Alexandria University
Faculty of Engineering
Computer and Systems Engineering Dept.

Prof. Dr. Mohamed Abougabal
Prof Dr. Nagwa M. El-Makky
Special Topics in Information Systems

# Assignment 5
# Data Clustering

## Objectives

1. Introducing different clustering techniques.
2. Learning how to apply, visualize and evaluate different clustering techniques.

## Problem Statement

Unlike classification, clustering does not rely on predefined classes. It is the process of grouping a set of data objects into multiple groups (clusters), so that objects within a cluster have high similarity, but are very dissimilar to objects in other clusters.

You are given the Wholesale customers Data Set; the dataset consists of 440 samples and 8 features. It refers to clients of a wholesale distributor. It includes the annual spending in monetary units (m.u.) on diverse product categories.

It is required to apply some clustering techniques and measure the validity of these clustering techniques for the given dataset. We will focus on partitioning methods (**K-Means**) and hierarchical methods (**Agglomerative Clustering**).

## Lab Session

### 1. Preprocessing

In this step, it is required to apply the following pre-processing methods

- Drop the nominal attributes; **`Region`** and **`Channel`**
- Apply non-linear scaling (e.g. logarithmic) to fix the normality of the data. This is used particularly for financial data (Hint: you may use **`log`** function from **`numpy`**).
- Check for Outliers and remove them if any.
- Apply PCA analysis on the normalized data set (set the number of components to 2)

### 2. Clustering

In this step, it is required to choose the best number of clusters for the following clustering techniques on the pre-processed data set.

- **K-Means:**
  **Parameters:** find the best value for `n_clusters`.
  Use **Silhouette coefficient score** as a metric to find the best `n_cluster` value.
  **Hint**: use **`KMeans`** from **`sklearn.cluster`**

- **Agglomerative Clustering:**
  **Parameters:** find the best value for `n_clusters` when setting the `linkage='average' and linkage='complete'`
  Use **Silhouette coefficient score** as a metric to find the best `n_cluster` value.
  **Hint**: use **`AgglomerativeClustering`** from **`sklearn.cluster`**

### 3. Visualization

In order to visualize generated clusters. You are required to use the 2-D data set ( after applying PCA). For this strep, it is required to do the following:

- For each cluster model generated above, use a scatterplot to plot the data set converted using PCA (use the labels in the clusters to color the plot, i.e. each cluster should have a specific color)

### 4. Evaluation

In order to validate the generated clusters and compare the three techniques. You are required to evaluate their cohesion and separation. For this purpose, it is required to compare the Silhouette coefficient for the selected models. Use **`silhouette_score`** from **`sklearn.metrics`** and report the value for each of the above models on the given dataset.

## Report Requirements

Your report should contain the following:

- Comments on the plots generated from visualizing the clustering models, as well as a comparison of the three plots.
- Detailed analysis of the evaluation of the clustering models on the given dataset.

## Notes

- **Use Google Colab.**
- You should work in groups of two. Each student should answer any question in the lab session.
  - You should deliver a well-documented code as well as a report showing all your work and conclusions.
  - Copied assignments will be penalized; so not delivering the assignment would be much better.
- You should write your code in python.

## References

[1] Chapter 8 of the second reference (P-N Tan, M. Steinbach, A. Karpatne and V. Kumar, “Introduction to Data Mining”, 2nd Edition, Pearson Addison- Wesley, 2018).
[2] Raschka, S. (2015). Python Machine Learning. Packt Publishing.