Q. 1 Assignment Summary

HELP International humanitarian NGO, committed to fight poverty and provide the people of backward countries with basic amenities and relief during the time of disasters and natural calamities.

Problem Statement:

During the recent funding programmes, we have been able to raise around $10 million. As an analyst, we have to come up with the countries list that are in dire need of aid on the basis of different socio-economic factors.

Solution Methodology:

1. Data Collection and cleaning
    a. Import the data
    b. Identifying the data quality issues and clean the data
2. Outlier analysis and removal
    a. Removing the outlier where ever required as per understanding the problem statement.
3. Visualizing the data
    a. Visualizing few original data variables to look for any pattern or correlation.
4. Scaling the data
    a. Standardizing all the continuous variables.
5. Hopkins Statistics
    a. To check if data has tendency to form clusters
6. K means clustering
    a. Identify the 'k' by silhouette analysis and sum of squared distances graph.
    b. Visualizing the clusters with various variables
    c. Analyzing the clusters
    d. Identifying the countries which requires aid.
7. Hierarchical Clustering
    a. Identify the 'n' via dendrogram.
    b. Visualizing the clusters with various variables
    c. Analyzing the clusters
    d. Identifying the countries which requires aid.
8. Decision Making
    a. Identifying the countries which requires aid by analyzing both K-means and Hierarchical Clustering results.

Exploratory Data Analysis:

We loaded the data, checked the null values and found no null values in the data. Then we checked for the outliers and found many columns had outliers, we handled the outliers in GDP and others columns outliers already handled. Then we did the univariate analysis for continuous variable and bivariate analysis between the variables. We also performed correlation between the variables and found that (total fertility, child mortality) , (income , gdpp) and (imports and exports) have high correlation.

After performing the standardization on data we implemented K-means and hierarchical clustering. For K-means, we did the silhouette analysis and elbow curve to know the optimum k values, k=3 as clusters. Similarly in hierarchical clustering we perform the single linkage and complete linkage to know the optimum k value, k=4 as clusters.

At the end we plot the scatter plot and barchart to know the clusters having countries in dire need of help. The countries which have low GDP, low income and high child mortality are the countries which need help.

Q. 2 Clustering

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

k-means is a method of cluster analysis using a pre-specified no. of clusters. It requires advance knowledge of 'K'.

Hierarchical clustering also known as hierarchical cluster analysis (HCA) is also a method of cluster analysis which seeks to build a hierarchy of clusters without having a fixed number of clusters.

Main differences between K means and Hierarchical Clustering are:

K-means Clustering	**Hierarchical Clustering**
k-means, 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.
One can use median or mean as a cluster centre to represent each cluster.	Agglomerative methods begin with 'n' clusters and sequentially combine similar clusters until only one cluster is obtained.
K- means clustering 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.
Methods used are normally less computationally intensive and are suited with very large datasets.	Divisive methods work in the opposite direction, beginning with one cluster that includes all the records and Hierarchical methods are especially useful when the target is to arrange the clusters into a natural hierarchy.

Advantages: 1. Convergence is guaranteed. 2. Specialized to clusters of different sizes and shapes.	**Advantages:** 1 .Ease of handling of any forms of similarity or distance. 2. Consequently, applicability to any attributes types.
Disadvantages: 1. K-Value is difficult to predict 2. Didn’t work well with the global cluster.	**Disadvantage:** 1. Hierarchical clustering requires the computation and storage of an n×n distance matrix. For very large datasets, this can be expensive and slow

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. Here the steps of the algorithm are:

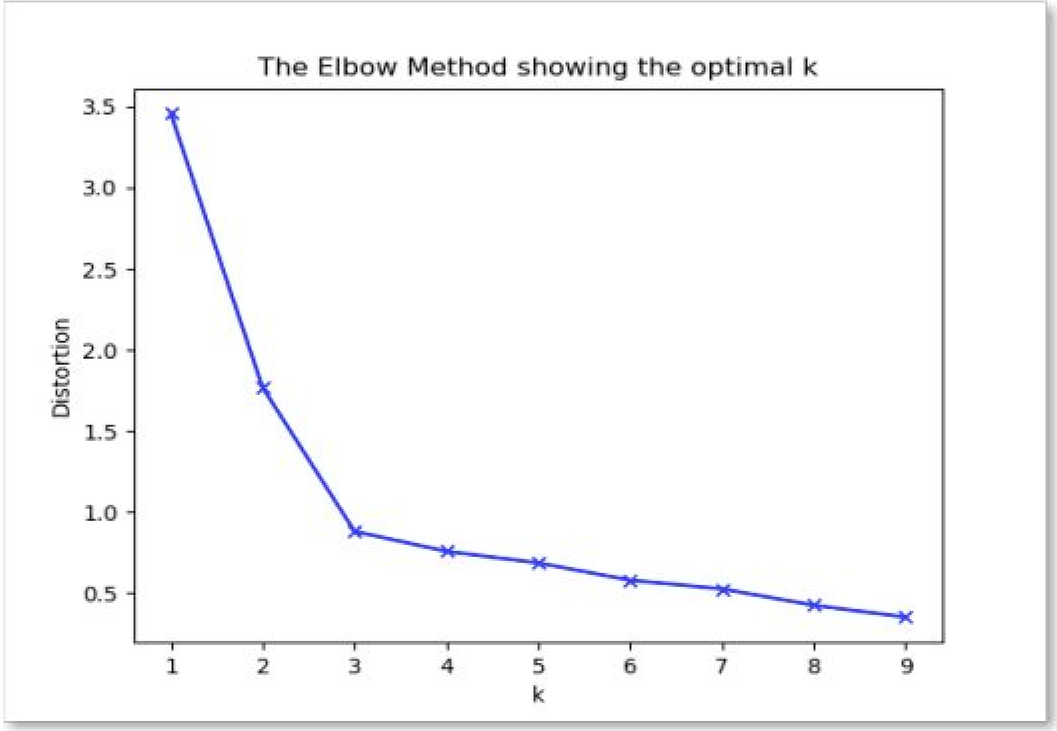
1. Start by choosing K random points as the initial cluster centres.
2. Assign each data point to their nearest cluster centre. The most common way of measuring the distance between the points is the Euclidean distance.
3. For each cluster, compute the new cluster centre which will be the mean of all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

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 a number of pointers that can help us decide the K for our K-means algorithm. Sometimes the business aspect is considered while deciding the clusters and the business team wants certain clusters/segments which may be useful for them to promote the products later. So while deciding the no of segments we always consider business and make sure no of clusters will always answer all business questions.
Statistically, we use two methods to choose the value of K.

1. Elbow method:-

- Compute clustering algorithm (e.g., k-means clustering) 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 (e.g., k-means clustering) for different values of k. For instance, by varying k from 1 to 10 clusters.
- For each k, calculate the average silhouette of observations (avg.sil).
- Plot the curve of avg.sil according to the number of clusters k.
- The location of the maximum is considered as the appropriate number of clusters.

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.

e) Explain the different linkages used in Hierarchical Clustering.

There are three common types of 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.