Clustering assignment – II

Question 1: Assignment Summary

Problem Statement: 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. 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 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.

Objective:

- To categorize the countries using some socio-economic and health factors that determine the overall development of the country.
- To suggest the countries which the CEO needs to focus on the most.

Steps followed:

- **Data Inspection, Cleaning and EDA:** As a part of this step, the following actions were performed:
  - There are 167 countries in the dataset with 9 socio-economic and health features related to each of them.
  - There are no null values or duplicates in the data set.
  - There are outliers present in the data which have been handled by putting a cap to them.
  - High correlation exists between the variables in the data.
  - The data has been standardized using the standard scaler method.
- **K-Means Clustering:** This method was used to cluster the country data, based on the SSD/elbow curve and silhouette analysis we concluded that the appropriate number of clusters should be 3.

  Clusters formed were analyzed using box and bar plots. Based on the analysis we found that **Cluster 2** seemed to have the countries which were in dire need of aid. (Burundi, Liberia, Congo, Dem. Rep, Niger, Sierra Leone, Madagascar, Mozambique, Central African Republic, Malawi, Eritrea)

- **Hierarchical Clustering:** This method was used to cluster the country data by using both single and complete leakage.

  Clusters formed were analyzed using box and scatter plots. Based on the analysis we found that **Cluster 0** seemed to have the countries which were in dire need of aid. (Burundi, Congo, Dem. Rep, Niger, Sierra Leone, Madagascar, Mozambique, Central African Republic, Malawi, Eritrea, Togo)

Conclusion: As by both K means and Hierarchical clustering method - we have got almost same countries which requires aid. The following are the countries which are in direst need of aid by considering socio – economic factor into consideration:

- Burundi
- Congo, Dem. Rep.
- Niger
- Sierra Leone
- Madagascar
- Mozambique
- Central African Republic
- Malawi
- Eritrea

Question 2: Clustering

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

1. K-means requires prior knowledge of the number of clusters whereas hierarchical does not.
2. With the same hardware configurations, hierarchical clustering cannot handle big data well, but k-means clustering can. This is because the time complexity of k-means is O(n) i.e., linear whereas of hierarchical its $O(n^2)$.
3. In K-means, since we start with a random choice of clusters, the results produced by running the algorithm multiple times may differ while in Hierarchical clustering they are reproducible.
4. K-means works well when the shape of the cluster is hyper spherical (like circle in 2D, sphere in 3D)

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

1. Randomly pick k points in space and take them as the cluster centers.
2. Assign each observation in your dataset to the cluster that minimizes the Euclidean distance.
3. Recompute the center of each cluster by taking the means of each of the observation in each cluster.
4. Repeat steps 2 & 3 until either the center doesn't change, or the maximum number of iterations is reached.

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

The methods to determine the optimal number of clusters(k) include both direct methods and statistical testing methods:

Direct methods: consists of optimizing a criterion, such as the within cluster sums of squares or the average silhouette. The corresponding methods are named **elbow and silhouette methods**, respectively.

Elbow Method:

1. Compute clustering algorithm (e.g., k-means clustering) for different values of k. For instance, by varying k from 1 to 10 clusters.
2. For each k, calculate the total within-cluster sum of square (wss).
3. Plot the curve of wss according to the number of clusters k.

4. The location of a bend (knee) in the plot is generally considered as an indicator of the appropriate number of clusters.

Silhouette Method:

1. Compute clustering algorithm (e.g., k-means clustering) for different values of k. For instance, by varying k from 1 to 10 clusters.
2. For each k, calculate the average silhouette of observations (avg.sil).
3. Plot the curve of avg.sil according to the number of clusters k.
4. The location of the maximum is considered as the appropriate number of clusters.

Statistical testing methods consists of comparing evidence against null hypothesis. An example is the **gap statistic**. The gap statistic compares the total within intra-cluster variation for different values of k with their expected values under null reference distribution of the data. The estimate of the optimal clusters will be value that maximize the gap statistic (i.e, that yields the largest gap statistic). This means that the clustering structure is far away from the random uniform distribution of points.

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

Scaling/Standardization is often performed as a pre-processing step, particularly for cluster analysis, standardization may be important when working with data where each variable has a different unit (e.g., inches, meters, tons and kilograms), or where the scales of each of the variables are very different from one another (e.g., 0-1 vs 0-1000). The reason this importance is particularly high in cluster analysis is because groups are defined based on the distance between points in mathematical space.

When working with data where each variable means something different, (e.g., age and weight) the fields are not directly comparable. One year is not equivalent to one pound and may or may not have the same level of importance in sorting a group of records. In a situation where one field has a much greater range of value than another (because the field with the wider range of values likely has greater distances between values), it may end up being the primary driver of what defines clusters. **Standardization helps to make the relative weight of each variable equal by converting each variable to a unitless measure or relative distance**.

e) Explain the different linkages used in Hierarchical Clustering.

The various types of clusters used in Hierarchical clustering are :

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.

Average-linkage: the distance between two clusters is defined as the average distance between each point in one cluster to every point in the other cluster.

Centroid-linkage: finds the centroid of cluster 1 and centroid of cluster 2, and then calculates the distance between the two before merging.