Part 2 Questions:

Question 1 : Assignment Summary

- 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.

Aim: is to categorise the countries using some socio-economic and health factors that determine the overall development of the country.

- Used heatmap to know the correlation initially, and then went for outlier treatment and did the treatment by capping the extremes. Not by using Elbow curve, even with the help of silhouette score, we finally decided the number of cluster = 4 would be optimum. Went for the modelling and modelled. And finally found out countries with cluster id ‘0’ to be most vulnerable.

Question 2: a) Compare and contrast K-means Clustering and Hierarchical Clustering

K-Means Clustering	Hierarchical Clustering
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.
Clustering needed to decide number of clusters one want to divide your data.	One can stop at any number of clusters, one can find appropriate by interpreting the dendrogram.
Since one start with random choice of clusters, the results produced by running the algorithm many times may differ.	Results are reproducible in Hierarchical clustering
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.
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	Hierarchical clustering don’t work as well as, k means when the shape of the clusters is hyper spherical.
Convergence is guaranteed.	Ease of handling of any forms of similarity or distance and consequently, applicable to any attributes types.

Question 2: b) Briefly explain the steps of the K-means clustering algorithm

Step 1: Choose the number of clusters *k*

The first step in k-means is to pick the number of clusters, k.

Step 2: Select k random points from the data as centroids

Next, we randomly select the centroid for each cluster. Let’s say we want to have 2 clusters, so k is equal to 2 here. We then randomly select the centroid.

Step 3: Assign all the points to the closest cluster centroid

Once we have initialized the centroids, we assign each point to the closest cluster centroid.

Step 4: Recompute the centroids of newly formed clusters

Now, once we have assigned all of the points to either cluster, the next step is to compute the centroids of newly formed clusters.

Step 5: Repeat steps 3 and 4

Stopping Criteria for K-Means Clustering

There are essentially three stopping criteria that can be adopted to stop the K-means algorithm:

1. Centroids of newly formed clusters do not change
2. Points remain in the same cluster
3. Maximum number of iterations are reached

Question 2: c) How is the value of 'k' chosen in K-means clustering? Explain both the statistical as well as the business aspect of it.

The **value of K** (number of clusters) can be determined using different methods like the *Elbow curve, Sillhouette curve, and Intercluster distances or other Clustering Algorithms.*

- ✓ *Elbow Curve* - The elbow method is a heuristic used in determining the number of clusters in a data set. The method consists of plotting the explained variation as a function of the number of clusters and picking the elbow of the curve as the number of clusters to use.
- ✓ *Sillhouette curve* - The silhouette coefficient calculates the density of the cluster by generating a score for each sample based on the difference between the average intra-cluster distance and the mean nearest-cluster distance for that sample normalized by the maximum value. We can find the optimal value of K by generating plots for different values of K and selecting the one with the best score depending on the cluster's assignment.

Question 2: d) Explain the necessity for scaling/standardisation before performing Clustering

Standardisation of data, that is, converting them into z-scores with mean 0 and standard deviation 1, is important for 2 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.

Question 2: e) Explain the different linkages used in Hierarchical Clustering

There are three types of Linkages:

- ✓ **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