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)

Ans:

Problem Statement:

HELP International is an international humanitarian NGO that is committed to fighting poverty and providing the people of back ward 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.

After the recent funding programmes, NGO have been able to raise around $ 10 million. Now the CEO of the NGO needs to decide how to use this money strategically and effectively.

Goal:

- Cluster the countries based on economic condition and significant factors provided in dataset.
- List down the top 5 countries which are in the direst need of aid.

SOLUTION METHODOLOGY:

To achieve the above mention goal, firstly I have done basic EDA which are as follows:

1. Inspected the dataset to get the information about it
2. Validated the missing value count for all the columns.
3. Converted the **exports, imports and health columns values from percentages to absolute values.**
4. **Fixed the outlier** for 'exports', 'health', 'imports', 'income' and 'gdpp' columns by capping technique
5. Done the **univariate and bivariate** analysis to understand the data better.
6. **Validated the collinearity between the variables** and observed that there is high collinearity between lot of variable pairs but most of them important for model building so **based the significance dropped 'imports' and 'health' columns**.

After performing the EDA **scaled the data using StandardScaler**. Before going to model building **Hopkins test** has been performed to examines whether data points differ significantly from uniformly distributed data in the multidimensional space. This test had given **Hopkins score as 0.931** which is pretty good.

After that find out the optimal value of K using **Elbow curve/SSD and Silhouette score** which came out as **K=3.**

Finally applied the K-means algorithm and build the model at **K=3.** Analyzed the clusters results by plotting the scatter plot against the label column for multiple variables.

Compared the clusters based on gdpp, child_mort , income columns and **identified the cluster which is having lowest average gdpp , lowest average income and highest mortality rate**. Then list down the top 5 countries which are in direst need aid.

Hierarchical Clustering:

Plotted the single/complete/average linkage dendrograms and observed that single and average linkage dendrogram not making much sense so **considered Complete Linkage dendrogram** for further analysis.

Cut the Complete linkage dendrogram to **obtain 3 clusters** and analyzed the clusters to identify the top 5 countries which are in dire need of aid. But observe that cluster which contain poor countries is having only one country **So cut**

the complete linkage dendrogram again to obtain 2 clusters and analyzed the cluster and identified the top 5 countries which are in dire need of aid.

Observed the top 5 countries obtain at k-means clustering are same as Hierarchical cluster when dendrogram is cut at 2 clusters. Countries are as follows:

- Burundi
- Liberia
- Congo, Dem. Rep.
- Niger
- Sierra Leone

Question 2: Clustering

1.Compare and contrast K-means Clustering and Hierarchical Clustering.

Ans:

k-means is method of cluster analysis using a pre-specified no. of clusters. It requires **advance knowledge of 'K'**.

Hierarchical clustering also known as hierarchical cluster analysis (HCA) is also a method of cluster analysis which seeks to build a hierarchy of clusters without having fixed number of cluster.

Main differences between K means and Hierarchical Clustering are:

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 wants to divide your data.	In hierarchical clustering one can stop at any number of clusters, one finds appropriate by interpreting the dendrogram.
One can use median or mean as a cluster centre 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.
K- means clustering a simply a division of the set of data objects into non- overlapping subsets (clusters) such that each data object is in exactly one subset).	A hierarchical clustering is a set of nested clusters that are arranged as a tree.
K Means clustering is found to work well when the structure of the clusters is hyper spherical (like circle in 2D, sphere in 3D).	Hierarchical clustering doesn't work as well as, k means when the shape of the clusters is hyper spherical.

2. Briefly explain the steps of the K-means clustering algorithm.

Ans:

The way k-means algorithm works is as follows:

1. Specify number of clusters *K*.
2. Initialize centroids by first shuffling the dataset and then randomly selecting *K* data points for the centroids without replacement.
3. Keep iterating until there is no change to the centroids. i.e assignment of data points to clusters isn't changing.
    - Compute the sum of the squared distance between data points and all centroids.
    - Assign each data point to the closest cluster (centroid).
    - Compute the centroids for the clusters by taking the average of the all data points that belong to each cluster.

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

Ans:

As per the statistical aspect, K value is chosen randomly in K-Clustering. And it can be performed using Elbow Curve and Silhoutte Score. If it's about business aspect, comprehension of dataset is important. In that sense, we can decide how many clusters to be taken.

4.Explain the necessity for scaling/standardization before performing Clustering.

Ans:

Though clustering techniques use Euclidean Distance, it will be wise to scale the variables. Ex:- heights in meters and weights in KGs before calculating the distance. It may create a big difference while calculating for K-Means and Hierarchical. This is because the cluster will tend to move variable having greater values or variances. In fact,

most clustering algorithms are even highly sensitive to scaling. Rescaling the data can completely ruin the results.

5. Explain the different linkages used in Hierarchical Clustering.

Ans:

In Agglomerative clustering, linkage plays a vital role in merging two clusters into one.

Different linkages that are used in Hierarchical clustering are as follows:

Single-Linkage: Single-linkage (nearest neighbour) is the shortest distance between a pair of observations in two clusters. It can sometimes produce clusters where observations in different clusters are closer together than to observations within their own clusters. These clusters can appear spread-out.

Complete-Linkage: Complete-linkage (farthest neighbour) is where distance is measured between the farthest pair of observations in two clusters. This method usually produces tighter clusters than single-linkage, but these tight clusters can end up very close together. Along with average-linkage, it is one of the more popular distance metrics.

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.

Centroid-Linkage: Centroid-linkage is the distance between the centroids of two clusters. As the centroids move with new observations, it is possible that the smaller clusters are more similar to the new larger cluster than to their individual clusters causing an inversion in the dendrogram. This problem doesn't arise in the other linkage methods because the clusters being merged will always be more similar to themselves than to the new larger cluster.