# ASSIGNMENT PART II

**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)

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

**ANS**: **Problem Statement**: Our main objective for this assignment is to find the countries that are in direst need of aid. To identify Top 10 countries which are in direst need of aid from the analysis work.

The steps ascertained for the Assignment are hereunder:

## a). 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.

## b). 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.

- Boxplots to view the outliers if any.
- Pair plots and heatmap to view the relation between individual features with each other.

## c). Outliers Detection and treatment:

The clustering process is very sensitive to the presence of outliers in the data. In this case, specifically the outliers need not to be dropped at all. As all the countries are to be considered for evaluation. The extremities should be considered, if any, and may form a characteristic of cluster. Thus, soft capping was implemented.

## d). Scaling data:

The distance metric used in the clustering process is the Euclidean distance therefore, standardized scaling was attempted as it is important for clustering and all the attributes of the countries data be on the same scale.

## e). Hopkins test:

To check if data has tendency to form clusters.

## f). 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.

## g). Hierarchical clustering:

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

- Single Linkage Clustering
- Multiple Linkage Clustering

## h). Decision making:

Successfully got same top 10 countries by analyzing both models i.e., **Kmeans clustering and Hierarchical clustering** which are in dire need of Aid. The Top 10 countries are as Follow:

- Sierra Leone
- Central African Republic
- Haiti
- Chad
- Mali
- Nigeria
- Niger
- Angola
- Congo, Dem. Rep.
- Burkina Fas

I prefer to choose Hierarchical clustering process is good for this reason is that:

- Eliminating the k Means limitation of predefined consideration of number of clusters.
- Data set is small.
- Slightly more data points considered in Un-developed Countries Segment (Cluster- labels = 0).

## 2. CLUSTERING

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

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

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

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

e) Explain the different linkages used in Hierarchical Clustering.

**ANS**: **a). K-means Clustering:**

- We need to have desired number of clusters ahead of time.
- It is a collection of data points in one cluster which are similar between them and not similar data points belongs to another cluster.
- Works very good in large dataset.
- The main drawback of k-Means is it doesn’t evaluate properly outliers.
- K-means only used for numerical.

## **Hierarchical Clustering**:

- We can decide the number of clusters after completion of plotting dendrogram by cutting the dendrogram at different heights.

- Clusters have tree like structures and most similar clusters are first combine which continues until we reach a single branch.

- Works well in small dataset and not good with large dataset.

- Outliers are properly explained in hierarchical clustering.

- Hierarchical clustering is used when we have variety of data as it doesn't require to calculate any distance.

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

## **ANS:**

- Randomly select K points as initial centroids.

- All the data points closet to the centroid will create cluster centre according to Euclidean distance function.

- Once we assign all the points to each of k clusters, we need to update the cluster centres or centroid of that cluster created.

- Repeat 2,3 steps until cluster centres reach 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.**

**ANS**: 'K' value is chosen randomly in K-Means clustering based on statistical aspect. From business aspect, we need to first understand the dataset and based on that we decide number of 'k'. for example, we have a dataset of variables like 'pen', 'pencil', 'books', 'notebooks', 'mobiles', 'charger', 'laptop'. Now if we want to have k values based on statistical aspect, we can use silhouette score to determine that but based on business aspect, after viewing the dataset we can easily make cluster = 2, one in electronics category and another non-electronics.

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

**ANS**: 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.**

**ANS**: Linkage is a technique used in Agglomerative Clustering.

Linkage helps us to merge two data points into one using below linkage technique.

**Single linkage**: The distance between two clusters is calculated by the minimum distance between two points from each cluster.

**Complete linkage**: The distance between two clusters is calculated by the maximum distance between two points from each cluster.

**Average linkage**: The distance between two clusters is the average distance between every point of one cluster to the another every point of other cluster.

**Ward linkage**: The distance between clusters is calculated by the sum of squared differences with all clusters.