Data Analysis for Decision-Making

Lecture Notes

October 16, 2024

Table of contents

List of figures

List of tables

Preface

About the notes

A PDF version of these notes is available here.

Please note that while the PDF contains the same content, it has not been optimized for PDF format. Therefore, some parts may not appear as intended.

- These notes aims to support my lecture at the HS Fresenius but are incomplete and no substitute for taking actively part in class.
- I appreciate you reading it, and I appreciate any comments.
- This is work in progress so please check for updates regularly.
- For making an appointment, you can use the online tool that you find on my private homepage: https://hubchev.github.io/
-

About the author

Figure 1.: Prof. Dr. Stephan Huber

I am a Professor of *International Economics and Data Science* at HS Fresenius, holding a Diploma in Economics from the University of Regensburg and a Doctoral Degree (summa cum laude) from the University of Trier. I completed postgraduate studies at the Interdisciplinary Graduate Center of Excellence at the Institute for Labor Law and Industrial Relations in the European Union (IAAEU) in Trier. Prior to my current position, I worked as a research assistant to Prof. Dr. Dr. h.c. Joachim Möller at the University of Regensburg, a post-doc at the Leibniz Institute for East and Southeast European Studies (IOS) in Regensburg, and a freelancer at Charles University in Prague.

Throughout my career, I have also worked as a lecturer at various institutions, including the TU Munich, the University of Regensburg, Saarland University, and the Universities of Applied

Sciences in Frankfurt and Augsburg. Additionally, I have had the opportunity to teach abroad for the University of Cordoba in Spain, the University of Perugia in Italy, and the Petra Christian University in Surabaya, Indonesia. My published work can be found in international journals such as the Canadian Journal of Economics and the Stata Journal. For more information on my work, please visit my private homepage at hubchev.github.io.

Contact:

Prof. Dr. Stephan Huber
Hochschule Fresenius für Wirtschaft & Medien GmbH
Im MediaPark 4c
50670 Cologne
Office: 4e OG-3
Telefon: +49 221 973199-523
Mail: stephan.huber@hs-fresenius.de
Private homepage: www.hubchev.github.io
Github: https://github.com/hubchev

I was always fascinated by data and statistics. For example, in 1992 I could name all soccer players in Germany's first division including how many goals they scored. Later, in 2003 I joined the introductory statistics course of Daniel Rösch. I learned among others that probabilities often play a role when analyzing data. I continued my data science journey with Harry Haupt's *Introductory Econometrics* course, where I studied the infamous Jeffrey M. Wooldridge [2002] textbook. It got me hooked and so I took all the courses Rolf Tschernig offered at his chair of Econometrics, where I became a tutor at the University of Regensburg and a research assistant of Joachim Möller. Despite everything we did had to do with how to make sense out of data, we never actually used the term *data science* which is also absent in the more 850 pages long textbook by Wooldridge [2002]. The book also remains silent about *machine learning* or *artificial intelligence*. These terms became popular only after I graduated. The *Harvard Business Review* article by Davenport and Patil [2012] who claimed that data scientist is "The Sexiest Job of the 21st Century" may have boosted the popularity.

The term "data scientist" has become remarkably popular, and many people are eager to adopt this title. Although I am a professor of *data science*, my professional identity is more like that of an applied, empirically-oriented international economist. My hesitation to adopt the title "data scientist" also stems from the deep respect I have developed through my interactions with econometricians and statisticians. Considering their in-depth expertise, I feel like a passionate amateur.

Ultimately, I poke around in data to find something interesting. Much like my ten-year-old younger self who analyzed soccer statistics to gain a deeper understanding of the sport. The only thing that has changed since then is that I know more promising methods and can efficiently use tools for data processing and data analysis.

About the Course

The course totals 125 hours over one semester at the Master's level, granting 5 ECTS points. It consists of 3 weekly contact hours (42 hours in total) and 83 hours of private study.

Abstract

This module provides essential skills for transforming data into actionable business insights. Upon completion of the module, students will be able to summarize the importance of analyzing business data and engage in discussions about different workflows for leveraging data analytics for new business trends. The module systematically develops skills to create plans for data collection and management. Students learn to recognize the challenges and opportunities of different quantitative empirical strategies. Decision principles, frameworks and tools, including decision trees and payoff tables, are covered in depth, focusing on the central role of decision support systems (DSS). Students will gain an in-depth understanding of the different roles in business analysis and a comprehensive overview of the entire data analysis workflow in a business context.

Learning outcomes/competences

After a successful completion of the module, the students are able to:

- summarize the significance of business data analysis for decision-making and demonstrate the ability to justify and articulate diverse workflows to convert data into actionable information,
- explain the role of data analytics to emerging trends in business and how organizations can use data analytics and decision support systems to solve problems,
- identify and contrast the competencies required to solve business problems with data and be able to assign the various tasks of a data scientific workflow to professionals with the appropriate profile,
- set up a plan for collecting, managing, analyzing, and applying data, reflectively applying quantitative methods,
- distinguish and discuss empirical research strategies to identify causal mechanisms, causes, and effects,
- identify the need of decision support systems and assess the possibilities of data-driven methods to improve decision-making of humans and organizations.

Module content

Introduction

- Significance of business data analysis for decision-making.
- Emerging trends: Evolution of computers and data processing, digitalization, artificial intelligence, machine learning, deep learning, big data, internet of things, cloud computing, and blockchain, industry 4.0, and remote working.
- The role of business analytics and intelligence in converting data into actionable information for decision-makers.
- Overview of various job roles for business analytics: data engineer, data analyst, machine learning engineer, business intelligence analyst, database administrator, data product manager, market research analyst, fraud analyst, …
- Analytics applications in strategic insights.
- Types of analytics: Descriptive, predictive, and prescriptive analytics.
- Workflows and Data science life cycles: OSEMN, DSLC, CRISP-DM, Kanban, TDSP, …

Data literacy competencies

- Types of data: Cross-section, panel, time-series, georeferenced, …
- Types of variables: Continuous, count, ordinal, categorial, …
- Conceptual framework: Knowledge and understanding of data and applications of data.
- Data collection: Identify, collect, and assess data.
- Data management: Organize, clean, convert, curate, and preserve data.
- Data evaluation: Plan, conduct, evaluate, and assess data analyses.
- Data application: Share, reflect, and evaluate results of analyses, comparing them with other findings while considering ethical issues and scientific standards.

Quantitative research design

- Techniques for measuring socio-economic and business phenomena.
- Strategies for identifying causes of effects and effects of causes.
- The fundamental problem of causal inference.
- Techniques to establish causation: matching, natural experiments, field experiments, and laboratory experiments.

Introduction to decision-making

- Overview of the principles guiding rational decision-making.
- Frameworks for structuring and representing the decision-making process.
- Utilization of decision trees, payoff tables, Lagrange multipliers, and expected utility theory.
- Exploration of bounded rationality in human decision-making behaviors.
- The necessity, concept, and evolution of

Assessment Methods and Criteria

Students complete this module with an **academic presentation**. The presentation takes place during the lecture period; the exact date is set by the lecturer. The presentation last for 10-15 minutes per student. In addition, a handout (3-5 pages per student) should be produced outlining the key features of the project and the literature on which these decisions are based (project outline). The handout should be submitted to the lecturer by the date of the presentation at the latest.

Group work is permitted. The maximum group size is 5 students. In case of group work, it must be possible to clearly define and assess each student's individual performance on the basis of specified sections, page numbers, or other objective criteria.

The presentation contributes 65% to the module grade, the handout contributes 35%. A passing grade in this module is achieved when the overall grade is greater than or equal to 4.0. To pass the course a student must give a 10 minute presentation and submit a handout of 3-5 pages.

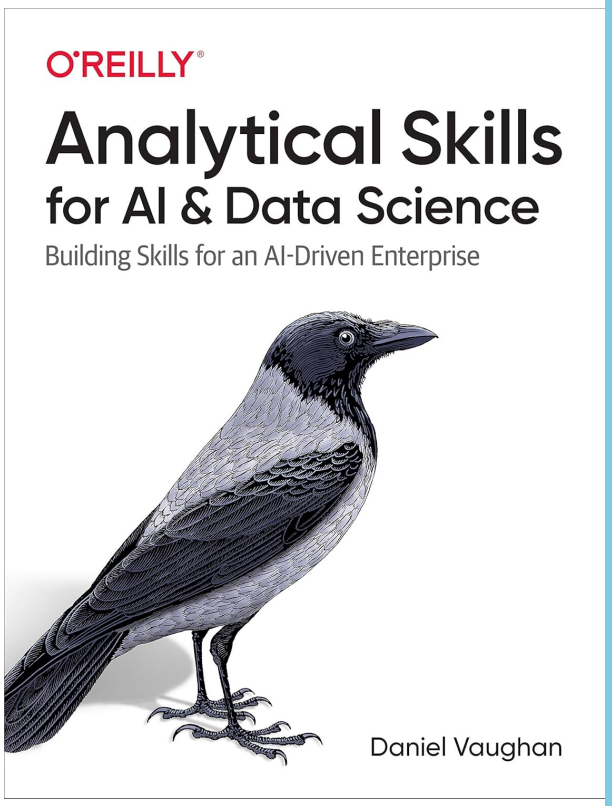
Literature

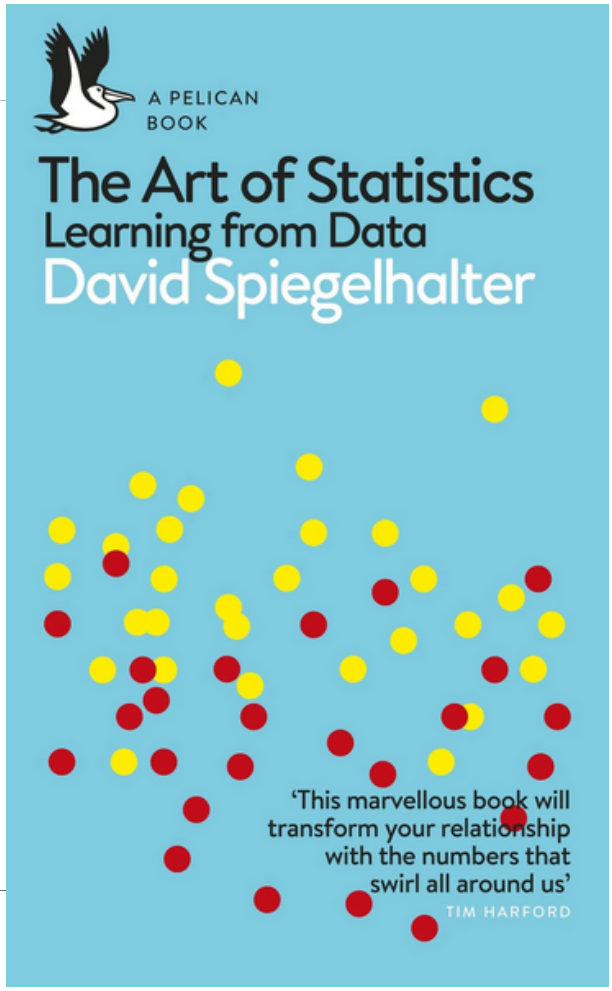
There are tons of books around that are both insightful and entertaining and support the lecture. In Figure 2, I present a short list of books I recommend: Bergstrom and West [2021], Chivers and Chivers [2021], Dougherty and Ilyankou [2021], Ellenberg [2015], Harford [2020], Huff [1954], Huntington-Klein [2022], and Jones [2020].

In addition, I highly recommend the two books mentioned in Figure 3. Vaughan [2020] teaches important skills for working effectively with data. For a more technical approach that focuses on statistics, Spiegelhalter [2019] is an excellent choice. Both books are easy to read even without advanced math skills. We will use these books in the course and some of your presentations will refer to them as well.

Figure 2.: Books for data literacy

Figure 3.: Books for skills and statistics





1. Introduction

1.1. What is the value that I can add in this course?

As a professor, my goal is to educate business graduates so that they can add value to their future organizations. Unfortunately, neither you nor I know what situations you will end up in in business where you can take advantage of using data to make good decisions. There are countless ways to contribute by understanding "data analysis for decision-making". Therefore, I will try to provide knowledge that is as general as possible. This way, I increase the likelihood that at least some parts of the lecture will be important for your future.

Since access to the internet is virtually constant and instantaneous these days, it's hard to add value by memorizing something. In seconds, we get answers to almost any sort of question using tools like Google and ChatGPT. Therefore, I don't see much added value to your education by teaching you facts and schemes to solve problems. Instead, I try to build up your skills and your data literacy that will help you ask the right questions, search for the right information, and use and analyze data in a meaningful way.

Perhaps you know the saying

> *"A man is well educated when he knows where to find what he doesn't know."*[1]

I would like to add something here. A man is well educated when he knows…

- … what he needs to know.
- … where to find what he doesn't know.
- … how to search efficiently.
- … how to verify the information being found.
- … how to transform information into insights.
- … how communicate the insight.

In the spirit of this lecture, my goal is to help you become an educated person who understands how data can support good decision-making.

1.2. What is data analysis for decision-making

For successful communication, intersubjectivity is essential. Once we agree on a clear definition of the words used in the title of this lecture, we can identify the things that we need to know, where to find them efficiently, how to evaluate and use the knowledge.

Data refers to all types of recordable information – facts about something or someone. Data analysis involves the thorough examination of this information with a specific goal in mind. Decision-making is defined as the process of reaching a decision. The word decision comes from the Latin verb decidere, which has various meanings, including

[1] Loosely translated and based on the well-known saying "Gebildet ist, wer weiß, wo er findet, was er nicht weiß." which is often attributed to George Simmel (1858 - 1918).

- make explicit,
- put an end to,
- bring to conclusion,
- settle/decide/agree (on),
- die,
- end up,
- fail,
- fall in ruin,
- fall/drop/hang/flow down/off/over,
- sink/drop,
- cut/notch/carve to delineate,
- detach,
- cut off/out/down,
- fell.

Let's agree on the following definition:

> Fitzgerald [2002, p. 8]: "A decision is the point at which a choice is made between alternative—and usually competing—options. As such, it may be seen as a stepping-off point—the moment at which a commitment is made to one course of action to the exclusion of others."

Exercise 1.1. Pass the course
To pass this course, you will be required to give a 10-minute presentation and produce a 3-5 page handout. Your performance will be measured by the quality of your work in relation to the time available. The value of the presentation lies not only in the content, but also in the impact it has on your audience – your classmates and your professor. Your goal is to add some value with your work.
Discuss:

- What do you need to know to do well in this course?
- Where can you find the information you need to be successful?
- What tools will help you find, store and use information efficiently?
- How can you verify the accuracy of the information you find?
- How can you analyze information to gain valuable insights?
- How can you communicate these insights effectively?

Exercise 1.2. Information is omnipresent
Facts are abundant and easily accessible, so education should not just focus on providing information. Instead, skills should be developed to deal effectively with the wealth of facts available. This includes mastering statistical methods and avoiding common pitfalls when working with data.
To find the appropriate statistical method and interpret the data correctly, you should have some knowledge about the type of data and variables that you look at. Use ChatGPT to inform yourself about the following points that are defined in the module description:
Data literacy competencies

- Types of data: Cross-section, panel, time-series, georeferenced, …
- Types of variables: Continuous, count, ordinal, categorial, …

Make a short presentation about it. At best, include definitions and examples to the different types of information.

Moreover, what tools would you pick to make the presentation? Discuss the pros and cons of the options available.

2. From anecdote to insight

Anecdotes are great. They are true stories—often intriguing, relatable, and easy to understand. They provide vivid examples that make abstract ideas more concrete and memorable. Whether it's a personal experience or a captivating story about a successful business leader, anecdotes resonate because they tap into our natural affinity for storytelling. Their simplicity and emotional impact can make them powerful teaching tools.

And importantly, anecdotes are hard to contradict. Take, for example, the argument that smoking can't be that harmful because your 88-year-old uncle has smoked his entire life and he is still in good health. It's a tough claim to refute, as it's a real-life example. However, the problem lies in extrapolating a single, isolated case to draw broader conclusions, which can be misleading.

However, while anecdotes can be persuasive, their strength is also their weakness. They represent isolated instances, and while it's hard to deny the truth of an individual story, the danger lies in overgeneralizing from it. Anecdotes lack the rigorous analysis and breadth of evidence necessary to draw reliable conclusions. They don't account for the full complexity of most situations, especially in business, where decisions are influenced by many interconnected factors.

In business, relying too heavily on anecdotes can lead to misguided conclusions. For example, a company might base its strategy on the success story of a famous entrepreneur without considering the countless failed ventures that didn't make the headlines. This is known as survivorship bias, where the successes are visible, but the failures are hidden.

The challenge, then, is to take anecdotes and go beyond them. Instead of drawing direct conclusions, use them as starting points for deeper investigation. They can provide valuable hypotheses but need to be supported by data, rigorous analysis, and an understanding of the underlying principles at play. Anecdotes can inspire curiosity and point us in interesting directions, but they should be tested against a larger body of evidence to ensure that the insights we draw are reliable and applicable in a broader context.

Exercise 2.1. Survivorship bias
Read "How Successful Leaders Think" by Roger Martin [2007] and the chapter "Identification" of "Quantitative Methods" by Huber [2024a].
Here is a summary of Martin [2007] taken from the Harvard Business Review Store:

> In search of lessons to apply in our own careers, we often try to emulate what effective leaders do. Roger Martin says this focus is misplaced, because moves that work in one context may make little sense in another. A more productive, though more difficult, approach is to look at how such leaders think. After extensive interviews with more than 50 of them, the author discovered that most are integrative thinkers–that is, they can hold in their heads two opposing ideas at once and then come up with a new idea that contains elements of each but is superior to both. Martin argues that this process of consideration and synthesis (rather than superior strategy or faultless execution) is the

hallmark of exceptional businesses and the people who run them. To support his point, he examines how integrative thinkers approach the four stages of decision making to craft superior solutions. First, when determining which features of a problem are salient, they go beyond those that are obviously relevant. Second, they consider multidirectional and nonlinear relationships, not just linear ones. Third, they see the whole problem and how the parts fit together. Fourth, they creatively resolve the tensions between opposing ideas and generate new alternatives. According to the author, integrative thinking is an ability everyone can hone. He points to several examples of business leaders who have done so, such as Bob Young, co-founder and former CEO of Red Hat, the dominant distributor of Linux open-source software. Young recognized from the beginning that he didn't have to choose between the two prevailing software business models. Inspired by both, he forged an innovative third way, creating a service offering for corporate customers that placed Red Hat on a path to tremendous success.

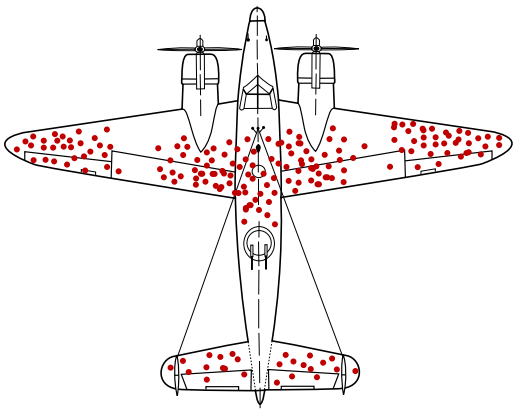
a) Discuss the concepts introduced by Martin [2007] critically:

- Does he provide evidence for his ideas to work?
- Is there a proof that his suggestions can yield success?
- Is there some evidence about whether his ideas are superior to alternative causes of action?
- What can we learn from the article?
- Does his argumentation fulfill highest academic standards?
- What is his identification strategy with respect to the *causes of effects* and the *effects of causes*?
- Martin [2007, p. 81] speculates:

  "At some point, integrative thinking will no longer be just a tacit skill (cultivated knowingly or not) in the heads of a select few."

b) If teachers in business schools would have followed his ideas of integrative thinkers being more successful, almost 20 years later, this should be the dominant way to think as a business leader. Is that the case? And if so, can you still gain some competitive advantage by thinking that way?

Figure 2.1.: Distribution of bullet holes in returned aircraft



Source: Martin Grandjean (vector), McGeddon (picture), Cameron Moll (concept), CC BY-SA 4.0, Link

c) Figure 2.1 visualizes the distribution of bullet holes in aircraft that returned from combat in World War II. Imagine you are an aircraft engineer. What does this picture teach you?

d) Inform yourself about the concept of survivorship bias explained in Wikipedia [2024].

e) In Martin [2007], the author provides an example of a successful company to support his management ideas. Discuss whether this article relates to survivorship bias.

Drawing insights from anecdotes is challenging, especially in business, for several reasons:

1. **Limited sample size**: Anecdotes are usually individual cases that do not reflect the full extent of a situation. In business, decisions often require data from large, diverse populations to ensure reliability. Relying on a single story or experience can lead to conclusions that are not universally valid.
2. **Bias and subjectivity**: Anecdotes are often influenced by personal perspectives, emotions or particular circumstances. Moreover, anecdotes often highlight success stories while ignoring failures. This is an example for the so-called *Survivorship Bias*.
3. **Lack of context and the inability to generalize**: Anecdotes often lack the broader context necessary to understand the underlying factors of a situation. Business problems tend to be complex and influenced by numerous variables such as market trends, consumer behavior and external economic conditions. Many of these variables change significantly over time. Without this context, an anecdote can oversimplify the problem and lead to incorrect decisions. Anecdotes are usually specific to a particular time, place or set of circumstances. They may not apply to different markets, industries or economic environments, which limits their usefulness for general decision-making. For example, learning only from the tremendous success of figures like Steve Jobs while ignoring the countless people who failed is like learning how to live a long life by talking to a single 90-year-old person. If that person happens to be obese and a heavy smoker, it doesn't mean those behaviors contributed to their longevity.
4. **Lack of data rigor**: Anecdotes lack the rigor and precision of data-driven analysis where the empirical model that allows to identify causality and to measure the effect of causes is formally described.

Conclusion

To make informed business decisions, it is critical to base insights on systematic data analysis rather than anecdotal evidence, as anecdotes are too narrow, subjective and unreliable to guide complex business strategies.

Exercise 2.2. Systematic analysis as an alternative to anecdotal analysis

- What defines a systematic analysis?
- When can we say that we have 'found evidence'?
- When can we claim to have identified a causal effect?
- When can we trust the size of an effect that we have measured?

3. Epistemic errors

Many things can go wrong when analysing data. Various issues can lead to misleading interpretations, whether due to data not measuring what it is intended to, human misinterpretation, or an overreliance on probabilistic reasoning. The following examples illustrate these pitfalls and are drawn from the insightful book by Jones [2020], which I highly recommend for further reading.

Epistemic errors, that are mistakes related to knowledge, understanding, or the acquisition of information, arise from cognitive biases, misunderstandings, and incorrect assumptions about the nature of data and the reality it represents. Recognizing and addressing these errors is crucial for accurate data analysis and effective decision-making.

When researchers analyse data they should be aware of the fact that data can represent reality but do not necessarily do so. For example, a survey on customer satisfaction that only includes responses from a self-selected group of highly engaged customers may not accurately reflect the overall customer base, see the box Hite Report. To avoid this pitfall, it is essential to ensure that your data collection methods are representative and unbiased, and to validate your data against external benchmarks or additional data sources.

Hite Report

In 1976, when the *The Hite Report* (see Hite [1976] and Figure 3.1) was published it instantly became a best seller. Hite used an individualistic research method. Thousands of responses from anonymous questionnaires were used as a framework to develop a discourse on human responses to gender and sexuality. The following comic concludes the main results.

Figure 3.1.: The Hite [1976] Report

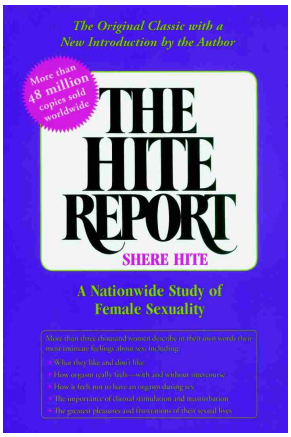
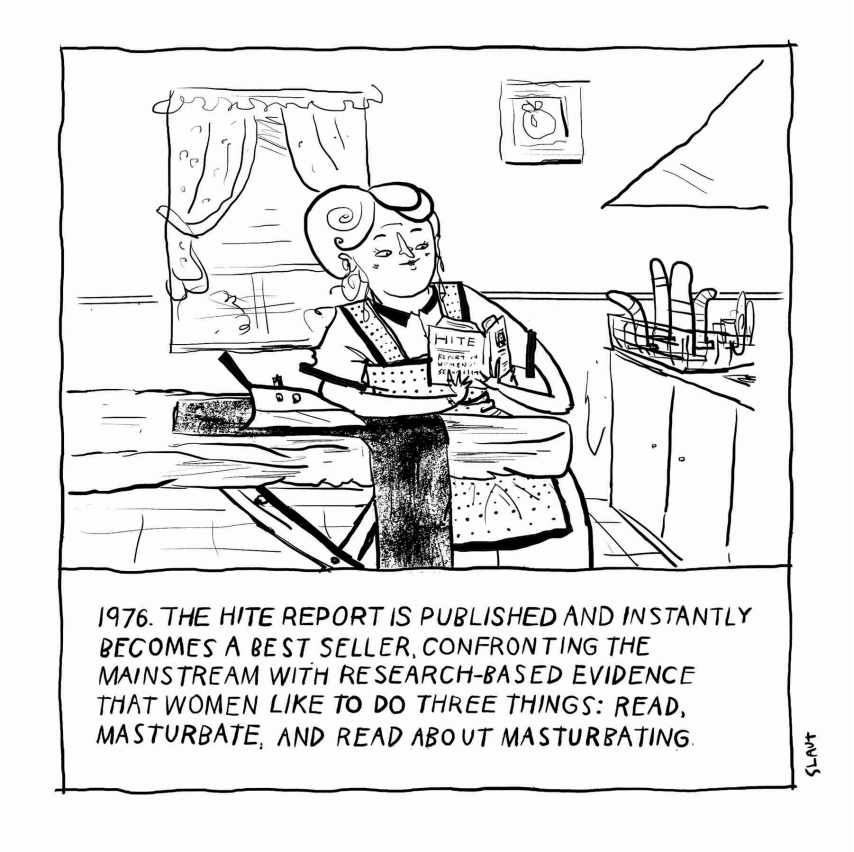


Figure 3.2.: Comic on the Hite Report



Source: Picture is taken from www.theparisreview.org.

The picture of womens' sexuality in Hite [1976] was probably a bit biased as the sample can hardly be considered to be a **random and unbiased** one:

- Less than 5% of all questionnaires which were sent out were filled out and returned (response bias).
- The questions were only sent out to women's organizations (an opportunity sample).

Thus, the results were based on a sample of women who were highly motivated to answer survey's questions, for whatever reason.

The Data-Reality Gap

The difference between the data we collect and the reality it is supposed to represent.

Another common epistemic error involves the influence of human biases during data collection and interpretation. Known as the all too human data error, this occurs when personal biases or inaccuracies affect the data. An example would be a researcher's personal bias influencing the design of a study or the interpretation of its results. To mitigate this, implement rigorous protocols for data collection and analysis, and consider using double-blind studies and peer reviews to minimize bias.

All Too Human Data

Errors introduced by human biases or inaccuracies during data collection and interpretation.

Inconsistent ratings can also lead to epistemic errors. This happens when there is variability in data collection methods, resulting in inconsistent or unreliable data. For example, different evaluators might rate the same product using different criteria or standards. To avoid this issue, standardize data collection processes and provide training for all individuals involved in data collection to ensure consistency.

Inconsistent Ratings

Variability in data collection methods that leads to inconsistent or unreliable data.

The black swan pitfall refers to the failure to account for rare, high-impact events that fall outside regular expectations. Financial models that did not predict the 2008 financial crisis due to the unexpected nature of the events that led to it are an example of this error. To prevent such pitfalls, consider a wide range of possible outcomes in your models and incorporate stress testing to understand the impact of rare events.

The Black Swan Pitfall

The failure to account for rare, high-impact events that fall outside the realm of regular expectations.

Falsifiability and the God pitfall involve the tendency to accept hypotheses that cannot be tested or disproven. This error might occur when assuming that a correlation between two variables implies causation without the ability to test alternative explanations. To avoid this, ensure that your hypotheses are testable and that you actively seek out potential falsifications. Use control groups and randomized experiments to validate causal relationships.

Falsifiability and the God Pitfall

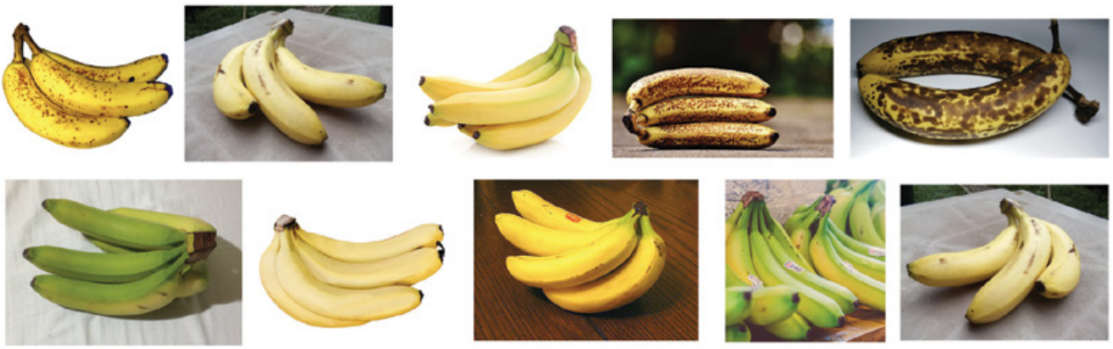
The tendency to accept hypotheses that cannot be tested or disproven.

To avoid epistemic errors, critically assess your assumptions, methodologies, and interpretations. Engage in critical thinking by regularly questioning your assumptions and seeking alternative explanations for your findings. Employ methodological rigor by using standardized and validated methods for data collection and analysis. Engage with peers to review and critique your work, providing a fresh perspective and identifying potential biases. Finally, stay updated with the latest research and best practices in your field to avoid outdated or incorrect methodologies.

Understanding and addressing epistemic errors can significantly improve the reliability and accuracy of your data analyses, leading to better decision-making and more trustworthy insights.

Exercise 3.1.

Figure 3.3.: Bananas in various stages of ripeness



Source: Jones [2020, p. 33]

a) Rate the ripeness level of the bananas pictured by Figure 3.3. Compare your assessment to that of a colleague and discuss any differences in your ratings. What might account for the variance in perception of the bananas' ripeness between you and your colleague?

b) Specify how you rated the second and the last bananas on the ripeness scale?

c) Upon reevaluation, it appears that the second and the last bananas are identical in ripeness. How would you justify your initial decision now? This scenario underscores an important lesson for interpreting polls and surveys: it illustrates how subjective assessments can lead to variance in results. It highlights the necessity of ensuring clarity and consistency in the criteria used for evaluations to minimize subjective discrepancies.

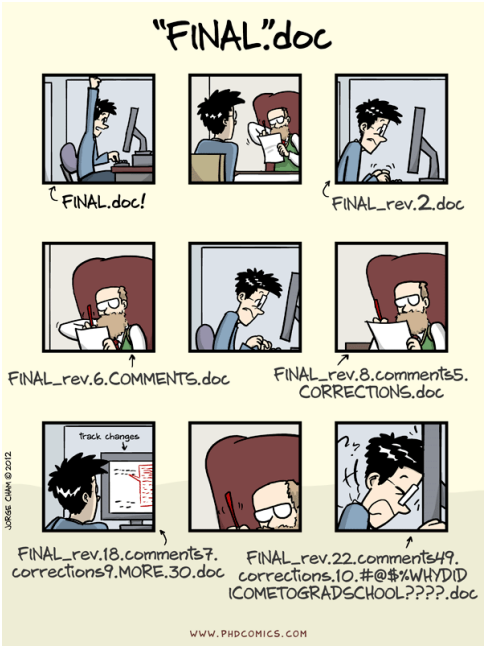
4. How to use R for data science

The programming language R is one of the major tools to do data science. I wrote some lecture notes on *How to use R for data science* [Huber, 2024b].

Please read these notes.

5. Collaborating with Git and GitHub

Figure 5.1.: The FINAL.doc problem



Source: phdcomics.com

5.1. Introduction

Git is open-source software for version control. It allows developers to track and manage changes to their codebase and files. Users can access a comprehensive history of their project and revert to previous versions of their data if necessary. It helps to overcome the *FINAL.doc problem* depicted in Figure 5.1.

Figure 5.2.: GitHub is big

100+ million	4+ million	420+ million	90%
Developers	Organizations	Repositories	Fortune 100

Source: https://github.com/about as of April 2024

GitHub is an incredibly popular (see statistics in Figure 5.2) online platform that implements Git's capabilities by providing a web interface for collaboration.

While you can use Git and GitHub independently, most developers integrate it with GitHub for enhanced project management and collaboration. This combination helps maintain local and remote copies of a project, facilitating teamwork and data backup as GitHub is sort of a

backup as data loss at your local machine do not matter if you have a remote version saved on GitHub.

Git and GitHub support simultaneous multi-user access, unlike systems that are optimized for single-user like Dropbox.

5.2. Install Git

To install the version control system Git, follow the instructions here.

Figure 5.3.: Memorizing six git commands

The popular approach to version control

Source: DEV Community on GitHub

Familiarize yourself with Git by using the resources available here. Specifically, work through the resources listed in the box below. Although Git may appear complex, it is generally not too challenging for most users. Many people use Git primarily to track their work and to host and share files conveniently with just a handful of commands. While Git is a robust system with many capabilities, you don't need to memorize all the commands (see Figure 5.3). In fact, you typically use only a few basic ones as shown in Table 5.1.

In the upcoming sections, I will demonstrate some use cases both in the terminal Section 5.3 and within RStudio Section 5.4. In Section 5.5, I show how to contribute to a repository using Git and GitHub.

Learning resources

Plenty books and tutorial exist that introduce Git and GitHub. I'd like to highlight the following sources:

- The book comprehensive book *Happy Git and GitHub for the useR* by Bryan
- The much shorter book [*Version Control with Git and GitHub*] by Halbritter and Telford
- The online tutorial *How to Use Git/GitHub with R* of David Keyes who explains

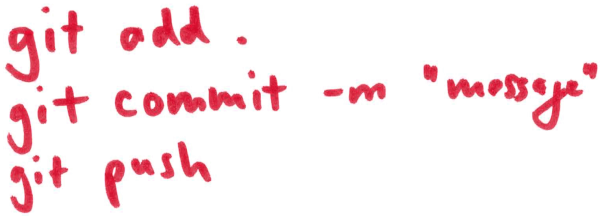
in short videos how to setup Git and GitHub in RStudio using among others the `usethis` package.

Table 5.1.: Most important git commands

Git Command	Description
`git init`	Initialize a new Git repository in the current directory.
`git clone <url>`	Clone a repository from a remote URL to your local machine.
`git add <file>`	Add a specific file to the staging area in preparation for committing.
`git add .`	Add all changed files in the current directory to the staging area.
`git commit -m "message"`	Commit the staged changes to the repository with a descriptive message.
`git status`	Display the status of the working directory and staging area.
`git push <remote> <branch>`	Push committed changes in your local branch to the remote repository.
`git pull <remote> <branch>`	Pull changes from the remote repository into your current branch and merge them.
`git branch <name>`	Create a new branch with the specified name.
`git checkout <branch>`	Switch to another branch and update the working directory.
`git merge <branch>`	Merge a specified branch into the current branch.

5.3. Using Git from the terminal

Figure 5.4.: Three git commands you really need



This tutorial will guide you through the basic Git operations using the Bash command line, commonly referred to as the terminal. Essentially, it focuses on the three Git commands illustrated in Figure 5.4.

5.3.1. Configuring Git

Before you start using Git, you need to configure your Git environment. Set your username and email address with these commands:

```
git config --global user.name "Your Name"
git config --global user.email "your.email@example.com"
```

5.3.2. Initializing a Repository

To create a new Git repository, use the git init command in the directory you want to version control:

```
cd /path/to/a/directory
mkdir my_project
cd my_project
git init
```

In case you are not familiar with using the terminal please consider Table 5.2 where I introduce the most basic commands that we use. For example, with `cd` you can change your directory and with `mkdir` you create a new directory. If you are not familiar with the file system of your computer please read the section *Navigating the file system* of Huber [2024b]. With `git init` you initialize the directory to be a git repository. This will create a hidden folder "`.git`" in which Git keeps track of all your changes.

ℹ Most common bash commands

Table 5.2.: Most common bash commands

Bash Command (macOS/Linux)	Windows Command Prompt Equivalent	Description
`pwd`	`cd`	Prints the current directory's path.
`ls`	`dir`	Lists all files and directories in the current directory.
`cd`	`cd`	Changes the directory.
`mkdir`	`mkdir`	Creates a new directory.
`rmdir`	`rmdir`	Removes an empty directory.
`touch`	`copy nul`	Creates a new empty file or updates an existing file's timestamp.
`rm`	`del` or `erase`	Removes files. `rmdir /s` is used for directories.
`cp`	`copy`	Copies files or directories.
`mv`	`move`	Moves or renames files or directories.
`echo`	`echo`	Displays a line of text/string.
`cat`	`type`	Concatenates and displays the content of files.
`grep`	`find` or `findstr`	Searches for patterns in files.

5.3.3. Staging Changes

To track changes in your repository, you need to stage them using the git add command. To stage a single file:

```
git add filename.txt
```

To stage all changes in the directory:

```
git add .
```

5.3.4. Committing Changes

After staging, you can commit it to the repository. A commit records changes to the repository and must include a message describing what changed:

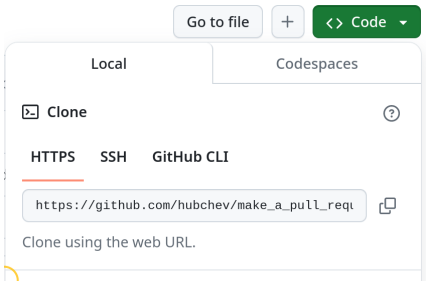
```
git commit -m "A message"
```

5.3.5. Pushing Changes

To share your commits with others or store them in a remote repository (GitHub), use `git push`. A prerequisite here is that you need to be connected to a remote repo. Therefore, you must add a remote repository by copying the URL of the GitHub repo as shown in Figure 5.5. Then you can add the remote repository and push it to the repo with these lines of code:

```
git remote add origin https://github.com/username/repository.git
git push -u origin main
```

Figure 5.5.: Copy the https URL of your repo



5.3.6. Undo changes

With `git reset` and `git revert` you can go back in time and undo specific changes, respectively. For example, with

```
git log
git reset --hard <commit_id_hash>
```

you can view the commit history and find the hash identifier of the commit to move the HEAD pointer to that commit. This effectively removes all commits after commit you choose from the current branch's history. Be cautious when using git reset –hard as it discards all changes made after the specified commit. Make sure you have backups or are certain you want to discard these changes before proceeding.

With

```
git revert <commit_id_hash>
```

you revert the changes introduced by that commit only. It will create a new commit that undoes the changes made in commit chosen while keeping the other commits that may have followed the chosen commit intact. It's a safer approach compared to `git reset --hard`, as it preserves the commit history and allows you to selectively undo changes without affecting the rest of the commits.

5.4. Using Git from RStudio

Integrating Git with RStudio enhances your project management by utilizing version control directly within the IDE. Here's how you can set up and use Git in RStudio using R code.

5.4.1. Set up Git in RStudio

First, ensure the `usethis` package is installed and loaded:

```
if (!require(pacman)) install.packages("pacman")
pacman::p_load(usethis)
```

Configure your Git settings in RStudio:

```
use_git_config(user.name = "Your Name", user.email = "Your@email.com")
```

You can change the configuration of your user name and email using the `edit_git_config()` function.

Start a new project in RStudio, which will also initialize a Git repository:

```
create_project("~/Music/")
use_git()
```

After restarting RStudio, you will notice a Git tab in the top right panel, indicating that Git is now active for your project.

5.4.2. Connecting RStudio Projects with GitHub repositories

To connect your RStudio project with GitHub, you need a Personal Access Token (PAT) on GitHub. If you haven't one already, you can use the `create_github_token()` function from `usethis` package, and store the PAT securely with `gitcreds_set` from the `gitcreds` package:

```
if (!require(pacman)) install.packages("pacman")
pacman::p_load(usethis gitcreds)
create_github_token()
gitcreds::gitcreds_set()
```

Now, the procedure depends on whether the project has bin initialized on your local machine and you want to create a repo on GitHub, or the repo already exists on GitHub and you want to connect that remote repo with your local PC. Both ways are described below.

5.4.2.1. Project exists on RStudio first

After initializing Git in your project, use the `use_github()` function from `usethis` to create a new GitHub repository and connect it directly:

```
use_github()
```

This creates a repo on your GitHub account.

5.4.2.2. Project exists on GitHub first

Alternatively, suppose you have created a repository on GitHub first, then start a new project in RStudio using the version control option, specifying your new repository's URL. Just click `File > New Project > Version Control` and then link the GitHub repo by putting the URL into the respective box of the menu. See Figure 5.5 how to get the URL of a repo.

5.5. Make a contribution using Git and GitHub

This is a guide for beginners on how to make a contribution using Git and GitHub. If you are looking to make your first contribution, follow the steps below.

> **Watch**
>
> this video where I do all the following steps in real time. It takes about 15 minutes.

1. Create an account on GitHub

It is for free and should just take some minutes.

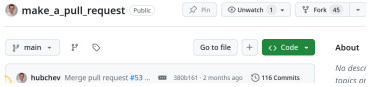
2. Install Git

Here is a tutorial on how to set up Git.

3. Fork this repository

Click on the fork button (see Figure 5.6) on the top of this page: https://github.com/hubchev/make_a_pull_request. This will create a copy of this repository in your account.

Figure 5.6.: Fork the repo



4. Clone the forked repository

Go to your GitHub account, open the forked repository, click on the code button and then click the *copy to clipboard* icon, see Figure 5.5.

Then, open a terminal and run the following git command:

```
git clone "url you just copied"
```

where "url you just copied" (without the quotation marks) is the url to this repository (your fork of this project). See the previous steps to obtain the url.

For example:

```
git clone https://github.com/this-is-you/make_a_pull_request.git
```

where `this-is-you` is your GitHub username. Here you're copying the contents of the first-contributions repository on GitHub to your computer.

5. Create a branch

Change to the repository directory on your computer (if you are not already there):

```
cd make_a_pull_request
```

Now create a branch using the `git switch` command:

```
git switch -c your-new-branch-name
```

For example:

```
git switch -c add-Stephan-Huber
```

6. Make changes.

Now open the`I_am_a_data_scientist.md` file in a text editor. (You find this file in the repository.) Add your name, your GitHub account and the project you are working on. You can put it anywhere in between. Now, save the file.

If you go to the project directory and execute the command `git status`, you'll see there are changes.

7. Add changes (staging). Add those changes to the branch you just created using the `git add` command:

```
git add .
```

8. Commit changes. Now commit those changes using the `git commit` command:

```
git commit -m "Add your-name to the list"
```

replacing `your-name` with your name.

9. Use Git Bash. Open Git Bash and set your email and your nickname on GitHub:

```
git config --global user.name "FIRST_NAME LAST_NAME"
git config --global user.email "MY_NAME@example.com"
```

10. Push changes to GitHub.

Push your changes using the command `git push`:

```
git push -u origin your-new-branch-name
```

replacing `your-new-branch-name` with the name of the branch you created earlier.

If you get any errors while pushing that refers to authentication failed something, go to GitHub's tutorial on generating and configuring an SSH key to your account. Alternatively, you can watch this YouTube tutorial

11. Submit your changes for review on GitHub.

If you go to your repository on GitHub, you'll see a `Compare & pull request` button. Click on that button.

Now submit the pull request.

Soon I'll be merging all your changes into the main branch of this project. You will get a notification email once the changes have been merged.

Congrats! You just completed the standard *fork -> clone -> edit -> pull request* workflow that you'll often encounter as a contributor!

6. Markdown, Quarto, and R Markdown

Verbal and non-verbal communication are crucial in business. This section focuses on writing and publishing texts, excluding aspects like body language and writing skills. I will introduce some tools commonly used by data scientists for writing and publishing their work, such as Markdown, RMarkdown, and Quarto. Unlike applications like Microsoft Word or Apple Pages, these tools use code to generate text. This concept may be unfamiliar to those who grew up after Windows 95; I will provide justification for its use in the following sections.

One notable advantage of a code-based approach to writing text is its seamless integration with version control systems like Git and platforms like GitHub. These tools are essential for most data science collaborations. Mastering them can greatly enhance your efficiency and make your presentations more impactful, even if you are not directly involved in data science.

What is Quarto?

Quarto, a modern documentation system, is an excellent choice for writing, especially for projects that require rigorous data analysis, visualization, and reproducibility. This tutorial will guide you through producing various forms of text with Quarto. You can write reports, articles, theses, books, websites and many more with Quarto.

Quarto and R Markdown

Quarto is a relatively new tool and can be considered as a successor to R Markdown. Most R Markdown documents are compatible with Quarto. However, Quarto offers some improved functionality over R Markdown, which enhances user-friendliness. A detailed overview of the differences and similarities between the two can be found in this article. For an introduction to R Markdown see Section 6.4.

6.1. Why Markdown and Quarto?

6.1.1. No-code vs. code-based writing applications

Students often use Microsoft Word, Apple Pages, or LibreOffice to write scientific texts. These word processing programs operate on the "What You See Is What You Get" (WYSIWYG) principle, displaying the document layout as you type. While this principle and its corresponding applications are widespread and may seem indispensable to many, this is far from true. Alternatives such as LaTeX, Markdown, R Markdown, and Quarto offer significant advantages. Many professional scientists and publishers prefer these alternatives for good reason. A large number of doctoral theses and scientific papers are authored using LaTeX, and nearly all publishers and editors work with code-based solutions that do not follow the WYSIWYG principle.

With code-based alternatives, layout specifications are either placed at the beginning of the text or embedded within the main text itself. The final document is only visible after converting (also called compiling or rendering) it into a format such as PDF. This may initially seem unusual

and less intuitive than a WYSIWYG interface, but the most intuitive solution is not necessarily the best or simplest. My experience supervising numerous student papers has shown that the intuitive features of MS Word and Pages often become time-consuming over the medium to long term and fail to adequately support users in avoiding errors when writing scientific work. Students who choose code-based applications tend to experience less frustration and greater success—at least, this has been true for the papers I have supervised.

Code-based applications allow writers to focus on the actual writing process, as formatting and adherence to citation rules are largely automated by the software. The necessary initial investment in learning a tool like Quarto quickly pays off, resulting in noticeable improvements in the quality of scientific texts.

In the following subsections, I will first outline typical usage of WYSIWYG applications, then discuss the advantages of code-based text creation using Quarto as an example, and finally explain how to successfully write texts using Quarto.

6.1.2. Typical (mis)usage of WYSIWYG applications

The use of traditional word processing software like Microsoft Word or Apple Pages for writing academic papers is pervasive among students. While these programs are user-friendly for everyday writing projects, they create a significant additional workload to meet the demands of academic work.

One of the first problems is integrating literature. Correct formatting according to various citation guidelines is often counter-intuitive, and errors occur easily. This is particularly true if the citation and bibliography functions provided by the software are not used or are used incorrectly. Instead of utilizing external citation managers and investing time to learn how to use them, many students manually create citations and bibliographies, which typically leads to numerous small and sometimes larger errors that could be avoided.

Another weak point in student work is adherence to specific formatting requirements. Academic institutions and journals often require strict adherence to formatting guidelines, including the design of title pages, headers and footers, page margins, and heading hierarchies. Although Word and Pages offer templates and styles, they must be individually adapted for each document and often modified due to minor text changes. Making a format adjustment can become a major effort.

The inclusion of empirical results such as statistical data and graphics presents an additional hurdle. With Word and Pages, the process is frequently manual: research data must be exported from statistical software, saved as images, and then embedded in the document. If the data changes, this time-consuming process must be repeated, significantly increasing the workload and risk of errors.

6.1.3. Advantages of Quarto for writing text

Writing academic texts using traditional tools such as MS Word or Pages can be time-consuming and error-prone for students. In the following section, I introduce Quarto (or R Markdown), a modern alternative that offers several advantages:

- **Versatile output formats:** Quarto makes it effortless to generate different output formats. The same text can be rendered as a website (HTML), manuscript (PDF, DOCX), book (EPUB, PDF), or slides (PDF). This flexibility allows you to focus more on the content than the format.

- **Simplified formatting changes:** Specific templates can be used in Quarto, simplifying the process of making formatting changes.
- **Seamless literature integration:** Quarto handles citation rules compliance and integrates seamlessly with citation management systems, enabling researchers to manage literature references and bibliographies more efficiently and consistently than in Word.
- **Easy cross-referencing:** Creating cross-references to sections, tables, and figures is straightforward.
- **Direct data analysis and output generation:** Data analysis and output generation occur directly within Quarto, ensuring that displayed graphics and tables are always up-to-date. This eliminates the need for manual post-processing and guarantees the reproducibility of results.
- **Embedded data visualizations:** Researchers can embed data visualizations directly in the text without manual intermediate steps.
- **Efficient collaboration with version control:** Version control systems like Git make collaboration on academic documents more manageable. Changes can be tracked and integrated without relying on complex and conflict-prone comparison tools.

Reading recommendation

For those interested, I recommend the online course Introduction to Reproducible Publications with RStudio, which explains explicitly how to work in an empirically reproducible manner. A somewhat more compact introduction is offered by Bauer and Landesvatter [2023], and the authoritative work on the subject is by Gandrud [2020].

6.2. Markdown

Markdown is a lightweight markup language with plain-text formatting syntax. It is very popular because it is easy to learn. It's an essential skill for using Quarto effectively. Start by learning enough Markdown to structure your thesis, including headings, lists, links, and code blocks.

You can learn Markdown (not R Markdown!) in 10 minutes. Just go to www.markdowntutorial.com and work through the interactive lessons. I also recommend the introduction offered in the section Markdown Basics on quarto.org.

6.3. Quarto

Recommended literature

Read Telford [2024]: Enough Markdown to Write a Thesis. This resource covers the basics and some advanced Markdown features that are useful for academic writing.
More extensive resources on how to do things with Quarto can be found at quarto.org.

6.3.1. Introduction

To set up Quarto on your machine do the following:

- Install R and R Studio.

- Install Quarto as follows:

```
install.packages("quarto")
```

- Install the tinytex package to generate PDF files:

```
install.packages("tinytex")
tinytex::install_tinytex()
```

- It is also advisable to install additional packages that might be needed later:

```
if (!require(pacman)) install.packages("pacman")
pacman::p_load(knitr, rmarkdown, papaja)
```

Exercise 6.1. First Quarto document

- Open RStudio.
- Select "File" -> "New File" -> "Quarto Document" and then "Create."
- Save the new file in an empty folder and set this folder as your working directory.
- Click "Render."
- Visit the Markdown Basics website, add some Markdown to your document, and click "Render" again.
- Click the arrow next to the "Render" button. Here, you can select and generate other file formats. Give it a try.
- Consult the PDF Basics website and supplement your header with the information found there.
- Try citing the paper by Huber and Rust [2016], which you can find here, in your document.
  - Click on "Visual,"
  - Go to the place in the text where you want to cite the paper and select "Insert" -> "Citation."
  - Search for the paper in the context menu using the corresponding DOI (https://doi.org/10.1177/1536867X1601600209) and insert it.
- To quote using APA Version 7 style, write the following in the YAML header:

```
csl: "https://www.zotero.org/styles/apa"
```

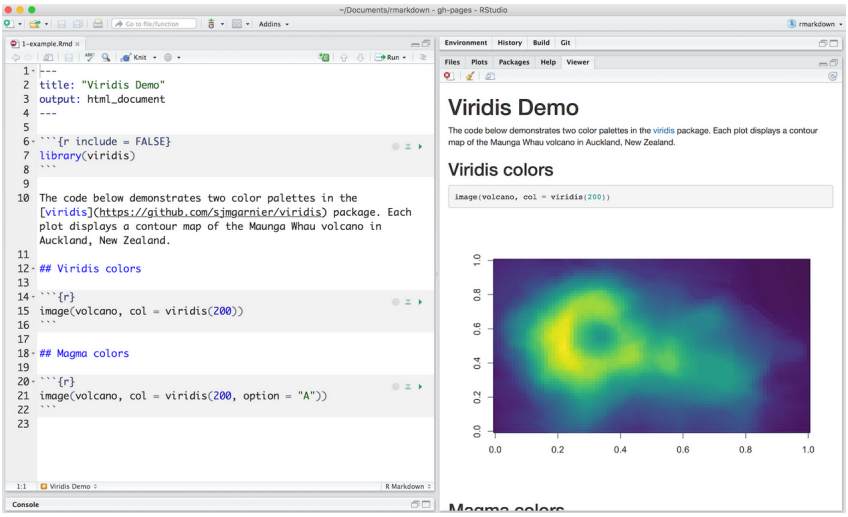
- Select a different citation style from www.zotero.org/styles. Then render the document again and observe the differences.

6.3.2. Create an APA compliant manuscript using Quarto

To create an APA compliant manuscript, it is recommended to use the *Quarto Extension* `apaquarto`. The extension that comes with nice templates is hosted on GitHub here. The installation and the usage is described in detail here. Using the template ensures that all APA rules are automatically considered. As APA allows a lot of leeway and every reviewer has specific preferences, `apaquarto` allows for manipulation of a variety of settings. For example, the language can be changed and the general style of the document can be modified in the Preamble (YAML header).

6.4. R Markdown

Figure 6.1.: Example of an R Markdown file

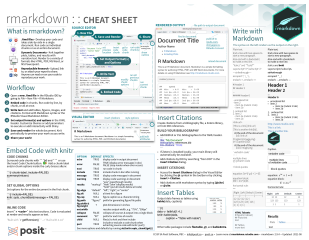


R Markdown provides an authoring framework for data science. You can use a single R Markdown file to transcript your work, run code, and generate high quality reports, books, websites, articles, theses, blogs, and many more (see Figure 6.1).

In contrast to Quarto (see Chapter 6), which is the more recent format, R Markdown is around for some time and hence there are uncountable resources to learn it. For example:

- The R Markdown Cheatsheet (see Figure 6.2) from Posit offers an overview on the most important features of R Markdown.

Figure 6.2.: R Markdown Cheatsheet from Posit



- The book *R Markdown Cookbook* by Xie et al. [2020] (see Figure 6.3) offers an introduction. The online version of the book is regularly updated and free of costs.

- The book *R Markdown: The Definitive Guide* by Xie et al. [2018] offers a comprehensive introduction. The online version of the book is regularly updated and free of costs.

Please watch the video What is R Markdown? and then study the R Markdown tutorial from RStudio.

> **Working directory in R Markdown**
>
> The working directory is by default set to the directory that contains the Rmd document. In case you want to use another directory you can do so by changing the working directory with `setwd()`. However, that is not persistent in R Markdown and only works for the

Figure 6.3.: Xie et al. [2020]: R Markdown Cookbook

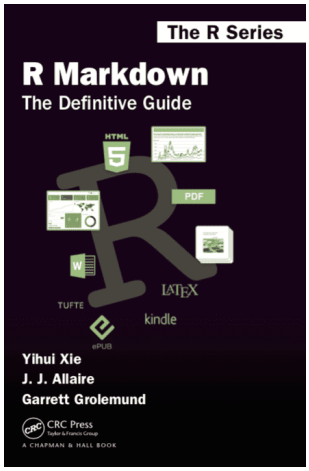
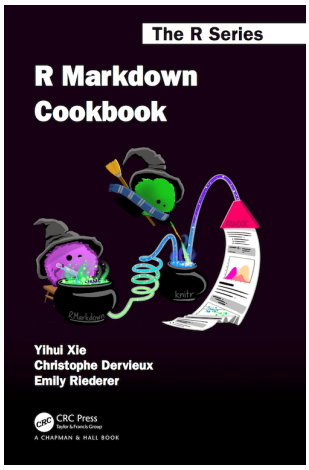


Figure 6.4.: Xie et al. [2018]: R Markdown: The Definitive Guide



current code chunk. After the code chunk has been evaluated, the working directory will be restored to the directory where the Rmd file is placed.

Exercise 6.2. Start Markdown and R Markdown

a) You can learn Markdown (not R Markdown!) in 10 minutes. Just go to https://www.markdowntutorial.com and work throught the interactive lessons.
b) Now create your first R Markdown file in 3 minutes by doing the following:

- click in RStudio on *File > New File > R Markdown*
- click *OK*
- look for a button entitled *Knit* and click it
- save your file (it will be saved with .Rmd file extension)

c) Play around with the file. For example, change the output format can you create a word file or a presentation. Play around with the code chunks. Add a picture that you find somewhere online.
d) Set your working directory to the folder where you have saved your first Rmd-file. Can you come up with a way to generate different output format with just one function.

Exercise 6.3. R Markdown cite literature

a) Create a new R Markdown file (*File > New File > R Markdown*), save the file in an empty folder, and knit it.
b)

- Make a new script with *File > New File > R Script*.
- Go to https://scholar.google.de/ and search for *osrmtime*.
- Click on *"cite"* and *"BibTeX"*. Copy and paste everything that you see into your script and save the script as *lit.bib*. R Studio will ask you if you confirm the file type change. Click yes. Your *lit.bib* file should look like this:

```
@article{huber2016calculate,
  title={Calculate travel time and distance with OpenStreetMap
    data using the Open Source Routing Machine (OSRM)},
  author={Huber, Stephan and Rust, Christoph},
  journal={The Stata Journal},
  volume={16},
  number={2},
  pages={416--423},
  year={2016},
  publisher={SAGE Publications Sage CA: Los Angeles, CA}
}
```

- Add the text *"bibliography: references.bib"* to your YAML header of your R Markdown file so that it looks somehow like that:

```
---
title: "Untitled"
author: "Stephan Huber"
date: "`r Sys.Date()`"
output: html_document
bibliography: lit.bib
---
```

- Now you can cite the OSRMTIME paper with `@huber2016calculate` somewhere in the text of your R Markdown file.
- Knit the R Markdown file and you should see the paper cited and a reference list at the end of the html report.
- You can manipulate the citation style you can specify a CSL (Citation Style Language) file in the YAML header. For example the APA style can be chosen with:

```
csl: "https://www.zotero.org/styles/apa.csl"
```

Many more citation styles can be found on github.com/citation-style-language and on the Zotero Style Repository.

Exercise 6.4. Preparing APA journal articles (`papaja`)
There is an easy way to write a manuscript that follows all the APA rules using the package `papaja` written by two psychologists from Cologne. Please read their manual and consider their repository on GitHub.
Now, install and load the package:

```
install.packages("papaja")
library("papaja")
```

Then, click "File > New File > R Markdown" and choose the "APA-style manuscript" from the section "from template". Knit the R markdown template and you will have a template for a APA manuscript.
Apart from the obvious adjustments, I recommend to make at least two general adjustments: Change `classoption` to "doc" and `linenumbers` to "no".

Exercise 6.5. R Markdown template
Please follow the instructions below to access the file "23-09_ds-project-desc.Rmd" from my GitHub account:

1. Download the file from my GitHub account by clicking on the link provided here.
2. Save the file in your working directory.
3. Use the knit function to run the file, but be aware that it may not work properly at first. If you encounter any issues, troubleshooting may be required. Don't worry, error messages will usually provide guidance to help you resolve the issue. Please note that the YAML header is sensitive to spacing, so be careful when setting it up to avoid breaking the code.
4. In the project template, I have used BibTeX to cite literature. This method is excellent for automating tedious tasks such as citing papers and generating reference lists based on citation styles, saving time and reducing the likelihood of citation

errors. The literature cited is in a separate file, which can be found on one of my GitHub repositories.

7. Create and host a website

7.1. Creating a website with Quarto

This tutorial guides you through creating a simple, yet professional-looking website using Quarto.

Step W1: Install Quarto

Ensure Quarto is installed on your system. If not, download and install it from Quarto's official website.

Step W2: Create a website

Follow the tutorial that you find here.

Step W3: Copy the `_site` directory

After you have rendered your website a directory "`_site`" appears in the project folder that contains your website. Copy all files of that directory to a directory where you want to save your website. Let's say `my_website`.

In the terminal you can do this with

```
mkdir /home/sthu/my_website/
cp -r /home/sthu/quarto_website/_site/* /home/sthu/my_website/
```

7.2. Hosting the website on GitHub

R Studio and Quarto offers you various ways to publish the website. I explain you a way that worked out well for me.

Step G1: Create a GitHub account

GitHub will host your thesis website and manage version control for your thesis project. If you don't already have a GitHub account, you'll need to create one: Sign up at GitHub.

Step G2: Create a repository

Create a repository. Name the repo with your username followed by *github.io*. You find a tutorial here.

Step G3: Obtain a personal access token

A personal access token (PAT) is required to authenticate with GitHub from Quarto and RStudio. This token allows you to push changes to your repository securely. Follow the instructions to create a personal access token on GitHub. Alternatively, you can do the following in R:

```
if (!require(pacman)) install.packages("pacman")
pacman::p_load(usethis)
create_github_token()
```

Make sure to note down your token and keep it secure. You'll use this token in RStudio and Quarto to authenticate your GitHub operations.

Step G4: Install and Learn Git

See Section 5.2.

Step G5: Upload the website to GitHub

Use the Terminal of R Studio. Go to the directory with your website that you have copied in Step W3. Then initiate a git repository on the command line, connect it to the repository created in Steph G2 on GitHub and finally push it:

```
cd /home/sthu/my_website/
echo "# test" >> README.md
git init
git add README.md
git commit -m "first commit"
git branch -M main
git remote add origin https://github.com/test-hsf/test.git
git push -u origin main
```

Alternatively, you can clone a repository, make some changes, and then push those changes back to GitHub. Here are the Bash commands to accomplish this:

```
# Clone the repository
git clone https://github.com/your-username/your-repository.git

# Make changes, here adding a new file as an example
echo "Some content for the new file" > newfile.txt

# Add the new file to the repository
git add newfile.txt

# Commit the changes
git commit -m "Add new file"

# Push the changes back to GitHub
git push origin main
```

References

APA. *Publication Manual of the American Psychological Association*. American Psychological Association, Washington, D.C., 7 edition, 2020.

Paul C. Bauer and Camille Landesvatter. Writing a reproducible paper with rstudio and quarto. Technical report, 2023. URL https://doi.org/10.31219/osf.io/ur4xn.

Carl T Bergstrom and Jevin D West. *Calling Bullshit: The Art of Skepticism in a Data-Driven World*. Penguin Books, 2021.

Dominik Brendel. Purpose: A foundational theory for an up-and-coming topic. *Marketing Review St. Gallen*, 37(4):10–16, 2020.

Jennifer Bryan. Happy git and github for the user. URL https://happygitwithr.com/.

Tom Chivers and David Chivers. *How to Read Numbers: A Guide to Statistics in the News (And Knowing When to Trust Them)*. Weidenfeld & Nicolson, 2021.

John Cochrane. Writing tips for Ph. D. students. online, 2005. URL https://www.johnhcochrane.com/s/phd_paper_writing.pdf.

Thomas H Davenport and DJ Patil. Data scientist: The sexiest job of the 21st century. *Harvard Business Review*, 90(5):70–76, 2012.

Jack Dougherty and Ilya Ilyankou. *Hands-On Data Visualization Interactive Storytelling from Spreadsheets to Code*. O'Reilly, 2021. URL https://handsondataviz.org/.

Jordan Ellenberg. *How Not To Be Wrong: The Power of Mathematical Thinking*. Penguin Books, 2015.

Stephen P. Fitzgerald. *Decision Making*. Capstone Publishing, 2002.

Christopher Gandrud. *Reproducible research with R and R studio*. Chapman and Hall/CRC, 3 edition, 2020.

Aud Halbritter and Richard J Telford. Version control with git and github. URL https://biostats-r.github.io/biostats/github/.

Tim Harford. *How to Make the World Add Up: Ten Rules for Thinking Differently about Numbers*. The Bridge Street Press, 2020.

Shere Hite. *The Hite Report. A Nationwide Study of Female Sexuality*. New York: Dell, 1976.

Stephan Huber. Quantitative methods, 2024a. URL https://hubchev.github.io/qm/.

Stephan Huber. How to use R for data science, 2024b. URL https://hubchev.github.io/ds/.

Stephan Huber and Christoph Rust. Calculate travel time and distance with openstreetmap data using the open source routing machine (osrm). *The Stata Journal*, 16(2):416–423, 2016.

Darrell Huff. *How to Lie with Statistics*. WW Norton & company, 1954.

Nick Huntington-Klein. *The Effect: An Introduction to Research Design and Causality*. CRC Press, 2022. URL https://theeffectbook.net.

Ben Jones. *Avoiding Data Pitfalls: How to Steer Clear of Common Blunders When Working with Data and Presenting Analysis and Visualizations*. John Wiley & Sons, Hoboken, New Jersey, 2020.

Robert Kabacoff. *Modern Data Visualization with R*. Chapman and Hall/CRC, 2024. URL https://rkabacoff.github.io/datavis/.

Roger Martin. How successful leaders think. *Harvard Business Review*, 85(6):71–81, 2007. URL https://hbr.org/2007/06/how-successful-leaders-think.

Plamen Nikolov. Writing tips for crafting effective economics research papers – 2023-2024 edition. Discussion Paper Series 16276, Institute of Labor Economics (IZA), 2023. URL https://hdl.handle.net/10419/278974.

David Spiegelhalter. *The Art of Statistics: Learning From Data*. Penguin UK, 2019.

Josh Starmer. *The StatQuest Illustrated Guide To Machine Learning*. Independently published, 2022.

William Strunk Jr. and E.B. White. *The Elements of Style*. Pearson, 4 edition, 1999.

Richard J Telford. Enough markdown to write a thesis, 8 2024. URL https://biostats-r.github.io/biostats/quarto/.

Daniel Vaughan. *Analytical skills for AI and data science*. O'Reilly Media, Sebastopol, CA, June 2020.

Hadley Wickham and Garrett Grolemund. R for data science (2e), 2023. URL https://r4ds.hadley.nz/.

Wikipedia. Survivorship bias, 2024. URL https://en.wikipedia.org/wiki/Survivorship_bias.

Jeffrey M. Wooldridge. *Introductory Econometrics: A Modern Approach*. South-Western, 2nd edition, 2002.

Yihui Xie, Joseph J. Allaire, and Garrett Grolemund. *R Markdown: The Definitive Guide*. Chapman and Hall/CRC, 2018.

Yihui Xie, Christophe Dervieux, and Emily Riederer. *R Markdown Cookbook*. Chapman and Hall/CRC, 2020. available at https://bookdown.org/yihui/rmarkdown-cookbook.

William Zinsser. *On Writing Well: The Classic Guide to Writing Nonfiction*. Harper Perennial, 30th anniversary, paperback reprint edition, 2016.

A. Presentation and Handout Guidelines

A.1. Assessment methods and criteria

Students complete this module with an academic presentation [Brendel, 2020]. Mode of the presentation is a in class presentation that takes place during the lecture period. The exact dates are set by the lecturer and communicated via ILIAS. The presentation lasts for a minimum of 10 minutes per student. In addition, a handout should be produced outlining the key theses of the presentation and the literature on which it is based. The length of the handout should be 3-5 pages of text. The handout should be submitted to the lecturer by the date of the presentation at the latest.

The presentation contributes 65 % to the module grade, the handout contributes 35 %. A passing grade in this module is achieved when the overall grade is greater than or equal to 4.0.

A.2. Topics

The following topics primarily correspond to specific chapters from the book by Spiegelhalter [2019]. The dates when each presentation will take place are shown in Table A.1. Topics will be randomly assigned to students who have officially registered. I will communicate the assignment on October 17 in class and a bit later on ILIAS.

1. Why are we looking at data anyway? Populations and measurements [Spiegelhalter, 2019, ch. 3]
2. What causes what? [Spiegelhalter, 2019, ch. 4]
3. What is machine learning? [Starmer, 2022, ch]
4. Algorithms, analytics and prediction [Spiegelhalter, 2019, ch. 6]
5. Data visualization with R [Kabacoff, 2024]
6. How sure can we be about what is going on? Estimates and intervals [Spiegelhalter, 2019, ch. 7]
7. Data manipulation with the R package dplyr [Wickham and Grolemund, 2023]
8. Probability the language of uncertainty and variability [Spiegelhalter, 2019, ch. 8]
9. Putting probability and statistics together [Spiegelhalter, 2019, ch. 10]
10. Answering questions and claiming discoveries [Spiegelhalter, 2019, ch. 11]
11. Learning from experience the Bayesian Way [Spiegelhalter, 2019, ch. 12]
12. How things go wrong [Spiegelhalter, 2019, ch. 13]
13. How we can do statistics better [Spiegelhalter, 2019, ch. 14]

Table A.1.: Topics and dates

Date	Topic 1	Topic 2
14.11.2024	Why are we looking at data anyway?	What causes what?
21.11.2024	What is machine learning?	Algorithms, analytics and prediction

Date	Topic 1	Topic 2
28.11.2024	Data visualization with R	How sure can we be about what is going on?
05.12.2024	Data manipulation with the R package dplyr	Probability the language of uncertainty and variability
12.12.2024	Putting probability and statistics together	Answering questions and claiming discoveries
19.12.2024	Learning from experience the Bayesian Way	How things go wrong
tba.	How we can do statistics better	

A.3. Content of the presentation and handout

Improve the understanding of your fellow students by providing insights into the topic assigned to you. Given your limited presentation time, it is important that you choose the content carefully. One of your challenges is to identify the most important elements of your topic and prepare them in a way that they can be communicate and understood within the time given. In some cases, your presentation may serve to enlighten the curiosity of the audience and encourage them to study the topic in greater depth on their own.

Your handout should complement the presentation.

While you are responsible for the content overall, your submission should include an exercise related to the content you have covered at best with an application of the programming language R. This task should be completed by the students during class immediately following your presentation. This interactive element will enhance engagement and reinforce the material discussed.

A.4. Form of the presentation and the handout

Create a presentation and a handout using Quarto in the format of a standalone website (HTML) and a PDF file. Publish both the website and the handout on GitHub, and share the URLs with the audience so they can access both during your talk. Additionally, please submit the four files on ILIAS. For guidance on creating a standalone HTML file, refer to this resource. Additionally, I will explain how Quarto and GitHub work in the lecture.

The design and the layout of the presentation slides are your choice. However, please avoid trying to impress with elaborate layouts or extraneous details. This is an academic presentation, and distracting decorations are inappropriate. Your primary focus should be on effectively communicating information, facts, and insights to the reader. Feel free to include any elements that support this goal. In the presentation, tables and figures don't need to be numbered.

The layout of the handout should be like it is a convention in academic writing as simple as possible. In the handout, all tables and figures should be numbered and the text should refer to them. In case of doubts, please apply the rules and conventions explained in APA [2020]. You can use the QMD file of this guide as a template. We will discuss that in the class in greater detail. Moreover, for your handout, please adhere to the following formatting rules:

- Use the documentclass set to article,
- on A4 paper,

- set the font size to 11pt, and
- specify the margins with 30mm at the top and bottom and 20mm on the left and right.

These guidelines will help maintain a consistent and organized appearance throughout your document.

In any academic work, all sources must be cited. In the presentation and the handout, use the APA [2020] citation style, following the 7th Edition. Specifically, use the CSL file (Citation Style Language) available here.

A.5. General tips

I refrain from specific advice on good writing, structuring your work, or adhering to academic rules and conventions. Therefore, I recommend the following resources:

- Nikolov [2023]: *Writing Tips for Crafting Effective Economics Research Papers*
- Cochrane [2005]: *Writing Tips for Ph. D. Students*
- Zinsser [2016]: *On Writing Well: The Classic Guide to Writing Nonfiction*
- Strunk Jr. and White [1999]: *The Elements of Style*

Nikolov [2023] and Cochrane [2005] offer excellent and concise guides with numerous tips you shouldn't miss. Zinsser [2016] elaborates on how to write in a convincing, clear, and appealing manner. The long-seller Strunk Jr. and White [1999] focuses on grammar and language and is available online here or slightly abridged here.

Moreover, I recommend reading "My Five Cents on How to Write a Thesis" that you find here.