Data Management and Programming

ECTS : 3

Description du contenu de l'enseignement :

This course provides an introduction to programming and to data management, with a data- oriented point of view. The course contains two parts. The data management part introduces the data life cycle in data oriented projects from data collection to data exploration. While the main focus of the course is tabular data, it contains also an introduction to entity-relationship models and to relational databases. The programming part of the course introduces the fundamental aspects of imperative programming and the use of the main R data structures. The two aspects of the course are tightly integrated: each aspect of data management is illustrated by adapted programming constructs and uses specific data structures from R. In addition, an introduction to reproducible research is provided, using active documents (in quarto) and git.

Compétence à acquérir :

The first objective of the course is to introduce students to data-driven projects, by presenting the first steps of such projects from data collection to data exploration. Acknowledging the strong limitations of integrated software that rely solely (or mostly) on graphical user interfaces, the second major objective of the course is to provide all the programming knowledge and tools needed to implement all those data management steps, relying on the R language.

After having attended the classes, the students will be able to:

- specify a data management chain adapted to a data-driven project;
- identify the potential data value increase at the different steps of the chain;
- implement those steps in R: data cleaning, data storage, data aggregation and other requests, data exploration;
- more generally implement non-obvious data manipulation schemes in R;
- write active documents using quarto;
- use git and github at a basic level.

Mode de contrôle des connaissances :

The final grade will be made of two types of grading: A continuous assessment grade, made mostly of grades obtained to quizzes (approximately 50 % of the grade) and integrating oral participation during the class and regular attendance; A grade obtained on a full data-oriented project from data collection to data exploration (preferably done in groups of 2 students).

Bibliographie, lectures recommandées :

R for data science: https://r4ds.hadley.nz/