ASSIGNMENT – PART II

ASSIGNMENT SUMMARY

Q: **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)**

A: Our main objective for the assignment is to find the countries which are in dire need of help.

Our job is to find these countries using socio- economic and health factors which represents the overall situation of that country.

The dataset provided has features such as child_mort (child mortality rate), exports and imports (in percentage of gdpp), money spending on health(in percentage of gdpp), income, inflation, life expectancy, total fertility and finally the gdpp (gross domestic product) of 167 different countries. There are no multiple entries in country column and no missing values in dataframe, I transformed some of the variables like health, imports and exports from % of GDP to actual values as it makes more sense to the dataset.

I took child_mort, income, gdpp for cluster profiling and all features are used for clustering. The given data is an unsupervised learning dataset.

Then checked for outliers in each feature then decided to go with the model which suites the Business Problem better. That is, not to treat the Outlier, and check the model with different K values to see which one gives a better business outcome.

Then, did EDA by visualising pair plots. Then scaled the data using Normalization.

Now, Cluster tendency using Hopkins Statistics. The Hopkins statistic, is a statistic which gives a value which indicates the cluster tendency, in other words: how well the data can be clustered.

- If the value is between {0.01, ...,0.3}, the data is regularly spaced.
- If the value is around 0.5, it is random.
- If the value is between {0.7, ..., 0.99}, it has a high tendency to cluster.

Hopkins Statistic got good score of 90%.so it indicates that data is highly clustered. Started doing hierarchical clustering. But didn't get good result.

Then performed k means clustering with k =3 value, got good result. By using descriptive statistics, analysed top 10 underdeveloped countries which are in dire need of the Financial Aid from the NGO.

Q2: 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. Compare and contrast K-means Clustering and Hierarchical Clustering.

K-Means 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
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.

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

Step 1: Randomly select K points as initial centroids.

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

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

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

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

'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 K-Means 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. 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.

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

It's definitely a good idea to do scaling/standardisation because our variables may have values 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.

e) Explain the different linkages used in Hierarchical Clustering.

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.