Assignment: Part II

Question 1: Assignment Summary

Briefly describe the "Clustering of Countries" assignment that you just completed within 200-300 words. Mention the problem statement and the solution methodology that you followed to arrive at the final list of countries. Explain your main choices briefly (what EDA you performed, which type of Clustering produced a better result and so on)

Answer:

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 programmes, 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.

Objective:

The requisite is:

- To categorise the countries using some socio-economic and health factors that determine the overall development of the country.
- To suggest the countries which the CEO needs to focus on the most.

Method followed:

Data Processing:

- No null values were found
- No duplicate values for country
- Only few outliers and they were handled later on during PCA
- The data was standardized for Principal Component Analysis (PCA)

Screeplot: To get a 95% variance of data, 4 components are good enough. Hence, PC is selected to be 4.

Clustering:

- Both the methods K means and Hierarchical Clustering was used on the 4 PCA components
- For K means, K= 3 was taken using the elbow dip and silhouette analysis.
- While doing the Hopkins Statistics a value of 0.77 was attained.
- If the Hopkins Statistics values are:

  0.3 -- Low chase of clustering
  Around 0.5 – Random
  Between 0.7 to 0.99 -- High chance of clustering
- Finally using all these values clusters of 3 were formed and the countries are split into clusters.

Question 2: Clustering

Compare and contrast K-means Clustering and Hierarchical Clustering.

Answer:

K-mean Clustering	Hierarchical Clustering
We need to have desired number of clusters ahead of time.	We can decide the number of clusters after completion of plotting dendrogram by cutting the dendrogram at different heights
It is a collection of data points in one cluster which are similar between them and not similar data points belongs to another cluster.	Clusters have tree like structures and most similar clusters are first combine which continues until we reach a single branch.
Works very good in large dataset	Works well in small dataset and not good with large dataset
The main drawback of k-Means is it doesn't evaluate properly outliers.	Outliers are properly explained in hierarchical clustering
K-means only used for numerical.	Hierarchical clustering is used when we have variety of data as it doesn't require to calculate any distance.

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

Answer:

Step 1: Randomly select K points as initial centroids.

Step 2: All the data points closet to the centroid will create cluster centre according to Euclidean distance function.

Step 3: Once we assign all the points to each of k clusters, we need to update the cluster canters or centroid of that cluster created.

Step 4: Repeat 2,3 steps until cluster canters reach convergence.

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

Answer:

‘K’ value is chosen randomly in K-Means clustering based on statistical aspect. From business aspect, we need to first understand the dataset and based on that we decide number of ‘k’. for example, we have a dataset of variables like ‘pen’, ‘pencil’, ‘books’, ‘notebooks’, ‘mobiles’, ‘charger’, ‘laptop’. Now if we want to have k values based on statistical aspect, we can use silhouette score to determine that but based on business aspect, after viewing the dataset we can easily make cluster = 2, one in electronics category and another non-electronics.

Explain the necessity for scaling/standardisation before performing Clustering.

Answer:

It is necessary to do scaling/standardisation because our variables may have units at different scale and as our method stresses more on calculation of direction of space or distance, so if we have one variable with high scale units then while calculating for k-Means or hierarchical it will create a big difference as the clusters will tend to move with the variables having greater values or variances. By applying standardisation/scaling will increase the performance of our model.

Explain the different linkages used in Hierarchical Clustering.

Answer:

Linkage is a technique used in Agglomerative Clustering.

Linkage helps us to merge two data points into one using below linkage technique.

- **Single linkage**: The distance between two clusters is calculated by the minimum distance between two points from each cluster.
- **Complete linkage:** The distance between two clusters is calculated by the maximum distance between two points from each cluster.
- **Average linkage**: The distance between two clusters is the average distance between every point of one cluster to the another every point of other cluster.
- **Ward linkage:** The distance between clusters is calculated by the sum of squared differences with all clusters.