Clustering Countries Assignment

Question 1: Assignment Summary

Problem Statement:

HELP International is an international humanitarian NGO. It has raised funds through fund raising program. The funds have to be utilised to aid the countries that are lacking basic socio economic infrastructure.

Objectives:

Main objective is to cluster the countries. The clustering will be done based on the socio economic state of the countries. 5 countries will be shortlisted who will be given aid.

- To begin with the data set provided exploratory data analysis was done. EDA revealed that there are 167 countries in the list. There are no null records. The data frame has 9 columns. The country name is the only object data type. The rest of the columns are numerical.

Univariate analysis

- The columns exports,imports and health are converted and represented as % of gdpp. The univariate analysis on the data set reveals that there are outliers.

Outlier Analysis:

- Columns child_mort, exports, imports, gdp, and total_fer have outliers. The univariate analysis was done to inspect each of the columns. The columns income,imports exports and gdp have outliers that are greater than Q3. They are soft capped instead of deleting them as we have very few records. Deleting them will result in a loss of data. Child mortality is very high in a few countries; it cannot be considered as outliers since they may be countries that need the funds.
- Life expectancy has outliers below the 25 percentile. These may be countries which need help to better their healthcare facilities. Hence they are also not removed.

Scaling

The columns have verif different numerical ranges hence a standard scaler is used to scale the numerical columns.

Hopkins Test

Hopkins test gives a score of 87% indicating clustering can be done for the data set.

Modelling

The Kmeans and Hierarchial methods of clustering are used.

Choosing number of clusters

To get to know the optimal number of clusters elbow curve method and silhouette methods are used.

Elbow curve method

The elbow curve method indicates an elbow at 3 clusters hence 3 or 4 clusters are chosen as optimal.

Silhouette method

The silhouette method of choosing the number of clusters reveals that the score reduces with 6 clusters hence reassuring that 3 or 4 clusters is the most appropriate number of clusters.

Now the models are built.
First KMeans and Hierarchical methods are carried out with 3 clusters. The results are analysed and then both methods are carried out with 4 clusters.

Kmeans with 3 clusters

The resulting clusters are analysed. When the boxplots are drawn for the three clusters for gdpp, child mortality and income, there is a slight overlap of clusters 1 and 2. and the whiskers of the boxplots of cluster 0 and 2. The countries in cluster 0 have a high child mortality rate, low gdpp and income hence will be a candidate for aid.

Hierarchical method with 3 clusters

The resulting clusters are analysed. When the boxplots are drawn for the three clusters for gdpp, child mortality and income, there is a slight overlap of clusters The countries in cluster 2 have a high child mortality rate, low gdpp and income hence will be a candidate for aid. The drawback was that there is only 1 country in this cluster hence the next cluster that was considered had 125 countries. This number was too large considering ther are only 167 countries in total hence Kmeans with4 clusters were also tried as below.

The scatter plots were drawn for both hierarchical and Kmeans and Kmeans show a better classification.

Kmeans with 4 clusters

The number of clusters are then chosen as 4 and the model is built and the resulting clusters are analysed. When the boxplots are drawn for the three clusters for gdpp, child mortality and income, there is a clearer clustering now .

Hierarchical method with 4 clusters

The resulting clusters are analysed. When the boxplots are drawn for the three clusters for gdpp, child mortality and income, there is a slight overlap of clusters The countries in cluster 2 have a high child mortality rate, low gdpp and income hence will be a candidate for aid. The drawback was that there is only 1 country in this cluster hence the next cluster that was considered had 125 countries. This number was too large considering there are only 167 countries in total.
The scatter plots were drawn for both hierarchical and Kmeans and Kmeans show a better classification.

The final cluster that was chosen was from Kmeans method with 4 clusters as that was best cluster separation as seen in the scatter plots.
The countries were then sorted on basis of high child mortality , low gdpp and low income.
The final list of 5 countries were deduced as
1. Haiti,
2. Sierra Leone
3. Chad
4. Central African Republic
5. Mali

Question 2: Clustering

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

K Means	Hierarchical
Using a predefined number of clusters assigns data points to clusters based on distances to the centroids.	Either assigns all data points to separate clusters and then joins them based on similarity(Agglomerative) or puts all data points in one cluster and divides as it proceeds(Divisive).
Iterative method	Hierarchical
Can handle big data since computation complexity is O(n)	Cant handle big data since computation complexity is O(n squared)
Results are dependent on the number of clusters that are chosen.	Results are reproducible.
The result of K-means is unstructured	Result is more interpretable and informative

Useful when clusters are intuitively known	Useful when naturally can cluster data but no hint of the number of clusters.
The cluster number is chosen before clustering	The number of clusters can be chosen after looking at the dendrogram

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

K means is an iterative clustering algorithm.

The algorithm works in 5 steps.

1. Chose the number of clusters K
    a. In this step a number is chosen either randomly or with some knowledge about the case study or by using statistical methods.
2. Randomly choose K centroids
    a. Randomly choose K centroids or K points from the data set .
    b. Assign each data point to a cluster based on distance to these centroids.
        i. Calculating distance from the centroid to each data point the data points are assigned to each cluster.
    c. Different measures of distances are used
        i. Euclidean Distance
        ii. Squared euclidean distance
        iii. Manhattan distance
        iv. Cosine distance
3. Cluster centroids are calculated now based on the clusters.
    a. reposition the centroid based on the cluster formed
4. Re-assign each point in the data set to the new closest cluster centroid.
5. Repeat steps 3 and 4 until centroids are not changing any more.

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

The statistical methods to choose the value of k are

1. Elbow Method.
2. Silhouette Method.

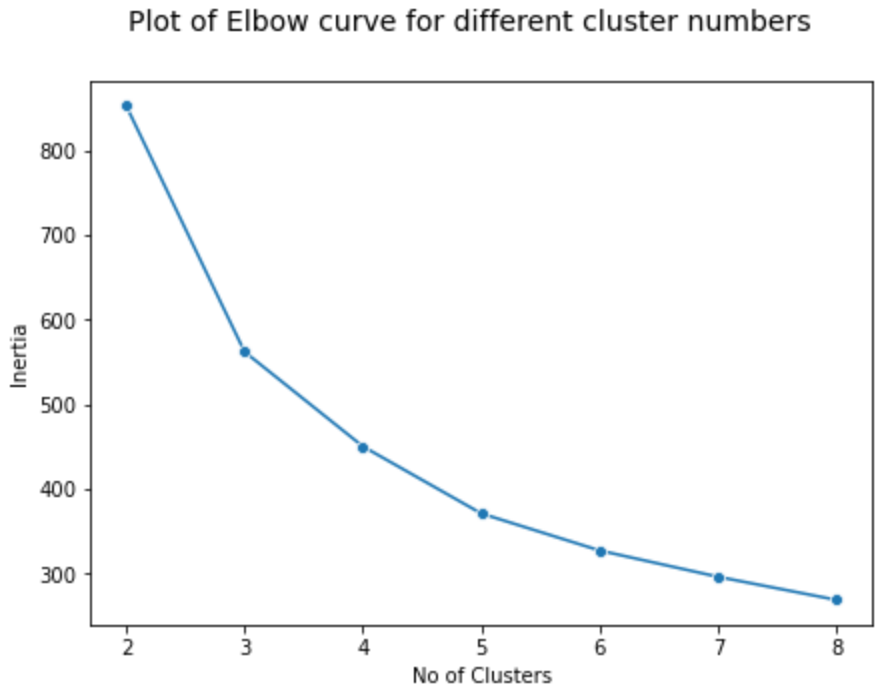
Elbow Method

Elbow method is used to select the right value of k.

For a range of values for k the elbow method runs k-means clustering on the dataset. For each value of k the algorithm computes an average score for all clusters formed with that value of k. By default, the score is computed as the distortion score. Which is the sum of square distances from each point to its assigned center.

Now the number of clusters k and the sum of square distances or any other distortion score is plotted. An "elbow" is visible in the plot as shown below. The value of distortion reduces sharply at the elbow and there after the distortion remains almost the same or keeps decreasing. Since increasing the cluster number will reduce the distance of the points to the centroid. Hence the value of k has to be chosen such that the distortion is not too high and the cluster number satisfies the purpose of the case study.

For the graph below the elbow is formed at k=3.



Silhouette Score

The idea behind this score is that the clusters can be classified as good if the inter cluster distances are minimum and the intra cluster distances have to be maximum.
That is the data points within a cluster have to be alike and the cluster in two different clusters must be very different.

$$\text{silhouette score} = \frac{p - q}{max(p, q)}$$

p is the average distance to the points in the nearest cluster that the data point is not a part of, q is the average intra-cluster distance to all the points in its own cluster.

- The value of the silhouette score range lies between -1 to 1.
- A score closer to 1 indicates that the data point is very similar to other data points in the cluster,
- A score closer to -1 indicates that the data point is not similar to the data points in its cluster.

Based on the silhouette score the number of clusters are chosen.

From the business application wise we need to segment the data. It may be segmenting people, products, markets etc. The technique used to segment is called the clustering. The main goal of segmenting is to group the data points into buckets that can be addressed to as a unit. For example with consumer data to address each customer is difficult now with the help of clustering techniques the customers can be segmented and each segment can be addressed individually. The points to consider while clustering keeping in mind the business aspect is that

- the segments have to be stable.
    - That is when a clustering technique is done and the consumers are segmented. These segments should not change when the technique is repeated within a short period of time. The clustering techniques hence should be mentioned with a time bound limit until which they are applicable.
- Inter and Intra Cluster homogeneity
    - The data points within a cluster should be very alike, that is the inter cluster homogeneity should be high. Example customers within a segment should be alike.
    - The data points from different clusters have to be very different, that is the intra cluster heterogeneity has to be high. For the customers between two segments they have to be highly different.

Clustering in a business perspective depends on the business also for example while trying to segment customers they can be segmented based on

- Behaviour
- Attitude
- Demographic

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

Scaling is a necessity before performing clustering when the features are of different magnitudes like for example grouping customers based on money spent on an online store and the number of items purchased. Since the numerical values are not on the same scale the

standardization helps in comparing them on a standardised scale and the clustering gives better results.
The clusters without standardization will be dominated with features that have higher numerical values. Standardization helps to make the relative weight of each variable equal by converting each variable to a unitless measure or relative distance.

e) Explain the different linkages used in Hierarchical Clustering.

The distances used to compare clusters for merging or splitting are often called Linkage Methods. Some of the common linkage methods are:

Complete-linkage: the distance between two clusters is defined as the *longest* distance between two points in each cluster.

Single-linkage: the distance between two clusters is defined as the *shortest* distance between two points in each cluster.

Average-linkage: the distance between two clusters is defined as the average distance between each point in one cluster to every point in the other cluster.

Centroid-linkage: finds the centroid of cluster 1 and centroid of cluster 2, and then calculates the distance between the two before merging.