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 humanitarian 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. After the recent funding programmes, they 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. The significant issues that come while making this decision are mostly related to choosing the countries that are in the direst need of aid.

Objective:

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:

1. **Data Reading and understanding**
2. **Preparing the data & Data cleaning**
3. **Data Visualization:**
    - **Univariate Analysis**
    - **Bivariate analysis**
4. **Prepare the data for modelling (Outliers Treatment)**
5. **Cluster tendency Check (Hopkins Statistics)**
6. **Modeling using KMeans & Finding the optimal k :**
    - **Elbow curve &**
    - **Silhouette Analysis Techniques**
7. **K-mean clustering**
8. **Hierarchical clustering**
9. **Conclusion**

1. **Data Reading and understanding:** Imported the dataset and understood the data . Read the datatypes well. Also read the shape and the statistical point of view.

2. **Preparing the data & Data cleaning:** This process performed the cleaning of the data. First checked any null values in any of the columns then after checking that found none and further proceeded with the duplicate rows. The dataset was very clean with no null values and duplicates rows. Converted the Three features export, import and health to the actual values

3. **Data viualization:**
Univariate Analysis: Performed the displot & Box plot for all the features
Bivariate Analysis: Visualizing numeric variables by plotting Pairplot and Heatmap.

4. **Outlier treatment:**

- Since my data is less I will not go for removing the data from the dataset. Outlier capping is one option.
- For all columns except "child_mort" and "total_fer" we will perform capping on the upper range, because we do not wish to loose critical information about the countries that are in need of help.

5. Hopkins Statistics

- Every time we run the above cell we get different hopkins value but if we observe we see its above 80 that means our dataset has a very good tendency to form clusters

After treating the outliers we did the scaling which is the most important step before clustering.

6. Modeling using KMeans & Finding the optimal k :

The elbow curve and the Silhouette Analysis help in finding the optimal k for clustering. it is good to proceed with 3 clusters.

7. **K-mean clustering**: When cluster with the k=3 we got three clusters in which the cluster 0 showing the behaviour of the countries which are in the dire need of the aid. Country In cluster 0 are exhibiting the following behaviour :

- **High child mortality**
- **Low gdpp**
- **Low income**

Countries from k-mean are :

1.Haiti	6.Nigeria
2.Sierra Leone	7.Niger
3.Chad	8.Angola
4.Central African Republic	9.Congo, Dem. Rep.
5.Mali	10.Burkina Faso

8. **Hierarchical clustering:**
The countries in the cluster 0 shows the behaviour of high child mortality, low income and low gdpp . So we can go ahead with the countries in this cluster. After clustering and sorting with the features high child mortality and low income and low gdpp we get countries similar to the K-mean clustering.

Countries from Hierarchical clustering are:

1.Haiti	6.Nigeria
2.Sierra Leone	7.Niger
3.Chad	8.Angola
4.Central African Republic	9.Congo, Dem. Rep.
5.Mali	10.Burkina Faso

9. Conclusion: After analysing the cluster through k-mean and hierarchical clustering the k-mean cluster provided better clarity on the countries exhibiting similar behaviours like high child mortality, low gdpp and low income.
Also, if notice countries from k-mean and the hierarchical clustering provides similar countries.

And low income and low gdpp indicates the country is poor country still developing and thus needs the aid.
The final countries are been considered from k-mean as the neat behaviour is been seen in the k-mean clustering.(However the hierarchical clustering too displays the similar countries)
.

Final Countries:

- **Haiti**
- **Sierra Leone**
- **Chad**
- **Central African Republic**
- **Mali**
- **Nigeria**
- **Niger**

Question 2: Clustering

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

Answer: -

K-Means Clustering	Hierarchical Clustering
We need to have desired number of clusters ahead of time.	We can decide the number of clusters after completion of plotting dendrogram by cutting the dendrogram at different heights
It is a collection of data points in one cluster which are similar between them and not similar data points belongs to another cluster.	Clusters have tree like structures and most similar clusters are first combine which continues until we reach a single branch.
Works very good in large dataset	Works well in small dataset and not good with large dataset
The main drawback of k-Means is it doesn't evaluate properly outliers.	Outliers are properly explained in hierarchical clustering
K-means only used for numerical.	Hierarchical clustering is used when we have variety of data as it doesn't require to calculate any distance.
Method to find the optimal number of clusters is by The Elbow method, Silhouette analysis.	Method to find the optimal number of clusters is by Dendrogram.
Python Library used is sklearn – KMeans	Python Library used is sklearn-AgglomerativeClustering
Category of K-Means Clustering is- Centroid based , Partition-based	Category of Hierarchical Clustering is- Hierarchical , Agglomerative

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

Answer: -

- Step 1 : Initialization - First thing in K-means clustering algorithm is to assign random centroid to the clusters, choosing K number of centroids as designated by business problem.

- Step 2: Cluster Assignment - After initialization of the centroids, then the data points close to that centroid forms a cluster. If we use the Euclidean distance as a measure to calculate the distance between the centroid and the data point, then distance is calculated and the minimum distance is used in assigning the datapoint to that particular cluster.

- Step 3 : Recalculation of Centroid - After the datapoints converge to form a cluster, the centroid is recalculated for the newly formed cluster. This created new centroid is the mean of all the newly assigned datapoints belonging to a particular cluster.

- All the above steps are iterated over and over again till they converge, i.e., till the calculated centroid doesn't change in value or remains almost the same.

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 basic idea behind partitioning methods, such as k-means clustering, is to define clusters such that the total intra-cluster variation [or total within-cluster sum of square (WSS)] is minimized. The total WSS measures the compactness of the clustering and we want it to be as small as possible. The Elbow method looks at the total WSS as a function of the number of clusters: One should choose a number of clusters so that adding another cluster doesn’t improve much better the total WSS.

- Average silhouette method briefly measures the quality of a clustering. That is, it determines how well each object lies within its cluster. A high average silhouette width indicates a good clustering.

- Average silhouette method briefly measures the quality of a clustering. That is, it determines how well each object lies within its cluster. A high average silhouette width indicates a good clustering.

- These above methods are the **Statistic methods** of finding out the value of K in accordance to the higher inter-cluster distance and low intra-cluster distance.

- The **Business aspect** usually deals with the requirement of the client. Sometimes, the requirement might be to have a certain number of clusters to understand the business problem. Then the clustering

assignment would involve that number of clusters.

- For example, if a retailer asks to classify their customers into 3 clusters. He has to classify the customers as always buying, mostly buying and very rarely buying from their store, that client would specify to have 3 as the number of customers in a general sense.

- The decision should be based on the purpose of the analysis. Sometimes forming many clusters doesn't make any sense instead making many clusters may confuse the main objective of the analysis and thus end up with the data which might not make any sense.
  So keeping the business point of view and also the purpose on what data is been analysed the k value should be decided.

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()

- The possible disadvantages of not standardizing is increased computational time to converge and misclassifications.

e) Explain the different linkages used in Hierarchical Clustering.

Answer: -

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

1. Single Linkage: Here, the distance between 2 clusters is defined as the shortest distance between points in the two clusters

2. Complete Linkage: Here, the distance between 2 clusters is defined as the maximum distance between any 2 points in the clusters

3. Average Linkage: Here, 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.