Question 1: Assignment Summary

1.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)

Problem Statement : To choose the countries that are in the direst need of aid Based on socio-economic and health factors.

Using K-means Clustering & Hierarchical Clustering Following 5 countries are identified as the countries which aid at the earliest.

1.Congo, Dem. Rep.

2.Liberia

3.Burundi

4.Niger

5.Central African Republic

Based on elbow curve & silhouette analysis 3 cluster is the optimum number of cluster.Out of all the features 3 features (gdpp,income & child_mort) are chosen as important feature to form cluster as the development of country can be decided based on these factors and if they need aid or not based on how developed a country is. The countries having low income & low gdpp & high child_mort are the countries which are in need of aid.So in K- means clustering cluster 1 & in Hierarchical clustering cluster 0 is having such case and by sorting values we get final list of above countries.

Dist plot & box plot has been used for EDA.K- means clustering produced better result.

Question 2: Clustering

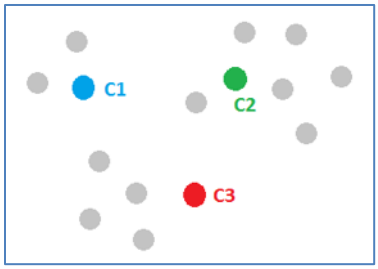
a)Compare and contrast K-means Clustering and Hierarchical Clustering.

K- Means Clustering	Hierarchical Clustering
Based on Silhouette Analysis & elbow curve we need to find K(no. of clusters) and give to K- means clustering.	We don’t need to provide value of K.
It takes less time so can handle big data well	Takes more time so can’t handle big data well.

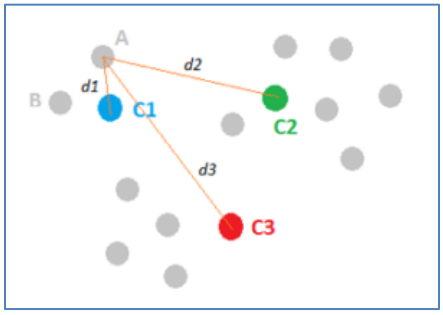
b)Briefly explain the steps of the K-means clustering algorithm.

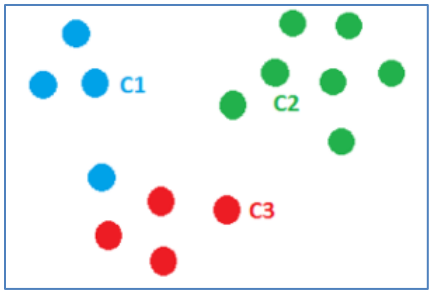
k-means clustering aims to find the set of k clusters such that every data point is assigned to the closest center, and the sum of the distances of all such assignments is minimized.

1. Initialize cluster centers.

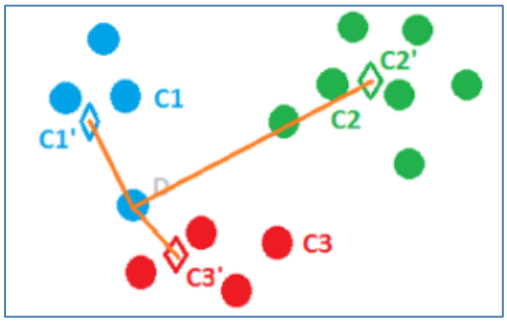


2. Assign observations to the closest cluster center.

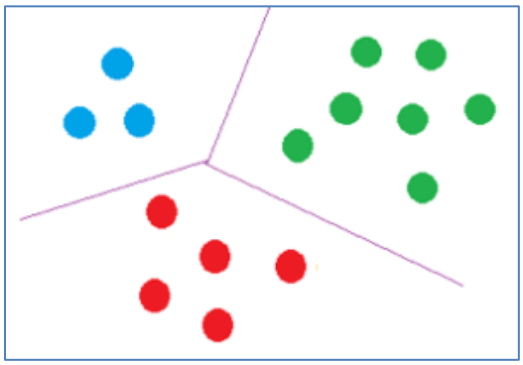




3. Revise cluster centers as mean of assigned observations



4. Repeat step 2 and step 3 until convergence



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

The value of K in K-means clustering chosen by elbow curve & Silhouette Analysis.

In elbow curve we select that k value from which further there is not much change in elbow curve.

In Silhouette Analysis we select that value of k from which further there is not much change in Silhouette Score

From business point of view if we understand the no. of cluster then we don't need to do analysis using elbow curve & Silhouette Score directly we can give the value of K.

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

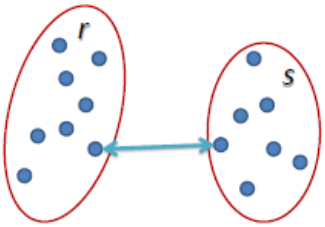
Since the distance metric used in the clustering process is the Euclidean distance, we need to bring all our attributes on the same scale.

The standard example is considering age (in year) and height (in cm). The age may range in [18 50], while the height may range in [130 180]. If we use the Euclidean distance, the height will have disproportionately more importance in its computation with respect to the age.

e) Explain the different linkages used in Hierarchical Clustering.

Single Linkage

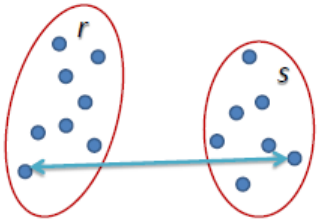
In single linkage the distance between two clusters is defined as the shortest distance between two points in each cluster.



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

Complete Linkage

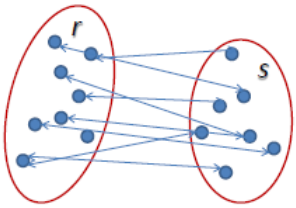
In complete linkage the distance between two clusters is defined as the longest distance between two points in each cluster.



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

Average Linkage

In average linkage the distance between two clusters is defined as the averagedistance between each point in one cluster to every point in the other cluster.



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