Clustering Assignment

Part-II

Q-1 Assignment Summary: Problem Statement

After the recent funding programs, 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. The significant issues that come while making this decision are mostly related to choosing the countries that are in the direst need of aid. And this is where you come in as a data analyst. Your job is to categorize the countries using some socio-economic and health factors that determine the overall development of the country. I need to suggest the countries which the CEO needs to focus on the most.

A-1

1. I started my analysis of the data set by importing the relevant libraries such as pandas, NumPy and matplotlib, and seaborn. Once, the data set was there I analyzed all the data types and the no. of columns present in the data set.

2. Then the data set was observed to check whether we have any irrelevant values present in our data set so, that I can remove them from there. But the analyses gave me the result that there are no such values that need to be removed from the data set. **There were no null values present in the data**.

3. Data set was prepared for modeling in which any **I selected to remove the values where the values were less than the 5th percentile and greater than the 95th percentile to remove the outliers from the data set**. Rescaling was done on certain columns to bring them a readable range.

4. Modelling Process was carried out on the data set in which for selecting the K-value we had 2 methods one is the **SSD Method** and the other is the **Silhouette Analysis Method**. By obtaining the elbow curve obtained from the SSD method I compared it with the Silhouette score obtained for each cluster. **I found out that the best-suited value for K is 3.So, instead of beginning with 4 clusters my model would be best suited with 3 clusters.**

5. Then I used another clustering method such as the **Hierarchical Clustering Algorithm**. In which the approach is to take all the points which are a close distance from each other and form a cluster with them. In this, I presented both single as well as complete linkage but got a better a much readable plot "**Dendrogram**" in the complete linkage. **Now here I cut 3 clusters and then got the rest of the analysis to be taken forward**.

6. In **K-Means** as per the cluster Id's got **48 countries** in that and around 126 **countries** in the **Hierarchical Clustering Algorithm**.

7. I went ahead with K-Means Algorithm as I found it to be more simple and reasonable without any complexity involved. **The no. of clusters that I got (K=3) at the end of my analysis was almost near to the no. of clusters that I started with (K=4). There was less computation involved in this method and was suited for this data set.**

Q-2 Clustering

A-2

1. Compare and contrast K-means Clustering and Hierarchical Clustering?

Ans-1

S. No.	K Means	Hierarchical Clustering Algorithm
1.	No. of clusters in this clustering are pre-decided.	Here the No. of clusters are not pre-decided.
2.	Elbow Curve and Silhouette Score helps us in finding out the No. of clusters	No. of clusters here are decided based on the cut that we make on the Dendrogram.
3.	Only, the centroid is used to find out the clusters.	Here we follow either the bottom-up approach or the top-down approach.
4.	Python libraries that are used here are sklearn-Kmeans	Here we use the scipy cluster.
5.	The shape of the cluster here is hyperspherical. (2D-circle and in 3D it will be a sphere).	Here the shapes of the cluster are neither 2D nor 3D.

2 Briefly explain the steps of the K-means clustering algorithm?

A-2 Steps of K-Means Clustering:

1. The No. of data points (e.g. **N=10**) which are given to us we divide into some pre-defined clusters. Let’s say **K=2** is the number of clusters.
2. Then we have to select 2 random cluster centers.
3. Now we calculate the distance of each data point to the 2 cluster centers and allocate the point to the center which is having the least distance to the center.
4. Now we have to recompute the center of the 2 clusters which will be the mean of the individual points of that cluster. This will give us the new cluster center.
5. Now we will update the position of the cluster center again using steps 3 and 4 again.
6. This process will stop when don’t see any change in the position of the data points.

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

A-3 There are 2 methods by which we select the value of K. Those are:

1. Statistical Method
2. Business Aspect

Statistical Method: This method involves the use of 2 techniques which are:

1. SSD-Sum of Squared Distances
2. Silhouette Analysis

In the **SSD Method,** we plot an **elbow curve** and see the dip in the curve where there was a significant change and then beyond that point it became stagnant. That gives us the K-Value from there.

In **Silhouette Analysis**, We calculate the **P** (Mean Distance to the points in the nearest cluster) and the **Q value** (Mean Intra Cluster Distance). The score obtained then gives us the good possible values of K from which we can select the more suitable one.

Business Aspect: By Business Aspect what we mean is that we if we knowledge of the concerned area on which we need to select the No. of clusters then we can ideally do that by having an interaction with the POC of the firm as well. In case we don't know about that then we should go ahead with Statistical Method.

4. Explain the necessity for scaling/standardization before performing Clustering?

A-4 Scaling process helps us in bringing the variables to a comparable scale when we plot any graph over them. I mean we can't plot variables where we have dominating variables suppressing other important variables that's why we use the scaling/standardization process before we perm Clustering.

5. Explain the different linkages used in Hierarchical Clustering?

A-5 Different types of linkages that we use are:

1. Single Linkage
2. Complete Linkage
3. Average Linkage

Single Linkage: It is the smallest distance between 2 points in the 2 clusters.

Complete Linkage: Distance between 2 clusters is defined as the maximum distance between any 2 points in the cluster.

Average Linkage: Distance between 2 clusters is defined as the average distance between every point of one cluster to every other point of another cluster.