Exercise 1

Riccardo De Bin and Vindi Jurinovic

Today we will start to analyze the NHANES dataset from your R Data Project. Choose one sub-sample to work with and think of a name for your data set (say, `nhanes`, or `dataproject`, or just `tab` for table, or any other name you like. . .). Load the data into your R workspace with the command:

```
> name_of_your_data <- read.table('name of the file', sep='/t', header=TRUE)
```

Here, `sep` is the field separator character. Values on each line of the file are separated by this character (you can check this by opening the file in a text editor, or you can just try out some separators until you get the right one). `header=TRUE` means that R should interpret the first row of the data set as the variable names.

Try to answer the questions by yourself. If you don't know the name of some function, say sequence, try `??sequence`. To find out more about a function and its arguments, use `?function_name`.

1 Getting familiar with the data set

- What is the dimension of the data set? How many rows (samples), and how many columns (variables) does the data set contain? What are the variable names of the data set?

- All the variables in the data set are either of a class `integer`, `numeric` (i.e., they are all interpreted as numbers by R) or `boolean` (i.e., logical). However, some of the variables should be factors rather than numerical variables. Change the class of these variables with the function `as.factor`. Save the new data set as an `.Rdata` file. Attach the data frame so you can assess the variable names directly.

- How many women and how many men are there in your data set?

- What is the mean BMI in the overall population? What is the mean BMI for men and women?

- Who has a higher mercury level in blood: men or women? People with chronic bronchitis or people without it? 'Hispanic', 'White', 'Black' or 'Other/Mixed' people?

- Use the function `summary` to get summarized information on all the variables in the data set.

2 Plots

- Plot the variable `rr_sys` as a function of `bmi`. Try out different types o point characters (function argument `pch`) and colors (argument `col`) and choose the prettiest ones ☺. Label the x- and y-axis in the plot with the corresponding variable names (function arguments `xlab` and `ylab`). Think of a suitable title and add it to your plot. Type `?plot` for more information on the function and its arguments and try out some of these.

- Now we want to plot the variable `rr_sys` against `diab_lft`. Which plot should we use here? Create the plot and choose different colors for the boxes. Give each box a suitable name and add a label for the y-axis as well as the main title. Do you see a difference in systolic blodd pressure between diabetics and non-diabetics?

- Plot the variable `bmi` against `educ`. Interpret the picture you see.

- Plot the histogram of the high-density lipoprotein (HDL) cholesterol levels. How does the distribution look like? Can you convert the variable `hdl` so that its distribution looks more normal? Create such variable and add it to your data set. Save the new data set in a new file.