CLUSTERING CASE STUDY KEY NOTES

Sabyasachi Parida

Question 1: Summary and Steps taken.
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: The objective of the Assignment was following:

I. categorize the countries using some socio-economic and health factors that determine the overall development of the country
II. Determine the top 5 countries which are in direst need of aid.

In order to do that I followed the following steps and got the following inferences to reach the conclusion that I will be mentioning in the conclusion section.

EDA

Observation on Data distribution.

I. Child mortality and export are right skewed distribution.
II. Life expectancy is left skewed.
III. Child mortality, Income, total fertility and GDPP doesn't seem to be following normal distribution which can be used for cluster profiling later.

After viewing the distribution, it was required to see where most of the data points are present for each feature.

Observation from Plots

I. Most of the values of Life expectancy is above 50
II. GDPP of most of the countries is less than 60000.
III. Inflation of most of the countries lie below 20.

Observation from the plot of Lowest performing nations

IV. Imports and Exports of most of the countries are below 50000.
I. Haiti has the highest child mortality rate and lowest Life expectancy.
II. Myanmar does the least import as well as export.
III. Eritria spends the least in health per person.
IV. Congo has the least income.
V. Nigeria has the highest Inflation.
VI. Burundi has the least GDPP

Observation of features with respect to income.

I. With increase in income child mortality decreases.
II. With increase in income Export, import, health spending per person, Life expectancy, GDPP increases.

Observation of features with respect to Health spend per person

III. Child mortality decreases with increase in spending in health.
IV. Life expectancy increases with increase in spending in health.
V. GDPP also increases with increasing in spending in health.

After performing the EDA, the outlier analysis was performed by plotting box plots for all the features.

Observations and considerations from Outlier Analysis

I. There are outliers for all the variables or Features.
II. We will not be capping Life expectancy, GDPP and child mortality rates as those might be the data for countries which would be in direst need of Aid.
III. As very less data is present, we will not delete any outliers.
IV. Soft capping was performed.

Hopkins Score Check

I. A Hopkins score above 0.7 indicates a clustering tendency at 90% confidence level. Our average score is 0.91 which indicates very good clustering tendencies.

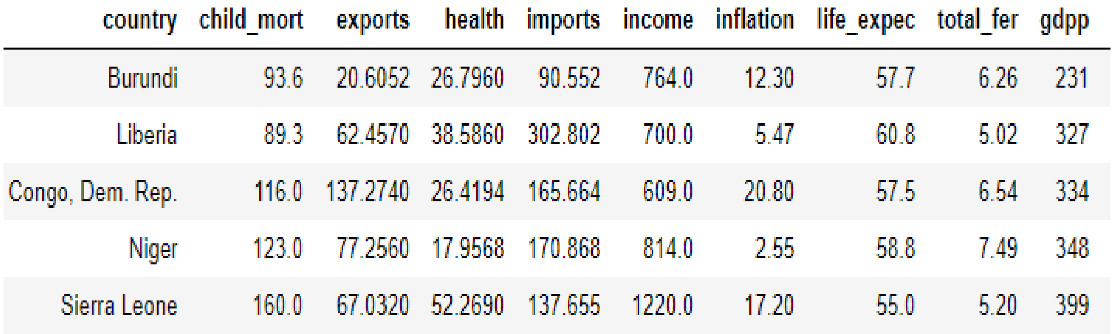
Scaling was performed

As it prevents variables with larger scales from dominating how clusters are defined. It allows all variables to be considered by the algorithm with equal importance.

Then **Hierarchical clustering** was performed.
The Hierarchical clustering was performed before K means in order to visualize the clustering and to find out how many distinctive clusters can be formed.
With Hierarchical Clustering -Single linkage we couldn't find distinctive clusters so went ahead with Complete linkage.

Observations from complete linkage

We can Cut the dendrogram at 8 i.e. for 4 clusters, as we can visualize 4 distinctive clusters.
After Plotting and observing the data obtained from Hierarchical clustering on the basis of cluster labels and then sorting on the basis of low GDPP ,low Income and high child mortality rate we obtained the following result .



country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
Burundi	93.6	20.6052	26.7960	90.552	764.0	12.30	57.7	6.26	231
Liberia	89.3	62.4570	38.5860	302.802	700.0	5.47	60.8	5.02	327
Congo, Dem. Rep.	116.0	137.2740	26.4194	165.664	609.0	20.80	57.5	6.54	334
Niger	123.0	77.2560	17.9568	170.868	814.0	2.55	58.8	7.49	348
Sierra Leone	160.0	67.0320	52.2690	137.655	1220.0	17.20	55.0	5.20	399

Then K means Clustering was performed.

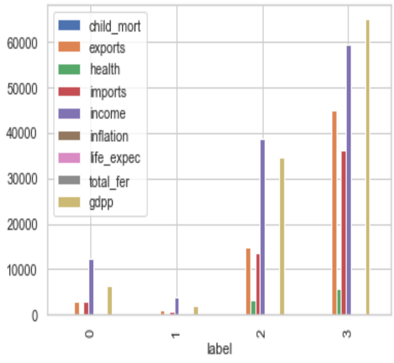
From the Inertia or the Elbow curve and the silhouette score it was statistically recommended to take 3 clusters. But 3 clusters would be too less clusters for the 167 countries present hence we tried with n=4 and n=5 and took the best.

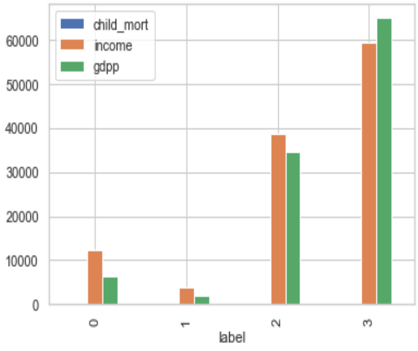
Observations

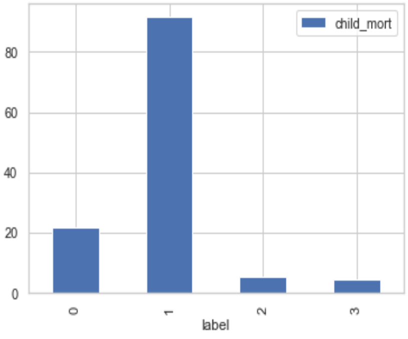
Cluster with n=4 gave better results as it was observed that when n=5, a lot of data points from cluster 0 and 4 seemed to be very nearby and seemed like they should belong to same cluster.

Cluster Profiling and Conclusion

In order to do cluster profiling few bar graphs were plotted .

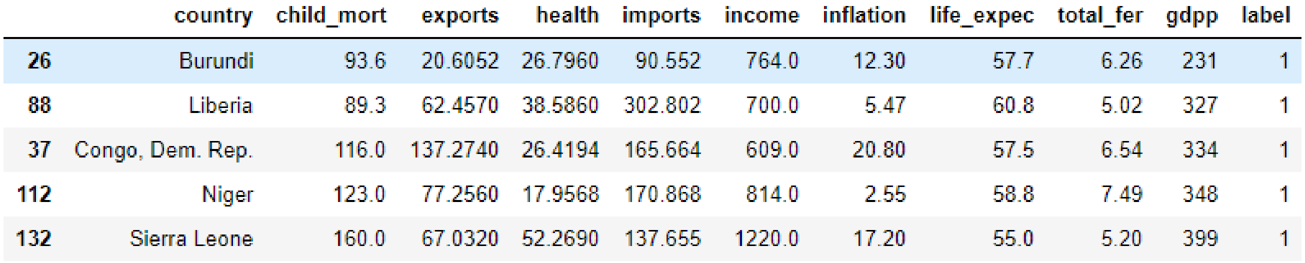






After observing the above plots I was able to come to conclusion that cluster1 is in direst need of the aid as it has lowest GDPP, Lowest Income and highest child mortality rate.

The following are the 5 countries which are in the direst need of the aid.



	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	label
26	Burundi	93.6	20.6052	26.7960	90.552	764.0	12.30	57.7	6.26	231	1
88	Liberia	89.3	62.4570	38.5860	302.802	700.0	5.47	60.8	5.02	327	1
37	Congo, Dem. Rep.	116.0	137.2740	26.4194	165.664	609.0	20.80	57.5	6.54	334	1
112	Niger	123.0	77.2560	17.9568	170.868	814.0	2.55	58.8	7.49	348	1
132	Sierra Leone	160.0	67.0320	52.2690	137.655	1220.0	17.20	55.0	5.20	399	1

From both hierarchical as well as K means we got the same set of countries.

Question 2: Clustering

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

Ans: Clustering is the task of dividing datapoints into a number of groups such that the data points in the same group are more similar than that of another group. In other words, clustering is the process of segregating data and putting them into clusters where all the data points in a specified cluster have similar characteristics.

Comparison between K means and Hierarchical Clustering

Sl no.	K - Means Clustering	Hierarchical Clustering
1	K-Means uses a prespecified number of clusters. This method assigns record to each cluster to find the mutually exclusive cluster based on distance.	Hierarchical methods can be either divisive or agglomerative
2	K Means clustering needed advance knowledge of K i.e. no. of clusters one wants to divide your data.	In hierarchical clustering one can stop at any number of clusters when the person finds it appropriate by interpreting the dendrogram.
3	One can use mean as a cluster centre to represent each cluster.	Agglomerative methods begin with 'n' clusters and sequentially combine similar clusters until only one cluster is obtained.
4	This method usually uses lesser resource and is less computationally intensive and hence can be used for larger data set	This method uses high computational resource as it is a divisive method which work in the opposite direction, beginning with one cluster. It can be implemented in top - bottom or bottom- top method
5	In K Means clustering, since one starts with random choice of centroids, the results are obtained by running the algorithm many times. So, the output depends on the initial choice of the centroids.	The decision of merging two clusters is taken on the basis of closeness of these clusters. There are multiple metrics for deciding the closeness of two clusters like Euclidian distance, Manhattan distance etc.
6	In K Means clustering, since we start with random choice of clusters, the results produced by running the algorithm multiple times might differ.	Results are reproducible in Hierarchical clustering
7	time complexity of K Means is linear i.e. O(n)	time complexity of hierarchical clustering is quadratic i.e. O(n2).
8	K- means clustering a simply a division of the set of data objects into non- overlapping subsets (clusters) such that each data object is in exactly one subset).	A hierarchical clustering is a set of nested clusters that are arranged as a tree (Dendrogram).
9	K Means is found to work well when the shape of the clusters is hyper spherical (like circle in 2D, sphere in 3D)	This method is doesn't work well as compared to K means when the shape of cluster is hyper -spherical

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

Ans: K-means is an approach for partitioning dataset into K distinct, non-overlapping clusters.
The algorithm works as follows:

1. First, we initialize k points, called means, randomly.
2. We categorize each item to its closest mean and we update the mean's coordinates, which are the averages of the items categorized in that mean so far.
3. We repeat the process for a given number of iterations and at the end, we have our clusters.

Mathematical Explanation: -

Let's consider a Dataset where
N is the total no of observations
K is the total number of desired clusters
Ck denotes the set containing the observations in *kth* cluster

The following steps are Applied to create K clusters:

1. **Initialisation**: The cluster centres for each of the *K* clusters are randomly picked. These can either be from any of the *N* observations or totally a different point.

2. **Assignment**: Each observation *n* is assigned to the cluster whose cluster center is the closest to it. The closest cluster center is found using the squared Euclidean distance. Equation for assignment:

$$C_k = Argmin \left\{ \sum_{k=1}^{K} \sum_{i=1}^{N} \left(x_i - \mu_k \right)^2 \right\}$$

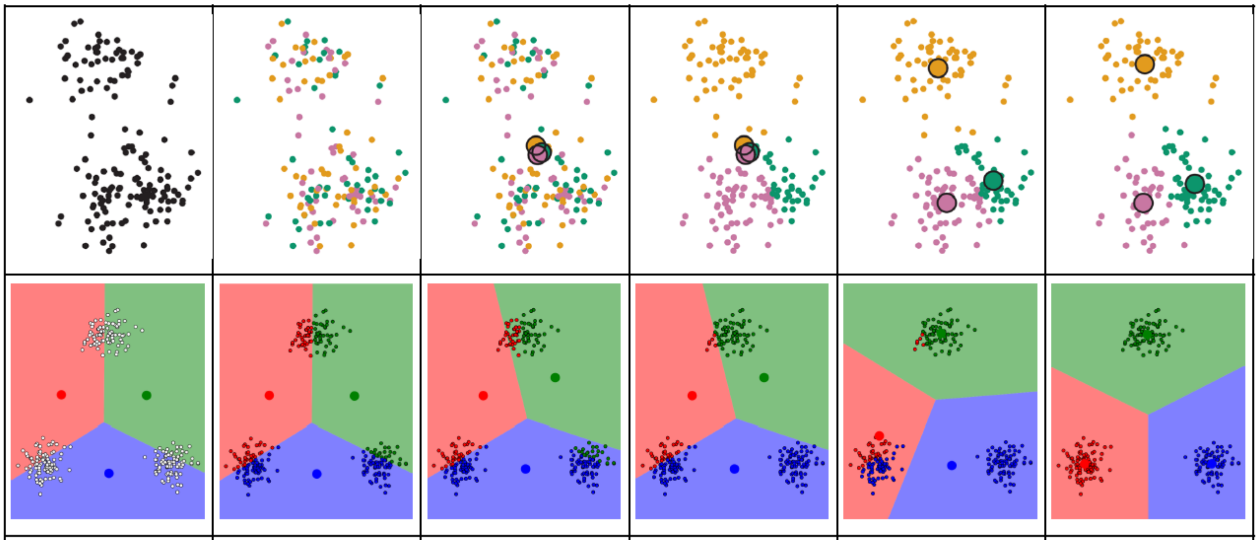
   The observations are to be assigned in such a way that it satisfies the below conditions:
   - C1 U C2 U ... *Ck* = {1, 2, ..., *N*} i.e. each observation belongs to at least one of the *K* clusters
   - *Ck ∩ Ck' = ϕ where k! = k' i.e. no observation belongs to more than one cluster.*

3. **Optimisation:**
   For each of the clusters, the cluster center is computed *K* such that the *kth* cluster center is the vector of the *p* feature means for the observations in the *kth* cluster.

$$\mu_k = \frac{1}{\left| C_k \right|} \sum_{i \in C_k} x_i$$

4. **Iteration:**
   The process of assignment and optimisation is repeated until there is no change in the clusters or possibly until the algorithm converges.

The following is graphical representation of the same

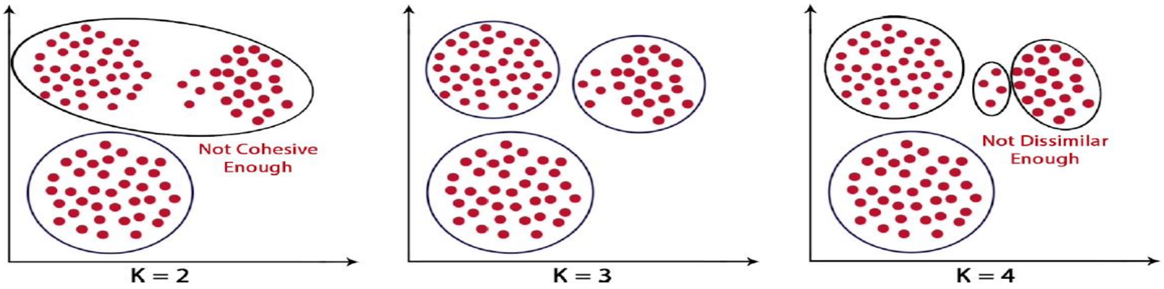


Initialization	Iteration 1, Step 1	Iteration 1, Step 2	Iteration 2, Step 1	Iteration 2, Step 2	Final Result
K-means algorithm with K=3. The 3 cluster center are randomly selected	Each observation is assigned to the nearest cluster center	Cluster centers are computed using squared euclidean distance	Each observation is assigned to the nearest cluster center	Cluster centers are again computed leading to new centers	After N iterations the result is found when the clusters stop changing

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: Choosing the **K** value in case of **K-means** algorithm has the utmost importance as **lower values** of K may lead to **less cohesive** clusters and **higher value** of K may lead to clusters which are **not dissimilar enough**.

E.g.-



The following are the methods that we usually follow to find the number of clusters:

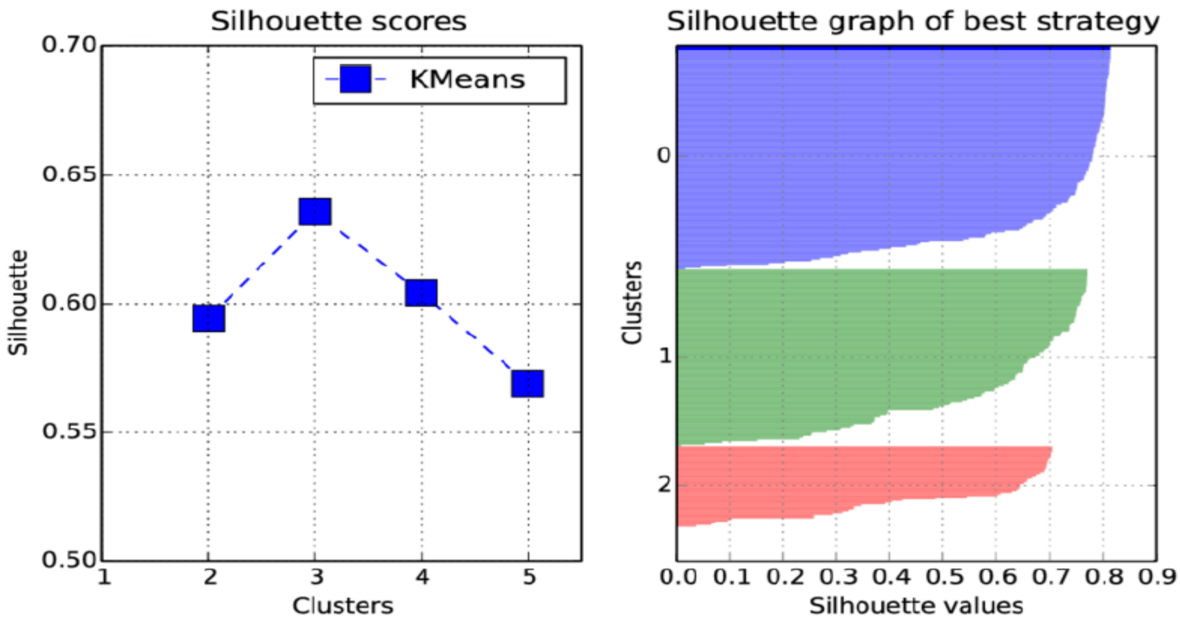
Silhouette Analysis:

It is an approach to chose the value of K where the silhouette coefficient is a measure of how similar a datapoint is to its own cluster compared to other clusters.

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

- **p** is the mean distance to the points in the nearest cluster that the data point is not a part of.
- **q** is the mean intra-cluster distance to all the points in its own cluster.
- The value of the silhouette score range lies between -1 to 1.
- A score closer to 1 indicates that the data point is very similar to other data points in the cluster.
- A score closer to -1 indicates that the data point is not similar to the data points in its cluster.

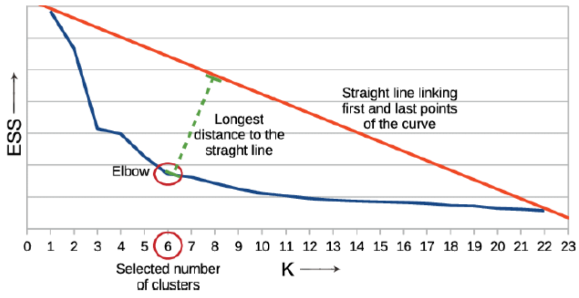
The average ***S*(*i*)** over all the points of a cluster measures how tightly grouped all the points in the cluster are. Thus, the average ***S*(*i*)** over all data of the entire dataset is a measure of how appropriately the data has been clustered. If there are too many or too few clusters, some of the clusters will typically display much narrower silhouettes than the rest.



Elbow Curve Method:

The elbow curve is an approach for finding the appropriate number of clusters in a dataset by interpreting and validating the consistency within a cluster.

The idea of the elbow method is to run K-Means clustering on the dataset for a range of values of *K* while calculating the Sum of Squared Errors (SSE) for each value of *K* . On plotting a line chart of the SSE for each value of *K*, The elbow on the arm is the value of *K* , that is the best.



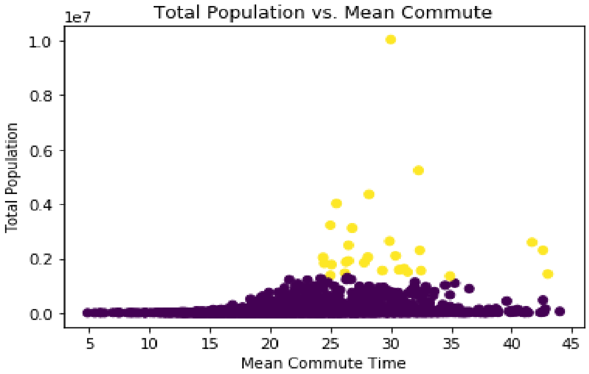
It tries to have SSE Low, but SSE tends to decrease towards 0 with increase in K. The goal is to choose a value of K which has still low SSE and the elbow usually represents that point. If the data is not well clustered the Elbow curve doesn't work well in such cases, we should go for other approaches such as Silhouette analysis etc.

d. Explain the necessity for scaling/standardization before performing Clustering.

Ans: Standardization is the process of rescaling the values of the variables in our data set so that they share a common scale. This is often performed as a preprocessing step. Standardization may be important for us when we are working with data where each variable has a different unit (e.g.- inch, meters) or where the scales of each of our variables are very different from one another (e.g., 0-1 vs 0-1000).
The reason this importance is particularly high in cluster analysis is because groups are defined based on the distance between points.

When we are working with data where each variable means something different, (e.g., age and weight) the fields are not directly comparable. One year is not equivalent to one pound, and may or may not have the same level of importance in sorting a group of records. In a situation where one field has a much greater range of value than another (because the field with the wider range of values likely has greater distances between values), it may end up being the primary driver of what defines clusters. Standardization helps us to make the relative weight of each variable equal by converting each variable to a unitless measure or relative distance.

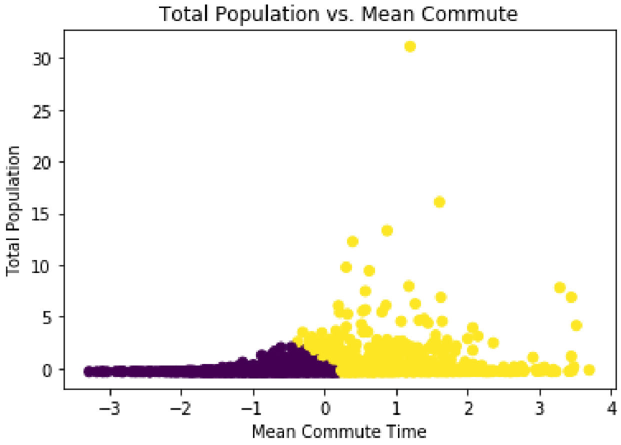
E.g. The following is the example of Total vs mean commute time using (https://www.kaggle.com/muonneutrino/us-census-demographic-data) This dataset includes different demographic variables for counties in the United States, including population, race.



In the above figure when we performed cluster analysis on Total Population and Mean commute time. We would like to use these two variables to split all of the counties into two groups. The units (number of people vs. minutes) and the range of values (85 - 10038388 people vs. 5 - 45 minutes) of these attributes are very different. It is also worth noting that Total Population is a sum, and Mean Commute Time is an average.

When we create clusters with the raw data, we see that Total Population is the primary driver of dividing these two groups. There is an apparent population threshold used to divide the data into two clusters as shown above.

But after standardization both Total Population and Mean Commute seem to have an influence on how the clusters are defined.



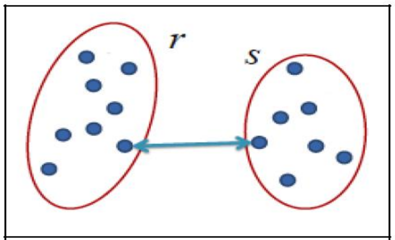
Hence, we can conclude that Standardization prevents variables with larger scales from dominating how clusters are defined. It allows all variables to be considered by the algorithm with equal importance.

e. Explain the different linkages used in Hierarchical Clustering.

Ans: The process of Hierarchical clustering involves either clustering sub-clusters into larger clusters in bottom up manner or dividing a larger cluster into smaller sub cluster in a top down manner.
In case of both the approaches, the distance between each cluster needs to be computed before any clustering is performed. In other words, it is required to determine the proximity matrix using a distance function.

This proximity or linkage between any 2 clusters can be measured by the following:

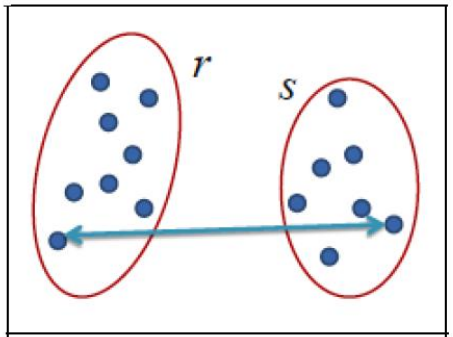
Single linkage: The distance between two clusters is defined as the shortest distance between two points in each cluster.



$$L(r,s) = min\left(D\left(x_{ri}, x_{sj}\right)\right)$$

It suffers from chaining. In order to merge two groups, only one pair of points is required to be near to each other, irrespective of the location of all other points. Therefore, clusters can be too spread out, and may not compact enough.

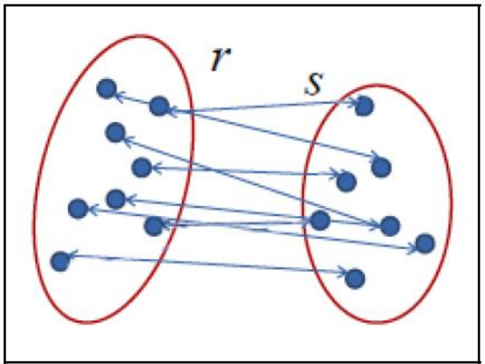
Complete Linkage: The distance between two clusters is defined as the longest distance between two points in each cluster.



$$L(r,s) = max\left(D\left(x_{ri}, x_{sj}\right)\right)$$

It suffers from crowding. The score is based on the worst-case dissimilarity between pairs, making a point to be closer to points in other clusters than to points in its own cluster. Thus, clusters are compact, but not far enough apart.

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



$$L(r,s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} D\left(x_{ri}, x_{sj}\right)$$

In this case it tries to strike a balance between the two by using mean pairwise dissimilarities. Hence the clusters tend to be compact as well as far apart.