Question 1: Assignment Summary

Briefly describe the "Clustering of Countries" assignment that you just completed within 200-300 words. Mention the problem statement and the solution methodology that you followed to arrive at the final list of countries. Explain your main choices briefly (what EDA you performed, which type of Clustering produced a better result and so on)

Answer:

The CEO of HELP International wants to decide how the recently acquired fund of 10 million dollars should be utilized in the most efficient manner. In particular, the countries which are in dire requirement of aid need to be identified.

The dataset contains socio-economic features of around 167 countries. There were no missing values but a boxplot revealed a few outliers. They were not treated as "outliers" as they either represented very developed/under-developed countries.

After EDA, all the features were scaled. The Hopkins Statistic had a value of 0.89 which proved that there was clustering tendency in the data.

Using the elbow curve, 3 clusters were identified. The Silhouette score (which represents how concrete the clusters are) for k = 3 was also the highest among all the other values of k which was desirable.

Using KMeans for k=3, some variation in the boxplots of GDP/capita, net income/person and child mortality/1000 was seen.

After KMeans, Hierarchical Clustering was used. At first using single linkage, the dendrogram was cut at k=3. The resulting clusters obtained were vague. Then using complete linkage, the dendrogram was cut at k=4 which resulted in distinct clusters. Then KMeans was run again for k=4 by initializing the cluster centres that were calculated by Hierarchical Clustering.

This model was finalized and used to identify the countries which required the aid the most. The top 5 countries that were returned were Burundi, Liberia, Republic of Congo, Niger and Sierra Leone.

Question 2: Clustering

a) **Compare and contrast K-means Clustering and Hierarchical Clustering.**
b) **Briefly explain the steps of the K-means clustering algorithm.**
c) **How is the value of 'k' chosen in K-means clustering? Explain both statistical as well as the business aspect of it.**
d) **Explain the necessity for scaling/standardisation before performing Clustering.**
e) **Explain the different linkages used in Hierarchical Clustering.**

Answer:

a) Similarities:

I. K-means clustering and Hierarchical clustering are both unsupervised machine learning algorithms which are used to cluster numerical data.
II. Both the algorithms use the metric of Euclidean distance to cluster nearby points.

Differences:

I. K-means requires a beforehand knowledge of K whereas there is no such requirement for Hierarchical Clustering.
II. The original positioning of cluster centres have an impact on the final positioning of cluster centres in K-means but in Hierarchical clustering there is no such condition.
III. K-means algorithm is faster and memory efficient than Hierarchical clustering which requires more computational memory.

b) The following steps are implemented while performing K-means clustering.

I. We start by randomly choosing the value of K and initializing K cluster centres.
II. Then based on the metric of Euclidean distance, the distance between each point and all the cluster centres is calculated and the point is assigned to that cluster centre from which the distance is minimum. This is called the assignment step.
III. In the optimization step, the cluster centres are recomputed by using the centroid formula (mean) of the points assigned to it in the assignment step.
IV. The assignment step and the optimization step are repeated till the cluster centres do not update (convergence).

The cost function for the K-means algorithm is as follows.

$$J = \sum_{i=1}^{n} (x_i - \mu_{k(i)})^2$$

where,

x_i is the co-ordinate of the ith data point.

$k(i)$ is the cluster centre to which the ith data point is assigned to.

μ_k is the co-ordinate of the kth cluster centre.

The goal is to minimize the cost function.

c) The value of 'k' is chosen based on both statistical aspect using the elbow curve and silhouette score and also using business aspect. The optimal value of k returned by the elbow curve is analysed to see if it can be interpreted from the business point of view also.

d) It is very important to standardize the numerical columns before clustering otherwise one numerical feature having a high range of values will overpower in determining the cluster centres. Standardization helps in equalizing the contribution of each feature towards the cluster centre by converting all the features to a common scale/range.

e) In hierarchical clustering, we first start with n distinct clusters and then combine clusters based on the distance metric. To calculate the distance of a cluster with another cluster containing more than one data point, we define linkage. There are three types of linkages in hierarchical clustering. They are as follows:

I. Single Linkage: This type of linkage uses the minimum distance metric to combine clusters.
II. Complete Linkage: This type of linkage uses the maximum distance metric to combine clusters.
III. Average Linkage: This type of linkage uses the average distance metric to combine clusters.