Assignment Part 2

Question 1: Assignment Summary

Ans:

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. They have raised 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.

We need to analyze the data and cluster the countries based on -some socio-economic and health factors that determine the overall development of the country. -provide the list of top 5 countries to which the financial aid should be given.

Steps followed for solving this assignment:

- **Data Reading and Understanding** –
    - This included reading the dataset, inspecting for the shape of data etc and checks for missing values, null values, duplicates etc. This data set was clean and did not have any such cases.
    - Columns like imports,exports and health were expressed in terms of percentages so they were converted to absolute values.
- **EDA:**
    - In EDA, we have plotted a heatmap to check the correlation of all columns and find out which ones are highly corelated and in the positively and negatively.
    - Then we have plotted a distribution plot to understand for various numerical columns like gdpp, income, inflation, child mortality rate etc to see how the data is distributed
    - We have also plotted graphs of country v/s cols like child mortality, income, health and life expectancy to understand the trend. This covers the bivariate analysis of categorical and continuous type
    - Scatter plots have been plotted for income vs child mortality and health. This tells us if there is any relationship between income and child mortality and health. This is bivariate continuous type of analysis.
    - Then we have checked for outliers in the various columns and capped the higher range values at a 99 percentile in order to not loose valuable information for our analysis. Also the lower range outliers are untouched as they are not really outliers but actual values which would facilitate formation of good clusters

3) Data Preparation for Building The model:

- Calculated the Hopkins statistic, we got a high value of around 0.91 which indicates that we can expect good clusters from our data

- Then we did Data Scaling using the standard scaling technique on the numerical columns

4) Building the model using K means:

- For K means , we need to pre determine the value of k , this was done by calculating the silhouette score and elbow curve . After checking the value , k =3 was a suitable value to proceed , so k means clustering was done using k=3 and the clusters formed were inspected based on socio economic factors like gdpp, income and child mortality. Below categories could be derived from the clusters formed. 48 countries were a part of Underdeveloped Countries cluster
- - Underdeveloped (Cluster with low gdpp and low income having high mortality rate)
- - Developing (Cluster with moderate gdpp and moderate income and moderate mortality rate)

- - Developed (Cluster with high gdpp and high income and low mortality rate)

- So based on this , the top 10 countries were found which had high mortality rate and low gdppa and low income - 'Haiti' 'Sierra Leone' 'Chad' 'Central African Republic' 'Mali' 'Nigeria' 'Niger' 'Angola' 'Congo, Dem. Rep.' 'Burkina Faso'

5) Building the model using Hierarchical clustering:

- For this , we do not have to determine value of k before hand. We started with single linkage method , but the dendrogram did not seem to give any meaningful results so we went ahead to check the complete linkage. In complete linkage we could easily spot clusters from the dendrogram which were 4,3,2. 3 seemed to be a meaningful value so we chose 3 and formed the clusters. The clusters formed here also derived the same categories as we saw in K means based on gdpp, mortality rate and income. Here also we could see 48 countries were a part of Underdeveloped Countries cluster.

- So based on this , the top 10 countries were found which had high mortality rate and low gdppa and low income - 'Haiti' 'Sierra Leone' 'Chad' 'Central African Republic' 'Mali' 'Nigeria' 'Niger' 'Angola' 'Congo, Dem. Rep.' 'Burkina Faso'

6) Conclusion:

we could see that both the clustering techniques gave same results. So based on the dataset and business problem, clusters would change depending on how we treat the outliers and what columns are chosen for analysis. So it is better to use both techniques and check the types of clusters we are getting before making any recommendations to business partners

Question 2- Clustering

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

Ans:

K means	Hierarchical
You need to know the number of clusters in advance	In hierarchical clustering one can stop at any number of clusters, one find appropriate by interpreting the dendrogram.
It can handle high volume of data. The time complexity is O(n)	It is not suitable for high volume of data. The time complexity is $O(n^2)$.
In K Means clustering, since one start with random choice of clusters, the results produced by running the algorithm many times may differ.	In Hierarchical Clustering, results are reproducible
K- means clustering a 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
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.

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

Ans:

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 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 clustermembers.
4. Now re-assign all the data points to the diffrent clusters by taking into account the new cluster centres.
5. Keep iterating through the step 3 & 4 until there are no further changes possible. At this point, you arrive at the optimal clusters.

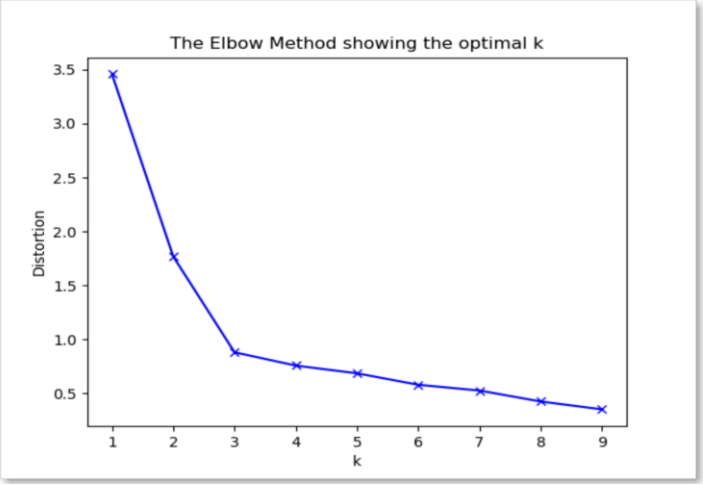
c) How is the value of 'k' chosen in K-means clustering? Explain both the statistical as well as the business aspect of it.

Ans:

There are a number of pointers that can help us decide the K for our K-means algorithm:- 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.

From a business perspective, the number of clusters would depend on the domain of application. For example , for the clustering countries dataset , the value of k =3 seemed ideal as this ensured all categories of clustering were covered. So it would largely depend on the dataset and industry in which the clustering is to be performed.

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

Ans:

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.

Ans: 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.