Assignment Part II:-

Question 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) ?

Answer:

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 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. 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:
The requisite is:

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

Method followed:

-Data Processing:

- It was found that no null values.
- There was also no duplicate value for the country.
- Then we have done EDA and Outlier treatment on the data.
- And after EDA & Outlier treatment was done on the data.

Assignment Part II:-

-Clustering:

1. Both the methods KMeans and Hierarchical Clustering was used on the 4 PCA components
2. For K means, K= 3 was taken using the elbow dip and silhouette analysis .
3. While doing the Hopkins Statistics a value of 0.87 was attained.
4. If the Hopkins Statistics values are:
   - 0.3 : Low chase of clustering
   - around 0.5 : Random
   - 0.7 - 0.99 : High chance of clustering

Finally using all these values clusters of 3 were formed and the countries are split into clusters.

Question 2

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

Answer:

1. K Means needs a prior knowledge of number of centroid (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 compared to hierarchical clustering.
4. Hierarchical clusters need more ram to run.

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

Answer:

Kmeans algorithm is an iterative **algorithm** that tries to partition the dataset into Kpre-**defined** distinct non-overlapping subgroups (**clusters**) where each data point belongs to only one group. ... Keep iterating until there is no change to the centroids. i.e assignment of data points to **clusters** is not changing.

Assignment Part II:-

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

Answer

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

When we use a k-means clustering algorithm, we will need to select the number of clusters we would like to work with. Selecting the optimal number of clusters is important because this will fall somewhere between full localisation or standardisation. There are two statistical approach.

- **The Elbow method:** To determine the optimal number of clusters, we will need to run the k-means algorithm for different values of k (number of clusters). For each value of k, we will then need to calculate the total within-cluster sum of squares (wss). You can then plot the values of wss on the y-axis and the number of clusters (k) on the x-axis. The optimal number of clusters can be read off the graph at the x-axis.
- **The Silhouette coefficient:** To determine the optimal number of clusters, we will need to measure the quality of the clusters that were created. This value determines how closely each data point is to the centroid of its cluster. A high average silhouette coefficient indicates successful clusters. This method checks the silhouette coefficient for different values of k. The optimal number of clusters is, therefore, the maximised silhouette value for the data set.

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

Answer

This will impact the performance of all distance-based model as it will give higher weightage to variables which have higher magnitude. Hence, it is always advisable to bring all the **features** to the same scale for **applying** distance-based **algorithms** like KNN or **K**-**Means**

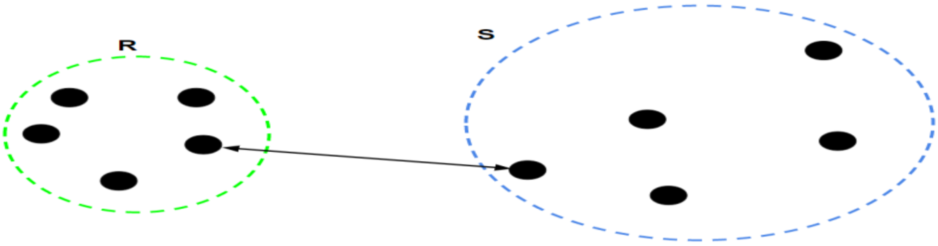
Assignment Part II:-

e) Explain the different linkages used in Hierarchical Clustering.

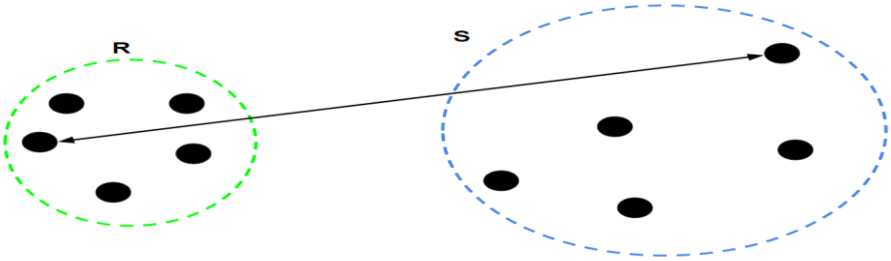
Answer

The process of Hierarchical Clustering involves either clustering sub-clusters (data points in the first iteration) into larger clusters in a bottom-up manner or dividing a larger cluster into smaller sub-clusters in a top-down manner.. The different types of linkages are:-

1. **Single Linkage:** For two clusters R and S, the single linkage returns the minimum distance between two points i and j such that i belongs to R and j belongs to S.



2. **Complete Linkage:** For two clusters R and S, the single linkage returns the maximum distance between two points i and j such that i belongs to R and j belongs to S.



3. **Average Linkage:** For two clusters R and S, first for the distance between any data-point i in R and any data-point j in S and then the arithmetic mean of these distances are calculated. Average Linkage returns this value of the arithmetic mean.where

– Number of data-points in R – Number of data-points in S