Assignment 1

Summary of assignment

Problem Statement:

HELP international NGO has raised fund to fight poverty, providing basic amenities to the people of countries who are in direst need of it. Data for all countries with some essential variables is given. We need to cluster countries best on various factors such as income, gdpp, child mortality rate to get countries which are in need of fund.

Step 1. Reading and understanding data:

- First step is reading and understanding data. Data contains 167 rows and 10 columns.
- We then checked summary and basic info such as data types and null values.
- All columns except country are numerical.

Step 2. Data Cleaning and Visualisation:

- In data dictionary we have given that columns 'export', 'import', 'health' are as percentage of 'gdpp', so we converted them in values.
- We then checked for null values present in data but luckily there were no null values in data.
- Next we visualized data by plotting correlation heatmap between columns and we got to know that income and gdpp is highly correlated. On other hand total fertility and child mortality is highly correlated.
- We plotted boxplot to understand outliers in data respectively. There were outliers in almost all columns.

Step 3. Data Preparation:

- We capped outliers in all the columns by quantile method.
- We then Scaled variables in the data so as to standardize values by using StandardScaler from Sklearn
- We also run Hopkins test and score came out to be > 0.5 so data has cluster tendency.

Step 4. Model Building:

Here we have two methods to build a model-

a) K-Means:

- we calculated optimum number of clusters by:
    - Elbow curve method
    - Silhouette method
- We got 3 as optimum cluster number from it.
- Then using KMeans from sklearn we formed clusters of countries and assigned label to them.

b) Hierarchical clustering:

- In this method we plotted dendrograms for different methods such as single linkage, complete linkage, average linkage. Since complete linkage method was giving better insights we chose that method and we chose 4 as optimum clusters.
- Then using cut_tree from scipy we formed clusters of countries and assigned label to them.

Step 5. Final Analysis and inferences:

a) For K-means clustering:

- We used scatter plot to see how different cluster groups are located.
- We plotted boxplot for (income, gdpp, child mortality rate) vs clusters groups to get insights which countries/cluster group is need to target.
- We got to know following top 10 countries are in real need which belonged to cluster group 2:

1) 'Burundi'
2) 'Liberia'
3) 'Congo, Dem. Rep.'
4) 'Niger'
5) 'Sierra Leone'
6) 'Madagascar'
7) 'Mozambique'
8) 'Central African Republic'
9) 'Malawi'
10) 'Eritrea'

b) For Hierarchical clustering:

- Here also we plotted scatter plot to see how cluster groups are located.
  Also we plotted boxplot for (income, gdpp, child mortality rate) vs clusters to get insights that which group of clusters are in real need.
- We get to know that following countries are in real need which belongs to cluster group 0:

1) 'Burundi'
2) 'Liberia'
3) 'Congo, Dem. Rep.'
4) 'Niger'
5) 'Sierra Leone'
6) 'Madagascar'
7) 'Mozambique'

8) 'Central African Republic'
9) 'Malawi'
10) 'Eritrea'

Note : As data was not that big so choosing hierarchical method would have been a good choice to go with as it gives better results but in our case both the methods (K-Means and Hierarchical) gives same set of top 10 countries.

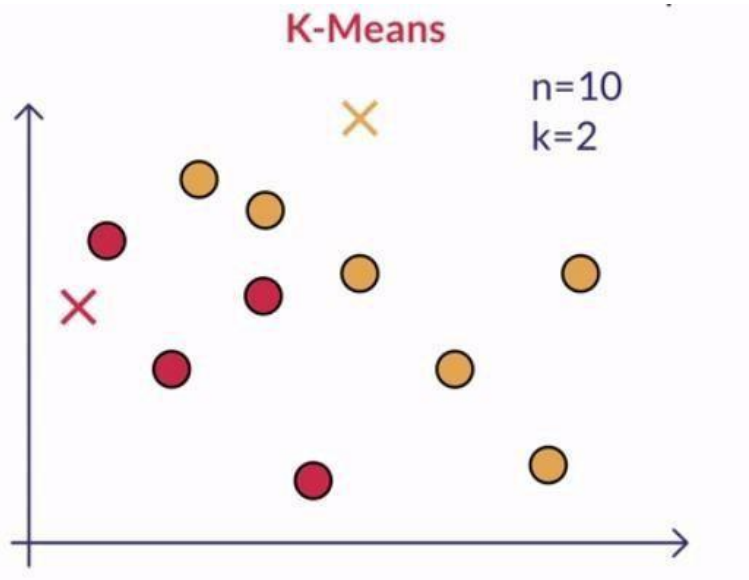
Assignment 2

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

- **<u>K-means clustering:</u>**

- K-means is one of the clustering method. K-Means algorithm is the process of dividing the N data points into K groups or clusters. This is done by assigning each of the data points to their nearest cluster centres based on the Euclidean distance.
- Nearest cluster centres is nothing but a centroid of all the points in the given data sets. Number of clusters needs to be pre decided based on two approaches i.e.
    - Elbow curve method
    - Silhouette score method

- Once we get number of clusters lets say n to be formed we then assign n cluster centres randomly at start. Then we assign all data points to the cluster centres and we get k-clusters based on Euclidean distance. Now we update the position of each of the cluster centres to reflect the mean of each cluster.
- This process continues iteratively till the clusters converge; that is, there are no more changes possible in the position of the cluster centres. At this point, we achieve the n optimal clusters.
- Following image shows 10 data points clustered into 2 clusters red and yellow by K-Means



- **Hierarchical Clustering:**

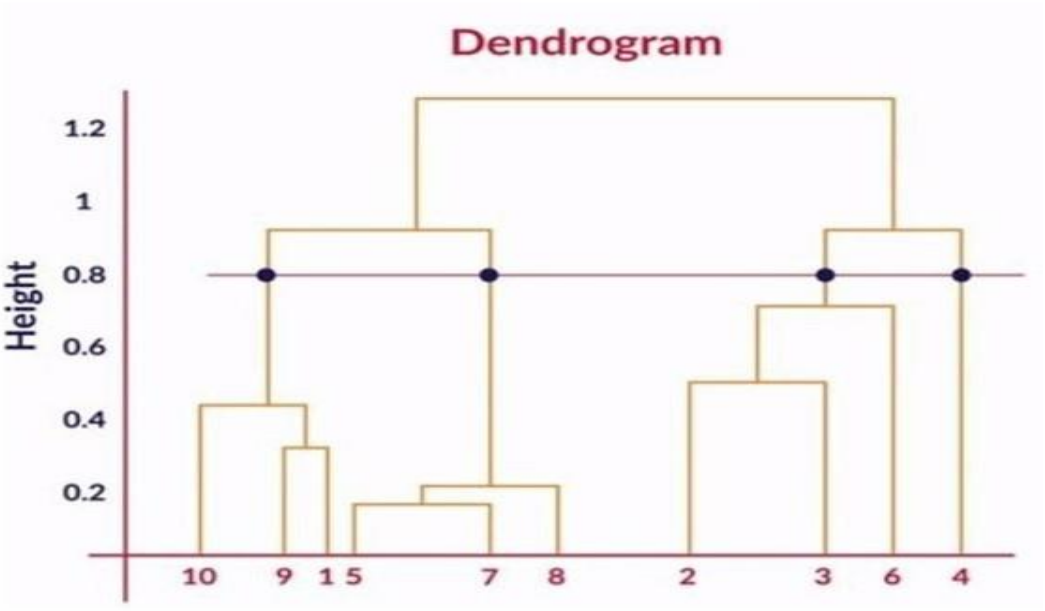
- Hierarchical Clustering is another method of clustering. Hierarchical Clustering proceeds a bit differently from the K-means Clustering method in the following ways :

  Given a set of N items to be clustered

  1. Calculate the NxN distance (similarity) matrix, which calculates the distance of each data point from the other.
  2. Start by assigning each item to its own cluster, so that if you have N items, you now have N clusters, each containing just one item.
  3. Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now you have one less cluster.
  4. Compute distances (similarities) between the new cluster and each of the old clusters.

  5. Repeat steps 3 and 4 until all items are clustered into a single cluster of size N.

- This is how finally you get a tree like structure which is called dendrogram, which shows datapoints grouped together in which clusters at what distance.
- Once we obtain the dendrogram, the clusters can be obtained by cutting the dendrogram at an appropriate level.
- Following image shows dendrogram cut at 0.8 height to form 4 clusters.



Difference between K-Means and Hierarchical clustering

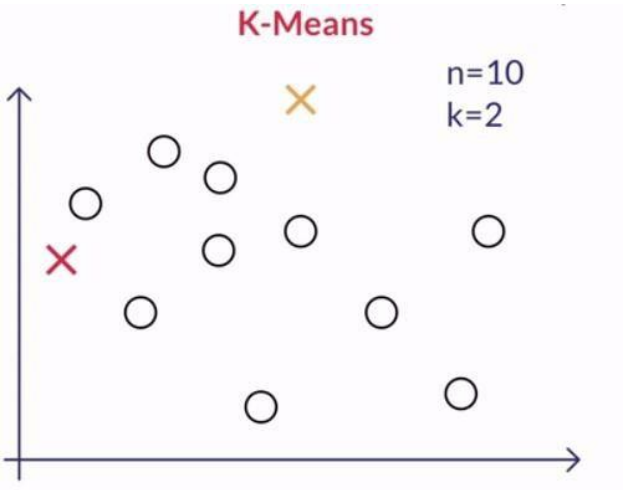
As we can clearly see in hierarchical clustering, you don't need to specify the number of clusters as we do in K-means clustering. This is one of the key differences between K means and Hierarchical Clustering.

- 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).
- Generally, for large datasets, it is preferred to used K-means clustering whereas for smaller ones we use Hierarchical Clustering. The reason for this is that Hierarchical Clustering is computationally expensive. In each iteration, it runs on every cluster that has been formed previously and store them in the memory as well. Therefore, it uses a lot of RAM and if one has limited memory bandwidth, then it becomes a problem to get good clusters.
- Thus, the time complexity of K Means is linear i.e. O(n) while that of hierarchical clustering is quadratic i.e. O(n2).

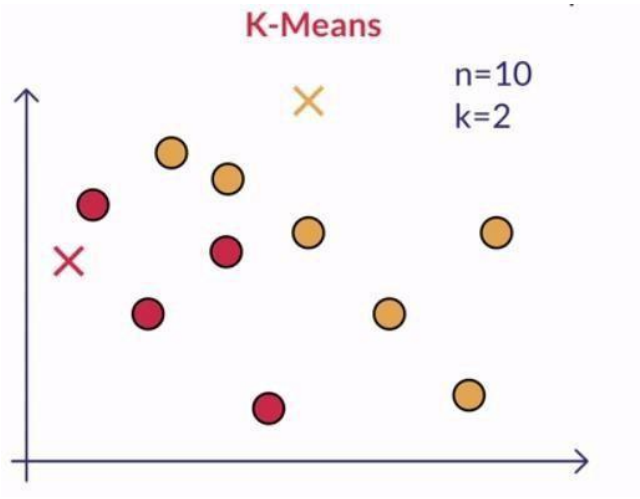
b) Briefly explain the steps of the K-means clustering algorithm.

In K-means algorithm we divide n points in k clusters. Lets take n = 10 and k = 2 for our analysis and understand the steps involved in this process :

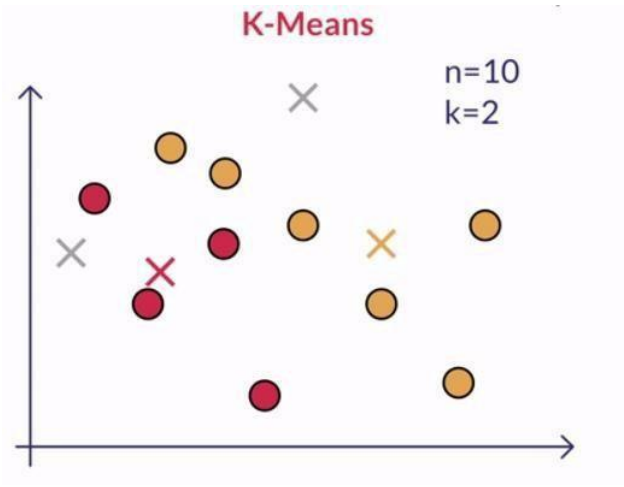
1) Choose K random points as initial clusters (here K=2, n =10). Yellow and red crosses are the two cluster centres chosen randomly for 10 data points.



2) **Assignment Step**: We then assign each of the data points to their nearest cluster centres based on the Euclidean distance. Euclidean distance is the distance calculated from data points to all the cluster centres and the data points are assigned to cluster centres based on minimum distance from the cluster centres. This way points are divided into k clusters here K=2

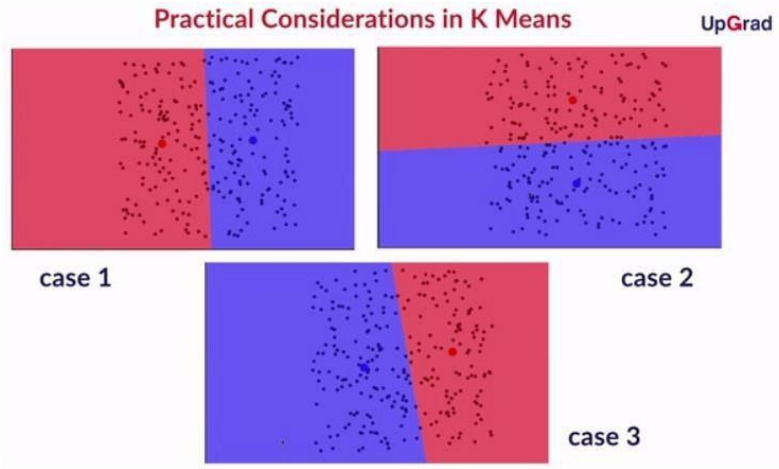


3) **Optimisation Step**: Now mean of cluster points for respective cluster centre is calculated and updated as new cluster centre. This process is goes on till cluster centre converges i.e. there is no further change in cluster centre.



4) Now re-assign all the data points to the different clusters by using the new cluster centres.
5) Keep iterating through step 3 & 4 until there are no further changes possible.

6) Here the choice of initial cluster centres has a impact on the final results as follow:



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

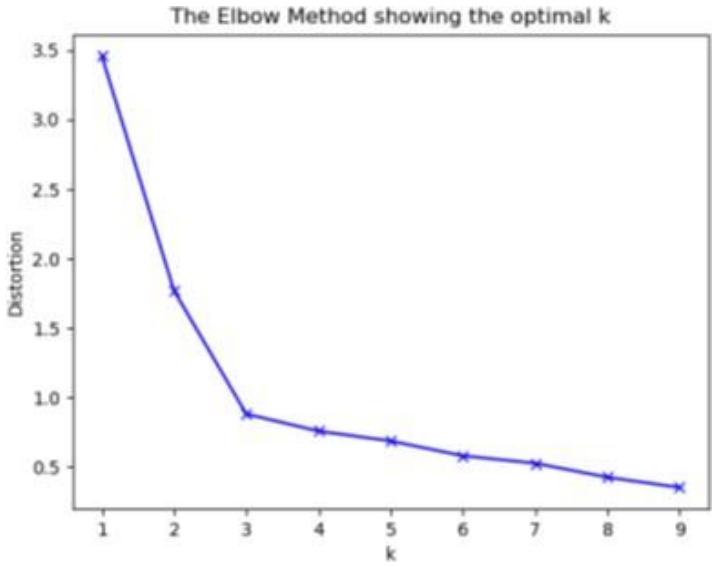
The value of K can be chosen in 2 WAys. Either by using Statistical Analyses or by using the business aspect. Both are explained below:

- **<u>Statistical Analyses</u>**

We can use either the Silhouette score or the elbow curve to find the optimal value of K. Both are described below :

1. Elbow curve method:

- In this method we calculate within-cluster sum of squares (wss) for K clusters.
- WSS is calculated by calculating distance of data points from its cluster centre and we do this for all data points and then you calculated average of the sum by dividing number of points. Then we calculate average across all clusters
- Plot the curve of (wss) for respective cluster numbers (K)
- The location of bend in the plot is considered as optimum number of clusters here 3.



2. Silhouette score method:

Silhouette score is calculated as -

$$\text{silhouette score} = \frac{p - q}{max(p, q)}$$

p = mean distance of points with the nearest clusters of which it is not a part of.

q = mean distance of points with the same clusters of which it is a part of.

- Silhouette score is calculated for all data points and then average silhouette score is calculated for K clusters
- Silhouette score ranges from -1 to 1
- -1 means points are not tightly gathered with its own clusters and they are close to nearest cluster, while 1 means points are tightly gathered with its own clusters and they are not close to nearest cluster.
- So cluster number is selected best on higher silhouette score.

- **<u>Business aspect:</u>**

  - This is mostly oriented for situations where the business understanding of the problem dictates as to how many clusters need to be created.
  - Although we can find number of clusters by statistical methods but these are not always feasible, because in some of the cases number of clusters obtained maybe 2, but business needs more cluster groups so as to improve

focus on specific customers or values.

- So in these cases we need to create clusters according to Business aspect, irrespective of whether the statistical analyses give a different result.

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

- In k means clusters are formed by assigning data points to specific clusters centre based on Euclidean distance between the cluster centres and data points.

- In hierarchical clustering clusters are formed by joining data points based on their distance from other data points.

- So in both the cases it is very much important to have a standard scale for values of different variables as clusters are formed based on distance. So distance may vary uncertainly if they are not in proper scale and clusters formed may incorrect.

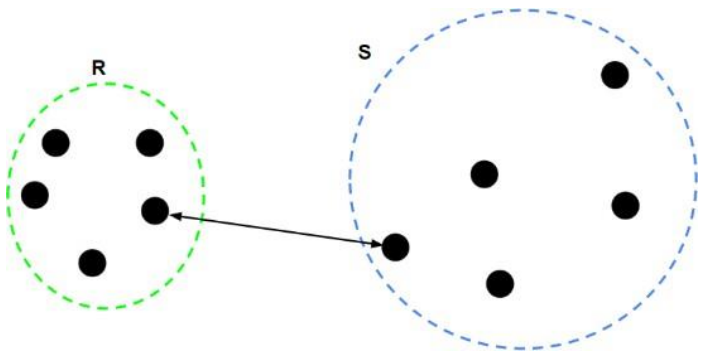
- Standardisation is the process of converting all the columns in any dataset to a comparable scale before applying the method of clustering. This is done to offset the effect of very large-scale variables on those having low scale.

- Suppose we have 3 variables namely income, cars owned, flats owned. So when we calculate Euclidean distance income will dominate other two variables easily because it has greater values and clusters formed will be redundant. So we will scale all variables using standardization or normalization method so that they all receive equal weightage.

e) Explain the different linkages used in Hierarchical Clustering.

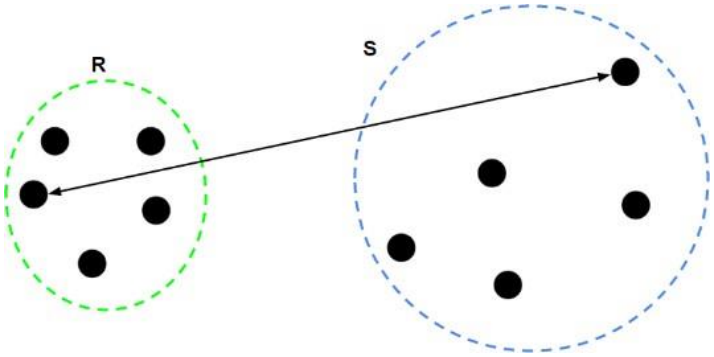
In hierarchical clustering clusters are formed by calculating distance of each points with all other points and then grouping data points which are close to each other. After forming clusters, distance between two clusters can be represented in dendrogram's height axis in following way:

1. Single linkage:
   In this method distance between two clusters is measured as shortest distance



between the points in the clusters.

2. Complete linkage:
   In this method distance between two clusters is measured as longest distance between the points in the clusters.



3. Average linkage:
   In this method distance between two clusters is measured as average distance between the all points in the clusters.