Question 1: Assignment Summary

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 project that included a lot of awareness drives and 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.

And this is where you come in as a data analyst. Job is to categorise the countries using some socio-economic and health factors that determine the overall development of the country. Then need to suggest the countries which the CEO needs to focus on the most.

Solution Methodology

Below are the steps are basically followed:

Step 1: Reading and Understanding the Data

Step 2: Data Cleansing

- Missing Value check

Step 3: Data Visualization

Step 4: Data Preparation

- Rescaling

Step 5: Hopkins Statistics Test

- Hopkins Score Calculation

Step 6: Model Building

- K-means Clustering
- Elbow Curve
- Silhouette Analysis
- Hierarchical Clustering

Step 8: Final Analysis

- Final Country list Preparation

The following approach is suggested:

- The necessary data inspection and EDA tasks are done like, data cleaning, univariate analysis, bivariate analysis etc.

- **Outlier Analysis:** Performed Outlier Analysis on the dataset. However, there is no flexibility of removing the outliers if it suits the business needs or a lot of countries are getting removed. Hence, need to find the outliers in the dataset, and then choose whether to keep them or remove them depending on the results We get.

Question 2: Clustering

a) Compare and contrast K-means Clustering and Hierarchical Clustering.
   i. In K means clustering we have to define the number of clusters to be created beforehand, which is sometimes difficult to say. Whereas in Hierarchical clustering data is automatically formed into a tree shape form (dendrogram) and we can choose which trees are significant
   ii. Hierarchical clustering can't handle big data well but K Means clustering can. This is because the time complexity of K Means is linear, while that of hierarchical clustering is quadratic
   iii. In K Means clustering, since we start with random choice of clusters, the results produced by running the algorithm multiple times might differ. While results are reproducible in Hierarchical clustering.
   iv. K Means is found to work well when the shape of the clusters is hyper spherical (like circle in 2D, sphere in 3D).

b) Briefly explain the steps of the K-means clustering algorithm.
   i. Clusters the data into *k* groups where *k* is predefined.
   ii. Select *k* points at random as cluster centres.
   iii. Assign objects to their closest cluster canter according to the *Euclidean distance* function.
   iv. Calculate the centroid or mean of all objects in each cluster.
   v. Repeat steps 2, 3 and 4 until the same points are assigned to each cluster in consecutive rounds.

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

1) Statistical Aspect:

- **Elbow method**
   i. Compute clustering algorithm (e.g., k-means clustering) for different values of k. For instance, by varying k from 1 to 10 clusters.
   ii. For each k, calculate the total within-cluster sum of square (wss).

iii. Plot the curve of wss according to the number of clusters k.

iv. The location of a bend (knee) in the plot is generally considered as an indicator of the appropriate number of clusters.

- Average silhouette method

i. Compute clustering algorithm (e.g., k-means clustering) for different values of k. For instance, by varying k from 1 to 10 clusters.

ii. For each k, calculate the average silhouette of observations (*avg.sil*).

iii. Plot the curve of *avg.sil* according to the number of clusters k.

iv. The location of the maximum is considered as the appropriate number of clusters.

2) Business Aspect: categorise the countries using some socio-economic and health factors that determine the overall development of the country.

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

Feature scaling is essential for machine learning algorithms that calculate distances between data. Since the range of values of raw data varies widely, in some machine learning algorithms, objective functions do not work correctly without normalization. *K*-means needs to compute means, and the mean value is not meaningful on this kind of data which is given for assignment. the k-means minimizes the error function using the Newton algorithm, i.e. a gradient-based optimization algorithm. Normalizing the data improves convergence of such algorithms.

e) Explain the different linkages used in Hierarchical Clustering.

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

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

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.