Project Submission Guidelines

Data Science for Business (WS 2024)

September 30, 2024

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

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 sections of the chosen paper to replicate. This project is an opportunity for students to demonstrate their mastery of empirical research methodologies using 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 Submit your preferences

Below is a list of research papers. Your task is to replicate the empirical results of one of these studies. Please let me know your choice as soon as possible. Selections are on a first-come, first-served basis. If another student has already chosen a paper, I will notify you so that you can select an alternative.

1. Jack and Oster [2023]: *COVID-19 and Educational Attainment: Learning Loss in the Covid Era*
2. Kearney et al. [2022]: *Labor Market Challenges and Opportunities in the Post-Pandemic Era*
3. Morgan et al. [2023]: *Economic Considerations for Health Policy Post-Pandemic*[1]
4. Okunogbe and Tourek [2024]: *How Can Lower-Income Countries Collect More Taxes? The Role of Technology, Tax Agents, and Politics*
5. Price and Viceisza [2023]: *What Can Historically Black Colleges and Universities Teach about Improving Higher Education Outcomes for Black Students?*
6. Rogoff [2022]: *Emerging Markets and the Global Economy in the Post-COVID World*
7. Sloane et al. [2021]: *College Admissions in America: Challenges for Diversity and Inclusion*

2.2 Conduct the reproduction study

Your task is to reproduce the paper's results using the programming language R. Recognizing the constraints of time, fully reproducing every statistic, table, and graph from the paper might not be feasible, and that's perfectly acceptable.

[1] If you choose this paper, I will provide the dataset for your convenience.

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.

2.3 Give a presentation and publish it on GitHub Pages

Create a presentation using (R) Markdown and **Quarto**, and subsequently **publish it as a website through GitHub**. How to use Markdown and Quarto as well as how to publish a website on GitHub is explained here [see Huber, 2024].

The presentation takes place on 20.12.2024 in 4e OG3.

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, 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.

2.4 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 in
  - html standalone format and
  - PDF format.

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 and ask yourself if it is concise, clear, and coherent with what was said before and after.

I recommend writing the report as a Quarto book. Telford [2023] is a good tutorial on how to write with Markdown and Quarto. Additionally, I recommend reading Huber [2024]. For guidance on creating a standalone HTML file, refer to this resource.

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 document).
- 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 2.5.
- A section where you explain briefly, how you made the pull request, see Section 2.5.
- The Affidavit, see Section 2.6.

2.5 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 [2024]. Remember to reference this pull request in your report.

2.6 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 like this:*

☒ 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 have read the Handbook of Academic Writing by Hildebrandt and Nelke [2019] and have endeavored to comply with the guidelines and standards set forth therein.
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 (+/- 500).
☐ A title page with personal details (name, email, matriculation number).
☐ An abstract.

- ☐ A bibliography, created using BibTeX with APA citation style.
- ☐ The complete R code required to reproduce the results.
- ☐ Detailed instructions on data acquisition and importation into R.
- ☐ An introduction to guide the reader and a conclusion summarizing the work and discussing potential future extensions.
- ☐ 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: make_a_pull_request.
- ☐ A concise description of the presentation published on GitHub.

The project submission includes:

- ☐ The .qmd file(s) of the report.
- ☐ The _quarto.yml file of the report.
- ☐ The .pdf file of the report.
- ☐ The standalone .html file of the report.
- ☐ All necessary files (not available online) to reproduce the report and the R code.
- ☐ The standalone .html file of the presentation.

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

2.7 Submit via ILIAS

- Please consider the deadline for academic papers and written assessments!
- Upload **one .zip file** containing the following:
  1. the paper as (a) .pdf and a (b) .html file.
  2. the .qmd file
  3. the presentation as .html file,
  4. additional files, if needed, so that I can evaluate your work.

3 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.

4 Literature

Jens Hildebrandt and Matthias Nelke, editors. *Handbook of Academic Writings*. VNR Verlag für die Deutsche Wirtschaft, Bonn, Germany, 2019.

Stephan Huber. Data science for business leaders, 2024. URL https://hubchev.github.io/dsbl/.

Rebecca Jack and Emily Oster. Covid-19, school closures, and outcomes. *Journal of Economic Perspectives*, 37(4):51–70, November 2023. doi: 10.1257/jep.37.4.51. URL https://www.aeaweb.org/articles?id=10.1257/jep.37.4.51.

Melissa S. Kearney, Phillip B. Levine, and Luke Pardue. The puzzle of falling us birth rates since the great recession. *Journal of Economic Perspectives*, 36(1):151–76, February 2022. doi: 10.1257/jep.36.1.151. URL https://www.aeaweb.org/articles?id=10.1257/jep.36.1.151.

T. Clifton Morgan, Constantinos Syropoulos, and Yoto V. Yotov. Economic sanctions: Evolution, consequences, and challenges. *Journal of Economic Perspectives*, 37(1):3–30, February 2023. doi: 10.1257/jep.37.1.3. URL https://www.aeaweb.org/articles?id=10.1257/jep.37.1.3.

Oyebola Okunogbe and Gabriel Tourek. How can lower-income countries collect more taxes? the role of technology, tax agents, and politics. *Journal of Economic Perspectives*, 38(1): 81–106, February 2024. doi: 10.1257/jep.38.1.81. URL https://www.aeaweb.org/articles?id=10.1257/jep.38.1.81.

Gregory N. Price and Angelino C. G. Viceisza. What can historically black colleges and universities teach about improving higher education outcomes for black students? *Journal of Economic Perspectives*, 37(3):213–32, September 2023. doi: 10.1257/jep.37.3.213. URL https://www.aeaweb.org/articles?id=10.1257/jep.37.3.213.

Kenneth Rogoff. Emerging market sovereign debt in the aftermath of the pandemic. *Journal of Economic Perspectives*, 36(4):147–66, November 2022. doi: 10.1257/jep.36.4.147. URL https://www.aeaweb.org/articles?id=10.1257/jep.36.4.147.

Carolyn M. Sloane, Erik G. Hurst, and Dan A. Black. College majors, occupations, and the gender wage gap. *Journal of Economic Perspectives*, 35(4):223–48, November 2021. doi: 10.1257/jep.35.4.223. URL https://www.aeaweb.org/articles?id=10.1257/jep.35.4.223.

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