Data Management Course

The objective of this course is double:

- To provide the students with a toolbox for data management. The language used in the course is R, but brief talks on Python and Stata are also included.
- To disseminate the reproducible research idea, that is, that data analyses, and more generally, scientific claims, are published with their data and software code so that others may verify the findings and build upon them.

A fundamental point of the course is that the students should start with data in a text file format, such as CSV, and organize all the manipulations of their data in the code, never doing them manually in Excel sheets or similar. The professor will provide some data sets for the students to develop and test their skills. They can also try their own data.

Although proficiency in R is not assumed, the students are expected to have R and RStudio installed in their computers. Jupyter notebooks will also appear briefly in class.

The course has ten sessions, two sessions per week, (probably) from 9:30 to 12:15. The plan (so far) is:

- Week 1: objects in R, importing data from CSV files, for loops, if-then-else arguments, tabulation and plotting.
- Week 2: import/export from/to Stata, SAS, Excel and SQL databases, merging tables and aggregating data. Packages: foreign, readr, DBI, RSQLite, RMySQL, RPostgreSQL.
- Week 3: data wrangling (duplicates , missing values, variables selection, etc).
- Week 4: text data and regular expressions. Packages: stringr.
- Week 5: importing data from HTML, XML and JSON formats. Packages: RCurl, XML, twitteR.

References

1. BC Boehmke (2016), *Data Wrangling with R*, Springer.
2. R Kabacoff (2011), *R in Action*, Manning.
3. J Kazil & K Jarmul (2016), *Data Wrangling with Python*, O'Reilly.
4. S Munzert, C Ruba, P Meissner & D Nyhuis (2015), *Automated Data Collection with R*, Wiley.
5. H Wickham (2014), *Advanced R*, CRC Press.
6. H Wickham & G Grolemund (2016), *R for Data Science*, O'Reilly.