Homework Assignment 2 – K-Means

Dr. Gail Gilboa-Freedman
Dr. Naveh Eskinazi

Submission: 28/04/2022

GENERAL INSTRUCTIONS

In the current assignment you will analyze google users' reviews with the k-means algorithm.

The work will be based on a CSV named **"google_reviews.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_karin_tenne.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 **“google_reviews.csv”** from Moodle.
3. Upload the CSV file to Jupyter (Note: make sure the file is placed ~~places~~ 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?
6. Based on how many **dimensions** will the algorithm perform the clustering?
7. How are the **data points** represented in the data?

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 5 for the **number of clusters (K’s)**
9. (For the K value chosen in the previous question) show and identify the **allocation to clusters** of the first 2 and last 2 data points.
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.

Good Luck!