***Question – 1 Assignment Summary***

HELP International is an international humanitarian NGO which is committed to fight 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.

In recent fund program NGO raised $10M and CEO of the NGO wants to use this money strategically and effectively to provide to the countries which are very poor and require the aid urgently.

- The data is first read and its structure/dimension is determined so as to understand the rows and columns of the dataframe.
- Name short forms are converted to easy, readable & understandable format i.e they are standardized.
- Missing values are checked in the dataframe if any are present.
- Few column values which were % of GDP were converted to actual values.
- Statistical understanding of the data is build by describing the data.
- Heatmap of correlation matrix is plotted to understand the relation between the variables of the dataframe.
- Various bar plot is plotted to understand the count of countries (in %) between GDP, income, child mortality, life expectancy, total fertility. So as to understand how many countries are underdeveloped, underprivileged and require to fund the meet the basic amenities of the life.
- The continuous variable/columns are standardized to bring them to same scale.
- Few checks are performed on the dataframe like to Hopkins score to check whether the data is feasible for clustering or not, then silhouette & elbow analysis is done to what is the feasible no. of clusters will be formed & then followed by model building using K Means clustering.
- Once model is build using K Means, clusters are visualized to understand the countries distribution and filtered according to child mortality, GDP & income.
- Another the hierarchical clustering is performed and clusters are obtained to identify the countries according to child mortality, income & GDP.
- Both the clustering produced similar results.

---

***Question – 2 Clustering***

**a)** Compare & contrast K-means clustering & hierarchical clustering ?

**Ans)** Differences between k means & hierarchical clustering are as below:-

- Hierarchical clusters can't handle big data well but k means clustering can. This is because the time complexity of K Means is linear i.e. O(n) while that of hierarchical clusters is quadratic i.e. O(n2).
- In K Means clustering, since we start with random choice of clusters, the results produced by running the algorithm multiple times might differ. While results are reproducible in hierarchical clustering.

- K Means is found to work well when the shape of the clusters is hyper spherical (like circle in 2D, sphere in 3D).
- K Means clustering requires prior knowledge of K i.e. no. of clusters you want to divide your data into. But you can stop at whatever number of clusters you find appropriate in hierarchical clustering by interpreting the dendrogram.

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

**Ans)** first step, is to choose the number of clusters, k.

Second step, is to randomly select the centroids for the clusters from the datapoints present.

Third step, is to assign the closest datapoints to the nearest cluster centroid.

Fourth step, once all the datapoints are assigned to the clusters, centroids of those clusters are computed again.

Fifth step, is to assign the nearest datapoints to the newest centroid & form the new clusters.

Step 3 & 4 are repeated till

- Centroid of the newly formed clusters do not change even after multiple iteration.
- Points assigned in the cluster are not changing i.e remains same.

**c)** 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' chosen depends on the 2 methods i.e Elbow method & other is Average Silhouette Method.

Elbow method: -

- Compute clustering algo for different values of k. for instance by varying from 1 to 10 clusters.
- For each k, calculate the total within cluster sum of squares (wss).
- Plot the curve of wss according to the number of cluster k.
- The location of a bend (knee) in the plot generally considered as an indicator of the appropriate number of clusters.

Average silhouette method: -

- Compute clustering algo for different values of k. For instance by varying from 1 to 10 clusters.
- For each k, calculate the average silhouette of observations (avg.sil).
- Plot the curve of avg.sil according to the number of clusters k.
- The location of the maximum is considered as the appropriate number of clusters.

**d)** Explain the necessity for scaling/standardization before performing clustering.

**Ans)** There are 2 reasons why standardization (converting into z-scores whose mean is 0 & standard deviation is 1) is necessary for kmeans algorithim:-

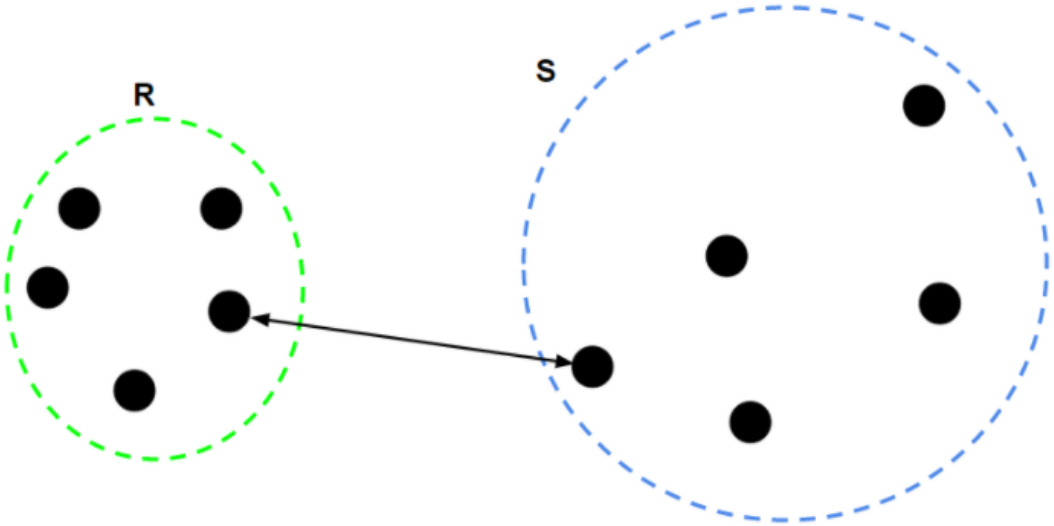
- As the Euclidean distance has to be calculated between the data points it is necessary to scale down all the attributes to same normal scale which helps in the process so that the attributes with large values do not outweigh the smaller values.
- Standardization helps in making the attributes unit free & uniform.

**e)** Explain the different linkages used in hierarchical clustering.

**Ans)** Following types of linkages are present in hierarchical clustering :-

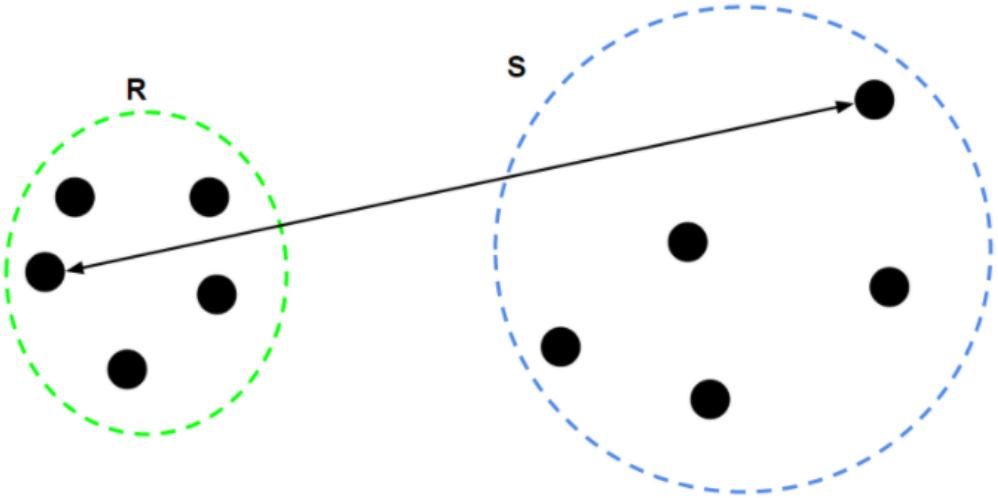
- Single linkage :- In single linkage hierarchical clustering, the distance between two clusters is defined as the shortest distance between two points in each cluster. For two clusters R & S, the single linkage returns minimum distance between 2 points i & j such that i belongs to R & j belongs to S.

$$L(R,S) = min(D(i,j)), i \epsilon R, j \epsilon S$$




- Complete linkage :- In complete linkage hierarchical clustering, the distance between two clusters is defined as the longest distance between two points in each cluster. For two clusters R & S it returns the maximum distance between two points i & j such that i belongs to R & j belongs to S.

$$L(R,S) = max(D(i,j)), i \epsilon R, j \epsilon S$$




- Average linkage :- the distance between two clusters is defined as the average distance between each point in one cluster to every point in the other cluster. For two clusters R & S, first for the distance between any data point i in R & any data point j in S & then arithmetic mean of these distances are calculated. Average linkage returns this values of the arithmetic mean.

$$L(R,S) = \frac{1}{n_R + n_S} \sum_{i=1}^{n_R} \sum_{j=1}^{n_S} D(i,j), i \epsilon R, j \epsilon S$$

where

$n_R$ – Number of data-points in R

$n_S$ – Number of data-points in S