Statistics for Data Science Exercise 6
Wintersemester 2023/24
Please complete these problems before the exercise session on Tuesday 28 November, 2023, 8:30. Please be prepared to present your solutions to any problems that you completed successfully.

The goal of these programming exercises is to become familiar with accessing data given in *comma separated values* (CSV) format, converting the data into a *data frame*, working with data frames, and visualizing data. In Python, you may need to install (some of) the following libraries using `pip install` or `pip3 install`:

```
numpy
scipy
matplotlib
pandas
openpyxl
statsmodels
seaborn
```

Naturally, please feel free to use other Python libraries or any other programming language of your preference (such as R) to complete the tasks.

1. (a) Create a data frame that has 20 observations of height. Simulate the heights from the uniform distribution between 140 and 200; you can use `numpy.random.uniform(low=140,high=200,size=20)` and store the results as vector `h`. Then, simulate a gender for each of the observations such that the probability of being female is 60%; you can use `numpy.random.choice(['female','male'],20,replace=True,p=[0.6,0.4])` and store the results as vector `g`. Finally, the data frame can be constructed using `df = pandas.DataFrame({'height': h, 'gender': g})`.

   Use the `print` command to display the contents of the data frame you constructed. What is the output of `df['gender']` and `df.height`? What about `df.loc[integer]`, `df.columns`, and `df.index`?

   (b) Replace the height of the first entry with the numerical value 120. Visualize the heights for each gender as a box plot using the command `df.boxplot(by='gender')` or `seaborn.boxplot(data=df,x='gender', y='height')` (note that you may also need to add `import matplotlib.pyplot as plt` in the preamble and execute `plt.show()` after creating the box plot object). How do you interpret the box plot?

   (c) Draw a quantile-quantile plot (Q-Q plot) of the female heights and male heights, respectively. You can use, e.g., the `statsmodels.api.qqplot` function with the option `line='q'`. How do you interpret the Q-Q plots?

The exercises continue on the next page!

2. Download the data sets `mtcars.txt` and `mtcars.xlsx` from the course website.

   (a) Import both files as data frames using the commands `pandas.read_csv` and `pandas.read_excel`. The tab key has been used as a separator in the TXT file, so you will need to use the option `sep='\t'` to import the data correctly. Check that both data frames have exactly the same data using, e.g., the function `pandas.DataFrame.equals`.

   (b) Save one of the data frames as variable `cars`. Make a scatter plot with the variables `wt` and `mpg`. You can use, e.g., `cars.plot.scatter` or `seaborn.scatterplot`. Rescale the axis so that x takes values from 0 to 6 and y takes values from 10 to 40.

   (c) Create a scatter plot for each pair of two different variables using, e.g., the function `pandas.plotting.scatter_matrix` or `seaborn.pairplot`.

   (d) What does `cars['cyl'].value_counts()` do?
   Create a pie chart for the number of cylinders in the different cars using, e.g., `cars['cyl'].value_counts().plot(kind='pie')`.

   (e) Create a sample correlation matrix of the data set. You can use, e.g., the command `cars.corr(numeric_only=True)`. Visualize the sample correlation matrix as a heat map using, e.g., the function `seaborn.heatmap`.

3. Download the data set `HW.txt` from the course website and import it as a data frame.

   (a) Draw a scatter plot of the variables `ageYear` and `heightIn` such that the males and females have different color and shape in the plot.

   (b) You can add a regression line to scattered data using, e.g., the `seaborn.lmplot` function. Note that this function also draws the 95% confidence region by default.
   Draw a regression line (and optionally also the confidence region) to the `(ageYear,heightIn)` plot in the following cases:
   (i) consider only the heights corresponding to males;
   (ii) consider only the heights corresponding to females;
   (iii) consider the heights regardless of gender.

4. (a) Download the data `FT.txt` from the course website and make a histogram. First plot the histogram such that you have binwidth=5. After that, plot the histogram such that you have 15 bins.
   Hint: The `seaborn.histplot` function supports sophisticated binning operations.

   (b) There is something wrong with the data set `BP2.txt` on the course website. What are the indices of the problematic elements? How would you repair the data set?