ASSIGNMENT 3: CLUSTERING

PART A (50 points)

Description of Data

The file **census.csv** in the "Assignments" folder contains information from the USA census data collection. The goal of the analysis is to group people in the United States into distinct subsets based on urbanization, household size, and income factors.

In this cluster analysis, you will do the following:

a) Clean the data set and select variables for clustering. Are all attributes important for this dataset?

b) Use TABLEAU/PowerBI's clustering feature to identify some useful insights . Provide 2 interesting visualizations in this context.

c) Apply **k-means** with different number of clusters (k). Which size of k leads to the most insightful or meaningful clusters? After examining the results, summarize the nature of the clusters. Visualizing the clusters might yield some interesting patterns.

d) Use the EM algorithm to cluster the data and comment on the nature of the clusters produced. Are the clusters produced by EM similar to k-means? Which algorithm would you prefer and why?

PART B (50 points)

Description of Data

The file **cereals.xlsx** in the "Assignments" folder contains customer rating of breakfast cereals. The dataset Cereals.xls includes nutritional information, store display, and consumer ratings for 77 breakfast cereals.

a) Pre-process the data and explain what you did with missing values.

b) Use TABLEAU/PowerBI's clustering feature and present 2 visualizations from this dataset that you find interesting.

c) Apply **k-means** with different number of clusters (k). Which size of k leads to the most insightful or meaningful clusters? Also apply the **EM** algorithm and compare this with the clusters produced by K-means.

d) The elementary public schools would like to choose a set of cereals to include in their daily cafeterias. Every day a different cereal is offered, but all cereals should support a healthy diet. For this goal you are requested to find a cluster of "healthy cereals." Use any clustering algorithm to find and justify your cluster of healthy cereals. Using the Cereals dataset explain why the data should be standardized?