Task 1. Create a directory on your account/laptop, that you will use for this course. Create new *.R script for this seminar, save it in this directory and set this directory as your working directory.
Hints: setwd()

Task 2. Download files skulls.txt, newborns.csv and height.txt from Study materials in IS, save them in your created directory.

Task 3. Load file skulls.txt. This data file contains the height of skulls (in mm) of people. We also know their sex and their assigned ID number.

1. Check the structure of data.
2. Check the dimensions of this data frame. How many variables are there? How many observations?
3. Look at the first couple of observations and few last observations.
4. Are there any missing values?
5. If there are missing values, remove corresponding observations.
6. Check that there aren't any duplicities (= no person was entered into the database twice).
7. Find out the number of men and women in the database.
8. Select just the women.
9. Select men with skull height bigger than 125 mm, how many are there?
10. Select everyone with skull height bigger or equal to 130 mm, but smaller than 140 mm, how many people are there? How many women? How many men?

Hints: read.table(), str(), dim(), head(), tail(), is.na(), na.omit(), unique(), duplicated(), table()

Task 4. Load file newborns.csv. This data file contains the birthweight of children (in g, column weight.C), their sex (column sex.C), their assigned ID (column id) and the highest level of education their mother achieved (column edu.M). The level of education is coded in the following way: 1 is elementary school, 2 is high school, 3 is university degree.

1. Check the structure of data.
2. Check the dimensions of this data frame. How many variables are there? How many observations?
3. Find the frequencies of different levels of mother's education.
4. Variable edu.M was loaded as a numerical variable, recode it as a factor with levels labeled as specified above.
5. Are there any missing values?
6. If there are missing values, remove corresponding observations.
7. Check that there aren't any duplicities (= no person was entered into the database twice).
8. If there are any duplicities, remove them (keep the first observation for each child)
9. Find out the number of boys and girls born to mothers of different education levels.

Hints: read.csv(), str(), dim(), table(), is.na(), na.omit(), duplicated(), unique()

Task 5. Load file height.txt, which contains height of people (in *cm*) and their gender.

1. Check the structure of data.
2. How many different levels of variable gender can you see?
3. We would like to have men codes as 'M' and women as 'F'. What percentage of observations is classified incorrectly or not at all?
4. Extract the vector of genders (so you don't change the original data) and try changing incorrect values to 'M' or 'F' and missing values to NA. How many levels of gender can you see now? What is going wrong?
5. Again extract the vector of genders and try relabelling levels through argument labels in function factor(). What levels do you see now?
6. Can you thing of a solution that will lead to the right classification?

Hints: levels(), factor()