**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

**Answer-** I Started with Loading the Dataset and then explored about the various features using commands like .shape, .info and .describe() after Analyzing the Dataset. I started Checking for Null Values in the dataset, but there were No Null values present in the dataset. In Data Preparation I have Converted the export, import, health variables into actual values ,as they are given as %age of GDP per capita .After that I plotted the Heatmap representation of correlation matrix of Numeric Features of Dataset .now, its time to perform Univariate Analysis on the Features Plotted the distribution of the Variables and Analyzed the pattern. after that Bivariate Analysis is performed by plotting the Pair plot of the dataset. Next, I did the Visualization of the Outliers and some of the Outliers in the Dataset are handled by Capping the upper values of import, export, income and gdpp. After that I Calculated the Hopkins statistic To check the Randomization in dataset. Data Scaling using standard scaler is performed next, Then I have Plotted the Silhouette score and Elbow curve-ssd. Which helped me in Choosing the value of K for K-means Clustering. After that I visualized the Cluster with GDP, INCOME AND CHID_MORT Only as they the Major contributing Features And Top 5 Countries which are in Direst need of Aid are selected Which are –

1. Congo, Dem. Rep.
2. Liberia
3. Burundi
4. Niger
5. Central African Republic

After that I Performed Hierarchical Clustering . I have made tha dendrogram for Single Linkage And Performed Hierarchical Clustering On Complete Linkage. I Selected the value of K=4 for Hierarchical Clustering. After that I Started Visualizing the Cluster with GDP, INCOME AND CHID_MORT Only as they the Major contributing Features. when I filtered out the data of child_mort, income and gdpp for different clusters .I saw that cluster 2 is having the lowest income and gdpp distribution ,Also I can see that the child_mort is very high in this cluster. This is the cluster we were looking for which is in Direst need of help And after Sorting the Top 5 Countries which are in Direst need of Aid. I got these countries –

1. **Congo, Dem. Rep.**
2. **Liberia**
3. **Burundi**
4. **Niger**
5. **Central African Republic**

## Question 2: Clustering

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

**Answer-** In K-means Clustering we need to Choose the number of clusters k. then Select k random points from the data as centroids and Assign all the points to the closest cluster centroid.then, Recompute the centroids of newly formed cluster.whereas in Hierarchical Clustering we don't need to pre define the value of K ,but by looking at the dendrogram we can decide the value of K carefully. Hierarchical Clustering Take computationally more time than K-means Clustering.

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

**Answer-** In k-means clustering in the beginning we determine number of cluster K then Select k random points from the data as centroids and Assign all the points to the closest cluster centroid.then, Recompute the centroids of newly formed cluster till there is no change in the positions of the Centroids and we assume these centroid or center of these clusters. We can take any random objects as the initial centroids or the first K objects in sequence can also serve as the initial centroids.

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-** The value of K in k-means clustering can be choosen with various types of statistics like using Silhouette score Statistics or Elbow-curve Statistics. 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. Clusters can also be formed by by keeping the business aspect in mind if you want more variety of customers you can cluster them in more numbers .

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

**Answer-** When we standardize the data before performing cluster analysis, so that there is no affect of outliers on clusters. We find that with more equal scale. Standardization prevents variables with larger scales from dominating how clusters are defined.

e) Explain the different linkages used in Hierarchical Clustering.

**Answer-** There are three types of Linkages used in hierarchical Clustering-

1. Single Linkage - In statistics, single-linkage clustering is one of several methods of hierarchical clustering. It is based on grouping clusters in bottom-up fashion (agglomerative clustering), at each step combining two clusters that contain the closest pair of elements not yet belonging to the same cluster as each other.

2. Complete Linkage - At the beginning of the process, each element is in a cluster of its own. The clusters are then sequentially combined into larger clusters until all elements end up being in the same cluster.

3. Average Linkage - Average-linkage is where the distance between each pair of observations in each cluster are added up and divided by the number of pairs to get an average inter-cluster distance. Average-linkage and complete-linkage are the two most popular distance metrics in hierarchical clustering

.