Clustering Assignment

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).

Ans-) HELP International is an international humanitarian NGO that is committed to fighting poverty and providing the people of backward countries with basic amenities and relief during the time of disasters and natural calamities.

Problem Statement - It runs a lot of operational projects from time to time along with advocacy drives to raise awareness as well as for funding purposes. After the recent funding programmes, they have been able to raise around $ 10 million. Now the CEO of the NGO needs to decide how to use this money strategically and effectively.

Objective of this is to categorise the countries which are in direct need of aid by considering socio – economic factor into consideration.

Solution Methodology - Firstly, we read , clean and prepare data for EDA.

Secondly, we performed and visualized EDA with lowest 10 countries for each category.

Thirdly , we did outlier analysis, scaling of data and Hopkins statistics.

Fourthly, we did silhoutte analysis and elbow curve and perform K-Means using k value.

Fifth - Using K-means clustering , found 5 countries which has lowet gdpp , lowet income and highest child_mort which requires more concentrations.

Sixth - Hirerchical clustering - Indentify n using dendograms, formed cluster , visualize and analysed the same and identified 5 countries which need aid.

Results found from both hirerchical and K-Means clustering are used for decision making.

Question 2: Clustering

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

Ans**) K-Means Clustering** - 1. Identify k-values using silhoutte analysis. 2. Form the cluster and visualize the same. 3. Analyze the cluster. 4. identify countries which requires aid.

As per K-Means Clustering, the countries which require aid are:

Burundi ,Liberia, Madagascar, Central African Republic,Guinea-Bissau.

Hierarchical Clustering - 1. Identify n-values using dendograms. 2. Form the cluster and visualize the same. 3. Analyze the cluster. 4. identify countries which requires aid.

As per Heirarchical Clustering, the countries which require aid are:

Burundi,Liberia,Congo, Dem. Rep.,Sierra Leone,Madagascar.

---

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

Ans) Steps are **-**

1) Randomly select 'c' cluster centers.

2) Calculate the distance between each data point and cluster centers.

3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers..

4) Recalculate the new cluster center and represents the number of data points in ith cluster.

5) Recalculate the distance between each data point and new obtained cluster centers.

6) If no data point was reassigned then stop, otherwise repeat from step 3).

---

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

Ans) There is a popular method known as elbow method which is used to determine the optimal value of K to perform the K-Means Clustering Algorithm. The basic idea behind this method is that it plots the various values of cost with changing k. As the value of K increases, there will be fewer elements in the cluster. So average distortion will decrease. The lesser number of elements means closer to the centroid. So, the point where this distortion declines the most is the elbow point.

---

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

Ans) Standardization is the central preprocessing step in data mining, to standardize values of features or attributes from different dynamic range into a specific range. When we standardize the data prior to

performing cluster analysis, the clusters change. We find that with more equal scales, the Percent Native American variable more significantly contributes to defining the clusters.Standardization prevents variables with larger scales from dominating how clusters are defined. It allows all variables to be considered by the algorithm with equal importance.

---

e**) Explain the different linkages used in Hierarchical Clustering.**

Ans) Different linkage are -

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.

Thank You!