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 : The final clustering of countries are

- Haiti
- Sierra Leone
- Chad
- Central African Republic
- Mali

The methodology that was followed to get the result of the countries that needed help are K Means Algorithm and Hierarchical Clustering alogirithm.

The EDA performed during the analysis was

Step1 :Checked if the data has any missing values

Step2:Converted the columns exports,health imports that are in % to actual values

Step3:Performed Univariate and Bi variate analysis to see which all countries have less Income and GDP followed by high child mortality rate

Step4:Checked the correlation of the variables.

Step5:Performed outliers

Step6:Instead of dropping outliers , capping outliers was performed as dropping might loose the data of the countries that are required for further analysis.

After above steps clustering was performed, both K means/Hierarchical clusterial has provided similar results in terms result of top countries that need aid

However based on hopskin score K means shows that the clustering data is good enough to proceed further

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.

Answer:

a)k-means, using a pre-specified number of clusters, the method assigns records to each clusters.

Hierarchical methods can be either divisive or agglomerative.

K Means clustering needed number of clusters one want to divide data.

In hierarchical clustering ,one can stop at any number of clusters, one find appropriate by interpreting the dendrogram.

In K -means , one can use median or mean as a cluster centre to represent each cluster.

In hierarchical clustering ,Agglomerative methods begin with 'n' clusters and sequentially combine similar clusters until only one cluster is obtained

b) In K means , following are the steps that to be performed.

Choose the value of K

Silhouette score

Elbow curve-ssd

Here we need to choose the value of K based on the Silhouette score/ Elbow curve-ssd.

Once the K value is been choosen we need add labels to the data frame .This label slusters the data points based on K value .

Once the above is done , need to perform the clustering profiling as it gives the result of the countries that required aid.

c) K value in K means choosed in 3 .This is choosen based on the Silhouette score , Elbow curve-ssd .K value 2 cannot be choosen as its quite not correct in terms of business aspective too.Choosing more clusters will create the complexity .Hence K =3 is the best fit for clusters in K means algorithm.

d) Scaling/Standardization to be performed before clustering as this allows all the values to fit and transform on the same standards.

e)Simple and Complete linkages are used in Hierachical alogorithms.

Using only simple linkage will be complex to determine , hence comple linkages is used which in turn gives the number of clusters that to be used.