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)

1. **Problem Statement:** HELP International is an international humanitarian NGO that is committed to fighting poverty and providing the people of backward countries with basic amenities and relief during the time of disasters and natural calamities. After the recent funding programmes, they have been able to raise around $ 10 million. Now the CEO of the NGO needs to decide how to use this money strategically and effectively. The significant issues that come while making this decision are mostly related to choosing the countries that are in the direst need of aid.
2. **Solution Methodology:**
    a. Data inspection
    b. EDA tasks suitable for this dataset
        i. Data cleaning
        ii. Univariate analysis
        iii. Bivariate analysis
    c. Data preparation for clustering
        i. Outlier Treatment
        ii. Feature Scaling
        iii. Hopkin Check
    d. Perform Clustering
        i. KMean Clustering
            - choose K using both Elbow and Silhouette score
            - Run K-Means with the chosen K
            - Visualize the clusters
            - Cluster profiling using “gdpp, child_mort and income”
                - Hierarchical Clustering
            - Use both Single and Complete linkage
            - Choose one method based on the results
            - Visualise the clusters
            - Clustering profiling using “gdpp, child_mort and income”
    e. Final Model
        i. Final Model Selection and labelling
        ii. Select model based on cluster results
    f. Conclusion
        i. Top 5 countries selection for financial aid

ii. Top 5 countries selection on some socio-economic and health factor

Question 2: Clustering

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

→ k-means Clustering:

1. k-means, using a pre-specified number of clusters, the method assigns records to each cluster to find the mutually exclusive cluster of spherical shape based on distance.
2. K Means clustering needed advance knowledge of K i.e. no. of clusters one want to divide your data.
3. One can use the median or mean as a cluster centre to represent each cluster.

Hierarchical Clustering:

1. Hierarchical methods can be either divisive or agglomerative.
2. In hierarchical clustering one can stop at any number of clusters, one finds appropriate by interpreting the dendrogram.
3. Agglomerative methods begin with ‘n’ clusters and sequentially combine similar clusters until only one cluster is obtained.

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

→ The algorithm for the K-means algorithm is as follows:

- Select initial centroids. The input regarding the number of centroids should be given by the user.
- Assign the data points to the closest centroid
- Recalculate the centroid for each cluster and assign the data objects again
- Follow the same procedure until convergence. Convergence is achieved when there is no more assignment of data objects from one cluster to another, or when there is no change in the centroid of 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.

→ ‘K’ value is chosen randomly in K-Means clustering based on the statistical aspect. From the business aspect, we need to first understand the dataset and based on that

we decide a number of ‘k’. for example, we have a dataset of variables like ‘pen’, ‘pencil’, ‘books’, ‘notebooks’, ‘mobiles’, ‘charger’, ‘laptop’. Now if we want to have k values based on statistical aspect, we can use silhouette score to determine that but based on the business aspect, after viewing the dataset we can easily make cluster = 2, one in the electronics category and another non-electronics.

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

→ It is a good idea to do scaling/standardisation because our variables may have units at different scale and as our method stresses more on the calculation of direction of space or distance, so if we have one variable with high scale units then while calculating for k-Means or hierarchical it will create a big difference as the clusters will tend to move with the variables having greater values or variances. By applying standardisation/scaling will increase the performance of our model.

e) Explain the different linkages used in Hierarchical Clustering.

→ Linkage is a technique used in Agglomerative Clustering. Linkage helps us to merge two data points into one using below linkage technique

Single Linkage

In single linkage hierarchical clustering, the distance between two clusters is defined as the shortest distance between two points in each cluster.

Complete Linkage

In complete linkage hierarchical clustering, the distance between two clusters is defined as the longest distance between two points in each cluster.

Average linkage

The distance between two clusters is the average distance between every point of one cluster to the another every point of another cluster.

Ward linkage

The distance between clusters is calculated by the sum of squared differences with all clusters.