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 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.

Business Goal:

Our job is to categories 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.

Solution Methodology:

Below are the steps as follow:

Step 1: Reading and Understanding the Data.

- Reading the Data_df file, and inspect the like shape, info, describe, index, etc.

Step 2: Exploratory Data Analysis.

- Clean the data if any missing values.
  - No missing values for the above imported data set.
- Standardising the values.
  - We identified that a few variables are given in percentage format. So, converted those variable values into actual values.
- Handling outliers.
  - We observed that there are outliers for few variables and we do have the flexibility of not removing the outliers, so used capping technique to treat the outliers.
- Univariate analysis.
  - Visualizing single/individual variables.
- Bivariate analysis.
  - Visualizing multiple variables.

Step 3: Hopkins Statistics Test.

- Hopkins Score Calculation
  - We could observe that the average Hopkins value for 10 iterations is 88% which is best identify of cluster tendency.

Step 4: Feature Scaling.

- Standardisation scaling
  - The standardization scaling transforms the data to have a mean of zero and a standard deviation of one.

Step 5: Clustering.

There are 2 clustering techniques:

- K-means Clustering:
  - K-means clustering is one of the simplest and popular unsupervised machine learning algorithms, to achieve the optimal K value.
  - To identify the optimal number of clusters, we have 2 methods.
    - Elbow Curve/Sum of Squared Distance(SSD)
    - Silhouette Analysis.
  - Finally using above optimal number of clusters methods, 3 clusters were formed and countries are split into clusters.
- Hierarchical Clustering:
  - Hierarchical cluster analysis is an algorithm that groups similar objects into groups called clusters.
  - To minimise the pairwise distance between the data points, we use the other method called linkage.
  - There are 2 types of linkages.
    - Single/complete Linkage
      The distance between 2 clusters is defined as the shortest distance between points in the clusters called single linkage and the distance between 2 clusters is defined as the maximum distance between any two points in the clusters is called complete linkage.
  - The best of linkage is complete linkage, because it is the maximum distance between any two points in the clusters.
  - Dendrogram plots.
    - A dendrogram is a diagram that shows the hierarchical relationship between objects.

- Finally by observing dendrogram we choose clusters as 3 and clusters were formed and countries are split into clusters.
- Both clustering techniques give the same list of countries. However, Hierarchical clustering is time consuming and needs a lot of processing power because of we compute the distance of each point from each other point.

Step 6: Final Analysis

- Final Country list Preparation based on high child_mort, low income and low gdpp.
- So, it is highly recommended the below listed countries are required quick aid.
  - Haiti
  - Sierra Leone
  - Chad
  - Central African Republic
  - Mali

Question 2: Clustering

1. Compare and contrast K-means Clustering and Hierarchical Clustering?

K Mean clustering techniques needs a prior knowledge of number of centroid (K) whereas hierarchical cluster do not need this kinds of parameters. Cut_tree () function is used to create the number of clusters of any choice.

Both clustering techniques give the same list of countries. However, Hierarchical clustering is time consuming and needs a lot of processing power because of we compute the distance of each point from each other point.

2. Briefly explain the steps of the K-means clustering algorithm?

K-means clustering is one of the simplest and popular unsupervised machine learning algorithms, to achieve the optimal K value.

Step by step of the K-means clustering algorithm:

- Step-1: Initialize the cluster centres.
- Step-2: Assign observations to the closest cluster centres.
- Step-3: Revise cluster centres as mean of assigned observations.
- Step-4: Repeat step 2 and step 3 until convergence.

To identify the optimal number of clusters, we have 2 methods.

1. Elbow Curve/Sum of Squared Distance(SSD)
2. Silhouette Analysis.

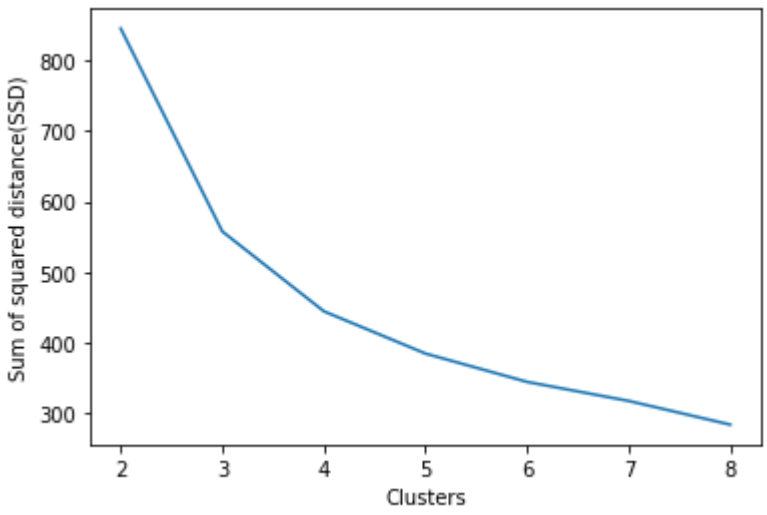
3. How is the value of 'k' chosen in K-means clustering? Explain both the statistical as well as the business aspect of it.

K-means is a simple unsupervised machine learning algorithm that groups a dataset into a user specified number of clusters. The algorithm is somewhat naive--it clusters the data into k clusters, even if k is not the right number of clusters to use. Therefore, when using k-means clustering, users need some way to determine whether they are using the right number of clusters.

One method to validate the number of clusters is the elbow method. The value of k at which improvement in distortion declines the most is called the elbow, at which we should stop dividing the data into further clusters.

Elbow method is to determine the optimal number of clusters for k-means clustering. The idea of the elbow method is to run k-means clustering on the dataset for a range of values of K.

Below is example of Elbow Curve/Sum of Squared Distance (SSD) graph.



4. Explain the necessity for scaling/standardisation before performing Clustering?

When we standardize the data prior to performing cluster analysis, the clusters change.

We find that with more equal scales, and variable more significantly contributes to defining the clusters.

Standardization prevents variables with larger scales from dominating how clusters are defined. It allows all variables to be considered by the algorithm with equal importance.

5. Explain the different linkages used in Hierarchical Clustering?

There are 3 types of linkages.

1. Single Linkage

The distance between 2 clusters is defined as the shortest distance between points in the clusters are called single linkage.

2. Complete Linkage

The distance between 2 clusters is defined as the maximum distance between any two points in the clusters are called complete linkage.

3. Average Linkage

The distance between 2 clusters is defined as the average distance between every points of one cluster to every other point other clusters are called complete linkage.

**********The END**********