Exercises 1

1.1 Data structures

Consider the following three vectors:

- `A`: 5, 8, 7, 6, 8, 4
- `B`: 1.3, 2.1, 1.8, 1.2, 1.4, 2.3
- `C`: y, y, n, y, n, n

Combine the vectors into a data frame with 3 columns and 6 rows.

- Print the element (3, 2).
- Print the 4th row.
- Create a subset that consists of the last two columns and the rows 2 - 5.
- Transpose the data frame so that it has 6 columns and 3 rows.

1.2 Thyroid Disease

- Download the file `allbp.data` from the Thyroid Disease data set available at https://archive.ics.uci.edu/dataset/102/thyroid+disease.
- Load and preprocess the data so that it is ready for analysis. (Check categorical variables, missing values, variable names and so on.) Use the file `allbp.names` to your advantage.
- How many observations and how many variables are there in the data?
- Which variables have missing values? How many?

1.3 Thyroid Disease (continued)

Continue to analyse the data you prepared in the earlier exercise.

- For each variable that has only yes/no values, calculate the number of yes values divided by the number of observations.
- For each of the `TSH`, `T3`, `TT4`, `T4U`, `FTI` and `TBG` variables, calculate the sum of the squared values divided by the number of non-NA values.
- Calculate the mean ratio (i.e. the mean of ratios) between `T3` and `TT4`.

1.4 Purchases

- Load the data available in the file `purchases.csv`.
- Find invalid values in the data and replace them either with a correct value (if possible) or with NaN.
- Replace all missing values of the `purchases` variable with zero.
- Use median imputation to fill in all missing values of the `retention_time` variable.
- (BONUS) Group the observations by `sex` and `location` before calculating the substitute median(s).

1.5 Electric bikes

This exercise begins a series of exercises that spans throughout the course. The same data is analysed from various perspectives in order to illustrate how a statistical analysis project can proceed. The exercises concern a hypothetical scenario, but the data is derived from real-world data.

The file `bikes.data` contains data that was collected from the recording devices of commercial electric bikes. You can assume that no preprocessing or filtering has been done. The data may therefore contain irrelevant records, such as customers only trying out how bikes can be rented with an Android app and cancelling the transaction without actually riding the bike. You can also assume that there have been quite a few technical problems with the bikes, which may have resulted in invalid values in the data.

- Load the data from the file `bikes.data`.
- Can you find any irrelevant records or invalid values? If you do, explain why the records are irrelevant or the values invalid.
- Process the irrelevant records and invalid values you found. Explain why the modifications you made are the correct way to fix the problems in the data.

(NOTE: At this stage of the course, it might be challenging to effectively explore the data, so it okay if you do not find anything interesting. We will return to this task in a later exercise when studying descriptive statistics.)