Data Analytics – Exercises

(Week 07)

In these exercises, you will learn:

- to statistically analyze contingency tables using the Chi-squared test
- to perform a Pearson correlation including a significance test

In the data analytics process model, these exercises cover part of the steps "Statistical data analysis and/or Modeling" and "Evaluation & Interpretation" (see figure 1). Results of the exercises must be uploaded as separate files (**no .zip files!**) by each student on Moodle. Details on how to submit the results can be found in the tasks below.

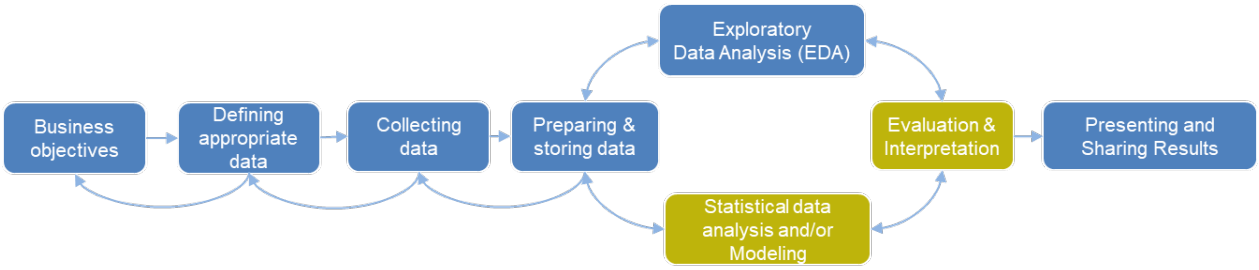

Figure 1: Data analytics process model (see slides of week 01)

Task 1

In this exercise, you will learn to statistically analyze a contingency table using the Chi-squared test. The tasks are:

a) Run the Jupyter notebook 'analysis_of_contingency_tables.ipynb' and try to find out, what the Python code does.
b) Look at the Chi-squared test for the contingency table showing the number of apartments for different living area and price categories. The Chi-Square Test is commonly used to test statistical association between two categorical variables.
c) Do you fully understand, the output and interpretation of the results of the Chi-squared test? If not, look at the slides provided in the large class of this week.

To be submitted on Moodle: nothing 😊!

Task 2

In this exercise, you will learn to calculate a Chi-squared test "by hand".

a) Follow this link to a research study in which a Chi-squared has been used to study animal behavior. In detail, the research is based on the social interactions of two spotted hyenas (*Crocuta crocuta*) which were observed over 40 hours in a U.S. zoo. The 4x2 contingency table below has been taken from this research study.

	Female	**Male**	***Total***
Fear	0	4	*4*
Greeting	7	6	*13*
Aggression	9	0	*9*
No Response	15	10	*25*
Total	*31*	*20*	*51*

b) In the Jupyter notebook of task 1, you will find a section 'Applying the Chi-squared test to animal behavior'. Go to this section and perform a Chi-squared test based on the table above (note that the table is already included in the Jupyter notebook).
c) Write a Python function which calculates the expected frequencies of the contingency table. Input must be a numpy array or data frame with observed frequencies. To write the function, ask ChatGPT for help. Apply the function to your data.
d) Compare your calculated expected frequencies with those from the chi2_contingency() method in Python.
e) Extent the Python function under c) to additionally provide the Chi-square test statistic. Again, ask ChatGPT for help.
f) Compare your calculated Chi-square test statistic with the Chi-square test statistic from the chi2_contingency() method in Python.
g) In the Jupyter notebook, state in one sentence whether the results of the original study are correct or not.

To be submitted on Moodle:

- The Jupyter notebook as html-file 'analysis_of_contingency_tables.html' extended according to b), c), d), e), f) and g).

Task 3

In this exercise, you will learn to perform correlation analyses. The tasks are:

a) Run the Jupyter notebook 'correlation_analysis.ipynb' and try to find out, what the Python code does.
b) In the Jupyter notebook section 'Correlation analysis based on car data', part of the car data set 'autoscout24_data.csv' has already been imported to a data frame.
c) Explore the numerical car data (i.e.: Price, Kilometer, PS) exploratively using a paired scatterplot (see slides and exercises of previous weeks).

d) Perform correlation analyses inclusive significance tests for the variables:
    - Price versus Kilometer
    - Price versus PS
    - Kilometer versus PS

To be submitted on Moodle:

- The Jupyter notebook as html-file 'correlation_analysis.html' extended according to c) and d).