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).

Note: You don't have to include any images, equations or graphs for this question. Just text should be enough.

Solution:

The main objective of the case study is to find countries that are in direst need for financial aid. I analysed the socio-economic and health factors of all the countries and divided them into clusters.

Overall, the data was passed through basic EDA such as, checking the missing values, data transformation, outliers checking, checking of correlation and multicollinearity. The dimensions of the dataset were standardised using standard scaler and then PCA was applied to correct the multicollinearity. Using this method, 9 PCA components were obtained (0-8).

After this, x-y axes plot of first two principal components is plotted to understand the feature sense. Using scatter plot, it was found that features like GDP, life expectancy are in the direction of 5^{th} component of first principal component.

To select the optimum number of structures, elbow curve was used and found that 5 clusters will yield the best results with more than 95% of variance.

In order to check if outliers have the countries with direst need for financial aid, 2 models were developed – one with the outliers and the other without it.

Now, 2 method were used to identify the list of countries namely k-means and hierarchical clustering. It was observed that, there were differences in countries obtained when clustering was done by k-means and hierarchical clustering. Though the k-means output for less records hierarchical clustering has higher advantage than k-means but k-means is preferred as it provides precise information as per the business requirement.

These methods were repeated on the data without outliers and it was observed that countries with low GDP and low income were neglected. Therefore, data with the outliers is preferred. Hence, from K-means method with outliers, 47 countries are selected for the financial aid.

Question 2: Clustering

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

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

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

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

e) Explain the different linkages used in Hierarchical Clustering.

a)

I. K-means Clustering:
- We need to decide desired number of clusters before the process.
- Works well with large number of datasets.
- K-means only used for numerical values.
- Creates clusters with inter-homogeneity and intra-heterogeneity.
- It does not evaluate outliers properly.

II. Hierarchical Clustering
- We can decide the number of clusters by cutting the dendrogram at different heights.
- Works well for small datasets.
- Can be used in wide variety of data values.
- Combines similar clusters and creates a tree like structure.
- Outliers are properly explained.

b)

Step 1: Randomly select K points as initial centroids.
Step 2: A cluster is formed with points closest to the selected centroid.
Step 3: After assigning the clusters, update the cluster's centroid having minimum distance from each point.
Step 4: Repeat Step 2 and Step 3 till the centroids of each cluster are stable.

c)

The k value for K-Means clustering is randomly selected when looked through statistical aspect also, silhouette score can also be used. From business point of view, the data is understood to select the 'k' value.

d)

It is advisable to scale/standardise because the variables may have units at different scales. This increases the stress on calculation. For instance, if we have one variable with high scale units then while calculating k-means or hierarchical clustering it will create huge differences as the clusters will tend to have wide variation in the clusters. The performance of the model shall significantly improve by standardisation.

e)

Linkages is a technique used in Agglomerative clustering. It helps to merge two data points of different clusters.

The different types of linkages are:

i) **Single Linkage:** For two clusters R and S, the single linkage returns the minimum distance between two points i and j such that i belongs to R and j belongs to S.

ii) **Complete Linkage:** For two clusters R and S, the single linkage returns the maximum distance between two points i and j such that i belongs to R and j belongs to S.

iii) **Average Linkage:** For two clusters R and S, first for the distance between any data-point i in R and any data-point j in S and then the arithmetic mean of these distances is calculated. Average Linkage returns this value of the arithmetic mean.

iv) **Ward Linkage:** The distances between clusters is calculated by the sum of squared differences with all clusters.