Data management project

Fabrice Rossi

You final data management project should tell a story about the collection of data sets you decided to analyse. The main result of the project is a report, preferably in pdf format, that tells this story. The report must be produced entirely from a quarto document with a simple use of the quarto rendering system. No human intervention should be necessary to produce the report. I should be able to render the document directly from the zip file you will upload on moodle as a result of your work.

The report must not exceed 5000 words (you can use the quarto extension wordcount to count them).

The expected organisation of the report is the following one:

1. research question presentation
2. data sets descriptions (size, production condition, main variables)
3. data analysis including visual representations, tests, model fitting, etc.
4. conclusion
5. a short annex on your data import work (joining, recoding, etc.), including links to the sources (and a link to your github repository)

The code must be entirely hidden from the report. No R warning, message, error, etc. can be included in the report.

The data analysis part is highly project dependent, so I can only provide generic guidelines. Most of the research questions are of a "causal" nature and you want to characterise the links between two variables. This can be investigated graphically first with adapted representations. Then fitting a linear model with `lm()` maybe be a good idea. You can also test dependencies between categorical variables via a χ^2 test (`chisq.test()`). Linear models are interesting as their coefficients can be interpreted in a causal way if all confounding variables have been accounted for. If you want to use the project as a way to learn more about causal inference, I recommend the following resources:

- the book *"The Effect; An Introduction to Research Design and Causality"* is probably the most adapted resource for this project. It describes exactly the standard process to follow when addressing research questions of the type you are facing in the project. It is available on line and contains examples in R.
- the book *"Causal Inference: What If"* is an excellent treatment of the subject but in a much more formal way. It is also available on line. It comes with R code.
- the book *"Causal inference: The Mixtape"* is also very good with an in between formal content. It is also available on line with R code.

The deadline for uploading you project is the 14th of January 2024. You have to upload a zip archive of your project including the quarto document, the data files, the .Rproj file, and the rendered document, preferably in pdf.