# Assignment: Part II

## Question 1: 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)

**Answer:**

- Our main objective for this assignment is to find the countries that are in direst need of aid. Our job is to find those countries using socio-economic and health factors which will show overall development of country.

- So in order get those countries we need to analyze by stepwise .it has 167 countries with no missing values. And in data description they clearly mentioned about the health, imports and exports are % of GDP so transformed it to absolute values.

- Then started visualization outlier analysis using box plots and found many features having outliers and with respect to variable preferences capped the outliers to extreme limits of box plot , for example child_mort , inflation and total_fer deal only about lower outliers and capped to lower extreme point. And performed heat map for given data frame and found high level of multi colliniority and hence performed some insightful actions. And distribution plot for to see groups in the features with respect to values inside it found maximum groups in distribution of features.

- Before concluding about the clustering we need one more test to get whether the data set is suitable for clustering or not by conducting HOPKINS Statistics test it found at 0.85 and more likely to clustering.

- Scaled the data before going for cluster analysis using standardscaler and performed silhouette and sum of squared analysis to get suitable cluster numbers.

- Started clustering with Kmeans with 3 clusters and found cluster 2 is having high child_mort rate and low gdpp , income so taken consideration by visualizing and calculating the clusters comparisons found top 10 countries that direst need of aid ,sorted with respect to child mortality in descending and gdpp, income in ascending order.

- In the same way we performed the hierarchical clustering to with 2 clusters since we don’t want to fall in short of countries in some clusters if wee deep down in number of clusters ,using to clusters again we compared the variables and got cluster 0 as final selection and sorted same as kmeans result and stored in a variable.

- **Top 5 countries :1.Haiti ,2.Sierra Leone , 3.Chad, 4.Central African Republic, 5.Mali**

## Assignment: Part II

# Question 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/standardization before performing Clustering.
e) Explain the different linkages used in Hierarchical 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 want to divide your data. | In hierarchical clustering one can stop at any number of clusters, one find appropriate by interpreting the dendrogram. |
| One can use median or mean as a cluster center 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. |
| In K Means clustering, since one start with random choice of clusters, the results produced by running the algorithm many times may differ. | In Hierarchical Clustering, results are reproducible in Hierarchical clustering |
| **Advantages:**<br>1. Convergence is guaranteed. | **Advantages:**<br>1 .Ease of handling of any forms of similarity or distance.<br>2. Consequently, applicability to any attributes types. |

## Assignment: Part II

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

**ANS)**

K-Means clustering is an unsupervised learning algorithm. There is no labeled data for this clustering, unlike in supervised learning. K-Means performs division of objects into clusters that share similarities and are dissimilar to the objects belonging to another cluster.

The term ‘K’ is a number. You need to tell the system how many clusters you need to create. For example, K = 2 refers to two clusters. There is a way of finding out what is the best or optimum value of K for a given data.

For a better understanding of k-means, let's take an example from cricket. Imagine you received data on a lot of cricket players from all over the world, which gives information on the runs scored by the player and the wickets taken by them in the last ten matches. Based on this information, we need to group the data into two clusters, namely batsman and bowlers.

**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)**

When you use a k-means clustering algorithm, you will need to select the number of clusters ‘k’ you would like to work with. Selecting the optimal number of clusters is important because this will fall somewhere between full localization or standardization

- The Elbow method: To determine the optimal number of clusters, you will need to run the k-means algorithm for different values of k (number of clusters). For each value of k, you will then need to calculate the total within-cluster sum of squares (wss). You 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, you 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. A high average silhouette coefficient indicates successful clusters. This method checks the silhouette coefficient for different values of k. The optimal number of clusters is, therefore, the maximized silhouette value for the data set.
- The Gap Statistic: To determine the optimal number of clusters, you will need to know the variation between clusters for different values of k with their expected values of distribution with no clusters.

## Assignment: Part II

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

**Ans)**

When you are working with data where each variable means something different, (e.g., age and weight) the fields are not directly comparable. One year is not equivalent to one pound, and may or may not have the same level of importance in sorting a group of records. In a situation where one field has a much greater range of value than another (because the field with the wider range of values likely has greater distances between values), it may end up being the primary driver of what defines clusters. Standardization helps to make the relative weight of each variable equal by converting each variable to a unitless measure or relative distance.

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

**Ans)**

1. **Single Linkage:** For two clusters R and S, the single linkage returns the minimum distance between two points i and j such that i belongs to R and j belongs to S.
2. **Complete Linkage:** For two clusters R and S, the single linkage returns the maximum distance between two points i and j such that i belongs to R and j belongs to S.
3. **Average Linkage:** For two clusters R and S, first for the distance between any data-point i in R and any data-point j in S and then the arithmetic mean of these distances are calculated. Average Linkage returns this value of the arithmetic mean.