R for Data Science - Syllabus

Andrea Gilardi

Last modified on: 2024-02-27

The objective of this course is to introduce the students to effective and modern tools for data analysis, version control, and development of R packages. All the materials for the lessons can be found at the following link: https://github.com/agila5/R4DS-PhD-Unimib.

Program

We will cover the following topics.

1. The `tidyverse` and some of its most important packages for data manipulation (such as `dplyr`, `tidyr` and `purrr`) (1.5h) (Wickham et al. 2019, 2023b);
2. Debugging techniques provided by R and Rstudio (e.g. `debug()`, `browser()`, `traceback()`), and `try()`/`tryCatch()`. (1.5h) (Wickham 2019);
3. Git and Github: After creating our first git project, we will explore the most important `git` commands (e.g. `clone`, `status`, `push`, `pull`, `merge`, `diff`, …) either via the shell or an R package (e.g. `usethis`). Then we will review common errors that may occur when dealing with `git` and `Github` and, finally, we will review a powerful tool for (big) data management named `git lfs` (4.5h) (Bryan 2023; Chacon et al. 2024);
4. R packages: We will create our first R package and discuss the most important aspects (e.g. Imports vs Depends vs Suggests or documentation). Finally, I will show you how to upload that R package on github and present the most important tools for collaborative package development (issues, comments, and PR) (4.5h) (Wickham et al. 2023a);

(Tentative) Schedule

The lessons will be held in presence at the DEMS seminar room U7-2062 according to the following calendar:

- Friday, March 1st, 14:30-17:30;
- Friday, March 8th, 09.00-12:00;
- Friday, March 15st, 09.00-12:00;

- Friday, March 22st, 09.00-12:00;

Please notice that we are going to have class in a seminar room. Therefore, the students are kindly requested to bring their own laptop to enjoy hands-on coding sessions. Moreover, please try to install R and Rstudio before the beginning of the lessons. Any version of those two software is ok. If you have any doubt, feel free to contact me (`andrea.gilardi@polimi.it`).

Prerequisites

The students are expected to be already familiar with the basics of computer programming (e.g. for-loops, if-clauses, ...) and the R language. If you want to briefly recall the most important topics, I would recommend reading the first few chapters of Micheaux et al. (2013).

Final exam

The students will be divided into two groups, each one consisting of at least two people. The two groups are expected to work together to develop a project that consists of two parts:

1. Exploratory Data Analysis (EDA) with the Tidyverse. You will be required to analyse a given dataset and answer a set of questions;
2. Create an R package that answers a simple statistical problem, providing the necessary testing and documentation.

The detailed assignments for the two parts will be uploaded on the webpage of the course before the end of the term. More details will be provided during the classes.

References

Bryan, Jennifer (2023). *Happy Git and GitHub for the useR*. URL: https://happygitwithr.com/.

Chacon, Scott and Ben Straub (2024). *Pro Git*. URL: https://git-scm.com/book/en/v2.

Micheaux, P. Lafaye de, Rémy Drouilhet, and Benoit Liquet (2013). *The R software*. Springer. URL: https://link.springer.com/book/10.1007/978-1-4614-9020-3.

Wickham, Hadley (2019). *Advanced R*. CRC press. URL: http://adv-r.had.co.nz/.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, et al. (2019). "Welcome to the Tidyverse". In: *Journal of open source software* 4.43, p. 1686.

Wickham, Hadley and Jennifer Bryan (2023a). *R packages (2e): organize, test, document, and share your code*. "O'Reilly Media, Inc.". URL: https://r-pkgs.org/. Forthcoming.

Wickham, Hadley and Garrett Grolemund (2023b). *R for data science (2e): import, tidy, transform, visualize, and model data*. "O'Reilly Media, Inc.". URL: https://r4ds.hadley.nz/. Forthcoming.