THINK LIKE A Data Analyst

Mona Khalil

MEAP

MANNING

Think Like a Data Analyst MEAP V02

MEAP Edition

Manning Early Access Program

Think Like a Data Analyst

Version 2

Copyright 2023 Manning Publications

https://livebook.manning.com/#!/book/think-like-a-data-analyst/discussion

For more information on this and other Manning titles go to

manning.com

welcome

Thank you for purchasing the MEAP for *Think Like a Data Analyst*. 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 achieiving that goal.

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

In this book

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

Analytics has long been a core function of businesses or organizations. You will almost always see dedicated efforts to working with data in a well-structured organization under various titles and functions such as business analytics, business intelligence, product analytics, data science, and more. Leveraging data effectively enables you to understand your users, customers, or stakeholders and iteratively improve and guide your work over time. In the age of big data, maximizing value from data is more critical—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

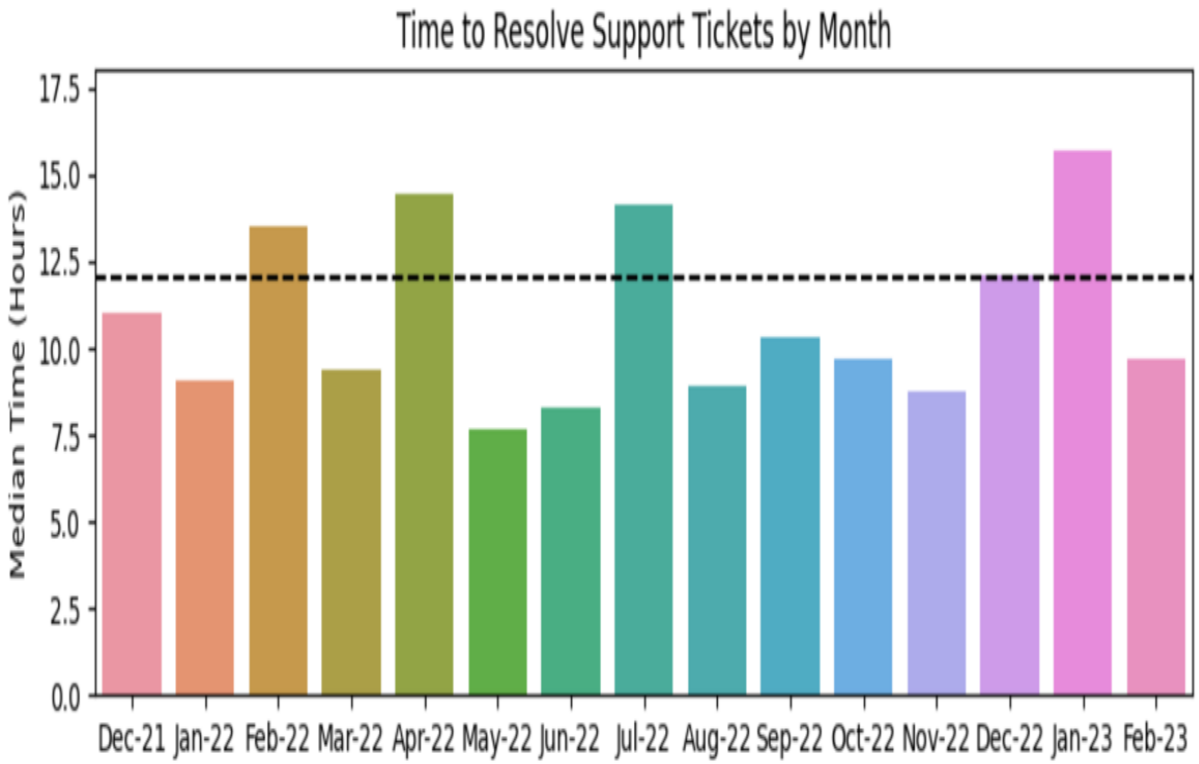
A business intelligence or business analytics team enables tracking and analyzing data about an organization’s performance and makes informed and strategic operational decisions across various functions. This type of team can employ a wide variety of methods of synthesizing data and communicating results 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 and key performance indicators (KPIs) is foundational to the success of a data-informed organization. Many business intelligence teams will track a combination of standard (recommended data points to track used widely within an industry or field) and custom (developed for a specific unique function within an organization) metrics over time. 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 bandwidth to generating novel insights about the function and operation of the business. They may identify areas of inefficiency and revenue-generating opportunities or answer questions from stakeholders to enable them to make increasingly strategic decisions. These results are often shared in the form of ad-hoc reports or presentations.

Developing Dashboards for Ease of Information Consumption

The business intelligence team will frequently work within a business intelligence or dashboard *tool* maintained in-house or purchased as software. These tools enable stakeholders to consume data in a self-service capacity, requiring minimal work from the team on a routine basis. The deliverable for metrics and insights most often includes charts and graphs, which are usually developed and shared as part of a dashboard or report.

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

Win Rate

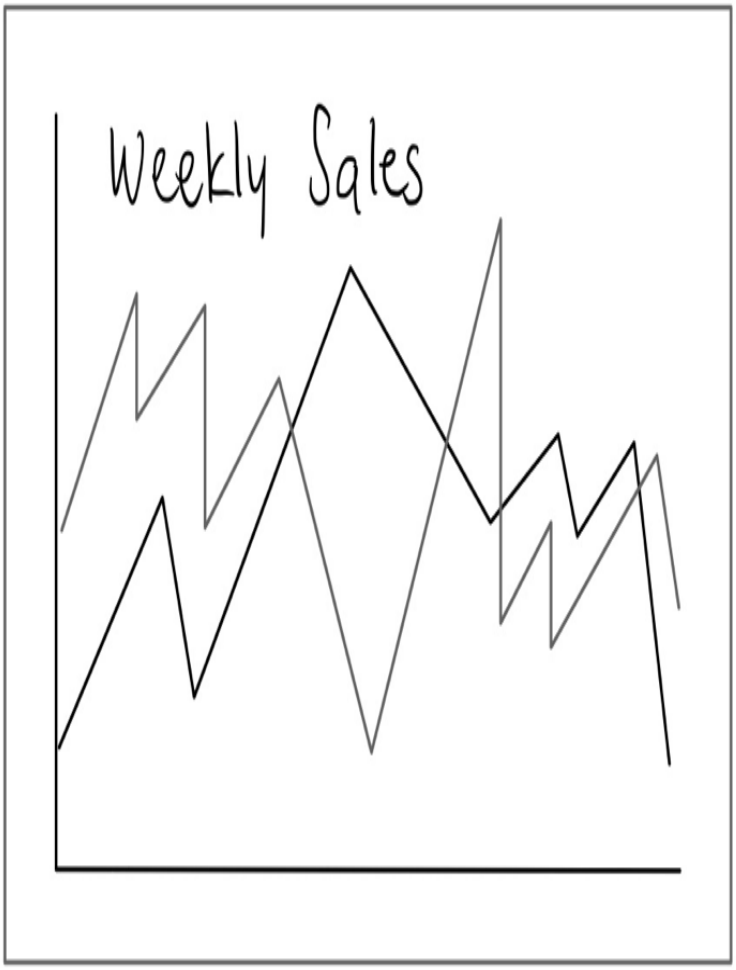
36% -

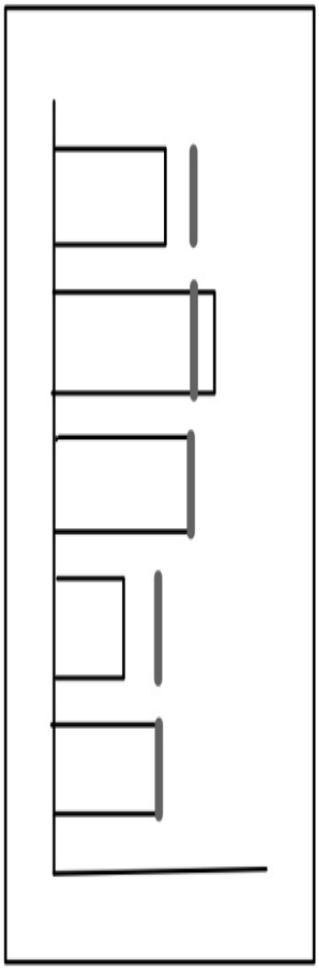
Revenue this Year

$1.5m ↑

Total Customers

3,000 ↑





Distilling and Communicating Results to Business Stakeholders

A business intelligence team often prepares for work that requires them to be highly flexible in delivering insights to stakeholders. They must frequently adjust the granularity, statistical and mathematical terminology, and formatting to meet business stakeholders' various needs and data literacy levels. 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.

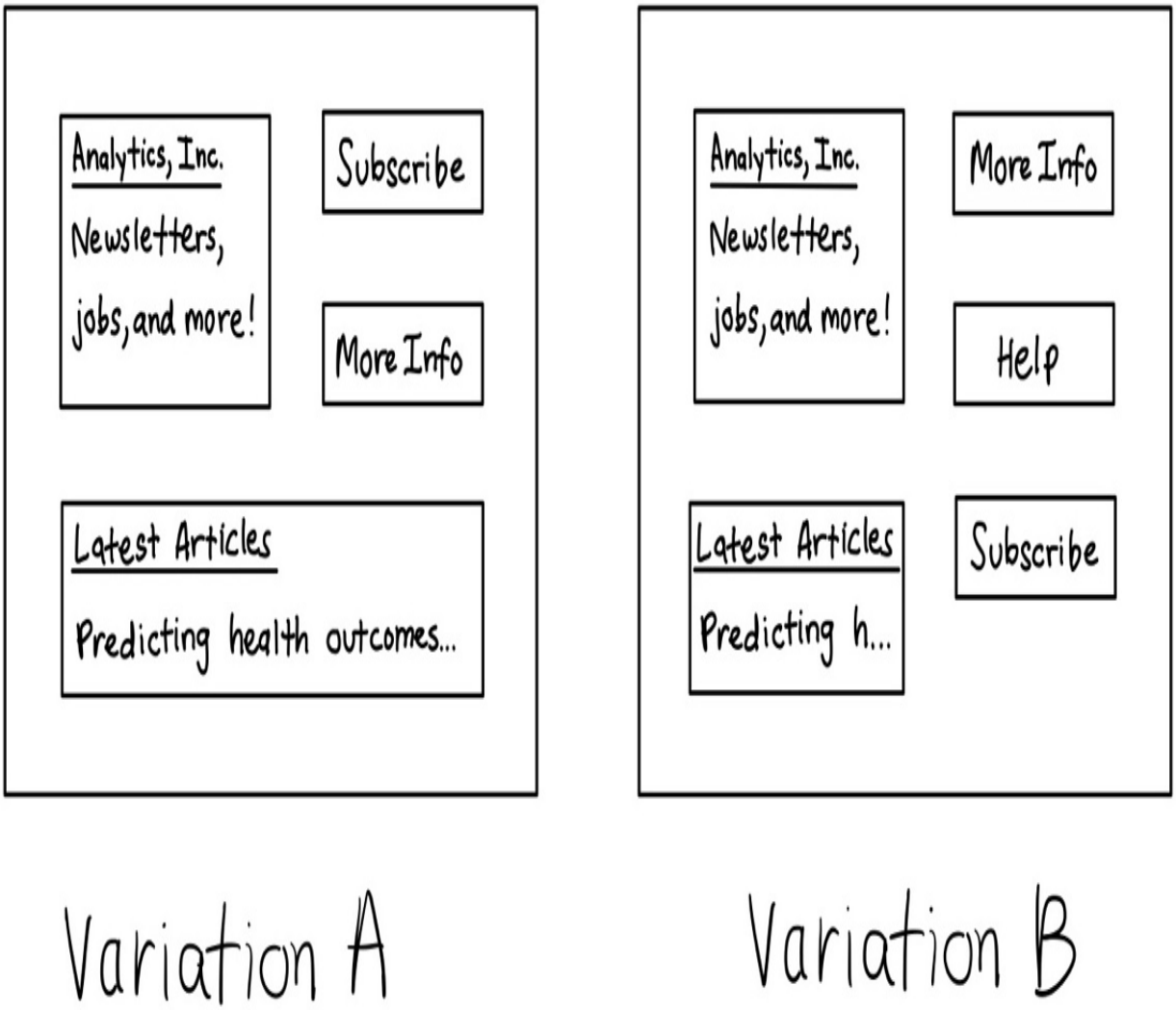
Experimentation

Marketing analytics teams often employ experimental methods or A/B tests to iterate on and optimize for opportunities to engage with prospective customers or users, invite them to enter a pipeline, and convert them to users or subscribers.

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. The methods used to evaluate these tests are often the same tests covered in a college statistics course and are discussed 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.

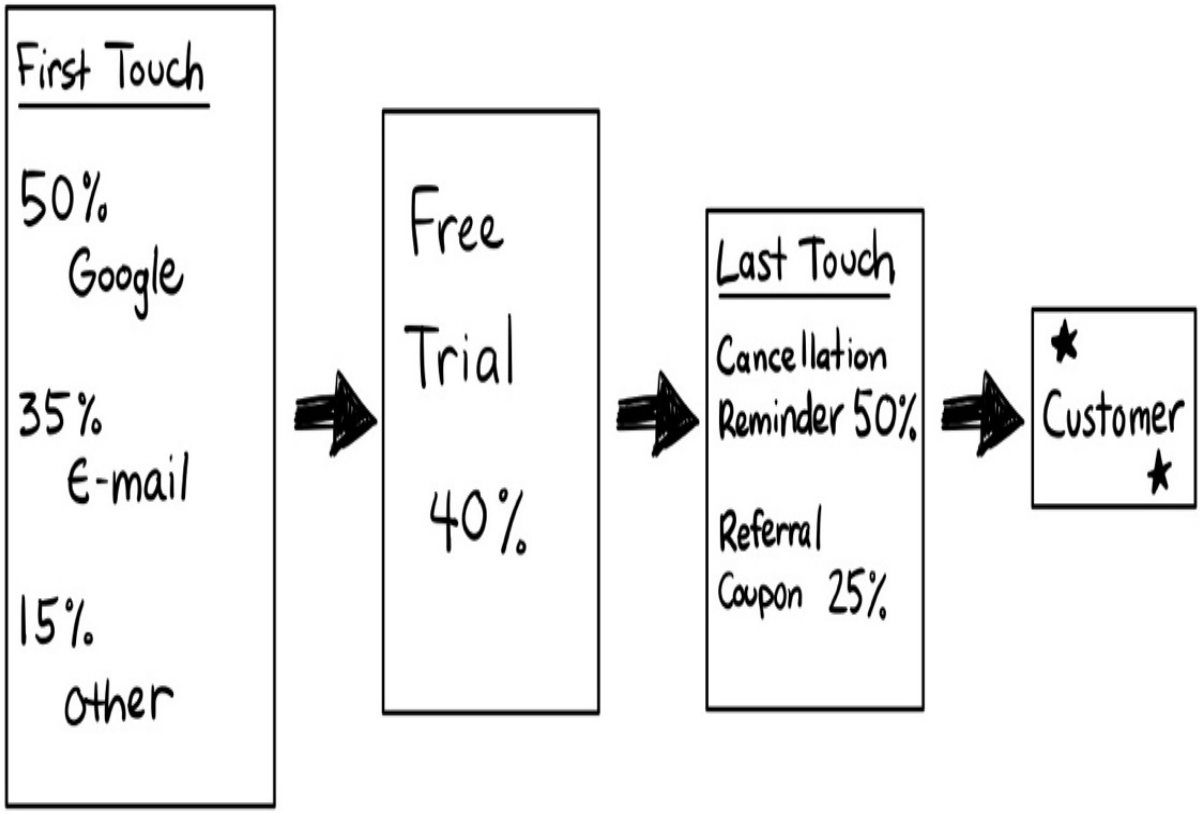


Attribution Modeling

Attribution modeling is the analysis of each touchpoint prior to a purchase or subscription (e.g., visits to the marketing page, requesting a product demo, attending a webinar). These approaches seek to *attribute* a value 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 obtain publicly available data on industry 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 research and compiling information or be involved in the process when a quantitative comparison is necessary. 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, customer lifetime value, etc. Each of these 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.

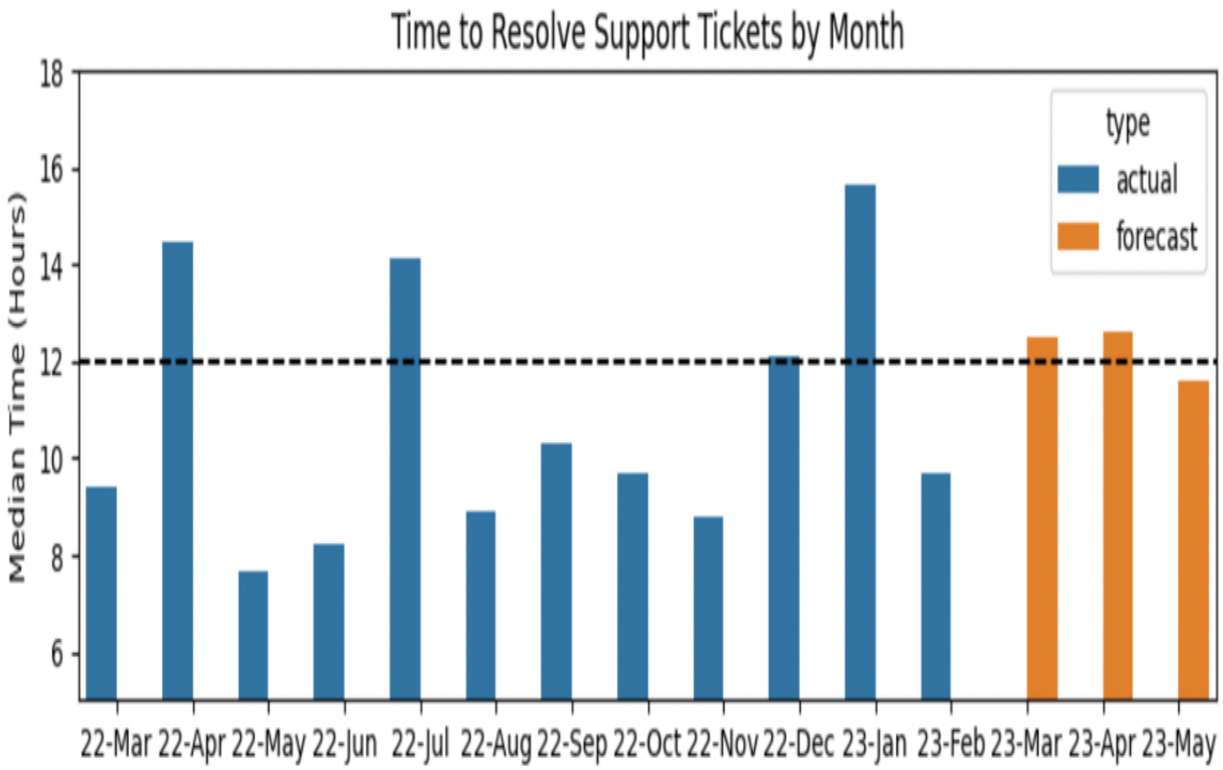
Risk Analysis

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

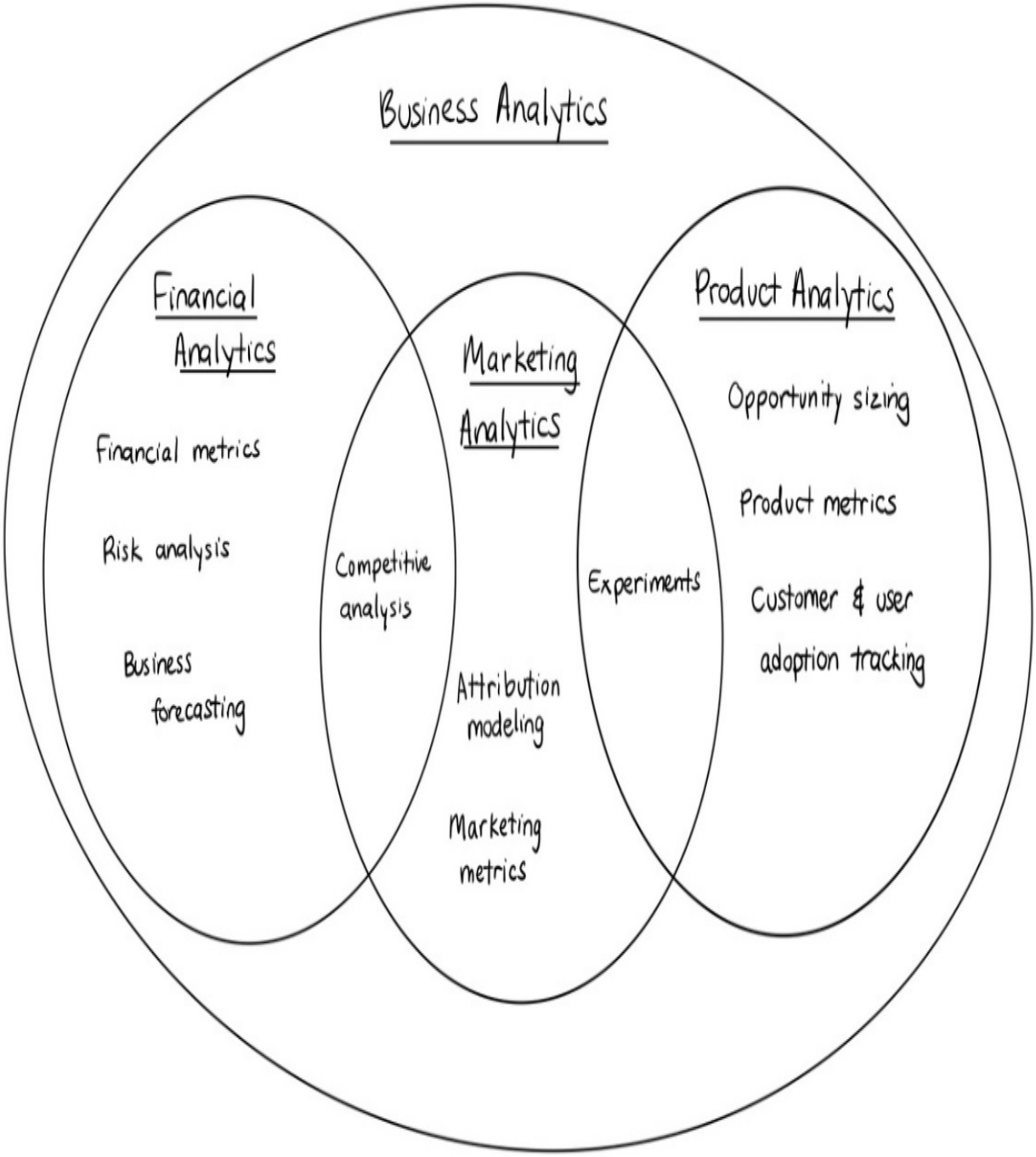
The experimental procedures that marketing analytics teams use are also frequently owned and performed by product analytics or *growth* teams. In addition to simple iterations on layouts, buttons, text, etc., product analytics teams will use a broad range of methods to design and evaluate more sophisticated 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.

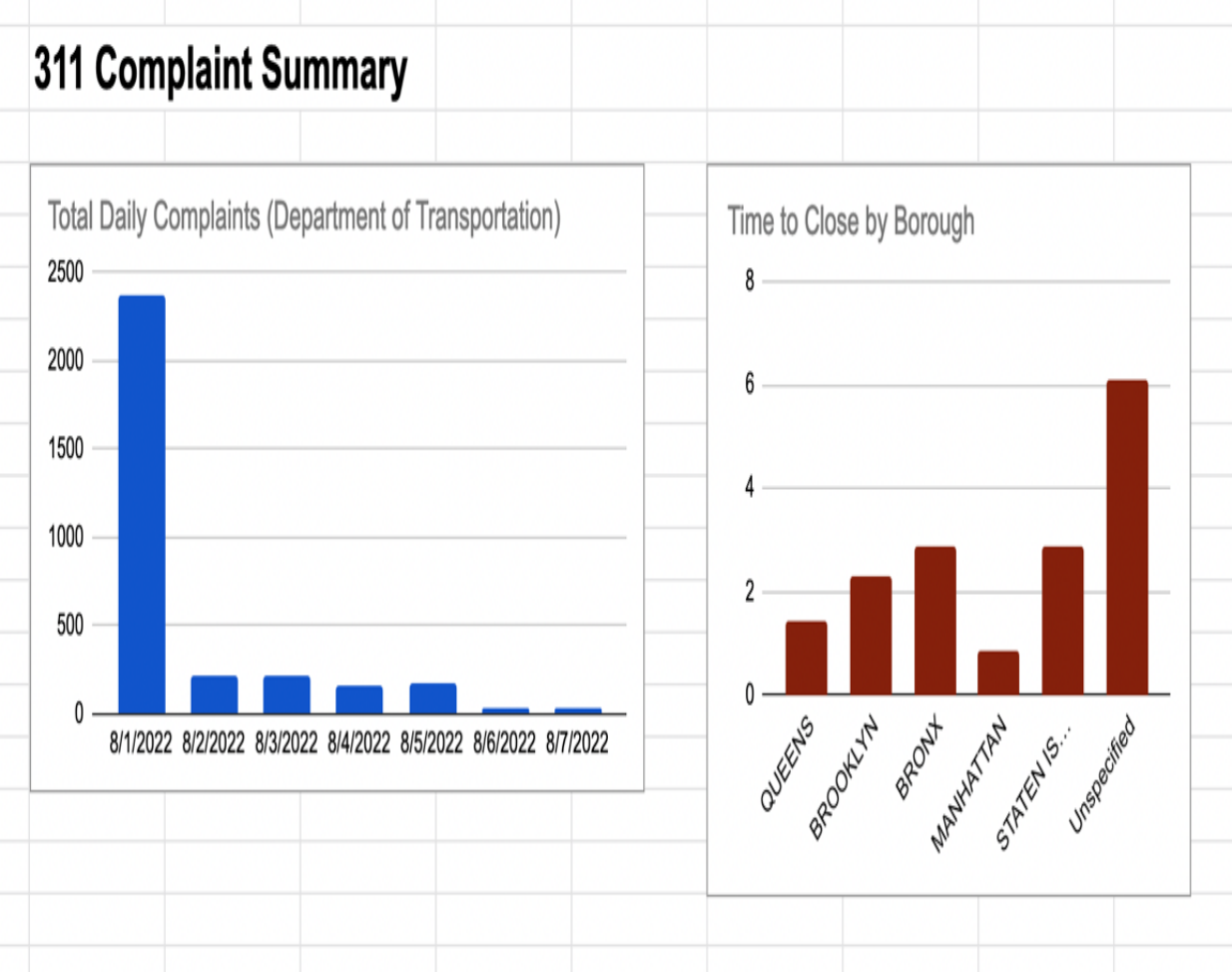
Regardless of your organization's previous investment in data tooling, you will benefit from accessing the following categories of tools for your work.

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!



1.2.2 Querying Language

The majority of organizations store data in tabular format across multiple sources. It’s common to have access to a data warehouse [1] containing structured data about the processes relevant to the organization. In this case, an analyst is usually expected to draw from those data sources using an appropriate *querying language*. This is almost always a dialect of SQL similar to the ANSI-standard SQL [2] taught in the classroom. Even without previous exposure to the specific data warehouse used within an organization, an analyst familiar with SQL can quickly access and manipulate data in that warehouse, using the documentation as

necessary where differences occur (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, as you can see below:

```
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
```

You can also directly write SQL queries in either language to manipulate a data frame.

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. Eventually, you will likely be required to do so 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 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 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 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 available tools and strategies for investing in a data toolkit in chapter 10.

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.

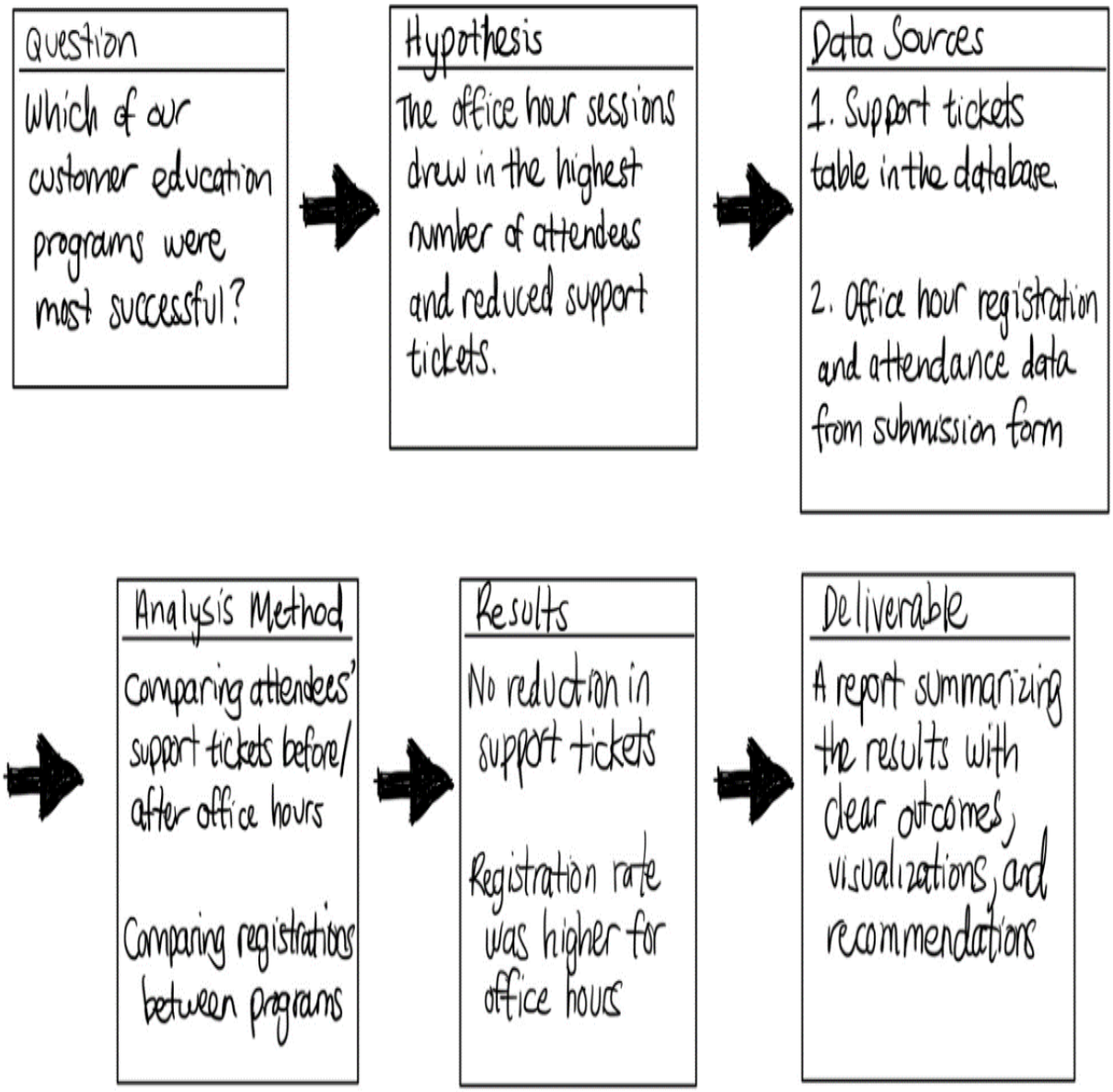
Most analytics teams have a range of responsibilities beyond statistical analysis and presenting results. Analysts are usually expected to consult on the appropriate application and usage of findings and guide stakeholders through asking valuable questions. Much of the success of a project depends on 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

Access to data is often the primary driver of how you ultimately operationalize the original question presented to you. Data is typically available to analytics teams in various formats (e.g., CSVs, relational databases), and each data source has strengths and limitations guiding what you can report on in your work.

We will discuss data formats and sources in greater depth in chapters 6 and 9. For this purpose, we will follow up on the example above with Sam’s product analytics team, highlighting the strengths and limitations of various data sources and how to identify them.

Tracking Page Views

Clickstream and behavioral data are common sources of information for analytics teams. This is often available for product and marketing analytics teams since they’re built into standard tooling and frameworks used by those departments (e.g., Google Analytics, an A/B testing framework, etc.).

Here is an example of a table in the data warehouse that captures data from a third-party data source 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.

Tracking Support Center Calls

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 unique 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.

Help Center Satisfaction

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.

Identifying Dataset Characteristics

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.

The specific examples of datasets in the above section cover a range of items I recommend checking for when searching for appropriate data to answer questions:

- Do you have raw access to *event-level* data (e.g., one row per page view, click, call)?
- Do the different datasets allow you to connect users between various sources?
- 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 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.

2.1.3 Measures and Methods of Analysis

Statistical tests and research methods are *tools* in your analyst's toolkit. Specific methods of measurement, testing, and their responsible application will be discussed at length throughout this book (Chapters 3, 4, 5, 6, and 8). Concerning the rest of the analytics lifecycle, we'll continue with the steps appropriate to the example scenario of Sam's team.

Identifying Limitations

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.

Choosing Methods and Statistical Tests

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, 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 academics. 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

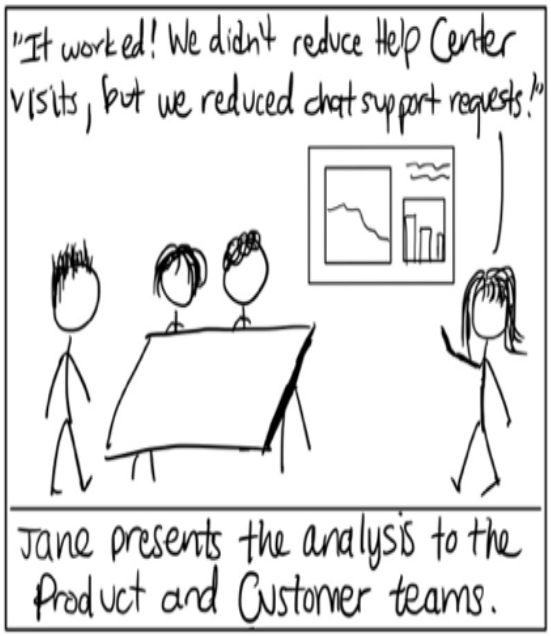
each question, suggesting appropriate datasets and metrics, choosing statistical tests, and identifying appropriate levels of detail for the stakeholders in marketing.

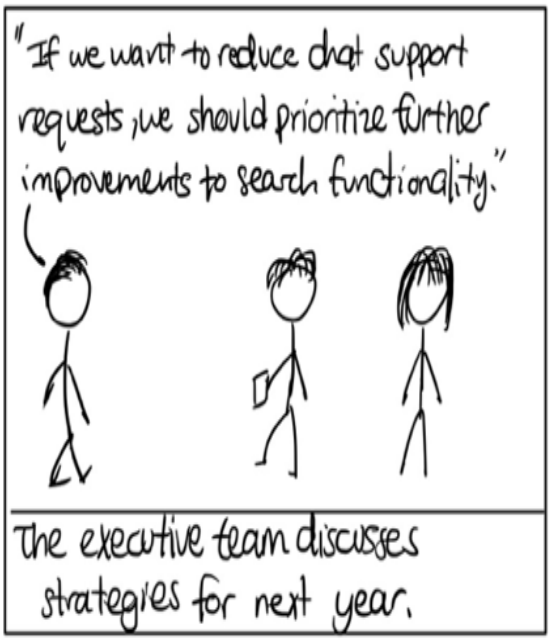
2.2 Communicating Back to Stakeholders

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

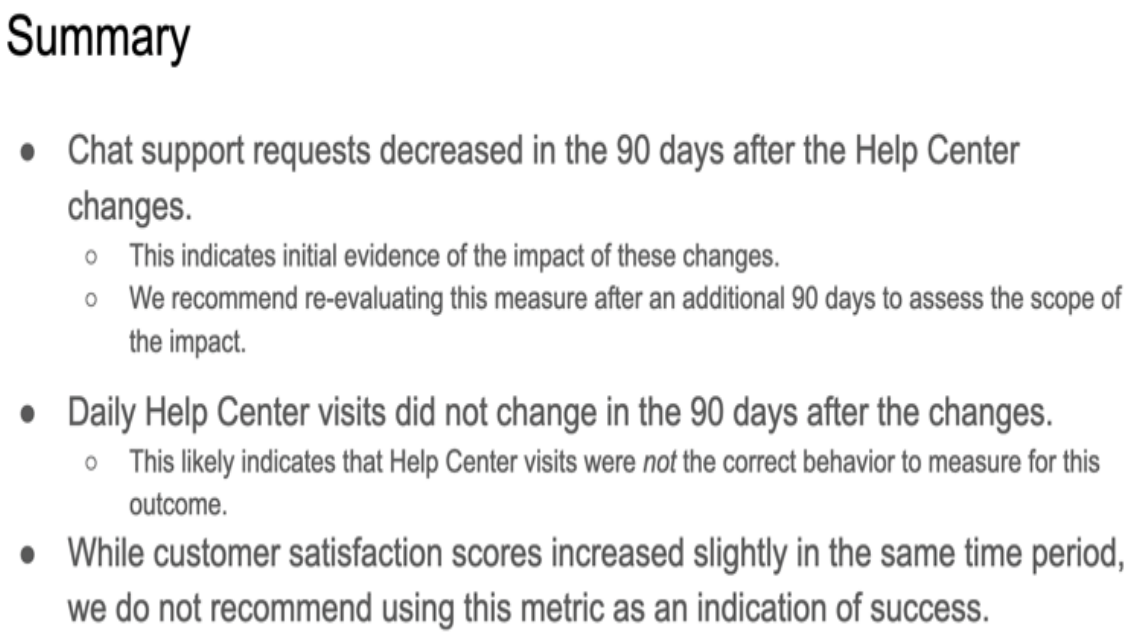
Analysts benefit from the strategic communication of two concepts in their findings:

- *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.



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.

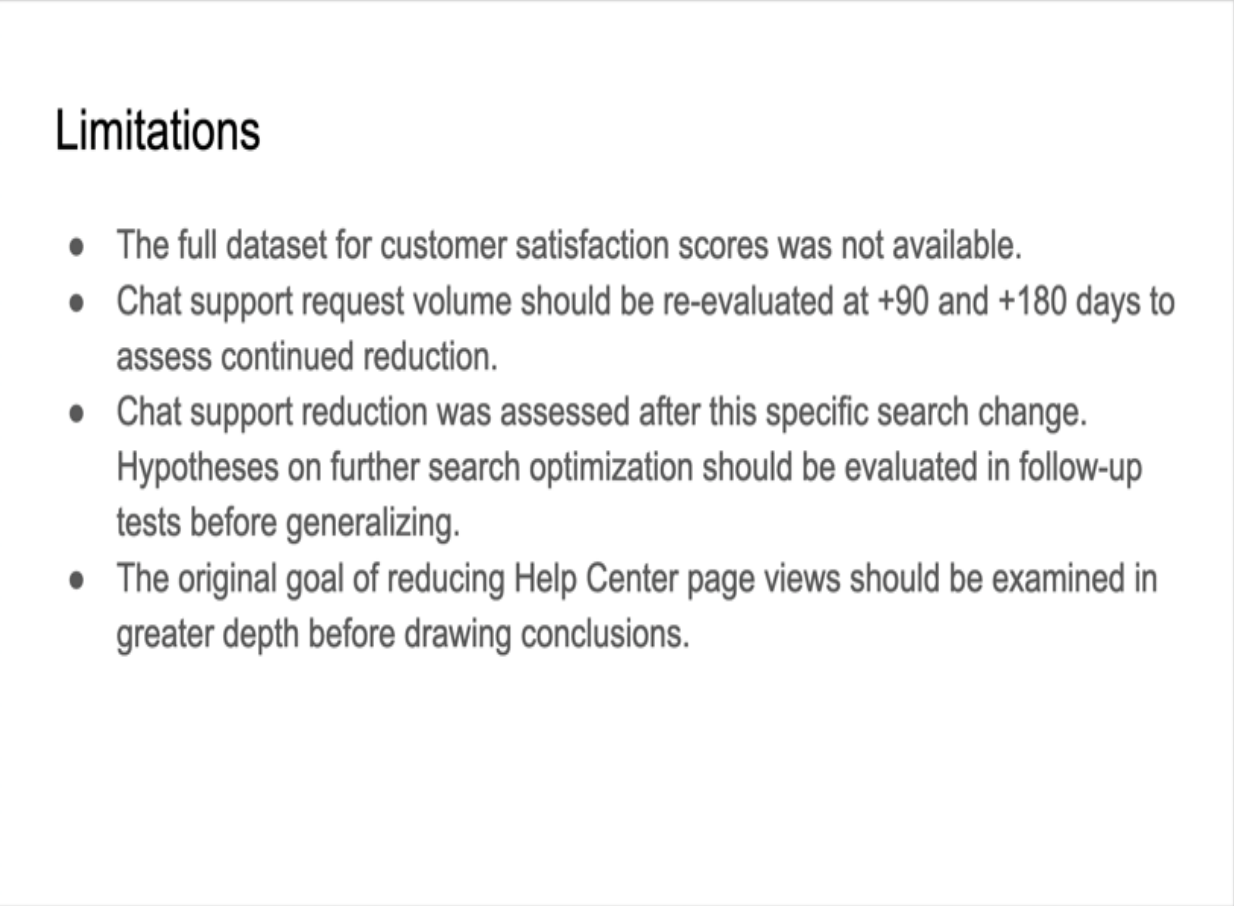
- **Include a limitations section in your report or presentation:** this is a standard section in peer-reviewed papers that is valuable to communicate in your reports as a slide or page for stakeholders to read. Include a list of bullet points of data unavailable for in-depth analysis, the scope of interpretation, 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.



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?

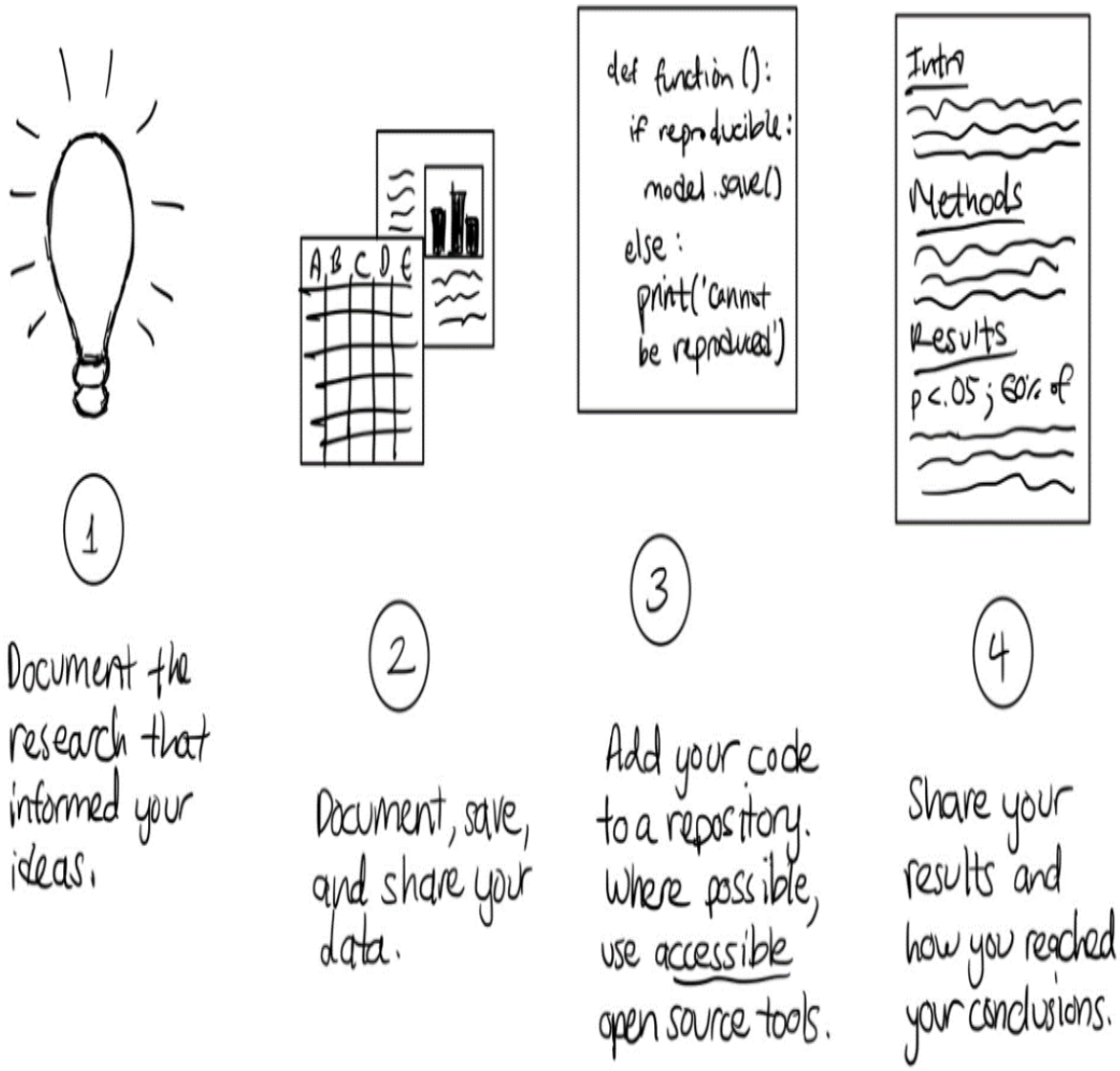
As with the previous exercises, it's important to note that there is no *single* correct answer to each question. It's valuable to document and be prepared to explain your rationale for any interpretation of the results.

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.



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

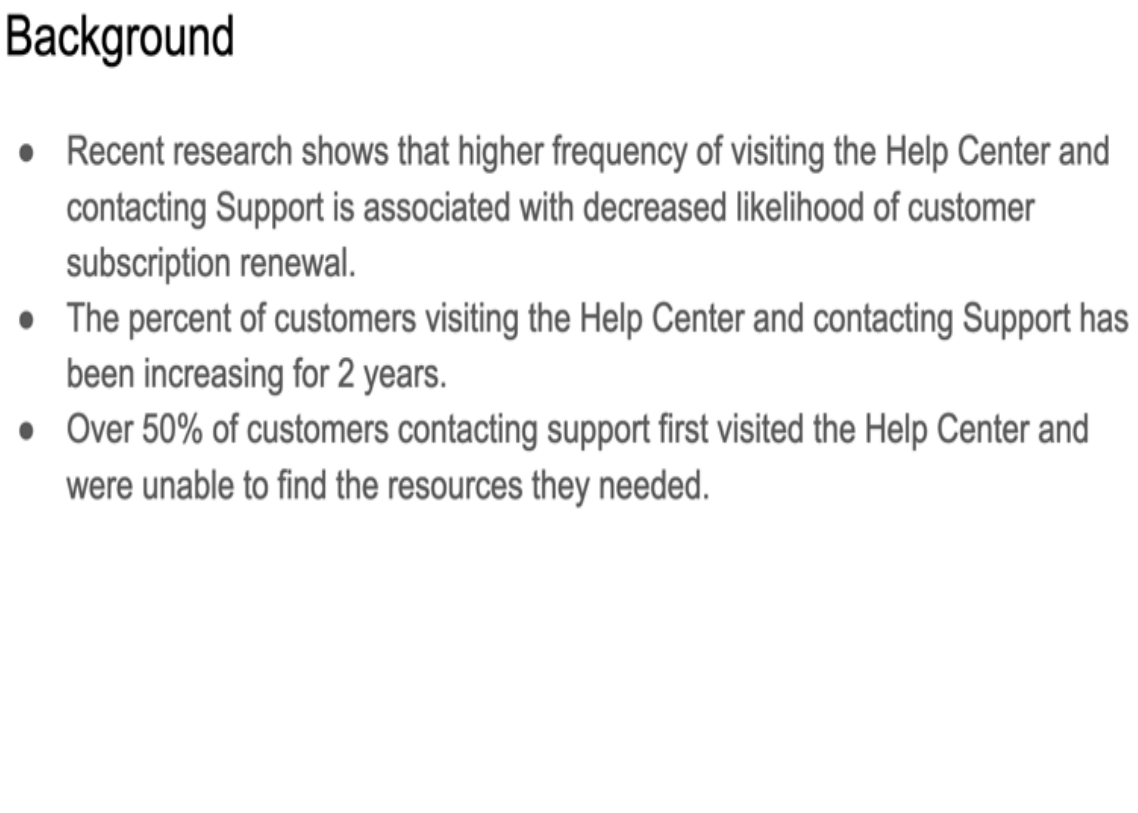
Creating a seamless customer experience is one of the company's core values. For years, we have sought to provide insights into the customer experience and understand what behaviors indicate a positive experience or may instead indicate friction when using the product.

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.



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 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.
- 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

The previous chapter walked through the complete lifecycle of an analytics project and strategies to create success at each step. This chapter will zoom in on the foundational approaches to operationalizing questions, developing hypotheses, and choosing an appropriate method for evaluating the hypothesis. Each of these steps supports you in synthesizing your results and presenting an appropriate analysis for your final deliverable.

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 chapter 4. 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.

Example Case Study: Non-Profit

Let's look at our example 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.

Example Case Study: Non-Profit

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 understandings 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?

Example Case Study: Non-Profit

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.

Example Case Study: Non-Profit

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.

Summary

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

H_0: The test scores in the treatment group (individual tutoring) are equal to or less than in the control group (no tutoring).

H_1: 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

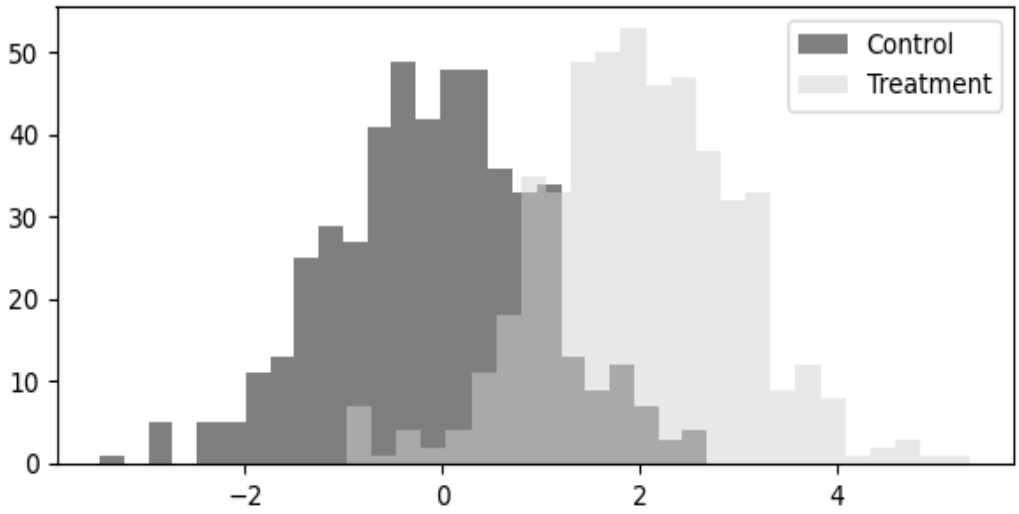
Quantifying a hypothesis is a process aligned with *Bayesian hypothesis testing* instead of the commonly used framework of *frequentist hypothesis testing* we've discussed so far. We won't go into depth on probability and Bayesian methods – there are fantastic resources available that you can leverage to build that knowledge. However, quantifying a hypothesis can be valuable in learning to *think 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 3.1 Overlapping 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.

Example Case Study: Non-Profit

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:

H_0: There is no significant difference in the registration rate between adult and adolescent volunteering events.

H_1: 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:

H_0: The is no significant difference in the registration rate between adult and adolescent volunteering events.

H_1: 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).

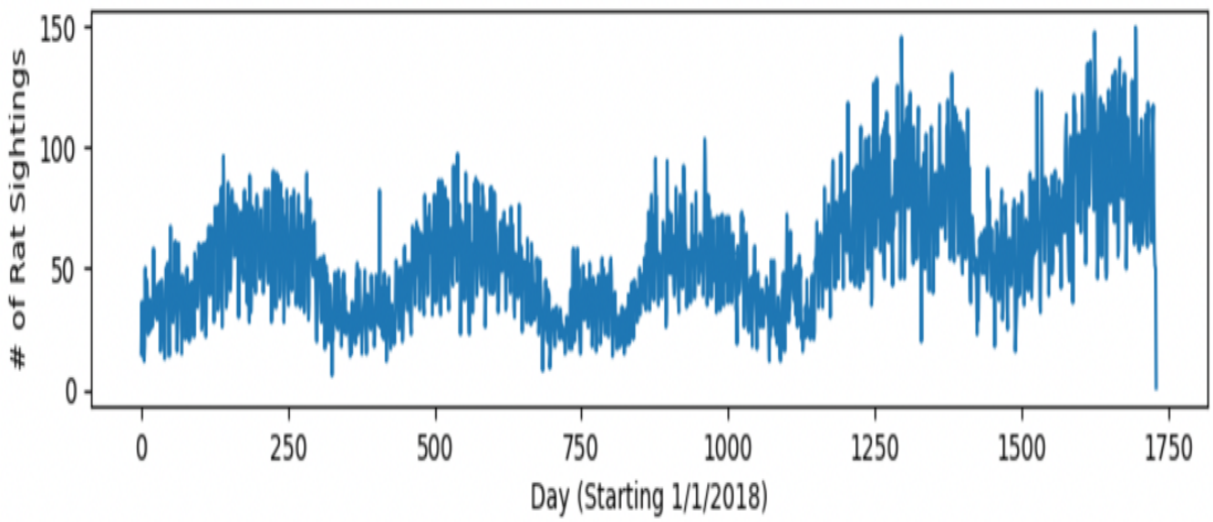
How might we answer the question, *have the number of rat sightings changed over time?* Our independent variable is the *number of rat sightings*, and our dependent variable is *time*; no group differences or pre/post comparisons are necessary for this question.

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 3.2 Time series plot of daily reports of rat sightings in NYC.



From this graph, we can see the following:

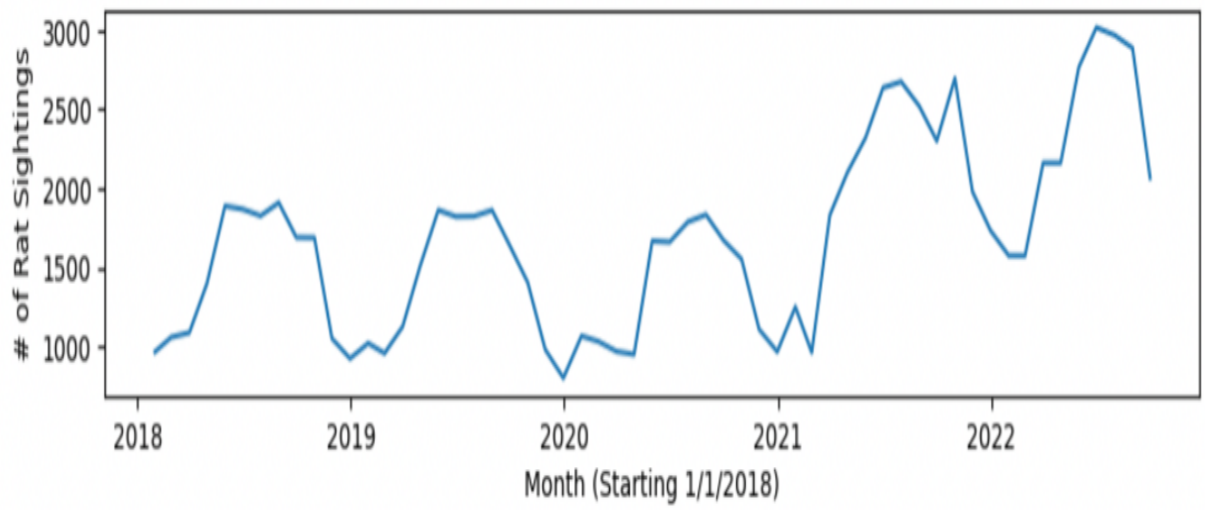
- The number of daily rat sightings has been trending upward in recent years
- There are consistent weekly and monthly seasonal trends in the number of rat sightings

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")).agg(
    ["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 3.3 Time series plot of total monthly rat sightings in NYC.

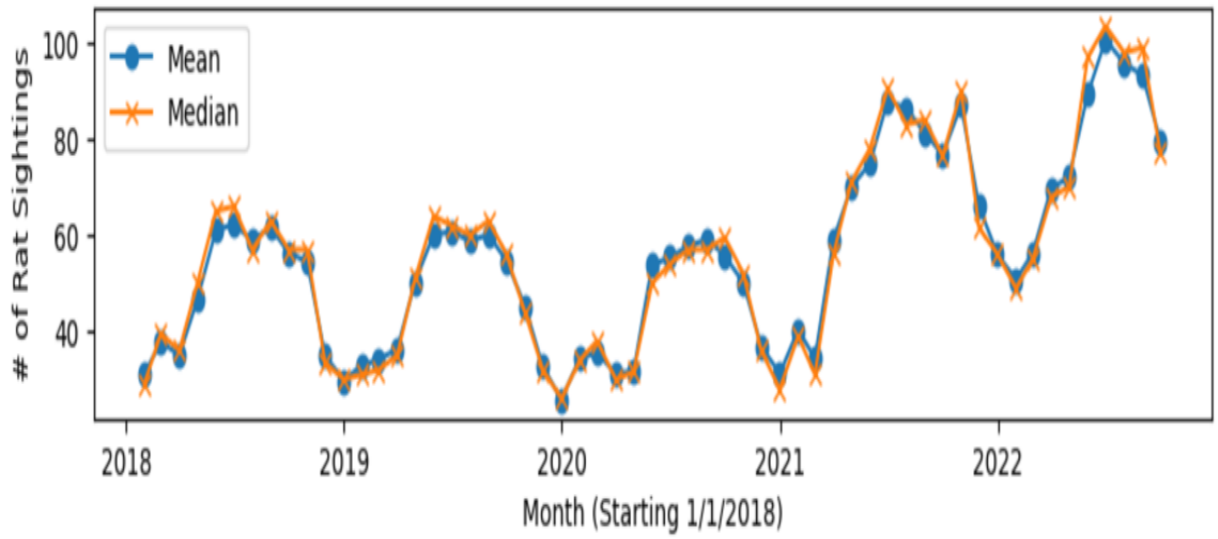


This visual makes it easier for stakeholders to understand the *volume* of rat sightings over time. The weekly seasonality causing the dense spikes has been removed, highlighting the monthly seasonality associated with colder months (approx. November through March) and warmer months (April through October) 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 3.4 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.

Example Case Study: Non-Profit

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 3.5 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

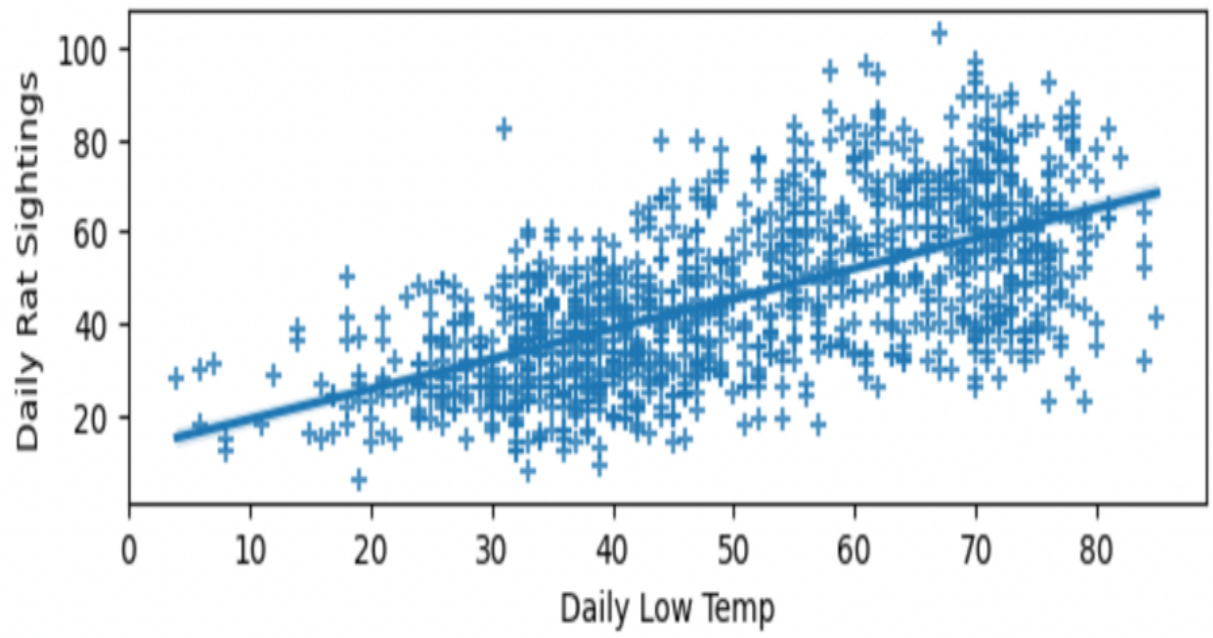
The daily sightings of rats have a strong positive linear correlation with high and low temperatures on the same day. We can see that the high and low temperatures are correlated at nearly `r = 1`, indicating they are likely not independent, and we should select *one* of the variables to highlight the 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 3.6 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.

Example Case Study: Non-Profit

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

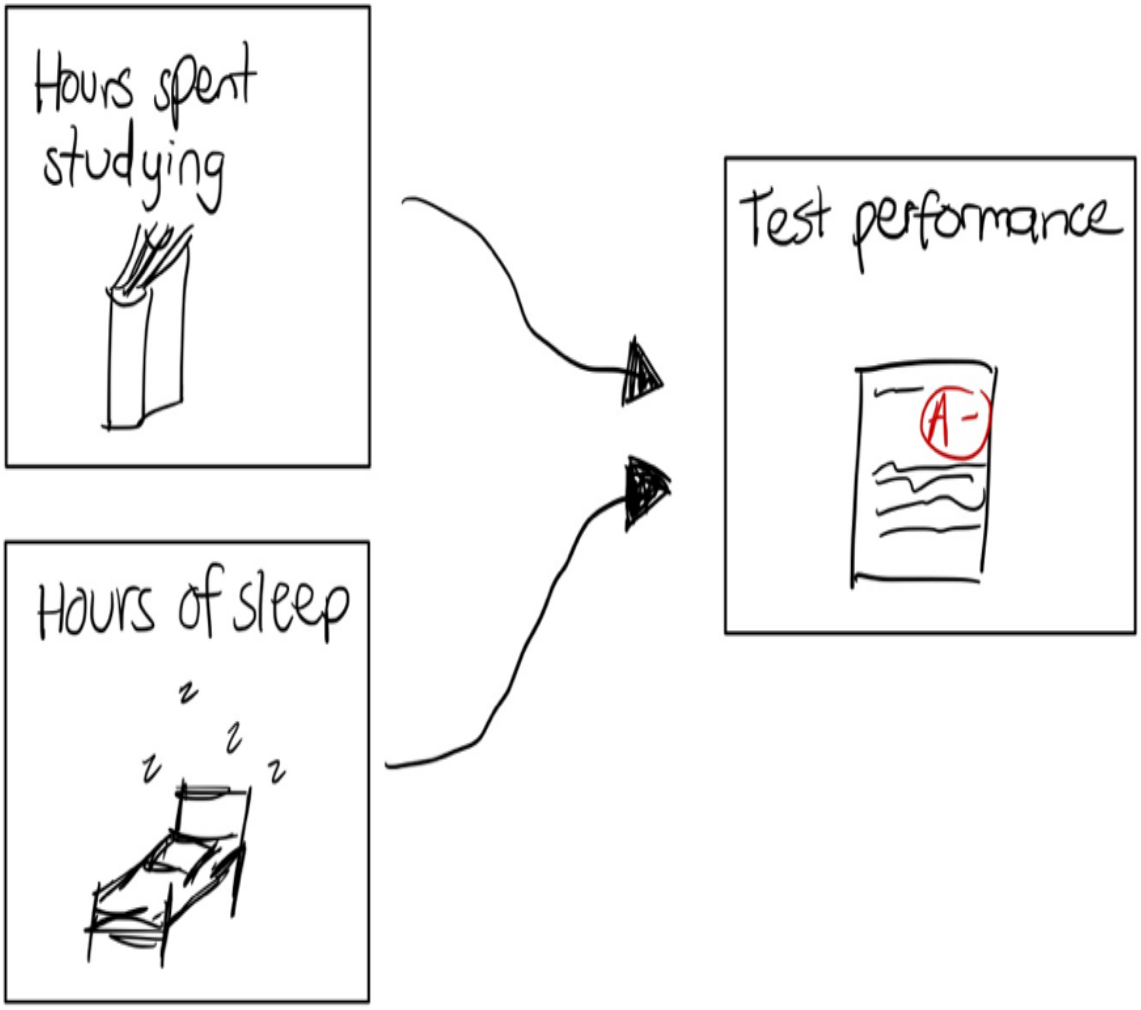
change the dependent variable (the number of rat sightings)?
- Are rat sightings representative of the rat **population** or the **visibility** of the rat population?
- Is the goal to reduce the rat **population** or the **visibility** of the rat population? (We will expand more on this in chapter 6.)

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 3.7 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.

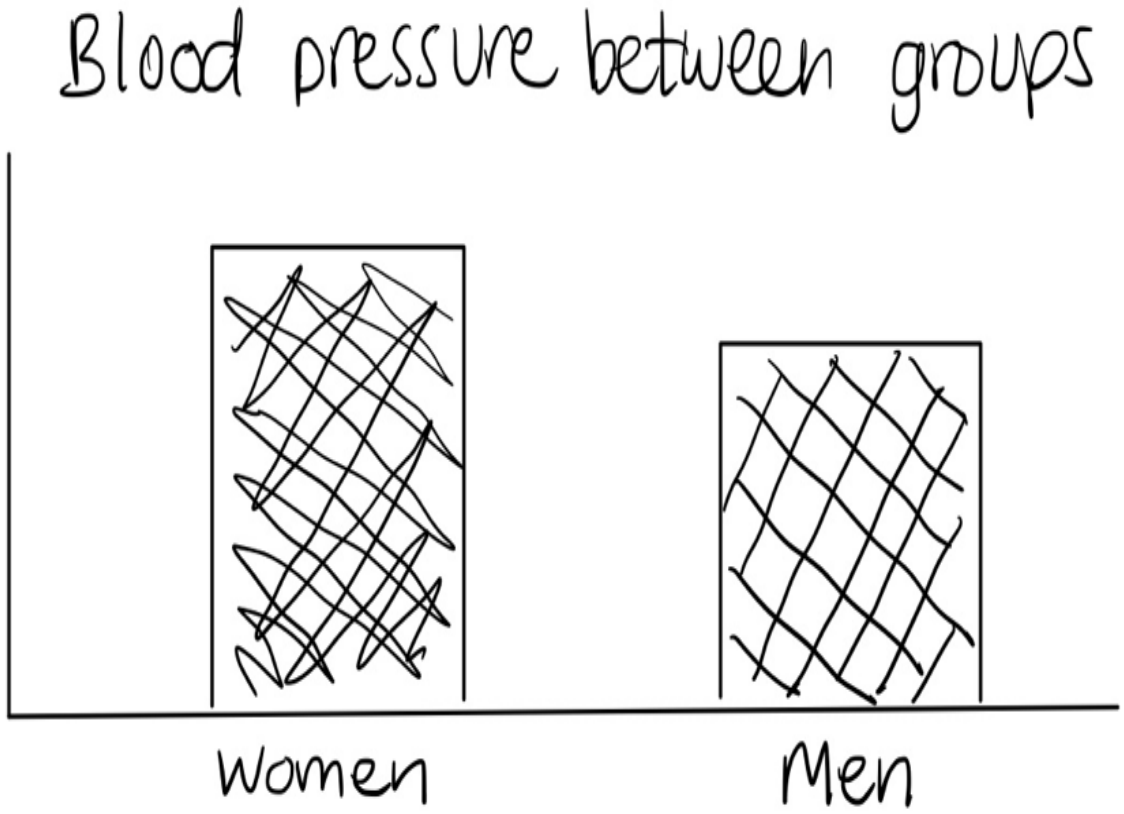
Quasi-Experiments

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 3.8 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

Researchers and analysts typically perform an evaluation using statistical tests with standardized criteria to determine if group differences are statistically meaningful. The types of tests used depend on the *study design* or method of assignment and comparison of groups.

A single dataset collected for an evaluation can include parameters that allow for multiple study designs; however, a single statistical test without interaction effects will typically involve only one design.

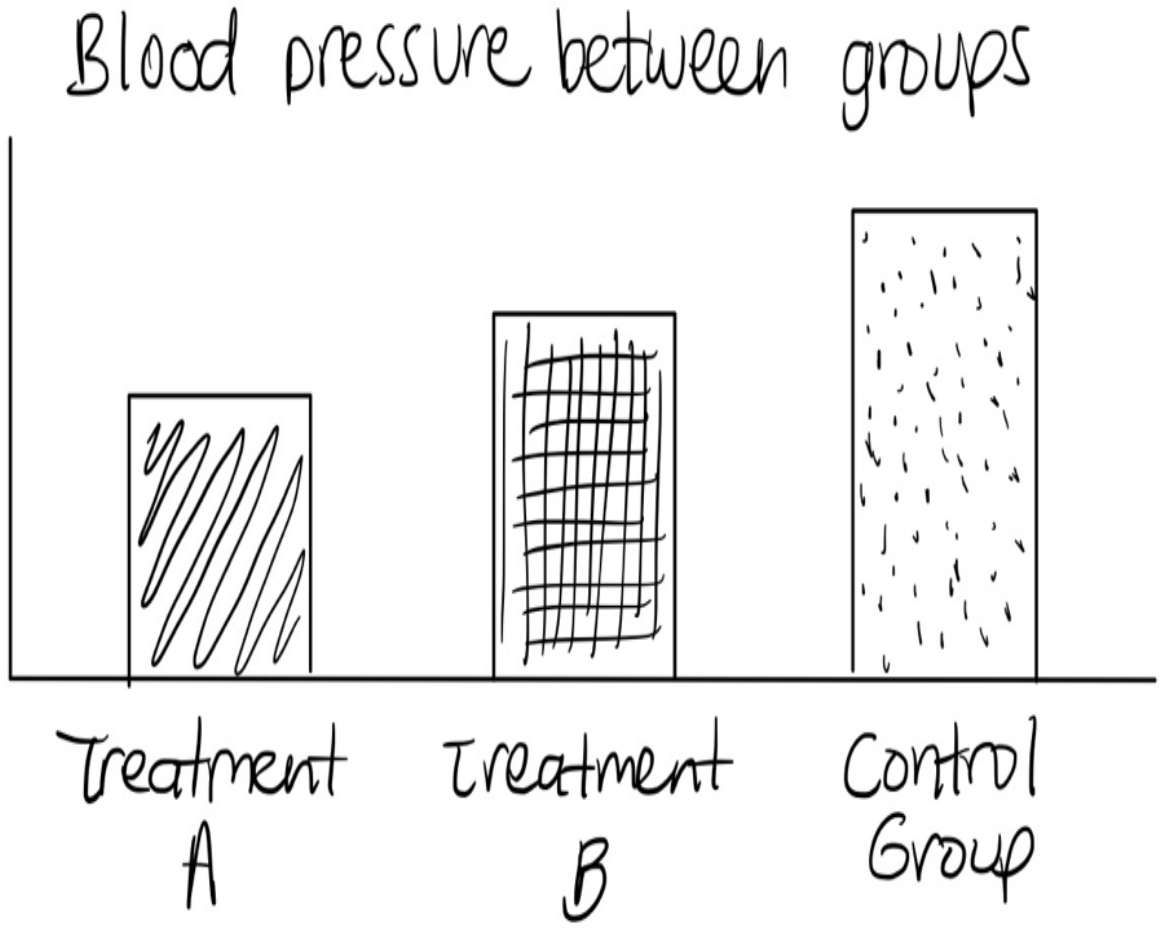
This section will cover some of the most common study designs used in academic, business, and other organizational settings. Keep in mind this list is far from exhaustive; dozens of study designs are used to answer specialized questions in different domain areas. You can expect the designs covered here will cover 90% or more of the use cases you will encounter in your early to middle career as an analyst.

Between-Subjects

A *between-subjects design* compares the dependent variable between *separate groups of participants*. This is the method used in the majority of experiment examples provided. Participants are either randomly assigned or inherently

belong to *mutually exclusive groups*, and statistical tests compare a single data point for the dependent variable between members of each group.

Figure 3.9 Between-subjects comparisons evaluate differences between mutually exclusive characteristics or assignments.

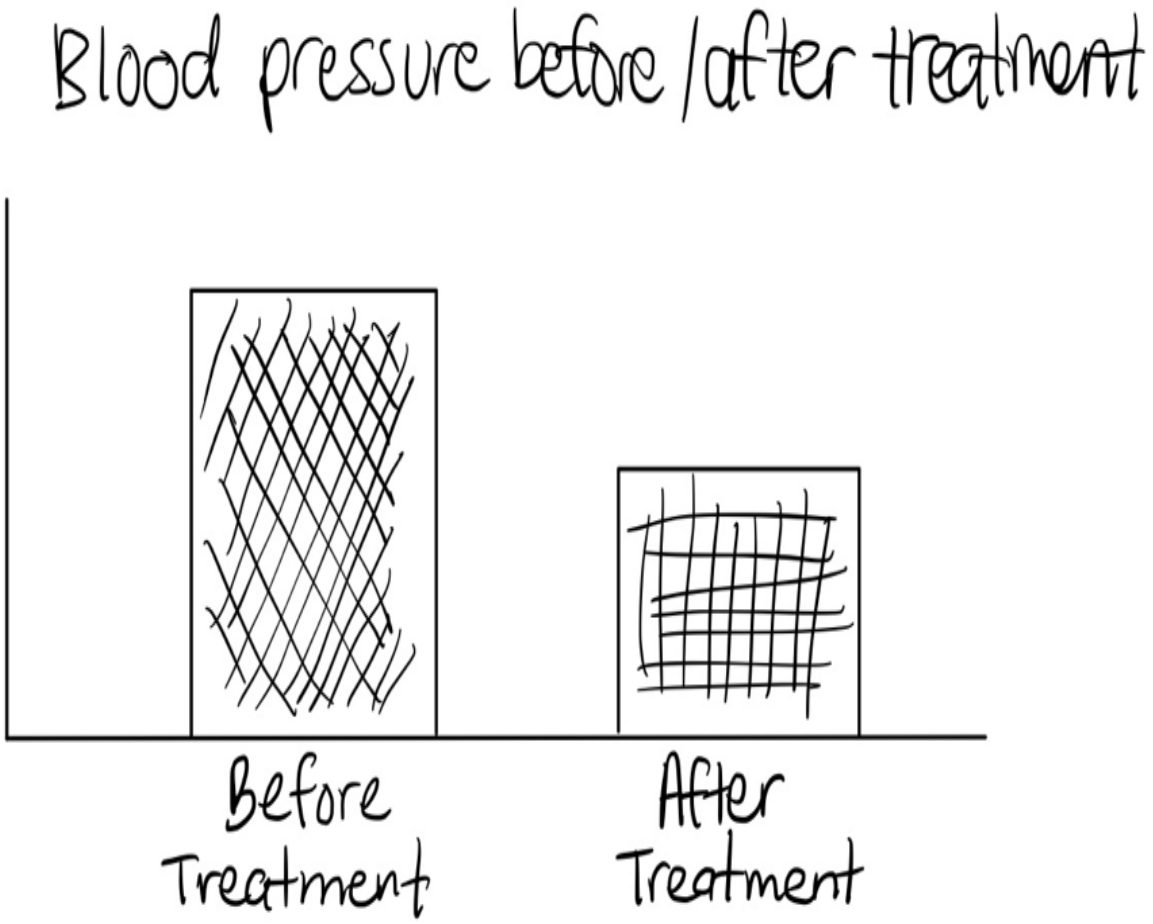


Within subjects

A *within-subjects design* or *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 3.10 Within-subjects comparisons evaluate differences before and after treatment.

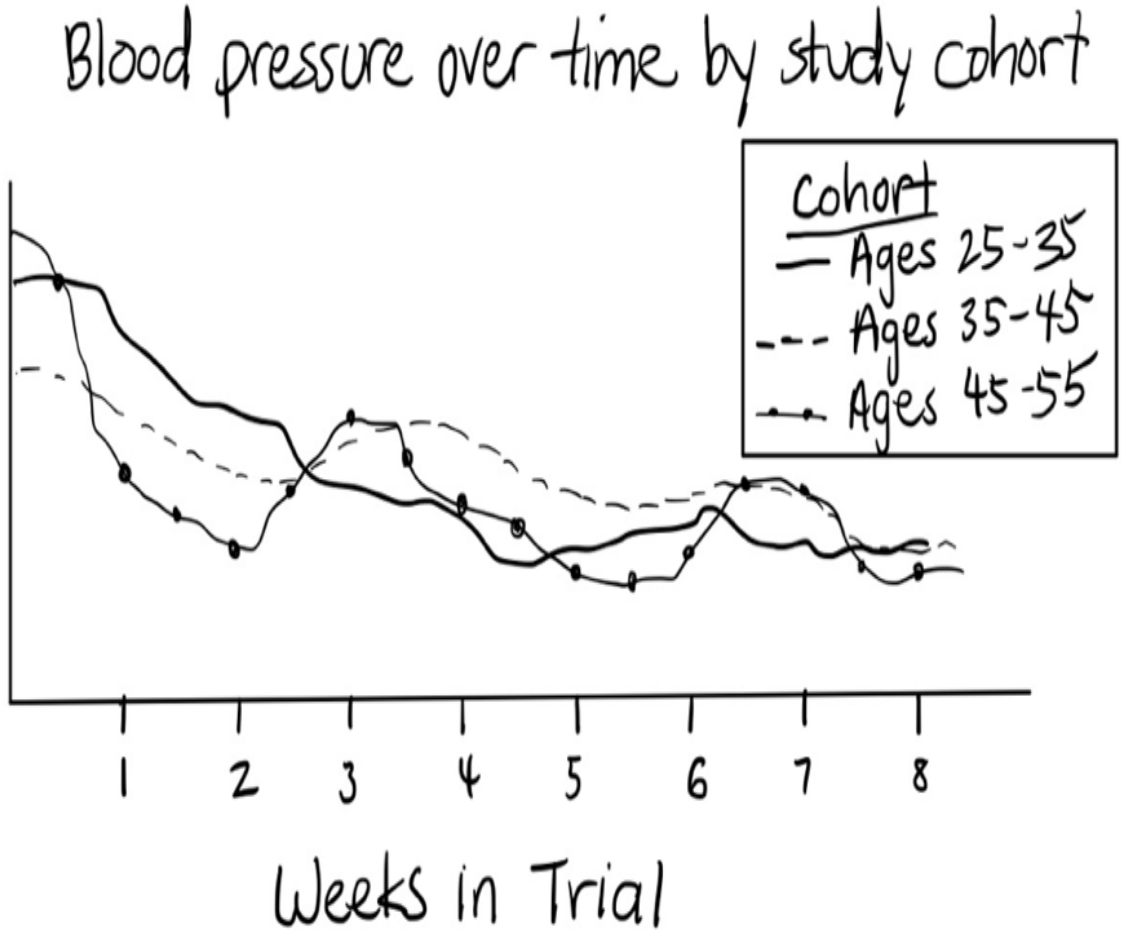


Cohort Comparisons

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 by characteristics that change over time (e.g., age or the year they subscribed

to a service) or static characteristics (e.g., school attended, subscription tier). This is a combined within-subjects and between-subjects approach to designing a study. This design also excludes random assignment and instead seeks to identify whether measurable trend differences occur over time between mutually exclusive groups.

Figure 3.11 Cohort comparisons evaluate meaningful cohort groups over time.



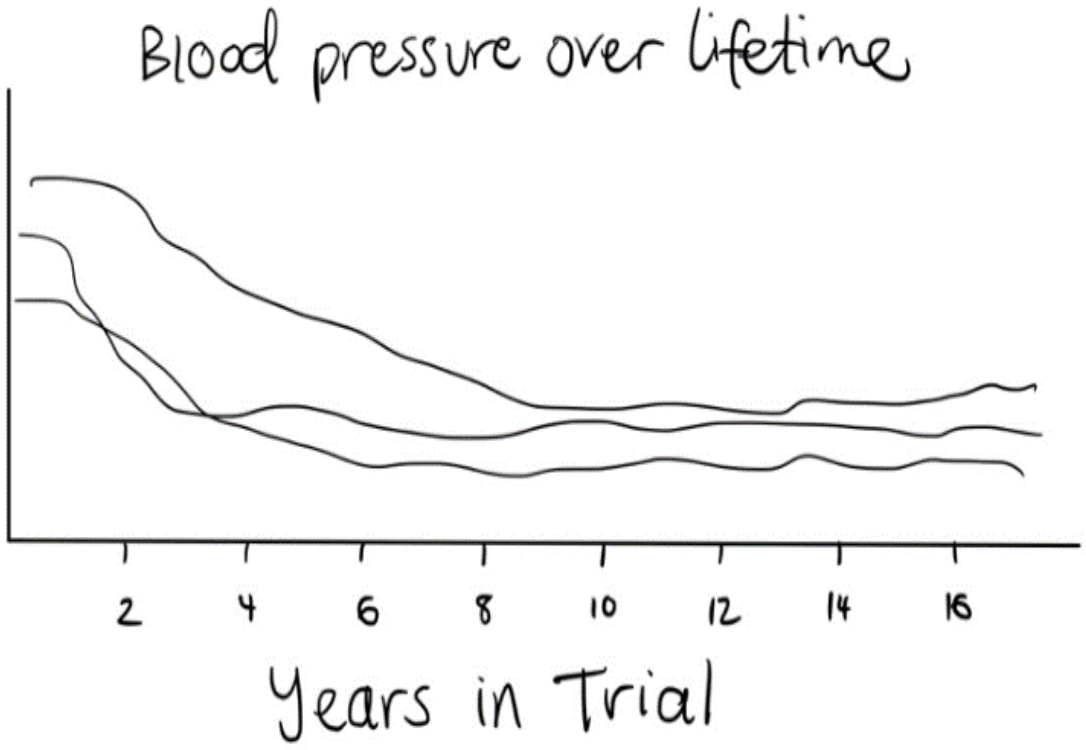
Longitudinal Comparisons

The previous two methods discussed are types of longitudinal analysis, though many longitudinal studies assess participants repeatedly over months, years, or

lifetimes. A longitudinal study design evaluates participants over multiple time points, usually across a long span of time. These studies can include random or cohort assignments and perform a between-subjects comparison and an expected within-subjects comparison.

Longitudinal studies are often assessed using repeated measures univariate tests, and others are assessed using specialized statistical methods tailored to the question and expected trends in the data (e.g., survival analysis, individual growth curve analysis).

Figure 3.12 Longitudinal comparisons evaluate participants over months, years, or even lifetimes.



Example Case Study: Non-Profit

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

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.

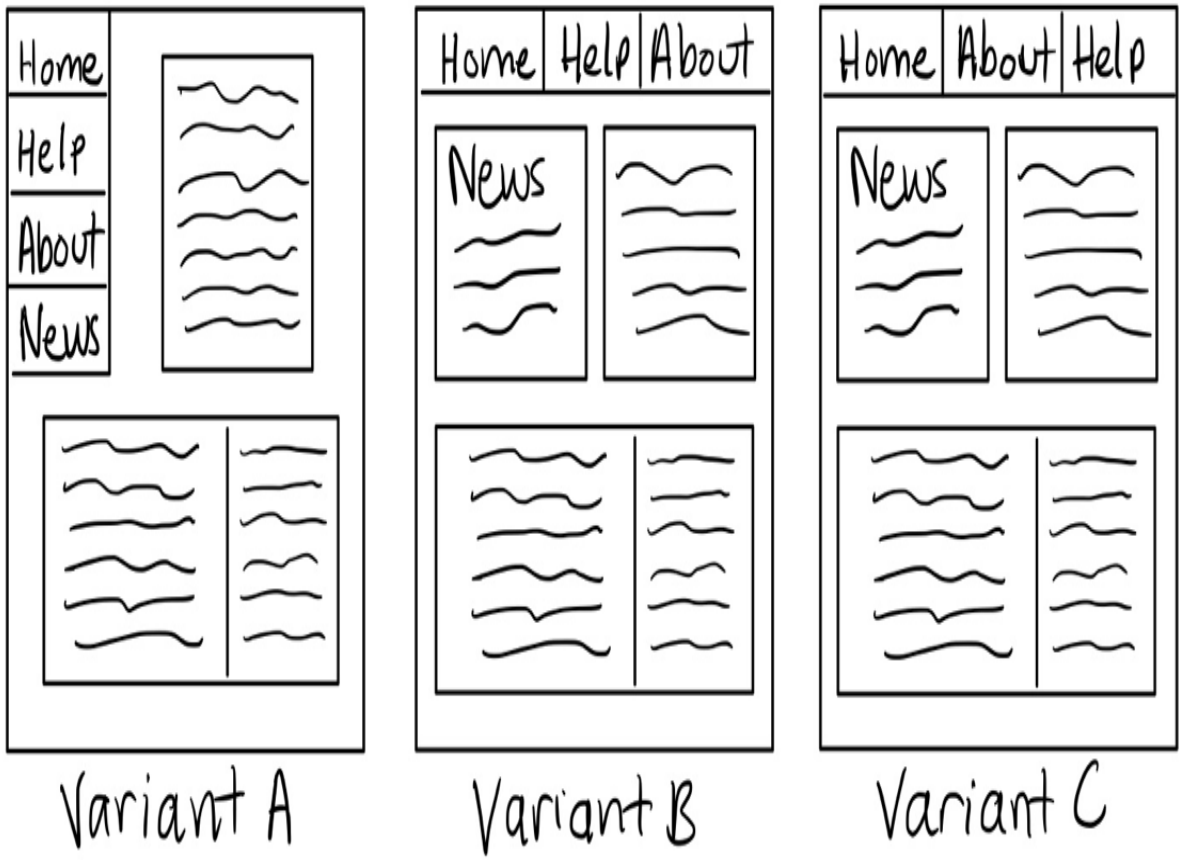
While applied research programs don’t have a primary goal of contributing to basic theories within the field of study, they will usually affect an accrual of knowledge about distinct subsets of a population (e.g., children in a geographic area, second-generation immigrant youth, persons with a specific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

3.3.2 A/B Testing

In business settings, experiments are often designed as *A/B tests*. An A/B test is a method of 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 3.13 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.

Example Case Study: Non-Profit

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.
- Data can be evaluated as **descriptive information**, **correlations** between variables, or **causal/predictive analyses to test a hypothesis**. The hypothesis and design of your study will dictate which of these methods are most appropriate.
- 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** are standard in academic settings, **A/B testing programs** are common in businesses and product analytics, and

program evaluations are typical in non-profit and government environments.

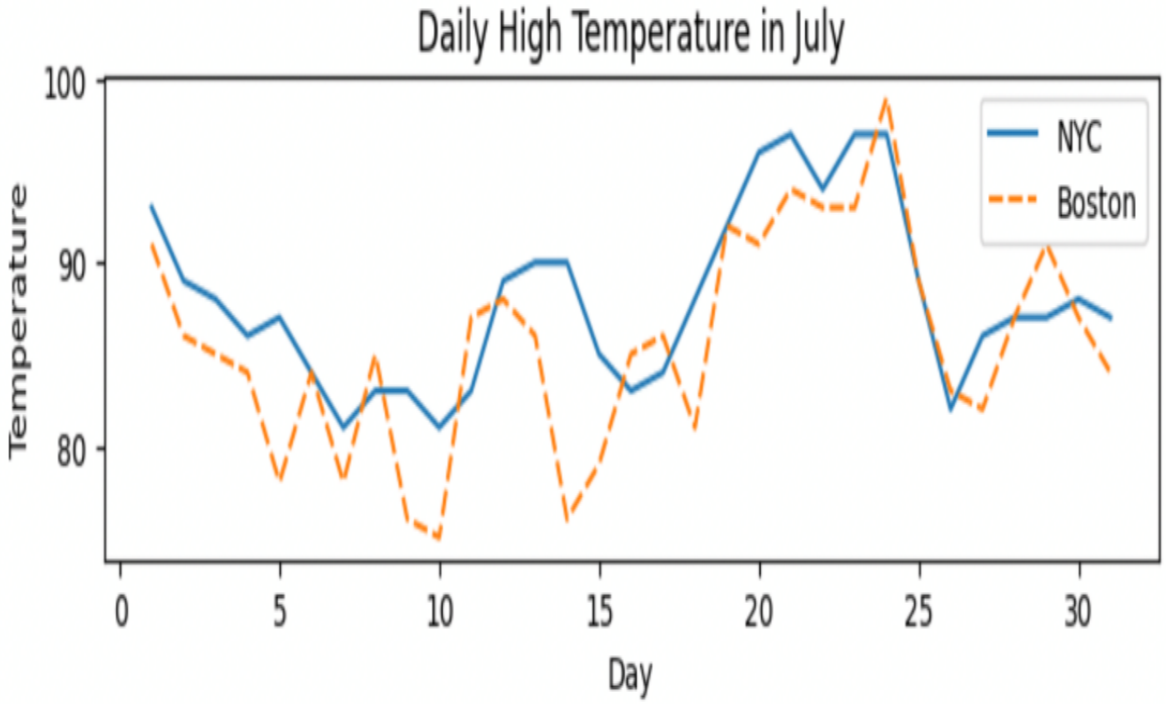
4 The Statistics You (Probably) Learned: T-Tests, ANOVAs, and Correlations

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:

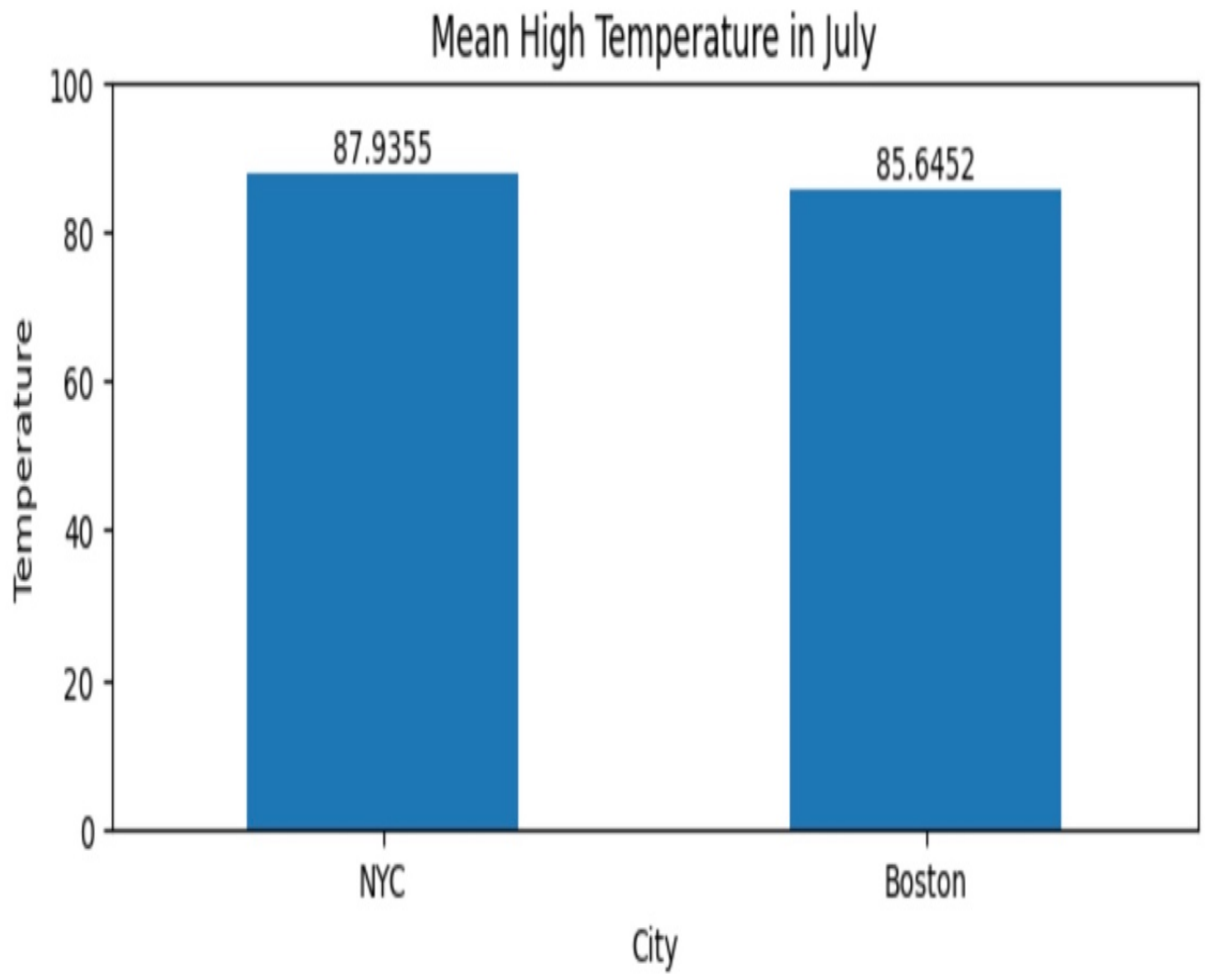
Figure 4.1 Comparison of temperatures in July between New York City and Boston



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
```

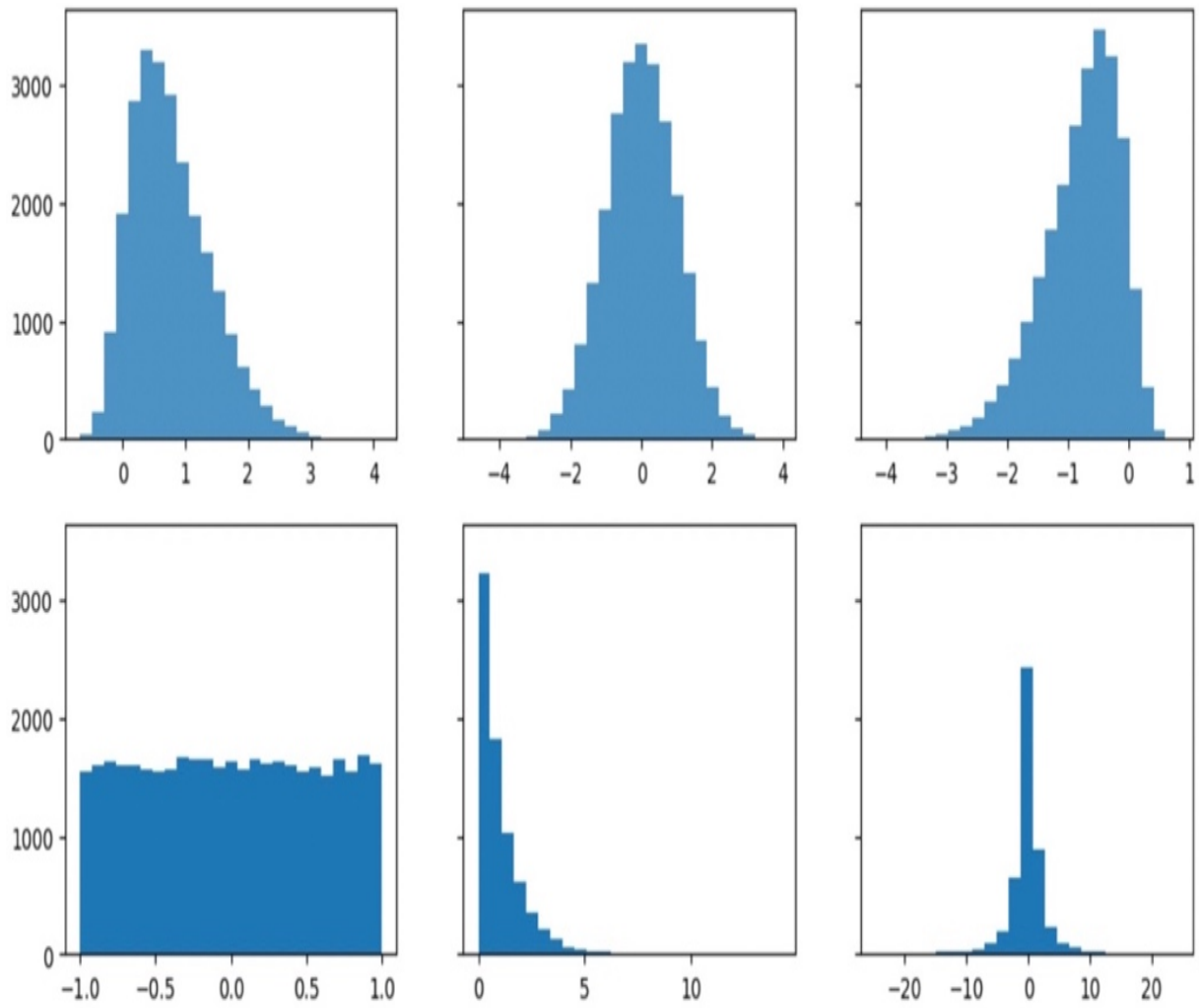
If you had only seen Figure 4.2 and not Figure 4.1, you would not know about the range, fluctuations, and heat waves depicted in the raw dataset. By visually inspecting the list of 31 values, you can see that an average of 87.9 degrees provides a limited view of the dataset. The temperature ranges between 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

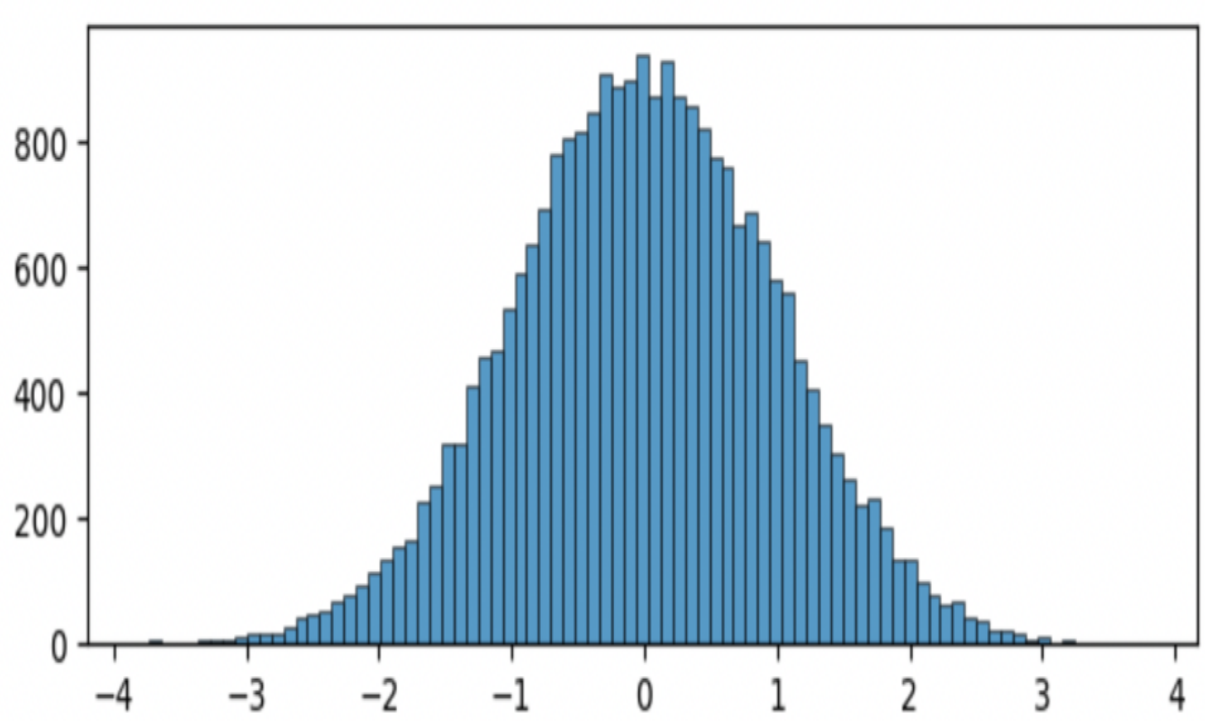
pros and cons of using measures as part of the *metrics* you report in chapter 6.

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.

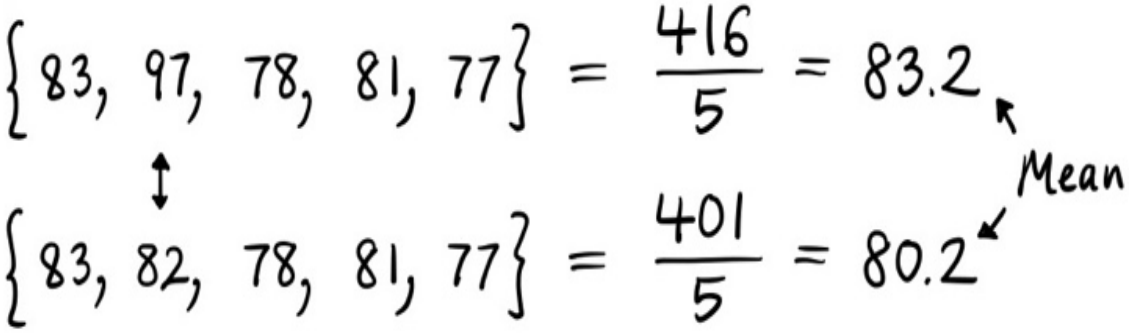
The *arithmetic mean* is the most common and widely used statistic for summarizing numerical data. Analysts will often start their work by taking means of their data. Using this measure has a *lot* of benefits:

- 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.

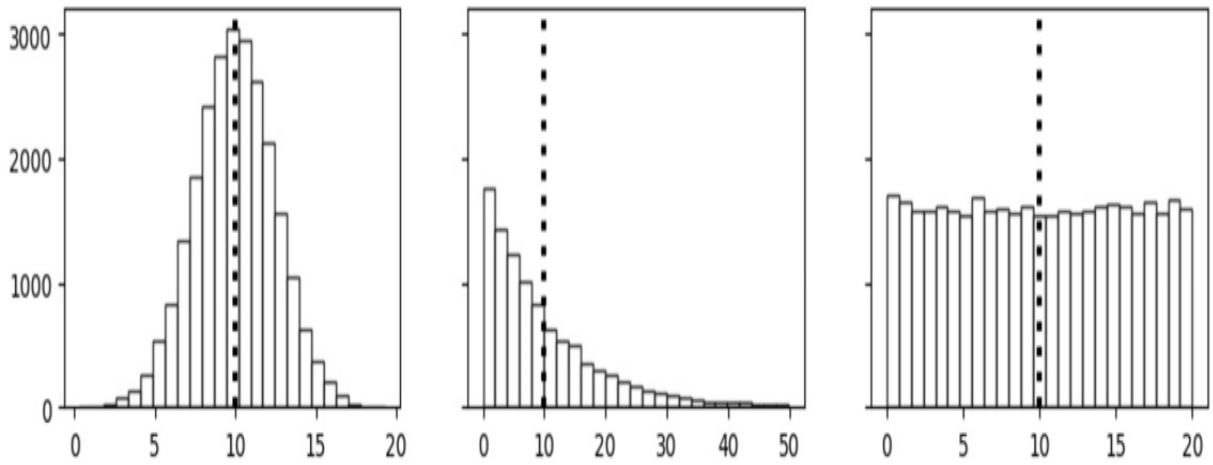
Figure 4.5 A mean calculation is highly sensitive to skewed data and outliers. The mean noticeably decreases when the highest outlier value of 97 is replaced with a value closer to the rest of the set.



- 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.

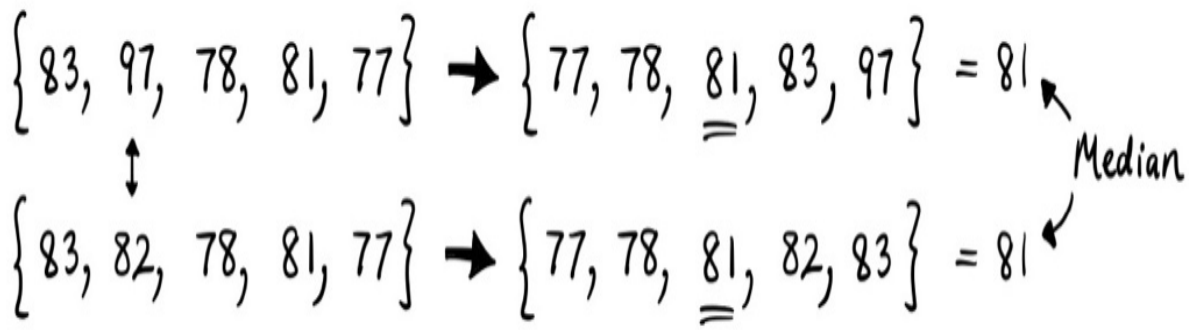
Figure 4.6 Three distinct distributions with an identical mean of 10. The interpretation of the mean or “average” is very different for each distribution.



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.

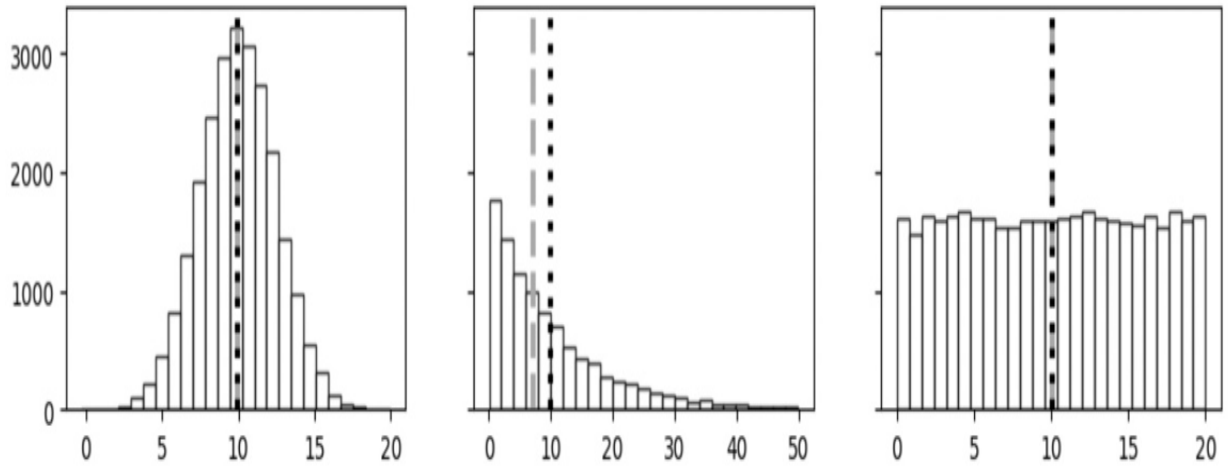


As with the mean, there are key limitations to note about the median:

- 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.

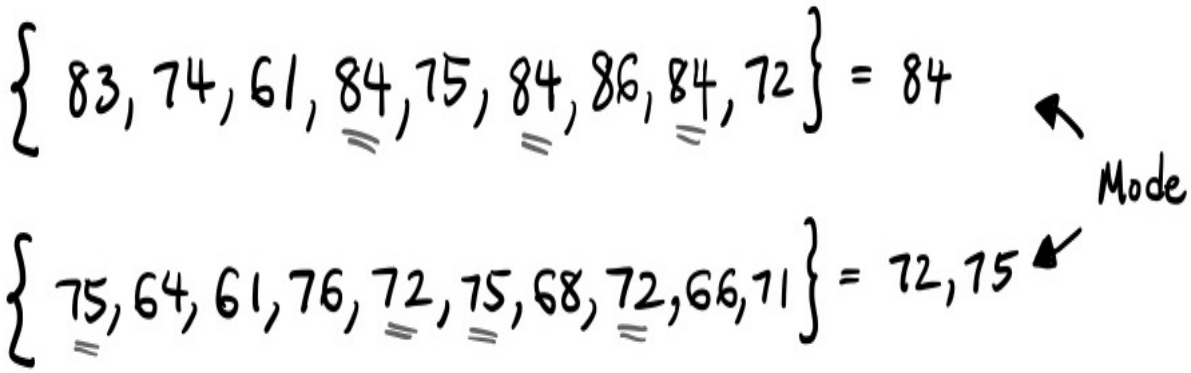
When preparing a report, you often choose between the mean and median to share with stakeholders based on the measure that provides the greatest clarity and value. Outside of direct stakeholder reporting, observing the differences between the mean and median indicates that your dataset is likely skewed 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.



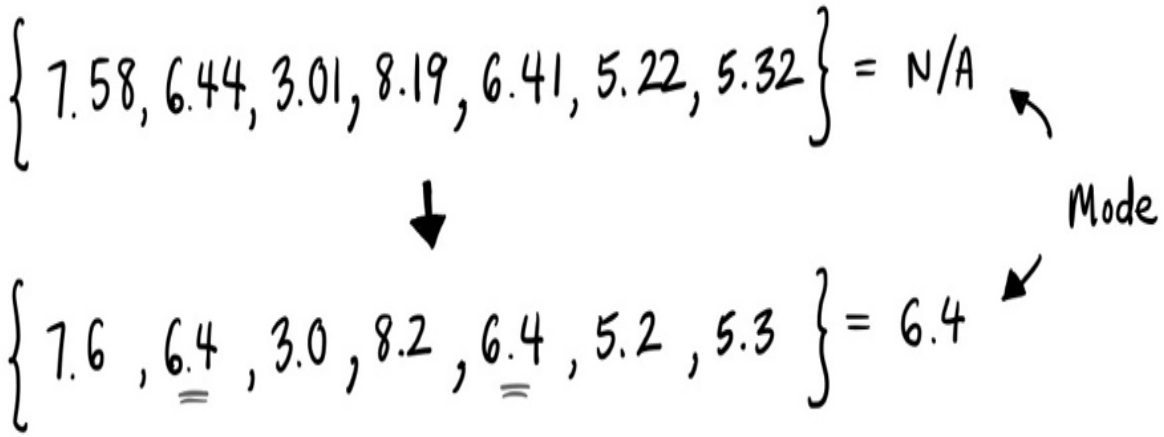
The *mode* is the most frequent value that occurs in your dataset. In practice, it is leveraged far less often than the mean and median and requires more context to appropriately explain its importance in reports.

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



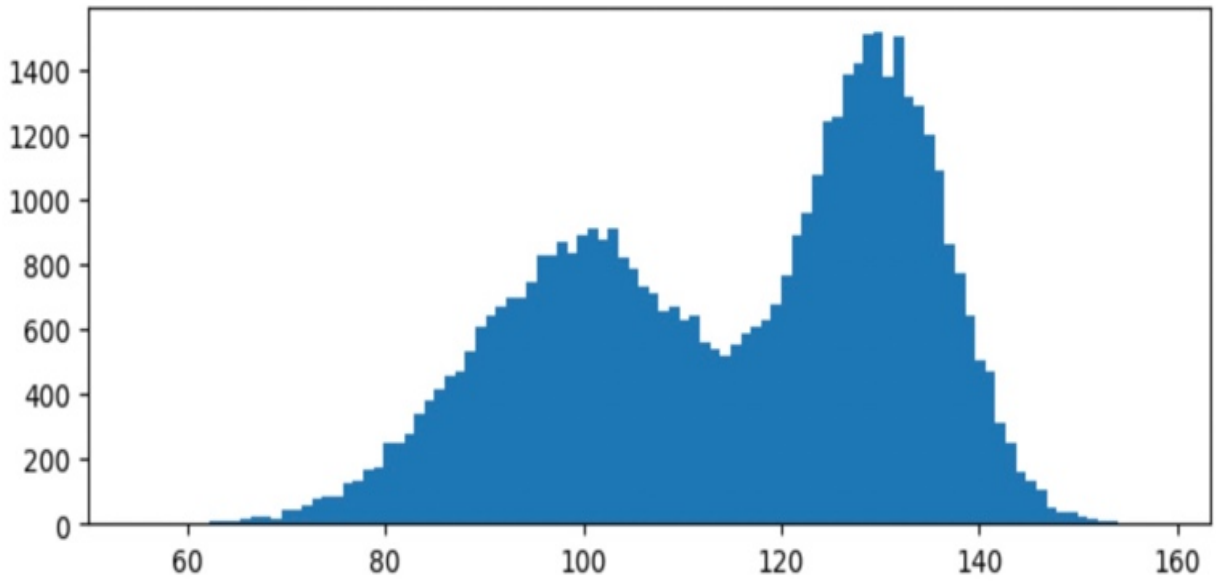
You may sometimes need to round or bin values to derive a meaningful mode. When testing a rounding calculation on a series with a large range or floating-point continuous data, your choice of bins or decimal point to round to can drastically change your outcome.

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



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.

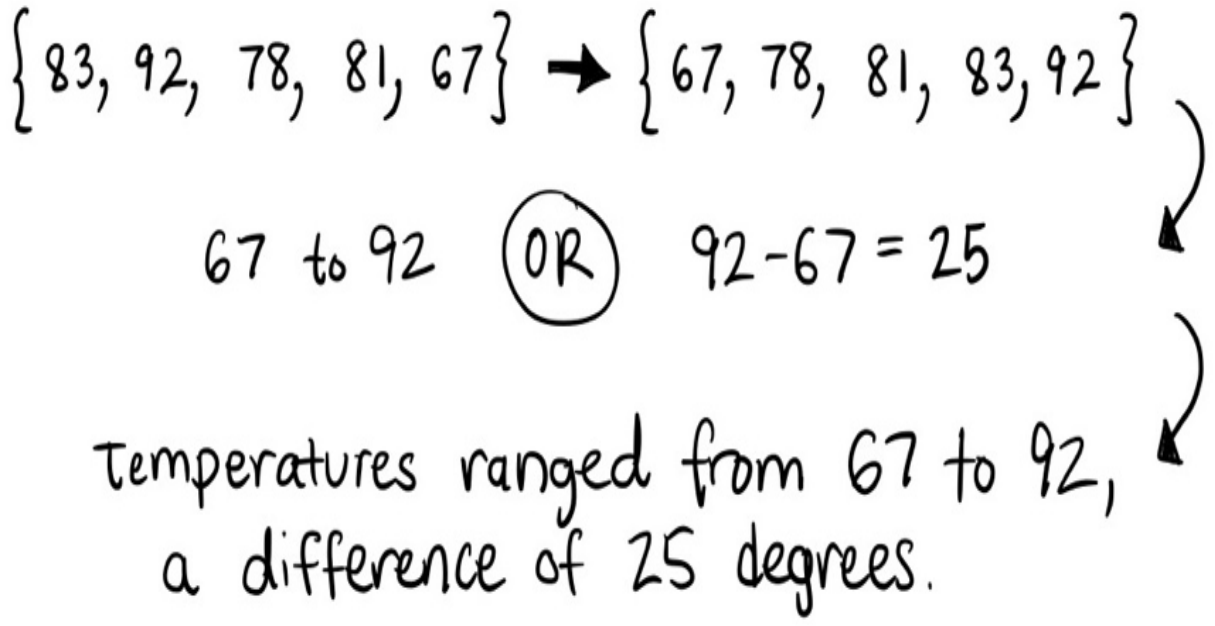
Naomi begins the summary of her descriptive statistics for a paper to be submitted for peer review with the mean and median for both measures and the mode for the number of sleep interruptions.

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.

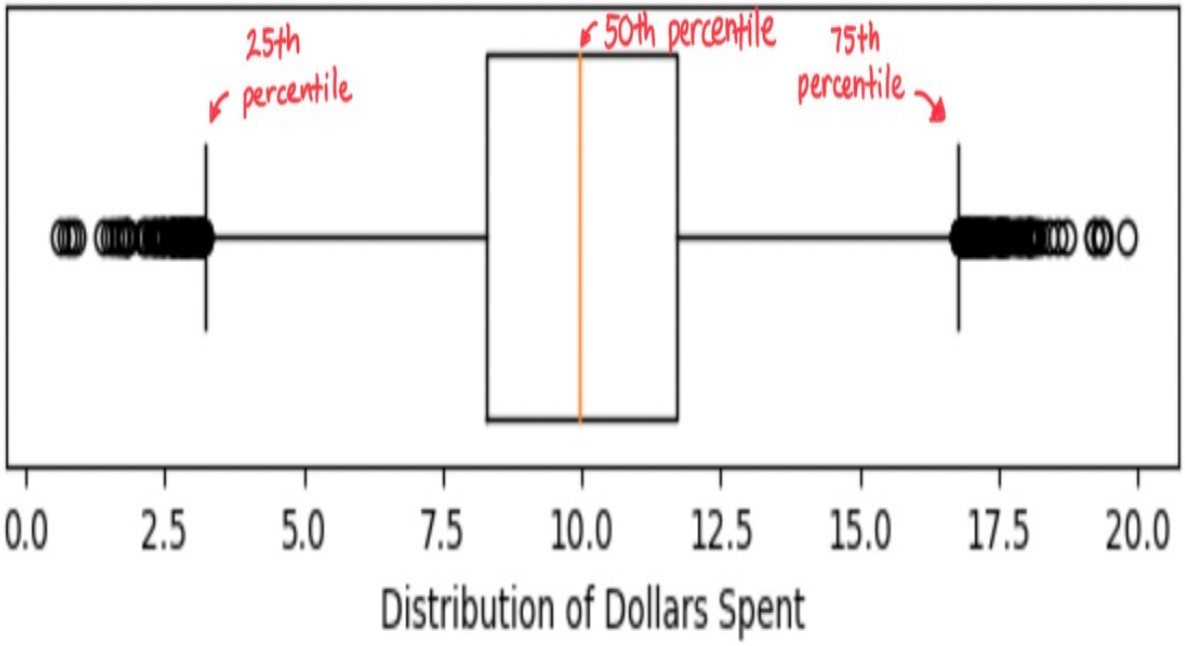


In addition to the full range, the interquartile range (IQR) shows the spread of the middle 50% of your data points from the 25th to the 75th percentile. This can be compared to the overall range to better describe the spread of your dataset

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 5th/95th percentiles in the whiskers by default. Values outside of the whiskers are typically treated as outliers.



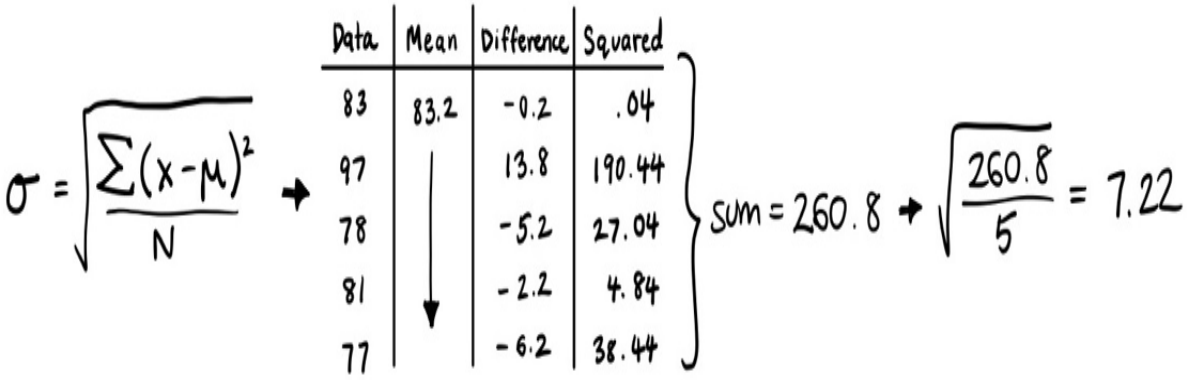
Effectively communicating the results of a range calculation requires appropriate context to clarify its importance to your stakeholders in addition/in place of other measures. If you decide it's valuable to include in your findings, you can consider contextualizing it with statements such as the following:

- *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.

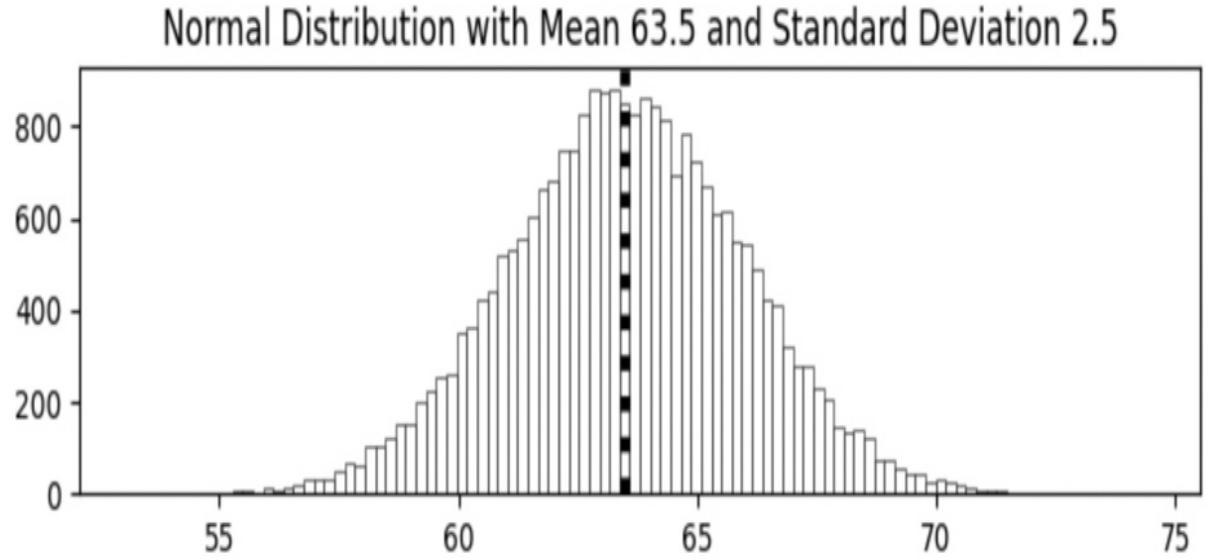


If you have a mean and standard deviation and assume your data is normally distributed, you can easily approximate the shape of the dataset. Similar to the range and IQR, the standard deviation can be used as a coordinate system to estimate the proportions of data points between two values. To demonstrate, 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
Deviation {sd}")
```

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.



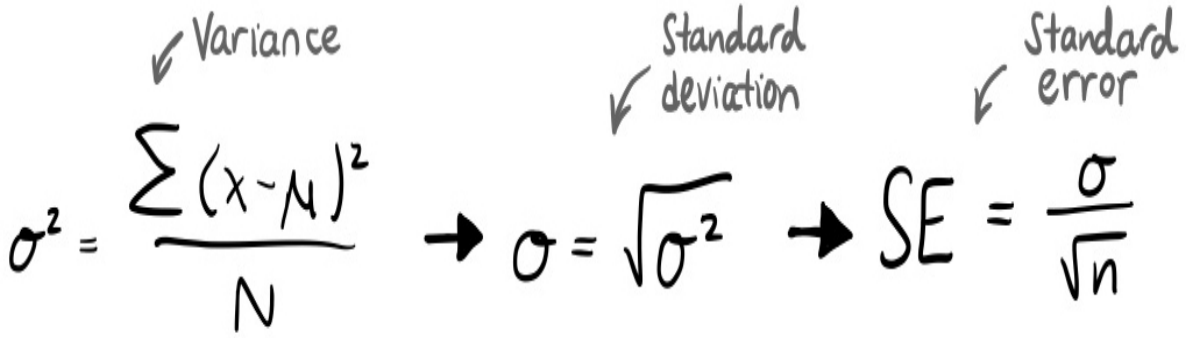
In peer-reviewed papers and technical reports, the standard deviation and the mean are almost always included in the summary statistics. If you include the standard deviation in your reporting to less technical stakeholders, you will likely need to provide a layperson’s explanation to minimize confusion.

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.

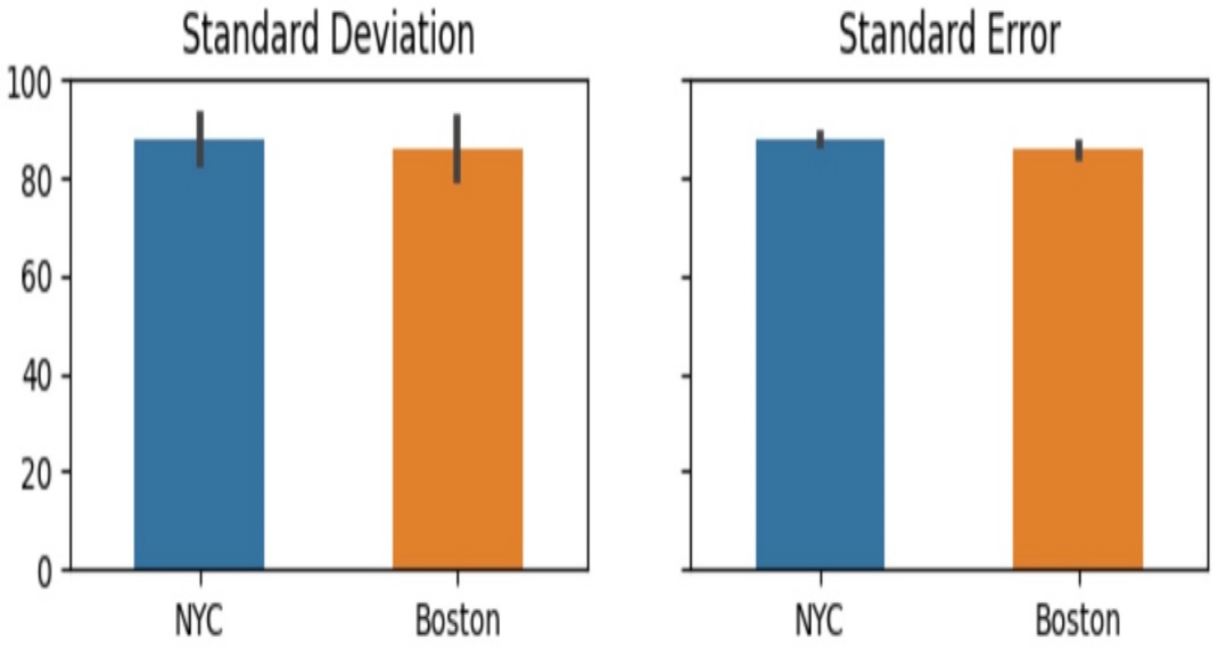


The standard error is a common choice for augmenting visualizations such as bar graphs to add context on variability within/between groups. It's an option in the `seaborn`

`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

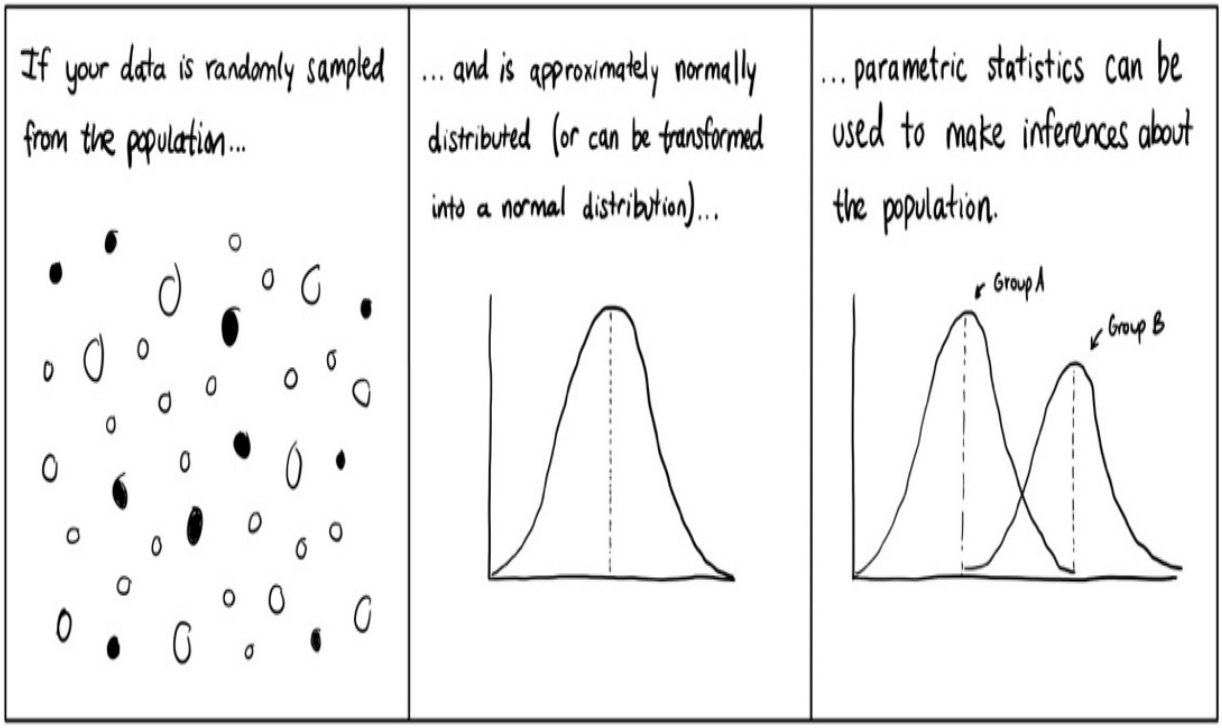
* p<.05
** p<.01

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

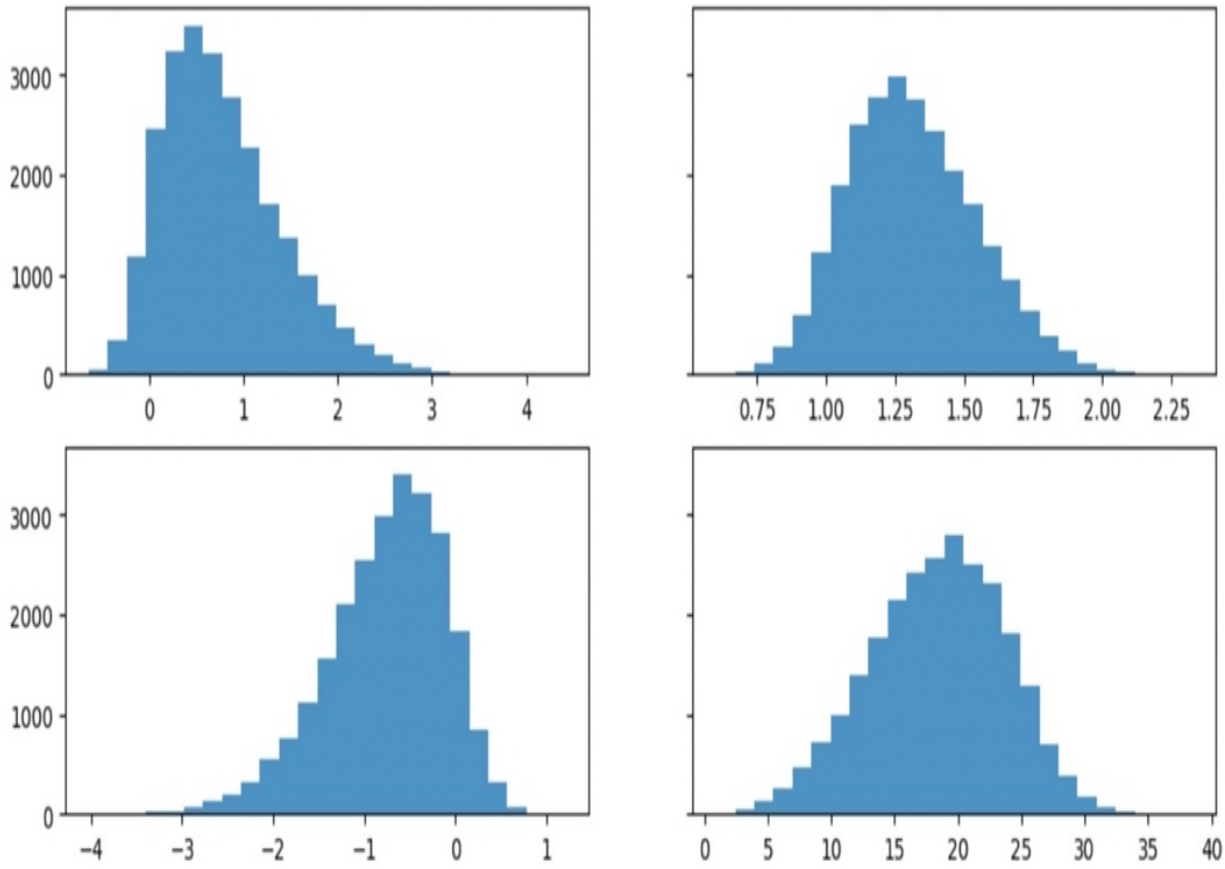
In addition to the assumptions of the measures of central tendency and variation discussed in the previous section, parametric statistics have strict assumptions about the shape of the data within and between groups. Meeting these assumptions is *necessary* for making accurate inferences about your data.

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.

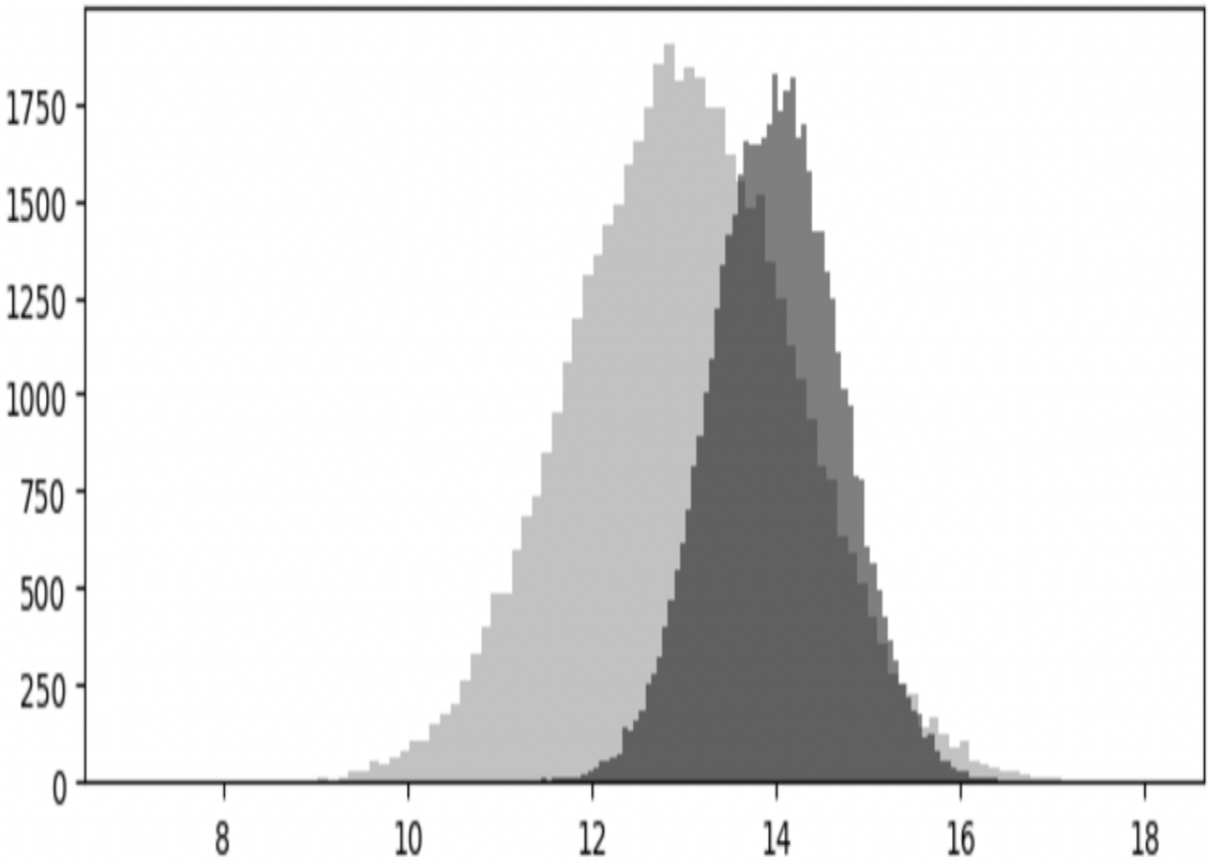
The second assumption is the *independence* of data points in your dataset. Unless otherwise indicated with the statistical test you use (e.g., a repeated measures t-test or ANOVA), the probability of events in your dataset are assumed *not* to impact the probability of other events. In practice, a lack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

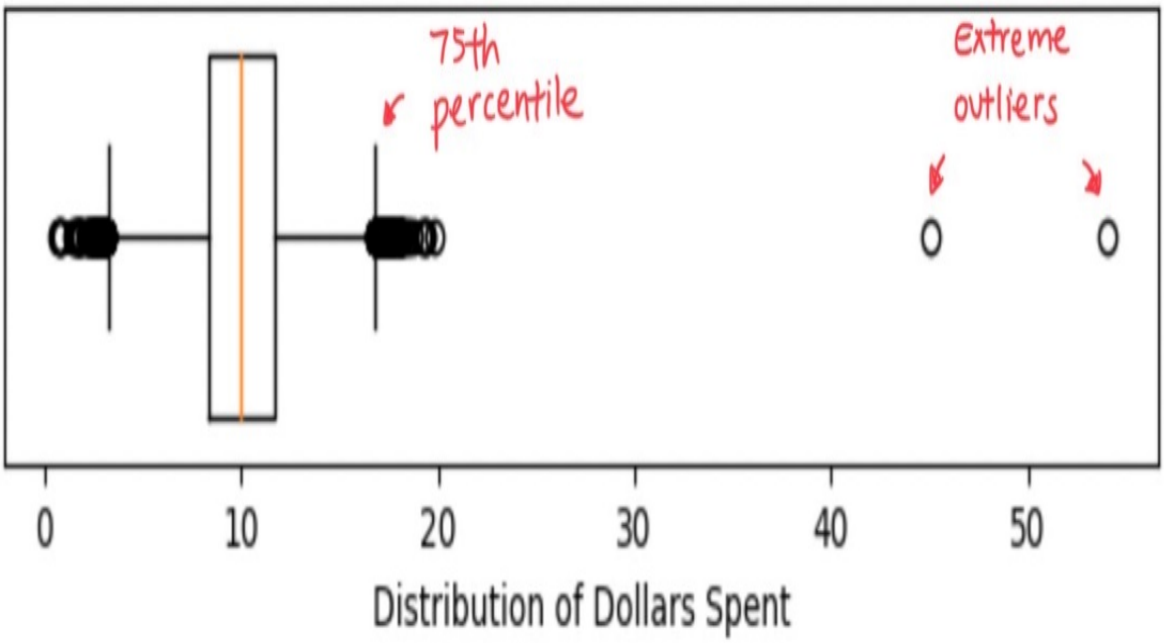


If you determine that your samples have unequal variances, you can use adjusted versions of t-tests and ANOVAs (Welch's tests) that are more robust to violations of this assumption. Non-parametric tests for group comparisons may also be better choices for your work.

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:

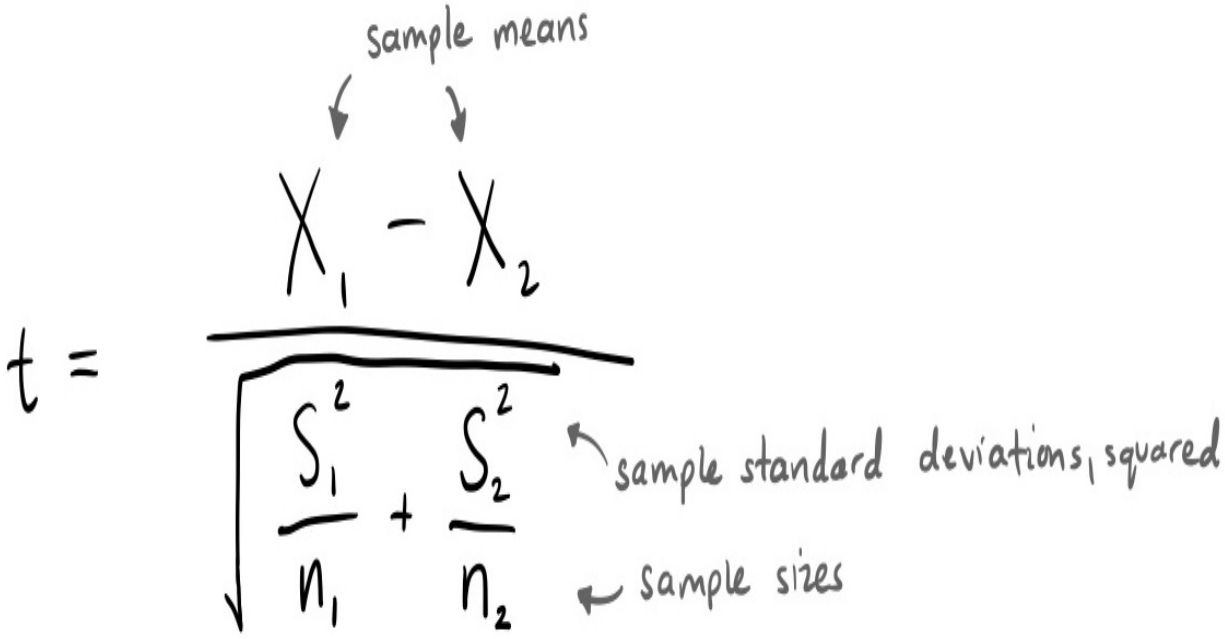
Table 4.1 Abbreviated t-distribution table showing that increasing the degrees of freedom has diminishing returns on the t-critical values at each alpha level/p-value threshold.

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



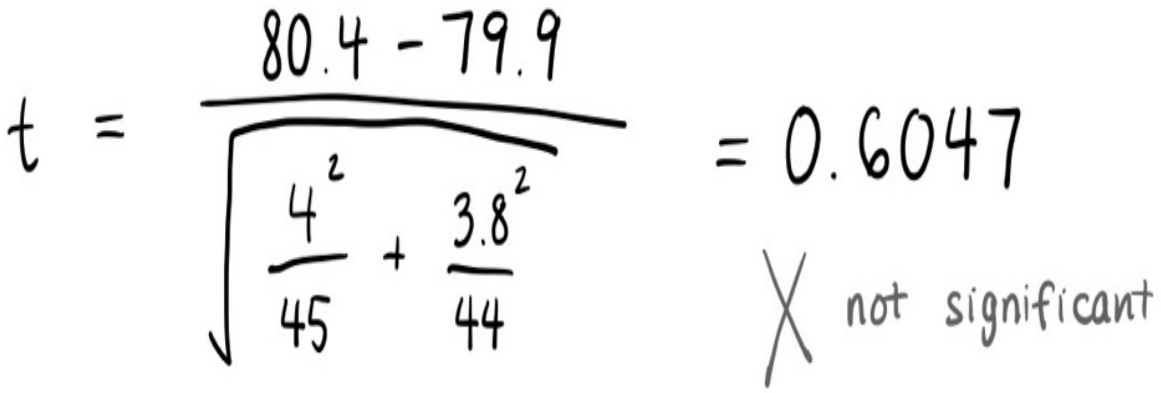
When the sample sizes n_1 and n_2 increase, the size of the overall t-value increases with no other changes to the mean or standard deviations. Let's take two samples with the following summary information:

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

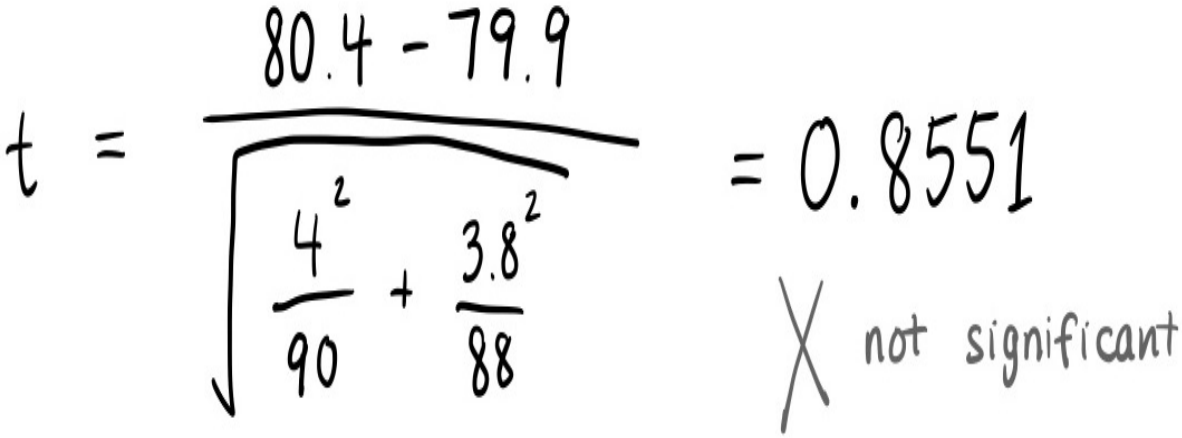
If you calculate the t-value for these two groups, your t-value is far below the critical threshold at the current sample size.

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



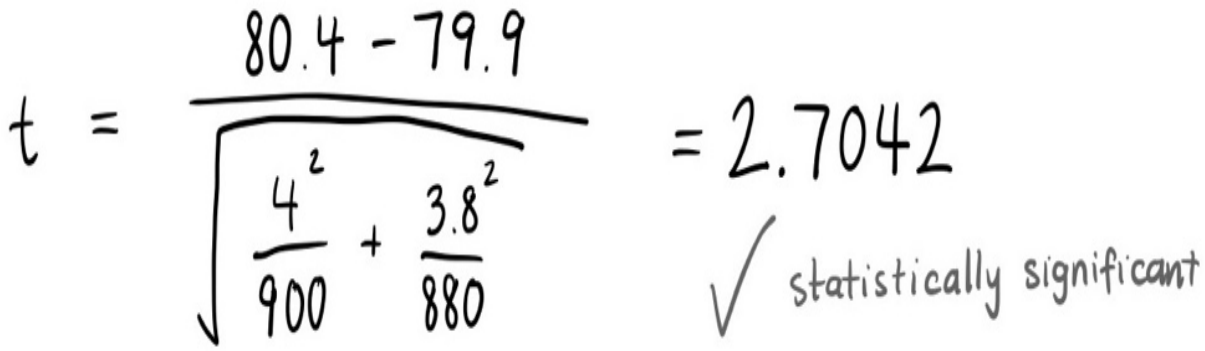
If you double the sample size for each group to 90 and 88, respectively, you get the following result:

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



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.



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.

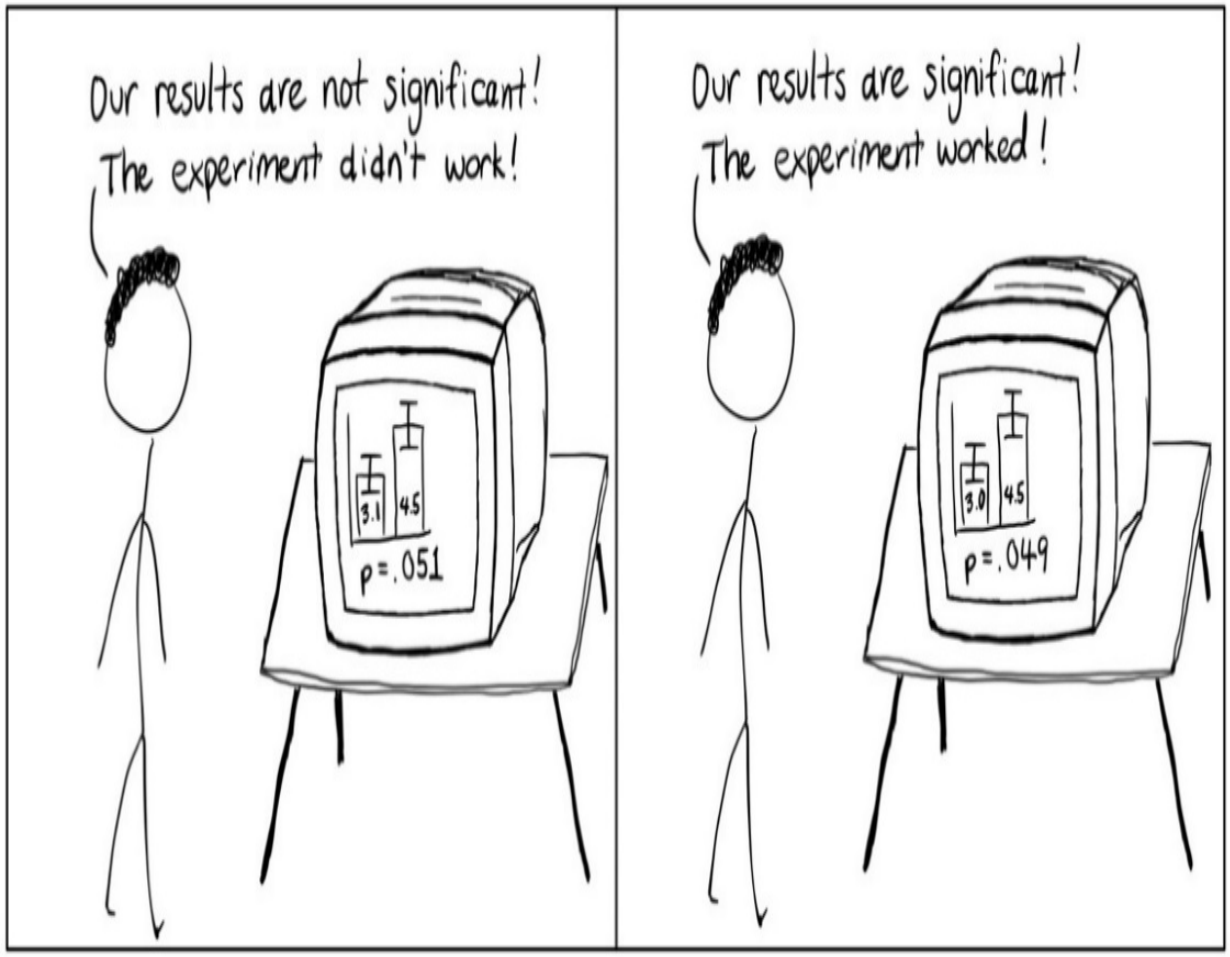
The p-value is **NOT** defined as any of the following, though you may frequently encounter these interpretations:

- 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 20th 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 ? `.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 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=.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:

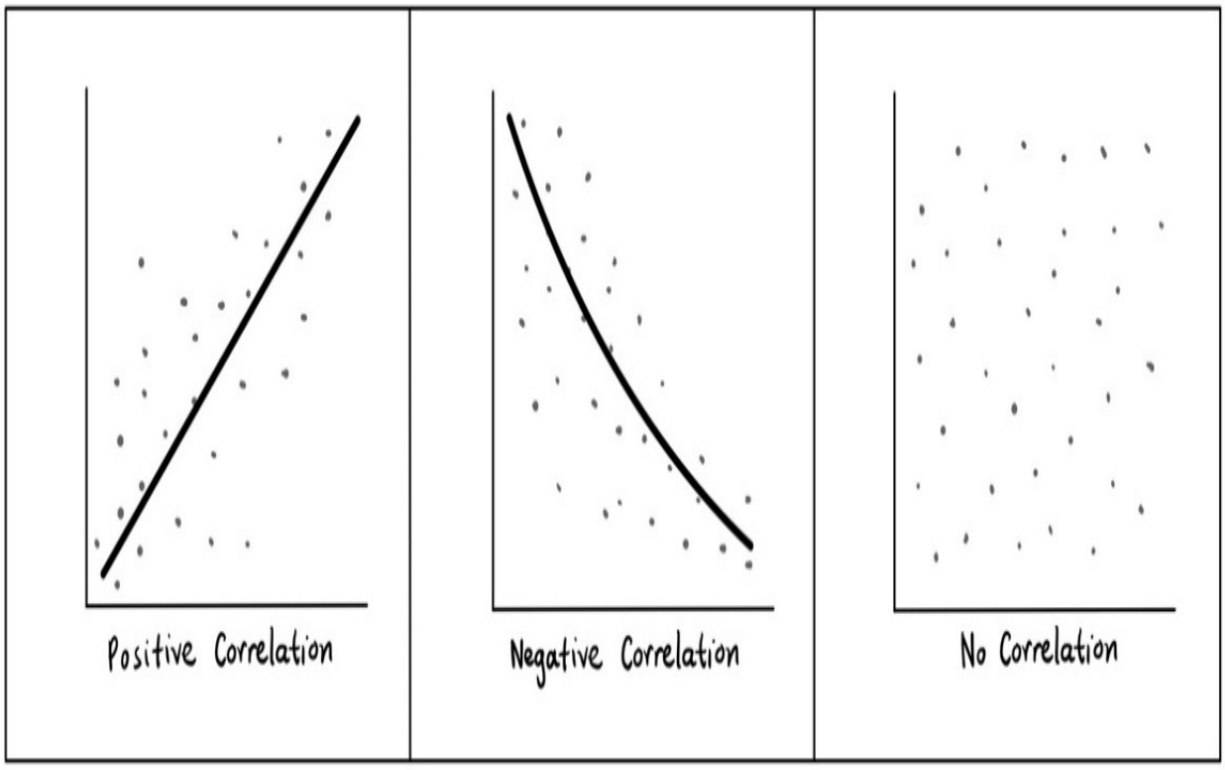
- **Pearson's correlation** measures linear relationships between two continuous variables. It's the most commonly used among correlational methods. To effectively leverage this coefficient, your data must meet the assumptions of other parametric statistics and represent a *linear* trend. If data is *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.

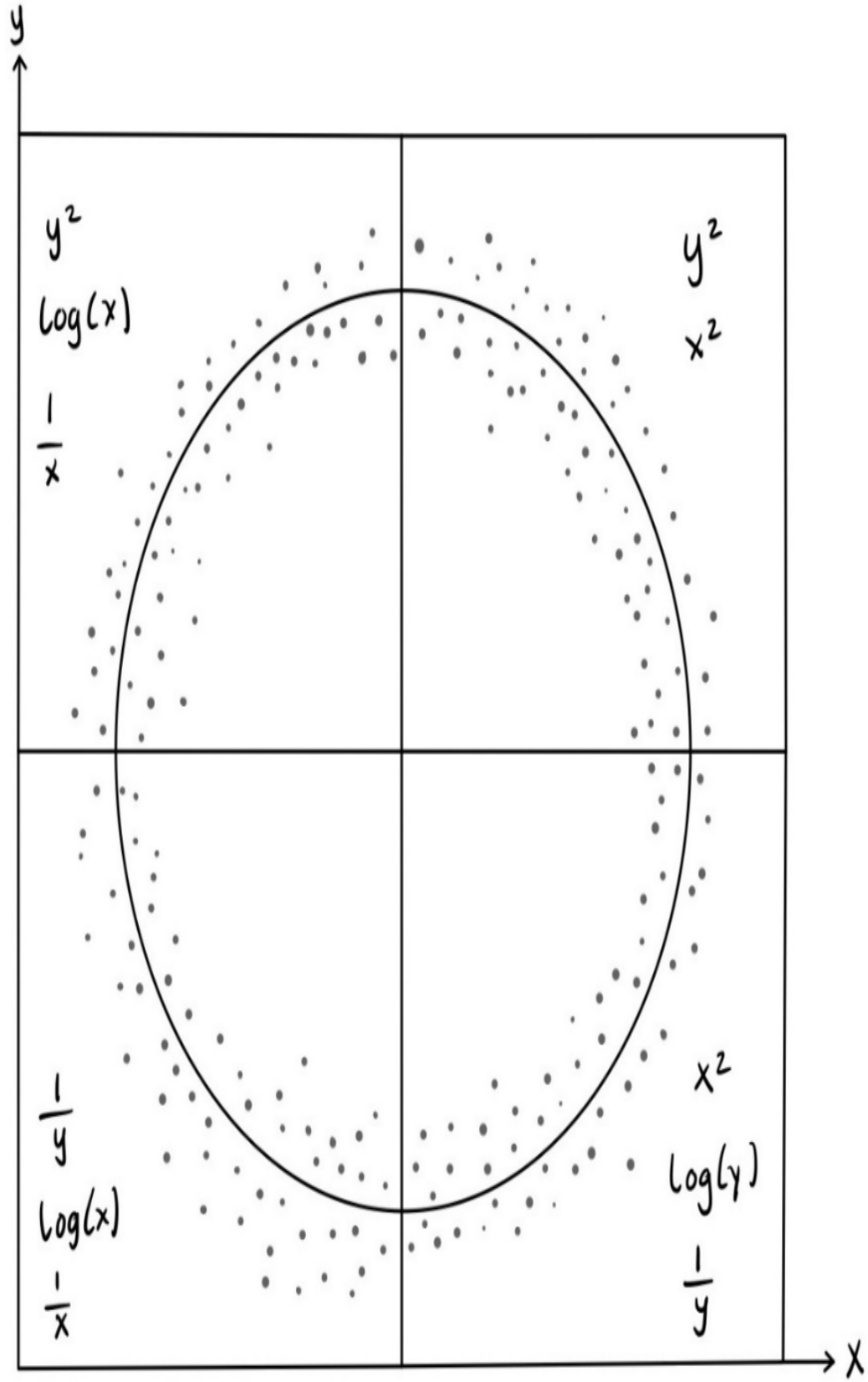


Choosing a correlation coefficient is often dictated by the type of data you are working with (e.g., when relationships are not linear or one variable is not continuous). When measuring associations between two continuous variables, you will generally benefit from visually observing a scatterplot of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


y
y^2
$\log(x)$
$\frac{1}{x}$
y^2
x^2
$\frac{1}{y}$
$\log(x)$
$\frac{1}{x}$
x^2
$\log(y)$
$\frac{1}{y}$
X

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, β_0 is the y-intercept, and β_1/β_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.

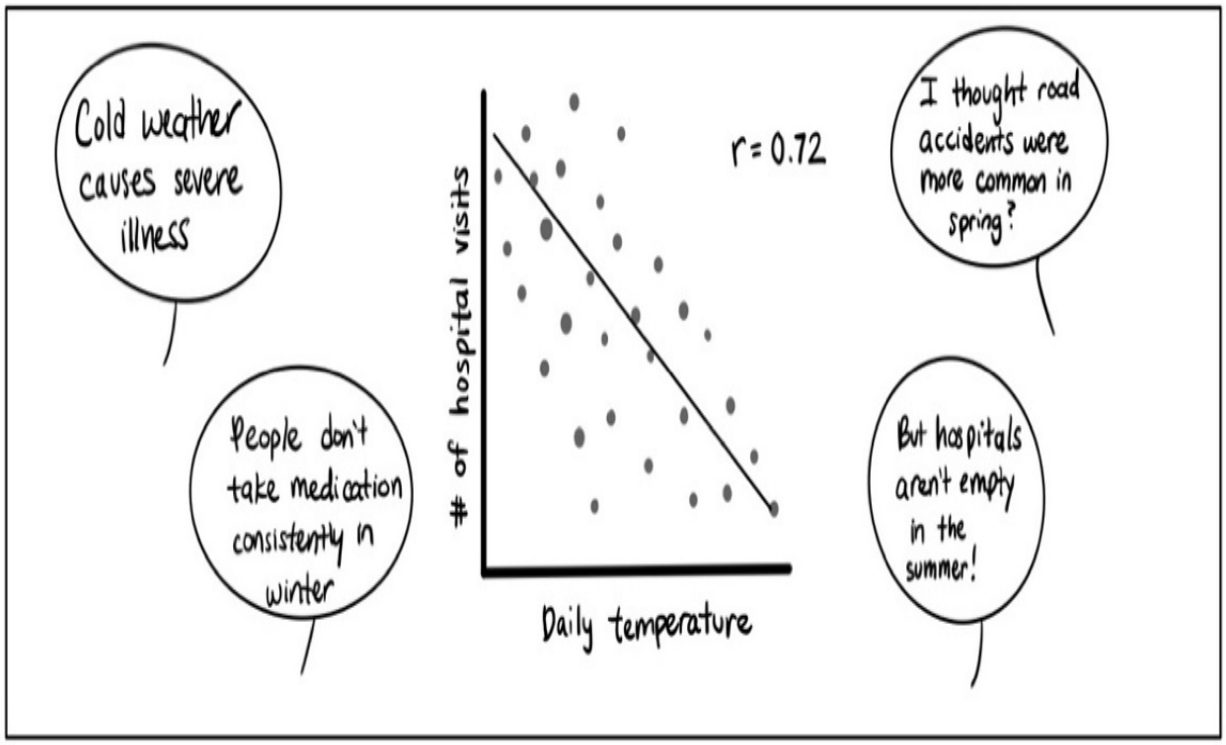
We will discuss the Python implementation of regression models in depth in chapter 9 when we cover approaches to statistical modeling.

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.

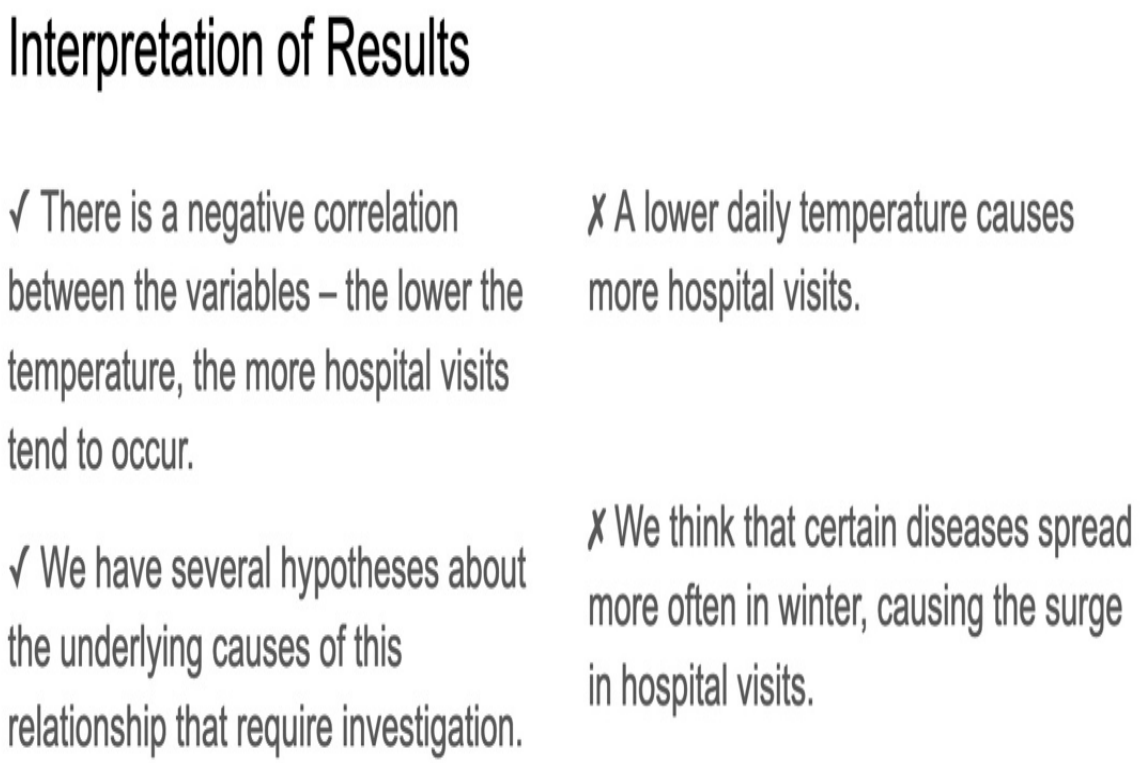


In one of my roles in data science, our team discovered this issue when reporting on correlations to various stakeholders. We identified some patterns of misinterpretation:

- 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.



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

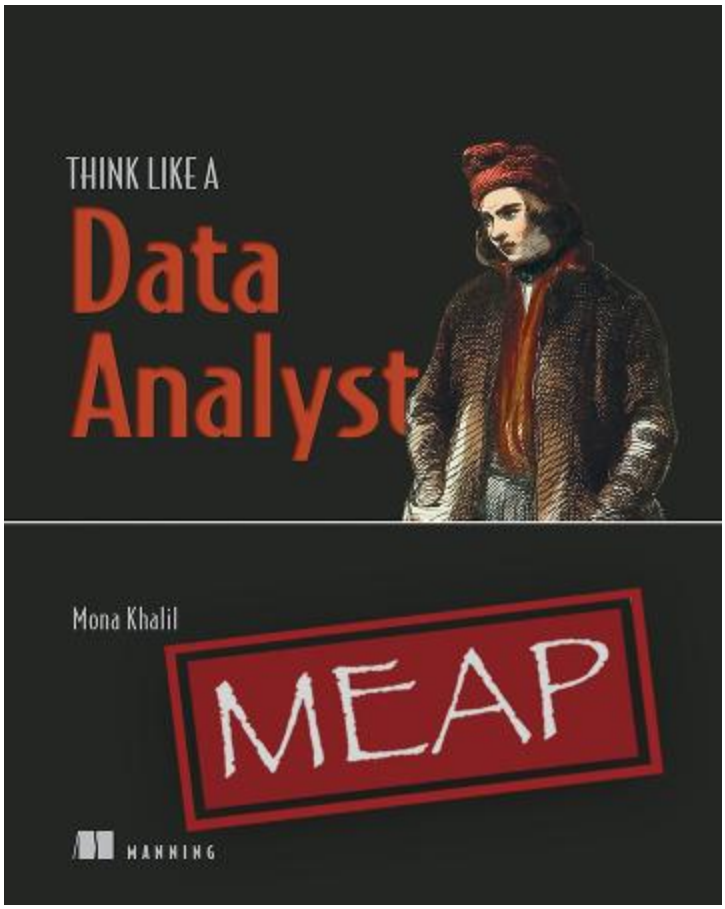
We haven't yet answered the first question of this chapter – is Boston or New York City warmer in July?

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.


THINK LIKE A
Data
Analyst
Mona Khalil
MEAP
MANNING