Exercise 4

TKO_7093 Statistical Data Analysis

1.

Download the file `allbp.data` from the `Thyroid Disease Data Set` available at https://archive.ics.uci.edu/ml/datasets/Thyroid+Disease (see the files in the `Data Folder`).

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.

1. How many observations and how many variables are there in the data?
2. Which variables have missing values? How many?

2.

Using the data you prepared above,

1. calculate the mean and standard deviation for each quantitative variable.
2. calculate the frequency and relative frequency of the `yes` value for each indicator variable (i.e. a variable with only `yes` / `no` values).
3. (BONUS) calculate the frequency and relative frequency of each observed value for each other categorical variable.

3.

Download and unzip the `DataSet SOFT file` from https://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS5037 . The SOFT file format contains tabular gene expression data accompanied by its metadata. The rows that start with either `^`, `!`, or `#` are metadata, which can be ignored in this exercise.

Some useful pieces of information about the file `GDS5037.soft`:

- The first 160 rows and the last row are metadata.
- The `ID_REF` column contains probe ids. Use them as row names.
- The `IDENTIFIER` column contains gene names.
- The other columns contain gene expression levels for subjects.

Load and filter the data such that your data frame only contains the probe and subject ids listed in the files `43-probes.ids` and `43-subjects.ids`, respectively.

1. For each subject, calculate the mean expression level over the genes. Do you notice any obvious differences between the subjects?
2. For each probe, calculate the mean expression level over the subjects. Do you notice any obvious differences between the probes?
3. How many probes are there with a mean greater than 10.0? Which genes do they correspond to?

4.

The files `44-helsinki.csv` and `44-espoo.csv` contain daily numbers of cyclists spotted on selected streets in Helsinki and Espoo. Load the files and merge the data into a single data frame.

1. For how many days were observations made in total?
2. How many observation days were there for each street?
3. On how many days were all streets observed simultaneously?
4. Which street was the busiest in terms of the total number of cyclists?
5. Filter out the dates which have one or more missing values. Does this affect your conclusion about the busiest street? Why or why not?