Clustering Assignment: Part II

Question 1 :
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.

Answer:
Our main objective for this assignment is to find the countries that are in direst need of aid. Our job is to find those countries using socio-economic and heath factors which will show overall development of the country.
So, in order to do so, I followed below steps and got below insights as per the steps:

- Read and Understand Data
- Check for missing value, and their treatment
- Check for outlier and their treatment
- Perform the basic EDA to find the variability and distribution of the data, so as to identify if we need t scaling the data
- Data Scaling if necessary
- Use Hopkins Method to check if the dataset is good enough for a cluster analysis
- Using Hierarchical clustering to identify the optimal cluster value.
- Use Silhouette and Elbow method to validate the optimal cluster values.
- Use K-Means Cluster method to build the final cluster model.
- Analyse the cluster that is representing the countries that will solve the Business Problem.
- Present the final report

And below are the insights summary which I have observed in my analysis.

In this Dataset, I have found 167 rows and 10 columns in total based on various countries and their socio-economic factors, wherein I have checked for missing and duplicate country name but there are no duplicity and missing values in the dataset.

Then I observed that the 'imports', 'exports' and 'health' variable seems to be in percentage of GDP per capita, and this can sometimes give an incorrect insight in our EDA. for example, the health spending of 'United states' is 17.9 and that of 'Sierra Leone' is '13.1', both of which are very close to each other in health spending in terms of their % of GDP per capita. But these figures do not actually tell us the real story of how rich and poor are 'USA' and 'Sierra Leone' is. So, to tackle the problem in the best way I converted % values in absolute values.

Then I did outlier Analysis wherein I found there are outliers in every variable. So as per the business understanding I chosen not to treat outlier and do the analysis with different K values to get the better business outcome.

In EDA, I visualize the data distribution through Pairplot where I found most of the Data points are not normally distributed. Variance were also different. Data indication was to be done data scaling. So I did Standard scaling.

Then I did Hopkins Statistics which is the process to evaluate the data to check if the data is feasible for clustering or not and with this I found good score which is more than 70% so I proceeded further with clustering.

In both the K means, and Hierarchical clustering I found optimal K value = 3 with SSD (elbow curve) analysis and visualization.
With all the visualization and thorough analysis on data I chosen final model with k = 3 and with this I implemented the model on my final data.

Got the final result and displayed top 10 **'Under Developed Countries' .**

Question 2:

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.

Ans. 2. a)

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.

Ans. 2. b)

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.

Ans. 2. c)

'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.

Ans 2. d)

It is definitely a good idea 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.

Ans. 2. e)

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.