# Clustering Assignment Part – II

**Question 1: Assignment Summary**

**Briefly describe the "Clustering of Countries" assignment that you just completed within 200-300 word. 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)**

## Answer: -

**Problem Statement:**

HELP International is an international 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. It runs a lot of operational projects from time to time along with advocacy drives to raise awareness as well as for funding purposes.

**Project goal:-**

The requisite is:

- To categorise the countries using some socio-economic and health factors that determine the overall development of the country.
- To suggest the countries which the CEO needs to focus on the most.

**Method followed:-**

- It was found that there were no null values.
- There were also no duplicate values in the dataset.
- Converted the value of the variables to the absolute value which are before in percentage.

**Exploratory Data Analysis:-**

- Try to find the values of variables are normally distributed or not using dist plot.
- There were a few outliers and they were treated with mid range quintiles.
- Used box plot to finding the outliers.
- Scaled the data using Standard Scalar.

**Clustering (K-Means):-**

- Tried to find the Hopkins Statistics score it was around 88 – 91 %.
- To find the optimal number of k used Elbow curve and got k = 3 is good to go with it.
- To find the goodness for clustering used silhouette score.

- Build final K-Means model with k = 3
- Assigned the Clusters with the data frame and plotted using Scatter plot
- Done Cluster profiling with respect to the Clusters.

**Clustering (Hierarchical):-**

- Used Single Linkage and Complete Linkage to find the dendograms.
- Using cut-tree method got the clusters and assigned to the data frame and plotted with Scatter Plot.
- Done Cluster profiling with respect to the Clusters.

Finally using all these values clusters of 3 were formed and the countries are split into clusters.

**Question 2: Clustering**

**a) Compare and contracts K-means Clustering and Hierarchical Clustering.**

**Answer: -**

- K Means needs prior knowledge of number of centroid (K) whereas hierarchical cluster do not need this kinds of parameters. The cut_tree () function is used to create the number of clusters of any choice.
- In K Means the algorithm will calculate the centroid in each time.
- K Means is faster as compare to Hierarchical clustering.
- Hierarchical clusters need more RAM to run.

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

**Answer: -**

- K Means algorithm is the process of dividing the N data points into K groups or clusters.
- Start by choosing K random points the initial cluster centres.
- Assign each data points to their nearest cluster centre. The most common way of measuring the distance between the points to the Euclidean distance.
- For each cluster, calculate the new cluster centre which will be mean of all cluster members.
- Now re-assign all the data points to the different clusters by talking into account the new cluster centres.
- Keep iterating through the step 3&4 until there are no further changes possible.

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

**Answer: -**

- There are two ways we can choose what should be the number of k.

1) **Elbow method**:-
   - Compute K Means clustering algorithm for different values of k, e.g. (1-10) clusters.
   - For each k, calculate the total within-cluster sum of square distance.
   - Plot the curve of SSD according to the number of clusters k.
   - The location of bend (Knee) in the plot is generally considered as an indicator of the appropriate number of clusters.
2) **Average Silhouette Method:-**
   - Compute K Means clustering algorithm e.g. (1-10) clusters.
   - For each k, calculate the average silhouette of observations.
   - Plot the curve of avg.sil according to the number of clusters k.
   - The location of the maximum is considered as the appropriate number of clusters.

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

**Answer: -**

- Standardisation of data, that is, converting them into z-scores with mean 0 and standard deviation 1, is important for two reasons in K-Means algorithm:
  - Since we need to compute the Euclidean distance between the data points, it is important to ensure that the attributes with a larger range of values do not out-weight the attributes with smaller range. Thus, scaling down of all attributes to the same normal scale helps in this process.
  - The different attributes will have the measures in different units. Thus, standardisation helps in making the attributes unit-free and uniform.
  - We create the instance of the standard scalar like this,
    standard_scaler = StandardScaler()

**e) Explain the different linkages used in Hierarchical Clustering.**

**Answer: -**

- There are three types of linkages are present in Hierarchical Clustering.

1) **Single Linkage:-**
    - In Single Linkage, the distance between 2 clusters is defined as the shortest distance between points in the two clusters.
2) **Average Linkage:-**
    - In Average Linkage, the distance between 2 clusters is defined as the average distance between every point of one cluster to every other point of the other cluster.
3) **Complete Linkage**:-
    - In Complete Linkage, , the distance between 2 clusters is defined as the maximum distance between any 2 points in the clusters.