Exercise 5

TKO_7093 Statistical Data Analysis

1.

Load the data available in the file `51-data.csv`.

1. Find invalid values in the data and replace them either with a correct value (if possible) or with NaN.
2. Replace all missing values of the `purchases` variable with zero.
3. 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).

2.

Load the file `GDS5037.soft` (see the exercise 4.3) and filter the data such that it only contains the probe ids listed in the file `52-probes.ids` and the subject ids listed in the files `52-control.ids` and `52-asthma.ids`.

1. Create a box plot of gene expression levels for each subject. Do you notice visible differences between the subjects or between the groups?
2. Use a T-test to find differentially expressed genes (i.e. probes for which the means are different between the control and asthma groups). Adjust the p-values with the Benjamini-Hochberg method.
3. Create histograms of the unadjusted and adjusted p-values. Why do these two histograms differ?
4. How many differentially expressed genes (i.e. statistically significant differences) are there at the false discovery rate of 0.05?
5. Sort the adjusted p-values in ascending order. Which genes do the first ten probes correspond to?

3.

Load the data available in the file `53-data.csv`, which contains daily numbers of cyclists spotted on selected streets.

1. For each weekday (Monday - Sunday), calculate the mean daily number of cyclists for each street. Plot the means as a bar plot.
2. For each month (January - December), calculate the mean daily number of cyclists for each street. Plot the means as a bar plot.
3. Calculate the Spearman's correlation coefficient and create a scatter plot for each pair of streets.
4. What information do these plots reveal?

4.

An advertisement company followed online customers to discover how effectively advertisements lure them to spend money. The data files `54-image.csv` and `54-video.csv` contain information on how much customers spent in total after clicking on advertisements.

Is there statistical evidence to claim that the total amount spent by customers is different if they click on image advertisements than on video advertisements?