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:

Describing Clustering of Countries:

- 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.

Solution Methodology:

- To categorize the countries based on the socio-economic and health factors that determine the overall development of the country

- To suggest the countries which the CEO needs to focus the most υ The approach we used here to cluster the countries are:
  - ➢ Data Inspection and EDA
  - ➢ Outlier analysis
  - ➢ Hierarchical and K means clustering

- Analysis of these clustering techniques

My choices:

- Both Univariate and Bivariate analysis where done for the variables in the data.
- Converted columns which has percentage values to absolute values.
- Did capping to treat outliers.
- Visualized the data using boxplot, bar plot, scatterplot etc.
- Performed Hopkins statistics
- Applied StandardScalar() to the data which will help us in clustering
- Used Silhouette and Elbow curve to find the optimal cluster
- Used K-means and hierarchical clustering techniques
- Finally using k=3, clusters of countries were formed.

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.

1. K -means needs prior knowledge of number of centroids (K) whereas hierarchical cluster do not need this kind of parameters, cut_tree() function is used to create the number of clusters of any choice

2. In K-means clustering the algorithm will calculate the centroid each time
3. K-means is fast as compared to hierarchical clustering

b.

1. Specify number of clusters *K*.

2. Initialize centroids by first shuffling the dataset and then randomly selecting *K* data points for the centroids without replacement.

3. Keep iterating until there is no change to the centroids. i.e assignment of data points to clusters isn't changing.

4. Compute the sum of the squared distance between data points and all centroids.

5. Assign each data point to the closest cluster (centroid).

6. Compute the centroids for the clusters by taking the average of the all data points that belong to each cluster.

c.

1) A cluster center is the representative of its cluster. The squared distance between each point and its cluster center is the required variation. The aim of k-means clustering is to find these k clusters and their centers while reducing the total error**.**

2) Two methods that can be useful to find this mysterious k in k-Means.

   I. Elbow curve:
   Calculate the **Within-Cluster-Sum of Squared** Errors (WSS) for **different values of k**, and choose the k for which WSS becomes first starts to diminish. In the plot of WSS-versus-k, this is visible as an **elbow.**

II. Silhouette analysis:
The silhouette value measures how similar a point is to its own cluster (cohesion) compared to other clusters (separation).

3) The Elbow Method is more of a decision rule, while the Silhouette is a metric used for validation while clustering. Thus, it can be used in combination with the Elbow Method. Therefore, the Elbow Method and the Silhouette Method are not alternatives to each other for finding the optimal K. Rather they are tools to be used together for a more confident decision.

d.

1. It refers to the process of rescaling the values of the variables in your data set so they share a common scale. Often performed as a pre-processing step, particularly for cluster analysis, standardization may be important if you are working with data where each variable has a different unit (e.g., inches, meters, tons and kilograms), or where the scales of each of your variables are very different from one another (e.g., 0-1 vs 0-1000).
2. The reason this importance is particularly high in cluster analysis is because groups are defined based on the distance between points in mathematical space.
3. When you are working with data where each variable means something different, (e.g., age and weight) the fields are not directly comparable
4. 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
5. Standardization helps to make the relative weight of each variable equal by converting each variable to a unitless measure or relative distance.

e.

1. Single Linkage: Single-linkage (nearest neighbor) is the shortest distance between a pair of observations in two clusters. It can sometimes produce clusters where observations in different clusters are closer together than to observations within their own clusters. These clusters can appear spread-out.

2. Complete Linkage: Complete-linkage (farthest neighbor) is where distance is measured between the farthest pair of observations in two clusters. This method usually produces tighter clusters than single-linkage, but these tight clusters can end up very close together. Along with average-linkage, it is one of the more popular distance metrics.

3. Average Linkage: Average-linkage is where the distance between each pair of observations in each cluster are added up and divided by the number of pairs to get an average inter-cluster distance. Average-linkage and complete-linkage are the two most popular distance metrics in hierarchical clustering.

4. Centroid Linkage: Centroid-linkage is the distance between the centroids of two clusters. As the centroids move with new observations, it is possible that the smaller clusters are more similar to the new larger cluster than to their individual clusters causing an inversion in the dendrogram. This problem doesn’t arise in the other linkage methods because the clusters being merged will always be more similar to themselves than to the new larger cluster.