Clustering Case Study:

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)

Ans:

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 - Our job is to categorise the countries using some socio-economic and health factors that determine the overall development of the country. Then you need to suggest the countries which the CEO needs to focus on the most.

Solution -

1. Understanding the data
2. Data Cleansing : a.) Missing value check

   b.) Data type check

   c.) Check for duplicates
3. Data Visualisation : Use different plots such as pairplot , heatmap and box plots .
4. Data Preparation : Scaling of Data
5. PCA Application : Selection of principal components followed by analysis and treatment.
6. Hopkins Test
7. Building the model : Application of K-means and Hierarchal clustering with Elbow curve and Silhouette Analysis
8. Final Analysis : Analysing and preparing the final list of counties along with cluster ID and drawing insights by using plots.

Question 2: Clustering

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

Ans : K-Means is method of cluster analysis using a pre-specified no. of clusters. It requires advance knowledge of 'K'.

Advantages
a. Convergence is guranteed

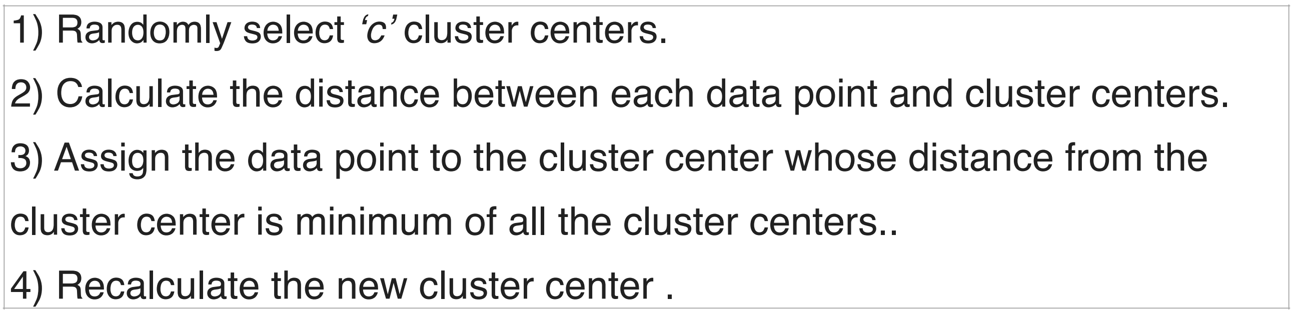
Hierarchical Clustering also known as hierarchical cluster analysis (HCA) is also a method of cluster analysis which seeks to build a hierarchy of clusters without having fixed number of cluster.

Advantages:
a .Ease of handling of any forms of similarity or distance.

b. Consequently, applicability to any attributes types.

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

Ans :

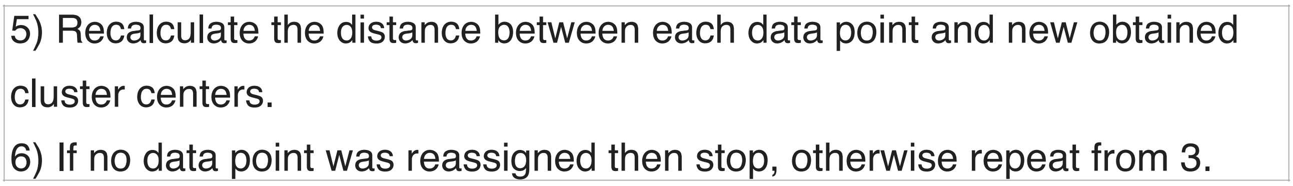


1) Randomly select *‘c’* cluster centers.

2) Calculate the distance between each data point and cluster centers.

3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers..

4) Recalculate the new cluster center .



5) Recalculate the distance between each data point and new obtained cluster centers.

6) If no data point was reassigned then stop, otherwise repeat from 3.

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

Ans : 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.But there are methods to decide this

- The Elbow Method
- The Silhouette Method

4. Explain the necessity for scaling/standardisation before performing Clustering.

Ans : When we standardize the data prior to performing cluster analysis, the clusters change. We find that with more equal scales, the Percent Native American variable more significantly contributes to defining the clusters. Standardization prevents variables with larger scales from dominating how clusters are defined.

5 . Explain the different linkages used in Hierarchical Clustering.

Ans : 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.