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

"Clustering of Countries" assignment deals with 167 countries and their socio-economic factors. Based on this data set, top 10 countries are to be identified, so that HELP international NGO can provide aid to such countries and help these countries to overcome.

As part of this case study, I have looked at data set, it has 9 factors which needs to be considered while identifying the countries. Few variables like imports, exports, health are given as % of GDP, so first I converted these values in actual based on the % provided.

Looking at data set there are no missing values, but there are lot of outliers in each variable, because we have to find the countries which are in real need of aid, I performed capping of 1-95 percentile for factors which are at higher end like GDP, Income, Import, Exports & Health these are high for good performing countries, so capped the data.

Once capping is done, performed the EDA to look at countries by Child Mortality, Income & GDP to know which are low performing countries and can be right candidate for the aid. Bar Graphs are plotted for top such 20 counrties. Also look at correlation between these factors to know which all impacting others.

Once EDA is performed, Hopkins cluster tendency is checked, to make sure if data is good for clustering, almost got more than 80% score, which is a good indicator that clustering can be better performed on data set.

Performed standard scaling and plotted silhouette score to get the right number of clusters, looking at this plot, we can very well see 3 is the right number of clusters in which these countries can be divided.

Using heretical clustering – complete and single linkage as well performed clustering, plotted dendrogram. Complete linkage dendrogram easily indicates that 3 clusters are optimal where in single linkage shows only 1 cluster, which is kind of unbalanced.

So, went ahead with K means cluster profiling and chosen top 10 countries based on high child mortality rate & low on income and gdpp.

Question 2: Clustering

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

Answer:

In K-means clustering, we should be aware or sure in how many clusters we need to divide our data points where in Hierarchical Clustering with the help of dendrogram we can look at data points and see what the best size of cluster is.

In our case study using value of k=3 , gives almost optimum group of data , data points seems to be well groups and cluster is also looking balanced.

In case of hierarchical clustering when performed single linkage, data points looks to be converged in single cluster, which is not good performer, where-in complete linkage performed well and gave balanced cluster.

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

Answer:

K-means clustering aims to find the set of k clusters such that every data point is assigned to the closest center, and the sum of the distances of all such assignments is minimized. Below are steps to be performed as part of k-means algorithm

1. *Initialize cluster centers*
2. *Assign observations to the closest cluster center*
3. *Revise cluster centers as mean of assigned observations*
4. *Repeat step 2 and step 3 until 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.

Answer:

Choice of k value purely depends on business needs and problem we are trying to solve. Anything above 2 or less than 15 can be choice of k in k-means clustering. Many clusters may not make sense and difficult to analyze. While choosing the right value of k , we should discuss with business and finalize how we want to cluster our data.

Statistically, we can plot silhouette score or elbow curve to get the optimal number of clusters.

c) Explain the necessity for scaling/standardization before performing Clustering.

Answer:

Scaling is really important in model building, since each variable is having different values and sometime their unit of measurement is also different, so we need to make sure all the variables which are to be used for cluster building are at same scale, one of most commonly used scaling method is standard scaling in which each data point is standardize, its basically normal distribution of data .

d) Explain the different linkages used in Hierarchical Clustering.

Answer:

Hierarchical clustering can be divided into two main types: agglomerative and divisive.

- Agglomerative clustering: It works in a bottom-up manner
- Divisive hierarchical clustering: It works in a top-down manner.

There are different ways to measure the distance between clusters to decide the rules for clustering, and they are often called Linkage Methods.

Complete linkage: the distance between two clusters is defined as the longest distance between two points in each cluster.

Single linkage: the distance between two clusters is defined as the shortest distance between two points in each cluster. This linkage may be used to detect high values in your dataset which may be outliers as they will be merged at the end.