Data Science for Business Leaders

Lecture Notes (currently under revision)

Prof. Dr. Stephan Huber

April 19, 2024

Table of contents

List of figures

List of tables

Preface

This is a Quarto book.

To learn more about Quarto books visit https://quarto.org/docs/books.

```
1 + 1
```

```
[1] 2
```

1. (Extended) Syllabus

Some time ago, I wrote the syllabus of this course, see Section 1.1. that was successfully accredited by the state. Since the syllabus is quite short, I thought about an extended version to help students familiarize themselves with the buzzwords I used in the original version. As a data scientist who is supposed to teach you how to use the tools of data science in business, I wondered how I could write it without devoting too many resources to it, that is, being economical. I thought it would be a good idea and a fun exercise to let ChatGPT from OpenAI [2024] write it for me. ChatGPT uses a large language model to generate text based on a vast corpus of text data. I provided ChatGPT with specific prompts to craft the extended abstract, which you can find in Section 1.2. The final extended syllabus, actually written by ChatGPT, is available in Section 1.3.

1.1. Syllabus

Scope and Nature of Data Science

- Defining data science as an academic discipline (informatics, computer science, mathematics, statistics, econometrics, social science)
- Importance of data science in businesses

Emerging Trends in a Data-Driven Business Environment

- Evolution of computers, computing, and data processing
- Business intelligence (performance marketing, etc.)
- Artificial intelligence, machine learning, deep learning, and algorithms
- Big data
- Internet of things, cloud computing, blockchain
- Industry 4.0 and remote working

Data Science Process in Business

- Workflows and data science life cycles (OSEMN, CRISP-DM, Kanban, TDSP, …)
- Types of data science roles (data engineer, data analyst, machine learning engineer, business intelligence analyst, database administrator, data product manager, …)

Data Literacy

- 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 and compare them with other findings considering ethical issues and scientific standards)

Overview of Data Science Methods

- Data exploration and data mining
- Supervised and unsupervised learning
- Regression and classification
- Predictive analysis
- Causal analysis

Introduction to Data Scientific Tools

- Writing and publishing reports (Markdown, Quarto)
- Collaborating in teams using a version control system (git)
- Overview of programming languages (R, Python, SQL, …)
- Overview of no-code and low-code tools for data science (makeML, PyCaret, Rapidminer, KNIME, etc.)
- Development environments (Unix-like systems, containers, APIs, Jupyter, Rstudio, etc.)

1.2. How I let ChatGPT wrote the extended syllabus

Here are the seven prompts to ChatGPT 4.0:

1. Act as a professor of data science. Write lecture notes for management students. The first chapter of the notes should contain the following:

Scope and nature of data science

- Defining data science as an academic discipline (informatics, computer science, mathematics,statistics, econometrics, social science)
- Importance of data science in businesses

2. The chapter of the lecture notes should contain the following:

Emerging Trends in a Data-Driven Business Environment

- Evolution of computers, computing, and data processing
- Business Intelligence (Performance Marketing, etc.)
- Artificial intelligence, machine learning, deep learning, and algorithms
- Big data
- Internet of things, cloud computing, blockchain
- Industry 4.0 and remote working

3. The third chapter of the lecture notes should contain the following:

Data science process in business

- Workflows and Data science life cycles (OSEMN, CRISP-DM, Kanban, TDSP, …)
- Types of data science roles (data engineer, data analyst, machine learning engineer, business intelligence analyst, database administrator, data product manager, …)

4. Do the same for the fourth section, which contains:

Data literacy

- 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 Kdata)

- Data evaluation (plan, conduct, evaluate, and assess data analyses)
- Data application (share, reflect, and evaluate results of analyses and compare them with other findings considering ethical issues and scientific standards)

5. And the next section:

Overview of data science methods

- Data exploration and data mining
- Supervised and unsupervised learning
- Regression and classification
- Predictive analysis
- Causal analysis

6. And the final section:

Introduction to data scientific tools

- Writing and publishing reports (Markdown, Quarto)
- Collaborating in teams using a version control system (git)
- Overview on Programming languages (R, Phyton, SQL, …)
- Overview on no-code and low-code tools for data science (makeML, PyCaret, Rapidminer, KNIME, etc.)
- Development environments (Unix-like systems, containers, APIs, Jupyter, Rstudio, etc.)

7. I am not happy with the third section. In particular the section "Workflows and Data Science Life Cycles". Can you re-write it? Consider the learning objective a bit better, that is, summarize the strengths, weaknesses and use-cases of data science processes and workflows and demonstrate the ability to justify and explain tools and methods used along the process.

I manually cut some parts of the last answer and pasted it into the respective section.

I assembled everything by copy and paste it step-by-step into Quarto and render it. That was all I did to create the notes.

Exercise 1.1. What do think about the extended syllabus and my approach
What do you think about the extended syllabus? Is it well-written? Does it include a lot of details or errors? What aspects do you think we should work on manually? How could my prompts be improved to yield better results?

1.3. The ChatGPT generated extended syllabus

The following text was generated as described in Section 1.2 using OpenAI [2024].

1.3.1. Scope and Nature of Data Science

Welcome to the introductory chapter on Data Science, designed specifically for management students. In this chapter, we will explore the multifaceted discipline of data science, understanding its definitions, scope, and the pivotal role it plays in the business world today.

1.3.1.1. Defining Data Science

Data Science is an interdisciplinary field that uses scientific methods, processes, algorithms, and systems to extract knowledge and insights from structured and unstructured data. It is a confluence of several disciplines including informatics, computer science, mathematics, statistics, econometrics, and social science. This integration allows for a comprehensive approach to solving complex problems, making informed decisions, and generating actionable insights through data analysis.

1.3.1.1.1. Informatics

It is the study of computational systems and the way humans interact with these systems. In data science, informatics plays a critical role in the management and processing of data.

1.3.1.1.2. Computer Science

This foundational pillar focuses on the development of algorithms and software that are used to process, store, and analyze data.

1.3.1.1.3. Mathematics and Statistics

Mathematics provides the theoretical foundation for algorithms and models, while statistics is crucial for understanding and interpreting data. Together, they enable data scientists to perform quantitative analysis and predictive modeling.

1.3.1.1.4. Econometrics

Econometrics applies statistical methods to economic data, enabling the analysis of economic relationships. In data science, econometrics helps in forecasting economic trends and evaluating economic policies through data.

1.3.1.1.5. Social Science

The incorporation of social science principles allows data scientists to understand human behavior and societal trends, which is invaluable for businesses looking to market their products or services effectively.

1.3.1.2. Importance of Data Science in Businesses

Data Science has become a cornerstone in the business world, offering unprecedented opportunities for growth, efficiency, and innovation. Here's how:

1.3.1.2.1. Data-Driven Decision Making

Businesses can make more informed decisions by analyzing vast amounts of data. This leads to better strategic planning, operational efficiency, and competitive advantage.

1.3.1.2.2. Customer Insights

Through data analysis, companies can gain deep insights into customer behavior, preferences, and trends. This information is critical for tailoring marketing strategies, improving customer experiences, and boosting customer loyalty.

1.3.1.2.3. Risk Management

Data science enables businesses to identify and mitigate risks by analyzing historical data, detecting patterns, and predicting future outcomes. This is particularly useful in finance, insurance, and cybersecurity.

1.3.1.2.4. Innovation and Product Development

By understanding market needs and customer feedback through data analysis, businesses can innovate and develop new products and services that meet or exceed customer expectations.

1.3.1.2.5. Operational Efficiency

Data science can optimize business operations, reduce costs, and improve service delivery through the analysis of operational data, leading to streamlined processes and better resource management.

1.3.1.2.6. Competitive Edge

In a data-driven world, businesses that leverage data science effectively can stay ahead of the curve, identifying trends and opportunities before their competitors.

In conclusion, data science is an essential academic discipline that integrates various fields to analyze and interpret complex data. Its importance in the business sector cannot be overstated, as it enables organizations to make informed decisions, understand their customers, manage risks, foster innovation, improve operations, and maintain a competitive edge in the market. As we delve deeper into this course, we will explore the methodologies, tools, and applications of data science in greater detail, equipping you with the knowledge and skills to leverage data science in your future management roles.

1.3.2. Emerging Trends in a Data-Driven Business Environment

In this chapter, we explore the dynamic landscape of a data-driven business environment. The rapid evolution of technology and its integration into business processes have led to the emergence of new trends that are shaping the future of industries across the globe. Let's dive into these trends and understand their significance.

1.3.2.1. Evolution of Computers, Computing, and Data Processing

The journey from the first generation of computers to the modern era of quantum computing marks a significant evolution in technology. Initially, computers were large, expensive, and limited in their capabilities. Over the decades, advancements in semiconductor technology, the invention of the microprocessor, and the development of personal computers transformed computing into an accessible and essential tool for businesses. Today, cloud computing and edge computing have further revolutionized data processing, allowing for more efficient data storage, access, and analysis.

1.3.2.2. Business Intelligence

Business Intelligence (BI) refers to the use of data analysis in business to support decision-making processes. BI tools analyze historical and current data to provide actionable insights, helping businesses to improve their performance. Performance marketing, a subset of BI, focuses on analyzing marketing campaigns in real time to optimize marketing strategies and expenditures for better ROI.

1.3.2.3. Artificial Intelligence, Machine Learning, Deep Learning, and Algorithms

Artificial Intelligence (AI) and its subsets, Machine Learning (ML) and Deep Learning (DL), are at the forefront of technological innovation. AI involves creating systems capable of performing tasks that typically require human intelligence. ML and DL are about teaching computers to learn from data, improving their accuracy over time without being explicitly programmed. These technologies are transforming business operations, from customer service automation and predictive analytics to personalized marketing and decision-making processes.

1.3.2.4. Big Data

Big Data refers to the vast volumes of data generated every second from various sources like social media, business transactions, and IoT devices. The ability to process and analyze this data has unlocked new opportunities for businesses to gain insights into customer behavior, market trends, and operational efficiency. Big data analytics is now a crucial tool for strategic planning and competitive analysis.

1.3.2.5. Internet of Things, Cloud Computing, Blockchain

- **Internet of Things (IoT)**: IoT technology connects everyday devices to the internet, enabling them to send and receive data. This interconnectivity offers businesses real-time insights into their operations, asset tracking, and supply chain management.
- **Cloud Computing**: Cloud computing provides businesses with scalable computing resources over the internet, facilitating remote data storage, processing, and management. It supports flexibility, reduces IT costs, and enhances collaboration.
- **Blockchain**: Blockchain technology offers a secure, decentralized platform for transactions. It's particularly valuable for enhancing transparency, security, and efficiency in business operations, supply chain management, and financial services.

1.3.2.6. Industry 4.0 and Remote Working

Industry 4.0, also known as the fourth industrial revolution, integrates digital technologies into manufacturing and industry, including IoT, AI, and robotics. It represents a shift towards smart, automated production processes and data exchange. Coupled with the rise of remote working, Industry 4.0 technologies enable businesses to operate more flexibly, with teams collaborating effectively from various locations, leveraging digital tools and platforms for communication and project management.

In conclusion, the evolution of technology and its integration into business practices have brought about significant changes in the way companies operate. From the way data is processed and analyzed to the automation of manufacturing processes and the flexibility of remote working, these emerging trends are shaping the future of a data-driven business environment. As we move forward, understanding and adapting to these trends will be crucial for businesses looking to thrive in the digital age.

1.3.3. Data Science Process in Business

This chapter delves into the structured approach behind the application of data science in business settings. We will explore various data science workflows and life cycles that guide the process from raw data to actionable insights. Additionally, we will outline the different roles within a data science team and their contributions to this process.

1.3.3.1. Workflows and Data Science Life Cycles

Data science projects in business environments follow structured workflows and life cycles to ensure that the analysis is efficient, reproducible, and scalable. Several frameworks guide these processes, each with its strengths and applications.

1.3.3.1.1. OSEMN Framework

OSEMN (Obtain, Scrub, Explore, Model, iNterpret) is a streamlined approach to data science projects:

1. **Obtain**: Acquiring the data from various sources.
2. **Scrub**: Cleaning the data to ensure it is accurate and usable.
3. **Explore**: Analyzing the data to find patterns and relationships.
4. **Model**: Applying statistical models to predict or classify data.
5. **Interpret**: Drawing conclusions and making recommendations based on the model's results.

- **Strengths**: The OSEMN (Obtain, Scrub, Explore, Model, iNterpret) framework is straightforward and easy to understand, making it accessible for teams of all skill levels. It covers the essential steps of a data science project in a logical sequence.
- **Weaknesses**: Its simplicity may overlook the complexity of certain stages, such as model validation or deployment.
- **Use-Cases**: Ideal for small to medium-sized projects where the primary goal is to gain insights from data through exploration and modeling.

1.3.3.1.2. CRISP-DM

CRISP-DM stands for Cross-Industry Standard Process for Data Mining. It's a comprehensive framework that includes six phases:

1. **Business Understanding**: Define the project objectives and requirements.
2. **Data Understanding**: Collect and explore the data.
3. **Data Preparation**: Clean and preprocess the data.
4. **Modeling**: Select and apply modeling techniques.
5. **Evaluation**: Assess the model's performance.
6. **Deployment**: Implement the model in a real-world setting.

- **Strengths**: CRISP-DM (Cross-Industry Standard Process for Data Mining) is industry-agnostic and provides a detailed structure that includes understanding the business problem and deploying the solution. It encourages iterative learning and refinement.
- **Weaknesses**: Can be perceived as too rigid for projects requiring rapid development and deployment. The model doesn't explicitly address the updating or maintenance of deployed solutions.
- **Use-Cases**: Suitable for projects that require close alignment with business objectives and thorough consideration of deployment strategies.

1.3.3.1.3. Kanban

Kanban is a lean method to manage and improve work across human systems. In data science, it helps in visualizing work, limiting work-in-progress, and maximizing efficiency.

- **Strengths**: Kanban is highly flexible and promotes continuous delivery. It allows teams to adapt quickly to changes and prioritize tasks effectively.
- **Weaknesses**: Without strict stages or phases, projects might lack direction or oversight, potentially leading to inefficiencies.
- **Use-Cases**: Best for dynamic environments where priorities shift frequently and teams must remain agile to respond to business needs.

1.3.3.1.4. TDSP (Team Data Science Process)

TDSP is a standardized approach to data science projects that helps teams to improve quality and efficiency. It includes:

- **Strengths**: TDSP offers a structured approach with a strong emphasis on standardized documentation and project management methodologies, facilitating collaboration and scalability.
- **Weaknesses**: Its comprehensive nature might introduce overhead and slow down smaller projects.
- **Use-Cases**: Ideal for larger teams working on complex projects that require coordination across different roles and departments.

1.3.3.2. Types of Data Science Roles

In a business environment, a data science team might consist of various specialized roles, each contributing uniquely to the data science process.

1.3.3.2.1. Data Engineer

Focuses on the design, construction, and maintenance of the systems that data analysts and data scientists use for their work. They ensure that data flows smoothly from source to database to analytics.

1.3.3.2.2. Data Analyst

Works on processing and performing statistical analysis on existing datasets. They interpret the data to help the business make more informed decisions.

1.3.3.2.3. Machine Learning Engineer

Develops algorithms and predictive models to solve specific business problems using machine learning techniques.

1.3.3.2.4. Business Intelligence Analyst

Analyzes data to provide insights that help businesses with strategic planning. They use BI tools to convert data into understandable reports and dashboards.

1.3.3.2.5. Database Administrator

Responsible for managing, backing up, and ensuring the availability of the data stored in an organization's databases.

1.3.3.2.6. Data Product Manager

Oversees the development of data-driven products or services, ensuring that they meet the users' needs and the business objectives.

In summary, the data science process in business involves a structured approach to turning data into actionable insights. This process is supported by various frameworks and relies on the collaboration of professionals in specialized roles. Understanding these aspects of data science is crucial for anyone looking to leverage this discipline in a business context.

1.3.4. Data Literacy

Data literacy is the ability to read, understand, create, and communicate data as information. It encompasses a broad range of skills necessary for effectively working with data, from the initial stages of data collection to the final stages of analyzing and sharing findings. In this chapter, we will break down the conceptual framework of data literacy and explore its various components in detail.

1.3.4.1. Conceptual Framework

At the heart of data literacy is a deep knowledge and understanding of how data can be used to make decisions, solve problems, and communicate ideas. This conceptual framework involves:

- **Understanding the nature of data**: Recognizing different types of data (quantitative vs. qualitative) and their sources.
- **Comprehending the applications of data**: Knowing how data can be used in various contexts to derive insights and inform decisions.

1.3.4.2. Data Collection

The first step in the data lifecycle involves identifying, collecting, and assessing data:

- **Identify**: Determining the data needed to answer a question or solve a problem.
- **Collect**: Gathering data from various sources, whether they are existing datasets or new data collected through surveys, experiments, or observations.
- **Assess**: Evaluating the quality of the data, including its relevance, accuracy, and completeness.

1.3.4.3. Data Management

Once data is collected, it must be managed effectively:

- **Organize**: Arranging data in a structured format that facilitates analysis.
- **Clean**: Removing errors or inconsistencies in the data.
- **Convert**: Transforming data into a format suitable for analysis.
- **Curate**: Selecting, annotating, and maintaining valuable data for current and future use.
- **Preserve**: Ensuring that data remains accessible and usable over time.

1.3.4.4. Data Evaluation

Evaluation is critical to understanding what the data signifies:

- **Plan**: Designing a methodology for analyzing the data.
- **Conduct**: Performing the analysis using appropriate statistical methods and tools.
- **Evaluate**: Assessing the quality and reliability of the analysis.
- **Assess**: Interpreting the results in the context of the research question or business problem.

1.3.4.5. Data Application

The final step involves applying the insights gained from data analysis:

- **Share**: Communicating findings to stakeholders through reports, presentations, or visualizations.
- **Reflect**: Considering the implications of the results and how they can inform future actions.
- **Evaluate results**: Comparing findings with those from other studies or data analyses to draw broader conclusions.

- **Ethical considerations**: Ensuring that the use of data respects privacy, confidentiality, and ethical standards.
- **Scientific standards**: Adhering to rigorous standards of validity, reliability, and objectivity in data handling and analysis.

In summary, data literacy is a comprehensive set of skills that enable individuals to navigate the complex world of data from collection to application. By understanding and applying the concepts outlined in this chapter, individuals can enhance their ability to make informed decisions, solve problems, and communicate effectively using data.

1.3.5. Overview of Data Science Methods

Data science encompasses a wide array of methods and techniques for analyzing data, drawing insights, and making predictions. This chapter provides an overview of some core data science methods, including data exploration, data mining, machine learning approaches, and various types of analyses.

1.3.5.1. Data Exploration and Data Mining

1.3.5.1.1. Data Exploration

Data exploration involves analyzing data sets to find initial patterns, characteristics, and points of interest without making any prior assumptions. It typically includes summarizing the main characteristics of the data through visualizations and statistics.

1.3.5.1.2. Data Mining

Data mining is the process of discovering patterns and knowledge from large amounts of data. The data sources can include databases, data warehouses, the internet, and other sources. Data mining techniques include clustering, classification, regression, and association rule learning.

1.3.5.2. Supervised and Unsupervised Learning

1.3.5.2.1. Supervised Learning

Supervised learning is a type of machine learning where the algorithm learns from labeled training data, helping to predict outcomes for unforeseen data. It is divided into two main categories: regression and classification.

1.3.5.2.2. Unsupervised Learning

Unsupervised learning involves training on data without labeled responses. The system tries to learn the patterns and the structure from the data without any supervision. Common unsupervised learning methods include clustering and dimensionality reduction.

1.3.5.3. Regression and Classification

1.3.5.3.1. Regression

Regression methods are used to predict a continuous outcome variable based on one or more predictor variables. The goal is to find the relationship between variables and forecast an outcome. Linear regression is one of the most basic types of regression analysis.

1.3.5.3.2. Classification

Classification methods are used to predict or identify the category to which a new observation belongs. Examples include spam detection in email service providers and customer churn prediction.

1.3.5.4. Predictive Analysis

Predictive analysis uses statistical algorithms and machine learning techniques to identify the likelihood of future outcomes based on historical data. It's used in various fields, including finance, healthcare, marketing, and weather forecasting, to make more informed decisions.

1.3.5.5. Causal Analysis

Causal analysis seeks to identify and understand the cause-and-effect relationships between variables. Unlike correlation, which merely indicates that two variables move together, causation establishes that a change in one variable is responsible for a change in another.

In conclusion, these methods and techniques form the backbone of data science, enabling professionals to extract valuable insights, make predictions, and inform decision-making processes. Understanding these methods is crucial for anyone looking to delve into data science or apply its principles in their field.

1.3.6. Introduction to Data Scientific Tools

The practice of data science requires not only a solid understanding of theories and methodologies but also proficiency in a variety of tools and technologies. This chapter introduces essential tools for writing and publishing reports, collaborating in teams, programming, as well as no-code and low-code platforms, and development environments.

1.3.6.1. Writing and Publishing Reports

1.3.6.1.1. Markdown

Markdown is a lightweight markup language with plain-text formatting syntax. Its simplicity and ease of conversion to HTML and other formats make it an ideal choice for writing and publishing reports, documentation, and articles.

1.3.6.1.2. Quarto

Quarto is an open-source scientific and technical publishing system built on Pandoc. It enables users to create dynamic and reproducible reports and articles that can include executable code from various programming languages, such as R and Python.

1.3.6.2. Collaborating in Teams Using a Version Control System

1.3.6.2.1. Git

Git is a distributed version control system that enables multiple developers to work together on the same project efficiently. It tracks changes in source code during software development, supporting collaboration and fostering code integrity.

1.3.6.3. Overview of Programming Languages

1.3.6.3.1. R

R is a programming language and free software environment for statistical computing and graphics, widely used among statisticians and data miners.

1.3.6.3.2. Python

Python is a high-level, interpreted programming language known for its simplicity and versatility. It has a wide range of libraries for data analysis, machine learning, and data visualization, making it a popular choice in data science.

1.3.6.3.3. SQL

SQL (Structured Query Language) is the standard language for managing and manipulating relational databases. It allows users to query, update, and manage data.

1.3.6.4. Overview of No-Code and Low-Code Tools for Data Science

1.3.6.4.1. makeML

A no-code platform for machine learning, makeML simplifies the process of training and deploying ML models without writing extensive code.

1.3.6.4.2. PyCaret

PyCaret is a low-code machine learning library in Python that automates machine learning workflows. It enables data scientists to perform end-to-end experiments quickly and efficiently.

1.3.6.4.3. Rapidminer

Rapidminer is a data science platform that provides an integrated environment for data preparation, machine learning, deep learning, text mining, and predictive analytics.

1.3.6.4.4. KNIME

KNIME is an open-source, graphical workbench for the entire analysis process: data access, data transformation, initial investigation, powerful predictive analytics, visualization, and reporting.

1.3.6.5. Development Environments

1.3.6.5.1. Unix-like Systems

Unix-like operating systems, including Linux and macOS, provide powerful tools and environments for software development and data science.

1.3.6.5.2. Containers

Containers, such as Docker, allow for the packaging of applications and their dependencies in a virtual container that can run on any Linux server, enabling easy deployment and scalability.

1.3.6.5.3. APIs

Application Programming Interfaces (APIs) enable different software applications to communicate with each other, facilitating data exchange and integration.

1.3.6.5.4. Jupyter

Jupyter Notebook is an open-source web application that allows you to create and share documents containing live code, equations, visualizations, and narrative text.

1.3.6.5.5. RStudio

RStudio is an integrated development environment (IDE) for R. It provides a user-friendly interface for coding, debugging, and visualizing data.

In summary, the array of tools and technologies available to data scientists is broad and varied, catering to different aspects of the data science workflow. From data manipulation and analysis to collaboration and report writing, mastering these tools is essential for effective data science practice.

Part I.

Writing and publishing

2. Markdown and Quarto

Verbal and non-verbal communication is important in business. This section is about writing and publishing texts, leaving out body language and writing skills. I will introduce some applications (Markdown, RMarkdown, Quarto) that data scientists often use to write and publish their work. I will also discuss the version control system *git* and the online platform *GitHub*, which can be used to create, store, manage and share files. These tools are the backbone of most data science collaborations. Once you master these tools, they can significantly enhance your efficiency and make your presentations more impactful, even if you are not directly involved in the field of data science.

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.

Step 1: Learn Markdown

Markdown is a lightweight markup language with plain-text formatting syntax. 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 https://www.markdowntutorial.com and work throught the interactive lessons.

Step 2: Learn Quarto

Read Telford [2023]: 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.

> **Quarto and R markdown**
>
> Quarto is a relatively new tool. It can be considered the successor to R Markdown, as it is built upon R Markdown. Consequently, almost all R Markdown documents are compatible with Quarto. However, Quarto includes several improvements over R Markdown that enhance its ease of use. For a detailed description of all the differences and similarities between the two, you can read this article.

3. Create and host a website

3.1. Creating a website with Quarto

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

Step 1: Install Quarto

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

Step 2: Create a website

Follow the tutorial that you find here.

3.2. 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 1: 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 2: Create a reopository

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

Step 3: 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.

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.

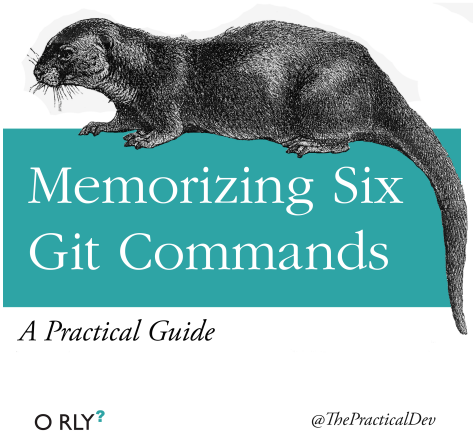
Step 4: Install and Learn Git

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

Familiarize yourself with Git using the resources available here. Although Git might seem complex, it's actually not too challenging for most users. Many people use Git primarily to track their work and to host and share files conveniently.

A humorous figure Figure 3.1 illustrates that while Git is a robust system with many capabilities, you don't need to remember all the commands. In fact, you typically use just a few basic commands. The table Table 3.1 lists the most important ones.

Figure 3.1.: Memorizing six git commands



Source: DEV Community on GitHub

Table 3.1.: 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.

Step 5: Upload the website to GitHub

Use the Terminal of R Studio. Go to the docs folder that is in the folder in which you have created the website. Then create a new repository on the command line:

```
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

OpenAI. Chatgpt (april 18, 2024). Large language model, 2024. URL https://chat.openai.comt.

Richard J Telford. Enough markdown to write a thesis, 9 2023. URL https://biostats-r.github.io/biostats/quarto/.