Reichman University

Introduction to Data Science

Homework Assignment 2 – K-Means

Dr. Gail Gilboa-Freedman
Dr. Naveh Eskinazi
Mr. Asaf Lev

Submission: 04/05/2023

GENERAL INSTRUCTIONS

In the current assignment you will analyze air traffic passenger statistics by the airline with the k-means algorithm.

The work will be based on a CSV named **“air_traffic.csv”** located on the course’s Moodle site.

SUBMISSION:

Through the assignment box within the course Moodle, submit a **Jupyter Notebook file named HWA2_<student name>.ipynb** (e.g. HWA2_avia_malka.ipynb)
Should include all the relevant code needed to perform the assignment’s tasks along with the code's output.
(Recommendation: Add headers and sub-headers using the Markdown option)

Good Luck!

PART 1: PREREQUISITES

TASK 1: SETTING THE FOLDER

1. Create a Jupyter Notebook named **HWA2_<student name>.ipynb.**
2. Download from the CSV file named **“air_traffic.csv”** from Moodle.
3. Upload the CSV file to Jupyter (Note: make sure the file is placed in the same location as your Jupyter Notebook)

TASK 2: IMPORT LIBRARIES & MODULES

4. Import the following libraries and modules within your notebook: **scipy, numpy, matplotlib, pandas, matplotlib.pyplot, KMeans (from sklearn.cluster), and silhouette_score (from sklearn.metrics)**

TASK 3: EXPLORE THE DATA

Use Python commands (e.g., head, columns, and shape) to plot the answers to the following questions:

5. Based on how many **cases** will the algorithm perform the clustering? (Provide a numerical answer).
6. Based on how many **dimensions** will the algorithm perform the clustering? (Provide a numerical answer).
7. How are the **data points** represented in the data? (Provide a verbal answer that will also include an explanation on what the values ‘0’ and ‘1’ represent)

PART 2: BUILDING A K-MEANS MODEL

TASK 4: BUILDING THE MODEL

Use Python commands (i.e., KMeans and fit_predict) to build a K-Means model.

8. Use the Silhouette measure to make a wise selection of a number from 2 to 10 for the **number of clusters (K’s).** (Provide an answer that will also include the value of the silhouette measure)
9. (For the K value chosen in question 8), show and identify the **allocation to clusters** of the first 3 and last 3 data points.(Provide a numerical answer that will include the cluster’s number
10. (Use your own words along with useful statistics like mean values and visual plot of the allocation to clusters) Describe the **main characteristics** of each of the clusters obtained by the model.
11. (Use your own words) Suggest a possible practical and business application that can be used based on the results found in Question 10.