Clustering Assignment Part 2

Q.1) Assignment Summary

Problem Statement: We need to categories the countries using socio-economic and health factors that determine the overall development of the country. Then we need to suggest the countries which the CEO (HELP International which is an International NGO) needs to focus on the most.

Methodology:

- Firstly, the Data Extraction/Reading step was performed over the dataset by reading the data from the .csv file into Dataframe i.e. **country_df.**
- After reading the data into dataframe, we performed Preprocessing part by understanding its shape (no of rows and columns), checking data types and outliers. On the basis of above steps and visualization using Box Plot, we treated some of outliers as data size was low, we didn't cap all the outliers.
- After cleaning data, we ran visualization using Distribution Plot and Box Plot to generate some insights from various variables.
- Now, we are ready to perform Modelling but before that we ran Hopkins Statistics to see how well the data can be clustered, then we rescaled the data for further inspection and finding 'k' (cluster) value.
- We performed Elbow Curve and Silhouette Score to find optimal 'k' value for model building.
- Now, we can build our model using KMean Clustering Algorithm and generate our first result of 'Top 5 Countries in need of urgent Aid'.
- Similarly, we performed Hierarchical Clustering Algorithm for model building, before that we used Linkages (i.e. Single Linkage, Complete Linkage and Average Linkage) to find optimal value for 'k' and among these three Complete Linkage gave suitable value.
- Lastly, we performed model building using Hierarchical Clustering Algorithm and generated our second result of 'Top 5 Countries in need of urgent Aid'.
- Though both Model gave us same result but I would prefer to choose KMean Clustering Algorithm because
  - Hierarchical Clustering clustered almost all data points under single label and very less under remaining labels and this makes to visualize the result accordingly.

Q.2) Compare and contrast K-means Clustering and Hierarchical Clustering

Ans.

K-means Clustering	**Hierarchical Clustering**
K-means can handle big data well.	Hierarchical clustering cannot handle big data very well
K-means require prior knowledge about value of k.	In hierarchical clustering, there is no need to prior knowledge of value of k and we can stop to whatever point once we select appropriate value of k.
K-means use two methods to decide k value- • Elbow Curve • Silhouette Score	Hierarchical clustering uses Linkage methods like- • Single Linkage • Average Linkage • Complete Linkage
K-means takes less time to build model because it uses non-linear methods to find optimal value for k.	As compare to K-means, Hierarchical clustering uses more time because it uses Linear approach to find optimal value for k.
Industry K-Means is used less frequently.	Whereas, hierarchical is used more frequently to find most optimal k value along with K-Means when data size is huge.
The time complexity for k-means clustering is O(n).	The time complexity for hierarchical clustering is O(n^2).

Q.3) Briefly explain the steps of K-means clustering Algorithm

1. Clusters the data into *k* groups where *k* is predefined.
2. Select *k* points at random as cluster centers or centroid.
3. Assign objects to their closest cluster centroid according to the **Euclidean distance** function.
4. Calculate the centroid or mean of all objects in each cluster.
5. Repeat steps 2, 3 and 4 until the same points are assigned to each cluster in consecutive rounds and there is no change in the centroid in step 2.

Q.4) How is the value of ‘k’ chosen in K-means clustering? Explain both the statistical as well as the business aspect of it?

Ans. Value of ‘k’ is chosen by using Elbow Curve and Silhouette Score methods which can be achieved under sklearn.cluster library.

- Statistical aspect is the one where industry decide the value of ‘k’ using statistical methods like Elbow Curve or Silhouette Score. In elbow curve, elbow like plotting is done, the value on which the elbow more bent’s is selected as optimal ‘k’ value and in Silhouette Score the score is generated for each range of clusters between -1 to 1 the point with higher score is selected.
- In business aspect, the plotting will be done to verify but the selection of value of ‘k’ will be done on the basis of domain and problem statement to be achieved. Also, in Silhouette Score value with 2nd or 3rd or so on can also be selected as per business requirement.

Q.5) Explain the necessity for scaling/standardization before performing clustering.

Ans.

- Scaling or Standardization is a method in which the values of each variables in the data is brought to almost same range.
- For example, when we want to visualize or cluster two variables say ‘Age’ and ‘Income’ we will find the variable ‘Age’ was unable to make any impact in generating result and insights because the minimum and maximum value for ‘Age’ is in two or three digits whereas for ‘Income’ it was in 4-7 (approx.) digits.
- Hence, we need to perform Scaling or Standardization before performing clustering or visualization.

Q.6) Explain the different linkages used in Hierarchical clustering.

Ans. Linkages in Hierarchical Clustering are used to find the optimal number of clusters by plotting a tree like structure with the help of Dendrogram. There are basically three types of Linkages in Hierarchical clustering- Single Linkage, Average Linkage and Complete Linkage. Among all three linkages Single Linkage is more likely to generate worst and improper tree structure of clusters, whereas, Complete or Average will generate more proper tree like structure.

1. Single Linkage:
    - Here, the distance between two clusters is defined as the shortest distance between points in the two clusters.
2. Complete Linkage:
    - Here, the distance between two clusters is defined as the maximum distance between any two points in the two clusters.
3. Average Linkage:
    - Here, the distance between two clusters is defined as the average distance between every point of one cluster to every other point of the other cluster.