Q1. 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.

After the recent funding programs, 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.

My job is to categorize the countries using some socio-economic and health factors that determine the overall development of the country. Then you need to suggest the countries which the CEO needs to focus on the most. The datasets containing those socio-economic factors.

First I load and inspect the data set and also check for null values then check for outliers in every numerical variables. After that use, the Winsorization technique at 5th and 95th percentile which implies values that are less than the value at the 1st percentile are replaced by the value at 5th percentile, and values that are greater than the value at the 95th percentile are replaced by the value at 95th percentile.

Then perform silhouette analysis and draw elbow curve to select an appropriate number of clusters and run the k means algorithm on our standardized data.

Then analyzed different clusters to identify the clusters that contain countries which are having a high rate of child mortality and low rate of income and GDP.

Also tried hierarchical clustering to obtain clusters and finally compare the result of two algorithms to give an ultimate solution.

I used a box plot and scatter plot to analyze different clusters with respect to child mortality, income, and GDP.

Q2. Clustering

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

The time complexity of K Means is linear i.e. O(n) while that of hierarchical clustering is quadratic i.e. O(n2)
Difference in execution was not experienced as dataset was very small.
K Means clustering requires prior knowledge of K i.e. no. of clusters we want to divide our data into. But, we can stop at whatever number of clusters we find appropriate in hierarchical clustering by interpreting the dendrogram
One can use median or mean as a cluster centre to represent each cluster in k means algorithm.
Hierarchical algorithm used Agglomerative methods that begin with 'n' clusters and sequentially combine similar clusters until only one cluster is obtained.
Kmeans follow a less computationally intensive process but hierarchical clustering is computationally very intensive.
Kmeans clustering is generally designed for large dataset on the other hand hierarchical clustering worked well with small dataset as it very computationally intensive process.

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

k-means clustering algorithm used to form k number of distinct clusters from a dataset.

1. Initialization and Assignment: First, we need to mention the number of k or the number of clusters we want to obtain from the dataset. Then we have to choose k random number as the cluster center those numbers can be the points belongs to the dataset or can be a totally different point.
2. Then we found the points which are near to the cluster center for each k clusters using squared Euclidean distance.

3. Optimization: Then in every cluster again cluster center is computed in such a way that for kth cluster center is the mean of a vector of p features for every data point.
4. Iteration: This process of assignment and optimization is repeated until there is no change in the cluster centers.

Q. 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 can be chosen from silhouette analysis.

The silhouette plot displays a measure of how close each point in one cluster is to points in the different like neighbouring clusters and hence provides a way to obtain apropriate number of clusters visually. This measure has generally a range of [-1, 1].

our silhouette analysis showing the highest silhouette score in k is equal to 2 but chose k as 2 is just divide the data set into two parts hence chose k value as 3 also draw elbow curve to verify our assumptions.

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

As we have different scales of data in our dataset hence if we don't do the scaling before making clusters one variable may have more weightage than other variables hence it is important to standardize our data before performing any clustering algorithm.

Q. Explain the different linkages used in Hierarchical Clustering.

There are two types of linkages used in hierarchical clustering. One is single linkage another is complete linkage

In a single linkage, the distance between two clusters is defined as the shortest distance between two points in each cluster.

In complete linkage, the distance between two clusters is defined as the longest distance between two points in each cluster.