Clustering Assignment

1. Assignment Summary

Problem Statement: To categorise the countries using some socio-economic and health factors (determining the overall development of the country) and analyse the clusters to identify the ones which are in dire need of aid so that the international humanitarian NGO named "HELP" which is committed to fighting poverty can providing the people of backward countries with basic amenities and relief during the time of disasters and natural calamities can utilise its funds efficiently.

Methodology:

- Data Quality Check: The dataset has no missing values thus, no use of missing values imputation/treatment needed. Nevertheless, there were certain columns ('export', 'import' and 'health') bearing values in percentage of their respective countries GDP, which were then converted into units. On further examination of dataset, it is clearly visible that the columns expressed values in sets of different scales thus, requiring the need of scaling the data.

- EDA: On analysing the data for outliers, it was clearly visible that certain countries were outperforming while some were underperforming with respect to each other in terms of income, imports and exports (to mention a few) thus, there was a need of soft capping those variables (1% - 2%) to minimise the effect of those outliers in model building.

- Scaling: As mentioned before the data needs to be scaled for further analysis and model building to bring the variables to a comparable scale. The method used for scaling was Standardisation.

- Clustering: The data was subjected to KMeans algorithm and Hierarchical Clustering to create clusters of developed and under-developed countries. To obtain the optimal number of clusters for KMeans model building we used 'Sum of Squared Distances/Elbow Curve' method and 'Silhouette Score'.
  For Hierarchical Clustering model both methods of single and complete linkages were used. Further on comparing various models of KMeans and Hierarchical Clustering the model that gives better outcomes in clustering under-developed and developed countries is *KMeans model with K=4*. The comparison was based on factors like GDPP, Health, Income and Child Mortality. The clusters formed were easily distinguishable from each other.

  The other models were rejected because they couldn't provide better clusters, therefore it was hard to bifurcate developed and underdeveloped countries. For e.g., Nigeria alone was taking a separate cluster for K=5 model and the Hierarchical models.

  Therefore, the Top 10 countries which are in direst need of aid are:

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

2. Clustering:

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

- In KMeans Clustering it is important to decide the value of 'K' beforehand, whereas there is no such restriction for Hierarchical Clustering.
- It is comparatively difficult to predict the value of K (i.e., cluster) for K Means. On the other hand, it is easy to predict the number of clusters in case of Hierarchical Clustering with the help of dendrogram.

- K Means Clustering can easily handle large sets of data while in case of Hierarchical Clustering it requires more storage and computational power as it calculates the NxN distance/similarity matrix (given N items to be clustered) which calculates the distance of each data points from the other. This step is a repetitive step which is repeated from forming N clusters to a single cluster.

- In Hierarchical Clustering the results produced are similar no matter how many times we run the algorithm whereas in terms of K Means Clustering the results may vary after running the algorithm many times because at first the centroid is chosen at random, so every time we run the algorithm it may produce different results until it hits the convergence point, and the centroids position doesn't change.

- In Hierarchical Clustering the algorithm can give results in a single run (although the time taken for that single run depends on storage and computational power) whereas in case of K Means clustering

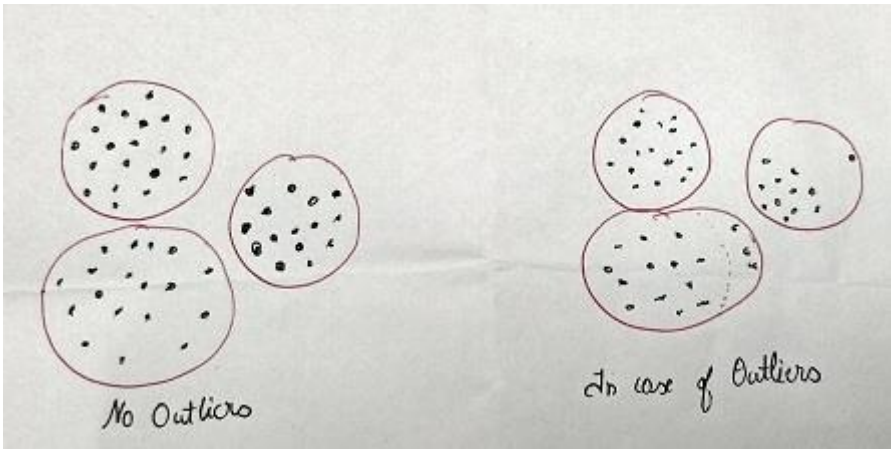
the algorithm should be run multiple times (till the convergence of centroids) to get better results.

- K Means is highly sensitive to outliers and can act as an hinderance in finding optimal clustering, in case of Hierarchical Clustering the single linkage method is said to be more sensitive to outliers than complete linkage.

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

K Means clustering is an approach of partitioning 'N' number of observations into 'K' distinct and non-overlapping clusters. Before implementing K-means clustering following steps are necessary:

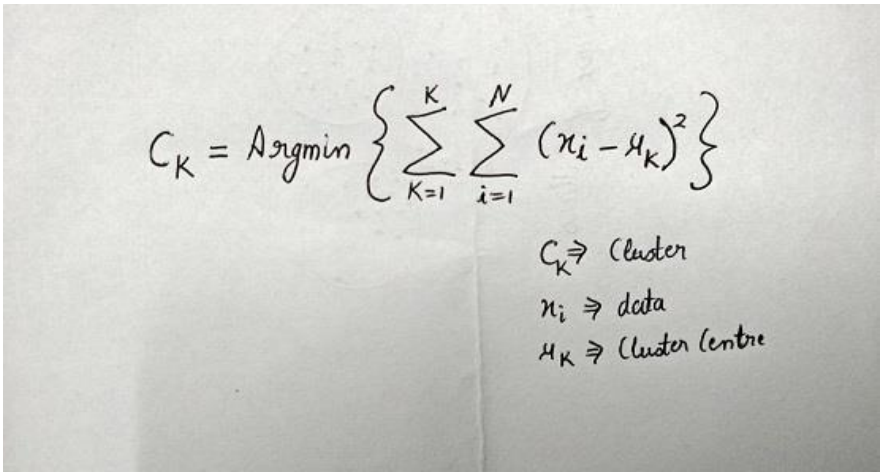
- To predetermine the number of clusters (value of K)
- If the observations are having different scale/range, then they must be scale down to make them uniform.
- If the are outliers in the data then they must be treated as per requirement as they may effect the final clusters because K-means tries to allocate each and every data point to one of the clusters.



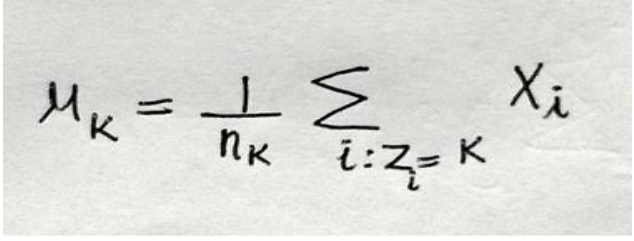
- One must run the algorithm multiple times as the algorithm may not converge at first, therefore we must run it multiple times until centroid positions of clusters do not converge.

K-means Algorithm: The steps are as follows –

- Initialisation: Random selection of cluster centres for each of the K clusters is done (They can be from any of the N observations or totally different).
- Assignment: Each of the N observation gets assigned to a cluster whose cluster centre is close to it, which is done using the squared Euclidean distance. The equation is as follows:



  K-means assigns the data in such a way that each data point gets assigned to at least one cluster and no observation belongs to more than one cluster.
- Optimisation: For the K clusters obtained new cluster centres are computed which will be the mean of all the cluster members. Now re-assigning of data points is done based on new cluster centres obtained. The equation for this step is:



n_k= no. of observation in Kth cluster

Now the assignment and optimisation steps are repeated until no new cluster centres are formed. This is the point were we get optimal clusters.

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

To obtain the optimal values of K for K-means clustering following statistical methods are used:

Silhouette Analysis: it tells us about the similarity of the observations to its own cluster (termed as cohesion) compared to other clusters (termed as separation). The value of Silhouette ranges between -1 and +1. For a Silhouette value close to +1 indicates well assigned cluster whereas a Silhouette value close to -1 indicates that the observation is not well matched with the cluster. On the other hand, a Silhouette value close to 0 means that the observation lies on the border of two clusters.

Silhouette value is given by :

$$S(i) = \begin{cases} 1 - \frac{a(i)}{b(i)} & ; \text{ if } a(i) < b(i) \\ 0 & ; \quad a(i) = b(i) \\ \frac{b(i)}{a(i)} & ; \quad a(i) > b(i) \end{cases}$$

$a(i) \Rightarrow$ avg distance from own cluster

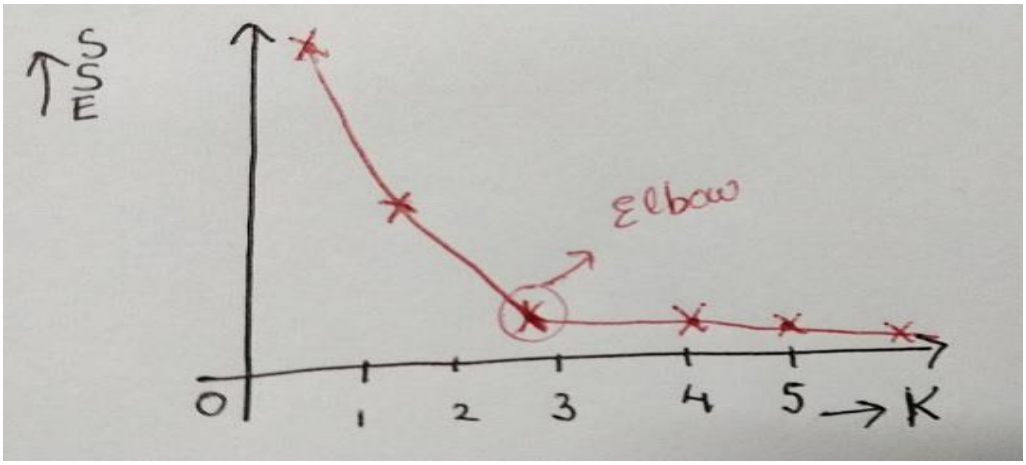
$b(i) \Rightarrow$ avg " " nearest neighbour cluster)

Therefore, a(i) should be as small as possible, and b(i) should be as large as possible to get better clusters.

Therefore, to obtain optimal number of clusters the K-means algorithm is computed for a set of different values of K and for each values of K the average silhouette is calculated. So, the K with maximum score can be used as optimal values of K.

Elbow Curve method: Like Silhouette method here too the K-means algorithm is computed for different values of K along with sum of squared errors (SSE) for each K. On plotting K and SSE values of each K an arm like structure can be visible. The point where elbow of the arm is located is interpreted as optimal values of K.

The graph looks like:



The objective is to find a value of K which has a small SSE and the elbow usually determines that point.

In terms of business aspect of choosing optimal value of K, we must not completely rely on the statistical methods as they may be statistically correct interpretation of optimal values of clusters but may stand useless as per the business understanding so we must select at least a few optimal values of K either with the help of elbow curve or Silhouette or a combination of both and then compare those models select the one which suits the business requirement.

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

Most of the time the data sets have variables with observation in varying ranges and scale or in different units of measure. Thus, if re-scaling is not performed over the data set then the variables having data in higher ranges or units may overshadow other important variables (having data in comparatively smaller range or unit). This makes it necessary to scale the data before clustering.

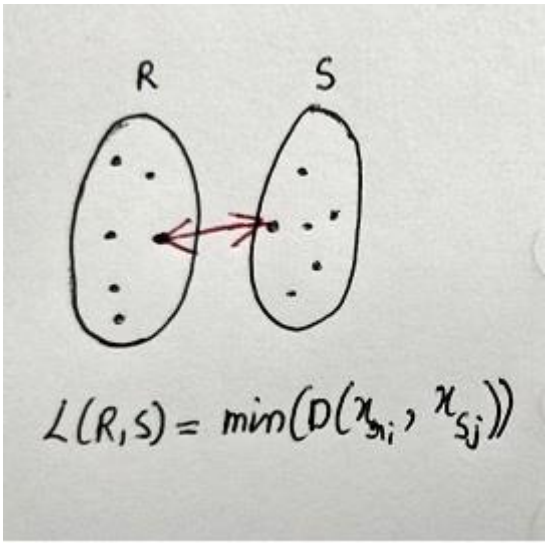
The method preferred to scale the data is called Standardisation. Standardisation is a process of converting observations into z-score with mean equal to 0 and standard deviation as 1. This way the attributes with larger range of values will not out-weight the attributes with values in smaller range.

Standardisation also helps in making the variables uniform and unit free.

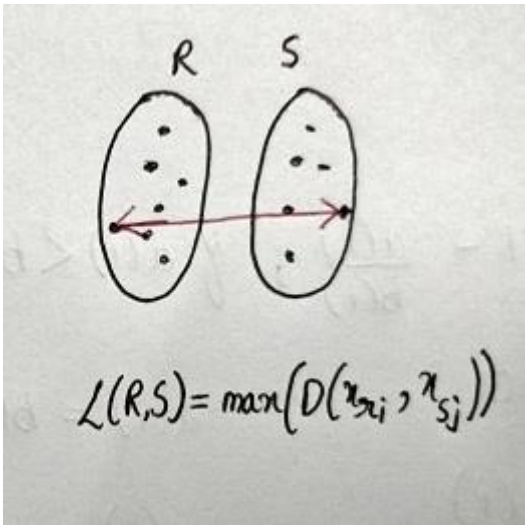
2.5. Explain the different linkages used in Hierarchical Clustering.

Different linkage methods are used in how the distance between each cluster is measured. The following linkages are commonly used in Hierarchical Clustering :

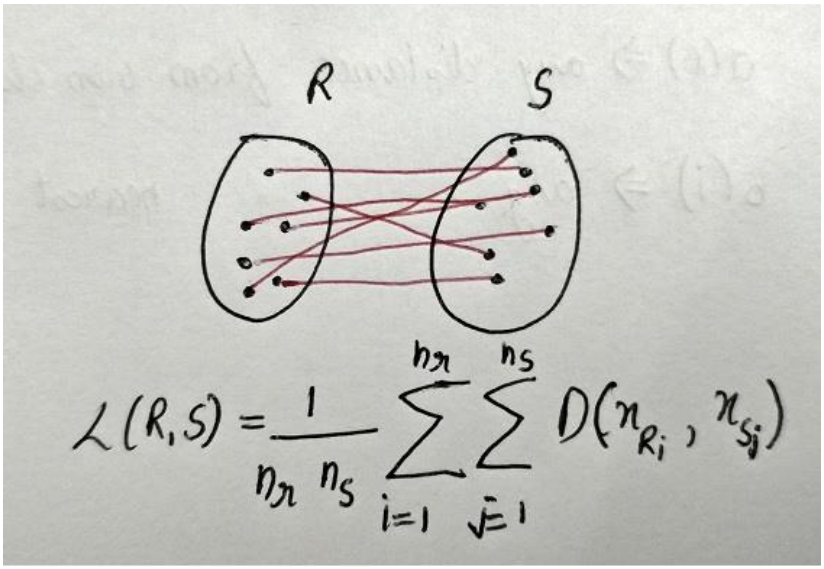
Single Linkage: In single linkage the distance between the clusters is computed as the shortest/minimum distance points in each cluster. Usually, the dendrograms produced by single linkage are not well structured.



Complete Linkage: In complete linkage the distance between the clusters is computed as the longest/maximum distance between any two points in each cluster. The dendrograms produced are well tree shaped structured.



Average Linkage: In average linkage the distance between the clusters is computed as the average distance between each point of a cluster to all the points of the other cluster. The dendrograms produced are well shaped tree structures. Commonly average linkage is used as the clusters are relatively compact and relatively distanced apart.



$$\mathcal{L}(R,S) = \frac{1}{n_R \, n_S} \sum_{i=1}^{n_R} \sum_{j=1}^{n_S} D(r_{Ri}, r_{Sj})$$