Question 1: Assignment Summary

Briefly describe the "Clustering of Countries" assignment that you just completed within 200-300 words.

Problem Statement: To categorise the countries using some socio-economic and health factors that determine the overall development of the country. From the analysis we need to suggest the countries which the CEO needs to focus on the most.

The solution methodology followed included following steps

- EDA activities done to the initial data
- Outlier Treatment
- K-Means clustering (Silhouette Analysis and Elbow curve)
- Hierarchical clustering (single and complete linkage)
- Plotting the cluster with respect to data
- Cluster Profiling
- Final inference of the list of countries which require aid

EDA performed include conversion of columns health, exports and imports to their actual values. Then bivariate analysis of different countries using 3 target factors child mortality, gdpp and income. Multivariate Analysis was performed using pairplots where the variation of different numerical variables with one another were analysed. Income directly proportional to gdpp, child mortality inversely proportional almost in case all developed countries.

In the Hierarchical clustering the problem faced was that the clusters were skewed. 1 cluster contained almost 100 points another less than 5 points (outlier effect). K- Means clustering was more stable and inferential compared to the hierarchical clustering.

Question 2: Clustering

a) Compare and contrast K-means Clustering and Hierarchical Clustering.

The K-Means clustering is parameterized by the value k, which is the number of clusters that you want to create. The algorithm begins by creating k centroids. It then iterates between an assign step (where each sample is assigned to its closest centroid) and an update step (where each centroid is updated to become the mean of all the samples that are assigned to it. This iteration continues until some stopping criteria is met.

Hierarchical clustering instead, builds clusters incrementally, producing a hierarchy. The algorithm begins by assigning each sample to its own cluster (top level). At each step, the two clusters that are the most similar are merged, the algorithm continues until all of the clusters have been merged.

The import difference between both K-means and hierarchical clustering is on the basis of scalability and flexibility. Hierarchical is flexible but cannot accommodate large data. K- Means is scalable but cannot use for flexible data.

b) Briefly explain the steps of the K-means clustering algorithm.

Choosing K includes the following steps.

Start by choosing K random points the initial cluster centres. Assign each data point to their nearest cluster centre. The most common way of measuring the distance between the points is the Euclidean distance. For each cluster, compute the new cluster centre which will be the mean of all cluster members. Now re-assign all the data points to the different clusters by taking into account the new cluster centres. Keep iterating through the step 3 & 4 until there are no further changes possible. Finally assign each of the data points to their nearest cluster centres based on the Euclidean distance. This way all the points are divided among the K clusters.

c) How is the value of 'k' chosen in K-means clustering? Explain both the statistical as well as the business aspect of it.

The value of best K is selected by **Elbow method** and **Average silhouette Method.**

Elbow Method

Compute K-Means clustering algorithm for different values of k. For instance, by varying k from 1 to 10 clusters. For each k, calculate the total within-cluster sum of square (wss). Plot the curve of wss according to the number of clusters k. The location of a bend (knee) in the plot is generally considered as an indicator of the appropriate number of clusters.

Average silhouette Method

Compute K-Means clustering algorithm for different values of k. For each k, calculate the average silhouette of observations (avg.sil). Plot the curve of avg.sil according to the number of clusters k. The location of the maximum is considered as the appropriate number of clusters.

Business aspect of the k value selection include the strategies or goal of clustering by the marketing team or the team responsible for same depends on the portfolio or target they have to achieve by clustering.

d) Explain the necessity for scaling/standardisation before performing Clustering.

Standardisation of data, that is, converting them into z-scores with mean 0 and standard deviation 1,

is important for 2 reasons in K-Means algorithm:

• Since we need to compute the Euclidean distance between the data points, it is important to

ensure that the attributes with a larger range of values do not out-weight the attributes with

smaller range. Thus, scaling down of all attributes to the same normal scale helps in this

process.

• The different attributes will have the measures in different units. Thus, standardisation helps

in making the attributes unit-free and uniform.

e) Explain the different linkages used in Hierarchical Clustering.

Single Linkage: Here, the distance between 2 clusters is defined as the shortest distance between points in the two clusters

Complete Linkage: Here, the distance between 2 clusters is defined as the maximum distance between any 2 points in the clusters

Average Linkage: Here, the distance between 2 clusters is defined as the average distance between every point of one cluster to every other point of the other cluster