Assignment: Part 2

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 analyze 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.

- ***Approach:***

Following EDA steps starting from inspecting the data frame 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.

- ***EDA: Univariate / Bivariate Analysis Inferences:***
  - Exports and imports are following a linear trend.
  - Income and life expectancy increases as GDP increase. Which is expected.
  - Life expectancy decreases with increase fertility and child mortality.

- ***Outlier Treatment:***

Capped Upper range outliers for Exports, health, imports, income, life_expect, gdp to 0.93 percentile. They are not removed fully as this would hamper the data.

- ***For 10 iteration of HOPKINS TEST value is greater than .86 Thus, data is good for clustering.***

- ***After scaling , plotting SSD and silhouette score for Kmeans Clustering, the best value of k came out to be k=3 & k=4.***

- ***Kmeans Clustering:***
  K ==3 & K==4 has cluster with almost 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 high 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
- Madagascar
- Mozambique
- Central African Republic
- Malawi
- Togo
- Eritrea

I find KMeans method to be more precise because in hierarchal method visual representation of clusters is not as visible as in KMeans. Although in KMeans is tough to choose the k value. In hierarchal clustering it's easier since dendrogram gives clear picture. There is almost no difference in countries of both method.

Question 2: Clustering

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

K-Means Clustering	Hierarchical Clustering
It is a collection of data points in one cluster which are similar between them and not similar data points belongs to another cluster.	Clusters have tree like structures and most similar clusters are first combine which continues until we reach a single branch.
We need to have desired number of clusters ahead of time.	We can decide the number of clusters after completion of plotting dendrogram by cutting the dendrogram at different heights.
Works very good in large dataset	Works well in small dataset and not good with large dataset
K-means only used for numerical.	Hierarchical clustering is used when we have variety of data as it doesn't require to calculate any distance.
The main drawback of k-Means is it doesn't evaluate properly outliers.	Outliers are properly explained in hierarchical clustering

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

Step 1: Choose the number of clusters k.

Step 2: Select k random points from the data as centroids
Next, we randomly select the centroid for each cluster. Let's say we want to have 2 clusters, so k is equal to 2 here. We then randomly select the centroid: let's say C1 and C2.

Step 3: Assign all the points to the closest cluster centroid
Once we have initialized the centroids, we assign each point to the closest cluster centroid: Let's say C1 to red and C2 to green.
If points which are closer to the red point are assigned to the red cluster whereas the points which are closer to the green point are assigned to the green cluster.

Step 4: Recompute the centroids of newly formed clusters
Now, once we have assigned all of the points to either cluster, the next step is to compute the centroids of newly formed clusters.

Step 5: Repeat steps 3 and 4

Note: Stopping Criteria for K-Means Clustering
There are essentially three stopping criteria that can be adopted to stop the K-means algorithm:

- Centroids of newly formed clusters do not change
- Points remain in the same cluster
- Maximum number of iterations are reached

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.

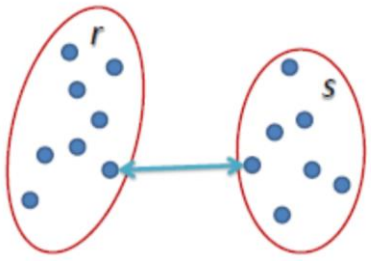
d) Explain the necessity for scaling/standardisation before performing Clustering.

It is definitely a good idea to do scaling/standardisation because our variables may have units at different scale and as our method stresses more on 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.

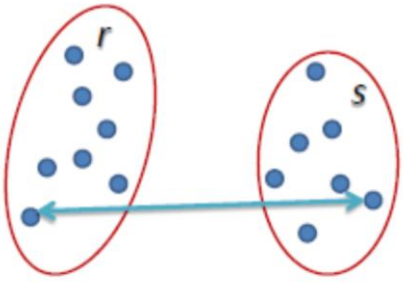
The process of Hierarchical Clustering involves either clustering sub-clusters (data points in the first Iteration) into larger clusters in a bottom-up manner or dividing a larger cluster into smaller sub-clusters in a top-down manner. During both the types of hierarchical clustering, the distance between 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.



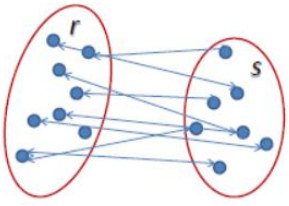
$$L(r,s) = \min(D(x_{ri}, x_{sj}))$$

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.



$$L(r,s) = \max(D(x_{ri}, x_{sj}))$$

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



$$L(r,s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} D(x_{ri}, x_{sj})$$