Handle your data!

Golden rules for data handling

The first three golden steps

1. Look at it
2. Look at it
3. LOOK at it

You should look at your data using any kind of software, try differnt software, look at it with R, with excel, with a text editor. Import it in R look at the types of the variable, look at the variable names, the values, is there some macroscopic property you did not expect, something that looks wrong ?

Use describe(). Use unique().

Look at each column separatly, count the number of NA ntrue(is.na(data$column)) (ntrue is a function of the pipeline package from the lab).

Keep the raw file untouched

You should not modify the raw file, or overwrite it. You load it, and every step of modification should be in a script that you can run anytime without risking overwriting or corrupting your raw data in any way. If it is really not possible, make sure to keep a copy of your initial raw data, make necessary manual modification, and consider this new file as raw data, but this should be kept to a minimum.

Rename your variables

For the lab prefer snake_case_names, and do not use abreviations like rt, use something that makes complete sense without any ambiguity like response_time. Do not fear to be overly clear, for example the mean of task 2 block 1 should not be name mt2b1, or mean_t2_b1, or avg_task2B1, but mean_correct_ratio_task2_block1. You can use the formatter of the pipeline to make sure all your variables are correctly named and documented.