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)

Note: You don't have to include any images, equations or graphs for this question. Just text should be enough.

Problem Statement:

- To analyse the facts and figures in dataset of Countries, 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. Now decision of how to use this money strategically and effectively providing countries who need AID.

Analysis Approach:

- Following EDA steps starting from inspecting the dataframe and doing data outlier treatment, Non-Graphical Analysis to Graphical Analysis continuing with Scaling, Checking the tendency of the data: Hopkins Test, finding the best value for K by SSD & silhouette method, Performing KMeans with the final value of k then Visualizing the clusters using scatter plot profiling: GDPP, CHILD_MORT, INCOME after than using Hierarchical clustering (single & complete) finding the required cluster and comparing list of countries from both clusters.

Inference of Univariate analysis: Child mortality, income, gdpp, exports, imports, health inflation have large outlier values. Multi - Peaks of variable shows more promising clustering and hence helping in outlier treatment as well.

Correlation Matrix gives better insights about data:

- Imports Exports increases simultaneously

- As GDP grows health and income increases

- Fertility and child moratlity gets infected directly by life_expectancy

Hopkin’s value for 10+ iteration is more than .85 henceforth, data is good for clustering.

Kmeans Clustering:
K= 3 and K=4 looks appropriate, according to SSD and Silhouette Curve. Hence, K ==3 and K==4 has cluster with same countries of highest child mortality, total fertility and low GDP.

Hierarchical Clustering:

Result of complete linkage depicts 3 clusters, when mapping cluster labels and countries with least GDPP and hight child mortality and total fertility. We get the same set of countries as from Kmeans.

List of Countries in both the clusters:

- Burundi
- Liberia
- Congo, Dem. Rep
- Niger
- Sierra Leone
- Mozambique
- Central African Republic
- Malawi
- Eritrea
- Madagascar

I find Hierarchical clustering better than Kmeans as there is no need to assume the value according to curve, we can directly see division of clusters in complete dendrogram and can proceed with clustering.

Question 2: Clustering

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

- Hierarchical clustering can’t handle big data well but K Means clustering can. This is because the time complexity of K Means is linear i.e. O(n) while that of hierarchical clustering is quadratic i.e. $O(n^2)$.
- In K Means clustering, since we start with random choice of clusters, the results produced by running the algorithm multiple times might differ. While results are reproducible in Hierarchical clustering.
- K Means is found to work well when the shape of the clusters is hyper spherical (like circle in 2D, sphere in 3D).
- K Means clustering requires prior knowledge of K i.e. no. of clusters you want to divide your data into. But, in case of hierarchical clustering we can find number of clusters by interpreting the dendrogram

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

Algorithmic steps for k-means clustering

- Let X = $\{x_1,x_2,x_3,\ldots\ldots..,x_n\}$ be the set of data points and V = $\{v_1,v_2,\ldots\ldots.,v_c\}$ be the set of centers.
- 1) Randomly select *‘c’* cluster centers.
- 2) Calculate the distance between each data point and cluster centers.
- 3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers.
- 4) Recalculate the new cluster center using:
- where, *‘c_i’* represents the number of data points in i^{th} cluster.
- $v_i = (1/c_i) \sum_{j=1}^{c_i} x_i$
- 5) Recalculate the distance between each data point and new obtained cluster centers.
- 6) If no data point was reassigned then stop, otherwise repeat from step 3).

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

Two methods that can be useful to find mysterious k in k-Means. These methods are:

The Elbow Method: Calculate the Within-Cluster-Sum of Squared Errors (WSS) for different values of k, and choose the k for which WSS becomes first starts to diminish. In the plot of WSS-versus-k, this is visible as an elbow. Within-Cluster-Sum of Squared Errors sounds a bit complex. Let’s break it down:

1. The Squared Error for each point is the square of the distance of the point from its representation i.e. its predicted cluster center.

2. The WSS score is the sum of these Squared Errors for all the points.

3. Any distance metric like the Euclidean Distance or the Manhattan Distance can be used.

The Silhouette Method: The silhouette value measures how similar a point is to its own cluster (cohesion) compared to other clusters (separation). The range of the Silhouette value is between +1 and -1. A high value is desirable and indicates that the point is placed in the correct cluster. If many points have a negative Silhouette value, it may indicate that we have created too many or too few clusters.

The Elbow Method is more of a decision rule, while the Silhouette is a metric used for validation while clustering. Thus, it can be used in combination with the Elbow Method. Therefore, the Elbow Method and the Silhouette Method are not alternatives to each other for finding the optimal K.

In business perspective, objectives and available resources must be evaluated when finalizing the clusters.

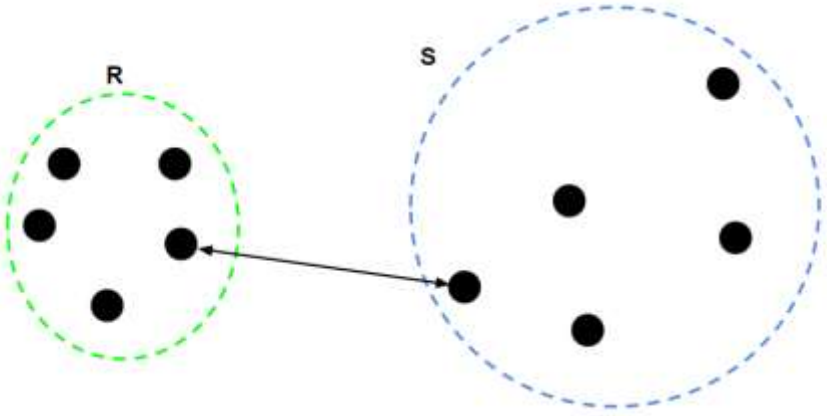
d) Explain the necessity for scaling/standardisation before performing Clustering.

All such distance-based algorithms are affected by the scale of the variables. Standardization prevents variables with larger scales from dominating how clusters are defined. **It** helps to make the relative weight of each variable equal by converting each variable to a unitless measure or relative distance. It allows all variables to be considered by the algorithm with equal importance. So that all the features contribute equally to the result.

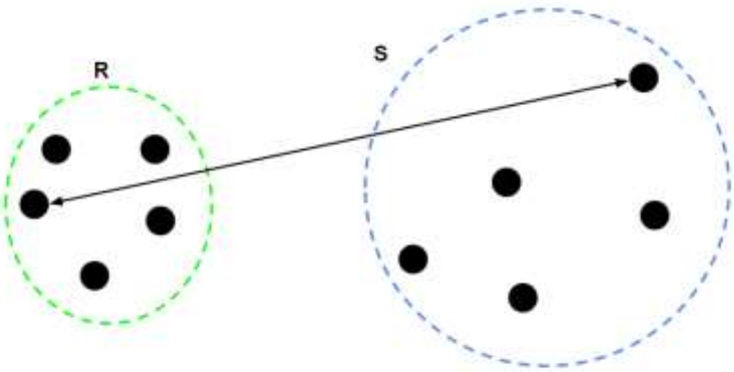
e) Explain the different linkages used in Hierarchical Clustering.

The process of Hierarchical Clustering involves either clustering sub-clusters(data points in the first iteration) into larger clusters in a bottom-up manner or dividing a larger cluster into smaller sub-clusters in a top-down manner. During both the types of hierarchical clustering, the distance between two sub-clusters needs to be computed. The different types of linkages describe the different approaches to measure the distance between two sub-clusters of data points. The different types of linkages are:-

1. **Single Linkage:** For two clusters R and S, the single linkage returns the minimum distance between two points i and j such that i belongs to R and j belongs to S.



2. **Complete Linkage:** For two clusters R and S, the single linkage returns the maximum distance between two points i and j such that i belongs to R and j belongs to S.



3. **Average Linkage:** For two clusters R and S, first for the distance between any data-point i in R and any data-point j in S and then the arithmetic mean of these distances are calculated. Average Linkage returns this value of the arithmetic mean.

where

– Number of data-points in R

– Number of data-points in S