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.

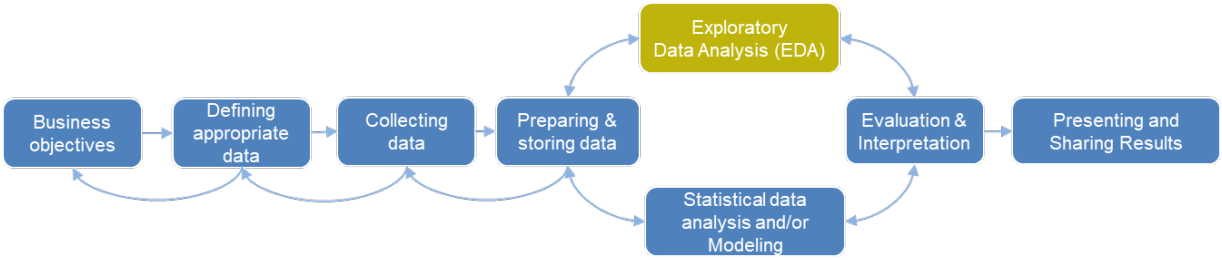


Figure 1: Data analytics process model (see slides of week 01)

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. 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) Go to the section 'Quantiles' and look for the 10% and 90% quantile of pop_dens.
c) Remember: pop_dens is the population density of municipalities.
d) Make a copy of the Jupyter notebook, and go to the section 'Filter apartments'. Filter 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).
e) Make another copy of the Jupyter notebook and filter 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).
f) Compare the results of the two previous notebooks. What are the differences?

To be submitted on Moodle:

- Jupyter notebook as html-file: rural_apartments.html
- Jupyter notebook as html-file: city_apartments.html
- PDF-file 'differences_rural_cities.pdf' with a comparison of rural versus cities.
- Use Power point to prepare the PDF – file.
- Divide each Power point slide into two parts (left part = rural; right part = cities).
- Use the following statistics for comparisons (screenshots from Jupyter notebook):
    - Lists of municipalities in the subset
    - Summary statistics of numeric variables
    - Boxplots of prices per m2
    - Boxplots of areas
    - Histograms of prices per m2
    - Histograms of areas

Task 3

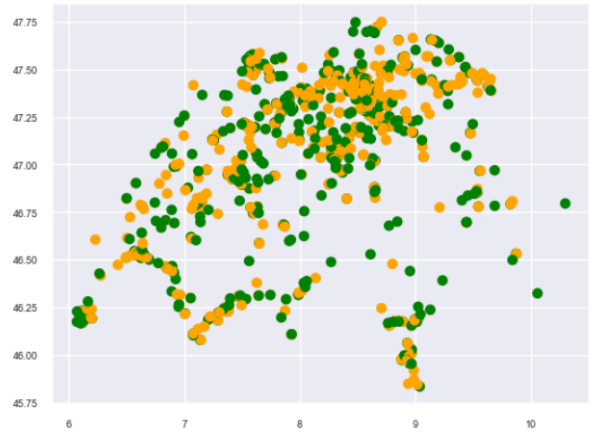
In these exercises, you will apply EDA methods to the supermarket data. The tasks are:

a) Create a Jupyter notebook exploratory_data_analysis_supermarket_data.ipynb.
b) Use the pandas library to import the file 'supermarkets_data_enriched.csv' from the materials for exercises folder on Moodle.
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 summary statistics of the numeric variables `lat, lon, pop, pop_dens, frg_pct, emp.`
- Create a histogram of the variable `pop_dens`. Is there skewness in the distribution? If yes, is it right-skewed or left-skewed?
- 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.` Do you see any relationships between the variables?
- Create a plot with the locations of supermarkets in different colors according to their brand. The plot must contain at least locations of the following brands: Coop, Migros, Denner, Volg, Landi. Use the following example code as the basis:

```
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))
```

```
<matplotlib.collections.PathCollection at 0x1ca80029d30>
```



To be submitted on Moodle:

- Jupyter notebook as html-file: exploratory_data_analysis_supermarket_data.html