Assignment 2

Q1: 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. Given dataset is the socio economic and health parameters of all countries. The case study requires me to find out the top backwards countries in direst need of relief/aid during the time of disasters and natural calamities using these parameters so that NGO called HELP International can actually help the right candidates
2. After the preparation steps of loading data I performed univariate and bi-variate analysis and reported findings
3. Next steps, would be to make the data ready for clustering modelling, which requires scaling all numerical variables using which we would model
4. Performed silhouette analysis and elbow curve analysis learnt that values after cluster 3 where somewhat similar so went with max three clusters
5. Went on and performed K means clustering and using the cluster labels drew visualizations which would be useful in drawing insights
6. Performed box plot with these three clusters and reported findings
7. The facts which I incurred where the countries with high child mortality rate, low income with low GDPP are the countries which required the utmost aid/relief.
8. Performed Hierarchical clustering with the scaled data available from above steps.

Q2: Clustering

A. Compare and contrast K-means Clustering and Hierarchical Clustering.

K-means Clustering

1. K-means can handle big data well as time complexity is linear.
2. In K Means clustering, since we start with random choice of clusters, the results produced by running the algorithm multiple times might differ.
3. K Means clustering requires prior knowledge of K i.e. no. of clusters you want to divide your data into.

Heirarchical Clustering

1. Hierarchical clustering cannot handle big data as time complexity is quadratic (not scalable)
2. Results are reproducible as there is no random number of clusters to be selected.
3. We can stop at whatever number of clusters you find appropriate in hierarchical clustering by interpreting the dendrogram.

B. Briefly explain the steps of the K-means clustering algorithm.

The dataset is partitioned into K clusters and the data points are randomly assigned to the clusters resulting in clusters that have roughly the same number of data points.

For each data point:

1. Calculate the distance from the data point to each cluster.
2. If the data point is closest to its own cluster, leave it where it is. If the data point is not closest to its own cluster, move it into the closest cluster.
3. Repeat the above step until a complete pass through all the data points' results in no data point moving from one cluster to another. At this point the clusters are stable and the clustering process ends.

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

When you use a k-means clustering algorithm, you will need to select the number of clusters you would like to work with.

Working with the optimal number of clusters for our data and market environment will facilitate the use of resources in a more efficient and effective manner. We can select the number of clusters using industry- related knowledge or three different statistical methods when we use the k-means algorithm.

The Elbow method: To determine the optimal number of clusters, we will need to run the k-means algorithm for different values of k (number of clusters). For each value of k, we will then need to calculate the total within-cluster sum of squares (wss). We can then plot the values of wss on the y-axis and the number of clusters (k) on the x-axis. The optimal number of clusters can be read off the graph at the x-axis.

The Silhouette coefficient: To determine the optimal number of clusters, we will need to measure the quality of the clusters that were created. This value determines how closely each data point is to the centroid of its cluster. The optimal number of clusters is, the maximised silhouette value for the data set

D. Explain the necessity for scaling/standardisation before performing Clustering.

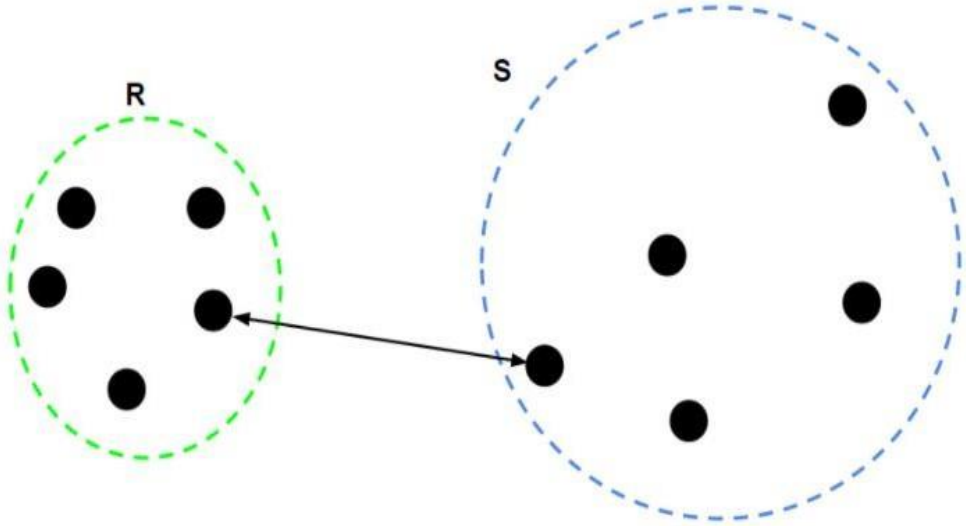
Algorithms are affected by the scale of the variables and since the units are not same if we do not scale the variables then they might not be correctly interpreted.

In our data set let's consider income and health column and health column now. The income is the average income of the whole country while health is average health of an individual person. Now income column consists of values in 10000 and health consists of values in double digits, so when we find the Euclidian distance between any two points the value will be majorly affected by income column as it has higher magnitude. We do not want our algorithm to be affected by the magnitude of these variables. The algorithm should not be biased towards variables with higher magnitude. To overcome this problem, we can bring down all the variables to the same scale.

So it becomes an essential step before clustering as Euclidean distance is very sensitive to the changes in the differences.

E. Explain the different linkages used in Hierarchical Clustering.

Single Linkage: For two clusters R and S, the single linkage returns the minimum distance between two points of the clusters



Complete Linkage: For two clusters R and S, the single linkage returns the maximum distance between two points of the clusters