Course Assignment D2.1 Modern Tools & Workflows for Data Quality

Over the course, you have learned about different tools and workflows that you can use to make your research (more) transparent and reproducible. For this course assignment, we would like you to create a fully reproducible research project using the tools and workflows you have learned about. More specifically you should:

1. Create a local project with a sensible folder and file naming scheme/structure
2. Add/create the relevant files: Data, Code, documentation, further materials as needed
3. If you cannot or do not want to use data from your own research, you can use
    - publicly available or "toy" data (e.g., exemplary data that is publicly available – e.g., from Kaggle or Tidy Tuesday; as well as examples included in an R or Python package)
    - Reuse existing data that is available for secondary use
    - Use aggregate-level data from sources like Our World in Data or Gapminder
    - Create simulated data (based on an existing dataset), e.g., using the synthpop package for R or Synthetic Data Vault for Python; *note:* this also makes sharing via GitHub (or similar platforms) easier (see below)
4. Use Git for version control
5. Connect the project with GitHub or any other public Git hosting service
    - You can start with a private repo. However, as you will need to share the results with the course instructors you will either need to make it public at some point or add the course instructors as collaborators (at least temporarily).
    - **Important! Please do not share any personal or sensitive data on any public Git repository.** Depending on the type of data you use, you should...
        - Use a suitable data repository and reference it (e.g., OSF, Zenodo...)
        - Properly cite the data source you use (and briefly explain where/how the data can be accessed)
        - Also, even if you can share the data via GitHub, If the data are too large, consider downloading it on the fly in your code/scripts rather than hosting it in the Git repository

   - If you can & want to share your own research data on a public Git repository, only provide processed and anonymized data in plain text formats (e.g., csv); i.e., not the full raw data
   - Only non-personal/anonymized/aggregate data with a CC-0 (or comparable) license should be shared in full publicly
   - Also remember that you can use a .gitignore file for things you do not want to be tracked by Git. This should also prevent you from accidentally sharing files you do not want to share.

6. Add or reference proper documentation to the public Git repo and your files
   - Comments in code files
   - Description in a README file
   - Codebook for data (if applicable; see above)
   - Information about the software, packages, etc. You have used (incl. version numbers). Note: If you opt for Binderizing your repository (see below), you can also use the options supported by Binder (e.g., install.R with or without renv, Pipenv, Docker).
7. Create & provide a shell script named "run.sh" that runs all required analysis scripts consecutively
8. Optional: Binderize the repository
   - Please note that full reproducibility may not be possible in cases where the data cannot be shared publicly (see above)
   - Also, it only makes sense to Binderize repositories/projects that use R and/or Python
   - If you want to Binderize your repo, make sure that you add the needed files as documented in the week 4 course materials and in the guide by The Turing Way.
   - Test whether you can run "run.sh" (see above) on Binder

Additional general information: We recommend that you use a CLI for creating project folders and moving files. In terms of programming languages, we suggest that you use either R or Python. If you work with R or Python, we suggest that you use literate programming tools/formats, such as Quarto or Jupyter. You can also use any other type of software that allows you to create scripts for analysis (e.g., SPSS or Stata) if you are more familiar/comfortable with this, but this will reduce/limit the reproducibility of your project (e.g., also Binderizing such content is not easily possible).

Submission of the assignment

Due date for the submission is **April 11th, 2025** (= 2 weeks after the end of the course)

There two ways in which you can submit your assignment:

1. Create a (Binderized) public repo on GitHub (or Git repo on any other public Git hosting service) and either share the link via the Moodle course forum (in the “Course Assignment” topic) or, if you are not comfortable with sharing it with the whole group, send an e-mail with the repo link to all course instructors.
2. Create a private repo on GitHub, GitLab, or any other Git hosting service where you can add (external) collaborators and add all course instructors as collaborators. If you use a private GitHub repo for this, our usernames are: jobreu (Johannes Breuer), yfiua (Jun Sun), chainsawriot (Chung-hong Chan), arnim (Arnim Bleier), lukasbirki (Lukas Birkenmaier)

Evaluation of the assignment

The assignment will be graded as pass/fail. To pass, you need to have created and shared a Git repo that includes all files/information required to reproduce an analysis you have come up with for your own data, exemplary data, or “toy data” (see above). Ideally, the analysis should be reproducible either interactively via Binder or using a shell script named "run.sh" that runs all required analysis scripts consecutively. If the underlying data cannot be shared directly/easily, the code should be properly structured and commented, and the source of the data should be indicated/cited correctly.