Data Analytics – Exercises

(Week 04)

In these exercises, you will apply Exploratory Data Analysis (EDA) methods to the data of the previous weeks. The objectives of EDA are to:

- Enable unexpected discoveries in the data
- Suggest hypotheses about the causes of observed phenomena
- Assess assumptions on which statistical inference will be based
- Support the selection of appropriate statistical tools and techniques
- Provide a basis for further data collection through surveys or experiments

In the data analytics process model, these exercises cover part of the step "Exploratory Data Analysis (EDA)" (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.

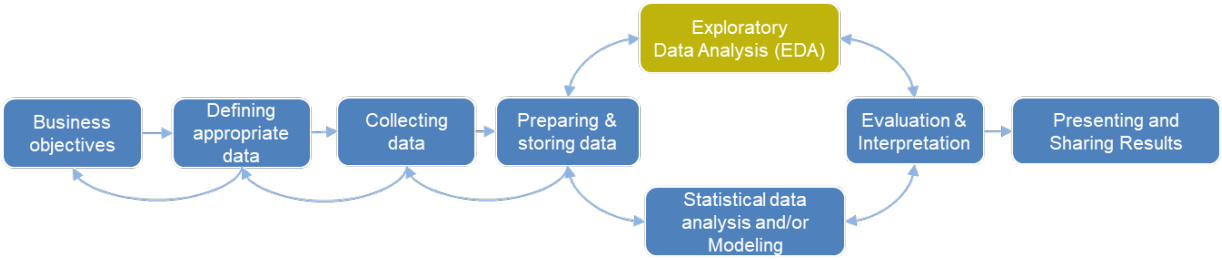


Figure 1: Data analytics process model (see slides of week 01)

Task 1

In these exercises, you will learn how to create and format graphics with matplotlib. The tasks are:

a) Run the Jupter notebook 'graphics_with_matplotlib.ipynb' step by step and try to understand what the code does.
b) Look how multiplots are created (section 'Creating multiplots with .subplots()').
c) Try to change marker symbols and colors in the plots.
d) Try to change the legend text.

To be submitted on Moodle: nothing 😊!

Task 2

In these exercises, you will learn to explore the apartment data and identify and remove outliers. The tasks are:

a) Run the Jupyter notebook 'exploratory_data_analysis_apartment_data.ipynb' step by step and try to understand what the code does.
b) Look at the statistics and graphics and identify outliers in the data.
   ➔ Note that there is no clear definition of what an "outlier" is. In the case of apartment data in the canton of Zürich, this could be, for example an apartment with an extremely large living area or very small / high rental price per m2.
c) Create a subset of the data without outliers (e.g. remove all apartments with very large living areas and/or very small prices per m2).
   ➔ To create a subset, use the .loc method as shown in the previous exercises.

   Run the Jupyter notebook without the outliers and look how the statistics and graphics change.

To be submitted on Moodle: Screenshots of tables with **a)** the summary statistics of data including outliers and **b)** the summary statistics of data without outliers.

Task 3

a) In the Jupyter notebook of Task 2, go to the section 'Quantiles original values' and look for the 10% and 90% quantile of pop_dens. Remember: pop_dens contains the population density of municipalities.
b) Make a copy of the Jupyter notebook and rename it 'rural_apartments.ipynb' and go to the section 'Filter apartments'. Filter the apartments in the more rural parts of the canton of Zuerich. Use the 10% quantile of the variable pop_dens as the threshold for the filter, i.e. only municipalities with a density equal or lower this value shall be considered in the analysis. Run the Jupyter notebook and save the result as html-file (rural_apartments.html).
c) Make another copy of the Jupyter notebook and rename it 'city_apartments.ipynb'. Filter the apartments in the denser parts of the canton of Zuerich. Use the 90% quantile of the variable pop_dens as the threshold for the filter, i.e. only municipalities with a density equal or higher this value shall be considered in the analysis. Run the Jupyter notebook and save the result as html-file (city_apartments.html).
d) Compare the results of the two previous notebooks. What are the differences?

To be submitted on Moodle:

- Use Power Point to prepare a slide set including a comparison of rural apartments with city apartments. For this, divide each Power Point slide into two parts (left part = rural; right part = cities).

- Use the following statistics/graphics for comparisons (screenshots from Jupyter notebooks):
    - Tables with summary statistics of numeric variables
      (Hint: you can use `df.describe()` for this where '`df`' is your data frame)
    - Boxplots of prices per m2
    - Boxplots of areas
    - Histograms of prices per m2
    - Histograms of areas

Task 4

In these exercises, you will apply EDA methods to the supermarket data. The tasks are:

a) Create a Jupyter notebook exploratory_data_analysis_supermarket_data.ipynb.
b) Use pandas to read the file 'supermarkets_data_enriched.csv'.
c) Use EDA methods to:
    - Count the number of supermarkets per brand (e.g. Migros, Coop, etc.). Note that the `.value_counts()` method from the pandas library could be used for this purpose, i.e. use: `df['brand'].value_counts()`.
    - Create a barchart with the number of supermarkets per brand.
    - Use the .PairGrid() method from the seaborn library to create a scatterplotmatrix of the numeric variables `lat, lon, pop, pop_dens, frg_pct, emp`.
    - Create a plot with the locations of supermarkets in different colors according to their brand. The plot must contain at least locations of the brands: Denner, Volg & Landi. Use the following example code as the basis:

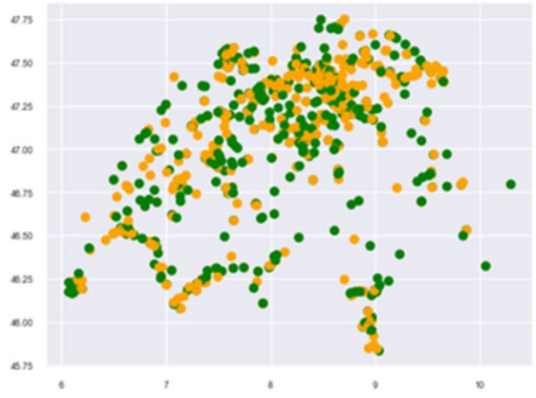
```python
import pandas as pd
import matplotlib.pyplot as plt

# Read and select variables
df = pd.read_csv("supermarkets_data_enriched.csv")[['id',
                                                    'brand',
                                                    'lat',
                                                    'lon']]

# Subset
df_sub = df.loc[df['brand'].isin(['Coop', 'Migros'])]
df_sub

# Colors
colors = {'Coop': 'green', 'Migros': 'orange'}

# Plot
plt.scatter(df_sub['lon'],
            df_sub['lat'],
            c=df_sub['brand'].map(colors))
```



To be submitted on Moodle:

- Jupyter notebook as html-file: exploratory_data_analysis_supermarket_data.html