Assignment-based Subjective Questions

Question 1: Assignment Summary

a) 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.

Problem Statement:

HELP International is an international humanitarian NGO that is committed to fighting 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. × After the recent funding programmes, HELP have been able to raise around $ 10 million. Now the CEO of the NGO needs to decide how to use this money strategically and effectively. The significant issues that come while making this decision are mostly related to choosing the countries that are in the direst need of aid.

Approach Followed:

The clustering of the countries was done by starting with the data understanding and data visualization step. The initial inspections and the duplicates were all checked. There were 3 variables whose values were given in the form of percentages. Those values were changed based on the GDP. Then the EDA was performed. The EDA was performed in steps like analysing the individual variable i.e. univariate analysis and then analysing the effect on one variable over the other i.e. bivariate analysis and then the correlation between each variables are checked. The outliers are found using the boxplot and the patterns of each variables are found using the pairplots. Once the outliers are detected using the boxplot visualization, the outliers are capped with the quantile range of 0.95-0.99 based on the variables. Almost 4 variables needed capping and the upper quantile range was capped to

0.95 i.e. 95% .After capping again the boxplot visualization was done to check for the outliers in the data. Once it was done, the Hopkins score was calculated and then scaling was performed. Once the scaling is done the first clustering was done based on K-Means Algorithm. The identical k-value was found using the elbow-curve and silhouette score and was taken as **3.** Then the cluster labels were generated and the label column was attached to the dataframe. Then the cluster profiling was done based on the visualizations it was evident that cluster of ID 0 had the highest child mortality rate and least income and gdpp rates. Then a separate dataframe was created with value only of cluster_id = 0. Then the newly created dataframe was sorted based on the highest child mortality rate and least income and gdpp rate. Then the countries list that requires AID was taken. After this the Hierarchical clustering was performed and the clustering were visualized using the single and complete linkage method. Based on that visualization the number of clusters were taken as **4.** Then the labels were generated and attached to the dataframe and cluster profiling was done. Based on the cluster profiling it was obvious that the cluster of ID 0 had 31 countries in it that needs the AID with high child mortality and low income and gdpp ranges. I choose Hierarchical clustering in a business perspective as the number of countries listed by the hierarchical clustering is less than the k-means. Less the number of countries good would ne quality and AID provided by the organization. The final list of countries that needs AID are Afghanistan, Angola, Benin, Burkina Faso, Burundi, Cameroon, Central African Republic, Chad, Congo, Dem. Rep., Congo, Rep., Cote d'Ivoire, Equatorial Guinea, Gabon, Ghana, Guinea, Guinea-Bissau, Haiti, Iraq, Malawi, Mali, Mauritania, Mozambique, Niger, Nigeria, Sierra Leone, Sudan, Timor-Leste, Togo, Uganda, Yemen, Zambia. The top 5 countries which needs an urgent AID are:

1. Haiti
2. Sierra Leone
3. Chad
4. Central African Republic
5. Mali

Question 2: Clustering

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

Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group than those in other groups. In simple words, the aim is to segregate groups with similar traits and assign them into clusters. Clustering helps identify two qualities of data:

1. Meaningfulness
2. Usefulness

Clustering can be divided into two sub-groups:

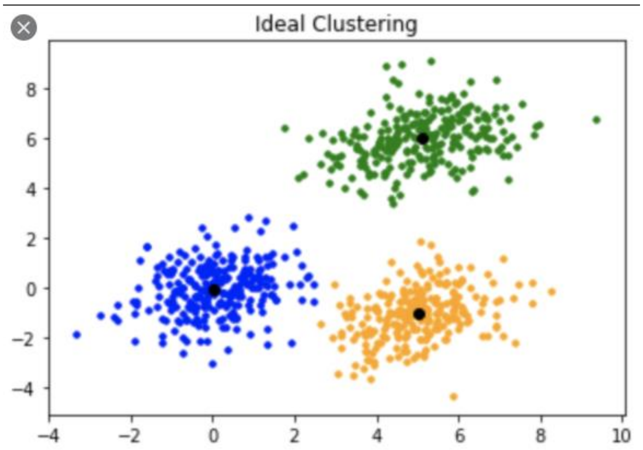
- **Hard Clustering:** In hard clustering, each data point either belongs to a cluster completely or not.
- **Soft Clustering**: In soft clustering, instead of putting each data point into a separate cluster, a probability or likelihood of that data point to be in those clusters is assigned.

K Means Clustering

K means is an iterative clustering algorithm that aims to find local maxima in each iteration. This 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.

The "points" mentioned above are called means, because they hold the mean values of the items categorized in it. To initialize these means, we have a lot of options. An intuitive method is to initialize the means at random items in the data set. Another method is to initialize the means at random values between the boundaries of the data set (if for a feature *x* the items have values in [0, 3], we will initialize the means with values for *x* at [0, 3])

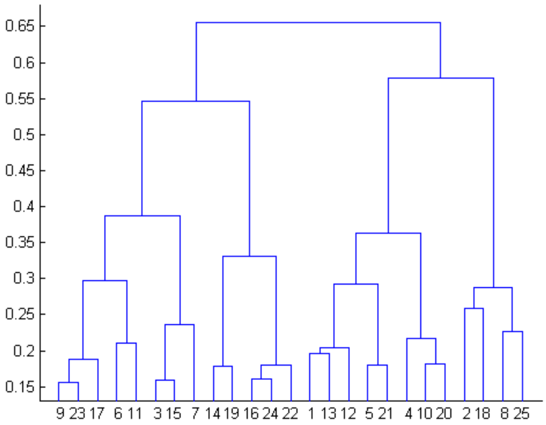


Hierarchical Clustering

Hierarchical clustering determines cluster assignments by building a hierarchy. This is implemented by either a bottom-up or a top-down approach:

- **Agglomerative clustering** is the bottom-up approach. It merges the two points that are the most similar until all points have been merged into a single cluster.
- **Divisive clustering** is the top-down approach. It starts with all points as one cluster and splits the least similar clusters at each step until only single data points remain.

These methods produce a tree-based hierarchy of points called a dendrogram. In hierarchical clustering the number of clusters (*k*) is often predetermined by the user. Clusters are assigned by cutting the dendrogram at a specified depth that results in *k* groups of smaller dendrogram. Hierarchical clustering is a deterministic process, meaning cluster assignments won’t change when you run an algorithm twice on the same input data.



K-Means	Hierarchical
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.
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.
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.
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.
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
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.
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.

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

K-means clustering is a type of unsupervised learning, which is used when you have unlabelled data (i.e., data without defined categories or groups). The goal of the algorithm is to find groups in the data, with the number of groups represented by the variable *K*. The algorithm works iteratively to assign each data point to one of *K* groups based on the features that are provided. Data points are clustered based on feature similarity. The results of the *K*-means clustering algorithm are:

1. The centroids of the *K* clusters, which can be used to label new data
2. Labels for the training data (each data point is assigned to a single cluster)

Steps in the iteration:

The *K*-means clustering algorithm uses iterative refinement to produce a final result. The algorithm inputs are the number of clusters *K* and the data set. The data set is a collection of features for each data point. The algorithms starts with initial estimates for the *K* centroids, which can either be randomly generated or randomly selected from the data set. The algorithm then iterates between two steps:

1. Data assignment step:

Each centroid defines one of the clusters. In this step, each data point is assigned to its nearest centroid, based on the squared Euclidean distance. More formally, if c_i is the collection of centroids in set *C*, then each data point *x* is assigned to a cluster based on where *dist*(·) is the standard (L_2) Euclidean distance. Let the set of data point assignments for each i^{th} cluster centroid be S_i.

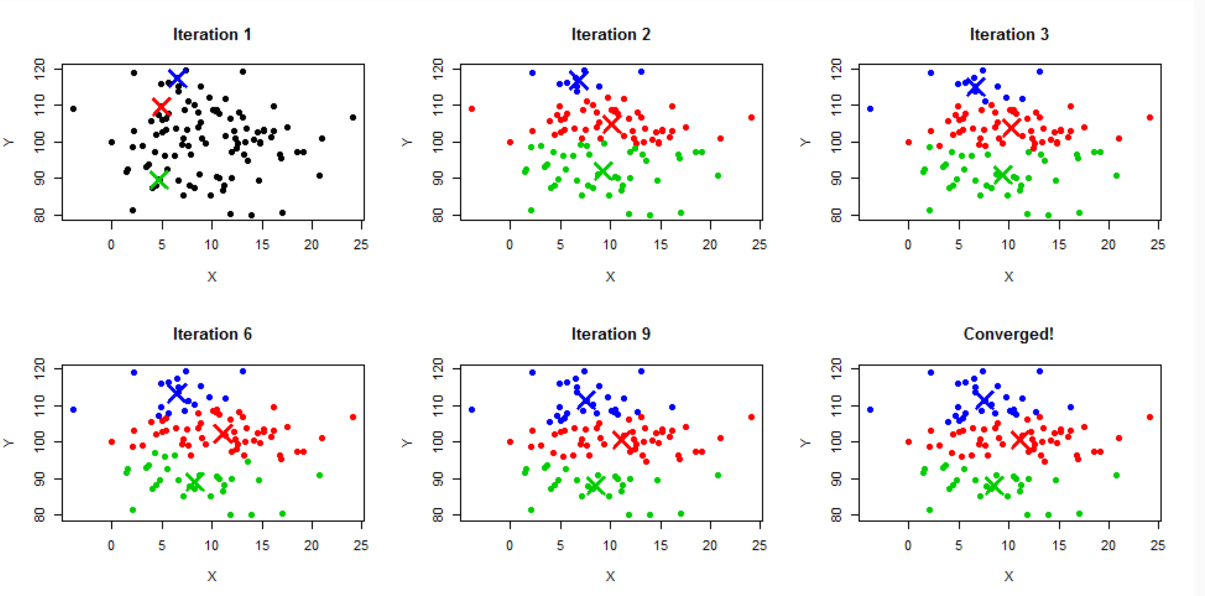
$$\underset{c_i \in C}{\operatorname{argmin}}\ dist(c_i, x)^2$$

2. Centroid update step:

In this step, the centroids are recomputed. This is done by taking the mean of all data points assigned to that centroid's cluster.

$$c_i = \frac{1}{|S_i|}\sum_{x_i \in S_i} x_i$$

The algorithm iterates between steps one and two until a stopping criteria is met (i.e., no data points change clusters, the sum of the distances is minimized, or some maximum number of iterations is reached).This algorithm is guaranteed to converge to a result. The result may be a local optimum (i.e. not necessarily the best possible outcome), meaning that assessing more than one run of the algorithm with randomized starting centroids may give a better outcome.



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

As we all know K-Means Clustering belongs to Partitioning Class (of Clustering), the principle rule is to keep the within-cluster variation or total within-cluster sum of square to be minimum. This actually means that, the variation within the cluster should be minimum. This can be done if we choose the right/optimal number of clusters.

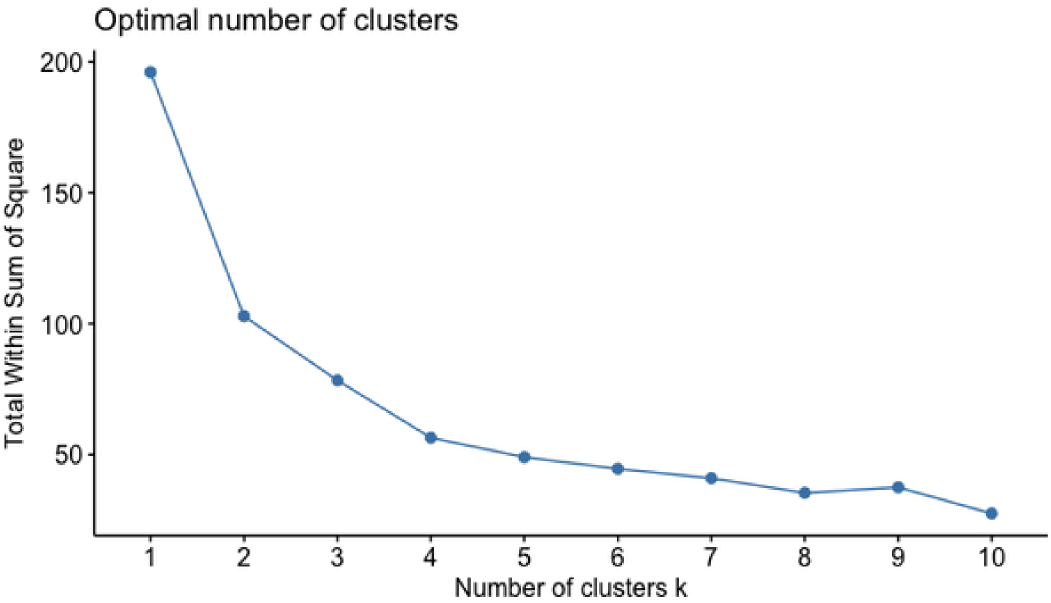
Determining the optimal number of clusters in a data set is a fundamental issue in partitioning clustering, such as k-means clustering, which requires the user to specify the number of clusters k to be generated. There are methods which include business aspect methods and statistical testing methods:

- **Direct methods:** consists of optimizing a criterion, such as the within cluster sums of squares or the average silhouette. The corresponding methods are named *elbow* and *silhouette* methods, respectively.
- **Statistical testing methods:** consists of comparing evidence against null hypothesis. An example is the *gap statistic*.

Elbow method:

The Elbow method looks at the total WSS as a function of the number of clusters: One should choose a number of clusters so that adding another cluster doesn't improve much better the total WSS.The optimal number of clusters can be defined as follow:

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



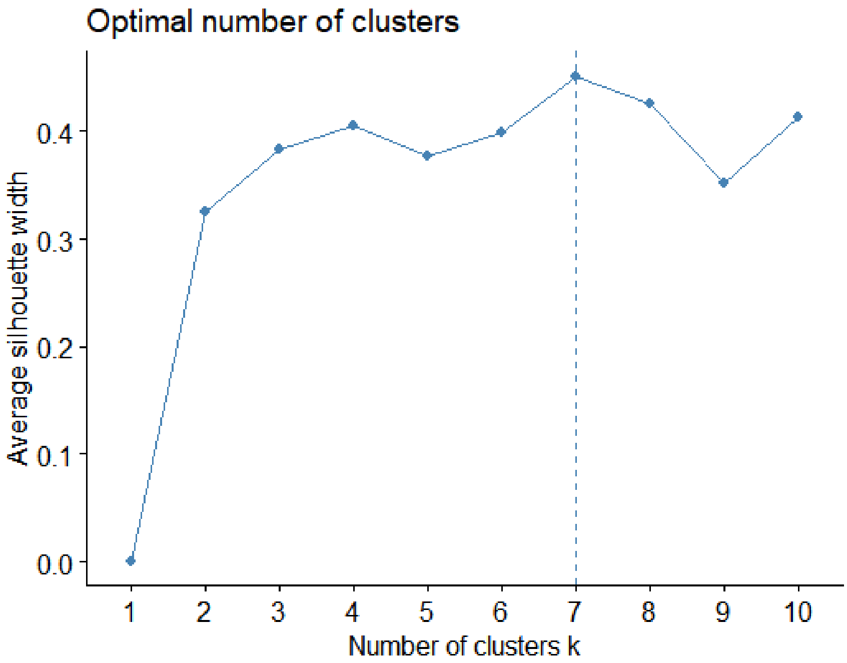
This is typical Elbow Plot that is drawn between Number of clusters(k) v/s Total Within Sum of Squares.

Silhouette Method:

Silhouette Method measures the quality of a clustering. That is, it determines how well each object lies within its cluster. A high average silhouette width indicates a good clustering.

Average silhouette method computes the average silhouette of observations for different values of k. The optimal number of clusters k is the one that maximize the average silhouette over a range of possible values for k. The algorithm is similar to the elbow method and can be computed as follow:

- Compute clustering algorithm (e.g., k-means clustering) for different values of k. For instance, by varying k 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.



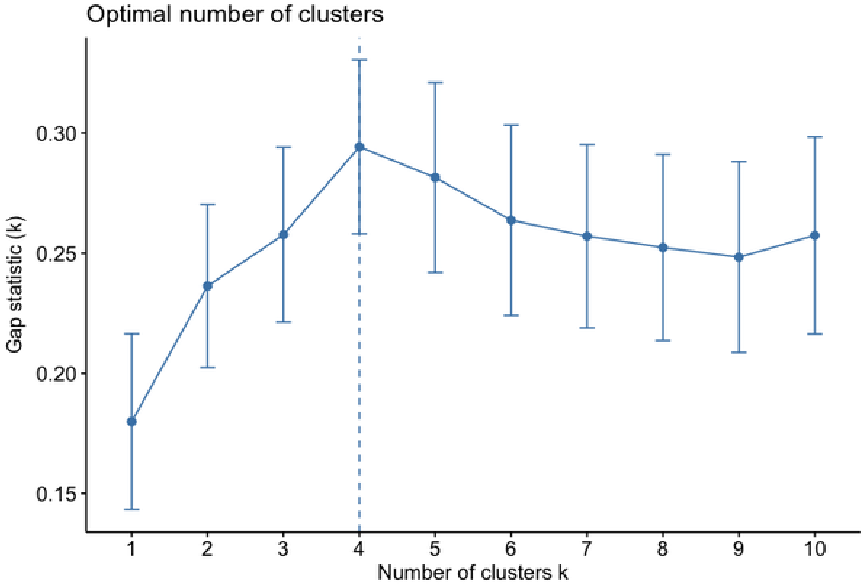
A good silhouette width indicates good clustering. Also it shows you the optimal number of clusters to be used.

Gap statistic method

The gap statistic compares the total within intra-cluster variation for different values of k with their expected values under null reference distribution of the data. The estimate of the optimal clusters will be value that maximize the gap statistic (i.e. that yields the largest gap statistic). This means that the clustering structure is far away from the random uniform distribution of points. The algorithm works as follow:

- Cluster the observed data, varying the number of clusters from k = 1, *kmax*, and compute the corresponding total within intra-cluster variation *Wk*.

- Generate B reference data sets with a random uniform distribution. Cluster each of these reference data sets with varying number of clusters k = 1, *kmax*, and compute the corresponding total within intra-cluster variation *Wkb*.
- Compute the estimated gap statistic as the deviation of the observed *Wk* value from its expected value *Wkb* under the null hypothesis: Gap(k)=1B∑b=1Blog(W∗kb)−log(Wk)Gap(k)=1B∑b=1Blog(Wkb∗)−log(Wk).
- Compute also the standard deviation of the statistics.
- Choose the number of clusters as the smallest value of k such that the gap statistic is within one standard deviation of the gap at k+1: $Gap(k) \geq Gap(k+1) - sk_{+1}$.



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

Feature scaling in machine learning is one of the most critical steps during the pre-processing of data before creating a machine learning model. Scaling can make a difference between a weak machine learning model and a better one. The most common techniques of feature scaling are Normalization and Standardization.

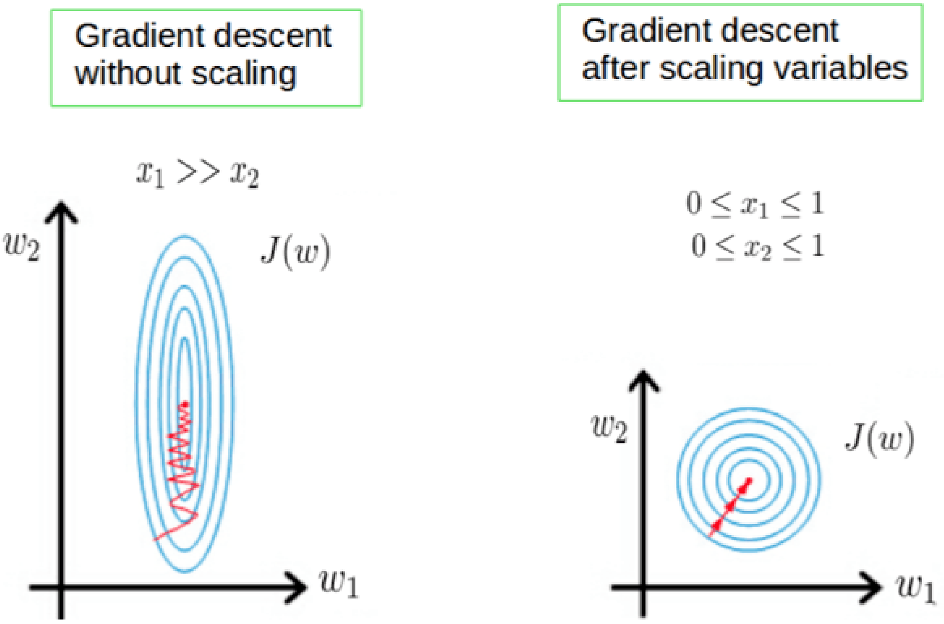
Normalization is used when we want to bound our values between two numbers, typically, between [0, 1] or [-1, 1]. While Standardization transforms the data to have zero mean and a variance of 1, they make our data unit less.

Machine learning algorithm just sees number — if there is a vast difference in the range say few ranging in thousands and few ranging in the tens, and it makes the underlying assumption that higher ranging numbers have superiority of some sort. So these more significant number starts playing a more decisive role while training the model.

The machine learning algorithm works on numbers and does not know what that number represents. A weight of 10 grams and a price of 10 dollars represents completely two different things — which is a no brainer for humans, but for a model as a feature, it treats both as same.

Suppose we have two features of weight and price, as in the below table. The "Weight" cannot have a meaningful comparison with the "Price." So the assumption algorithm makes that since "Weight" > "Price," thus "Weight," is more important than "Price."

So these more significant number starts playing a more decisive role while training the model. Thus feature scaling is needed to bring every feature in the same footing without any upfront importance. Interestingly, if we convert the weight to "Kg," then "Price" becomes dominant.



e) Explain the different linkages used in Hierarchical Clustering.

Clustering tries to find structure in data by creating groupings of data with similar characteristics. The most famous clustering algorithm is likely K-means, but there are a large number of ways to cluster observations. Hierarchical clustering is an alternative class of clustering algorithms that produce 1 to n clusters, where n is the number of observations in the data set. As you go down the hierarchy from 1 cluster (contains all the data) to n clusters (each observation is its own cluster), the clusters become more and more similar (almost always). There are two types of hierarchical clustering: divisive (top-down) and agglomerative (bottom-up).

Divisive

Divisive hierarchical clustering works by starting with 1 cluster containing the entire data set. The observation with the highest average dissimilarity (farthest from the cluster by some metric) is reassigned to its own cluster. Any observations in the old cl uster closer to the new cluster are assigned to the new cluster. This process repeats with the largest cluster until each observation is its own cluster.

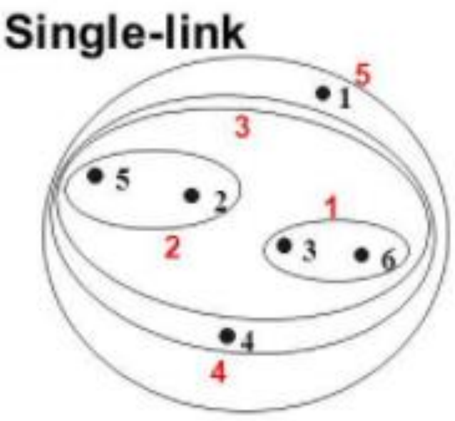
Agglomerative

Agglomerative clustering starts with each observation as its own cluster. The two closest clusters are joined into one cluster. The next closest clusters are grouped together and this process continues until there is only one cluster containing the entire data set.

There are a variety of possible metrics, the 4 most popular: single-linkage, complete-linkage, average-linkage, and centroid-linkage.

Single-Linkage

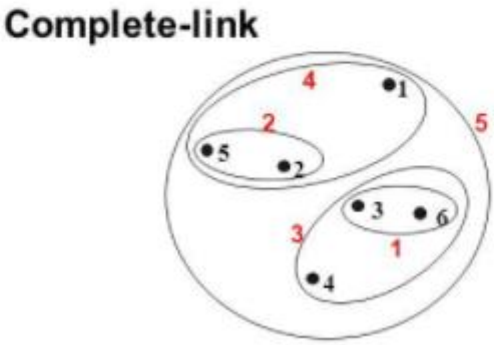
Single-linkage (nearest neighbour) is the shortest distance between a pair of observations in two clusters. It can sometimes produce clusters where observations in different clusters are closer together than to observations within their own clusters. These clusters can ap pear spread-out.



Single-link

Complete-Linkage

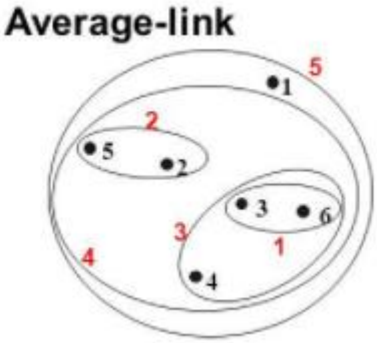
Complete-linkage (farthest neighbour) is where distance is measured between the farthest pair of observations in two clusters. This method usually produces tighter clusters than single-linkage, but these tight clusters can end up very close together. Along with average-linkage, it is one of the more popular distance metrics.



Complete-link

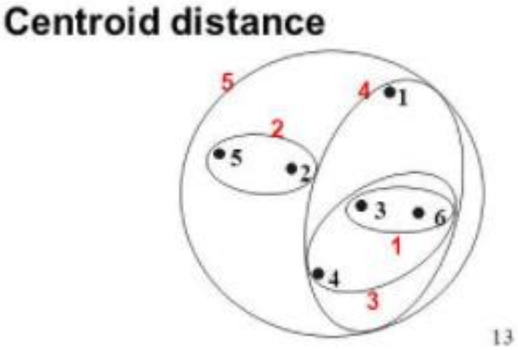
Average-Linkage

Average-linkage is where the distance between each pair of observations in each cluster are added up and divided by the number of pairs to get an average inter-cluster distance. Average-linkage and complete-linkage are the two most popular distance metrics in hierarchical clustering.



Centroid-Linkage

Centroid-linkage is the distance between the centroids of two clusters. As the centroids move with new observations, it is possible that the smaller clusters are more similar to the new larger cluster than to their individual clusters causing an inversion in the dendrogram. This problem doesn't arise in the other linkage methods because the clusters being merged will always be more similar to themselves than to the new larger cluster.


Centroid distance
5
4
1
2
5
2
3
6
1
4
3
13