R review – dataframe and graphics

You’ve got the dataset etubiol.csv

Notes on variable names: height_M: height of mother, height_F: height of father, n_siblings_F: number of female siblings, n_siblings_M: number of male siblings. The rest should be self-explanatory. Otherwise, just ask me ☺

Tips for good practice: comment on your scripts (with the #) what you do and why. In coding, you should be able to pass on your script to people as an example, and people should be able to understand your script. It would be good to have this habit in the early stage.

For example:

```
#Load dataframe

>read.csv(“etubiol.csv”, header = T)

#change the hair colour to factor as we will need it to….later

>etubiol$haircolour = as.factor(etubiol$haircolour)
```

Data examination

- Load and save the file as “df”
- Examine the file
    - How many observations and variable does df have?
    - What are names and types of the variables?
    - Summary statistics of “df”
- Calculate BMI of each person and add the extra variable “bmi” to a new dataframe called “df_bmi” (Google the BMI formula). (Sanity check: BMI values should be between 15 and 35). Save df_bmi to a csv file, with header.
- Make a scatter plot of some pairs of variables of your choice.

Data visualisation

- Use numerical summaries and visual tools, and statistical analyses, if possible, to test:
  - Is there a difference in bmi between males and females (hint: t-test)
  - Is there a difference in bmi between smokers and non-smokers?
  - How strong is linear (Pearson) correlation between shoe size and height? Is it a significant difference? (hint: use cor() function)
  - Find out a linear relationship: how much does shoe size increase per added cm of height? Try also to do it for male and female separately. (hint: lm() function)
  - Make plots to visualize the answers for these questions (either basic R or ggplot). For now, you can make as many plots as you want, so that you can present the results you find with the plot. We will discuss this ☺
  - Choose a variable or multiple variables and come up with a question of your own and investigate it as above.