Question1: Assignment Summary

ANS:

- *Analysis Approach:*

• Data Collection and Data Cleaning:

Importing the data then cleaning it, checking if there are any null values. We found out that some columns were in %of gdpp form so corrected them using the correct formulas.

• Visualising data:

We detected outliers by visualising the data, outliers were treated according to our problem statement. We also found out that some Variables are highly corelated to each other.

• Outliers Detection and treatment:

There were outliers in almost every column. Some outliers like in gdpp column for example, have outliers on high end of spectrum which we can remove safely because high gdpp countries won't need urgent aid. For this analysis we capped the gdpp column at .95 quantile at upper end and .1 quantile at lower end. We did not cap other variable because that would affect our analysis for finding poorest countries with high child mortality.

• Scaling data:

Standardizing all the continuous variables.

• Hopkins's test:

To check if data has tendency to form clusters.

• Kmeans Clustering:

Identifying the "k" through silhouette analysis and elbow curve. Then forming the cluster on scaled data, the adding the cluster id on original data for better interpretation of data. And visualizing the clusters.

• Hierarchical Clustering

Identifying optimal number for k by analysing dendrogram. Then forming the cluster on scaled data and adding the cluster label to original data for better interpretation. Visualisation of clusters was also done.

• Decision Making:

Successfully identified the top 10 countries by analysing both model which are in dire need of Aid.

Summary

We got same top 10 countries from both hierarchical and k-means model that are in dire need of Aid.

Following are the countries name requiring aid.

1. Sierra Leone
2. Central African Republic
3. Haiti
4. Chad
5. Mali
6. Nigeria
7. Niger
8. Angola
9. Congo, Dem. Rep.
10. Burkina Faso

Question 2: Clustering

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

ANS:

K-means Clustering	Hierarchical Clustering.
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.	Hierarchical methods can be either divisive or agglomerative.
K Means clustering needed advance knowledge of K i.e., no. of clusters one wants to divide your data.	In hierarchical clustering one can stop at any number of clusters, one finds appropriate by interpreting the dendrogram.
One can use median or mean as a cluster centre to represent each cluster.	Agglomerative methods begin with 'n' clusters and sequentially combine similar clusters until only one cluster is obtained.
Methods used are normally less computationally intensive and are suited with very large datasets.	Divisive methods work in the opposite direction, beginning with one cluster that includes all the records and Hierarchical methods are especially useful when the target is to arrange the clusters into a natural hierarchy.

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

ANS: The way K-means algorithm works is as follows:

1. Specify number of clusters K.

2. Initialize centroids by first shuffling the dataset and then randomly selecting K data points for the

centroids without replacement.

3. Keep iterating until there is no change to the centroids.

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

ANS: There is a popular method known as 'elbow method' which is used to determine the optimal value of K to

perform the K-Means Clustering Algorithm. The basic idea behind this method is that it plots the various

values of cost with changing k. As the value of K increases, there will be fewer elements in the cluster. So

average distortion will decrease. The lesser number of elements means closer to the centroid. So, the

point where this distortion declines the most is the elbow point.

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

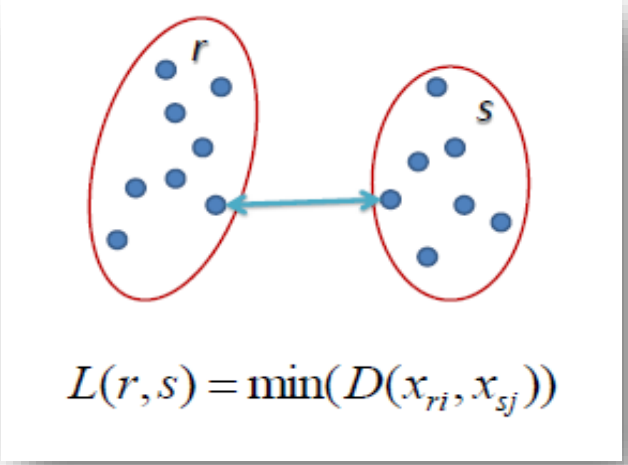
ANS: The issue is what represents a good measure of distance between cases. If you have two features, one where the differences between cases is large and the other small, are you prepared to have the former as almost the only driver of distance? So, for example if you clustered people on their weights in kilograms and heights in meters, is a 1kg difference as significant as a 1m difference in height? Does it matter that you would get different clustering's on weights in kilograms and heights in centimetres? If your answers are "no" and "yes" respectively then you should probably scale.

- k-nearest neighbours with a Euclidean distance measure are sensitive to magnitudes and hence should be scaled for all features to weigh in equally.

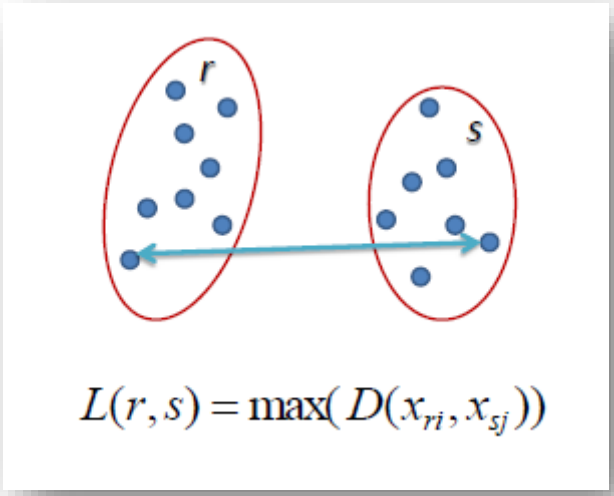
e) Explain the different linkages used in Hierarchical Clustering.

ANS:

- Single Linkages: In single linkage hierarchical clustering, the distance between two clusters is defined as the shortest distance between two points in each cluster. For example, the distance between clusters “r” and “s” to the left is equal to the length of the arrow between their two closest points.



- Complete Linkages: In complete linkage hierarchical clustering, the distance between two clusters is defined as the longest distance between two points in each cluster. For example, the distance between clusters “r” and “s” to the left is equal to the length of the arrow between thei two furthest points.



- Average Linkages: In average linkage hierarchical clustering, the distance between two clusters is defined as the average distance between each point in one cluster to every point in the other cluster. For example, the distance between clusters “r” and “s” to the left is equal to the average length each arrow between connecting the points of one cluster to the other.