Unsupervised learning:

In unsupervised learning models are learned from the ***unlabelled data*** where task is to analyse Pattern, Relationship or Structure of the data to make prediction or decision.

1. K means clustering:

K means clustering is an unsupervised learning algorithm where ***data points are classified into k different clusters***, where ***each cluster has same characteristics***.

Assumption:

1. Cluster canter is arithmetic mean of the all data points belong to cluster.

2. Each data point is closer to its own cluster canter than other cluster canter.

Process (Steps) in K means clustering algorithm:

1. Choose the number of clusters (K).

2. Initialize the centroids (Randomly select K points from the dataset as initial centroid).

3. Assign all the data points to respective centroid based on the Euclidean distance. **----->** ***E step***

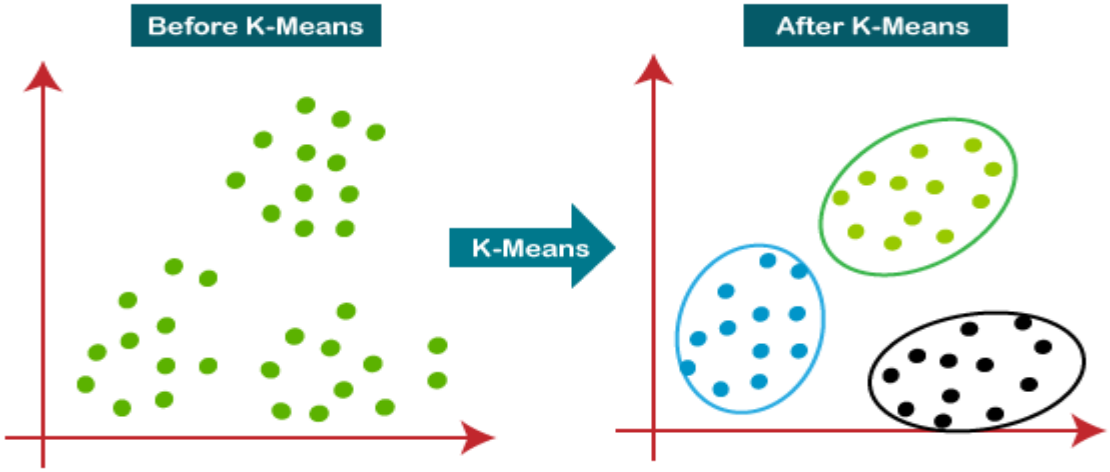
4. Recalculate centroid position by taking average distance of all data points in a cluster. **----->** ***M step***

Repeat this steps until the position of centroid remain constant.

Here's a brief overview of how K-means works:

- ***Decide the number of clusters (k)***
- ***Select k random points from the data as centroids***
- ***Assign all the points to the nearest cluster centroid***
- ***Calculate the centroid of newly formed clusters***
- ***Repeat steps 3 and 4***

Imp note: as value of K increases the number of centroid increases, so distance from centroid to data point will be decreased.



Select optimal value of K:

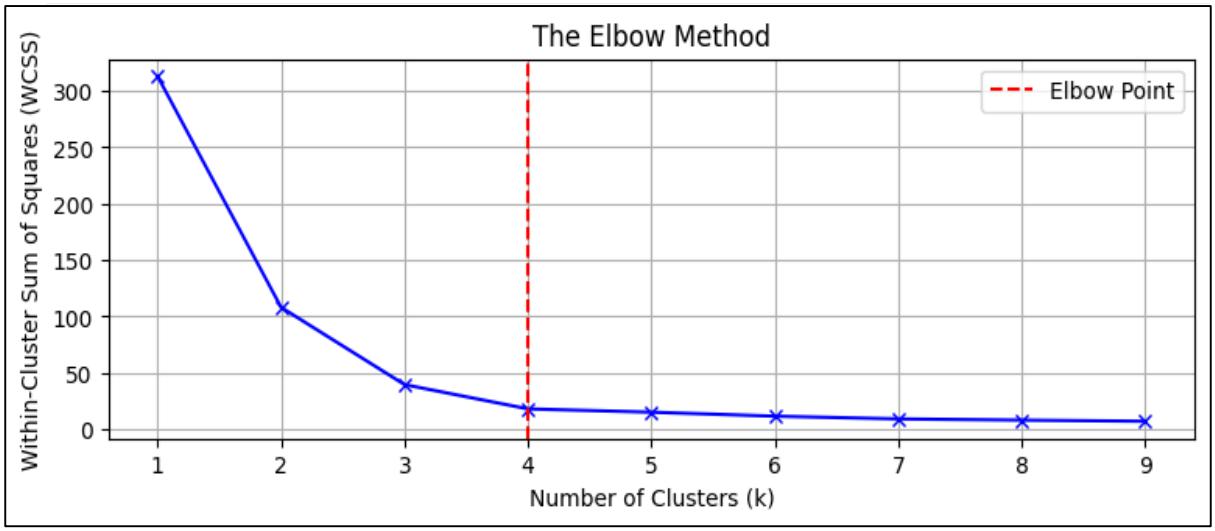
To select optimal value of K we used ***Elbow Method or silhouette score*** method.

1. Elbow method:

Elbow method is a graphical method for finding the optimum number of clusters within a k-means clustering algorithm.

To find the optimal value of clusters, the elbow method follows the below steps:

- It executes the K-means clustering on a given dataset for different K values (ranges from 1-10).
- For each value of K, calculates the WCSS value.
- Plots a curve between calculated WCSS values and the number of clusters K
- The point where adding more clusters results in only a marginal decrease in WCSS. This is where we observe an "elbow" shape in the graph.



At elbow point it almost captures up to 99 percentage of variability in the data. So the optimal value of K Is 4 here.

2. Silhouette method:

Let say we have initialize 3 cluster centres, where cluster 1 and cluster 3 are close to each other and cluster 2 is somewhat far from the cluster 1.

Let say we are calculating the silhouette score for the data point x1 belong to cluster 1.

Steps:

1. Calculate inter cluster distance (a). Which is average distance between X1 and all other data points in the same cluster which cluster 1.

2. Calculate intra cluster distance (b). Which is average distance between X1 and all other data points from the nearest cluster which cluster 3.

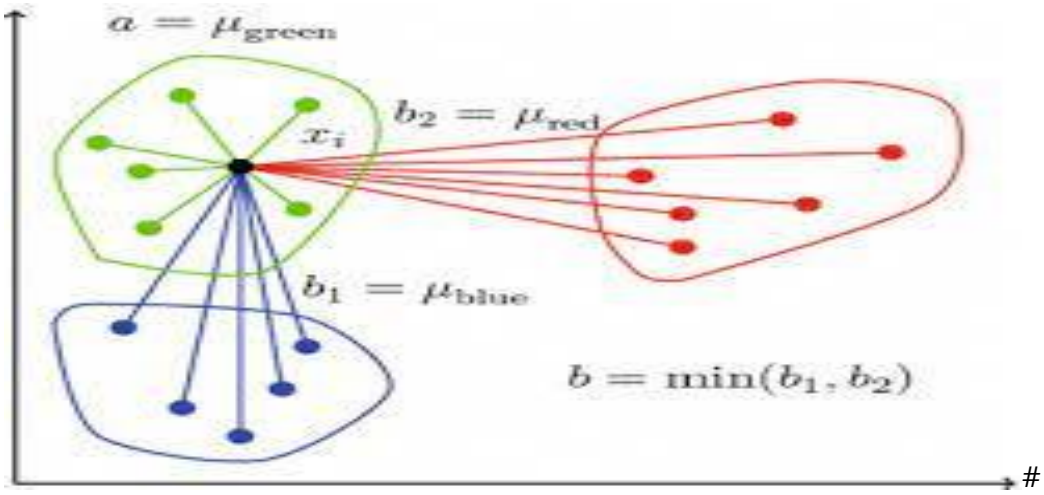
3. Calculate SC for x1 data point = b - a / max(a, b)

4. Similarly calculate SC for all the data points and then take average of SC

5. The value of k for which average of SC is maximum is best value of K

Imp Note:

The value of SC ranges from -1 to +1
SC = +1 ----→ Data point is well cluster.
SC = 0 ---→ Data point is near to boundary between two cluster.
SC = -1 ----→ Data point is assign to wrong cluster.



Disadvantage: Very sensitive to outliers

2. Hierarchical Clustering:

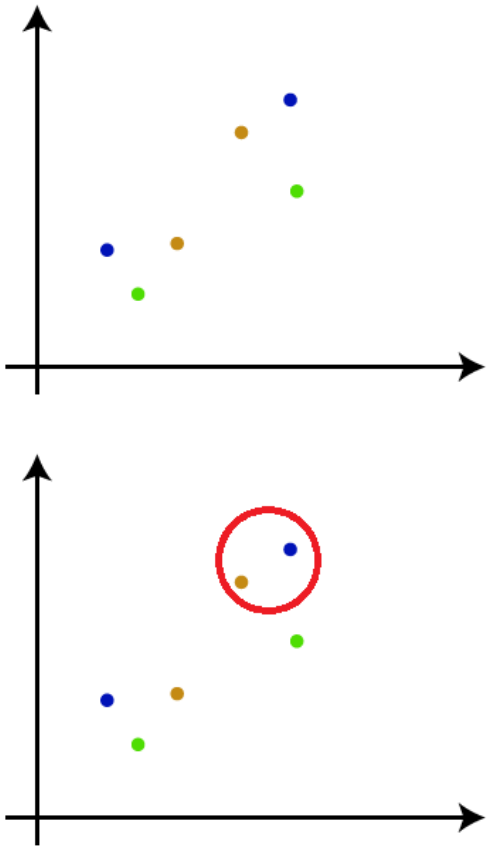
There are certain challenges with K-means. It always makes spherical clusters. Also, we have to decide the number of clusters at the beginning of the algorithm. Ideally, we would not know how many clusters should we have, in the beginning of the algorithm and hence it a challenge with K-means.

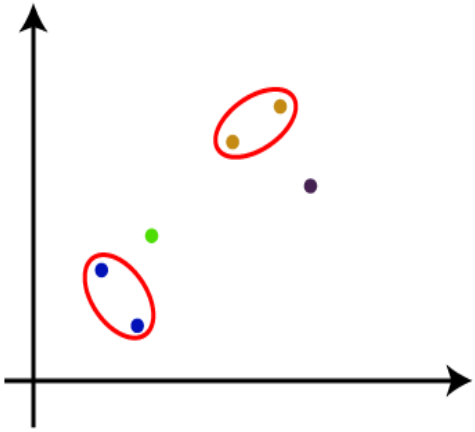
Hierarchical clustering takes away the problem of having to pre-define the number of clusters

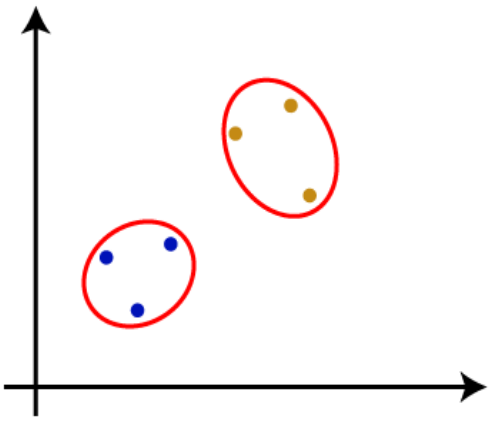
There are mainly two types of hierarchical clustering:

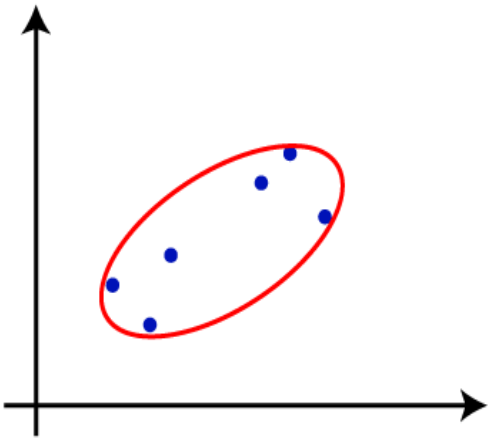
- Agglomerative hierarchical clustering
- Divisive Hierarchical clustering

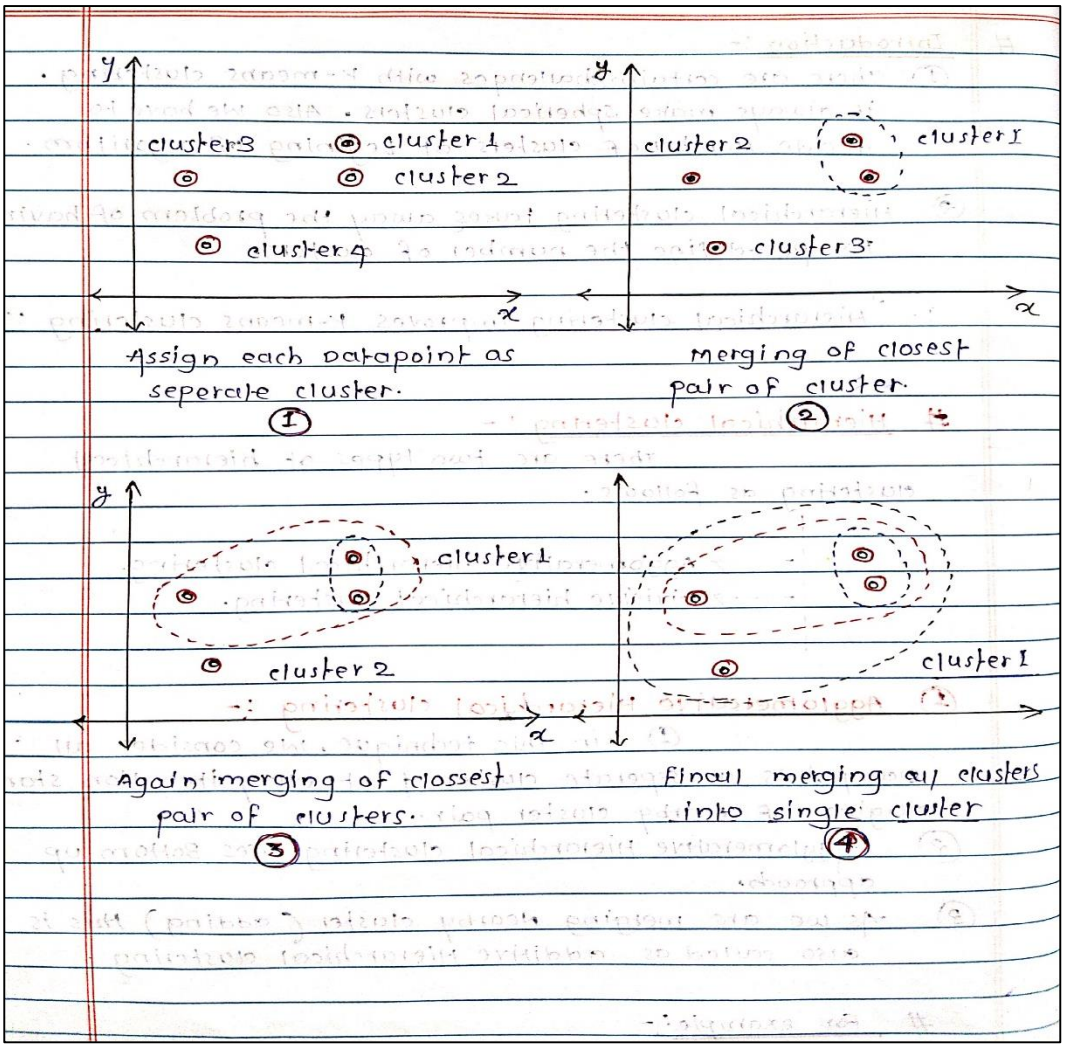
1. **Agglomerative hierarchical clustering :**
   Agglomerative hierarchical clustering uses bottom up approach where we consider every data points as a separate clusters and at every iteration we are merging the nearby clusters.
   This process continues till we get a single cluster.
   As we are merging the nearby clusters it is also called as Additive Hierarchical Clustering.











To decide nearby clusters we calculate Euclidean distance between the clusters and then we merge the clusters which are close to each other

The Euclidean distance formula says:

$$d = \sqrt{[(x_2 - x_1)^2 + (y_2 - y_1)^2]}$$

where,

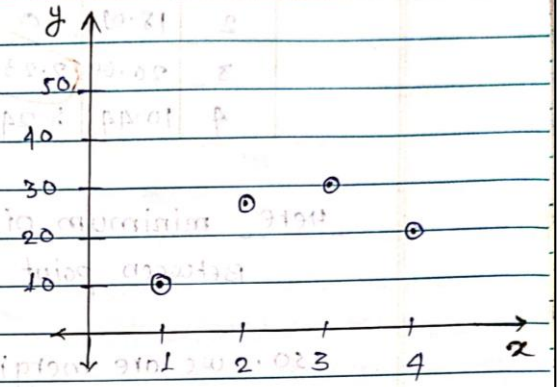
- (x_1, y_1) are the coordinates of one point.
- (x_2, y_2) are the coordinates of the other point.
- d is the distance between (x_1, y_1) and (x_2, y_2).

The Euclidean distance between each cluster is stored in the proximity matrix at each iteration.

Lets elaborate proximity matrix :-

let say we have 2 columns student ID and marks plotted in 2D Graph.

Student ID	marks
1	10
2	28
3	30
4	20



considering Agglomerative H.C.

(I) calculate proximity matrix.

> Eucledian Distance = $\sqrt{(x2-x1)^2 + (y2-y1)^2}$

$E.D = \sqrt{(2-1)^2 + (28-10)^2}$

$E.D = \sqrt{(18)^2 + 1}$

$E.D = \sqrt{324 + 1}$

$E.D = 18.01$ (Bet. 1,2).

> $(E.D)$ point $1,3 = \sqrt{(3-1)^2 + (30-10)^2}$
> $= \sqrt{(2)^2 + (20)^2}$
> $= \sqrt{4 + 400}$
> $= 20.09$

> $(E.D)$ point $1,4 = \sqrt{(3)^2 + (10)^2}$
> $= \sqrt{9 + 100}$
> $= \sqrt{109}$
> $= \sqrt{109} = 10.44$

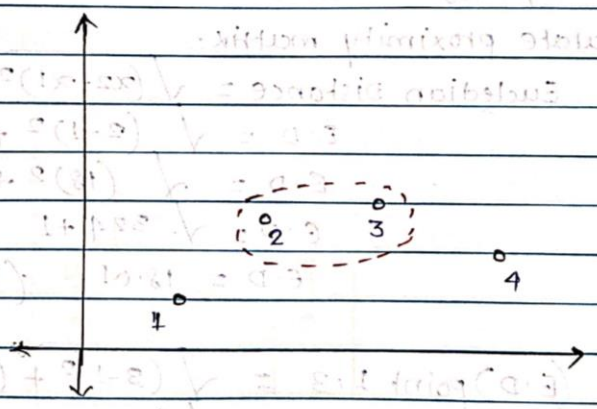
> $(E.D)$ point $2,3 = \sqrt{(1)^2 + (2)^2} = \sqrt{5} = 2.23$
> (ED) point $2,4 = \sqrt{(2)^2 + (2)^2} = \sqrt{68} = 8.24$
> (ED) point $3,4 = \sqrt{(1)^2 + (10)^2} = \sqrt{101} = 10.04$

Proximity matrix as follows.

	1	2	3	4
1	0	18.01	20.09	10.44
2	18.01	0	2.23	8.24
3	20.09	2.23	0	10.4
4	10.44	8.24	10.4	0

Here minimum distance is 2.23 which is between point 2 and 3.

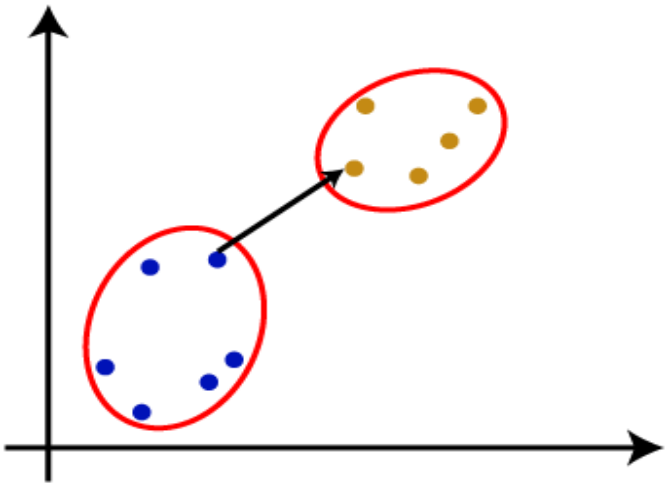
So, we are merging point 2 & 3 first.



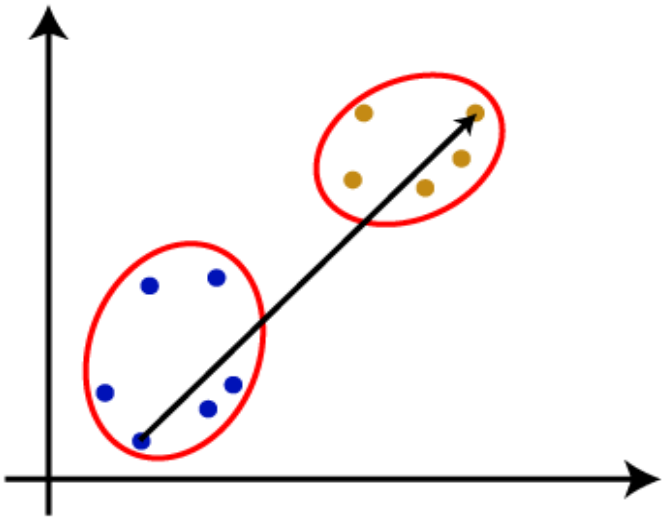
Here when we merge the closest clusters then to further calculate the proximity matrix we need to use linkage methods

Some of the popular linkage methods as follows

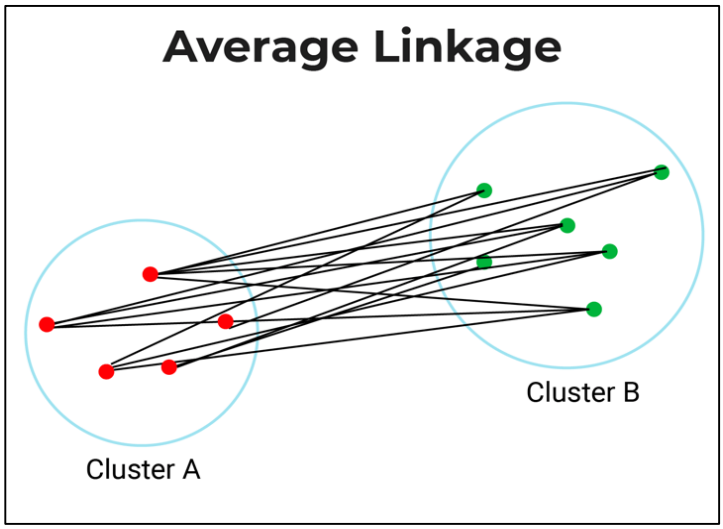
1. **Single linkage** : Distance between two cluster is define as shortest distance between any two points in a clusters



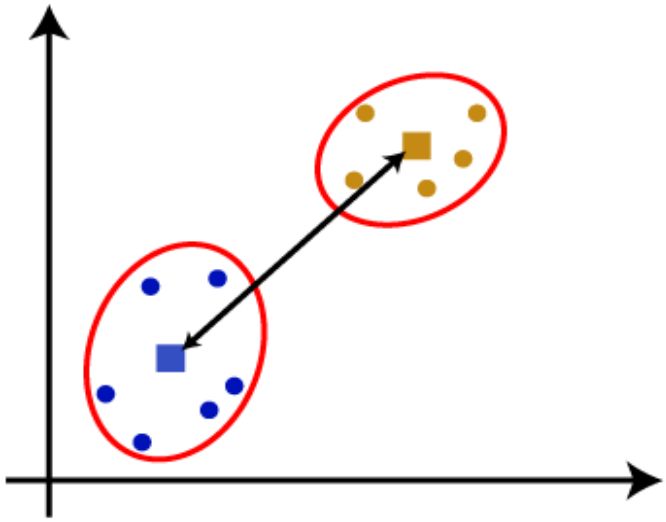
2. **Complete Linkage** : Distance between two cluster is the longest distance between any two points in a clusters



3. **Average Linkage** : The distance between two clusters is the average distance between all pairs of points in the clusters



4. **Centroid Linkage**: The distance between two clusters is the distance between their centroids (the mean of all points in each cluster).



5. **Ward's Method**: Ward's method is similar to average linkage method, only difference is that it uses sum of square distances.

Again calculate proximity matrix (update).

To update proximity matrix, select appropriate linkage method.

① Taking single linkage method (shortest distance).

We have merged closest clusters 2 and 3.
Now, min. distance Between clusters 2 and 3 is 18.01 For 1 & (2,3), and 8.24 For 4 & (2,3)
so, update matrix using 18.01

	1	(2,3)	4
1	0	18.01	10.44
(2,3)	18.01	0	8.24
4	10.44	8.24	0

Now, closest distance is 8.24 which is Between (2,3) and 4.
so, merge (2,3) & 4 New cluster created (2,3,4)

Again update proximity matrix.

	1	2,3,4
1	0	10.44
2,3,4	10.44	0

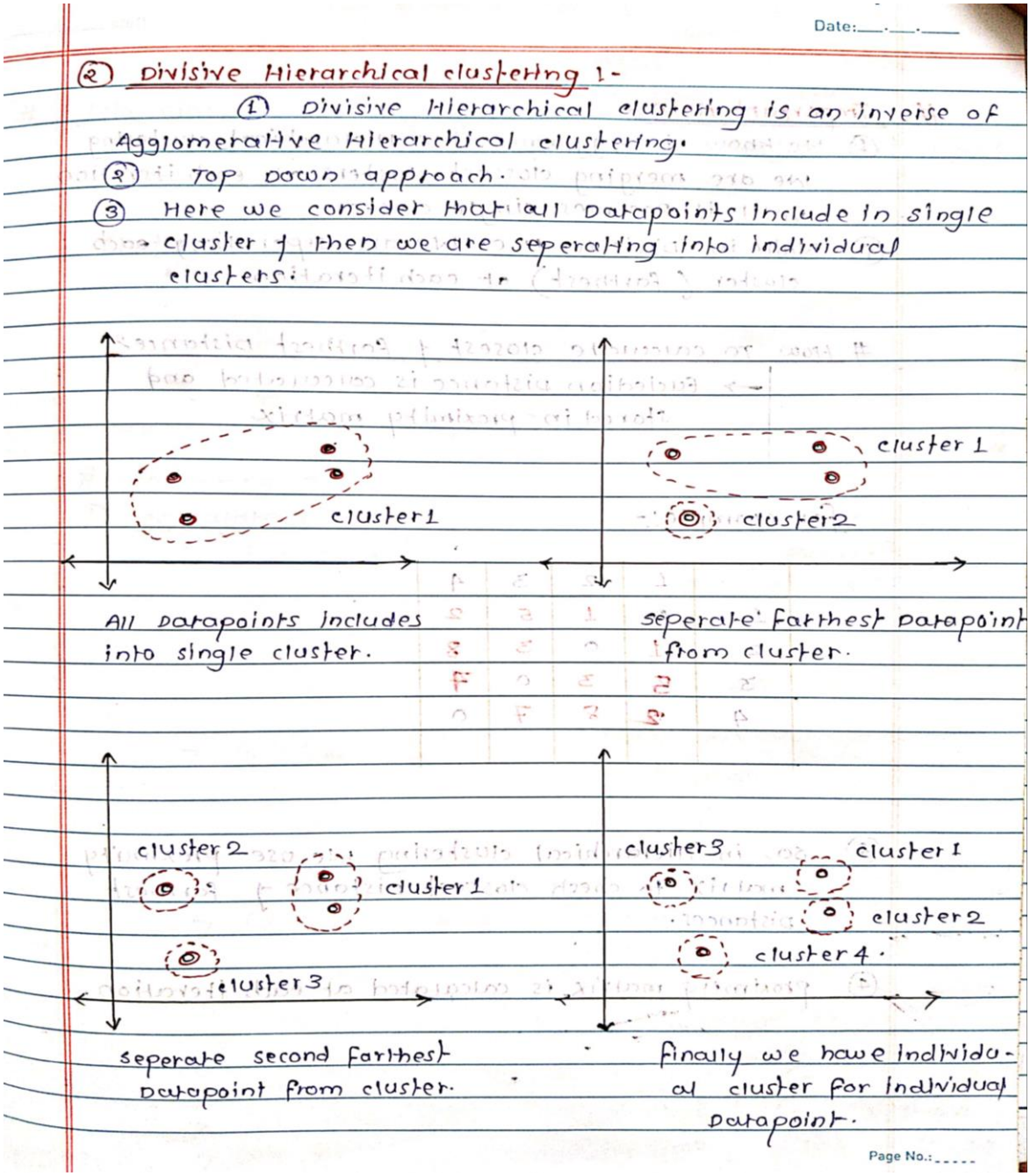
Here, min Distance Between 1 and 2,3,4 is 10.44 (18, 20, 10)

2. Divisive Hierarchical Clustering :

Divisive clustering is the opposite of agglomerative clustering.

Divisive hierarchical clustering uses top-down approach where we consider all data points in the single cluster and at every iteration we are splitting into the individual clusters.

This process continues till we get an individual cluster for every data point.



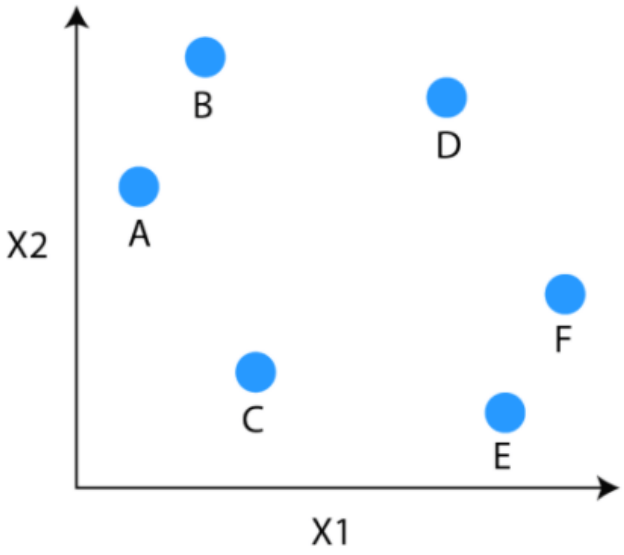
How to choose the number of clusters?

To choose the number of clusters in hierarchical clustering, we make use of a concept called dendrogram.

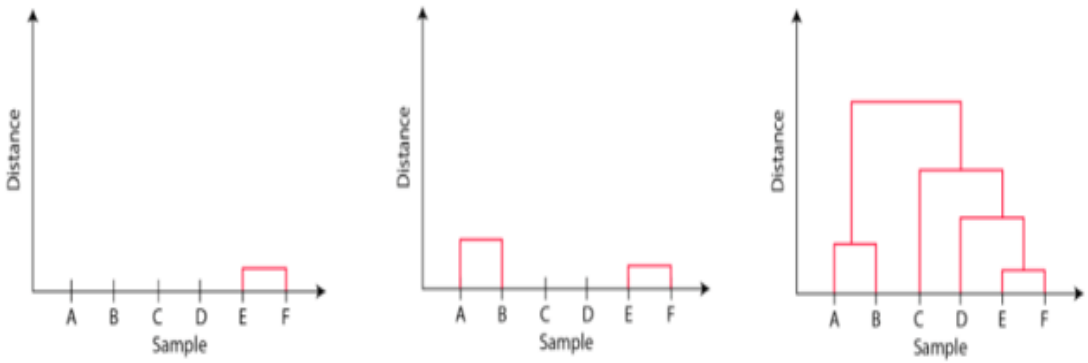
Dendrogram:

A dendrogram is a tree-like diagram that shows the hierarchical relationship between the observations.

Let's see how to form the dendrogram for the below data points.



1. Calculate Distance matrix of all the data points and then merge the data points into single cluster which are closest
   In this case we can say that E and F are closes to each other. So we will first merge E and F and so on
2. The next observations that are closest to each other are A and B which are combined together.
3. Similarly, D can be merged into E and F clusters and then C can be combined to that. Finally, A and B combined to C, D, E and F to form a single cluster.



 # The important point to note while reading the dendrogram is that:

1. The Height of the blocks represents the distance between clusters, and
2. Distance between observations represents dissimilarities.

But the question still remains the same, how do we find the number of clusters using a dendrogram or where should we stop merging the clusters?

Ans: Generally, we cut the dendrogram in such a way that it cuts the tallest vertical line. In the above example, we have two clusters. One cluster has observations A and B, and a second cluster has C, D, E, and F.