Assignment Summary :

Question 1: **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 assignment focused on to categorise the countries using some socio-economic and health factors that determine the overall development of the country and based on this data an organisation named HELP International will make decision to fund $ 10 million to countries that are in the direst need of aid.

So for this problem, our basic approach was to first categorize the country under different labels. This can be solved by using Clustering algorithm and for this assignment we have used KMeans clustering algorithm as well as Hierarchical clustering Algorithm.

We started with reading the data and understanding the structure and data type of each column and then converting each column values to relevant to data for e.g. converted exports, imports and health from percentage of gdpp to actual values. After processing the data we performed EDA to understand the relation between different features and for this we have used pair plot and heat map. Later we treated the outliers and by use of Hopkin Statistics concluded that since the score is more than 80%, we can proceed with clustering as data is appropriate for cluster formation. Scaling was applied on all numerical fields to give equal weightage to each feature.

Then we started with Modelling with KMeans algorithm. We use silhouette and Elbow curve together to find the value of K (number of clusters). We selected k = 3 for our further analysis. We found the cluster labels for each record and added that to the main data set. Using cluster profiling, we found out the characteristics of each clusters, and we concluded that cluster ID =2 is the clusters, which have list of countries that may require immediate AID and after sorting data we found top 5 countries that need immediate help.

To confirm our results we have also used Hierarchical clustering algorithms, we used the same scaled data that we used in Kmeans algorithms and build a dendogram using complete linkage and from the dendogram we selected K =3 for clustering. We build model based on number of cluster =3 and assign labels to each records. Using cluster profiling, we found out the characteristics of each clusters and we concluded that cluster ID =0 is the clusters, which have list of countries that may require immediate AID and after sorting data we found top 5 countries that need immediate help.

From above analysis, results from both the algorithms were same but from above analysis following points can be concluded:

- Hierarchical clustering 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 clustering 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.

Assignment Subjective Questions:

a) 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'.
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.

Main differences between K means and Hierarchical Clustering are:

S No.	k-means Clustering	Hierarchical Clustering
1	k-means, using a pre-specified number of clusters, the method assigns records to each cluster to find the mutually exclusive cluster of spherical shape 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 want to divide your data.	In hierarchical clustering one can stop at any number of clusters, one find appropriate by interpreting the dendrogram.
3	One can use median or 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	Methods used are normally less computationally intensive and are suited with very large datasets.	Divisive methods work in the opposite direction, beginning with one cluster that includes all the records and Hierarchical methods are especially useful when the target is to arrange the clusters into a natural hierarchy.

5	In K Means clustering, since one start with random choice of clusters, the results produced by running the algorithm many times may differ.	In Hierarchical Clustering, results are reproducible in Hierarchical clustering
6	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.
7	K Means clustering is found to work well when the structure of the clusters is hyper spherical (like circle in 2D, sphere in 3D).	Hierarchical clustering don't work as well as, k means when the shape of the clusters is hyper spherical.
	Advantages:	**Advantages:**
1	Convergence is guaranteed.	Ease of handling of any forms of similarity or distance.
2	Specialized to clusters of different sizes and shapes.	Consequently, applicability to any attributes types.
	Disadvantages:	**Disadvantage:**
1	K-Value is difficult to predict	Hierarchical clustering requires the computation and storage of an n×n distance matrix. For very large datasets, this can be expensive and slow
2	Didn't work well with global cluster.	

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

Ans: k-means is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters. The main idea is to define k centres, one for each cluster. These centres should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centre. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centre. A loop has been generated. As a result of this loop we may notice that the k centres change their location step by step until no more changes are done or in other words centres do not move any more. Finally, this algorithm aims at minimizing an objective function knows as squared error function given by:

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} \left(\left\| x_i - v_j \right\| \right)^2$$

where,

'$||x_i - v_j||$' is the Euclidean distance between x_i and v_j.
'c_i' is the number of data points in i^{th} cluster.

'c' is the number of cluster centers.

Algorithmic steps for k-means clustering

Let X = $\{x_1,x_2,x_3,\ldots\ldots\ldots,x_n\}$ be the set of data points and V = $\{v_1,v_2,\ldots\ldots\ldots,v_c\}$ be the set of centres.

1) Randomly select 'c' cluster centres.

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

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

4) Recalculate the new cluster centre using:
Where, 'c_i' represents the number of data points in i^{th} cluster.

$$v_i = (1/c_i) \sum_{j=1}^{c_i} x_i$$

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

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

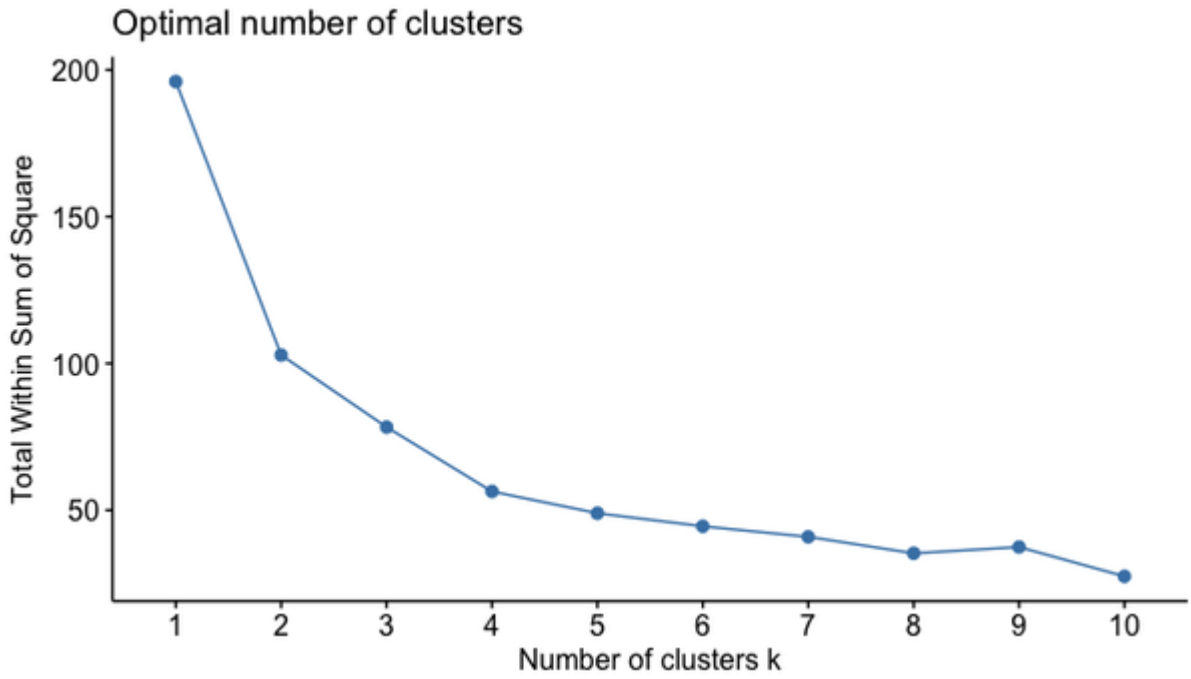
c) How is the value of 'k' chosen in K-means clustering? Explain both the statistical as well as the business aspect of it.

Ans: The basic idea behind k-means consists of defining k clusters such that total **within-cluster variation (or error) is minimum**.

A cluster centre is the representative of its cluster. The squared distance between each point and its cluster centre is the required variation. The aim of k-means clustering is to find these k clusters and their centres while reducing the total error.These methods are:

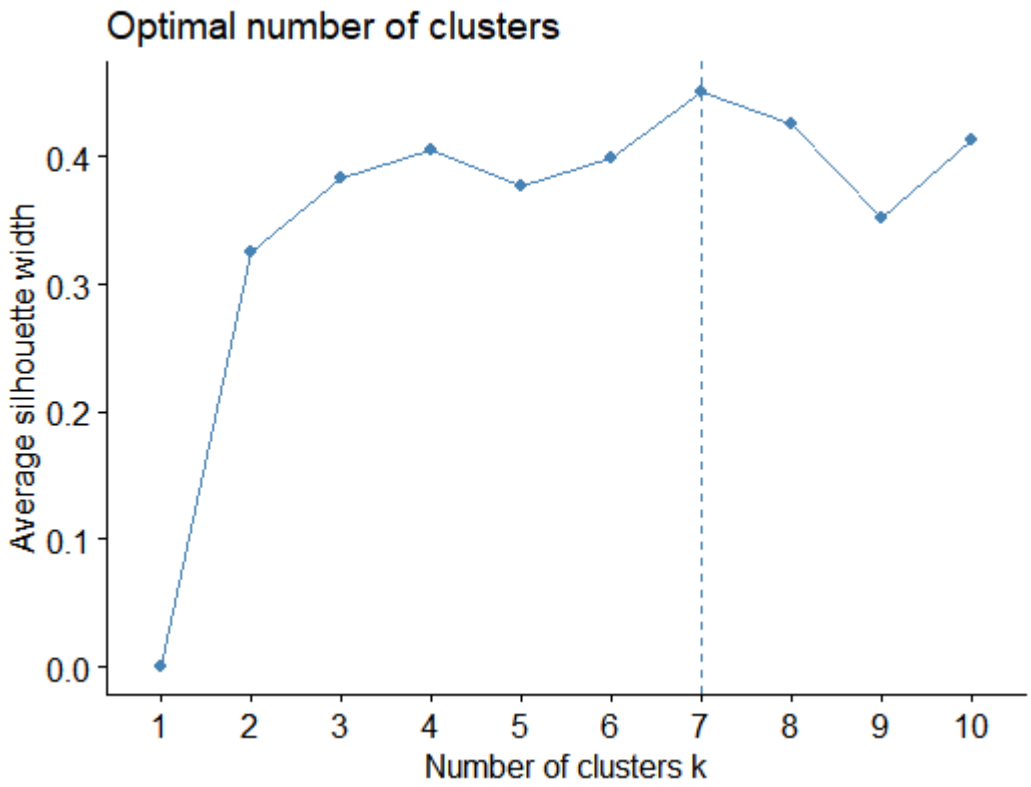
1. **The Elbow Method**
2. **The Silhouette Method**

The Elbow method: To determine the optimal number of clusters, you will need to run the k-means algorithm for different values of k (number of clusters). For each value of k, you will then need to calculate the total within-cluster sum of squares (wss). You can then plot the values of wss on the y-axis and the number of clusters (k) on the x-axis. The optimal number of clusters can be read off the graph at the x-axis.Elbow Method: The total within-cluster sum of square (wss) measures the compactness of the clustering and it should be minimum.



This is typical Elbow Plot that is drawn between Number of clusters(k) v/s Total Within Sum of Squares.

Silhouette Method : To determine the optimal number of clusters, you will need to measure the quality of the clusters that were created. This value determines how closely each data point is to the centroid of its cluster. A high average silhouette coefficient indicates successful clusters. This method checks the silhouette coefficient for different values of k. The optimal number of clusters is, therefore, the maximised silhouette value for the data set.
This method helps in measuring the quality of the clustering i.e. how well an object lies within its cluster.



A good silhouette width indicates good clustering. Also it shows us the optimal number of clusters to be used.

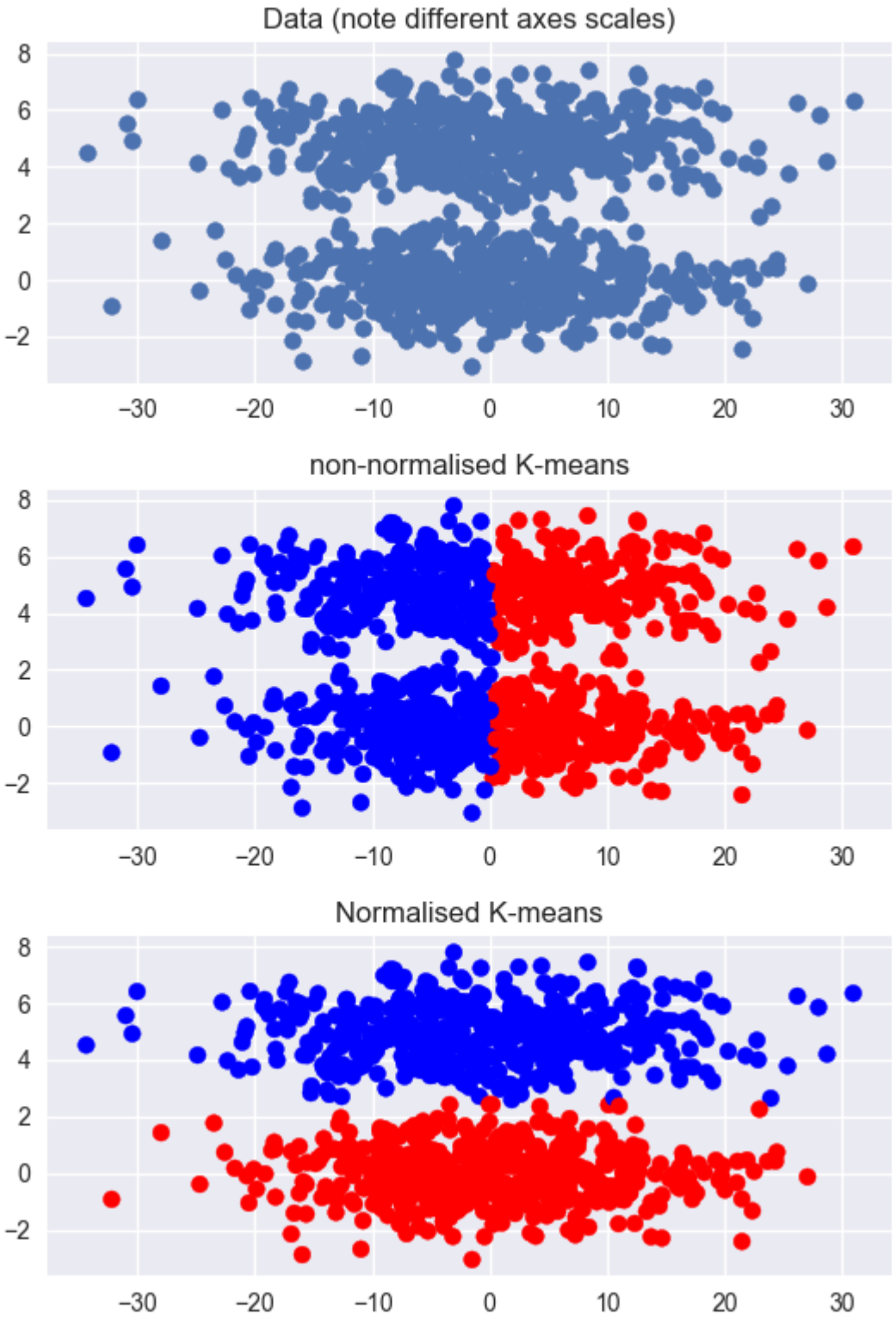
Business Aspect : from the business perspective, domain knowledge plays a vital role in determining number of clusters . For example we get 5 as the value of clusters from silhouette method but from business perspective we know that we need to have only 3 clusters than we will ultimately go with 3 not 5. Another example is that if Elbow curve return more than 30 clusters that means we will not go with 30 clusters that will be very difficult to bifurcate and build a model. Hence domain knowledge plays vital role in determining the value of clusters.

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

Ans. It **depends on our data**.

If we have attributes with a well-defined meaning. Say, latitude and longitude, then we should not scale your data, because this will cause distortion. If we have mixed numerical data, where each attribute is something entirely different (say, shoe size and weight), has different units attached (lb, tons, m, kg ...) then these values aren't really comparable anyway; The idea is that if different components of data (features) have different scales, then derivatives tend to align along directions with higher variance, which leads to poorer/slower convergence. z-standardizing them is a best-practise to give equal weight to them.

Below is a dataset that has two clear clusters, but the non-clustered dimension is much larger than the clustered dimension (note the different scales on the axes). Clustering on the non-normalised data fails. Clustering on the normalised data works very well.

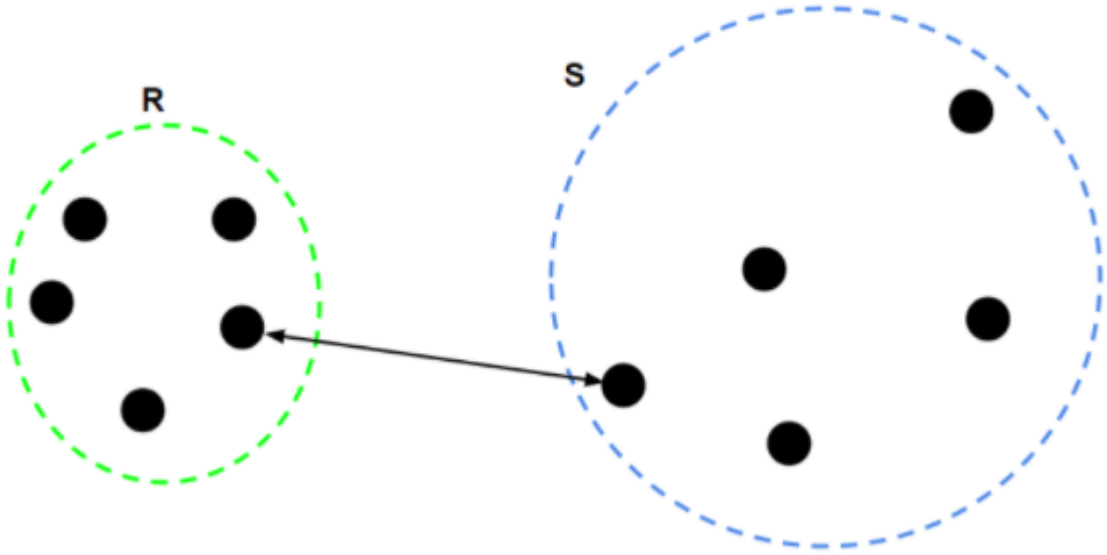


e) Explain the different linkages used in Hierarchical Clustering.

Ans: The process of Hierarchical Clustering involves either clustering sub-clusters(data points in the first iteration) into larger clusters in a bottom-up manner or dividing a larger cluster into smaller sub-clusters in a top-down manner. During both the types of hierarchical clustering, the distance between two sub-clusters needs to be computed. The different types of linkages describe the different approaches to measure the distance between two sub-clusters of data points. The different types of linkages are:-

- **Single linkage**: Also referred to as nearest neighbour or minimum method. This measure defines the distance between two clusters as the minimum distance found between one case from the first cluster and one from the second cluster. For example, if cluster 1 contains cases a and b, and cluster 2 contains cases c, d, and e, then the distance between cluster 1 and cluster 2 would be the smallest distance found between the following pairs of cases: (a, c), (a, d), (a, e), (b, c), (b, d), and (b, e). A concern of using single linkage is that it can sometimes produce chaining amongst the clusters. This means that several clusters may be joined together simply because one of their cases is within close proximity of case from a separate cluster. This problem is specific to single linkage due to the fact that the smallest distance between pairs is the only value taken into consideration. Because the steps in agglomerative hierarchical clustering are irreversible, this chaining effect can have disastrous effects on the cluster solution.

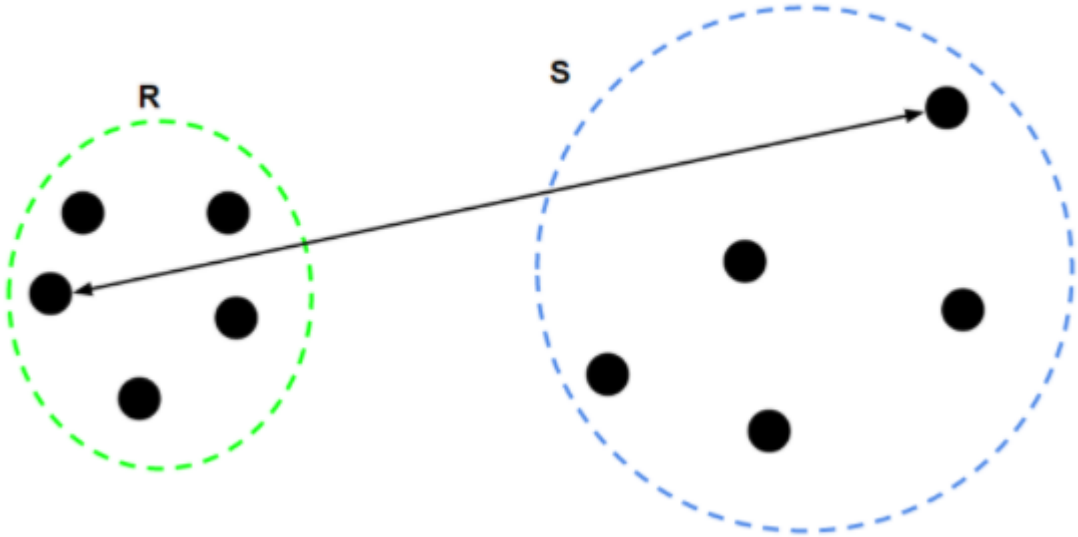
$$L(R,S) = min(D(i,j)), i \epsilon R, j \epsilon S$$



- **Complete linkage**: Also referred to as furthest neighbour or maximum method. This measure is similar to the single linkage measure described above, but instead of searching for the minimum distance between pairs of cases, it considers the furthest

distance between pairs of cases. Although this solves the problem of chaining, it creates another problem. Imagine that in the above example cases a, b, c, and d are within close proximity to one another based upon the pre-established set of variables; however, if case e differs considerably from the rest, then cluster 1 and cluster 2 may no longer be joined together because of the difference in scores between (a, e) and (b, e). In complete linkage, outlying cases prevent close clusters to merge together because the measure of the furthest neighbour exacerbates the effects of outlying data.

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



- **Average linkage:** Also referred to as the Unweighted Average linkage. To overcome the limitations of single and complete linkage, an approach to take an average of the distance values between pairs of cases was considered. This method is supposed to represent a natural compromise between the linkage measures to provide a more accurate evaluation of the distance between clusters. For average linkage, the distances between each case in the first cluster and every case in the second cluster are calculated and then averaged. This means that in the previous example, the distance between cluster 1 and cluster 2 would be the average of all distances between the pairs of cases listed above: (a, c), (a, d), (a, e), (b, c), (b, d), and (b, e). Incorporating information about the variance of the distances renders the average distance value a more accurate reflection of the distance between two clusters of cases. Each linkage measure defines the distance between two clusters in a unique way. The selected linkage measure will have a direct impact on the clustering procedure and the way in which clusters are merged together .

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

where

n_R – Number of data-points in R

n_S – Number of data-points in S