

may create difficulty deriving actionable insights. Theoretically, combining multiple metrics allows an accessible, heuristic understanding of the underlying processes being measured.

Indexes are commonly used to report economic and social information, such as the Gross Domestic Product (GDP) provided quarterly by the Bureau of Economic Analysis. Due to its simplicity and wide use, the percentage change in GDP is comfortably reported to general audiences. A shared understanding of the metric is expected; a percentage increase is considered positive, and a decrease is considered negative.

**Figure 7.20** The GDP is reported as a dollar value and a percentage value, showing relative change from the previous quarter.



Indexes like GDP require extensive research to combine metrics appropriately. Individual items are expected to vary together, representing the same underlying process. We discussed some ways to build composite scales in chapter 5 (e.g., inter-item reliability analyses) that can maximize the accuracy and comprehensiveness of the index. However, indexes tend to have serious limitations in the accuracy and actionability of the number it provides:

- For each item you add to an index, **you reduce the variability in scores** over time and between entities (e.g., customers, countries). By attempting to simplify the calculation of multiple metrics, you will likely obfuscate valuable trends that you'd otherwise notice in the individual metrics a score is comprised of.
- Including or excluding individual metrics in an index is ultimately **an arbitrary decision** you and your stakeholders have to make. By designing and using an index, especially for internal tracking, you are committing your time to continually iterating on the index based on new

findings, processes, or services offered at the organization.

- Depicting a **balance of competing processes** can be challenging. You can create and report on counter-metrics (more on that later in this chapter), but they're unlikely to carry the same perceived importance as the actual index.

Each of these limitations is true for the GDP calculation, which calculates the value of all goods and services produced within a specific period. It's considered a critical indicator of economic health. Still, it *excludes information such as wealth distribution*, the depreciating value of goods produced, or goods and services provided in the black market. Each of these potentially impacts the health of the economy and the quality of life of individuals and communities represented in the GDP calculation. In theory, any of these *can* be included in the calculation. However, no singular index will provide an entire picture of the economy and still be responsive to change.

Unless your organization has vast resources to dedicate to the research necessary to develop an appropriate, data-informed index, I strongly recommend avoiding creating them. Indexes and composite metrics sound appealing to stakeholders due to their perceived simplicity and ability to represent the concept or process you are measuring heuristically. Still, they often detract from the ability to make data-informed decisions.

## Counter Metrics

When tracking a comprehensive range of important processes in an organization, you will discover that many appear to be in conflict with each other. For example, a software company may have the following two goals: improving the speed in which new features are launched in the application, and reducing the number of bugs reported in the application by customers. These are examples of *counter metrics*, which are metrics that serve as checks and balances and help uncover potential negative consequences from singularly focusing on one goal.

Counter metrics are identified in the *context* of a new metric being designed. They tend to be known valuable processes in the organization that are often

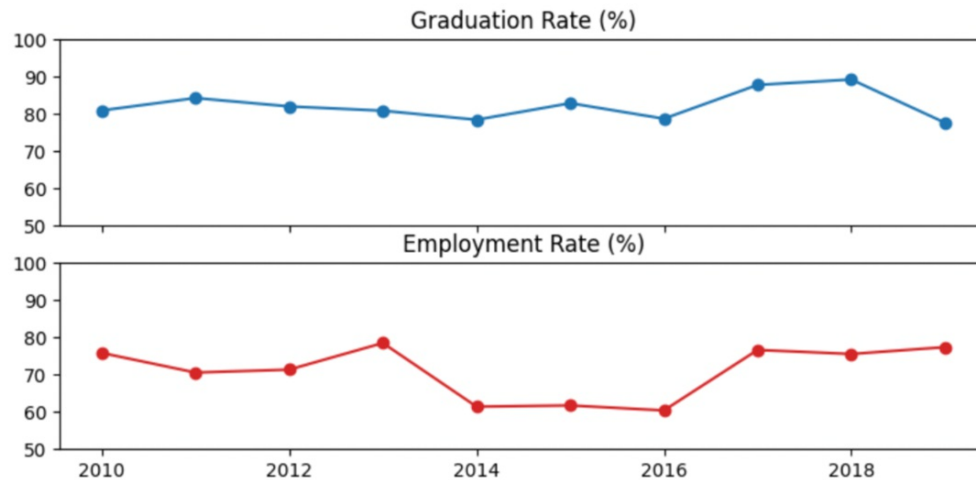
monitored for other goals. Identifying appropriate counter metrics is a qualitative process involving critical consideration of your new metric and its behavior or process. Ask your stakeholders questions: *What could go wrong if we focus only on this metric? Where might we be sacrificing quality in pursuit of quantity? What processes are potentially in opposition to this metric and still important?*

Consider the following examples:

- A customer support team has a metric tracking the *number of tickets resolved in under 24 hours*. As a counter metric, the team also reports on *customer satisfaction scores* to ensure they aren't sacrificing the quality of support in pursuit of quantity.
- A software engineering team tracks the *number of new features shipped*, which is considered one measure of team productivity. As a counter metric, the team also tracks the *number of new bugs introduced* to ensure that the quality of their code isn't sacrificed to increase productivity.
- A non-profit focusing on education tracks the college attendance rates of high school students participating in their programs. As a counter metric, they monitor college completion and student employment rates in their field of study. Although the non-profit does not have programs that directly impact these processes, they consider it essential to monitor students over the long term to ensure their success.

When preparing your metric deliverables (dashboards, reports, presentations), it's strongly recommended to incorporate a visualization of the chosen counter metric into that deliverable, explaining the counter metric and its rationale. Additionally, counter metrics should **always be evaluated alongside any metrics or measures for experiments, A/B tests, or evaluations**. Adding a section to a dashboard to visualize all counter metrics with an explanation of *why* they were chosen will empower you and your stakeholders to iterate on them over time and keep track of the complete picture of your goals.

**Figure 7.21** A simple visualization of a graduation rate metric and its chosen employment counter metric.



## 7.3.2 Visualization

At its core, a metric is a tool for telling a valuable story to your stakeholder, and how you visually depict it is a *massive* part of that narrative. A well-chosen visualization can speak for itself, highlighting trends that would otherwise take paragraphs to explain. However, poor representation of a metric can easily undo all the work you have done to ensure your metrics are interpretable and actionable.

Ultimately, the topics covered in this section apply to data visualization and storytelling. We will focus primarily on the representation of metrics as *time series data* (data presented over time, usually with the time variable as the x-axis). However, any type of visualization that isn't time-bound can still benefit from these considerations.

### Picking the wrong chart

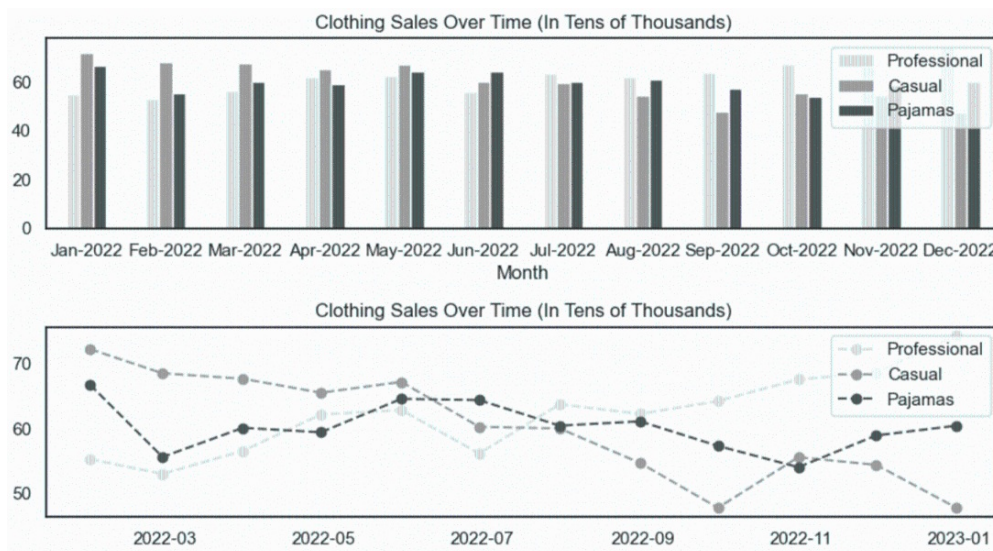
Inappropriate chart selection is *very* common. If you pay attention to news segments reporting survey results, studies, or other data, you will notice that the graphs chosen are not always the best choice for the data. Not all charts are appropriate for metrics, and many chart selection decisions can negatively impact your ability to interpret the data:

- A chart can easily **oversimplify** the trends you want to depict with your metric. A single bar or line graph may be helpful for an *overall* picture

of your metric across the organization (e.g., figures 7.8 and 7.16), but be prepared for different departments and teams to want a more granular view of the metric (e.g., by region, industry, etc.).

- Conversely, a chart can easily **overcomplicate** a trend if too much information is included or if it's not a visualization that your stakeholders are familiar with. If you plan to break out your metric by segments, I strongly recommend using a line graph instead of a bar graph. A grouped or stacked bar graph shown over time can be challenging to interpret, compared to following the growth trajectory of a set of lines, provided they are unique in color or representation.
- It's easy to pick the **wrong chart** when developing a dashboard or report to track metrics over time. Many chart types available in business intelligence tools are tempting, especially when looking to showcase your work. However, if you expect that you'll have to explain the chart's contents to your executives, then it's probably worth using a simpler visualization in deliverables. The type of analysis you perform should motivate your chart selection—not the other way around. I recommend reading Noah Iliinsky's paper, *Choosing a Successful Structure For Your Visualization* [4], to learn more.

**Figure 7.22** Line graphs are typically easier to interpret when visualizing trends between groups over time.



Usually, a grouped bar or line graph is the best choice for appropriately visualizing a metric. Depending on the calculation, you can easily group or

stack data, and most of your stakeholders can interpret the graphs without aid. Avoiding more complex charts will minimize the follow-up necessary for you and your team.

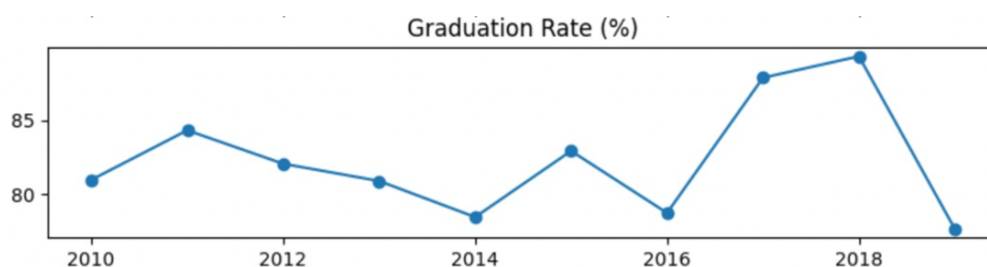
## Distorting the Axes

Another common visualization pitfall is the misuse of the y-axis. The y-axis represents the *scale* of your metric, which shows your stakeholders the size of fluctuations over time. It's incredibly easy to distort the meaning of change from one time period to another by manipulating the y-axis.

There are several key ways in which y-axis adjustments can distort metrics:

- **Truncating the y-axis**, or trimming the upper and lower limits, can exaggerate variations in the data. This can make trends appear more impactful than in reality, potentially motivating decisions based on only minor differences. This is important to monitor *even when it's not done intentionally*, as data visualization tools use the range of the dataset to set the y-axis limits. If the full range of values extends beyond the minimum and maximum, you must set the limits manually.

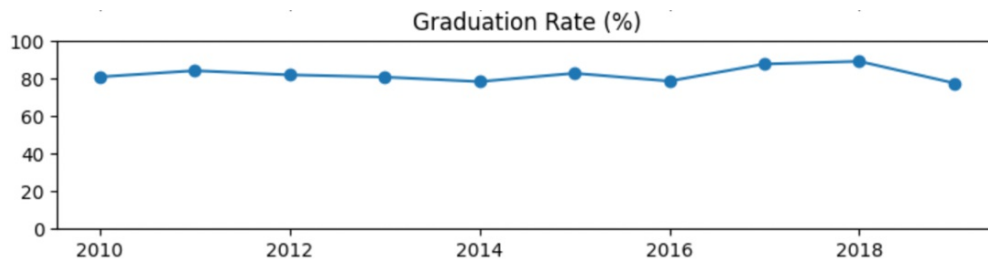
**Figure 7.23** This figure shows the same data as Figure 7.21, but with the default y-axis limits set using matplotlib based on the minimum and maximum values in the dataset. These defaults exaggerate the perception of fluctuation in the data.



- **Extending the y-axis** beyond the actual range of values can diminish the ability to perceive trends or fluctuations. Many metrics with an actual minimum and maximum (e.g., percentage values) will only ever return values for a limited subset of that entire range. The graduation rate metric in figures 7.21 and 7.23 can theoretically range between 0% and 100%, but a decade's annual data shows no values below 75%. If the

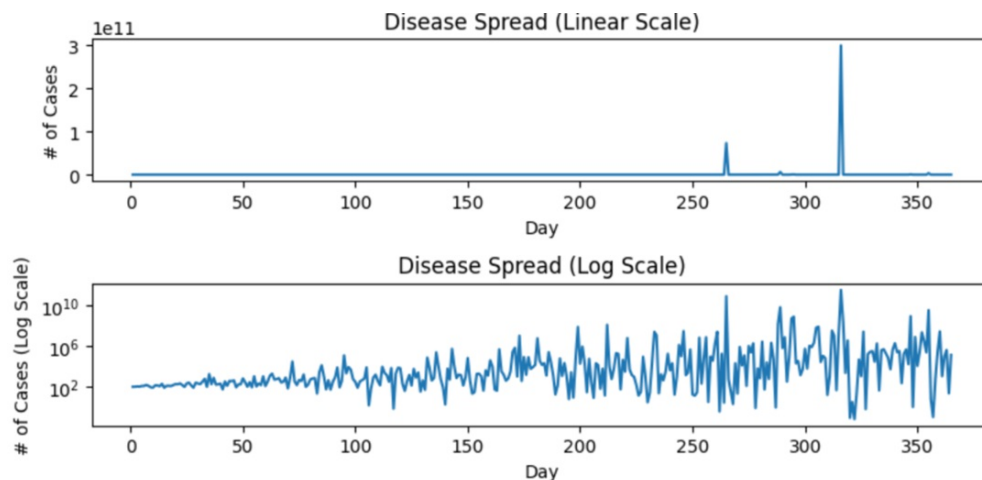
upper and lower bound of the y-axis are set at 0% and 100%, respectively, your stakeholders are unlikely to see fluctuations that may *actually* be meaningful and important to discover. In figure 7.21, the axis is set between 50% and 100%, which depicts the scope of fluctuations without over or under-representing changes over time.

**Figure 7.24** The same graduation rate metric appears to have little to no fluctuations when the y-axis limits are extended between 0% and 100%.



- **Using a dual y-axis** can be useful in some contexts, such as showing the relationship between two metrics on the same scale over time. More often than not, a dual y-axis will confuse your stakeholders and require far more time and effort to interpret than if you set up two separate charts. Your stakeholders will assume that the two metrics are on the same scale (e.g., inches, dollars, or percentage points) and may misinterpret the strength of correlations based on visual observation of fluctuations.
- **Non-linear/logarithmic scales** are often used to visualize data fluctuating over several orders of magnitude (e.g., population growth). Data presented on a logarithmic scale gives the appearance of being linear. While this can make it easier to depict the full range of values, logarithmic scales easily distort the perception of the variation that occurs. These should be used *sparingly* (only when it's impossible to interpret a linear scale), and should be accompanied with clear interpretations of the scale for your stakeholders.

**Figure 7.25** Logarithmic scales can help visualize metrics with exponential fluctuations in the data, but should be used with caution.



**Let's wrap up our case study for the chapter:**

As he wraps up the development of the Percentage of Sales Qualified Leads metric, Alex discovers several steps he needs to take to ensure the usability and effectiveness for his team.

Initially, Alex considers showing a cumulative view of the metric to highlight the company's overall growth of qualified leads. However, he realizes that the cumulative view masks a *seasonal* fluctuation, where more leads generated in the summer months are considered qualified. He decides to share the cumulative qualified leads as a single value to celebrate the team's work and an incremental view of the metric for each time period.

Next, Alex explores the best possible visualizations to show differences in Sales Qualified Leads from month to month. He uses a line graph because it allows him to easily visualize changes broken out by different segments valuable to the team (e.g., region, industry).

Beyond these final considerations, Alex knows the team could lose sight of other valuable processes if they focus solely on this new metric. To balance the indicator, he proposes a monthly counter metric to report with Sales Qualified Leads. The first is the *Lead Response Time*, which tracks the time it takes to contact a potential lead after being identified. This ensures that the team doesn't solely pursue qualified leads at the cost of their prompt response time for *all* leads.

Alex's diligent approach is clear in his presentation to the team, motivating them to adopt the new metric and counter metric to track the success of their efforts.

### 7.3.3 Activity

Let's continue with the SMART metric you designed in the previous section.

1. Conduct an online search of "website engagement metrics." What types of results do you get?
2. Conduct background research to understand other definitions of website engagement. Are there articles about how to measure this? How do they compare to your metric?
3. Identify some characteristics of metric examples you discovered in your search. Are they measured by day, week, or month? Are there any notable trends or benchmarks you can find that might be valuable to reference?
4. Are there potential counter metrics you should consider for your SMART metric?

Using the Python environment of your choice (terminal, Jupyter Notebook, etc.), import the dataset `website_engagement.csv`. The following columns are available for analysis:

- `website_engagement`: measured as a percentage of users active each week
- `session_duration`: the average duration of a website visit session in minutes
- `bounce_rate`: the percentage of visitors who navigate away from the site after viewing only one page
- `email_subscribers`: the cumulative total of email subscribers. There is an incremental version of the column, `new_email_subscribers`
- `social_media_followers`: the cumulative total of social media followers. There is an incremental version of the column, `new_social_media_followers`
- `avg_page_views_per_visit`: the average number of pages viewed per user session

- `total_items_purchased`: the total items purchased by users in the given week
- `total_sales`: the total dollar value of all sales in the given week

You must have `numpy`, `matplotlib`, and `pandas` installed to complete the following exercises.

1. Establish a thorough set of baseline information for website engagement. What trends (e.g., seasonal, distribution shape, longitudinal changes) exist in the website engagement metric that stakeholders should be aware of?
2. Are there relationships with other metrics valuable to note to stakeholders?
3. Identify one or more potential counter metrics in the dataset. How can focusing solely on website engagement negatively affect other business areas?
4. Create one or more visualizations for stakeholders to monitor `website_engagement`. Include information on any benchmarks they should be aware of, and watch the chosen counter metrics appropriately alongside the metric.
5. Using your baseline information, benchmarks, and valuable information you have gathered in this process, propose the first *achievable* goal for your stakeholders. How much should they strive to increase website engagement?

## 7.4 Summary

- **Metrics** are standardized quantitative measures tracked over time. They're frequently used to track progress, outcomes, and activities related to the organization's and its teams' performance.
- Metrics inform decision-making at multiple levels. **Performance metrics** are the broadest category an organization uses to understand its progress toward goals. Some examples include revenue, lifetime value of customers, and operational efficiency.
- **Organizational strategy metrics** track specific components of an organization's performance, such as product behaviors, customer sentiment, and market performance. Metrics in this category help

different teams at an organization understand the landscape they operate in and make appropriate data-informed decisions.

- **Accountability metrics** determine the effectiveness and productivity of individuals and teams. These metrics hold individuals *accountable* and guide performance reviews, bonuses, or training.
- The **SMART framework** is an essential guide for defining effective metrics. To ensure your organization can monitor and make strategic data-informed decisions, each metric at your organization should be **specific** (clearly defined), **measurable** (quantifiable), **achievable** (realistic and within reach), **relevant** (aligned with the organization's goals), and **time-bound** (aggregated at an appropriate timeframe).
- Understanding how to set goals for a metric involves gathering multiple sources of information to establish a **baseline** or a foundational understanding of the metric and its trends. These sources include:
  - Internal, **historical data** at the organization explored between segments, over time, and in relation to other business metrics
  - **Benchmarks**, or comparison metrics, often gathered from public sources such as government, industry surveys, or peer-reviewed literature
  - **Initial targets**, which allow you to test your ability to drive change in your metric and achieve goals
- How you **represent** your metric is as important as how you define it. There are many **pitfalls** in calculating and setting up a metric that can diminish the value an organization gains. Some examples include:
  - **Cumulative metrics** can create an illusion of continuous improvement despite declining performance. **Incremental views** of metrics are almost always a more appropriate way to provide a clear picture of trends in performance.
  - The **granularity**, or level of detail in your dataset, can significantly impact the interpretability and accountability of a metric. Metrics measured and grouped at a low grain (e.g., geography grouped by country) can obscure critical details, while a highly granular metric (e.g., zip code) can confuse your stakeholders with too much detail. Choosing an appropriate **grain** is a balancing act that requires understanding baseline information.
  - A **composite metric** combines two or more metrics to measure complex concepts. These tend to be challenging to interpret and act

on and obscure competing trends among the individual metrics.  
Limit efforts to develop these as much as possible.

- **Counter metrics** safeguard against the potential adverse effects of focusing on only one performance indicator. Each metric should include counter metrics to ensure your organization takes a balanced approach to strategic goal-setting.
- **Visual representation** is critical in communicating information about a metric and its performance. The **wrong chart type** can confuse your stakeholders and draw attention away from trends and differences. **Distorting the axes** on your charts can create the impression that there is a large trend where there is none, and vice versa. Your chart type and axes should be carefully selected and set to display your metric.

## 7.5 References

[1] T. Stobierski, "13 Financial Performance Measures Managers Should Monitor | HBS Online," *Harvard Business School Online*, May 05, 2020.

<https://online.hbs.edu/blog/post/financial-performance-measures>

[2] S. Cunningham, *Causal Inference: The Mixtape*. Yale University Press, 2021. Available: <https://www.jstor.org/stable/j.ctv1c29t27>

[3] J. Pearl, *Causality*. Cambridge: Cambridge University Press, 2009.

[4] N. Iliinsky, "Choosing a successful structure for your visualization."

Available: <https://newintelligence.ca/wp-content/uploads/2014/12/Choosing-a-successful-structure-for-your-visualization.pdf>

# 8 Navigating Sensitive and Protected Data

## This chapter covers

- The legal and regulatory landscape of sensitive data analysis
- Identifying and handling key types of protected information
- Applying techniques for anonymization to protect individuals in your datasets
- Analyzing sensitive data in a responsible and ethical manner

Let's talk about ethics. Whether intentional or not, data practitioners can produce unintended consequences for their users and the general population. A set of guidelines, ethical principles, and an understanding of the legal landscape will provide a framework for minimizing the likelihood of causing harm with your work and deliverables.

You may be thinking, "What harm could I possibly cause in my work? I work with spreadsheets, codes, and numbers. I write reports on operational metrics to improve my company's efficiency. How is my work potentially harmful to people?" Hear me out—there *are* countless tasks and specializations where your work has little to no implications outside of the daily operational functioning of your organization. Many roles (e.g., financial analysis, operational analysis) and tasks narrow in scope can be multiple degrees removed from the organization's relationship to its end users and target population. However, I invite you to consider that *far more areas of your work can impact and influence people in ways you may not know*.

Let's look at an example. In 2014, Amazon began developing a machine learning model to review job applicants' resumes and rank their qualifications to reduce time spent on the hiring process. The model was trained using a dataset of resumes received by the company over the previous decade. The training set showed an obvious bias toward male candidates, reflecting the overwhelming male pool of candidates in the tech industry. Thus, the model

actively penalized women—attendees of women's colleges and candidates who mentioned anything about "women" systematically received lower scores than those who didn't.

After several attempts to remove the underlying gender bias, Amazon abandoned the project. Due to their inability to control for all possible biases that could occur, the model couldn't be trusted to produce accurate and reliable rankings of candidate qualifications. Amazon ultimately discovered one of the issues with data analysis and data science—the quality and fairness of your output is only as good and unbiased as your input. Recruiters and hiring managers vary in their evaluation of candidate qualifications, and different companies, departments, and teams vary in the qualifications they prioritize in a candidate. In short, there is no singular measure of candidate quality.

Decisions made based on data or aided by data have tremendous implications for people's lives. Someone may be rejected from a job, evicted from an apartment, denied a loan, or identified as a suspect in crime using a facial detection model – all of which can have lifelong impacts on their quality of life. This is true regardless of the complexity of the data or model leveraged—a report you generate, a neural network, or a single bar graph are powerful tools that should be created, interpreted, and disseminated cautiously.

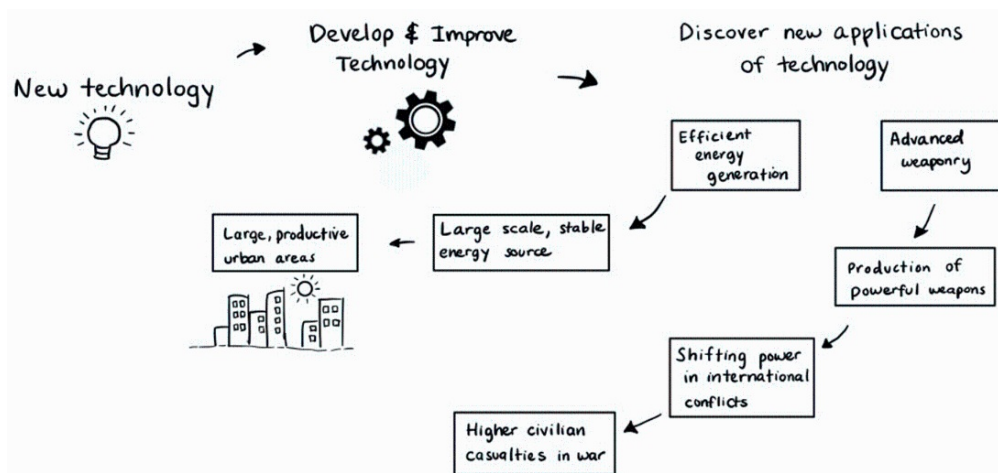
This chapter will cover the tools and knowledge you need to understand the scope of ethical considerations in an analyst's role. We'll include a history of legal precedent that guides research and analysis globally and recent laws in data privacy and machine learning. We will learn practical steps to protect your data, anonymize it, and responsibly leverage sensitive information in your analysis.

## **8.1 Consent in Research**

Discussions of ethics in science have existed for centuries, roughly coinciding with the Scientific Revolution that started in the 16<sup>th</sup> century. With rapid discoveries and advancement came questions about setting standards for scientific practice [1]. Several key questions were raised that impacted how we view science today:

- **How should research subjects be treated?** Scientists needed to develop guidelines for which investigative practices are permissible and which should be avoided at all costs. For example, vivisection (live dissection of animals) sparked debates about the ethical treatment of animals.
- **Who benefits from scientific knowledge and advancements?** Not everyone in society receives the benefits from science equally. Medical advancements and technology often take far longer to be available to those who need them in lower socioeconomic classes. Scientists had to reckon with their role and responsibility in making their discoveries (e.g., a cure to a disease) widely accessible.
- **How can scientists be proactive about the misuse of knowledge?** Science and technology can easily have negative, unforeseen consequences. Scientists continue to wrestle with the steps that need to be put in place to mitigate harm and their degree of responsibility when there are unintended use cases for findings.

**Figure 8.1** The outcomes of scientific discovery can lead society down a path of societal advancement or collective harm.



As analysts, we operate as the scientists of our organizations. We design and conduct experiments, collect data, and attempt to contribute to the organization's knowledge that enables it to progress toward its strategic goals. Many of our findings *can* impact our organization's stakeholders, users, and the broader society.

### 8.1.1 A Brief History Lesson

As part of the Nuremberg Trials held from 1945 to 1949, the "Doctors' Trials" dealt with war crimes related to medical experimentation conducted in Nazi Germany. The twenty-three doctors on trial were accused of conducting unethical and often deadly medical experiments on inmates in concentration camps. Many of the experiments lacked any scientific purpose and resulted in extreme suffering for people who had no opportunity to consent. A series of ethical principles for conducting research were developed out of the trials that guide scientific practice to this day.

#### Nuremberg Code

The *Nuremberg Code* is a set of ethical principles for research involving *human subjects* (people). The code actively guides scientific practice, even as methods and tools evolve. The ten principles of the code are summarized as follows:

1. The person involved in the research should be able to give voluntary consent. They should be provided with sufficient information to understand the research being done so that they can make an informed decision about whether or not to participate.
2. The purpose of the research should be to generate valuable results for the good of society.
3. Experiments should be designed carefully and have a strong justification for human research. Before conducting research with humans, animal studies should be performed in a laboratory where appropriate (e.g., mouse studies).
4. Research conducted should actively avoid physical or psychological harm to participants.
5. It should not be performed if an experiment or research study is expected to cause severe harm (e.g., death or disabling injury).
6. The expected risks from a study or experiment should never outweigh the expected benefits and importance to society.
7. Researchers should adequately plan study procedures to mitigate any possible risk of harm to participants.
8. Experiments should only be conducted by researchers and scientists with

appropriate qualifications in the field.

9. Participants should have the right to withdraw their participation in a study or experiment.
10. Researchers should be prepared to end an experiment if continuing may cause harm to the participants.

*Participants* generally refer to your organization's users, customers, or target population outside academic and clinical settings. Their satisfaction and positive experience are necessary for the organization to meet its goals, succeed, and grow. Much of the work we do as analysts can amount to a large-scale *field study* (conducted outside of a laboratory in a natural setting), which makes ethical guidelines such as the Nuremberg code essential to reference in our decisions. Some examples of its application include:

#### **Applications of Ethical Research Principles**

1. A non-profit is preparing to launch a new youth after-school program with its target population. Families are provided detailed information about the program's activities, time commitment, and expected outcomes to inform their participation decisions.
2. An analytics team at a financial institution carefully reviews results from statistical tests comparing loan repayment rates. If they discover differences in repayment rates by demographic groups or location, they make sure to thoroughly investigate other factors (e.g., an economic downturn or natural disaster) in their model that are known to correlate with individual characteristics among their user base. This diligence ensures that people are not systematically excluded from loan opportunities in the future.
3. A product analytics team collects tracking data about user visits and page views to their website. Each record includes metadata about the user, including their location and IP address, which can easily be used to identify individual users. To protect privacy, the team masks this personally identifiable information in their data warehouse.

While the Nuremberg Code was developed in the context of medical experimentation, its principles resonate powerfully in our data-driven work. Just as the principles in the Code emphasize dignity, autonomy, and well-

being, data analytics is responsible for upholding privacy, agency, and the rights and needs of those whose data we analyze.

### **8.1.2 Informed Consent**

At the beginning of most surveys, respondents are provided with summary information about what they can expect from their participation. These summaries usually include information about the survey length, the topics covered, and contact information for the research team. Some vary in length and the depth of information provided, but are all designed to give you the information you need to *inform your decision to participate*. In research settings, these are known as *informed consent documents*.

#### **Informed Consent Documents**

**Informed consent documents** in research are formal written agreements providing potential participants with the information they need to consent to participate in a study. This document's primary purpose is to protect participants by ensuring they have all the relevant information they need about the research, its potential risks and benefits (more in the next section), and their rights as participants.

Informed consent documents typically follow a standardized format, highlighting several key points that apply to most academic and non-academic research projects. This format is intended to give researchers the structure they need to ensure all critical information is shared with participants.

A typical document will include the following:

1. **Title and objectives:** begin the document by stating the name and objectives of the research project. The reason for pursuing this project should be apparent and easily understood by anyone reading the document.
2. **Description of the project:** provide a summary of the research project. Include precise details on all steps participants will take and the expected duration of their participation in the project. To respect their

time, be realistic about the expected time commitment and steps involved. Do not leave out key details or underestimate the time here!

3. **Potential risks:** explain any potential risks, comforts, or side effects that participants may experience from participating in the project. These should include risks to physical, emotional, and psychological well-being, regardless of how long any discomfort or negative impacts are expected to last. Be considerate of any sensitive topics you may be covering.
4. **Potential benefits:** explain any potential benefits that participants may receive from the research project, as well as potential benefits to others (e.g., society, the user base) that may be gained as a result of the research. As part of potential benefits, offering participants an opportunity to receive a copy of the project's results is often valuable.
5. **Alternative procedures:** where applicable, inform participants about any available alternative procedures, treatments, or interventions that can be advantageous to them and potentially pursued in lieu of your research project. This is primarily relevant in clinical and non-profit settings that provide a service to participants as part of the project.
6. **Confidentiality:** describe the measures your team, institution, or organization takes to keep their information secure. Be clear with participants about what personal information you will collect and store and any steps you will take to maintain their anonymity in your analysis. This should include details about your organization's data anonymization, storage, and security practices.
7. **Compensation:** include details on any compensation participants will receive for their involvement. This can be monetary, store credit, or otherwise, as appropriate to your organization.
8. **Voluntary participation and withdrawal:** ensure that participants know completing the research study is voluntary and that they can withdraw from the research at any time without penalty.

In addition to including each section, informed consent documentation should be well-tailored, readable, and **easily accessible** to the participants you are researching. It should be delivered *before* the research starts, and participants should have clear and reasonable options to opt out of participation without negative consequences. If the informed consent documentation is hidden in a lengthy and complicated agreement outlining the terms of service of using

your product or software, then participants will be unlikely to understand what they agree to.

## Obtaining Consent

In practice, many research projects outside clinical settings require only a summary to set expectations with your participants appropriately. What matters is the effort and intention you put into this process: be clear, stick to the procedures you outline, and honor your commitments to maintain the privacy and confidentiality of your participants. Your clear and honest communication goes a *long* way in establishing and maintaining the reputation of your team and organization.

**Figure 8.2 An informed consent document for a survey conducted by a product analytics team.**

User Experience Feedback on the Application Redesign

**Summary**  
Thank you for agreeing to participate in our survey. The goal of this project is to gather insights about the ease of use, look and feel, and areas of improvement for the application redesign that was recently launched. You will be asked to complete a survey containing 8 questions about each area of the application that was redesigned. The survey should take approximately 15-20 minutes to complete.

**Risks & Benefits**  
By participating in this survey, you are providing valuable feedback that will influence future iterations of the application design. We don't anticipate any risks associated with participating, but you are free to skip any questions that you are uncomfortable with or unable to answer. Your participation is voluntary, and you may also exit the survey at any time.

**Confidentiality**  
Your responses are anonymous unless you opt to provide your contact information at the end of the survey for follow-up or participation in other research. The results of this survey will be stored securely in our data warehouse and only accessed by the Product Analytics team. Only aggregated data without identifying information will be used in internal presentations and reports.

**Contact**  
If you have questions or concerns related to the survey, please contact Alex Johnson, the Lead Analyst on this project at [alex.johnson@company.com](mailto:alex.johnson@company.com).

If you agree to participate, please click "Next" to begin the survey.

Ideally, an informed consent process should be used where your users, stakeholders, or other target audiences are asked to participate in research. Consider the following example:

### Consideration for Participants

A marketing analytics team is researching trends in awareness, opinions, and experiences with recent advancements in artificial intelligence. The team hypothesizes that the responses they receive will vary widely, from excitement about its potential to fears about job loss and academic integrity. They know some participants may have had increased stress or changes associated with their employment, which can be tied to this topic. Given that they expect an emotionally charged set of responses, the team ensures that their informed consent document includes the following:

1. A clear description of the types of questions participants will be asked, such as their fears, concerns, and hopes about recent advancements in artificial intelligence.
2. A statement of risks and benefits outlines the potential discomfort or stress participants may experience when discussing their experiences with artificial intelligence.
3. An additional 5-10 minutes added to the estimated time to complete to account for potential long responses to free-text questions.

As we can see above, developing an appropriate informed consent document is a process of exercising *consideration and respect for your participants*. They may not leave your study with any lasting physical or emotional harm, but taking the time to understand their point of view and experiences goes a long way toward obtaining honest and high-quality results.

Next, consider the following example that outlines how the technology we use impacts the research we conduct and the data collected:

#### **Communicating Privacy and Security Practices**

A product research team is conducting customer interviews to test the usability of a new feature. The interviews will be conducted and recorded over a video call. Several dozen customers were emailed asking to participate, and each was given access to an informed consent document detailing the following:

1. Information about the procedures includes interacting with a sample application of the new features and answering questions while exploring its capabilities.

2. A clear, bolded statement informing customers that the video sessions would be recorded and transcribed. The team also includes a link to the security and privacy policy of the third-party services. If participants are uncomfortable being recorded, they are provided with an option to have the researcher take notes, with an understanding that the session will take 15 minutes longer.
3. The expected duration of the session (45 minutes) that is **strictly adhered to** by the team. Any unanswered or follow-up questions would require a separate consent process and scheduling at a later date.

While the informed consent practices we've covered were developed in clinical and academic settings, they easily apply to research leveraging new and emerging technologies. Your participants may be uncomfortable with their information being captured or stored by a third party, and their concerns should be treated as valid! In the above example, the third-party service being used may be a competitor to their company, may use their audio, video, or images to train machine learning models, or may suffer a data breach that puts sensitive information about them at risk. If participants have concerns, **respect them** and consider alternatives to capturing information that protects their privacy (we will discuss this in depth throughout the chapter).

Finally, consider the following example of a non-profit and the procedures it needs to put in place to protect program participants:

### **Protecting Anonymity**

A non-profit that aims to help youth in the foster care system is developing a new program for adolescents that aims to assist with job placement, financial literacy, and obtaining scholarships for higher education. The research team develops an informed consent document that outlines the expected procedures for the adolescent participants to read through before agreeing to participate. The document includes the following:

1. Language tailored to the age and education level of the prospective participants (adolescents ages 15-19). All technical terminology was replaced with clear, accessible descriptions.
2. Detailed information about the steps that would be taken to protect the personal information of participants. The non-profit research team

consulted with previous program participants to understand their concerns about their status as a foster youth being disclosed to teachers, potential employers, and colleges without their consent or potentially being revealed at a later date in adulthood. Per those concerns, all participants were given anonymized IDs in their database, and all identifying information was disconnected from the research program data. These steps were communicated to new prospective participants.

3. List all of the research questions that would be asked as part of the study and enumerate potential benefits *to them*. Previous program participants had told the research team that they were often sought out to participate in studies with little to no benefit.

Each of these examples shows how obtaining informed consent is more than just filling in the blanks on a structured document; it's a necessary step to understand the needs, concerns, and perspectives of anyone whose data you collect. Even if you are working with people's data that has already been collected (e.g., data about your users in a warehouse), it's essential to start by asking yourself, *would these users consent to this project if we asked them?* This is an important question regarding data collection in the current legal landscape, which we'll discuss in the next section.

### 8.1.3 Activity

You are a health data analyst at a large biotechnology company producing wearable devices that monitor vital signs (e.g., heart rate) and physical activity (steps, workouts, etc.). The company has recently embarked on a project to design features that detect early signs of disease using data collected from the wearable devices.

This project's first study will investigate potential predictors of hypertension (high blood pressure) and heart disease. The potential predictors are primarily activity data and user characteristics already collected by the company. The study will involve recruiting several hundred users who were diagnosed with hypertension or heart disease during the time they have been using the wearable device. Your team will be collecting detailed information about the diagnosis for the study.

1. Propose a written informed consent protocol to be shared with users interested in participating in the study.
  - a. What potential risks might users incur by participating in this study? Is there any potential for harm that you or your users need to know?
  - b. What benefits to participating users and society might you share in this document?
2. Consider that your team will collect sensitive information about participating users' health statuses. What steps might your company need to put in place to protect users from the risk associated with that information being disclosed? Write down your answer and consider it as you read the following sections of this chapter.
3. Suppose a user participating in the study requests to withdraw their consent to have their data included in your analysis. What should you do? What might you and your team need to put in place to honor that request?

## 8.2 The Current Legal Landscape

*Massive* amounts of data are collected and readily available on the internet. This data can often be purchased or otherwise obtained by individuals and organizations that we may not even be aware of—and for purposes we may not consent to. In the United States, research within a clinical or academic setting that works with human subjects must comply with a strict set of federal guidelines [2]. By comparison, data collection and storage practices outside of these settings aren't consistently regulated.

Many parts of the world are grappling with potential legislation to protect people's privacy and the ability to *consent* to data collection and usage to enable privacy in the era of big data. This landscape is changing *rapidly*; everything discussed in the following section reflects current and upcoming legislation as of December 2023.

### 8.2.1 Data Protection Regulations

Data protection regulations are sets of laws designed to protect the rights of individuals to have their data securely stored, processed, handled, and deleted

when requested. Organizations are expected to adhere to obligations set forth in regulations that are often similar to those in clinical and academic research.

Individual laws vary by region, as we will discuss in this section. Many current regulations typically cover the following topics:

- **Definitions of personal and sensitive data:** Each set of regulations specifies what constitutes personal data. This typically includes *any information that can be used to identify an individual*, such as their name, address, phone number, email, and IP address.
- **Expectations for handling data:** Organizations must follow guidelines for storing, processing, and using personal and sensitive data. These include explicitly limiting access to personal data, retaining data *only* as long as necessary, and limiting the data collected for a specific purpose.
- **Establishing the rights of individuals:** Data protection regulations enumerate individual rights concerning the collection and use of their personal data. These rights include being informed about how their data is collected and used, the right to have their data deleted, and the right to deny consent to its use in a specific manner.

Many organizations, especially those that operate out of multiple locations, fall under these emerging laws and are expected to determine appropriate practices for complying with these laws. In larger companies and organizations, analytics teams will likely partner with IT, security, and legal teams to enact practices that comply with these laws. In smaller organizations, you and your team may have more responsibility in ensuring that the people whose data you collect can exercise their rights under these laws. We'll discuss some of the most impactful regulations today and how they can impact the work of an analyst.

## GDPR

The General Data Protection Regulation (GDPR) in the European Union governs the collection, retention, and usage of personal data and individual rights related to how that data is used [3]. The regulation went into effect in May of 2018, becoming the first comprehensive set of laws of its kind. It's become a model for subsequent laws adopted in a dozen countries, including

a direct replica adopted in the United Kingdom in 2022.

If your organization collects data from offices, clients, or users in the European Union, they're likely required to comply with the GDPR. The specific responsibility of your team will vary, but analysts can expect that the GDPR will impact their work in one or more of the following ways:

- **Minimizing data collection:** When starting a project, an analyst may need to document the questions you're asking and the specific fields of data you need to answer those questions. This documentation will ensure you communicate the project's requirements and *only* collect that information without asking for extraneous personal information about your users or participants.
- **Anonymizing data:** To protect the privacy of your users, customers, or target population, your organization may need to mask all personal data queried or used for analysis (e.g., replacing full names with random strings of letters). This may require you to take extra steps (e.g., requesting access from your IT and security teams) to access personal information like email addresses, phone numbers, or locations for a specific project.
- **Retaining or deleting data:** Your organization may have a *data retention policy* requiring that all data is anonymized or deleted after a set period (e.g., one year). This might limit your time to complete an analysis project after data is collected or how far back in time you can access records with personal information.
- **Securely accessing data:** To minimize the likelihood of data breaches, you may be required to adhere to policies at your organization guiding how to access and share data. For example, you may not be permitted to download personal data onto your local machine or share it over email. It's strongly recommended that you review your organization's policies and available training on data security.
- **Contributing subject matter expertise:** As your organization seeks to remain compliant with evolving legislation, you and your team may be asked to collaborate with IT, security, and legal functions at your company to understand better the data you collect and use. Reviewing the key pieces of legislation that we'll discuss in the following sections will enable you to streamline the process of complying with these laws.

You'll also build valuable expertise in data regulations, benefitting your long-term career.

The interpretation, enforcement, and application of GDPR and similar laws are evolving worldwide. Analysts can expect to share the responsibility for complying with these laws and protecting the personal information of the users whose data you collect.

## CCPA

As of December 2023, no single federal law in the United States governs data protection across sectors. As we'll cover later in this chapter, specific sector laws (e.g., protected health information) require compliance with national laws. Data protection regulations are primarily being passed at the state level.

The California Consumer Privacy Act (CCPA) took effect in January 2020, defining rights to data privacy for California residents. It bears many similarities to the GDPR—the definitions of personal information and established rights to privacy are nearly identical. However, there are some notable differences in how the law was defined that may impact organizations operating in the state:

- CCPA applies to data collected about people who are *legal residents of California*. By comparison, GDPR applies more broadly to anyone in the EU regardless of their citizenship status. In practice, this may mean that fewer residents of California can exercise their right to have data modified, deleted, or restricted in use.
- While GDPR applies to *any* business or entity that processes data about EU residents, CCPA only applies to a subset of for-profit businesses that meet one or more of the following criteria:
  - \$25 million or more in annual gross revenue
  - Buys, sells, receives, or shares personal data from at least 100,000 consumers or households
  - Makes 50% or more of its annual gross revenue from selling consumers' personal information
- CCPA includes individuals and *households* in its definition of personal data. This might consist of data fields connecting multiple individuals to

the same family (e.g., a unique family ID for an address).

In general, if your organization complies with GDPR, it can easily comply with the CCPA. The CCPA is considered less restrictive than the GDPR. Still, you and your organization should be aware that this law is actively being modified (an amendment was passed later in 2020, adding additional privacy restrictions).

### Calculating Organizational Metrics

A business analytics team at a multinational company is creating a dashboard to track new metrics. The company has offices in the United States, EU, Canada, and Great Britain. Many of their customers are in the same regions, making their data and information regulated under the GDPR and CCPA.

The team is looking to calculate three new metrics calculated *monthly*:

1. Percentage growth in users
2. Count of users by region
3. Percentage of active users

Calculating these metrics is easy – each is a count or proportion aggregated by month and measured over time using a line or bar graph. However, the team is aware of some new steps taken to comply with the GDPR and CCPA that may impact the accuracy of calculations:

- Twelve months after users in the EU are deactivated, all identifying information about them is deleted (e.g., name, email address, specific location), leaving only null values for those fields in the `users` table. In addition, their records in the `user_activities` table are deleted entirely. This means the *percentage of active users* metric will *not* be accurate for EU users before the most recent rolling twelve-month period.
- If a user in the EU or a resident of California requests it, the company has to delete all information about them, including their anonymized record, in the `users` table. This is in accordance with the "right to be forgotten." This means that all metrics will be inaccurate for EU and United States users; this will be especially noticeable for the *count of users by region* metric, which will be artificially lower than their actual

values.

The business analytics team reaches out to the company's data engineering team. They recommend capturing a *snapshot* of the anonymized and aggregated data from the previous day in new, special tables in the data warehouse. The tables can then be used to create the dashboard tracking each of the three new metrics.

The case study above is a common example of how analytics, data science, and data engineering teams have had to adjust their practices in recent years to maintain their users' privacy in accordance with new laws. Later in this chapter, we'll cover strategies for striking this balance, including capturing anonymized and aggregated data.

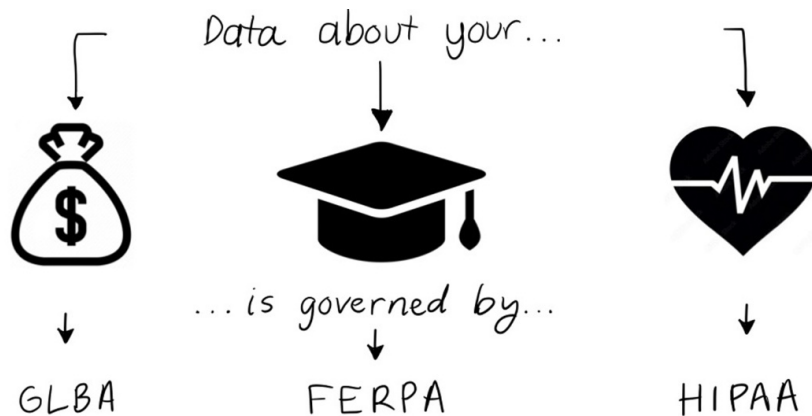
## Sector-specific laws

While GDPR and CCPA apply to a broad range of organizations, many sectors have their own regulations due to the nature of the data they handle. The mishandling of data in health, financial, education, and other fields can have specific implications for individuals whose data is breached or sold without consent. Here are some examples of sectors and laws that govern how organizations can store, access, and leverage people's data:

- The **healthcare** industry in the United States is governed by the Health Insurance Portability and Accountability Act (HIPAA), which strictly regulates protected health information (PHI) collected by healthcare providers, health insurance companies, and more [4]. This data is considered *strictly* confidential and often requires the person's explicit consent to share, access, or leverage for specific purposes.
- In **education**, the Family Educational Rights and Privacy Act (FERPA) provides students and parents the right to access and correct student records. It restricts the sharing and usage of personally identifiable information unless the student or parent consents.
- In **financial services**, several laws govern the privacy and security of financial information in the United States. For example, the Gramm-Leach-Bliley Act (GLBA) requires financial institutions to explain their data-sharing practices to customers and provides them with the right to

opt-out [5].

**Figure 8.3** Many types of sensitive information are governed by sector-specific laws



We'll discuss best practices for handling protected information (e.g., PII and PHI) later in this chapter.

## 8.2.2 Bias in Automated Systems

An *algorithm* is a sequence of computations used to process data, perform calculations, or solve a problem. In your work, this may be as simple as developing a set of repeated steps to clean a dataset, summarize it, and generate charts. *Artificial intelligence* is an advanced field that seeks to engineer systems that perform complex tasks typically requiring human intelligence to complete. These AI systems leverage complex algorithms and large datasets to identify underlying patterns in your data and attempt to replicate decisions based on those patterns continuing to exist in new data.

Algorithms of any complexity (e.g., a set of predefined rules or a neural network) can be used to create **automated systems** that perform tasks or make decisions that would otherwise need to be done by humans. An increasing number of these systems are designed to be tools that assist in our day-to-day work, automating mundane and repetitive tasks and increasing our overall efficiency.

Given an automated system's use of objective rules or large datasets, we can usually rely on them to make better, more accurate decisions than humans...

right?

Not necessarily! The example we outlined on page 1 is one of many tools shown to produce *biased outcomes* [6]. A simple rules-based algorithm can often have unintended consequences if the system's designers are unaware of how the rules are applied to people. Similarly, machine learning algorithms leveraged in AI systems are *trained* on patterns in the underlying datasets. Human biases are often an ever-present pattern in our data for these purposes.

Let's take a look at the following example:

### **Automating Labeling of Support Tickets**

A data science and analytics team at a large consumer goods company is looking to automate the categorization process for support tickets. Every day, the customer support team receives hundreds of support tickets on various topics that are manually triaged and categorized before a team member is assigned to each ticket.

Tickets have two categories assigned: a priority level (low, medium, and high) and a ticket type (Product issues, delivery, billing, inquiries, and suggestions). The customer support team then uses this information to determine how to prioritize them against other tickets in the queue and how quickly the problem needs to be resolved.

The team decides to build two prototypes to test the accuracy of automatically categorizing tickets—a rules-based model and a supervised classification model. They leverage key insights from an analysis performed by the Business Intelligence team to inform their approach:

- Most *Product Issues*, *Delivery*, and *Billing* tickets are also labeled "High" priority. Most *Inquiries* are labeled as "Medium" priority, and *Suggestions* are labeled as "Low" priority.
- An estimated 60% of tickets can be categorized by searching for one of 20 keywords in the text. However, the accuracy of this approach was not thoroughly tested.
- The volume of *Delivery* and *Billing* tickets is highly correlated with the number of sales, being more common during the holiday season from

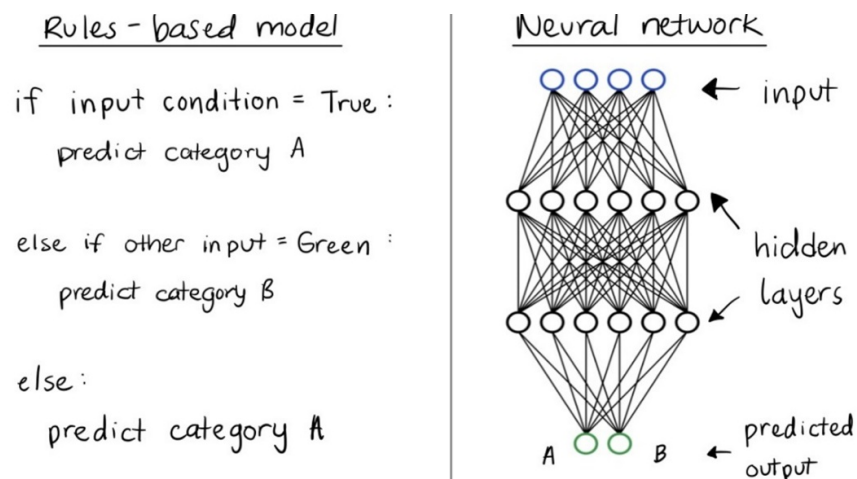
November to December (monthly seasonality). *Inquiries* and *Suggestions* are more commonly submitted on weekends (daily seasonality).

The team sets out to answer the following questions:

1. What percentage of tickets can they accurately categorize with the 20 keywords? How far can they increase the list to 30, 40, or 50 keywords?
2. Is a set of rules (e.g., keywords and explicit characteristics) sufficient to categorize most tickets, thereby reducing the volume of work for the support team?
3. Does machine learning (the proposed classification model) add value to the project? Is it more accurate than rules, or does it enable the team to perform less manual work?

Whether or not they actively leverage machine learning, analysts often participate in developing algorithms to perform calculations, automate tasks, and streamline the efforts of their stakeholders.

**Figure 8.4** An automated system for categorizing data can employ anything from simple rules to sophisticated machine learning algorithms.



Automated systems, regardless of complexity, are frequently used to make decisions about people (e.g., users, customers, and the general population). In the seemingly innocuous example above, the information provided by customers is used to **rank** the urgency of their request and **categorize** the submission. It's quite straightforward and can save a lot of time for the

support team. However, there may still be unintended long-term consequences on customers depending on their submission. Let's look at a few possible scenarios that the team may need to consider in the long term:

### Evaluating the New Automated System

With the addition of 30 more keywords, the rules-based algorithm classifies 80% of tickets into one of the five categories referenced above, while the remaining 20% are categorized as "Other." Those 20% still need to be manually triaged by the support team, which often adds up to 5 days before a team member is assigned to resolve the issue. After several months of manual triage, the support team discovers the following:

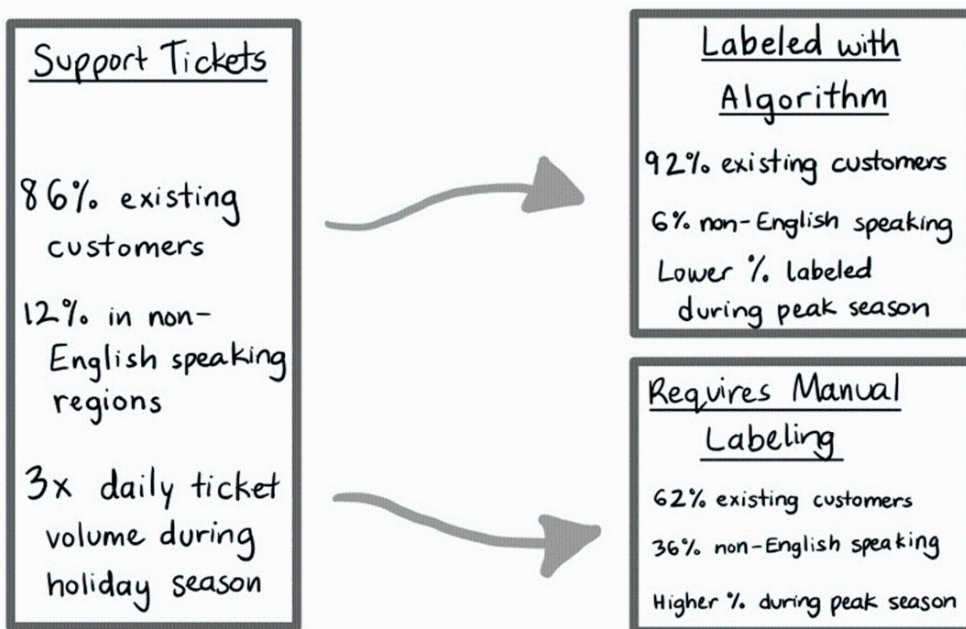
- The 20% of tickets classified as "Other" appear to disproportionately represent new customers with issues with their first-ever purchase. There are several hypotheses as to why; regardless, the team is aware that this can negatively impact the perceptions and loyalty of new customers.
- Two-thirds of the "Other" tickets are from customers based in non-English speaking countries. The entire support team and ticketing process is in English, potentially creating a barrier for international customers.

The data science and analytics team also trains a machine learning model to classify tickets into categories and urgency levels using inputs such as trends highlighted in the Business Intelligence team's analysis (e.g., week, month). At first glance, the model performs better than the rules-based approach. With further investigation, the team identifies the following phenomenon:

- The model accuracy isn't consistent throughout the calendar year. Specifically, *Delivery* and *Billing Inquiries* are more likely to be misclassified as *Product Issues* and ranked as "Medium" urgency during November and December. This can drastically reduce the quality of service received during the most profitable time of year.
- Similar to the rules-based approach, the model has lower accuracy among new customers and those in non-English speaking countries.

**Figure 8.5** The categorization system developed by the data science and analytics team can

streamline the work of the customer support team. Still, it may have an unanticipated negative impact on certain groups of customers.



The example scenario above has potential benefits for the stakeholder team, reducing repetitive manual work and freeing up resources to focus on other high-value tasks. However, the benefit that may be passed to customers is *not* even, and can even reduce the quality of service that certain sub-groups receive (e.g., new customers, international customers). This is an example of two forms of *bias* in an automated system—**disproportionate accuracy rates** and **disproportionate outcomes**.

## Bias in High-Stakes Decisions

Take the example we just covered and replace the following:

- Automating **job candidate qualification ranking and categorization** instead of **support ticket classification**
- **Candidates applying for a job** instead of **customers**
- **Open jobs** instead of **support tickets**
- **Candidate quality level** (e.g., not qualified, somewhat qualified, highly qualified) instead of **ticket urgency**
- **Candidate outcome** (interview, reject) instead of **ticket category**

Suddenly, the impact of inaccuracies in an automated system is not so harmless. We know human decisions are biased, and developing a model based on previous human decisions only codifies that bias. The issue with many automated systems is that they're sold as unbiased tools to aid or perform decision-making. Models attempting to replace human decisions often make selections that disproportionately favor certain subgroups while producing less accurate results for these or other groups.

A model that is less accurate or ranks certain groups of candidates as less qualified than others isn't just a meaningless side effect of automation; *it's potentially impacting the income, career trajectories, net worth, and other long-term outcomes for real people*. It doesn't matter if you're using a set of rules or a deep learning model—the potential for harm remains.

Analysts can have *a lot* of power throughout our careers. When stakeholders take action based on our recommendations, we can shape an organization's strategy and the people it serves. If we *don't* thoroughly audit our work, we risk downstream effects that we cannot otherwise anticipate.

## **Auditing Automated Systems in Human Resources for Bias**

In July 2023, a new law regulating automated systems in employment decisions took effect in New York City. Specifically, any company that uses AI or any other automated decision tool in hiring and promotion decisions must disclose the tool's use to candidates and employees, as well as participate in annual third-party audits evaluating the tools for bias [7]. These laws apply to businesses with employees or job candidates in New York City and may cover a wide variety of systems integrated into HR processes. Some examples of tools this may apply to include [8]:

- Third-party software that an HR team pays to rank the hundreds of candidates who apply to each job on their website using a machine learning algorithm.
- A complex set of rules developed internally that uses keyword matching to identify candidates whose skills match a given job.
- A vendor that generates a quality score for candidates and automatically rejects anyone below a certain threshold.

Bias in these automated systems is defined as differences in the scoring and selection of candidates by protected classes (e.g., race, ethnicity, sex, and intersecting groups). In essence, you are asking, "Is there any difference in the predicted ranking, rating, or selection rate for a job/promotion by race, ethnicity, sex, or any combination of the above?" The New York City government website provides specific metrics to calculate as part of the audit, ensuring standardized estimates of bias across different companies.

## **What's Next**

Mandated audits and regulation of automated systems are a new area of legislation. Given this law is the first of its kind, there will likely be iterations on the provisions and enforcement of the law in the coming years. It will also likely serve as a model for other laws being considered in and outside of the United States.

As an analyst, you have an opportunity to build real expertise in an emerging field by monitoring the changing landscape in regulation related to data collection, usage, and the impact of automated decisions. To keep your knowledge up to date, I suggest taking one or more of the following steps to synthesize information:

- Subscribe to data analytics and data science newsletters, many of which can be discovered with a quick Google search. Try looking at sources of information on Medium and Kaggle if you're unsure where to start.
- Set up alerts with news outlets that discuss technology (Tech Crunch, Forbes, Wall Street Journal, etc.) on topics related to data, AI, and machine learning ethics and regulations.
- Set up alerts on Google Scholar for similar keywords as news outlets—analytics, data science, AI, and machine learning ethics and regulations. These sources will be far more in-depth and technical than news outlets, and you'll often learn of key topics before they reach general audience news sources.

## **8.2.3 Activity**

The study conducted in your previous activity has been a huge success! You

and your team have discovered clear predictors of several common diseases that are all easily available from the activity data you collect. As such, your company will be launching a new AI-driven feature that alerts users to potential health issues.

1. To maximize the study's benefit to society, your company is partnering with clinical researchers to publish their findings. Your team plans to anonymize all user data before sharing it with external research partners. Which GDPR principle is the company adhering to with their research partnership?
  - a. Data minimization
  - b. Integrity and confidentiality
  - c. Accuracy
  - d. Transparency
2. Your company recently had a surge of new users in California. As such, what data policies will they need to have in place to comply with CCPA?
3. Six months after launch, your team discovers that the AI-driven feature is less accurate in predicting health issues for users over 65. What are the legal and ethical implications of the bias present in this algorithm?

## 8.3 Analyzing Sensitive Data

Now that you have a comprehensive understanding of the ethical foundations of research, data analysis, and experimentation, as well as knowledge of an emerging legal field, let's discuss some pragmatic skills you can leverage to work with sensitive data effectively and ethically.

*Sensitive data* refers to information about individuals or organizations that can have negative consequences if shared without consent or authorization. This type of data typically requires elevated security practices to protect against misuse and comply with regulations. There are *many* categories of sensitive data; some common ones include:

- **Personally identifiable information (PII)** can be used to identify a specific individual or sensitive characteristics about them. This category includes social security numbers, phone numbers, email addresses, and

demographic characteristics.

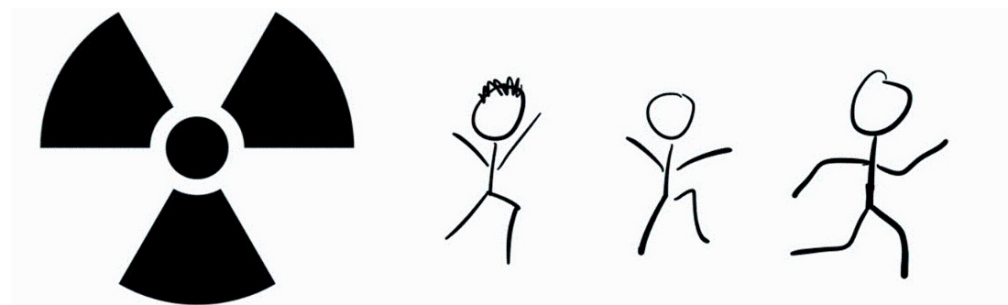
- **Protected health information (PHI)** includes all health status, medical history, or medical payment information that can be tied back to an individual.
- **Financial information** includes information about an individual's financial status, bank accounts, credit card numbers, and financial transactions.
- **Authentication data** includes usernames, passwords, and any other login credentials. Breaches of authentication data can easily lead to the leaking all other forms of sensitive data.

Throughout this section, we will use examples of PII and PHI given their prevalence in the work of an analyst. The best practices in data security, anonymizing data, and putting guardrails in place can apply to any form of sensitive data where there is a risk of identifying a person based on available information.

### 8.3.1 Data Minimization

Perhaps the easiest way to handle protected information is to *not* handle it at all—in many situations, this is the best-recommended approach to your work. Before determining the best approach to managing protected information, it's valuable to determine if you need it at all.

**Figure 8.6** When all else fails, treat protected information as radioactive.



**Data minimization** (as covered in section 8.2.1) is the practice of limiting the collection, storage, and use of personal data to *only what is directly relevant for a specific purpose*. In essence, this practice encourages you to think strategically about the purpose of any information you collect and *only*

store it for as long as needed.

There are several aspects to practicing data minimization as an analyst:

- **Limit the data you collect** to only what is necessary. Research and analysis have a long-standing tradition of collecting as much information as possible about users and participants "just in case" it might yield a statistically significant result when nothing else does. If you cannot justify the reason for collecting a piece of information, don't.
- **Limit the length of time in which you store data.** Many organizations collect data about their users, customers, and population for seemingly indefinite periods of time. However, due to its lack of relevance, most analyses don't end up using data from more than a few years prior to the current date. Work with your team to set a data retention policy and delete, scrub, or limit access to data after a certain amount of time has passed.
- **Limit access to raw data as much as possible.** This is a data minimization principle that analysts often have *a lot* of opportunity to enact. Where possible, do not share raw data with stakeholders that don't otherwise have permission to view it (e.g., in your data warehouse). **Do not** share sensitive data over insecure channels (e.g., a CSV of customer data attached to an email). Instead, use secure file-sharing methods at your organization (e.g., Google Drive, Proton).

Let's look at an example:

#### **Minimizing Data Collected in e-Commerce**

The analytics team at a large e-commerce company is currently rearchitecting its strategy for collecting and analyzing website user behavior. Previously, the data warehouse had available details about every user who visited their website – first name, last name, phone number, IP address, and specific location. Their security team recently flagged their data collection and retention practices as a risk for expanding into Europe and complying with the GDPR.

To better protect information about their users, the team opts to recommend the following strategy:

- Replace all personal information (e.g., name, email, phone number) with an anonymized user ID from the data warehouse.
- Remove the specific location fields (latitude and longitude) and replace them with more general location information (city, state, country).
- Set a retention policy for all website user behavior data: granular data (e.g., one row per user per visit) would only be available for the previous 12 months. After this point, the data was aggregated to a count of visits by user, page, and day and saved in a new table.
- Restrict access to the granular website user behavior data to *only* approved analytics team members who have completed security policies training.

These intentional steps enabled the analytics team to effectively comply with GDPR and mitigate the risk of data security breaches that might otherwise put customer data at risk.

Often, out of habit, we collect sensitive information in excess of what is necessary. Participants are used to filling out surveys asking for their age, race, gender, location, and more, so they rarely balk at these questions. However, having this information freely available and indefinitely stored at your organization can produce long-term risk for data breaches and compliance issues, generating the mistrust of your users. Minimizing the data you collect, retrieve, and include in your final analyses can help you avoid many situations where you must take extra measures to protect personal information.

### **8.3.2 Anonymizing and Pseudonymizing Data**

Now that you've eliminated every case where you don't need protected data, we can skip this section...right?

If only that were the case! While avoiding using protected information is preferable, there are some legitimate scenarios where access to that information is vital for completing your work effectively. In these situations, having the skills necessary to process that information safely and anonymize it where possible is valuable.

## When is using protected information necessary?

A data analyst may need to leverage protected information in one of the following scenarios:

- Conducting an analysis of health outcomes using protected health information (PHI) to understand patient diagnoses, treatment efficacies, and quality of care.
- Developing tailored and personalized marketing strategies often includes leveraging a combination of user demographic characteristics and behaviors to understand what products, goods, and services they might be interested in purchasing.
- When combining data across multiple third-party sources, sometimes the only unique key available is a name or email address to connect information.

In these cases where using protected information is unavoidable, countless tools and techniques are available to you for masking that information in your analysis. The first technique, **anonymization**, refers to the explicit removal of personal information so that an individual cannot be identified. Data that is *fully* and *irreversibly* anonymized is no longer considered personal data, meaning it falls outside the scope of current data protection regulations.

**Pseudonymization** refers to the deidentification of data using pseudonyms, fake identifiers, or other information that limits (but does not prevent) the identification of an individual. Pseudonymization is *reversible*; with the correct information, you can tie an anonymized ID back to the protected information about that record stored in a separate table. This approach is preferable when there's a need to anonymize data for analytics but store sensitive and protected information for specific use cases later on. For example, an e-commerce company likely needs to store customer contact, address, and credit card information in a database to enable them to make recurring purchases without re-entering that information with every new transaction. However, the information isn't necessary for analysis, so it won't be made available in the data warehouse for analysis.

## Anonymizing Protected Information in Python

Often, you can use the same tools and approaches in Python to anonymize or pseudonymize your data. The difference is in whether you save the protected information in a separate table, with the anonymized IDs available to be joined to the datasets used for analysis when necessary. Let's look at an example of a dataset at an online retailer. The following transactions table has information about each item purchased by a customer and a wealth of metadata about that customer.

```
import pandas as pd    #A

transactions = pd.read_csv("transactions.csv")    #B
transactions.head()
```

**Figure 8.7** This table contains numerous fields with protected information about the customer.

	transaction_id	name	email	phone	address	purchased_item	price
0	t-834207	Keith White	keith.white@example.com	001-490-356-3466x219	PSC 7000, Box 0663, APO AP 61947	Cool tablet	627
1	t-194507	Nicole Weaver	nicole.weaver@example.com	(581)642-5743	Unit 3524 Box 6820, DPO AE 31360	Retro mobile phone	584
2	t-310414	Joseph Perez	joseph.perez@example.com	297.447.3481	50270 Graham Alley, Lake Brittany, NH 15851	Yet another laptop	405
3	t-947522	Cheryl Salinas	cheryl.salinas@example.com	945.248.4798x675	51407 Jones Drive Suite 478, Wattsstad, CA 53037	Cool tablet	537
4	t-156647	Monica Foster	monica.foster@example.com	(738)712-8058x1408	81902 Roberts Route Suite 280, Port Andrewmout...	Retro mobile phone	943

This specific dataset has every piece of information necessary to contact a customer—names, emails, phone numbers, and addresses (*note: to protect the privacy of real people, this information has been generated with the faker library in Python*). In all likelihood, none of the sensitive and protected fields will be directly necessary for analysis. It may be possible to *derive* valuable information from some fields (e.g., standardized location information from the address), but it's unlikely that we will use the sensitive data directly. Let's drop the name and phone number fields from the dataset outright.

```
transactions.drop(["name", "phone"], axis=1, inplace=True)    #A
transactions.head()
```

**Figure 8.8** When sensitive data isn't necessary for analysis, you can drop the fields from your dataset.

	transaction_id	email	address	purchased_item	price
0	t-834207	keith.white@example.com	PSC 7000, Box 0663, APO AP 61947	Cool tablet	627
1	t-194507	nicole.weaver@example.com	Unit 3524 Box 6820, DPO AE 31360	Retro mobile phone	584
2	t-310414	joseph.perez@example.com	50270 Graham Alley, Lake Brittany, NH 15851	Yet another laptop	405
3	t-947522	cheryl.salinas@example.com	51407 Jones Drive Suite 478, Wattsstad, CA 53037	Cool tablet	537
4	t-156647	monica.foster@example.com	81902 Roberts Route Suite 280, Port Andrewmout...	Retro mobile phone	943

Next, let's extract non-protected standardized geographic information from the address field. The `usaddress` Python library offers the ability to extract and normalize addresses in the United States. We'll use this library for the transactions dataset since all of the addresses are in the United States; when working with international addresses, you can use more comprehensive libraries, such as `libpostal`, that can parse addresses worldwide. You'll first have to install the `usaddress` library at the command line or in your notebook as follows:

```
pip install usaddress    #A
```

To parse addresses in the entire dataset, we will define a function that extracts your choice's specific standardized geographic information. Keeping data minimalist principles in mind, we'll keep only the city, state, and zip code (assuming we know we will need them for our analysis!). Additional standardized information is available in the dictionary we get from this library, which will be dropped.

```
import usaddress    #A

def parse_addresses(address):    #B
    try:    #C
        parsed = usaddress.parse(address)
        parsed_dict = {item[1]: item[0] for item in parsed}
        city = parsed_dict.get("PlaceName", "")    #D
        state = parsed_dict.get("StateName", "")
        zip_code = parsed_dict.get("ZipCode", "")
        return city, state, zip_code
    except:
        return " ", " "    #E

transactions["city"], transactions["state"], transactions["zip_co
    *transactions["address"].apply(parse_addresses)
)    #F

transactions.head()    #G
```

**Figure 8.9 Standardized geographic information is easily extracted from the address column.**

	transaction_id	email	purchased_item	price	city	state	zip_code
0	t-834207	keith.white@example.com	Cool tablet	627	APO	AP	61947
1	t-194507	nicole.weaver@example.com	Retro mobile phone	584	DPO	AE	31360
2	t-310414	joseph.perez@example.com	Yet another laptop	405	Brittany	NH	15851
3	t-947522	cheryl.salinas@example.com	Cool tablet	537	Wattsstad	CA	53037
4	t-156647	monica.foster@example.com	Retro mobile phone	943	Andrewmouth	WV	50204

Finally, let's drop the original address column, effectively **anonymizing** your users' addresses. The resulting dataset is more valuable for analysis *and* less risky to your customers in the event of a data breach.

```
transactions.drop(["name", "phone"], axis=1, inplace=True)      #A
```

In most cases, standardized city and state information will be sufficient to understand widespread geographic trends among your users and customers. In cases where you need more granular neighborhood-level trends, the zip code is an appropriate unit leveraged in the majority of social research. Many publicly available data sources (e.g., the U.S. Census) have comprehensive information about residents of each geographic granularity that you can use to compare to your users.

Next, let's move on to the final column containing PII – the email address. This may prove to be more complicated, as email addresses can be a decent option to join data between sources within an organization where another unique identifier doesn't exist. For example, if you need to combine data from your application's database and two separate vendors, it's unlikely they share the same anonymized ID to link records. In cases like this, where the following criteria are met about a sensitive data field...

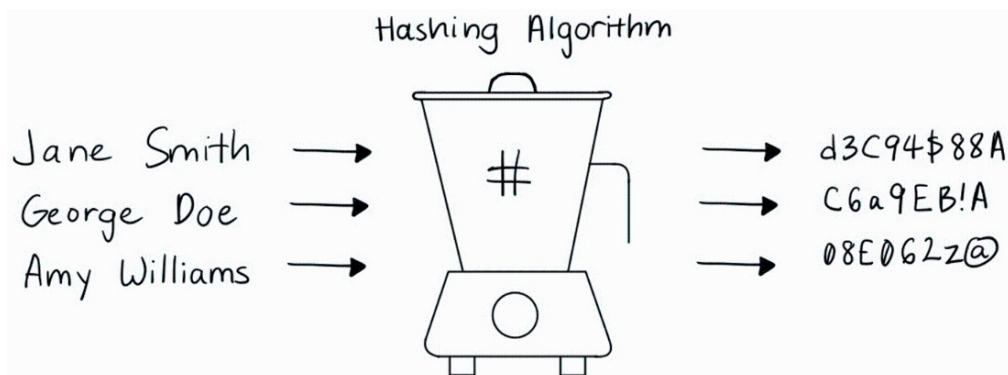
- It's needed in your data warehouse, such as to join data between sources
- The data itself is *not* used for analysis (or you can extract non-sensitive components) and can be hidden from analysts and stakeholders
- Analysts mainly need to identify unique individuals with a primary key,

such as a random string of numbers

...you likely need to **pseudonymize** this field and retain it in a separate, secure table not directly available for analysis. This is often done using a process called **hashing** the data. *Hashing* is a process that transforms your data into a fixed-size value (e.g., 256 bits) using an algorithm known as a *hash function*. The resulting *hash code* has the following characteristics:

- It's **deterministic**, meaning each unique value has a different hash code, and repeated values will have the same one. Thus, it *can* be used as an anonymized primary key in place of sensitive data.
- Slight differences in the original data result in very different hash codes; similarity between hash codes is *not* an indicator of similarity in the underlying data.
- At present, it's impossible to reverse-engineer the original data only from its hash code. However, this may not always be the case, so I strongly recommend keeping up to date on information security best practices.

**Figure 8.10** Hash functions are like a smoothie blender, taking the original data and scrambling it so you cannot undo it and retrieve your original ingredients.



Hashing values in a pandas dataframe is easily performed using the standard `hashlib` Python library, which contains several hashing algorithms to choose from when anonymizing your data. Let's use the **SHA-256 (Secure Hash Algorithm 256-bit)**: this is relatively quick to compute and widely used for encrypting data in a warehouse and sharing and transporting between individuals, organizations, and over networks.

```

import hashlib      #A

def hash_email(email):    #B
    return hashlib.sha256(email.encode()).hexdigest()

transactions["email_hash"] = transactions["email"].apply(hash_ema

transactions.head()    #D

```

**Figure 8.11** The resulting hash is a long sequence of random letters and numbers.

	transaction_id	email	purchased_item	price	city	state	zip_code	email_hash
0	t-834207	keith.white@example.com	Cool tablet	627	APO	AP	61947	e9a78df1d0bfbc8def1f0c0...
1	t-194507	nicole.weaver@example.com	Retro mobile phone	584	DPO	AE	31360	b1fd301be283df2e485c4408...
2	t-310414	joseph.perez@example.com	Yet another laptop	405	Brittany	NH	15851	702b9181b695c10f21dbcffc...
3	t-947522	cheryl.salinas@example.com	Cool tablet	537	Wattsstad	CA	53037	299d0bea719deb861a41d56e...
4	t-156647	monica.foster@example.com	Retro mobile phone	943	Andrewmouth	WV	50204	28776c6cd2896da2240017eb...

Finally, let's save the email and email\_hash as a separate dataframe and drop the email from the transactions table.

```

emails = transactions[["email", "email_hash"]]    #A
transactions.drop(["email"], axis=1, inplace=True)    #B
transactions.head()    #C

```

**Figure 8.12** The final transactions table can identify customers without sensitive or protected data.

	transaction_id	purchased_item	price	city	state	zip_code	email_hash
0	t-834207	Cool tablet	627	APO	AP	61947	e9a78df1d0bfbc8def1f0c0...
1	t-194507	Retro mobile phone	584	DPO	AE	31360	b1fd301be283df2e485c4408...
2	t-310414	Yet another laptop	405	Brittany	NH	15851	702b9181b695c10f21dbcffc...
3	t-947522	Cool tablet	537	Wattsstad	CA	53037	299d0bea719deb861a41d56e...
4	t-156647	Retro mobile phone	943	Andrewmouth	WV	50204	28776c6cd2896da2240017eb...

The new emails table retains the original PII in a separate location from the transactions data used for analytics. In situations where it's necessary to tie new data sources to your existing data (e.g., a new vendor that only contains the email address as a unique identifier), the following steps can be taken:

1. Join the new data source to the emails table

2. Add the `email_hash` to the new data source
3. Drop the original `email` column from the new data source, instead using the `email_hash` as a primary key for your users or customers

Where possible, this is a task that should be done in collaboration with the data engineering or IT team that manages your data warehouse. These teams will typically be able to restrict or permit access to PII, such as the `emails` table, and create highly curated sets of *views* for analysts to use in their work.

At a smaller organization (e.g., a non-profit with limited resources), you may not have these teams to collaborate with. If you don't have the resources available to manage sensitive and protected information at scale, you can use the steps in this section to anonymize, pseudonymize, and restrict access to PII as much as possible. Data sources such as the `emails` table can be stored and password protected in the most secure manner that your organization offers and *only* accessed when necessary by members of your analytics team.

### 8.3.3 Preventing Deanonymization

Believe it or not, the removal of PII is often not enough to prevent the identification of individuals. In 2006, Netflix released a dataset of 100 million movie ratings from 500,000 users as part of a competition to improve their movie recommendation system. The dataset had been stripped of all PII (e.g., names and email addresses) and released publicly to anyone interested in the competition.

A research team from the University of Texas developed a deanonymization algorithm using contextual information about the users from their reviews (e.g., the dates, ratings, and text of the reviews) to cross-reference with publicly available ratings on the movie database IMDb [9]. With the information in the original dataset, they were able to identify startling amounts of personal information—names, contact information, sexual orientation, and more.

This incident is one of many situations where anonymization can be insufficient to protect the privacy of individuals, who can easily be identified based on context. **Contextual information** refers to combinations of non-

sensitive data that provide sufficient clues as to an individual's identity. It's often shockingly easy to do so, given the volume of publicly available data available today. Take the following examples:

- A dataset of payments for a local bus in a small city can be cross-referenced with social media check-ins. Since many commuters use the same transit card, individual travel patterns can be identified and potentially cross-referenced with social media check-ins.
- Combinations of demographic characteristics like age, gender, race, ethnicity, and zip code can easily be used to narrow down individual identities. A 2000 paper estimated that 87% of the United States population can be uniquely identified using just the zip code, gender, and date of birth [10].
- In 2006, AOL publicly released a dataset of 20 million search queries performed by 657,000 users. While all user information was removed from the dataset, PII was still present in many searches. As such, the New York Times was able to identify one specific user from search queries alone [11].

Analysts are often responsible for the formatting, structure, and dissemination of data that, while anonymized, can easily be reverse-engineered to determine who a person is. Depending on your specialization, this can occur *very frequently*, putting your users, customers, or coworkers in a difficult position if they're identified. Let's take a look at the following case study:

#### **Protecting Anonymity in Employee Engagement Reports**

An HR analytics team at a non-profit with 1,200 employees has just completed its annual employee engagement survey. Every year, employees are given a 30-question survey about their workload, morale, and perceptions of management. The results are analyzed by the team and compiled into a comprehensive report, showing question scores by department, office, employee tenure, gender, and more.

Given that many questions were of a sensitive nature, the team enacted the following **guardrails** to maintain the anonymity of responses to encourage honest responses.

- The team configured the survey to only collect **anonymous responses**. Nobody would be able to connect a response to an individual employee.
- Individual responses were *never* reported on. Instead of sharing comments, the team synthesized overarching themes and sentiments in the qualitative data. This ensured that an employee couldn't be identified by their writing style or the topics they discussed.
- Since the report often included multiple breakouts (e.g., average satisfaction by department and age), the team set a **minimum aggregation** level of 5. If a sub-group had under five employees, the report did not show results.
- For sections of the report where the number of employees was consistently too low, the team avoided displaying multiple breakouts (e.g., average satisfaction only shown by department for small teams).

In addition, the team publicized this list of guardrails to the company to ensure employees felt confident in their anonymity when responding to the survey. This helped increase the response rate from 67% the previous year to 81%, enabling more comprehensive insights into the data.

**Guardrails** are a set of predefined criteria put in place to ensure the accuracy of your analyses. They're designed to maintain anonymity, protect individuals, and prevent biased or inaccurate interpretations that can result from your work in their absence. Teams will often develop these as a general checklist for their work. A typical set of guardrails might look like the following:

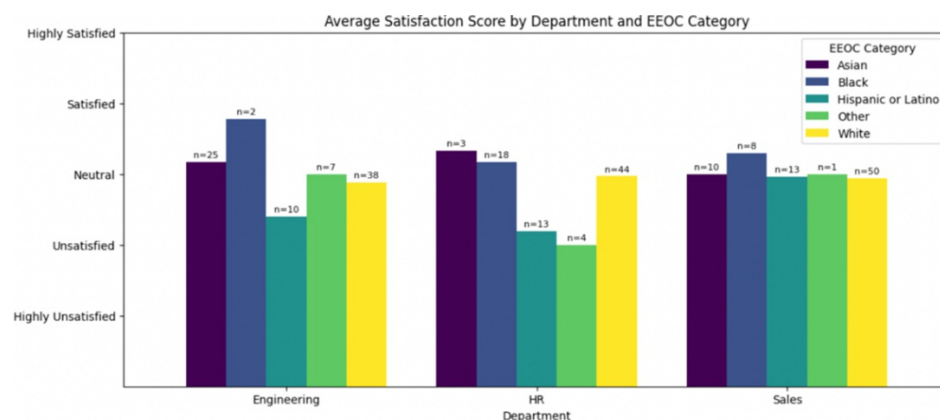
1. All datasets must have a minimum sample size chosen in accordance with the statistical test being used.
2. All groups, subgroups, and breakouts in reports must have a minimum sample size of 5 to show aggregate data. Any group with less than five individuals should be excluded or combined with another group.
3. Reports and other deliverables should not include access to the raw data unless absolutely necessary to prevent deanonymization. If necessary, the data should be shared securely and only with those who need it.
4. Any report or deliverable that includes breakdowns and interpretations based on demographic characteristics (e.g., gender, race, ethnicity, age, disability status, veteran status) should be approved by a peer-review

process with the team. The peer-review group should discuss the potential for harmful interpretations, inaccurate representations, and perpetuating stereotypes.

5. Every step in the analytics lifecycle should be well documented, from data collection to processing and results interpretation. Another team member should be able to reproduce an analysis based on the available documentation alone.

In practice, a chart like the one shown in figure 8.13 would violate these guardrails. The *n* value label shows the total number of employees in each department and EEOC group. Several subgroups have less than five employees represented (and one subgroup has only a single person!). It would be extraordinarily easy to identify which employees these are, creating a difficult situation should they be expressing negative sentiments or concerns about their work.

**Figure 8.13 An employee engagement survey with multiple breakouts can accidentally deanonymize individuals in groups with few members.**



An analytics team's efforts to prevent deanonymization will generally have multiple benefits for the quality of work produced. Many inaccurate and harmful interpretations can arise from beliefs that a response belongs to a specific person or that an aggregate value for a small sample size represents their broader group. Remembering when your data is about *people* and seeking to leverage their data responsibly will set you and your team up for success in your work.

### 8.3.4 Activity

Your biotechnology company has seen a surge in users leveraging the new AI-generated alert feature. To ensure the feature produces accurate data over time, users are asked to confirm whether or not they received a diagnosis for the health issue they were alerted to. This data is used to periodically re-train the underlying machine learning model powering the feature.

1. What type of data is being used to train the model powering the AI-generated health alert feature?
2. The company's product team is discussing whether they should erase all unnecessary identifying information about users or replace it with coded references while maintaining a separate list of actual data. What data protection methods are they considering? How would you weigh the pros and cons of each method?
3. Propose a set of guardrails for the analytics team to use when reporting the results of clinical trials. Include at least one guideline on data minimization, anonymization vs. pseudonymization practices, and minimum aggregations for reports and peer-reviewed journal submissions.

## 8.4 Summary

- The **Nuremberg Code** was established in the aftermath of harmful experiments in WWII. It set guidelines for ethical human subjects research, emphasizing voluntary consent and the rights of participants.
- The **informed consent process** in research ensures that participants have sufficient information about the purpose, procedures, and potential risks of a research project or study to make *informed decisions* about participating.
- The **General Data Protection Regulation (GDPR)** in the European Union provides individuals control over how their personal data is stored and used. It requires that companies have guardrails in place to protect the privacy and security of people's data and obtain consent to store and use personal data for specific purposes.
- The **California Consumer Privacy Act (CCPA)** is a regulation that offers data protection rights similar to the GDPR, specifically for California residents. It allows consumers to know about and opt out of the usage and sale of their personal data.

- Automated decision tools using rules-based or machine learning algorithms can potentially produce biased output that disproportionately impacts subgroups of your users, customers, or the general population. New York City recently passed a law requiring these decision tools in hiring to be **audited for bias**. This is a new category of law in the data world and will likely develop quickly in the coming years.
- **Sensitive and protected data analysis** refers to the analysis of personal information governed by law or that may risk producing harm if it were released without consent. This type of analysis requires special care to maintain the privacy and confidentiality of people and organizations.
- **Protected information** is data that, if exposed, could produce harm for the individual.
  - **Personally identifiable information (PII)** refers to individual characteristics such as names, addresses, phone numbers, and social security numbers. This type of information is governed by regulations such as GDPR and CCPA.
  - **Protected health information (PHI)** refers to medical records, transactions, diagnoses, and any other information that, if breached, could risk disclosing confidential medical conditions.
- **Data minimization** is a key principle in analyzing sensitive data, recommending that you *only* collect the data you need for a specific project and retain it for as long as necessary. This is intended to reduce the risk of data breaches and misuse of sensitive information.
- **Anonymization** is the process of completely removing personally identifiable information (PII) from your data to prevent the identification of individuals.
- **Pseudonymization**, in contrast, is the process of removing PII from specific datasets and replacing it with a pseudonym or key. The data is then retained elsewhere in case it's necessary for other purposes.
- Analysts have a responsibility to minimize the likelihood of **deanonymization** of data. Since anonymized datasets can be easily re-associated with individuals, a set of **guardrails** for your team should be established to protect from unintended consequences associated with re-identification.

## 8.5 References

- [1] B. E. Rollin, *Science and Ethics*. Cambridge University Press, 2006.
- [2] Office for Human Research Protections, "45 CFR 46," *HHS.gov*, Feb. 16, 2016. <https://www.hhs.gov/ohrp/regulations-and-policy/regulations/45-cfr-46/index.html>
- [3] GDPR, "General Data Protection Regulation (GDPR)," *General Data Protection Regulation (GDPR)*, 2018. <https://gdpr-info.eu/>
- [4] Centers for Disease Control and Prevention, "Health insurance portability and accountability act of 1996 (HIPAA)," *Centers for Disease Control and Prevention*, 2022. <https://www.cdc.gov/phlp/publications/topic/hipaa.html>
- [5] T. Klosowski, "The State of Consumer Data Privacy Laws in the US (And Why It Matters)," *Wirecutter: Reviews for the Real World*, Sep. 06, 2021. <https://www.nytimes.com/wirecutter/blog/state-of-privacy-laws-in-us/>
- [6] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine Bias," *ProPublica*, May 23, 2016. <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>
- [7] "Automated Employment Decision Tools (AEDT) | DCWP," [www.nyc.gov](https://www.nyc.gov). <https://www.nyc.gov/site/dca/about/automated-employment-decision-tools.page>
- [8] J. Kestenbaum, "NYC's New AI Bias Law Broadly Impacts Hiring and Requires Audits," *Bloomberg Law*, Jul. 5, 2023. <https://news.bloomberglaw.com/us-law-week/nycs-new-ai-bias-law-broadly-impacts-hiring-and-requires-audits>
- [9] A. Narayanan and V. Shmatikov, "How To Break Anonymity of the Netflix Prize Dataset," *arXiv:cs/0610105*, Nov. 2007, Available: <https://arxiv.org/abs/cs/0610105>
- [10] L. Sweeney, "Simple Demographics Often Identify People Uniquely," *Health (San Francisco)*, vol. 671, Jan. 2000, doi: <https://doi.org/10.1184/r1/6625769.v1>.

[11] M. Barbaro and T. Zeller, Jr., "A Face Is Exposed for AOL Searcher No. 4417749," *The New York Times*, Aug. 09, 2006. Available: <https://www.nytimes.com/2006/08/09/technology/09aol.html>

# 9 The World of Statistical Modeling

## This chapter covers

- The purpose and application of common classes of statistical models
- Evaluating the information available when fitting a model
- Fitting a statistical model to a dataset and iterating on it to improve performance
- Developing appropriate explanatory and predictive deliverables based on the results of a statistical model

Three data analysts walk into a bar.

The first says, "I bet I can figure out the top five traits that predict whether a person orders beer, wine, or spirits. We can use that model to better plan the inventory."

The second retorts, "Well, give me info on the next 100 patrons, and I'll use your model to forecast all their orders in advance."

The third smirks, "Why wait? Give me real-time data, and I'll use your model to predict a drink right as they're about to order it. Now, that will *really* impress the patrons!"

What's the difference between the type of model that each analyst is proposing and the other two? Are all three approaches valuable? Can they all genuinely use the same statistical model to predict the same phenomenon, with a different approach and desired output? And are the analysts *actually* proposing successively better alternatives?

The answer to the first three questions is yes—a small handful of **statistical models** can be leveraged for a wide variety of purposes and deliverables. The same model can often inform decisions *and* generate predictions in the most appropriate format. None of the examples above is inherently *better*—instead, there are different strategic approaches to deriving value from each

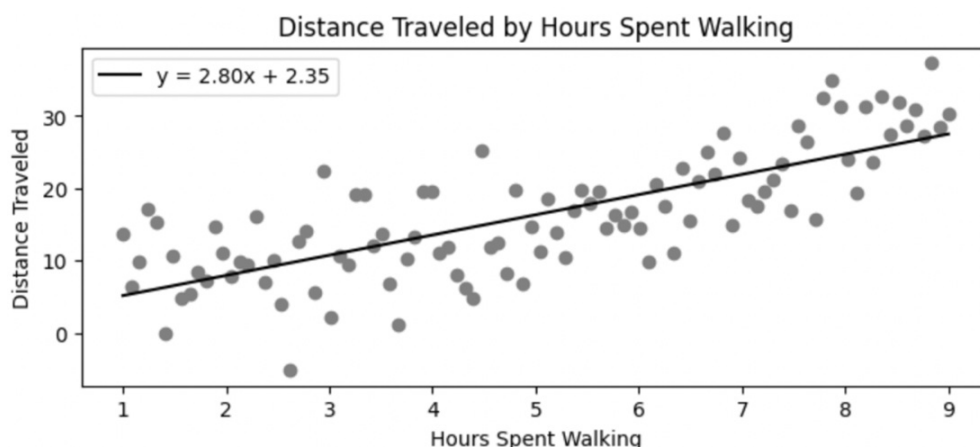
approach.

**Statistical modeling** is the process of mathematically representing relationships between one or more independent variables (X-variables) and one or more dependent variables (y-variables) in a dataset. Models are developed to describe and quantify your data's underlying structure and patterns. Through statistical modeling, we aim to understand the nature of complex processes that we would otherwise struggle to represent with descriptive (e.g., mean, median), univariate (e.g., t-test, ANOVA), and bivariate (e.g., correlation) statistics.

Analysts can use statistical modeling for many purposes, such as the following:

- **Making inferences** about the statistical significance of one or more independent variables and their shared relationship with a dependent variable.
- **Predicting outcomes for future values** that are within the same range as your dataset, but aren't actually part of the original sample used to fit the model
- Deriving **insights** to enable **data-informed decisions** at your organization. Statistical model results can increase confidence in strategic plans and actions.

**Figure 9.1** A linear regression with two variables (a predictor and outcome) is one of the simplest forms of statistical models to fit to your data.



In this chapter, we'll cover the process of fitting a statistical model, evaluating its performance, fine-tuning it, and producing the best deliverable(s) to meet the needs of your stakeholders. Each step will be covered using a single example dataset you may be familiar with from chapter 4 (remember rat complaints in New York City?), building upon the specific skills you need to create models for different purposes.

We'll demonstrate all of our concepts in this chapter with a *linear regression* model—however, you can follow along with examples using nearly any other class of model. If you need a refresher on the underlying math and logic of a linear regression model, you can return to chapter 4 for a review of the topic before we begin.

## 9.1 The Many Faces of Statistical Modeling

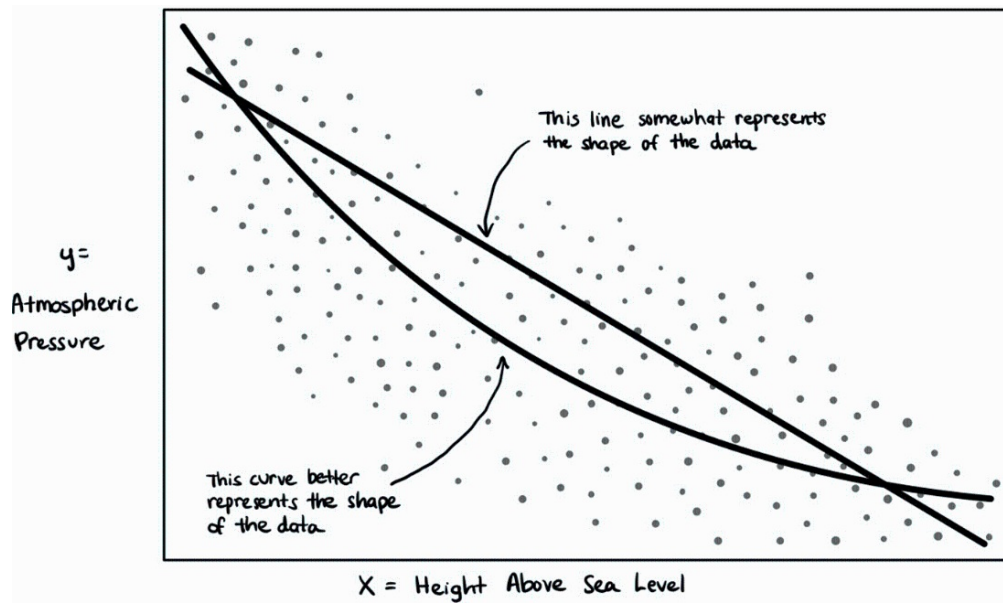
"All models are wrong, but some are useful."

- George Box, British Statistician

A **model** is a mathematical formula (equation) used to represent underlying patterns in your data. The modeling process aims to find a formula that appropriately *fits* or *represents* the relationship between one or more predictors and an outcome, allowing you to make inferences and predictions about the phenomena that the outcome represents.

In most cases, a good fit model can *approximate* patterns in your data but will rarely allow you to predict future values with 100% accuracy. The real world is complex, and most of what we measure is intertwined with so many processes that we can never completely figure out what predicts or causes an outcome. A *useful* or *good enough model* represents the general shape and trend of your data.

**Figure 9.2** Many mathematical formulas can represent the shape of relationships in your data, such as a linear formula (line) or a polynomial (curve) formula.



In general, the process of statistical modeling follows these steps:

1. Select an appropriate model for the problem you are attempting to solve (e.g., regression vs. classification). We will discuss this in depth later in this section.
2. Select a *formula* (e.g., a line) to fit the data. The line fits the data using an *objective function* to identify the model coefficients (e.g.,  $\beta_1$  and  $\beta_2$  in regression models). We'll discuss objective functions later in this section.
3. Evaluate the model's performance and ability to describe the patterns in the data. Optimize the model by adjusting the formula and model parameters.

This is shown in figure 9.2—a *linear* and *polynomial* model are fit to the same dataset and compared. From visual observation, the two variables appear to have a negative exponential relationship—as the height above sea level increases, the atmospheric pressure exponentially decreases. Thus, the polynomial model better represents the relationship between these two variables.

### 9.1.1 Classes of Statistical Models

When looking to make predictions, identify intricate underlying patterns, and

estimate future events, analysts have *countless* statistical models to choose from. Many models have variations in their formula and algorithms to help solve specific problems, which can be an intricate and complex web of options to navigate.

For the purpose of this chapter, we'll group many of the common models into **classes**, organized by the problem they're trying to solve. We'll define a **class** as a group of formulas and algorithms that share common characteristics and are designed to solve a *specific type of real-world problem*.

For example, *regression* models are a class designed to predict a *continuous outcome* based on one or more input variables. *Linear regression* assumes that a line can represent the relationship between your predictors and outcome; other formulas are available for data that cannot fit a line. By contrast, *logistic regression* uses a similar formula to solve a *classification problem*—the model is similar, but is used to predict a categorical instead of continuous outcome.

The problem you are looking to solve will typically guide your selection of a class of models. Some of these problems may look like the following:

- Predicting the dollar value spent during a store visit based on several properties about the customer, timing, and store location.
- Identifying factors contributing to whether or not a student passes their final exam (a binary, categorical outcome).
- Forecasting the business's sales for the coming nine months, accounting for seasonal changes and economic predictions.

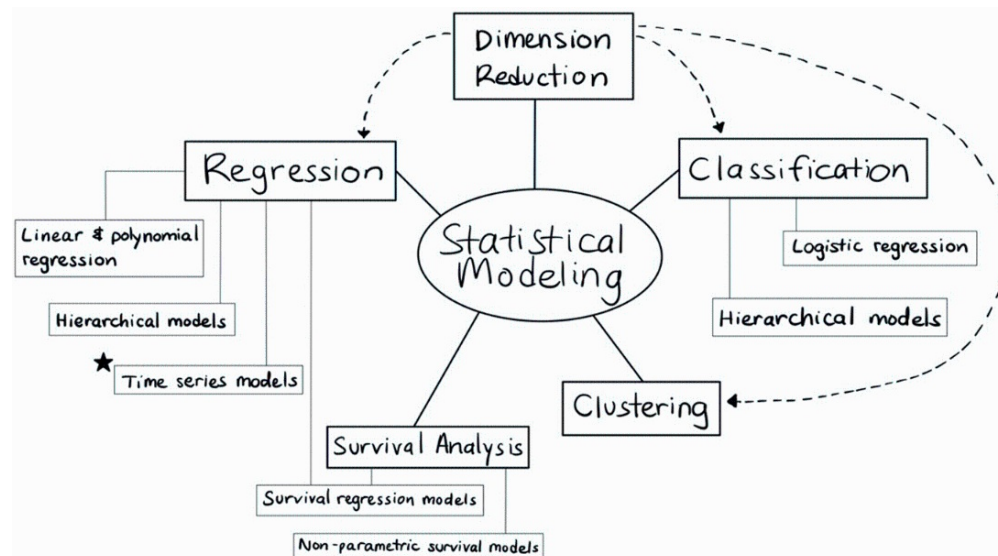
We'll discuss the following classes of models in this section:

- **Regression** models predicting continuous outcomes
- **Classification** models predicting categorical outcomes
- **Clustering** models, identifying underlying patterns and groups in data
- **Dimension reduction** models, identifying shared variance among predictors to simplify model input
- **Survival analysis** models, which predict the time until an event occurs

Figure 9.3 shows a map of each class listed, some examples, and their

connection. We'll cover each of these and their relationships later in this section.

**Figure 9.3** Map of common classes of statistical models organized by the problem they solve.



While most of your problems will likely fall into one of these classes, they are by no means the only approaches in this field. Within your specific domain, you'll probably discover that most of the problems you solve will belong to a limited subset of the available classes of models. Each can (and often does) comprise multiple books and cannot be given due attention in a single chapter. Thus, we'll focus on the strategies and approaches that can be applied to most of these categories while only using regression and classification models as examples.

## Regression

**Regression models** are fitted to datasets with a *continuous outcome*. The predictors in the model can be either continuous or categorical. If you're asked questions such as the following, a regression model may be a good choice for producing a deliverable:

- Can we predict the return on investment (ROI) for different marketing channels such as social media, search engines, and email?
- What factors contribute to employee tenure at our company? Can we use

data such as performance, satisfaction surveys, and training attendance to explain the tenure of employees?

- Can we predict the expected lifetime of machinery in a manufacturing company based on its usage, maintenance data, and other potential factors?

Figure 9.2 shows an example with a continuous variable for the outcome (atmospheric pressure), so you would want to fit a model *regressing* the atmospheric pressure (y-variable) onto the height above sea level (predictor/X-variable).

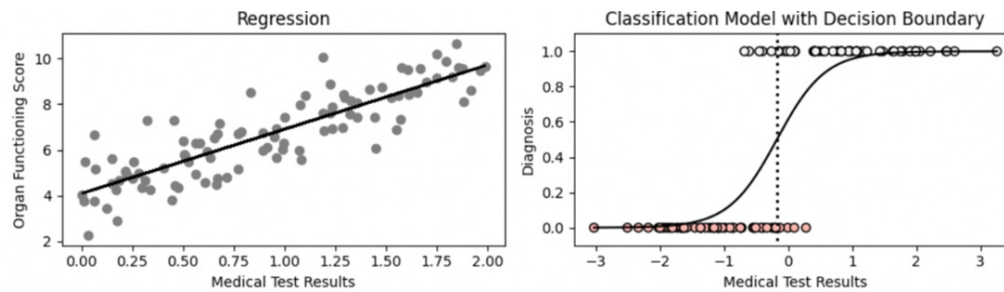
## Classification

Instead of a continuous outcome, **classification models** attempt to understand underlying patterns in your data that determine which *category* or *discrete outcome* a data point belongs to. A classification model can help you answer questions such as the following:

- Can we predict whether or not a patient has a specific type of cancer using a set of biological, psychological, and social factors?
- Can we use the same factors as the employee tenure regression model question to predict whether or not an employee will resign in the next quarter?
- Can we predict whether a crop will be healthy, at risk, or likely to be diseased based on weather conditions, soil quality, and crop type?

The underlying mathematical approach to regression and classification can be quite similar – predictors can be continuous or categorical, and many formulas can be used for both approaches (e.g., linear regression vs. logistic regression, random forest regression vs. random forest classifier). Figure 9.4 shows a simple example of how a regression and classification model is fitted to a dataset—instead of a line, a curve is fit to the model, and a *decision boundary* is estimated as a threshold for where data points are classified into each category.

**Figure 9.4 Regression and classification models are among the most commonly used in statistics.**



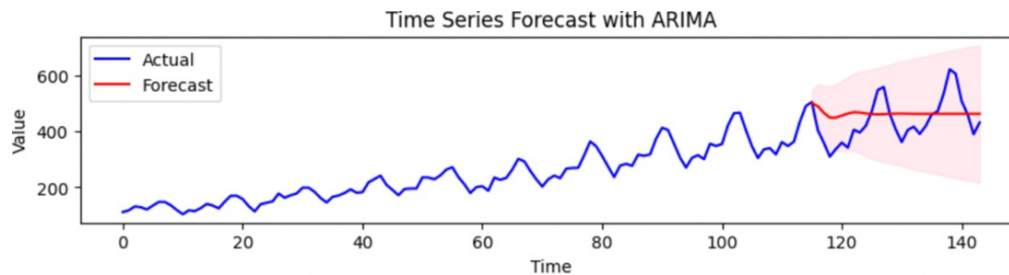
Often, the questions you receive from stakeholders can be answered by either a regression or classification model. The second question in this and the previous section can likely use the same data sources as inputs, and the resulting models will primarily differ in how the outcome variable is structured. Work with your stakeholders to determine whether a continuous or categorical outcome better meets their needs.

## Time Series

**Time series models** are a special type of model that uses historical data captured *over time* to predict future values. These models account for factors in previous data (e.g., seasonality, overall trends, recent trends) to provide an estimate for future time periods. For example, businesses often attempt to forecast their revenue over the coming weeks, months, and quarters. Questions that you answer using a time series model often look like the following:

- What is a city's forecasted daily electricity consumption over the next seven days?
- Can we predict weekly sales for a retail store while considering the upcoming holidays?
- Is it possible to predict hospitalizations for a specific disease over the coming season? What factors enable us to predict these values accurately?

**Figure 9.5** Time series models look for patterns in data captured over time to predict future values.



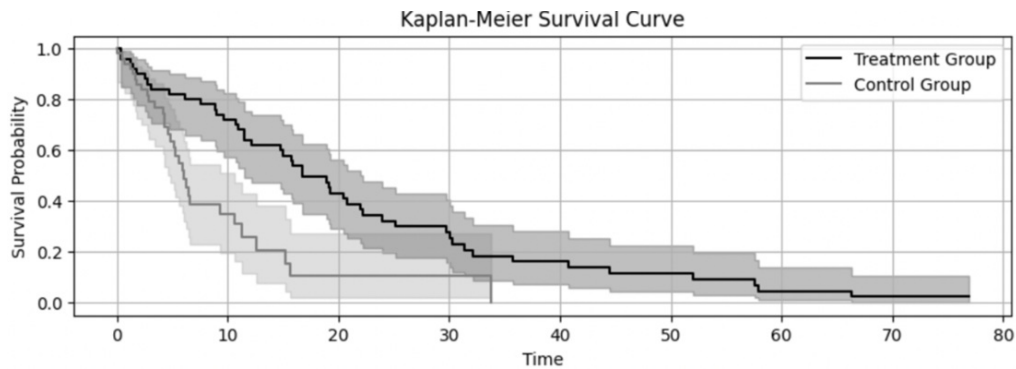
Time series models are invaluable in finance, economic forecasting, environmental studies, etc. These models involve unique challenges, where analysts need to account for the unique *components* of the model (e.g., seasonality, trend, and noise). For example, figure 9.5 shows a model with seasonality (patterns recurring over a time period) and an *increasing trend*. Each of these needs to be separately accounted for to forecast future values appropriately. Models like ARIMA (autoregressive integrated moving average) and Holt-Winters can identify and adjust for the specific trends you're working with.

## Survival Analysis

**Survival analysis models** attempt to predict *how much time will pass* until an event occurs. Essentially, these models seek to estimate how long individual records will *survive*. This model class partially overlaps with regression; however, powerful non-regression techniques are available for this type of problem. These are commonly used in clinical trials to predict the survival time of patients with serious illnesses. Outside of clinical research, survival analysis models can be used to answer questions such as the following:

- How can we model the lifetime of a machine in a factory based on usage, environmental conditions, and material properties?
- Can we predict the time until a customer unsubscribes from your service?
- What factors contribute to users dropping off the paid subscription sign-up process?

**Figure 9.6** Survival analysis models predict the probability of surviving an event after each time period.



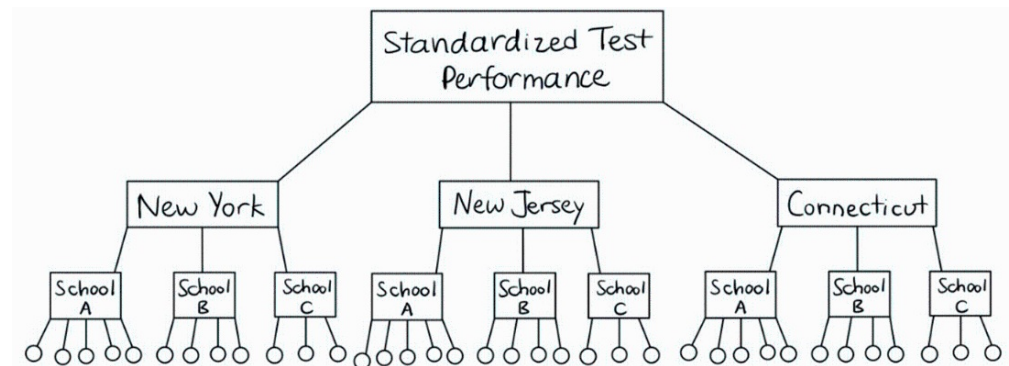
At each time interval, survival curves estimate the probability of survival for the sample used to fit the model. The curve in Figure 9.6 appears like a decreasing step, showing that the probability of survival drops after each subsequent time interval for the sample. The drop-off is more pronounced for the control group, suggesting that the treatment is associated with a prolonged probability of survival. The `lifelines` library in Python allows you to easily apply many common formulas for survival analyses, such as the Kaplan-Meier survival curve shown in the figure above.

## Hierarchical Models

**Hierarchical** or **multi-level models** are extensions of regression and classification models that account for situations where your factors have an inherent *hierarchical structure* (e.g., a city within a state within a country). These models look for patterns at the level of the individual you collect data from *and* the nested groups they belong to. Hierarchical models can help you answer questions that otherwise require separate models, such as the following:

- How can we understand the performance of different sub-contractors in our company, considering hierarchical relationships between each sub-contractor and the primary contracting company?
- Can we understand and predict an employee's tenure, considering individual and department-level properties?
- Can we model student test performance, considering individual student factors, classroom-level factors, and school-level factors that can each impact test scores?

**Figure 9.7 Hierarchical models allow you to capture multiple levels of underlying patterns in your data.**



Hierarchical models are valuable when studying complex real-world phenomena in the social sciences and related domains. If your work involves understanding people in their day-to-day environments (e.g., non-profit or public sectors), hierarchical models may be a common choice in peer-reviewed research that you can reference in your work and may also be an excellent choice to answer your stakeholders' questions. Many hierarchical models can also be used to better model *sparse* data (data where most values are zero or the sample size is low) with more than one level.

## Clustering

**Clustering models** look to categorize data points into *groups* or *clusters* based on their underlying similarity according to variables you select, where you don't have a designated outcome or y-variable. Most clustering algorithms are considered *unsupervised model* in machine learning, dealing with data has no pre-existing categories. In those cases, it's up to the user to interpret the model's outcome and determine if there is value in the clusters to which each data point was assigned.

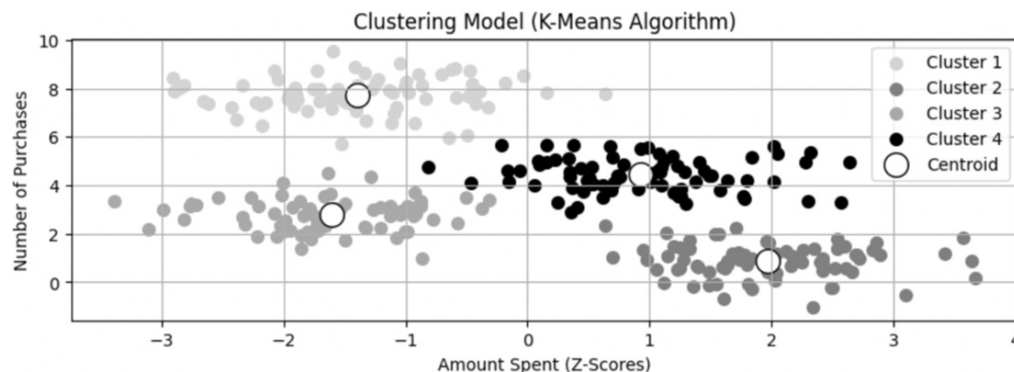
Questions answered using a clustering model tend to be more open-ended and may look like one of the following examples:

- Can we use patterns in customer browser behavior, purchase behavior, and feedback to identify distinct *groups* or *segments*? Can these segments help us improve how we meet customer needs?
- Can we group neighborhoods based on median income, school

performance, and population density to support urban planning and resource allocation efforts?

- What underlying patterns of customer behavior exist based on the usage pattern of our software? Can we discover underlying groups based on log-in frequency, number of actions taken in the app, and time spent on the app?

**Figure 9.8** Clustering models look for underlying patterns to create categories based on the data.



Clustering models can be a valuable tool for surfacing patterns you didn't know existed. However, these models are *not* magic – there's a lot of manual work necessary to determine if the clusters generated are meaningful, easily differentiable, and able to generate value for your organization. Be prepared to spend *a lot* of time understanding the differences between clusters after you fit the model.

## Dimension Reduction

**Dimension reduction models** aim to reduce the number of variables in a dataset by creating *composite variables* that represent the majority of the variability present in the data. The new composite *dimensions* correspond to *combinations* of the original set of variables and are then used as inputs in other models *to replace* those original sets. These techniques are used when working with a large number of inputs that would make it challenging to discern meaningful patterns in the data.

For example, the dataset below has ten columns to potentially use for predicting the amount of time until the next purchase or the amount of that

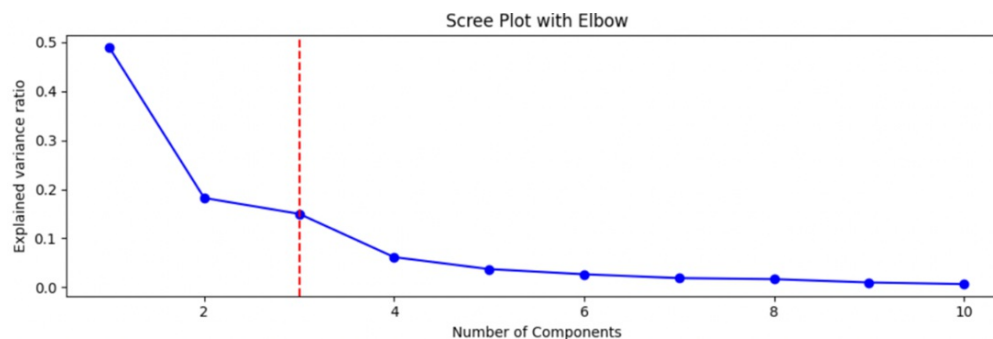
purchase.

**Figure 9.9** Dimension reduction models are used when models contain too many inputs to tease out the effect of any individual variable.

	age	income	avg_spend	visit_freq	last_purchase	n_purchases	loyalty_card	time_in_store	n_returns	feedback_score
0	56	103763	406	27	17	81	0	102	8	2
1	69	28680	21	49	1196	98	1	33	0	3
2	46	104896	616	13	1670	13	0	69	8	2
3	32	103879	311	33	1624	99	0	131	8	4
4	60	91295	907	34	325	68	1	140	8	2

Ten input variables are *a lot* to use in any kind of model, which tends to perform worse with increasing complexity. A dimension reduction model such as Principal Component Analysis (PCA) can potentially represent the variation of these ten inputs in three or four "components." For example, the plot below (known as a "scree plot") shows the proportion of total variance among the input variables explained by each additional component. A cutoff is selected at the "elbow," a visual point of diminishing returns from each additional component.

**Figure 9.10** Often, a large number of variables can be represented with a fewer number of "components."



## 9.1.2 Model Output and Diagnostics

Fitting a model in Python, R, or any proprietary software is straightforward. A few lines of code in Python summarize an *Ordinary Least Squares* (OLS) regression model, a common formula for estimating linear patterns in your data.

With this and any model you fit to your dataset, you can calculate several **diagnostics** that inform you about its behavior, fit, and performance. These include coefficients signifying the importance of individual predictors, metrics that evaluate the overall predictive accuracy, and *residuals* that assess discrepancies between observed and predicted values.

The statsmodels library gives you a detailed summary of a fitted regression model containing nearly every diagnostic you will need to evaluate your model:

```
import statsmodels.api as sm
import pandas as pd
housing_prices = pd.read_csv("housing_prices.csv")    #A

X = housing_prices[["sq_footage", "n_bedrooms"]]
y = housing_prices["price"]    #B
X = sm.add_constant(X)    #C

model = sm.OLS(y, X).fit()    #D
print(model.summary())
```

**Figure 9.11 A statsmodels summary output for an ordinary least squares regression predicting housing prices using the square footage and number of bedrooms.**

Model residuals  
(requires additional  
code to investigate)

③

Dep. Variable: price

Model: OLS

Method: Least Squares

Date: Wed, 06 Sep 2023

Time: 19:01:40

No. Observations: 1298

Df Residuals: 1295

Df Model: 2

Covariance Type: nonrobust

OLS Regression Results

R-squared: 0.836

Adj. R-squared: 0.836

F-statistic: 3307.

Prob (F-statistic): 0.00

Log-Likelihood: -9882.1

AIC: 1.977e+04

BIC: 1.979e+04

②

Evaluation  
metrics for the  
overall model

①

Model  
coefficients

coef

std err

t

P>|t|

[0.025

0.975]

const

900.5972

44.901

20.057

0.000

812.510

988.684

sq\_footage

1.2123

0.016

76.606

0.000

1.181

1.243

n\_bedrooms

250.1064

9.529

26.247

0.000

231.412

268.801

Omnibus: 0.350

Prob(Omnibus): 0.840

Skew: -0.002

Kurtosis: 3.068

Durbin-Watson: 2.009

Jarque-Bera (JB): 0.254

Prob(JB): 0.881

Cond. No. 7.33e+03

The model summary in figure 9.11 contains *a lot* of information. Each value you see contains unique information about the model's performance. Some of the information you see is unique to regression models, some are useful for interpreting classification models, and some have similar parallels to other models. We'll focus on the three critical pieces of output shown in the figure

– the model coefficients, evaluation metrics, and model residuals.

As discussed, we will spend most of this chapter applying these steps in regression models. We'll highlight where there are differences between model classes so you can follow up with additional resources specific to these models.

## Model Coefficients

A regression model aims to estimate or predict a dependent/y-variable using one or more independent/X variables. The model's **equation** allows us to input new data collected about our input variables and estimate the outcome. As we covered in chapter 4, the equation representing a regression model with a *linear formula* is depicted as follows:

**Figure 9.12** Linear multiple regression equation with two independent variables and an error term

The image shows the linear multiple regression equation  $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$  with handwritten annotations. An arrow points to  $\beta_0$  with the text "y-intercept (y-value when  $X=0$ )". An arrow points to  $X_1$  with the text "input value for first X-variable". An arrow points to  $X_2$  with the text "input value for second X-variable". An arrow points to  $\beta_1$  with the text "beta coefficient for the first  $X_1$ -variable". An arrow points to  $\beta_2$  with the text "beta coefficient for second X-variable". An arrow points to  $\epsilon$  with the text "random error that your model can't explain".

This formula provides you with a set of values to multiply the values for each independent variable by their respective coefficients, add them to the y-intercept, and get an *estimated* y-value. From figure 9.12, we can create a real-world example to estimate house prices using the price per square foot and the number of bedrooms.

After fitting a linear multiple regression model, we get a y-intercept and a set of coefficients that we can input into the equation. The first set of highlighted values in figure 9.11 are precisely the values we need—they only need to be multiplied by 100 to get a final predicted house price in the hundreds of thousands:

**Figure 9.13** A linear model estimating the price of a house based on the price per square foot and

number of bedrooms. The carat on the  $\hat{y}$  denotes that we are working with an estimate.

A handwritten regression equation:  $\hat{y} = 90,060 + 121.2 \cdot X_1 + 25,011 \cdot X_2$ . Annotations include: 'Estimated house price' pointing to  $\hat{y}$ ; 'y-intercept (y-value when  $X=0$ )' pointing to 90,060; 'total square feet' pointing to  $X_1$ ; 'beta coefficient for total square feet' pointing to 121.2; 'number of bedrooms' pointing to  $X_2$ ; and 'beta coefficient for number of bedrooms' pointing to 25,011.

From there, when considering a house with 1,500 square feet and three bedrooms we can input those values into the model to estimate the price the house would sell for in the dataset. This gives us the following price estimate:

**Figure 9.14** Estimating a house price with new inputs not included in the original sample.

A handwritten calculation:  $\$346,893 = 90,060 + (121.2 \cdot 1,500) + (25,011 \cdot 3)$ . Annotations include: 'Predicted house price' pointing to  $\$346,893$ ; 'total square feet' pointing to 1,500; and 'number of bedrooms' pointing to 3.

If we have input data about houses *not* in the original sample, we can use this equation to estimate the price the house will sell for. These *out-of-sample* predictions are the basis for so much of the value generated with a fitted model. **Predictive modeling**, as it's known, is the process of using statistical models and machine learning algorithms to *predict future outcomes based on historical data*.

Predictive modeling involves a rigorous process of fitting, fine-tuning, and rigorously testing a model to ensure its applicability in the real world. For example, our housing price equation will probably only apply to the geographic region from which its original sample came. An appropriate model will also need to include other inputs (e.g., money spent on renovations, commuting distance to an urban center) so that it doesn't falter when encountering new information not present in the original dataset. We'll cover this more in sections 9.2 and 9.3; what's important to remember now is that the model equation with its coefficients *is* the deliverable for a situation where we want to predict future values.

## Evaluation Metrics

On their own, model coefficients don't tell us about the ability to predict your

outcome variable. We have access to a wide range of diagnostics available for this purpose. These *metrics* are designed to be evaluated together, giving you a comprehensive picture of how well your fitted regression model performs.

Let's unpack the evaluation metrics for the overall model shown in figure 9.11. There are two types of metrics shown here—**absolute indicators** that tell you about the performance of the single model you are evaluating and **relative indicators** used to compare between models. The first three are **absolute indicators**, which we can interpret for one model:

- The **R-squared** ( $R^2$ ) value is the proportion of variance in your output variable explained by all input variables. When presenting the impact of your model to colleagues and stakeholders,  $R^2$  is your go-to number.  $R^2$  values range from 0 to 1, with a larger value indicating that your model explains a greater proportion of the variance in your output variable. For example, in our housing model from figure 9.11, an  $R^2$  of .836 captures about 83.6% of the variation in a housing price. This is a *very* strong score, indicating the model has a lot of potential for many possible deliverables.
- The **Adjusted R-squared** ( $R^2$ ) value modifies the original  $R^2$  score based on the number of predictors in the model. While the  $R^2$  score increases as you add more predictors, the adjusted  $R^2$  can decrease if those predictors don't sufficiently improve the model's fit. We *don't* want to add infinitely more predictors, and a more discerning metric like the adjusted  $R^2$  reminds us to be careful when increasing a model's complexity. With only two predictors, the model summary in figure 9.11 shows an adjusted  $R^2$  of identical value to the original score.
- The **F-statistic** is a coefficient from a test of the overall statistical significance of your model (remember this from chapter 4?). The F-test in a model assesses whether your predictors *collectively* influence the outcome. The accompanying **p-value** is interpreted the same way as any other statistical test—the probability that the true F-statistic is the same or larger than you generated, assuming your null hypothesis is true (e.g., "the predictors collectively do not influence the outcome"). Figure 9.11 depicts a very high F-statistic and a p-value less than 0.001, indicating that, collectively, the predictors are highly statistically significant

predictors of housing prices.

The other three in this section are **relative metrics** designed to be compared *between* models:

- The **log-likelihood** is an overall indicator of how well your model fits the data. It can range from negative to positive infinity, but the actual value produced for one model *cannot be evaluated independently*. It calculates the following two indicators: popular choices in predictive modeling and machine learning.
- **Akaike Information Criterion (AIC)** evaluates the overall *quality* between models, defined as a balance between fit and complexity—the lower the AIC, the better the model. You have many decisions to make as an analyst—transforming variables, adding predictors, and more. Performing each of these actions is a trade-off—you may increase your  $R^2$ , but you reduce your ability to make accurate and actionable predictions. AIC helps you balance these trade-offs and understand the impact of various changes to your model.
- **Bayesian Information Criterion (BIC)**, like AIC, is used to evaluate the overall quality of your model. It's interpreted similarly to AIC—the lower the value, the better the model. It does include a heavier penalty for the complexity of your model, so this indicator is a good choice when you need to err on the side of model simplicity. We'll discuss practical situations where AIC and BIC are better choices for model evaluation later in this chapter.

Let's increase the complexity of the model shown in figure 9.11 by adding a third predictor, the *number of floors*. If we fit a linear regression to the data once again and display the summary, the resulting model is shown below:

**Figure 9.15** Housing prices summary with an added predictor that introduces complexity to the model.

OLS Regression Results						
=====						
Dep. Variable:	price	R-squared:	0.836			
Model:	OLS	Adj. R-squared:	0.836			
Method:	Least Squares	F-statistic:	2203.			
Date:	Sat, 14 Oct 2023	Prob (F-statistic):	0.00			
Time:	18:13:12	Log-Likelihood:	-9882.1			
No. Observations:	1298	AIC:	1.977e+04			
Df Residuals:	1294	BIC:	1.979e+04			
Df Model:	3					
Covariance Type:	nonrobust					
=====						
	coef	std err	t	P> t	[0.025	0.975]
const	900.8486	53.029	16.988	0.000	796.816	1004.881
sq_footage	1.2123	0.017	72.438	0.000	1.179	1.245
n_bedrooms	250.1204	9.660	25.892	0.000	231.169	269.072
n_floors	-0.3320	37.225	-0.009	0.993	-73.360	72.696

Each evaluation metric we discussed differs slightly with the addition of a third predictor:

- The  $R^2$  and adjusted scores did not change with the additional predictor. This suggests that the third variable did not enable us to explain more variance in housing prices.
- The F-statistic is lower than the original model, though it's still statistically significant.
- Each relative metric is identical to the first model, suggesting we didn't add valuable information to the model with this new variable.
- Further, we can see that the t-value and p-value for the third predictor (in the bottom square of the output) are *not* statistically significant. Overall, this suggests that the number of floors in the house doesn't add any information not already captured in the original model and should probably be dropped. Thus, the best linear model we have fit contains two predictors.

Our example here is simplistic compared to an analyst's steps in the real world. Typically, you can expect to compare multiple models, testing different formulas and combinations of predictors to best represent your outcome variable. These steps are pretty similar across classes of models—the evaluation metrics differ. Still, your task is to compare absolute and relative models to produce the best output for your deliverable.

## Residuals

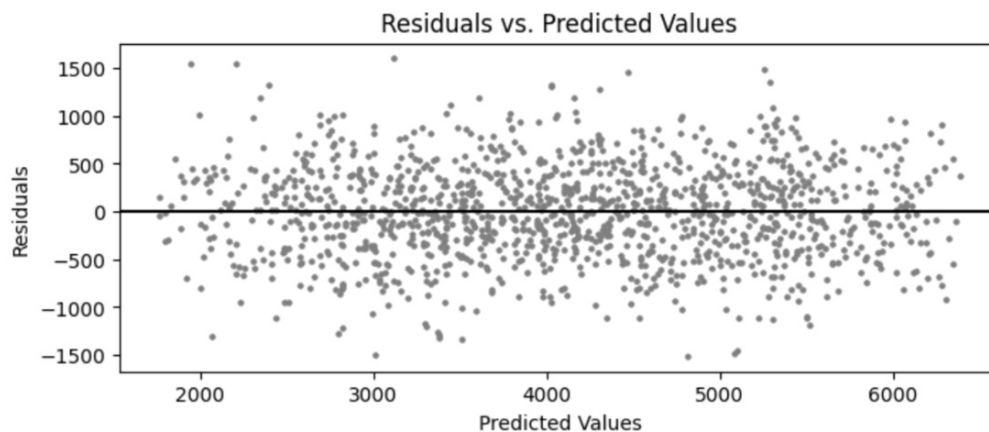
No model is perfect. Some examples come close (e.g., an  $R^2$  of .95), but you're far more likely to encounter situations where achieving a high-quality fit is challenging. Understanding the parts of your model you *cannot* explain is as important as what you can.

This is where residuals come into play, spotlighting the gaps between your model's predictions and the actual data. A **residual** is the difference between an observed outcome and the model's predicted value. By sifting through residuals, you can identify patterns or trends indicating quirks in your data that your model may be missing. This is usually done by plotting the *residual* values against the *predicted* values. Continuing with the model used to generate figure 9.11, we can save and plot our residuals as follows:

```
housing_prices["residuals"] = model.resid
housing_prices["predicted_values"] = model.predict()    #A

plt.scatter(
    housing_prices["predicted_values"],
    housing_prices["residuals"]
)    #B
plt.axhline(y=0, color= "black", linestyle="-")    #C
```

**Figure 9.16** Scatterplot of residuals vs. the predicted values for each record in the dataset used to fit the regression model.



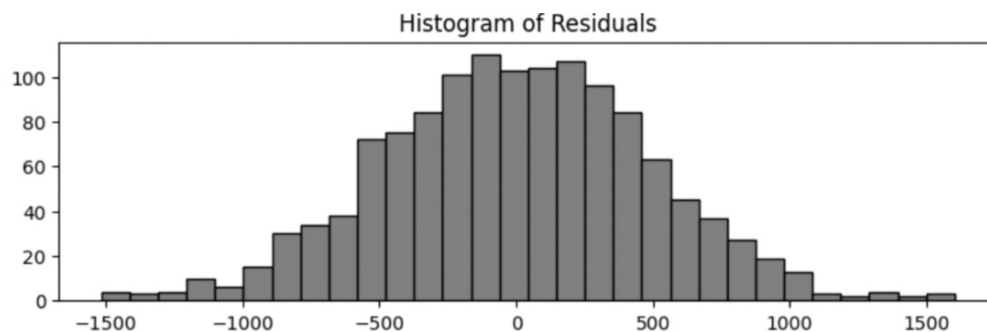
Ideally, the residuals should be equally spread out from zero about the x-axis, indicating low differences between predicted and actual values. The scatterplot should resemble a cloud across both axes. If any clear shape or trend is visible, you may have an issue with **heteroscedasticity** in your model

(a fancy term suggesting that your spread of residuals isn't consistent across your data). For example, if your residuals are positively correlated with your predicted values, your model is less accurate at predicting high or low values.

A histogram of residuals can be equally revealing. We expect residuals to be *normally distributed when working with linear regression models*. We can generate a quick histogram of our housing price prediction residuals below.

```
plt.hist(housing_prices["residuals"], bins=30)    #A
```

**Figure 9.17 Distribution of residuals for the housing prices model.**



The residuals for the housing price prediction model with two variables are normally distributed, indicating *no relationship* with the predicted values. Combined with the high  $R^2$  value, this model seems a good fit for the data! When these criteria *aren't* met, you can take the following steps to diagnose and improve your model:

- Check whether the formula you're using for your data (e.g., linear) is appropriate for the shape of the actual relationship between predictor(s) and outcomes. If the *actual* relationship is non-linear, you may find that your residuals are not normally distributed.
- If you have a small number of extreme outliers in your residual plots that represent erroneous data, you can try removing those records and re-fitting your model.
- Think about the domain you're working in—are there any predictors or explanatory variables that might be missing from your model? Sometimes, an obvious missing variable can cause residuals to capture patterns they shouldn't, and incorporating additional relevant data can remove those patterns. This is *necessary* if you plan on making causal

inferences from your model.

- If the above steps don't work, it might just be that your data isn't a good fit for a linear model. Consider trying more robust techniques such as generalized linear models (GLMs), support vector regression, or random forest regression. You have no shortage of options for trying to model your data.

### 9.1.3 Activity

1. A researcher is trying to predict the exact height of people based on their age, weight, and shoe size. What type of statistical model is most appropriate?
2. In survival analysis, what are we typically trying to predict?
  - a. The amount of time until an event occurs
  - b. The number of events occurring in a time period
  - c. The class label of a row in a dataset
  - d. How many components we can reduce our dataset into
3. A company wants to predict whether a user will click on an advertisement (yes or no) based on age and the number of times they've viewed it. What type of statistical model is best suited for this problem?
4. Write the general equation for a multiple linear regression model with four predictors. How does it differ from the equation shown in figure 9.12? What issues might arise as you continue adding more predictors?
5. Consider the output of a model fitted to answer Question 1. How would you interpret these metrics if the  $R^2$  value is 0.55 and the AIC and BIC are 810 and 920, respectively? What does each suggest about the model's performance and complexity?
6. If the model residuals for Question 1 show two extreme positive outliers, what steps can you take to handle these records? You can assume that the rest of the histogram of residuals appears normal, and the scatterplot of residuals and predicted values is otherwise shapeless and randomly distributed.

## 9.2 The Modeling Process

An effective predictive modeling strategy starts with (surprise!) a question.

Whether you and your stakeholders want to understand factors contributing to churn or predict whether users will complete a paid sign-up workflow, the steps to finding a good fit model are often similar. In this section, we'll cover the investigative techniques you can use to fit many predictive models.

First, let's note something distinct about the questions we'll seek to ask: they're often far more open-ended than those you ask with an experiment or clinical trial in controlled settings. When designing a study, you typically manipulate a limited set of independent variables to test for differences in an outcome variable. You *can* use these techniques in predictive modeling, but you often won't; instead, you can draw from various data sources to understand your variables of interest.

Let's look at an example. Do you remember the `rats` dataset from chapter 3? Fun stuff, right? In chapter 3, we discovered some noteworthy correlations between the total daily rat sightings reported in New York City and several daily weather parameters. If you're interested in *predicting* the number of daily rat sightings, you might start by asking the following question:

*What factors predict the number of daily complaints about rat sightings in New York City?*

As expected, this is a broad question that doesn't yet specify what you expect will predict your outcome. You probably have an outcome in mind—reducing rat complaints—and are open to whatever information will help you reach that desired end state. However, we have a starting point; we discovered several weather parameters with strong correlations to the number of rat sightings, which can potentially predict that variable.

First, let's import the `rats` dataset, join it to the weather dataset, and display Pearson's correlations. Since we may not have rat sightings *every* day, we'll left join `rats` to `weather` and fill in null records with a 0 value in the combined dataframe.

```
import pandas as pd      #A

rats = pd.read_csv("rat_sightings.csv")      #B
weather = pd.read_csv("weather.csv")
```

```

rats_weather = pd.merge(
    weather,
    rats,
    on="day",
    how="left"
).fillna(0)    #C

rats_weather.corr().round(2)    #D

```

**Figure 9.18 Pearson's correlations between daily weather and rat sightings, rounded to two decimal points.**

	high_temp	low_temp	humidity	wind_speed	precip	rat_sightings
high_temp	1.00	0.96	0.17	-0.22	-0.04	0.60
low_temp	0.96	1.00	0.19	-0.25	-0.03	0.61
humidity	0.17	0.19	1.00	0.04	0.23	0.15
wind_speed	-0.22	-0.25	0.04	1.00	0.21	-0.24
precip	-0.04	-0.03	0.23	0.21	1.00	-0.03
rat_sightings	0.60	0.61	0.15	-0.24	-0.03	1.00

The high and low temperatures strongly correlate with rat sightings, and the wind speed has at least a weak to medium correlation. While we can't claim that the weather *causes* rats to be present in visible parts of the city, we *can* assume that more rat sightings will be reported to the city's hotline on warmer days with less wind.

If we're working with a city agency looking to forecast the number of complaints they will respond to, this is precisely the type of model we can try to build. Each of these weather parameters is available as part of a seven or 10-day forecast, so a model with a good fit can be an excellent tool for planning purposes.

From these variables, we can develop a hypothesis that informs our initial model:

$H_0$ : No discernable factors predict the daily rat sightings in New York City.

$H_1$ : Hotter temperatures and lower wind speeds predict higher daily rat

sightings in New York City.

We'll iterate on this hypothesis several times as we discover more about the relationships between predictors and derive additional variables. For now, the strong correlations suggest that linear regression is an excellent place to start.

### 9.2.1 Exploratory Analysis

Exploratory data analysis is woven into every step of our work. Just as you'd explore the underlying trends when conducting a t-test or building a dashboard, a detailed exploration is necessary to understand the complex statistical relationships you're attempting to model. Regardless of your model's assumptions (e.g., linearity for linear regression), you'll perform many of the same steps: visually exploring your data, deriving additional variables, and transforming one or more variables to meet those assumptions.

#### Deriving Additional Variables

The first version of a dataset you synthesize for analysis will rarely enable you to fit a model with value to your organization or business. Most processes easily measured at your organization require some wrangling to get what you need for modeling. This work, also known as **feature engineering** in machine learning, combines your domain knowledge, mathematical insight, and analytical intuition to find the best possible predictors.

Take our `rats` dataset—this was synthesized using 311 data, which contains billions of detailed records for an incredible range of calls to the city's hotline. Since we want to predict the *volume* of rat sighting complaints at a given point, the data was filtered and aggregated to a count of sightings per a given time period. Since we're assuming that a city agency will use the data to forecast the number of complaints and how much staffing they require to respond in a timely manner, we chose a *daily* frequency to enable that process. If the city agency requires the data at an hourly grain, that can easily be updated.

Next, we gathered daily temperature data from a weather service that provides detailed daily historical data. We're combining this weather dataset

with rats for two reasons—local weather parameters are often great predictors of various phenomena within the same geographic region (**analytical intuition**). Additionally, as a lifelong resident of New York City, your author can confirm that rat sightings are far more common in warmer weather (**domain knowledge**).

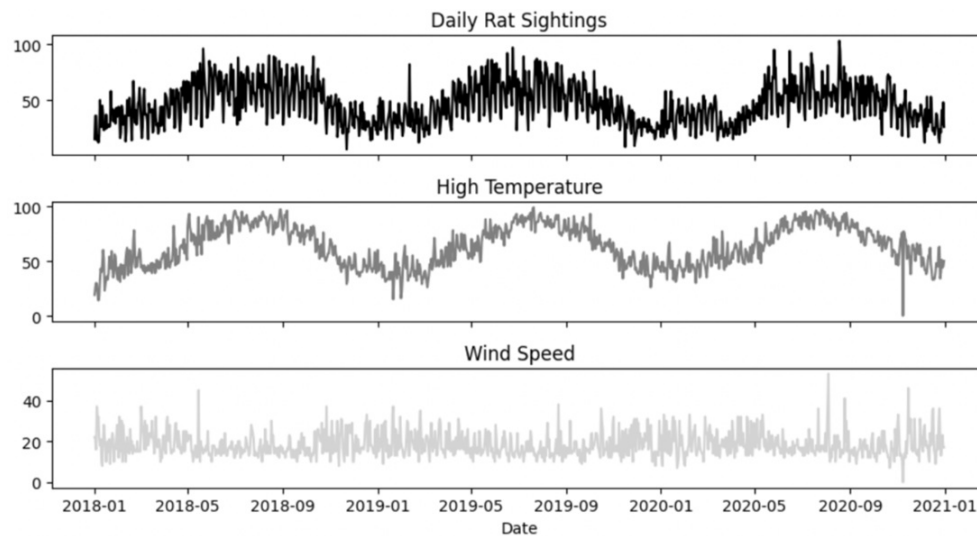
As a reminder, the resulting dataset is shown below:

**Figure 9.19** Preview of the combined rats and weather dataset

	day	high_temp	low_temp	humidity	wind_speed	precip	rat_sightings
0	2018-01-01	19.0	8.0	67.0	22.0	0.00	15
1	2018-01-02	26.0	14.0	59.0	21.0	0.00	36
2	2018-01-03	30.0	18.0	53.0	16.0	0.00	36
3	2018-01-04	29.0	20.0	92.0	37.0	0.02	14
4	2018-01-05	19.0	11.0	56.0	31.0	6.54	18

Chapter 3 reminds us that rat sightings are seasonally dependent, with repeated increases in summer months and decreases in winter months. The daily high temperatures vary with similar patterns, as shown in Figure 9.20 below. The time series plots tell us that daily high temperatures follow the same *yearly seasonal trend* as rat sightings. This isn't the case for humidity, which is still moderately correlated with rat sightings. The low correlation between temperature and wind speed also tells us these are distinct processes. If they *were* highly correlated, we'd want to pick between the two variables since they don't capture distinct variations in the outcome.

**Figure 9.20** Time series plots of our outcome variable and two selected predictors.



There's something else worth noting in these graphs—the daily rat sightings seem to have an additional, more frequent seasonal trend that could be worth capturing in our model. Since the data is captured once daily, we might see a weekly or monthly seasonal trend worth capturing in our model. This points us to our yet unexplored variable—the *day* field. You can extract *a lot* of valuable *ordinal data* from dates and timestamps to represent each type of seasonality—day of the week, day of the month, month number, week number, and day number are all easy features to derive and explore when time is an important variable in your model. Let's extract the month number and day of the week as new, separate columns.

```
rats_weather["dow"] = rats_weather["day"].dt.dayofweek      #A
rats_weather["month_num"] = rats_weather["day"].dt.month      #B
rats_weather.head()      #C
```

**Figure 9.21** First few rows showing the new integer columns with the day of the week and month number.

	day	high_temp	low_temp	humidity	wind_speed	precip	rat_sightings	dow	month_num
0	2018-01-01	19.0	8.0	67.0	22.0	0.00	15	0	1
1	2018-01-02	26.0	14.0	59.0	21.0	0.00	36	1	1
2	2018-01-03	30.0	18.0	53.0	16.0	0.00	36	2	1
3	2018-01-04	29.0	20.0	92.0	37.0	0.02	14	3	1
4	2018-01-05	19.0	11.0	56.0	31.0	6.54	18	4	1

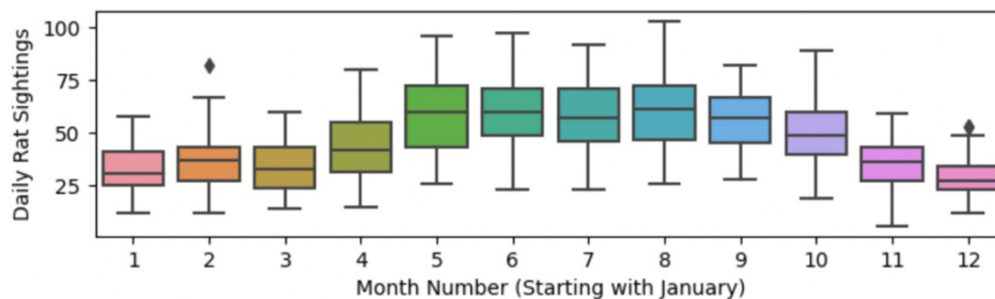
These two new columns have *integers* representing *discrete ordinal* rather than *continuous* data. This means that we cannot reliably run Pearson's

correlations on the data and use them in our model directly. Let's instead look at the distributions of rat sightings for each month number and day of the week, respectively.

```
import seaborn as sns    #A

sns.boxplot(data=rats_weather, x="month_number", y="rat_sightings")
```

**Figure 9.22** Boxplots showing the median and distributions of rat sightings for each month



Unsurprisingly, the median number of daily rat sightings shows a seasonal variation that matches the time series graphs we see in figure 9.20. It's unlikely that the month number provides any new information beyond the daily high temperature, so we can exclude it from the model. Next, let's look at the same graph grouped by the day of the week variable.

```
sns.boxplot(data=rats_weather, x="dow", y="rat_sightings")    #A
```

**Figure 9.23** Boxplots showing the median and distributions of rat sightings for day of the week

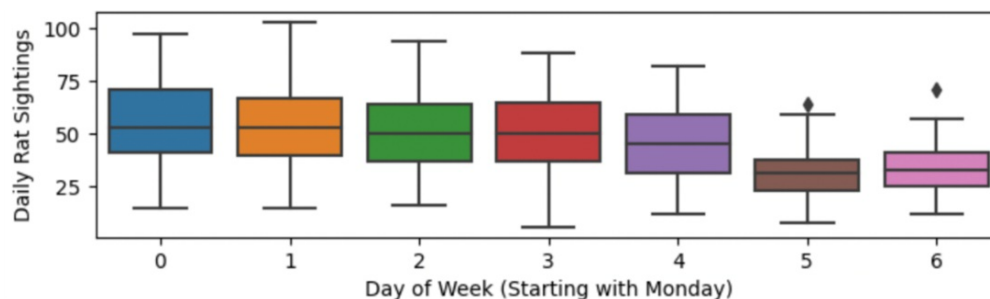


Figure 9.23 suggests that the dow (day of the week) column is a promising new variable—fewer rat sightings are reported on weekends compared to weekdays. This also introduces a new type of seasonality into the model (there's no reason to believe that daily temperatures and wind speed vary

based on the day of the week), so we're unlikely to risk introducing collinearity into our model. It's also a meaningful variable in a model where we want to generate actionable predictions because we can reliably include it for future days. It also represents an essential *behavioral* component of the process that we haven't yet captured—when residents are most likely to see rats and take action to inform the city.

Let's update our hypothesis accordingly to include the third variable:

H<sub>1</sub>: Hotter temperatures, lower wind speeds, and weekdays predict higher daily rat sightings in New York City.

In order to represent the day of the week, let's *dummy code* the `dow` column so we can include it in our model. **Dummy coding** is a technique used to represent categorical variables as a set of *binary* (0 or 1) variables indicating the presence (1) or absence (0) of each category. This method allows linear models to incorporate categorical data by treating each category as a separate variable.

If we were to dummy code each day of the week, we would create separate variables for each day of the week, such as the one shown below. Often, dummy coded variables exclude the final category (e.g., Sunday), which is implied when all other variables equal zero.

**Figure 9.24 Dummy coded variables for each day of the week. The final set of variables for a model will often exclude the last category, which is implied when all other categories are absent.**

	Day	Friday	Monday	Saturday	Sunday	Thursday	Tuesday	Wednesday
0	Monday	0	1	0	0	0	0	0
1	Tuesday	0	0	0	0	0	1	0
2	Wednesday	0	0	0	0	0	0	1
3	Thursday	0	0	0	0	1	0	0
4	Friday	1	0	0	0	0	0	0

However, we know from figure 9.23 that there's little difference in the number of rat sightings between each weekday. We may be able to better represent the variable by creating a simple binary variable indicating whether or not a given day is a weekday. Let's calculate the variable and examine our

Pearson's correlations again.

```
rats_weather["weekday"] = (rats_weather["dow"]<5).astype(int)
rats_weather[
    ["high_temp", "wind_speed", "weekday", "rat_sightings"]
].corr()      #B
```

**Figure 9.25** Correlation matrix including the new binary weekday column.

	high_temp	wind_speed	weekday	rat_sightings
high_temp	1.00	-0.22	0.01	0.60
wind_speed	-0.22	1.00	0.02	-0.24
weekday	0.01	0.02	1.00	0.47
rat_sightings	0.60	-0.24	0.47	1.00

Figure 9.25 shows a strong correlation between the new binary `weekday` column and the number of daily rat sightings—as well as *no* correlation with the temperature or wind speed. Therefore, it seems sensible to include it in our model.

There's no hard and fast rule for how much time or effort to spend deriving additional variables. You'll need to use your best judgment as an analyst for any project, estimating the time it takes to derive each new feature, the pros and cons of increasing its complexity, and when your model is *good enough* for its intended purpose. We've hit a stopping point for our example in this section—we found a new promising variable, and any other inputs would likely require identifying and combining our `rats` data with a third set of variables.

## Evaluating Assumptions

Since we've chosen to start by fitting a linear model, we'll need to evaluate the assumptions associated with a linear process. Specifically, we'll need to answer the following questions:

- Are the relationships between our predictors and the outcome of interest linear? Do we need to perform a transformation (see chapter 4) to better

represent the relationship?

- Are the predictors correlated with *each other*? This is known as **collinearity**, which makes it challenging to isolate the impact of any one predictor and can lead to inaccurate model coefficients and poor-quality predictions.

The seaborn library has a great option known as a `pairplot`, which visually represents a correlation matrix and includes histograms for individual variables on the identity (diagonal). A pairplot can be generated for an entire dataframe; however, these tend to be quite large, so let's subset this to the variables we include in our hypothesis.

```
sns.pairplot(
    rats_weather[["high_temp", "wind_speed", "weekday", "rat_sigh
) # A
```

**Figure 9.26** Seaborn pairplots are similar to a visual representation of a correlation matrix.

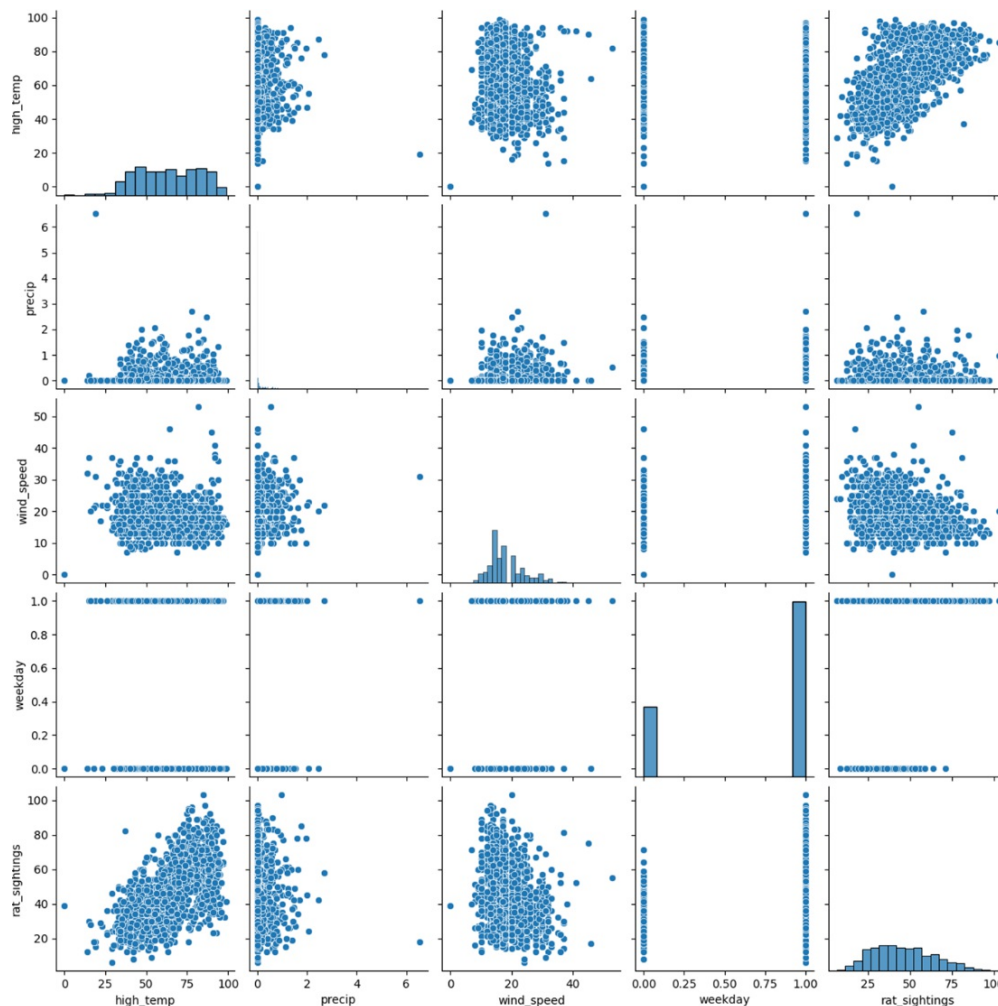


Figure 9.26 shows some interesting trends we should consider testing further before fitting a linear model:

- The relationship between wind speed and rat sightings may be *slightly* curvilinear. We won't know until we transform one of the variables since the scatterplot is densely packed.
- Unlike the other variables in our dataset, the day of the week variable follows a **uniform distribution**. You'll notice it's also a discrete and finite integer rather than a continuous value, which may impact the quality of the model. We'll see if including this variable affects the normality of our residuals later in this section.
- There are no strong relationships between predictors, which aligns with the  $r$ -values shown in figure 9.18.

We can quickly test the first consideration to determine whether a

transformation may improve our model. It's not immediately clear whether the scatterplot shows a curvilinear relationship in the bottom left or top right quadrant, so let's both square and take the square root of wind speed values to see if there's an impact on Pearson's correlation.

```
import numpy as np      #A

rats_weather["wind_speed_sq"] = rats_weather["wind_speed "]**2
rats_weather["wind_speed_sqrt"] = np.sqrt(rats_weather["wind_spee

rats_weather[
    ["wind_speed_sq", " wind_speed_sqrt", "wind_speed", "rat_sigh
].corr()      #C
```

**Figure 9.27 Pearson's correlation values between the transformed variables and the outcome**

	wind_speed_sq	wind_speed_sqrt	wind_speed	rat_sightings
wind_speed_sq	1.00	0.94	0.98	-0.23
wind_speed_sqrt	0.94	1.00	0.99	-0.23
wind_speed	0.98	0.99	1.00	-0.24
rat_sightings	-0.23	-0.23	-0.24	1.00

The correlation values are nearly identical—the squared and square root of wind speed are slightly less correlated with rat sightings than the actual value. Even if they were one point higher, ( $r=0.25$ ), it would not be different enough to suggest that the relationship isn't linear. Thus, on the whole, we can say that we meet the model's assumptions so far. We will again test this once we fit our model and examine the residuals.

## 9.2.2 Fitting a Model

Yes, we've finally made it! It's time to fit and evaluate the model! In practice, you'll probably test the model's fit through exploratory iterations. There are only so many times this chapter can show you the same output with minimal variations, and you'll get no judgment from your author if you choose to test something out quickly. If you perform your exploratory steps with due diligence, you can select whatever work pattern aids your strategic process.

Let's fit our model with the three predictors we've been working with:

```

X = rats_weather[["high_temp", "wind_speed", "weekday"]]
y = rats_weather["rat_sightings"]      #A
X = sm.add_constant(X)                 #B

model = sm.OLS(y, X).fit()              #C
print(model.summary())

```

**Figure 9.28 Summary of a linear regression model predicting daily rat sightings.**

OLS Regression Results						
Dep. Variable:	rat_sightings	R-squared:	0.588			
Model:	OLS	Adj. R-squared:	0.587			
Method:	Least Squares	F-statistic:	519.6			
Date:	Wed, 27 Dec 2023	Prob (F-statistic):	1.02e-209			
Time:	21:27:59	Log-Likelihood:	-4245.3			
No. Observations:	1096	AIC:	8499.			
Df Residuals:	1092	BIC:	8519.			
Df Model:	3					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
const	4.0289	1.949	2.067	0.039	0.205	7.853
high_temp	0.5598	0.020	28.559	0.000	0.521	0.598
wind_speed	-0.3766	0.061	-6.206	0.000	-0.496	-0.258
weekday	18.7393	0.781	24.000	0.000	17.207	20.271
Omnibus:	11.688	Durbin-Watson:	1.337			
Prob(Omnibus):	0.003	Jarque-Bera (JB):	14.968			
Skew:	0.135	Prob(JB):	0.000562			
Kurtosis:	3.505	Cond. No.	381.			

## Evaluating Output

What do you notice about the model output in figure 9.28? At first glance, would you say this is a good fit model? How do you know? A first glance at the evaluation metrics we've covered tells us the following:

- The  $R^2$  of the model is 0.588, indicating that we can explain nearly 60% of the variation in the number of rat sighting calls per day. The adjusted  $R^2$  value is only marginally lower, suggesting the complexity of the model isn't being heavily penalized at this point.
- The overall F-statistic is quite large and statistically significant, indicating that our chosen model fits the data well.
- Each predictor has a statistically significant t-value, indicating that each

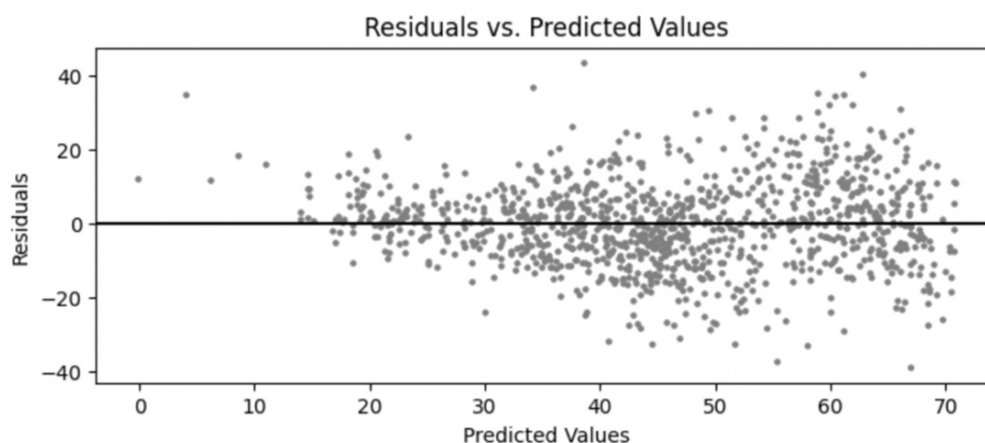
significantly contributes to the variation in daily rat sightings. The relative size and direction of the t-value align with what we saw in the Pearson's correlation matrix in figure 9.18, suggesting that any collinearity present is not impacting the predictions.

These initial findings suggest we can reasonably estimate the number of rat sightings per day with these 3 data points. Next, let's evaluate our residuals by plotting them against the predicted rat sighting values for the model:

```
rats_weather["residuals"] = model.resid
rats_weather["predicted_values"] = model.predict()    #A

plt.scatter(
    rats_weather["predicted_values"],
    rats_weather["residuals"], color="gray", s=5
)    #B
plt.axhline(y=0, color="black", linestyle="-")    #C
```

**Figure 9.29 Plot of residuals against the predicted rat sighting values**



The residual plot isn't as uniform as our example in figure 9.16. The majority of the cloud appears formless, but there's a noticeable gap of values on the far left of the plot—specifically, you can see a few outliers in the top left and an absence of values in the bottom left quadrant. Without one or more of these outliers, the cloud of residuals would seem far more "formless" than expected.

Let's examine the records in the dataset with low predicted values.

```
rats_weather[rats_weather["predicted_values"]<10]      #A
```

**Figure 9.30 Preview of dataset records with low predicted values**

	day	high_temp	wind_speed	rat_sightings	weekday	residuals	predicted_values
5	2018-01-06	14.0	32.0	12	0	12.18	-0.18
6	2018-01-07	18.0	21.0	18	0	11.80	6.20
13	2018-01-14	23.0	22.0	27	0	18.38	8.62
1042	2020-11-08	0.0	0.0	39	0	34.97	4.03

One issue is immediately evident in the data—row 1042 has no weather data! The null values have been filled with 0 for temperature and wind speed, potentially skewing predictions based on illegitimate values. The remaining records appear to be legitimate outliers, which we can consider handling in several ways. We'll discuss that in the next section as we iterate on our model.

## Iterating on the Model

We've come a long way, but don't forget that this is the first actual model we've fit together and evaluated from end to end. We can make many minor improvements to ensure we're meeting all necessary assumptions for linear regression and generating the best predictions for our stakeholders.

We identified two steps to take when evaluating our model:

- Handling the missing weather data by removing or replacing it as appropriate.
- Determining if we should perform a transformation on one or more variables to reduce the number of outliers in our residuals.

Let's start by removing the outlier from the dataset and fitting the model again.

```
rats_weather = rats_weather[rats_weather["high_temp"]!=0]      #A
```

When we fit the model a second time, we get the following overall model

summary:

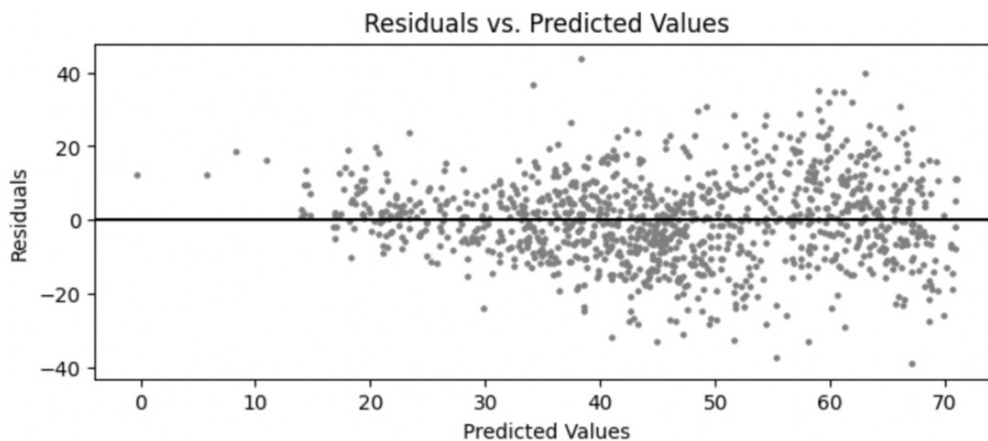
**Figure 9.31 Overall regression results for the model with bad data removed.**

OLS Regression Results			
Dep. Variable:	rat_sightings	R-squared:	0.591
Model:	OLS	Adj. R-squared:	0.590
Method:	Least Squares	F-statistic:	526.6
Date:	Wed, 27 Dec 2023	Prob (F-statistic):	1.66e-211
Time:	22:01:14	Log-Likelihood:	-4237.3
No. Observations:	1095	AIC:	8483.
Df Residuals:	1091	BIC:	8503.
Df Model:	3		
Covariance Type:	nonrobust		

The  $R^2$  has slightly increased by removing *a single* erroneous data point. You'll notice that the AIC and BIC values have also decreased.

Next, let's examine the residual plot for any changes:

**Figure 9.32 Residual plot for an adjusted model, showing the removal of the outlier**



As expected, the most extreme outlier we observed in figure 9.29 is no longer present. However, the change did little to alter the overall shape of the residual plot or remove the extreme outliers on the left-hand side of the plot. You'll recall in figure 9.30 that each of these outliers was present on days where the daily high temperatures were quite low—the weather itself was an outlier on each of these days, which is potentially throwing off the model and leading it to under-predict the number of rat sightings when it's extremely cold. We can try to handle this issue by taking a square root of the y-variable,

in case a curvilinear relationship is present that we could not detect visually.

This is done by adjusting the model fitting code as shown below:

```
X = rats_weather[["high_temp", "wind_speed", "weekday"]]
y = np.sqrt(rats_weather["rat_sightings"])    #A
X = sm.add_constant(X)

model = sm.OLS(y, X).fit()
print(model.summary())    #B
```

**Figure 9.33 Overall regression results for the model with a transformed y-variable.**

OLS Regression Results			
=====			
Dep. Variable:	rat_sightings	R-squared:	0.604
Model:	OLS	Adj. R-squared:	0.603
Method:	Least Squares	F-statistic:	554.8
Date:	Wed, 27 Dec 2023	Prob (F-statistic):	6.74e-219
Time:	22:04:56	Log-Likelihood:	-1382.8
No. Observations:	1095	AIC:	2774.
Df Residuals:	1091	BIC:	2794.
Df Model:	3		
Covariance Type:	nonrobust		
=====			

This second iteration of our model explains up to 60% of the variance in *the square root of* rat sightings, and the F-statistic is much larger. You'll also notice that the AIC and BIC of this model are *far lower* than the first two, indicating that the transformation produces a model of that better explains our outcome variable.

We'll conclude our iterations for this section; this model has served its purpose of demonstrating the model fitting, evaluation, and iteration process. However, determining an appropriate stopping point is ultimately a judgment call you will make as an analyst. Eventually, you will reach a point of diminishing returns on model quality, and spending vast amounts of time for little to no added value is easy. As you engage with the process of model iteration and improvement, consider the following questions:

- Do you have a strong justification for any new predictors you want to add? Is something about the process you're measuring missing from your model?

- What do you and your stakeholders achieve by producing a model with an additional 5% accuracy or variance explained? How about 10%? Is it necessary and worth it to derive value?
- Is there a point where the added complexity of additional predictors diminishes the value of your model? If your stakeholders want to take action based on the independent variables you include, how many different processes can they *take on simultaneously*?

### 9.2.3 Beyond Linear Modeling

Linear models are a great place to start in most modeling scenarios. However, these *aren't* the only models available when you're looking to predict a continuous or categorical outcome (regression and classification). Many of these options *don't* require you to meet assumptions of residuals' linearity, curvilinearity, or normality. Here are just a few examples that you can consider:

- **Generalized additive models** identify non-linear patterns between the predictors and outcome variables. This is done by adding *smoothing functions* such as a spline, which can indicate where a line or curve breaks from its previous pattern.
- While **support vector regression models** fit a line or curve to your data, they use a *kernel*—a function that computes the relationship between multiple dimensions of data—to capture complex non-linear patterns.
- **Random forest regression models** use multiple decision trees (think of a flowchart) to model complex relationships. These models *do not* assume normality, linearity, or any other shape to your data. Random forests and similar models are powerful tools for prediction; as such, they're a popular choice in machine learning.

In some cases, such as with random forests (and other tree-based models), you are potentially sacrificing the ease of explaining relationships between predictors and your outcome. We'll discuss some of these trade-offs and considerations in the next section.

**Table 9.1 Key benefits and considerations when using linear regression models.**

Benefits and Considerations of Linear Regression	
✓ Ease of identifying potential factors/predictors	✗ Assumptions need to be met: linearity of relationships with outcome, normality of residuals
✓ Ease of interpreting the relationships between factors and the outcome	✗ Factors should not be correlated with each other (colinear), as this can impact the accuracy of the output
✓ Relative ease of explaining relationships to your stakeholders	✗ A large number of factors can diminish the quality and interpretability of your model

## 9.2.4 Activity

You're an analyst working at a supply chain company. You've been provided with a dataset called `production_costs`, which contains several variables that may predict the profit margin of a specific item. The dataset includes the following variables:

- `production_scale` represents the scale in which the product is manufactured; the price per unit often decreases as the scale increases.
- `corporate_tax_rate` is the functional rate of taxes paid by the company manufacturing the product.
- `renewable_materials` is a binary variable indicating whether or not renewable materials are being used in manufacturing the product.
- `hourly_labor_cost` represents the cost per hour of all labor associated with the product's manufacturing.
- `product_margin` is the final profit margin on the sale price of the product.

In total, there are over 20,000 records in the dataset.

1. Use the provided dataset to fit a model of your choice. Perform the necessary exploration steps for that model, such as calculating Pearson's correlations and examining the shape of relationships. Based on your observations, is a linear regression appropriate for your model?
2. Which variables are you including in your model, and why? Are the

relationships that you see appropriate to include in your model, or do you need to transform any variables to best represent the shape of the trend?

3. Once you fit the model, interpret the coefficients. What does each coefficient tell you about the relationship between that predictor and the product price? Are there any relationships that changed from Pearson's correlations?
4. Notice the  $R^2$  and F-test results. Do they suggest that it's a good fit for the data? How would you explain the model's fit to someone unfamiliar with statistics?
5. Plot the residuals of your model against the predicted values. Do you see any patterns? What does this imply about your model's assumptions?
6. Perform at least one iteration on your model (e.g., add/remove a predictor, transform a variable). How do the AIC and BIC of your model change with these iterations?

## 9.3 The Statistical Model and its Value

Think back to the joke at the beginning of this chapter. Three analysts are arguing about creating what amounts to a model predicting the same outcome but focusing on a different type of deliverable. Which one is *actually* more complex, sophisticated, and valuable?

Truthfully, none of the three options presented are inherently *better* than the rest. Each type of deliverable is valid and valuable when it's the right choice for your stakeholders' needs. You'll generally want to work with your stakeholders as early as possible to determine their needs and the types of follow-up questions you might receive. Consider the following scenario:

### Modeling Click-Through Conversions from Ads

Anthony is a senior data analyst supporting the marketing team at a large meal plan subscription company. He has a vital role in steering the marketing strategies of the department by providing valuable and timely insights to leadership.

One of the team's marketing managers is looking to understand the

probability that each person who clicks an ad will eventually convert into a paying customer. Ads are placed in a wide variety of formats and on different websites, each of which attracts a subset of the target audience. When a person clicks an ad, the marketing team is provided some basic metadata about that person—browser type, location, and internet service provider. They also have information about the individual ad that was clicked and whether that customer then proceeded through the sign-up process that includes a free trial meal kit.

Anthony knows to start this project by narrowing down the question with the marketing manager – specifically, he asks the following questions:

1. What types of ads are run at the company? Are there classes of ads that need to be accounted for as hierarchies in the model?
2. We know that conversion rate to paying customers is very low (less than 1%), but conversion to the free trial is about 27%. Does it make sense to create a model that predicts free trial conversion, and a separate model predicting free trial to paid conversion?
3. How will the results be used? Does the team want an explanation of the most important factors contributing to conversions, or a live model that predicts the conversion likelihood of each person as they click an ad?

The marketing manager's responses to these questions tell Anthony what type of deliverable he needs to produce – an **explanatory model** that guides future strategy or a set of **predictions** to estimate which individuals clicking ads will become free or paying customers.

These are arguably the most important questions to answer in your modeling strategy. Your stakeholders' specific needs determine what models you can use. If you can anticipate how your model(s) will be used, you will likely minimize confusion and follow-up questions. And who knows – you may win a bet like our hypothetical analysts at the beginning of the chapter!

### 9.3.1 Explanatory Models

An **explanatory model** seeks to clearly describe the *fit* and *shape* of predictors and their relationships with the outcome variable [1]. By

definition, the deliverable of these models includes *explanations* of the strength, direction, and expected impact on the outcome variable if future actions are taken. If you're working with a stakeholder who asks questions such as the following, you may need to fit a model for explanatory purposes:

What recommendations should be made to patients with diabetes and chronic kidney disease to improve their daily blood sugar levels?

Which educational programs should the school district focus on to increase test scores?

Does users' webinar participation lead to an increase in their product usage activity?

Modeling for explanation typically requires you to model your data so that your stakeholders can build an appropriate mental representation of the relationships you're describing. In short, your model has to make sense to them. It has to provide output and recommendations with clear relationships so that they can develop strategic plans based on the findings. This means that your approach will need to meet the following criteria:

- There is a **strong theoretical and logical justification** for including each of your selected predictors in the model. These models are likely to be interpreted causally, so you will need to pair your work with comprehensive domain expertise to ensure you and your stakeholders can draw appropriate conclusions.
- You can **differentiate between input variables** by their importance in predicting the output variable (e.g., with significance testing in linear models or feature importance in tree-based models)
- You can explain the **shape and direction** of the relationship between each predictor and the outcome variable (e.g., time spent on an educational program).

You'll need to use a linear algorithm in most explanatory modeling scenarios to meet these criteria. You can see this in academic research—most peer-reviewed papers leverage linear and polynomial regression and classification models. The results sections of these papers include tables and a detailed description of the statistical significance of each predictor—including those

that did *not* contribute to the actual model.

## Creating an Explanatory Deliverable

Returning to our model predicting daily rat sightings in New York City, let's assume you have the following information from a research science team at a city agency:

- The city's hotline deals with most rat sighting complaints in warmer months, and agencies tend to allocate fewer resources in winter for mitigation efforts.
- Rats are primarily nocturnal, with most sightings occurring between sunset and sunrise the following day. Thus, they hypothesize that the *daily low temperature* has more of an impact than the daily high temperature.
- Rat sightings are unlikely to be impacted by very light precipitation, but they may be more likely to leave underground burrows when rainfall exceeds 0.1 inches.

Some of this information is unsurprising; the trends in rat complaints shown in figure 9.20 are easy to use for seasonal planning. Other information provided is about the biology and behavior of rats to help us choose or modify the model's predictors. With this knowledge, let's consider the following steps for our explanatory deliverable:

- We already know that the high temperature correlates *slightly* more with rat sightings than the low temperature. We can include a note in our deliverable showing the relationship between each temperature parameter, the outcome, and the relationship, indicating that they essentially represent the same process.
- We can add a new variable that turns the precipitation column into a Boolean (True/False) field, indicating whether or not the rainfall exceeded 0.1 inches.

Let's add this new column and examine the correlations:

```
rats_weather["high_precip"] = (  
    rats_weather["precip"] > 0.1
```

```

).astype(int)      #A

rats_weather[
    ["high_temp", "wind_speed", "weekday", "high_precip", "rat_si
].corr()           #B

```

**Figure 9.34 Correlation matrix including the new column for high vs. low precipitation.**

	high_temp	wind_speed	weekday	high_precip	rat_sightings
high_temp	1.00	-0.23	0.00	-0.03	0.60
wind_speed	-0.23	1.00	0.01	0.27	-0.24
weekday	0.00	0.01	1.00	0.01	0.47
high_precip	-0.03	0.27	0.01	1.00	-0.04
rat_sightings	0.60	-0.24	0.47	-0.04	1.00

The correlation between the new `high_precip` variable has an  $r$ -value of -0.04, which isn't meaningfully different from the original `precip` variable ( $r = -0.03$ ). If we're optimizing a prediction model, we would *not* consider including this in our model. However, we *can* iterate on our model to show the statistical significance of the variable and its impact on the  $R^2$  before removing it. If we rerun the model with the additional predictor, it gives us this summary:

**Figure 9.35 Regression results with the added high precipitation variable.**

OLS Regression Results						
Dep. Variable:	rat_sightings		R-squared:	0.604		
Model:	OLS		Adj. R-squared:	0.603		
Method:	Least Squares		F-statistic:	415.7		
Date:	Wed, 27 Dec 2023		Prob (F-statistic):	1.71e-217		
Time:	22:16:25		Log-Likelihood:	-1382.8		
No. Observations:	1095		AIC:	2776.		
Df Residuals:	1090		BIC:	2801.		
Df Model:	4					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
const	3.4041	0.146	23.337	0.000	3.118	3.690
high_temp	0.0429	0.001	29.482	0.000	0.040	0.046
wind_speed	-0.0274	0.005	-5.862	0.000	-0.037	-0.018
weekday	1.4292	0.057	24.870	0.000	1.316	1.542
high_precip	0.0069	0.064	0.109	0.913	-0.118	0.132
Omnibus:	30.941	Durbin-Watson:	1.294			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	39.559			
Skew:	-0.317	Prob(JB):	2.57e-09			
Kurtosis:	3.682	Cond. No.	388.			

If you recall, in chapter 2, we discussed the preparation of written deliverables for your stakeholders that follow the format of a peer-reviewed paper. An explanatory model is one type of those papers—your goal is to summarize the steps you took to find a model with significant predictors and *explain* it to stakeholders in a structured and accessible way. Thus, the methods and results sections of your deliverable can follow a structure such as this:

**Figure 9.36** Slides showing the summary of methods & results that document conclusions about individual input variables.

## Methods

We used a linear regression model due to strong Pearson's correlations between rat sightings and the following variables:

- Daily high temperature
- Daily high wind speed
- Weekday vs weekend, as a binary column

We took the **square root of daily rat sightings** due to a better model fit.

The following variables were excluded:

- **Daily low temperature:** near perfect correlation with the daily high
- **Precipitation:** no correlation with rat sightings

## Results

The three variables included were significant predictors of higher numbers of daily rat sightings reported:

- Higher temperatures
- Lower wind speeds
- Weekdays

Overall, we can explain ~60% of the variation in rat sightings reported

	coef	std err	t	P> t
const	3.4028	0.145	23.424	0.000
high_temp	0.0429	0.001	29.519	0.000
wind_speed	-0.0272	0.004	-6.067	0.000
weekday	1.4293	0.057	24.884	0.000

Since we're presenting our results to a team of research scientists, we're including more technical details directly in the presentation instead of an appendix. With the results we've yielded, we can provide the following

recommendations to the city agency:

- Leverage the seven- or ten-day temperature and wind speed forecasts to plan for staffing needs in the following week.
- Prepare for additional staffing needs to respond to complaints in warm weather, *including* unseasonably warm months.
- Prepare for additional staffing costs, including overtime, earlier in the week.

### 9.3.2 Predictive Models

Until this section, we've been using the term **predictive model** interchangeably with explanation. On some level, an explanatory deliverable still refers to *inferring* or *predicting* the impact of actions taken based on recommendations. You may track changes in your outcome measure and *infer* the impact of any actions you take; however, you are not making *specific* predictions on individual future events that match the shape of your original dataset.

**Predictive models** seek to estimate your outcome variable based on *individual records* you can capture before the outcome data is available. These granular out-of-sample predictions can be used for any number of planning and decision-making activities. With a high-quality model, you can predict an incredible range of processes for your business or organization. Some examples might include:

- Predicting the probability that each customer will churn at their next renewal
- Forecasting the aggregate sales over the coming 6 months
- Clustering new customers into predefined segments

In most cases, an explanatory model can be augmented to create a predictive deliverable. This requires two steps: adjusting your model's code, and ensuring you meet the criteria necessary to reasonably apply it in the real world. Starting with the criteria, let's ask ourselves the following questions about our model predicting rat sightings:

1. Can we reliably collect data about our predictors *far enough in advance*

of the outcome for it to bring value to the team?

2. Is the model accurate enough in the ways we need it to be (e.g., precision, recall, balanced accuracy) to meet the needs of your stakeholders?
3. Do we have the capacity to surface predictions in an appropriate setting for our stakeholders?

For the first question, we can capture the daily forecasted weather parameters included in the model approximately ten days into the future. However, these values are not the same as the actual weather parameters and introduce some unknown error to the model. We'd need to thoroughly test this on multiple samples of forecasted data to better understand the impact on prediction accuracy.

For the second question, we will need to work with our stakeholder to determine how we need to optimize the model. Is it worse if we overpredict the number of rat sightings vs. underpredicting? Where might inaccuracies cost the agency money if the model gets it wrong?

Finally, we'll need to discuss options for surfacing predictions to our stakeholders. Ideally, we will need to retrieve forecasted temperature parameters from an API and refresh the data once per day. We will input the forecasted daily high temperature, wind speed, and day of the week into the regression equation, providing a list or graph of projected rat sighting complaints for each of the following ten days. Depending on the resources we have, this can be done via a pipeline into a data warehouse, directly in a business intelligence tool, or in a manual tool such as a spreadsheet if no other options are available.

Let's assume for this chapter that we will need to use a spreadsheet in order to create the deliverable. We can start by adjusting the model fitting and training process as follows:

- Split the `rats_weather` dataset into a separate *training* and *testing* set, using 80% of the data for training and 20% for testing.
- Fit the model on the training data *only*.
- Generate predictions on the test set.
- Evaluate the accuracy of predictions on the test set compared to the

training set. Comparable accuracy between these datasets tells us that the model makes suitable predictions for data it has never seen before.

We'll use the `scikit-learn` module in order to split the data into training and test sets. In addition to data preparation, `scikit-learn` offers a wide array of algorithms for predictive modeling and machine learning. It provides a less comprehensive summary of individual predictors than `statsmodels`, so it's not particularly suitable for explanatory modeling. Since it's specifically designed for predictive modeling, we'll use it instead to retrain our model for prediction.

```
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error
from sklearn.linear_model import LinearRegression    #A

X = rats_weather[["high_temp", "wind_speed", "weekday"]]
y = np.sqrt(rats_weather["rat_sightings"])          #B

X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=99
)    #C

model = LinearRegression()
model.fit(X_train, y_train)    #D
```

Once we have fitted the model on the training set, we can use it to generate predicted  $y$ -values for the test set and compare them to the actual values. A model that's well suited for prediction will perform similarly on data it hasn't seen before.

In order to evaluate the model, we'll briefly introduce a new evaluation metric—the **root mean squared error (RMSE)**. RMSE tells us the square root of the average sum of squares between predicted and actual values. Essentially, this metric is the *standard deviation of predictions from the true  $y$ -values*. There are numerous evaluation metrics for each class of models that are uniquely valuable to prediction—since we won't be covering them in depth, I recommend additional resources on machine learning in order to appropriately grasp this topic.

Let's evaluate the performance of our rat sightings model on both the training

and test set:

```
from sklearn.metrics import mean_squared_error    #A

y_pred_train = model.predict(X_train)
y_pred = model.predict(X_test)    #B

r2_train = model.score(X_train, y_train)
r2_test = model.score(X_test, y_test)    #C

rmse_train = np.sqrt(mean_squared_error(y_train, y_pred_train))
rmse_test = np.sqrt(mean_squared_error(y_test, y_pred))    #D

print(f"Training Set RMSE: {rmse_train}")
print(f"Test Set RMSE: {rmse_test}")
print(f"Training Set R-squared: {r2_train}")
print(f"Test Set R-squared: {r2_test}")
```

```
Training Set RMSE: 0.84
Test Set RMSE: 0.9
Training Set R-squared: 0.61
Test Set R-squared: 0.57
```

If we evaluate the  $R^2$  and RMSE of the training vs. test set, we can conclude the following:

- Our model explains slightly more variance on the training set than the test set (61% vs. 57%). The difference is not drastic enough to raise concern, but we should rigidly test the model on a new *validation set* of predictions using forecasted weather data and monitor for further decreases.
- The RMSE of our model is quite low. The values tell us that, on average, predictions in both the training and test set are just under 1 rat sighting away from the true value. That's impressive!

If we're confident in our model's predictive validity, we can move on to determining the best deliverable type for your stakeholders. The output of your model can be delivered to your stakeholders as *batch* or *live predictions*.

**Batch predictions** are generated from a model in large, grouped sets at scheduled intervals. In this approach, data is collected over time and stored in

a database or data warehouse and then processed as a set (e.g., for all new records created on a given day). In the absence of access to a data warehouse or data engineering tools, a simple deliverable can be created using nothing more than a spreadsheet. This method of generating predictions is useful when dealing with large volumes of data that *don't require real-time analysis*, such as predicting customer churn, daily sales, or supply chain issues.

**Live predictions** involve generating predictions instantly as new information becomes available. These real-time predictions are essential where you need to make or influence decisions based on the latest available information. The tools necessary for these predictions vary greatly in complexity—a live prediction user interface can be as simple as a spreadsheet allowing user input (e.g., our model forecasting rat sightings), or require complex infrastructure so that products can be recommended as you browse an ecommerce website.

## Creating a Predictive Deliverable

For our final deliverable, let's pivot to our case study and discuss how Anthony might prepare predictions in a format that the marketing team can leverage when planning ad campaigns.

### Predicting Paid Conversions

After additional feedback from the marketing manager, Anthony decides to concentrate on building a robust **predictive model** that predicts conversion likelihood from the free trial to the paid tier. He decides that a set of *batch predictions* generated once per hour is appropriate for this purpose, because it will allow the marketing team to send tailored engagement emails based on the likelihood to convert to the paid tier.

Anthony takes the following steps to prepare his deliverable:

1. Anthony **gathers and preprocesses the data** on individuals who signed up for free trials. In addition to the available data, he derives variables such as the time of day and day of the week to determine if temporal factors influence conversion likelihood.

2. Since is working on a **classification problem**, he leans toward fitting a logistic regression and one or more tree-based models (e.g., random forest). These models are known to perform well on a wide variety of classification problems and offer insight into the most important variables included in the model. He prepares the **training** and **test sets** so he can appropriately **evaluate the model's performance** on out-of-sample predictions.
3. Anthony chooses metrics such as accuracy, precision, recall, and AUC (area under the curve) to **evaluate his model**. After several iterations on the model, he achieves an accuracy of 84%, a precision score of 74%, and a recall score of 65%.
4. Anthony works with the data engineering team to bring the model to production and running on an **hourly schedule**. The data will be available in a new table in the data warehouse that can be joined to the customer ID. Customers who recently signed up or whose information has changed will have new predictions generated with each scheduled run of the model.
5. Anthony reaches out to the product data science team to discuss additional applications for the model, such as in-app prompts to register and receive targeted discounts on the paid sign-up process. He also prepares a report for the marketing team to help them understand the most important inputs of the model, and what strategic decisions they can make based on the exploratory data analysis.

A predictive deliverable brought to production like the one Anthony developed can often be used by multiple teams across the organization. High-quality batch predictions are *new data* about your users or customers (e.g., a user segment) that can be used to understand them and better aid your strategy. These batch predictions can also be used in *live tools*, impacting the experience of using software even if they're not generated on the spot.

Ultimately, this type of deliverable will often necessitate a supplemental *explanatory model* for your stakeholders to understand the process and findings. At minimum, reporting on correlational relationships or statistically significant findings may aid in their understanding of your process even when you don't use a linear model. You'll also often need to consider how to make batch predictions available in a *live* user experience. It sounds like a lot, but

it's all part of the value that can be generated with predictive modeling!

### 9.3.3 Activity

Continuing the previous activity, you are asked to create an explanatory and predictive deliverable using the `production_costs` dataset.

1. The finance team is interested in an explanation of the factors that contribute to production costs and profit margins at the company. Perform the following steps to create the deliverable:
  - a. **Explore** each variable in the dataset with scatterplots and correlations with the `product_margin` outcome. Perform any necessary transformations to represent non-linear relationships.
  - b. Consider the factors that the finance team might hypothesize are significant predictors of profit margins. Are there input variables that align with those hypotheses, even if they're not highly correlated with the outcome? Should they be included in your deliverable? Which ones?
  - c. Fit a regression model that enables you to report on the direction and strength of the relationship between each predictor and the outcome.
  - d. Write a summary of the model's findings for the finance team. Explain which factors impact a product's profit margins. Include recommendations on which factors should be examined for further cost-saving opportunities.
2. The product operations team is interested in predicting the potential profit margin for new products. They want an interactive **predictive modeling tool** where they can adjust the inputs to estimate how the profit margins will change.
  - a. Develop a model to predict `product_margin`, focusing on accuracy and predictive power. You can try more advanced techniques (e.g., random forests) if you are interested.
  - b. Experiment with deriving new features that can improve the model's predictive ability.
  - c. To assess the model's accuracy, examine evaluation metrics like  $R^2$ , AIC, BIC, and RMSE (Root Mean Squared Error).
  - d. Consider what your deliverable to the product operations team

should look like. What would it look like if you *only* had a spreadsheet tool to create a resource for this stakeholder?

## 9.4 Summary

- There are many **classes of statistical models** available tailored to solving specific types of problems:
  - **Regression models** use one or more predictors to predict a continuous outcome (e.g., price).
  - **Classification models use one or more predictors to** predict a categorical outcome (e.g., purchased or not purchased).
  - **Time series models** use historical data collected over time to forecast values for future time periods.
  - **Survival analysis** is similar to regression modeling, where the outcome variable of interest is whether a participant or data point "survives" past a specific time period. These models are often used in clinical settings to understand the probability of surviving an illness for certain periods of time.
  - **Hierarchical models** account for the *nested* structure of your data, such as when participants belong to one or more hierarchical groups (e.g., a classroom, a school, a district) that may impact your outcome.
  - **Clustering models** identify underlying patterns in your data without an outcome variable (*unsupervised learning* in machine learning). These algorithms can help identify groups of individuals or data points that were not discoverable by visual observation alone.
  - **Dimension reduction** techniques consolidate large sets of predictor variables into a smaller set of *components*. Each component represents a significant pattern or aspect of variation within the original data, thus simplifying the dataset while retaining its most informative features.
- **Model evaluation metrics** are available for each class of models, giving us information we can use to diagnose the quality of its fit to the data. When working with regression models, we can use the following metrics to evaluate a model:

- **R-squared** ( $R^2$ ) represents the proportion of variance in your outcome variable explained by all of your input variables.
- **Adjusted  $R^2$**  adjusts the original value, penalizing the score for each additional variable you include in your model. This score enables you to balance model quality and complexity.
- The **F-statistic** tests the overall significance of the model.
- **The log-likelihood** is one of several **relative metrics** that tell you how well your model fits the data. These can be used to compare the relative quality between multiple model iterations.
- **AIC** and **BIC** are relative metrics used to evaluate a model's overall quality. Each score penalizes model complexity, enabling you to balance the fit and complexity to select a best-fit model.
- **Residuals** are the differences between the observed and predicted outcome values that help you understand the parts of your model *not* explained by your predictors. When residuals show patterns that aren't normally distributed, it hints at issues such as missing transformations or variables.
- **Statistical modeling** systematically involves data preparation, analysis, and iterative refinement. It begins with understanding your problem, gathering & cleaning data, and fitting one or more models to best represent that data.
  - **Exploratory analysis** is a crucial initial step that involves examining data through visualizations and summary statistics to identify patterns and spot anomalies, setting the stage for informed model building.
  - **Feature derivation** or **feature engineering** involves creating new variables from your existing dataset to better represent the underlying processes in your model. This can include transformations (e.g., square root), extracting components of a column (e.g., month number in a date), or creating combinations of multiple variables (e.g., combined profit).
  - **Evaluating assumptions** involves verifying the validity of foundational assumptions inherent to your chosen model. This can include checking linearity, normality, homoscedasticity, and the independence of residuals. These steps are necessary to ensure your model's reliability.
  - **Models are refined** using absolute and relative indicators to adjust

- variables, tune parameters, or change the modeling approach.
- **Explanatory and predictive models** are two overarching categories of model deliverables that guide your model fitting and optimization strategy.
    - **Explanatory models** showcase the relationships between individual variables and an outcome, emphasizing interpretability and the statistical significance of coefficients. The predictors included in your model are guided by domain knowledge and rigorously evaluate existing theories and hypotheses. Explanatory model deliverables are usually prepared as detailed reports with recommendations for the intended audience or stakeholder.
    - **Predictive models** focus on the accuracy of out-of-sample predictions and generalizability to new data. Predictions can be delivered in many formats, including interactive tools using the model's equation, batch predictions on large datasets, or live predictions that influence actions taken in real-time.

## 9.5 References

[1] C. Ismay and A. Kim, *Chapter 5 Basic Regression | Statistical Inference via Data Science*. Chapman and Hall/CRC, 2019. Available: <https://moderndive.com/5-regression.html>