HELP NGO

1) Assignment Summary

Problem Statement

HELP has raised $10 million through various funding programs. They plan to use this fund strategically as well as effectively to help the countries that are in dire need of help.

Please help the NGO in choosing the top 5 countries that they need to focus the most taking into consideration the various socio-economic and health factors of the respective countries from the provided data.

Approach to the solution

1) Data Reading and Understanding
   - Data Shape, Data type and Numeric characteristics of the data were checked
   - Converting the features (exports, imports and health) into their actual values which were actually presented as the percentage value of other feature (GDPP)
   - Missing values / Null value check

2) Data Visualization.
   - All the numeric variables were visualized using bar graph with respect to the countries so to check which top 5 countries might need help based on a individual factor
   - Univariate Analysis was performed in order to check the data pattern
   - Bivariate Analysis was performed in order to check the relationship among the variables

3) Outlier Check and treatment
   We used box plot to determine whether there were any outliers among the variables due to business needs the outliers were not removed instead soft capping was performed on the features (Export, imports, health, income, inflation, gdpp and total fertility).
   However higher outlier in child mortality and lower outlier in life expectancy are not capped as these countries may miss out on help.

4) Hopkins check and Scaling
   The clustering tendency was determined through Hopkins check where if the Hopkins value < 0.5 the clustering tendency of dataset is bad if > 0.75 the clustering tendency of dataset is bad.
   Since the Hopkins value for the dataset > 0.9 we proceeded with the clustering process

   Since the variables were of different unit so we use the scaling process to make sure the data is helpful in further analysis. We use standard scaler in our analysis where it scales features where it has a distribution of mean 0 and variance of 1

5) Clustering.

KMeans Clustering.

We performed silhouette score and the elbow curve to determine the optimal value of K to perform the Kmeans clustering. The 3 clusters were formed and visualized in order to determine the categories of the clusters. Cluster profiling was performed with three mentioned features (Child mortality, income and gdpp)

Hierarchical Clustering

Single linkage and complete linkage was performed and visualized using dendrograms .We did select complete linkage since it provided a more structured approach. Though the clusters were formed we were not able to categorize them in a distinguished way so the profiling was done in order to gain more insights

6) Conclusion

Through clustering we were able to find the top 3 countries that were in need of dire help where through Kmeans the top 5 in cluster label 2 and in Hierarchical the top 5 in cluster label 0

2) Clustering

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

K Means clustering

- We need to predefine the value of k i.e. the number of clusters you want to divide the dataset into.
- We can use mean or median to represent center of each cluster.
- The results may differ as we start with the random choice of clusters

Hierarchical Clustering

- We can find the number of clusters at any point by interpreting the dendrogram
- In the beginning n clusters will be formed and naturally the cluster will be similar features will be merged until we have single cluster in agglomerative method
- The results are reproducible since at the end we form a single cluster

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

In Kmeans algorithm we divide the N data points into K clusters, the steps are.

- We start by choosing the K random points which are considered to be cluster centers
- We assign data point to its nearest cluster center and check the distance from the data point to the cluster center euclidean distance is used
- Each time we adjust the cluster center which will be the mean of all the cluster members
- The data points will be reassigned to the clusters based on the new cluster centers
- The 3rd and 4th step will be repeated until there is no changes in the cluster centers/No changes is 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.

We determine the value of k from the below two methods

Elbow Curve Method.

- We determine the value of k by performing clustering algorithm in various values k (e.g. from 2-10)
- We calculate within cluster square of sum (wss) for each value of k
- We plot a wss curve according to the number of clusters of k
- The point of bend in the plot it is the it indicates the appropriate number of clusters

Average Silhouette Method

- We determine the value of k by performing clustering algorithm in various values k (e.g. from 2-10)
- We calculate average silhouette scores for each k
- We plot a avg silhouette score curve according to the number of clusters of k
- The location of maximum is considered to be the value of k

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

We perform scaling in order to standardize the independent data features. It is necessary to perform scaling else the algorithm tends to consider the greater values, higher and smaller values, lower regardless of the unit of the value.

e) Explain the different linkages used in Hierarchical Clustering.

There are 3 types of linkages

- Single Linkage : The single linkage can be defined as the shortest distance between two data points which are from two clusters
- Complete Linkage : It can be defined as the farthest distance between two data points which are from two clusters
- Average linkage : It is defined as the average distance between every point of each cluster with every point of other clusters