Data Analytics – Exercises

(Week 10)

In these exercises, you will learn:

- to perform cluster analyses using k-means clustering,
- to determine the optimal number of clusters (*k*) using the elbow method,
- to evaluate k-means clustering models,
- to visualize the results of k-means clustering.

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.

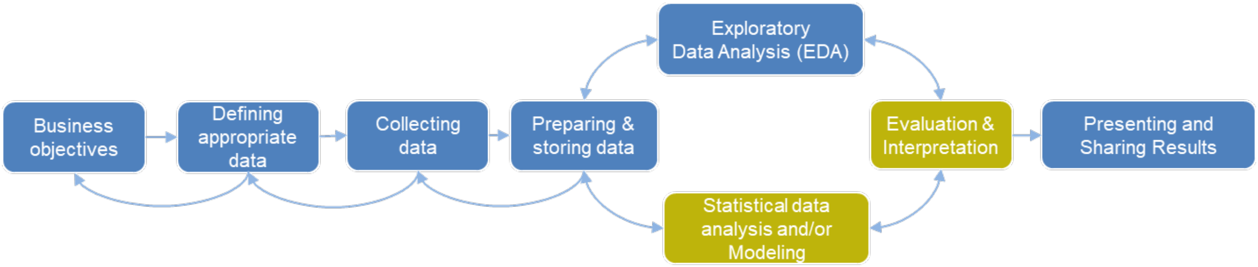
Business objectives
Defining appropriate data
Collecting data
Preparing & storing data
Exploratory Data Analysis (EDA)
Statistical data analysis and/or Modeling
Evaluation & Interpretation
Presenting and Sharing Results

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

Task 1

In this exercise, you will learn to determine the optimal number of clusters for a k-means clustering model by using the elbow method. In addition, you will learn to perform image segmentation and 3D point cloud segmentation based on k-means clustering.

a) Run the Jupyter notebook ‘k-means_clustering.ipynb’ step by step and try to find out, what the Python code does.

b) In section ‘1.) Simple k-means clustering example’, go to the subsection with the elbow method. In the Jupyter notebook, based on the elbow chart, state which is the optimal *k* value for this clustering model.

c) In the subsection ‘Perform Silhouette Analysis’, perform a Silhouette Analysis for different values of *k*. Use k = 2, 3, 4, 5, 6, 7, 8, 9, 10, 11. You can implement this in the `for loop` by replacing the code line:

```
for i in [2,3]:
```

by

```
for i in range(2,12,1):
```

d) Go to the section ‘2.) Image segmentation‘. In this example, an image with a parrot is used for image segmentation. Find another .jpg image (e.g., by using Google’s image search) and run the Jupyter notebook to perform image segmentation for your own image.
e) Use the elbow method to determine the optimal number of clusters *k* for the segmentation of your image.
f) Create 4 different image segmentations for your image based on different numbers of clusters *k* and plot the images.
g) Go to the section ‘3.) 3D point cloud segmentation’ and look at the segmentation example based on the airport data.
h) Open the file ‘3d_point_cloud_segmentation.txt’ and copy the content to a new cell in your Jupyter notebook. Run the cell, then change the value for *k* and look if you can improve the quality of the clusters (use the elbow-plot and the Silhouette Score to find the optimal number of clusters).

To be submitted on Moodle:

- The Jupyter notebook as html-file ‘k-means_clustering_task01.html’ with the changes and short explanations according to c), d), e), f), g) and h).

Task 2

In this exercise, you will learn to perform k-means clustering based on the apartments data.

a) Go to the section ‘4.) Finding clusters in the apartments data …’. In the prepared example, the variables `rooms`, `area` and `price_per_m2` are used to form the clusters.
b) Go to the subsection ‘Subset of the apartment data …’ and extent the data frame which is used for k-means clustering (name = ‘X3’) by including additional numerical variables (e.g. lat, lon, pop_dens, tax_income, …).
c) Use the elbow method to find the optimal number of clusters *k* for the extended data frame (X3).
d) Perform k-means clustering based on the optimal number of clusters. Use the (already included) name ‘kmeans_apmts’ as the name for the k-means model.
e) Use the following Python code to derive the attribute values from ‘kmeans_apmts’:

```
print(kmeans_apmts.labels_, '\n')
print(kmeans_apmts.inertia_, '\n')
print(kmeans_apmts.cluster_centers_, '\n')
print(kmeans_apmts.feature_names_in_)
```

f) In the Jupyter notebook, explain the meaning of the output of the code in e). You can find the required information in the sklearn.cluster documentation on this Link under "Attributes".

g) Calculate the Silhouette Score for ‘kmeans_apmts’. In the Jupyter notebook, state whether the optimal value for *k* suggested by the elbow method is consistent with the value for *k* showing the highest Silhouette Score.

To be submitted on Moodle:

The Jupyter notebook as html-file ‘k-means_clustering_task02.html’ with the changes and short explanations according to b), c), d), e), f) and g).