Question 1: Assignment Summary

HELP International is an international humanitarian NGO that is committed to fighting poverty and provides basic amenities to backward countries and with recent funding, CEO of the NGO needs the countries that are in dire need to aid.

We have been provided with a dataset that has income, child mortality, inflation, fertility rate, exports, imports, life expectancy, health along with the country names. We performed EDA on columns, GDPP and Income, Income and Child Mortality, GDPP and Child Mortality.

We have also capped the values at higher level for all columns expect for GDPP and Inflation and the values were scaled using StandardScaler.

KMeans and Hierarchical Clustering were performed to find the cluster of countries that are in dire need of aid and KMeans gave the best result. As CEO is expecting at least of 5 countries that need attention, Hierarchical Clustering isn't providing more than 1 country and KMeans 46 countries that can be considered.

Question 2: Compare and Contrast KMeans and Hierarchical Clustering

KMeans 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.

Question 3: Briefly explain the steps of KMeans Clustering

Step 1: Initialize the cluster centroids
Step 2: Assign observations to closest cluster center
Step 3: Revise cluster centers as mean of assigned observations
Step 4: Repeat step 2 and step 3 until convergence

Question 4: How is the value of 'k' chosen in K-means clustering? Explain both the statistical as well as the business aspect of it.

K is chosen by using either of the below two analysis

1. Elbow Curve Analysis
2. Silhouette Analysis

On plotting the score of above analysis for say 2 to 10 cluster, we choose the highest score cluster size.

Question 5: Explain the necessity for scaling/standardisation before performing Clustering.

Scaling is done to eliminate redundant data and ensures that good quality of clusters are generated. Clustering algorithm uses Euclidean Distance that is highly prone to irregularities in the size of various features and is why data is scaled before performing clustering.

Question 6: Explain the different linkages used in Hierarchical Clustering.

There are three types of linkages in the Hierarchical Clustering –

1. Complete
    - Distance between two cluster defined as the maximum distance between any two points in cluster.
2. Single
    - Distance between two cluster is defined as the shortest distance between points in two clusters.
3. Average
    - Distance between two cluster is defined as the average distance between every point of the one of the cluster to every other point of the other cluster.