Project Submission Guidelines

Data Science for Business (SS 2025)

© Prof. Dr. Stephan Huber

April 25, 2025

This paper outlines the project requirements for the Data Science for Business course. It provides guidance for efficient progress and success, and explains the components and files required for submission.

Table of contents

1 Project description

Students complete this module with a project that contains

- a written report (10-15 written pages per student) and
- a presentation, lasting for 10-15 minutes per student with a subsequent discussion.

Students show that they are capable of describing the status of their work, their approach, findings and results. The presentation and subsequent discussion take place during the lecture period; the exact date is set by the lecturer. Group work is permitted. 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.

This year's project focuses on demonstrating the reproducibility of an empirical academic paper by accurately reproducing some of its empirical results. Students must consult with the lecturer to determine which parts of the chosen paper to replicate. This project is an opportunity for students to demonstrate their mastery of empirical research methodologies using the programming language R, as well as their proficiency with essential data science tools, including Markdown, Quarto, git, GitHub, and BibTeX.

2 Details about the things to do

2.1 Form a team and plan the project

In Table 1 I list the matriculation numbers and initials of students, to identify those assigned to each team, along with the date of their presentation.

Table 1: Teams, students, and dates

Team	Date	Matriculation number and initials
A	May 30	400885389 - AA
A	May 30	400890897 - KM
A	May 30	400412707 - MA
B	June 6	400804174 - OT
B	June 6	400858482 - PV
B	June 6	400791457 - RC
C	June 13	400793389 - SA
C	June 13	400776861 - VA
C	June 13	400767209 - UM

If you do not see your number on the list but have registered, please contact me as soon as possible. Note that the presentation day is an official examination day. Failure to attend will result in a grade of 5.0, so please ensure you do not miss this important date.

Data science is typically conducted as a team-based research effort. In business settings, team formation is often beyond your control, and it is the same in this context. Since teamwork is an essential skill you will likely need in your future career, consider this project as valuable practice.

Students have been randomly assigned to teams. Group work often comes with challenges, such as difficulties in connecting with teammates or concerns about team members underperforming

and relying on others' efforts. These issues are inherent to teamwork. To mitigate these challenges and enhance the quality of your project, it is crucial to plan effectively. This includes creating a clear outline of tasks and assigning each task to individual team members, along with transparent deadlines for submission.

As previously mentioned, each student's individual performance must be clearly defined and assessed using specific criteria, such as designated sections or page numbers. In your project report, every individual is required to write a brief paragraph reflecting on the group dynamics and the challenges of collaborating as a team. If you experience significant issues with any team members, please consult me. I will do my best to assist you in resolving these problems. Overall, effective project management is key to success.

2.2 Find a paper

Your task is to reproduce specific parts of the empirical results from a scientific paper published in a journal listed in the Scopus database. To access this list, please refer to the SCImago Journal Rank (SJR) on www.scimagojr.com

Once you have identified a candidate paper that you believe is feasible to reproduce, I recommend consulting with me to obtain approval for your choice. Occasionally, I may identify challenges that you might overlook, suggesting you find an alternative paper.

2.3 Conduct the reproduction study

Your task is to reproduce the paper's results using the programming language R. I like to emphasize that you don't have to replicate the full empirical analysis from scratch. Recognizing the constraints of time, fully reproducing every statistic, table, and graph from the paper might not be feasible, and that's perfectly acceptable. However, the parts that you reproduce should show that you can successfully use the programming language R and have understood the purpose of replication studies.

To ensure a focused and achievable project scope, please talk with me at least once during your study to align on what aspects are essential to replicate or investigate further. This consultation will help us stay in sync and clarify the priorities for your work.

I encourage you to reach out to me proactively instead of waiting for me to initiate contact to ensure that we are fully aligned and understand each other's expectations and progress.

2.4 Prepare a presentation and publish it on GitHub Pages

Create a presentation using (R) Markdown and **Quarto**, and **publish it as a website through GitHub**. Ensure the HTML file is a standalone website (HTML), and also provide a PDF version of your presentation. Publish the presentation files on GitHub, sharing the URLs with the audience for access during your presentation.

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.

Given the limited presentation time, prioritize key points to ensure you stay within the given timeframe without overly promoting yourself. Briefly describe and present the research paper and your plan focusing more extensively on the dataset utilized in your study. The presentation should serve as a progress report, highlighting ongoing work rather than concluded results.

If you encounter weaknesses or challenges in conducting your reproduction study, the presentation is an appropriate platform to share these. The presentation is not the occasion for showcasing success stories. Similar to an internal business meeting, the interest lies in understanding the hurdles you face, as this opens the door for constructive feedback and suggestions that could help overcome these challenges.

2.5 Write the report

The report

- must be written with Quarto,
- should contain 4000-5000 words, or approximately 15 double-spaced pages, and
- should be published and submitted in

    - html standalone format and
    - PDF format.

The layout of the report should adhere to the style guide of APA [2020]. Therefore, I recommend using the Quarto extension, apaquarto, which you can find here: https://github.com/wjschne/apaquarto. Ensure the HTML file is a standalone website (HTML), and also provide a PDF version of your report. The PDF of your handout should be formatted for A4 paper. For guidance on creating a standalone HTML file, refer to Quarto HTML Publishing.

Please note that this report is different from an academic paper in that it should focus solely on documenting, discussing, and presenting your project. Its purpose is to introduce your work to me in a way that is similar to reports written in business settings, where you focus on explaining what you did. Additionally, you should

- motivate your work and your procedure,
- mention briefly obstacles you overcame,
- discuss what challenges, problems and weaknesses remain, and
- suggest a strategy proceeding with your work if you would have had more time and resources.

Please refrain from trying to impress me with a fancy layout or any extraneous details. Your primary focus should be on effectively communicating your current state of work to the reader. Feel free to include anything that can help achieve this goal.

Please put some emphasize on guiding and motivating the reader. For example, the introduction is a good place to introduce the scope and content of the report. To ensure conciseness and clarity, please eliminate all unnecessary repetition. Take the time to read each sentence multiple times, asking yourself if it is concise, clear, and coherent with the surrounding text. Avoid language issues such as incorrect punctuation, incomplete sentences, or grammatical errors. I recommend using software tools to assist you in identifying and correcting these issues.

Incorporate all R code relevant to reproducing the empirical findings directly into your Quarto file using code chunks. Your QMD file(s) must document the complete workflow, encompassing data import, cleaning, and analysis. While all code should be included, it's not necessary to display every message and output generated by the code in the PDF document.

The outline of the paper must contain at least the following building blocks:

- Title and all common personal details (name, email, …).
- Abstract of the paper (which highlights the content of the report).
- All the R code that is necessary to replicate your results.
- A section where you explain briefly how you published your presentation on GitHub, see Section 2.6.
- A section where you explain briefly, how you made the pull request, see Section 2.6.
- The Affidavit, see Section 2.7.

> **!** Important
>
> In group work, it is essential that every student demonstrates their proficiency in writing R code.

2.6 Make a pull request with Git and GitHub

As mentioned above, you should publish your presentation using GitHub pages. Furthermore, you are required to make a pull request to my Github repository: make_a_pull_request. What you should do here in detail is explained in the README of the repo and in Huber [2025]. Remember to reference this pull request in your report.

> **!** Important
>
> In group work, every student must make a pull request.

2.7 Add this affidavit to your report

*Your report should contain the following **Affidavit**. Simply, fill it out and put it at the end of your report. You can check the box in Quarto like this:*

```
-   [x] I checked this box
```

I hereby affirm that this submitted paper was authored unaided and solely by me. Additionally, no other sources than those in the reference list were used. Parts of this paper, including tables and figures, that have been taken either verbatim or analogously from other works have in each case been properly cited with regard to their origin and authorship. This paper either in parts or in its entirety, be it in the same or similar form, has not been submitted to any other examination board and has not been published.

I acknowledge that the university may use plagiarism detection software to check my thesis. I agree to cooperate with any investigation of suspected plagiarism and to provide any additional information or evidence requested by the university.

The report includes:

- ☐ About 4000 words per student (+/- 500).
- ☐ The submission contains the Quarto file of the report.
- ☐ The submission contains the Quarto file of the presentation.
- ☐ The submission contains the HTML file of the report.
- ☐ The submission contains the HTML file of the presentation.
- ☐ The submission contains the PDF file of the report.

☐ The submission contains the PDF file of the presentation.
☐ The title page of the presentation and the report contain personal details (name, email, matriculation number).
☐ The report contains an abstract.
☐ The presentation and the report contain a bibliography, created using BibTeX with APA citation style.
☐ The report contains R code that proofs the student(s) expertise in coding.
☐ The report includes an introduction to guide the reader and a conclusion summarizing the work and discussing potential further investigations and readings, respectively.
☐ All significant resources used in the report and R code development.
☐ The filled out Affidavit.
☐ A concise description of the successful use of Git and GitHub, as detailed here: https://github.com/hubchev/make_a_pull_request.
☐ The link to the presentation and the handout published on GitHub.

[Your Name,] [Date,] [Place]

2.8 Submit the project

Submit at least the following files on ILIAS:

- Quarto file of your report.
- Quarto file of your presentation.
- HTML file of your report.
- HTML file of your presentation.
- PDF file of your report.
- PDF file of your presentation.
- Additional files that are required to understand, evaluate, and reproduce your work.

At best, try to upload all files in one .zip file.

> ❗ Important
>
> Please consider the deadline for academic papers and written assessments!

3 General tips

I refrain from specific advice on good writing, structuring your work, or adhering to academic rules and conventions. Therefore, I recommend reading Huber [2024] and the literature cited therein.

Since this is an academic presentation and work, respectively, please choose your sources carefully and ensure that you cite them appropriately.

If anything is unclear, please let me know as soon as possible. It is your responsibility to be well prepared on the day of your presentation. If you encounter any problems, let me know immediately and I will do my best to help you. However, please understand that I may have limited availability in the days or weeks leading up to your presentation. Prioritise the essential tasks that are technically complex or where something is likely to go wrong.

4 Evaluation

- *65 % – Quality and execution of the project –* After your presentation, we will discuss your work in a personal meeting. The goal of this conversation will be that we agree on certain standards by which I will grade you. By this I mean that we define certain goals that you should achieve with your data set and your question. The goal is to create a transparent set of expectations on my part. So that you have an indication of what you need to accomplish at a minimum in order to pass the course.

- *35 % – Quality and execution of the presentation*

- I will try to evaluate your work as objectively as possible. In particular, I will

  - check whether your submission is complete, or not,
  - check whether your empirical work can be reproduced,
  - check if all formal criteria are met,
  - check for plagiarism,
  - check if the replication of the paper was already done with R by somebody else,
  - read your work and evaluate your writing skills (clarity, coherence, grammar, etc.),
  - review and evaluate the difficulty level of your project,
  - evaluate the technical level of use of the programming language R for your empirical goals,
  - assess whether your empirical reasoning makes sense and discuss your remaining weaknesses,
  - acknowledge your learning process.

5 Literature

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

Stephan Huber. My five cents on how to write a thesis: A guide for my students, 2024. URL https://hubchev.github.io/fivecents/fivecents.pdf.

Stephan Huber. How to use R for data science, 2025. URL https://hubchev.github.io/ds/.