Data Loading and Wrangling in R

Goal This first practical is a gentle introduction to the course and is designed to teach you R skills and specific packages to load and wrangle datasets. We will sometimes provide you references but you are expected to find your own resources to solve the problems.

Throughout this course, we will perform data wrangling and analysis using R. You can learn about R and install it from the official website. Although not necessary, we highly recommend to use RStudio's IDE, which gives you an ideal working environment for statistical analyses. In this practical, you will use the following packages, which you can install with the `install.packages` function:

RMySQL, **RODBC**, **pool**: to connect to external relational databases.The reference manual available on CRAN.

dplyr: for data wrangling. A cheat sheet can be found on RStudio's website.

Follow along this practical, learn by yourself about any function used and solve the exercises.

1 Loading Data into R

Colloquially, 'loading' a dataset means storing it onto your computer's Random-Access-Memory (RAM), which allows for a fast access of the CPU to the data. The RAM is extremely fast[1], but is typically of limited amount compared to what can be stored in a hard-drive or a SSD (usually 4 to 16 Gbs in a typical consumer-grade computer). The data stored in your computer's RAM is called *volatile*: the information it contains will disappear once the computer is shut down. This is the reason why the datasets must be loaded into R at the beginning of each session.

In order to get acquainted with the methods necessary for this course, please have a look at **section 1.4 Importing Data into R** on the e-book of the course. After having read and understood the concepts, solve the exercises of **section 1.4.5**.

[1] The amount of RAM in your computer does not determine the speed of computation: it merely sets an upper limit of storage available for these computations.

2 Dataframes and data wrangling

Many packages, and indeed R-base itself, are optimized to have the data organized in the following way:

- each **row** represent one observation
- each **column** represents a variable (or 'feature')

Data wrangling means to manipulate and prepare a dataset in such a way, that it becomes amenable to analysis. Minimally, a numerical dataset is stored as a Matrix object, a type optimized for computations. Preferably, however, a dataset is stored as a dataframe object. A dataframe is technically a list of columns, each containing data of a given type (e.g., integer, numerical, character, factor).
There are two packages dedicated to data wrangling in R:

- **dplyr** is a grammar of data wrangling, which focuses on efficient and elegant coding
- **data.table** is computationally extremely fast to manipulate very large datasets

Both create objects that are extensions of dataframes: dplyr uses a tibble, and data.table uses a data.table. We will not use data.table in this course, as it has a steeper learning curve.

In order to get acquainted with these instruments, please have a look at **section 1.5 Data Wrangling** on the e-book of the course. After having read and understood the concepts, solve the exercises of **section 1.5.1**.