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 funding programmes, they have been able to raise around $ 10 million.
As a data analyst, our job is to come up with the list of five countries that are in the direst need of aid.

Solution Approach:

- Read and Understand The Data
    - Import the Data
    - Find the data quality issues
- Data Cleaning and Preparation for Modelling
    - Converted the columns exports, health and imports from percentage to actual numbers
    - check the distribution of each columns
    - Outlier Treatment : My approach here is to cap the high value outliers for each column except child_mort.
    - Scaling/ Standardization
    - Hopkins Statistic to check cluster tendency - Hopkins statistic 0.84, that shows the formation of cluster tendency is good
- K Means Clustering
    - Find the optimal number of clusters by using elbow curve method and silhouette analysis – Optimal no. of clusters is 3
    - Perform K-Means Modelling
    - Visualization on the clusters: Cluster 0 contains countries with high gdpp, low child_mort and high income. Cluster 1 contains countries with low gdpp, medium child_mort and medium income. Cluster 2 contains countries with the lowest gdpp, highest child_mort and the lowest income. Therefore, the countries in cluster 2 are in the direst need of aid.
    - Cluster Profiling
    - Identify the countries which require aid
- Hierarchical Clustering
    - Create Dendrograms Using Single and Complete Linkage: Complete linkage dendrograms seems to easily interpretable and clear. Looking at the height of the dendrograms, the optimal number of clusters will be three.
    - Visualization on the clusters: Cluster 0 contains countries with very low gdpp,high child_mort and low income. Cluster 1 contains countries with high income, low child_mort and high income. Cluster 2 contains countries with low gdpp, highest child_mort and very low income. It seems cluster 2 contains less number of data points also. Therefore, cluster 0 and cluster 2 are in the direst need of aid.
    - Cluster Profiling
    - Identify the countries which require aid
- Conclusion: I found both the results from K Means and Hierarchical clustering to be same for the five countries that are in the direst need of aid. Those five countries are as follows.
    - Burundi
    - Liberia
    - Congo, Dem. Rep.

- Niger
- Sierra Leone

Question 2: Clustering

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

K-means Clustering	Hierarchical Clustering
This clustering technique uses a pre-specified number of clusters as k and then assign data points to each cluster	This clustering technique can be agglomerative or divisive.
Methods are Less computationally intensive. Therefore, it can handle large datasets well.	It cannot handle large datasets well.
K Means clustering needed advance knowledge of K, the number of clusters to divide the data.	In hierarchical clustering, one can stop at any number of clusters, which is appropriate by interpreting the dendrogram.

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

K-Means algorithm is the process of dividing the N data points into K groups or clusters. The steps of the algorithm are as follows.

1. Start by choosing K random points the initial cluster centers.
2. Assign each data point to their nearest cluster center. The most common way of measuring the distance between the points is the Euclidean distance.
3. For each cluster, compute the new cluster center, which will be the mean of all cluster members.
4. Now re-assign all the data points to the different clusters by taking into account the new cluster centers.
5. Keep iterating through the step 3 & 4 until there are no further changes possible.

At this point, we can arrive at the optimal clusters.

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

There are various methods, which can help us decide the K for our K-means algorithm. Two such methods are as follows.

1) **Elbow method**:
   - Compute clustering algorithm for different values of k. For instance, by varying k from 1 to 10 clusters.
   - For each k, calculate the total within-cluster sum of square (wss).
   - Plot the curve of wss according to the number of clusters k.
   - The location of a bend (knee) in the plot is generally considered as an indicator of the appropriate number of clusters.

2) Average silhouette Method:
   - Compute clustering algorithm for different values of k. For instance, by varying k from 1 to 10 clusters.

- For each k, calculate the average silhouette of observations.
- Plot the curve of average silhouette score according to the number of clusters k.
- The location of the maximum is considered as the appropriate number of clusters.

However, based on application for which the business is targeting, analyst can propose their own k value with some changes to above techniques.

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

Scaling/Standardization of data is important for 2 reasons in K-Means algorithm.

a) 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.

b) The different attributes will have the measures in different units. Thus, standardisation helps in making the attributes unit-free and uniform.

e) Explain the different linkages used in Hierarchical Clustering.

Linkage is the measure of dissimilarity or similarity between the clusters having multiple observations. There are 3 most common type of linkages.

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.