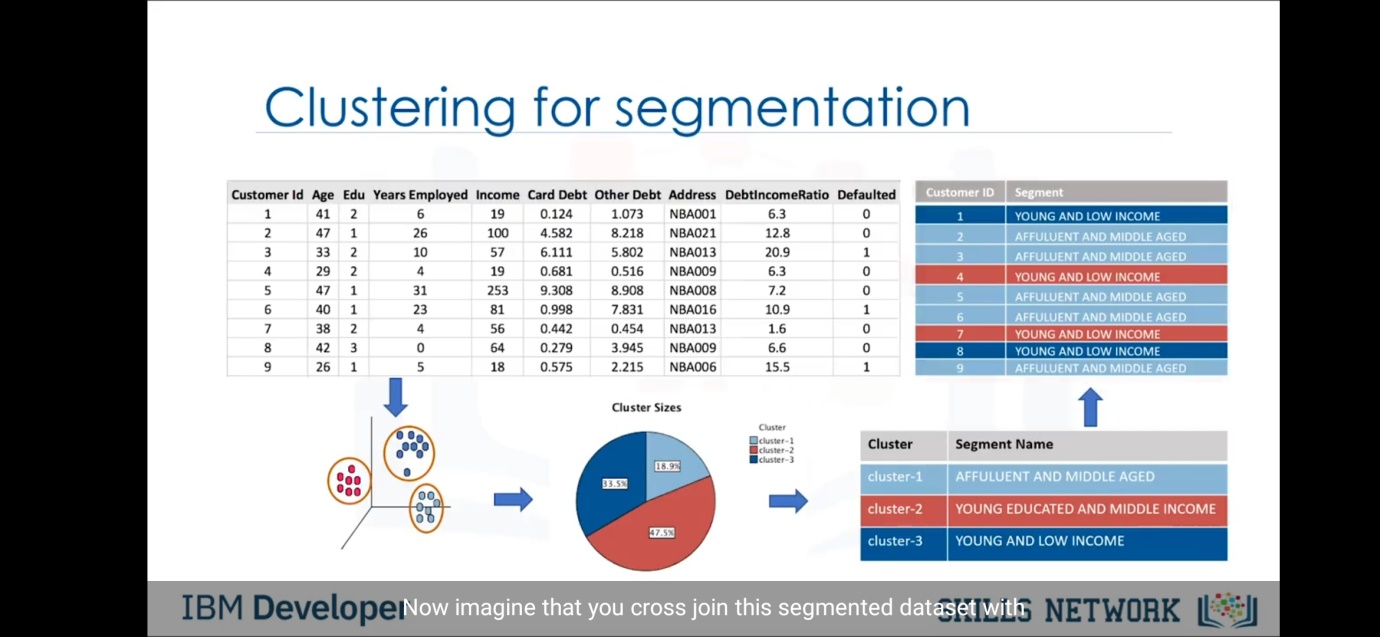
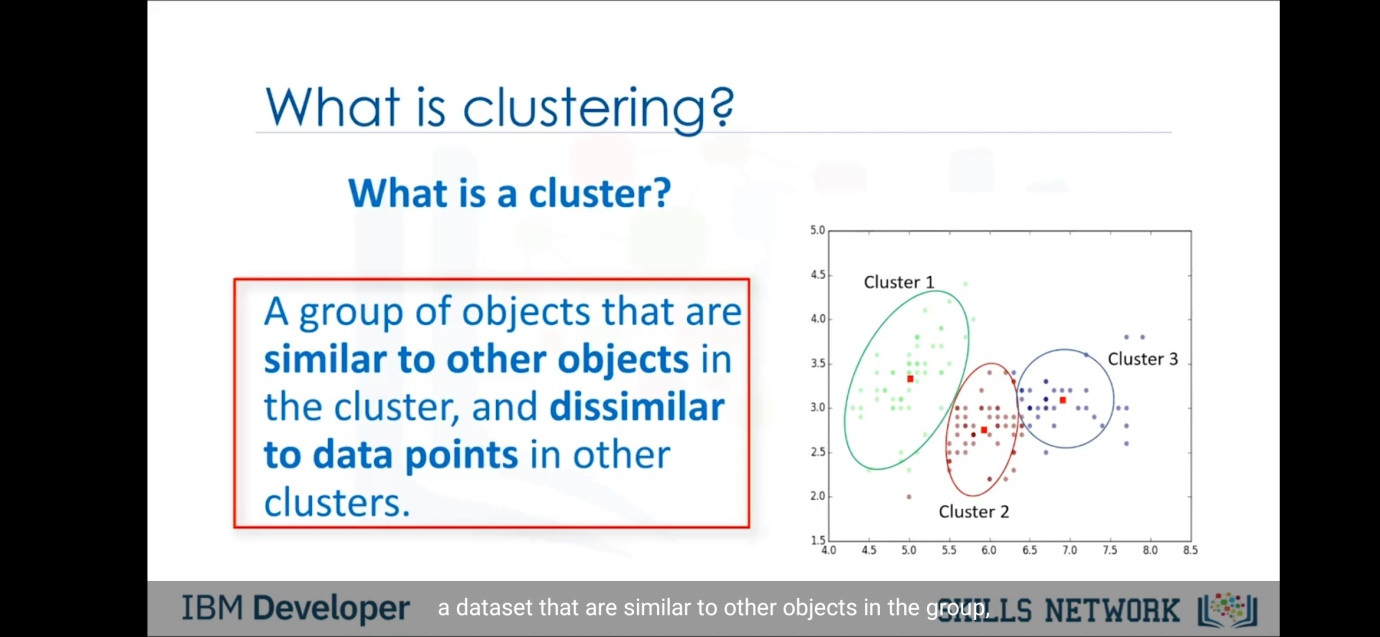
**CLUSTERING**

**INTRODUCTION TO CLUSTERING:**

Clustering is one of the important techniques of machine learning used in many routine areas. The understanding of clustering can be extend by the given simple example. Suppose the company(rather a bank) has the data of their customers as shown in the table below. Every customer data involves id , age , education ,years employed , income , card debit, other debit , address ,debt to income ratio, defaulted or not. Now the bank want to segregate the data into different categories. So clustering technique can be used for categorizing the data. By doing so the marketing team can approach the people of different fragments with different schemes and plans. For example we want to categorize the customers into 3 parts i.e. affluent and middle aged ,young educated and middle income, young and low income . So by segmenting the customers using clustering technique we can get the respective goals. Hence clustering is one of the important machine learning weapon to conquer such problems. 

So the basic definition of clustering is to make a group of objects that are similar to the other objects of the cluster and different from the other clusters. Technical definition is given in the image: 

We have gained knowledge about supervised learning from the further notes. To recap the basic idea, the supervised learning is generally A to B mapping . Given an input , it will give output of yes/no or 1/0. Also in supervised learning we have labelled dataset. On the other side of the coin, unseprvised learning helps us in generating groups or clusters of data. It does not gives output as yes/no but gives that this data set can be categorized with this particular group.

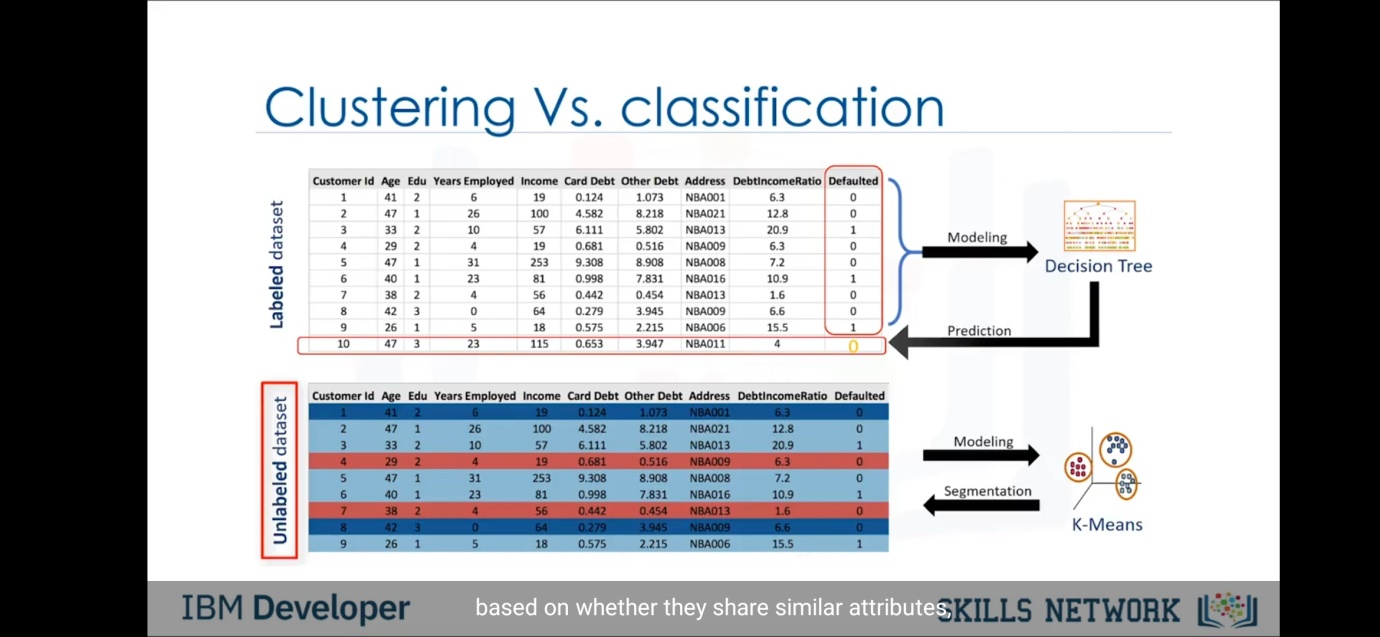
Let’s take an overview of difference between classification and clustering. Here we will take the same example we rented above . There is a dataset of customers of a bank with labels like age,income etc. Now for classifcation , the concept refers as a supervised learning . So using classification we can detect for a given new data , for example for given age,income and other things we can clssaify that the customer will be deaulter or not. Hence the data used in classification is labeled . So it generates output as 1/0 for default/ not default. On the other side clustering is used to cluter this customer into different segments like middle aged and affluent, young and low income etc. So clustering is an example of unsupervised learning as it doesn’t generates the output like 1/0 or yes/no.

So in classification the data set is given, the process can be simplified as follows:-

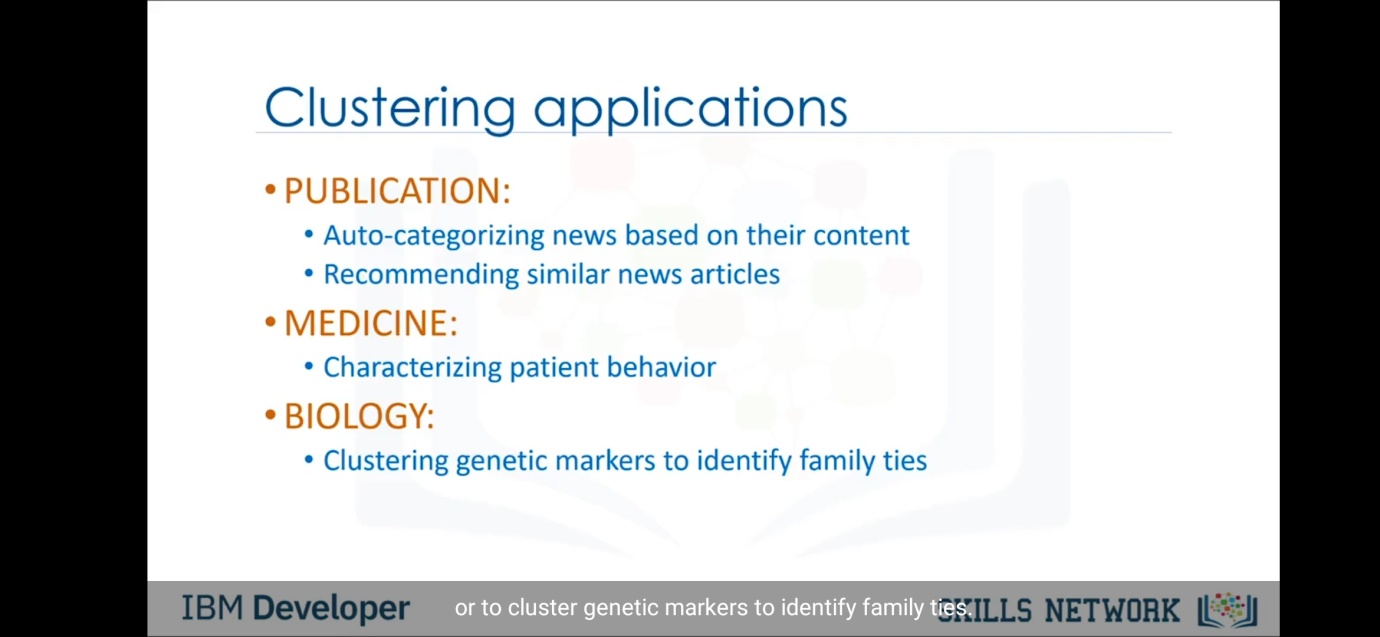
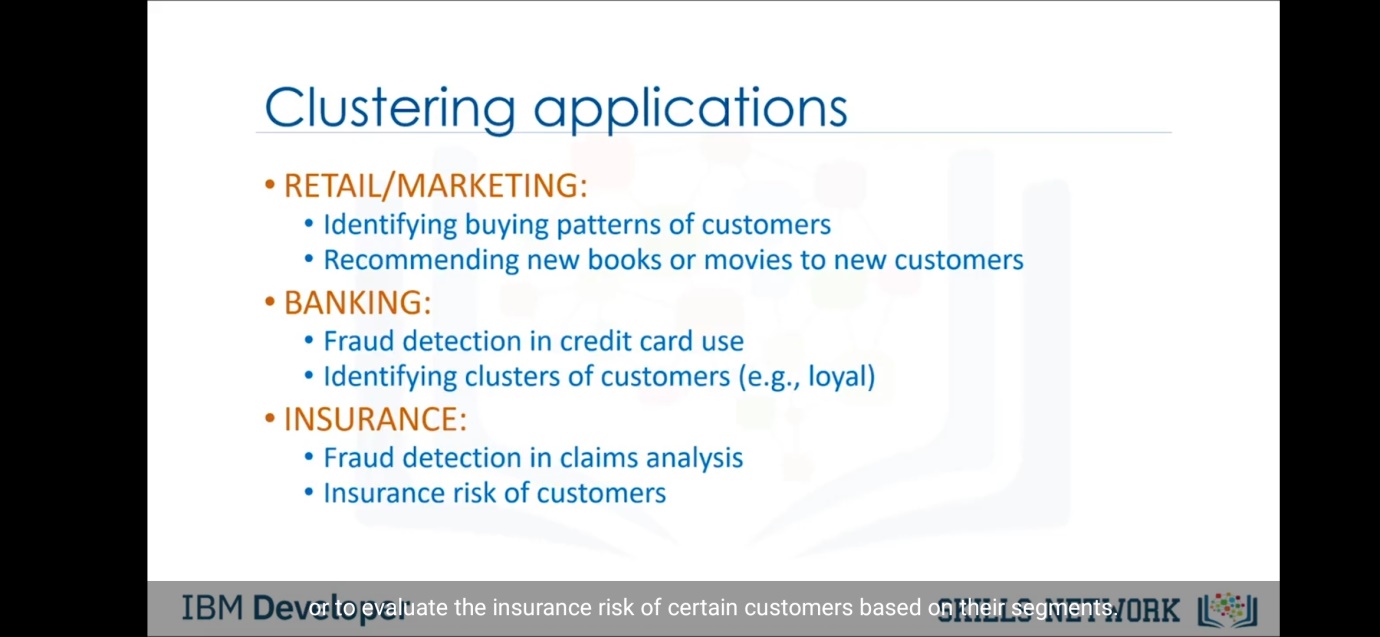
Collect data -> Modelling -> decision tree -> Prediction

Whereas the clustering process looks like this:-

Collect data -> Modelling -> k-means -> Segmentation



**Applications of Clustering:-**



Now clustering can be divided into mainly three parts.

1. Partitioned based clustering

-> Relatively efficient

-> k-means,k-median etc

1. Hierarchical Clustering

-> Produces trees of cluster

-> Agglomerative, Divisive

1. Density based clustering

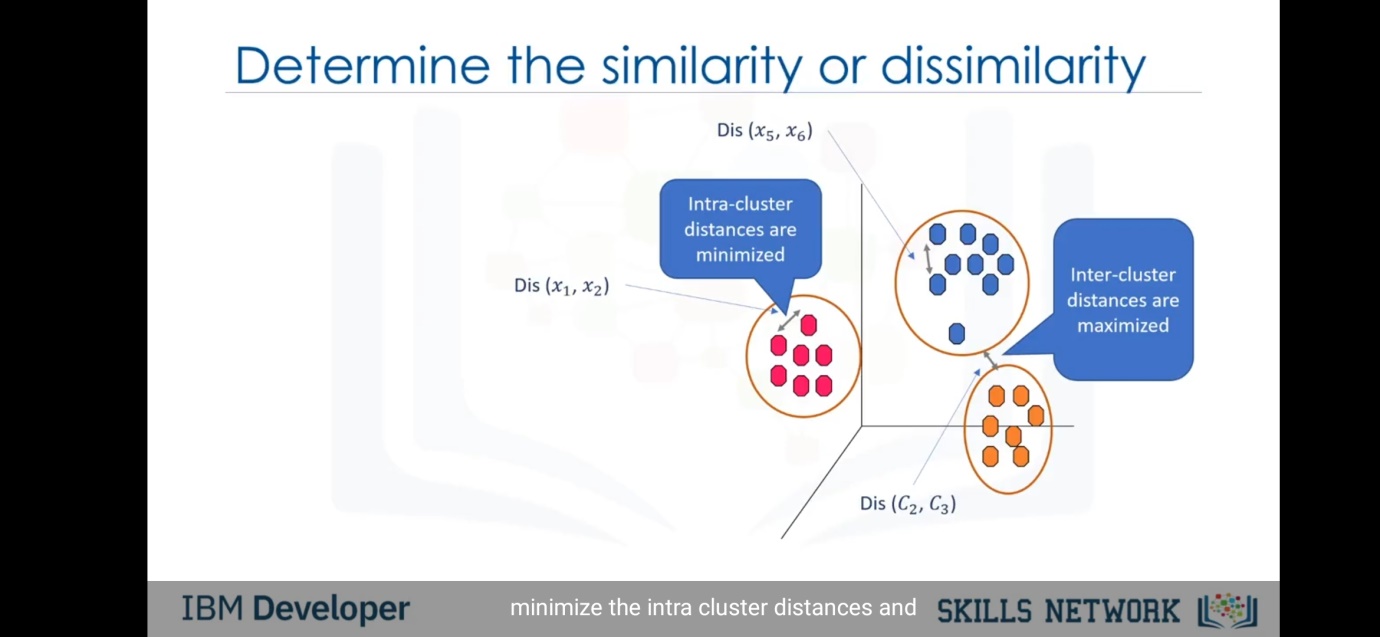
-> Produces arbitrary shaped clusters

-> DBSCAN

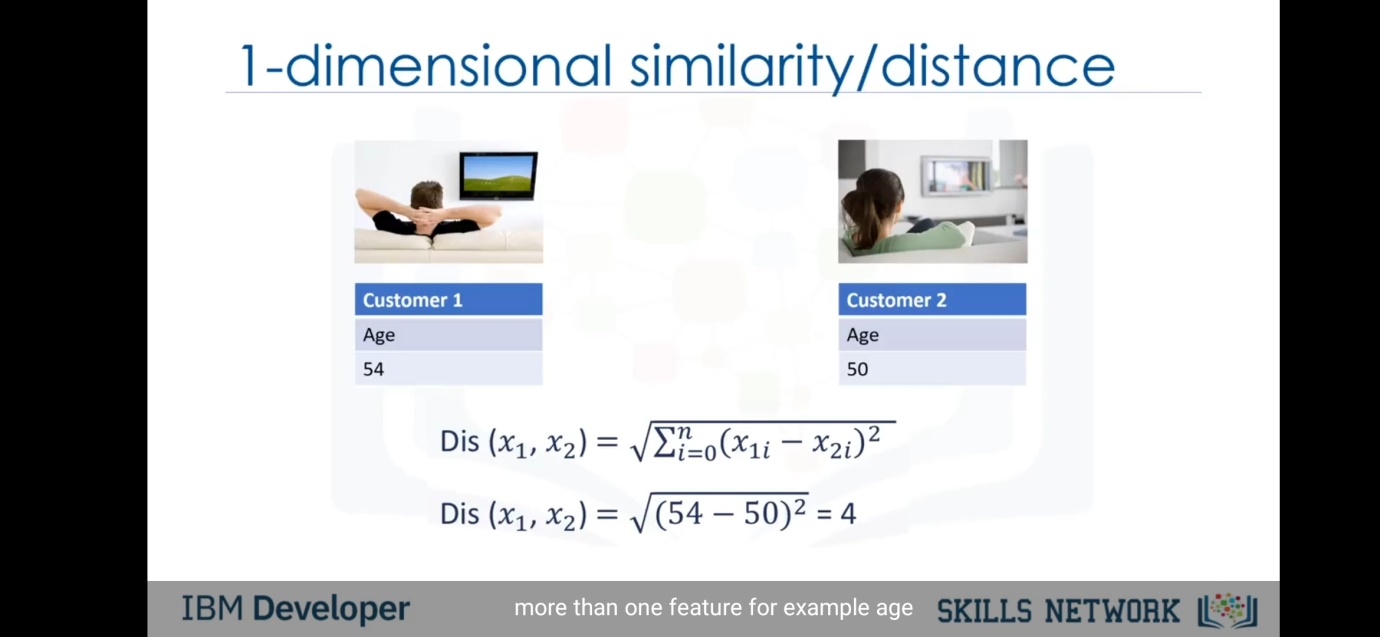
Now first we will study the k-means algorithm.

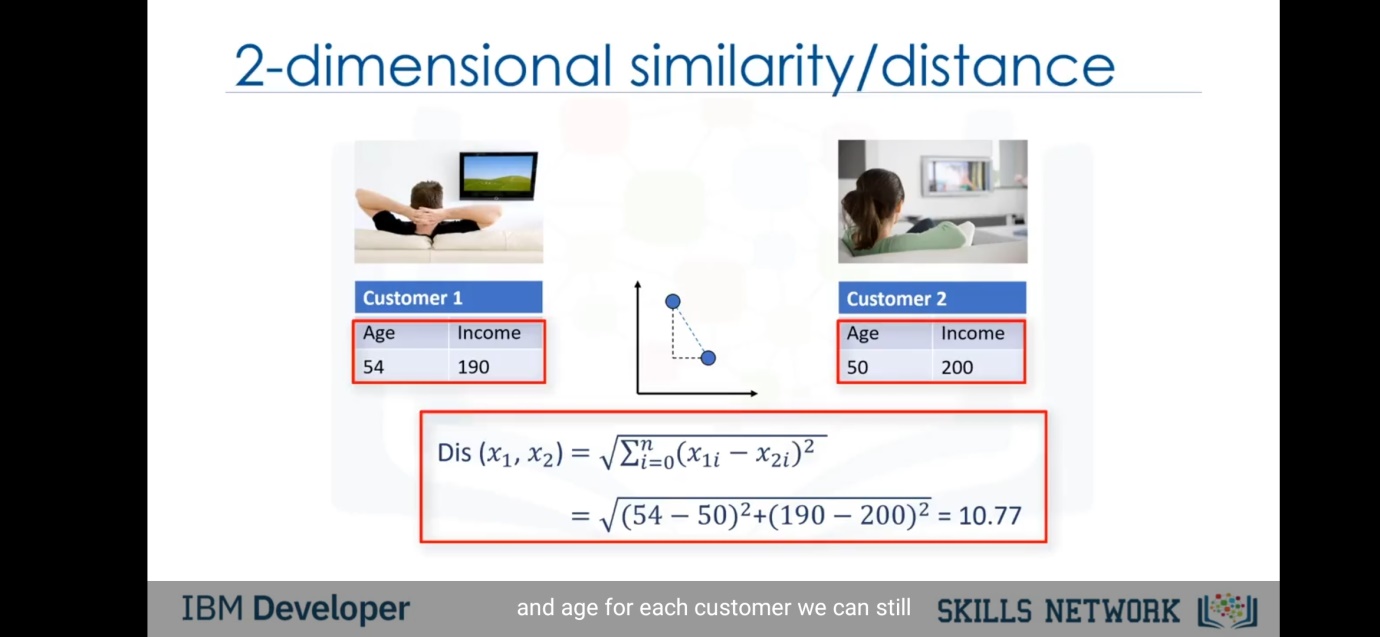
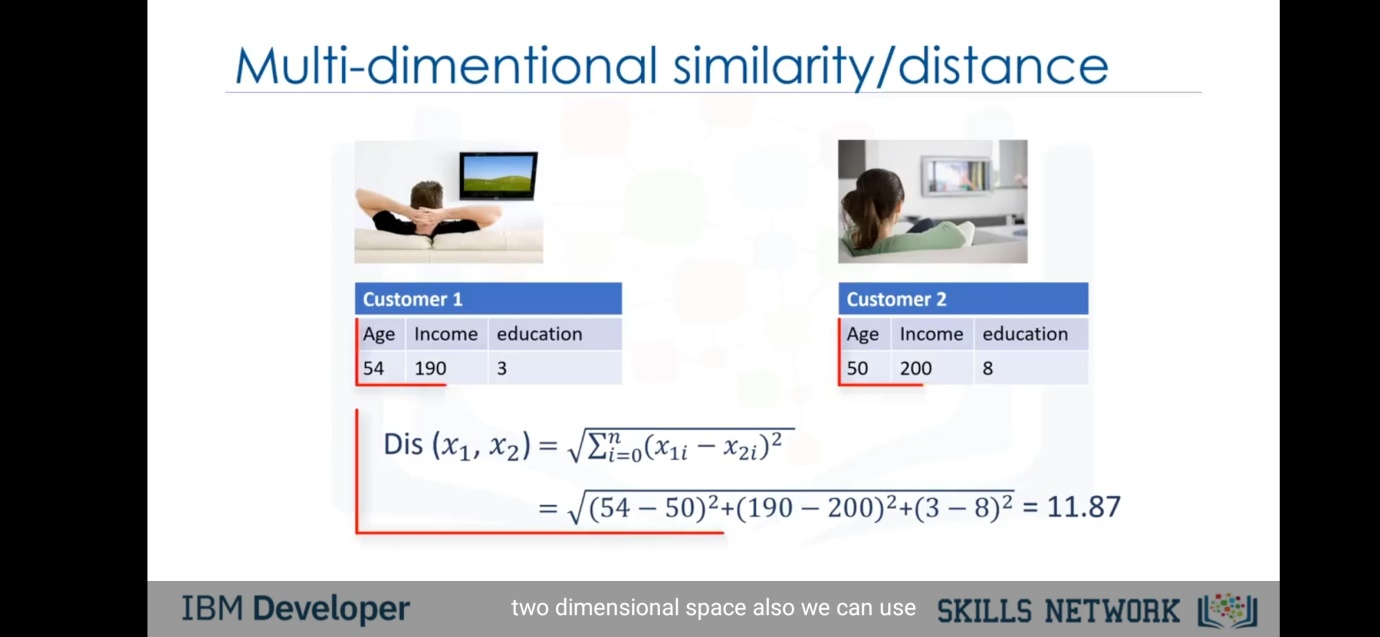
K-means algorithm is art of the partitioning clustering. In k-means we divide the given data into k-different clusters such that the data is similar to the other data of same cluster and is different from data of the other clusters.

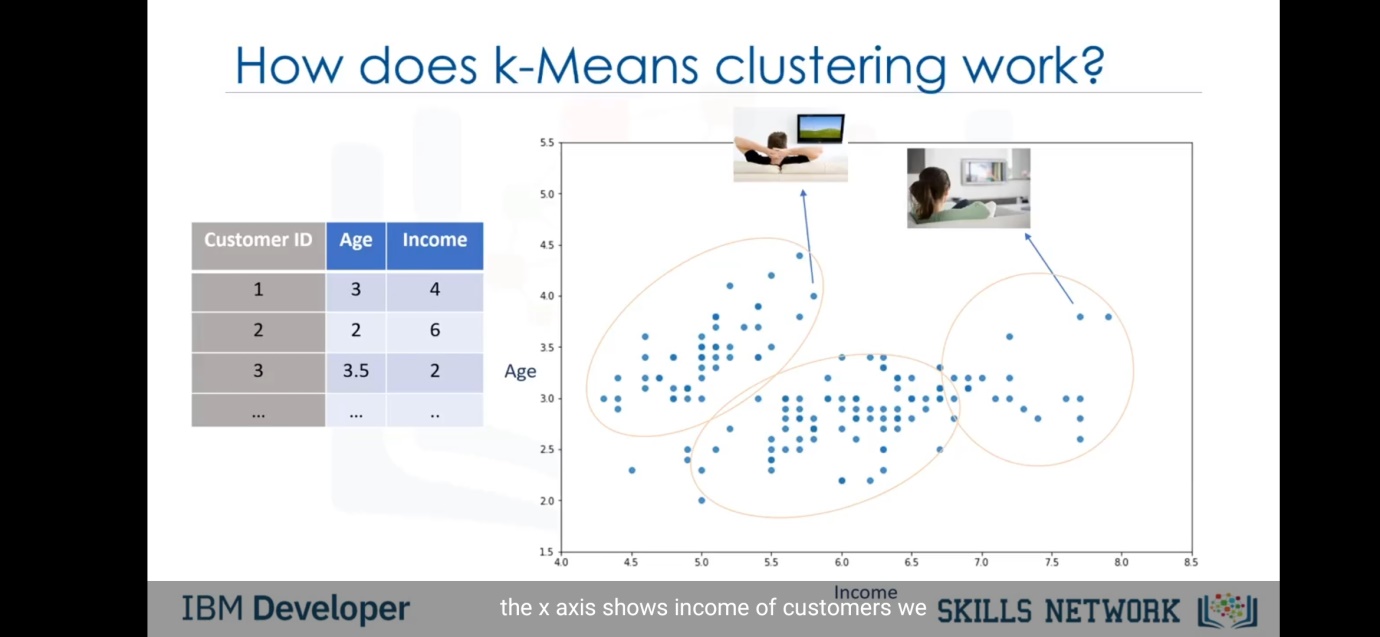
To measure the similarity or dis-similarity we measure the distance between the data points. For that we can used many methods like Euclidian distance, cosine distance etc .

The k-means clustering works on the principle as the clusters should be made in such a way that the intra-cluster distance should be minimized and inter-cluster distance should be maximized. 

Here are the examples of how the distance can be measures. For example if we have data of age of the customer then distance calculation is as follows.



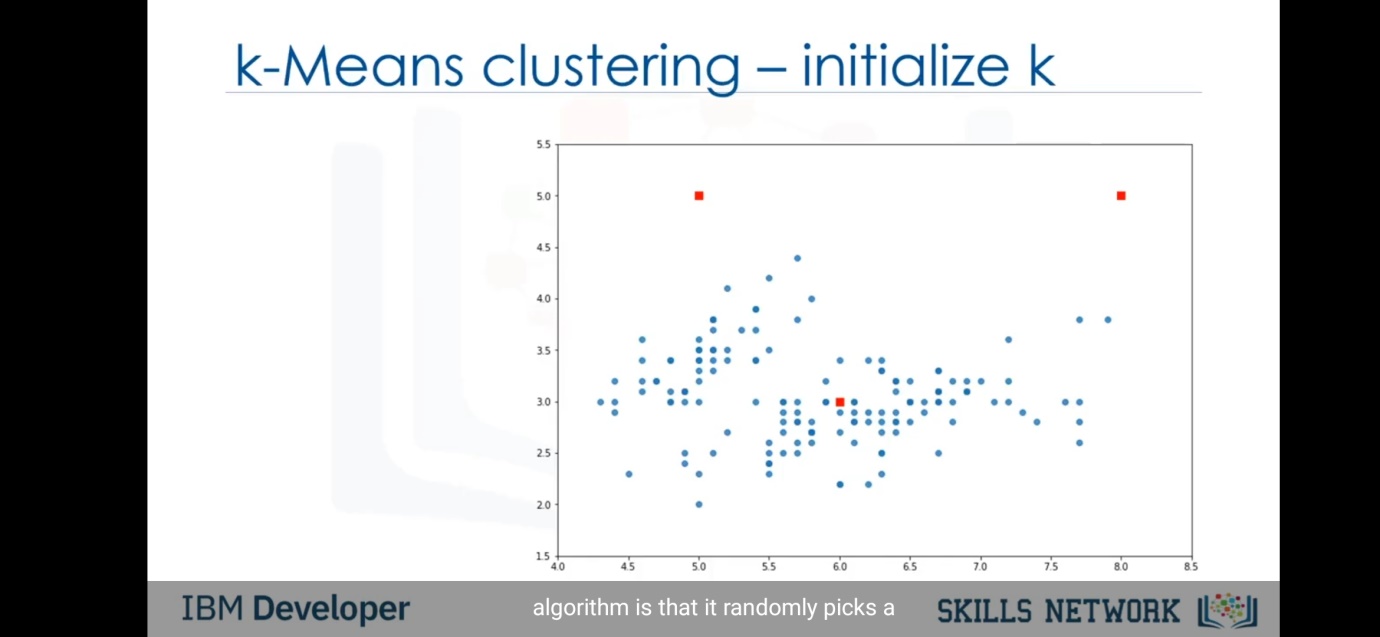
For the 2-d data the calculation is as follows:-Hence the general formula is given as :-

Now we shall discussed that how the algorithm of k-means worked in order to cluster the data into different segments. Suppose we have a data distributed into age v/s income graph as shown in the figure. 

So the algorithm is followed by 5 steps as follows :-

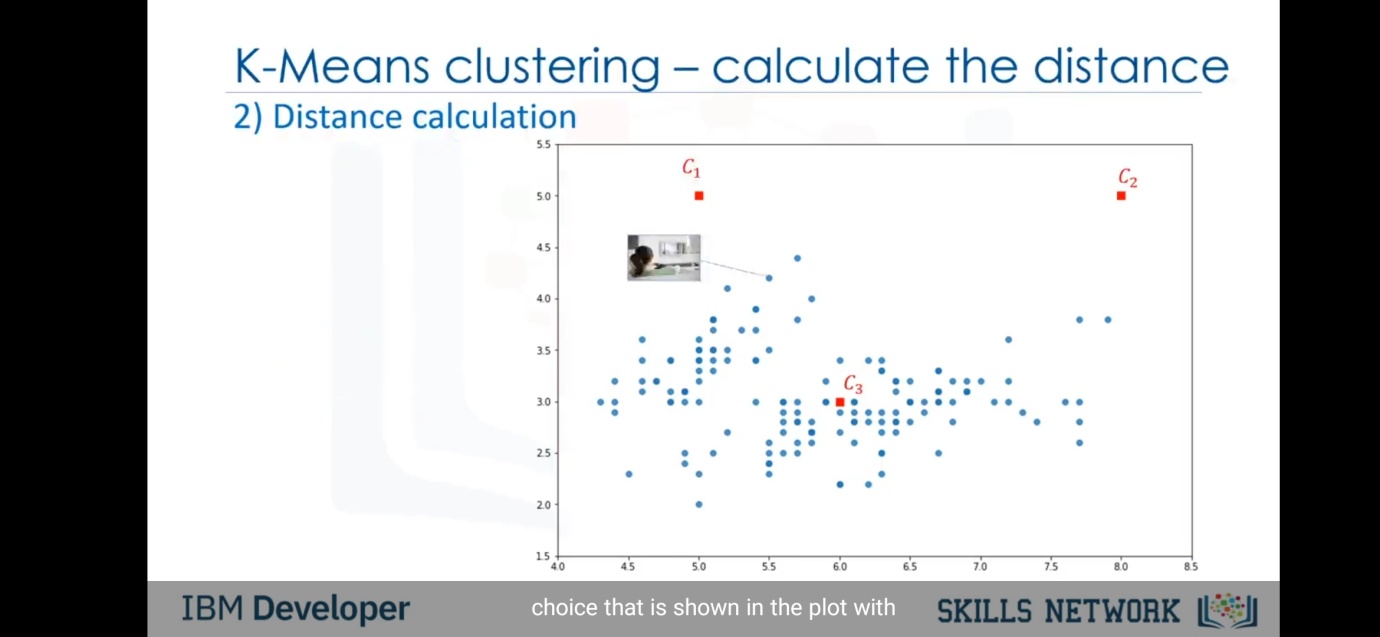
1) Initialize K.

Here k means into how many clusters we want to divide the data. The value of k plays an important role in this method. But for now we will take a random k. Let’s suppose k=3. So we will keep the centroid of clusters in the graph randomly. As there are three clusters to be made(k=3) we required 3 centroids. So the red points display here in the figure are the centroids which initially we kept at the random position.



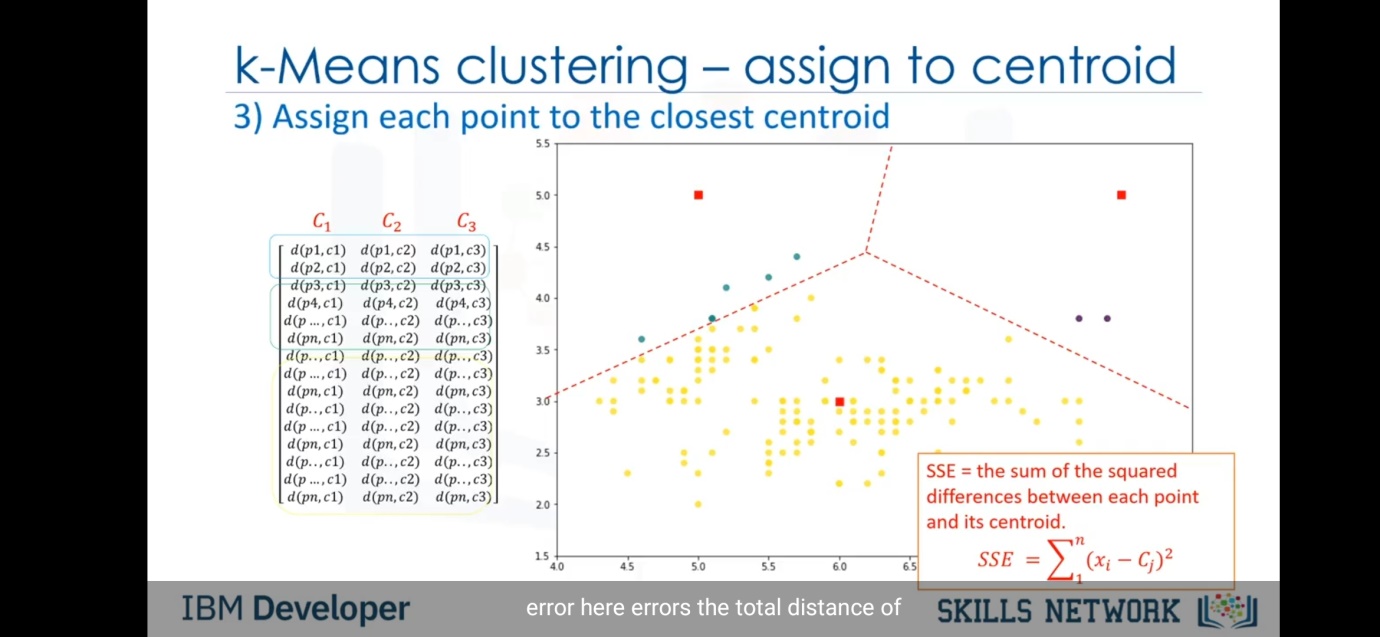
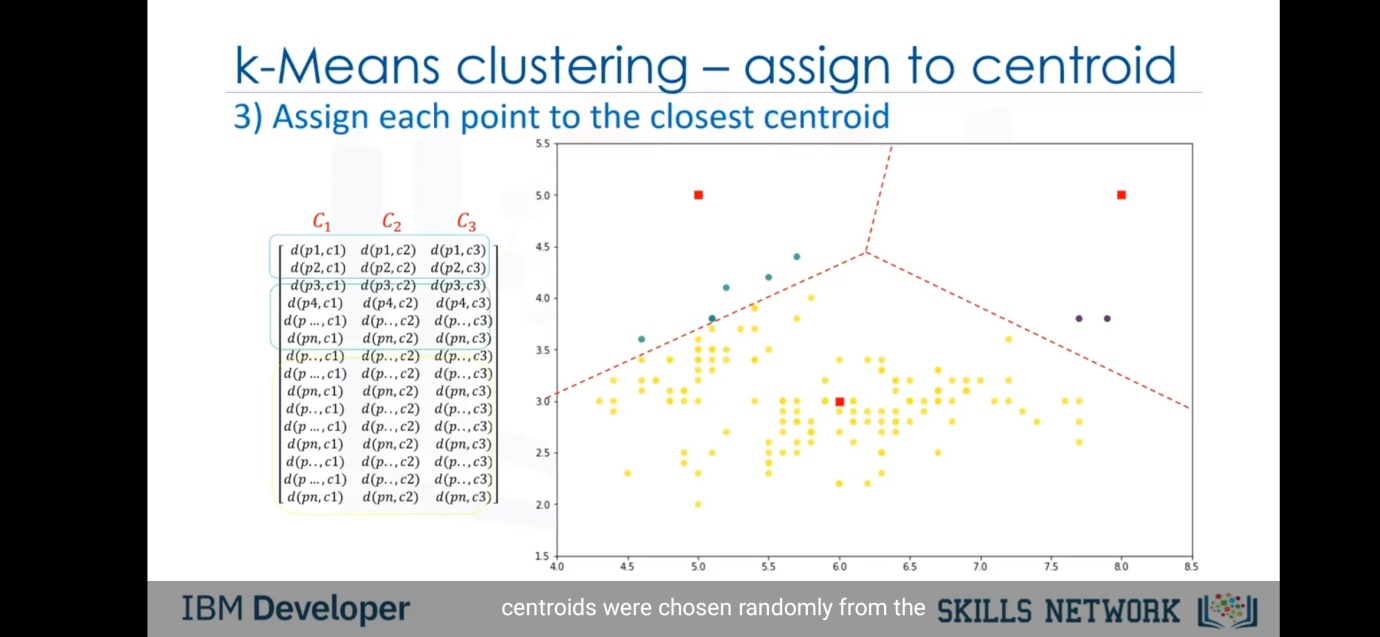
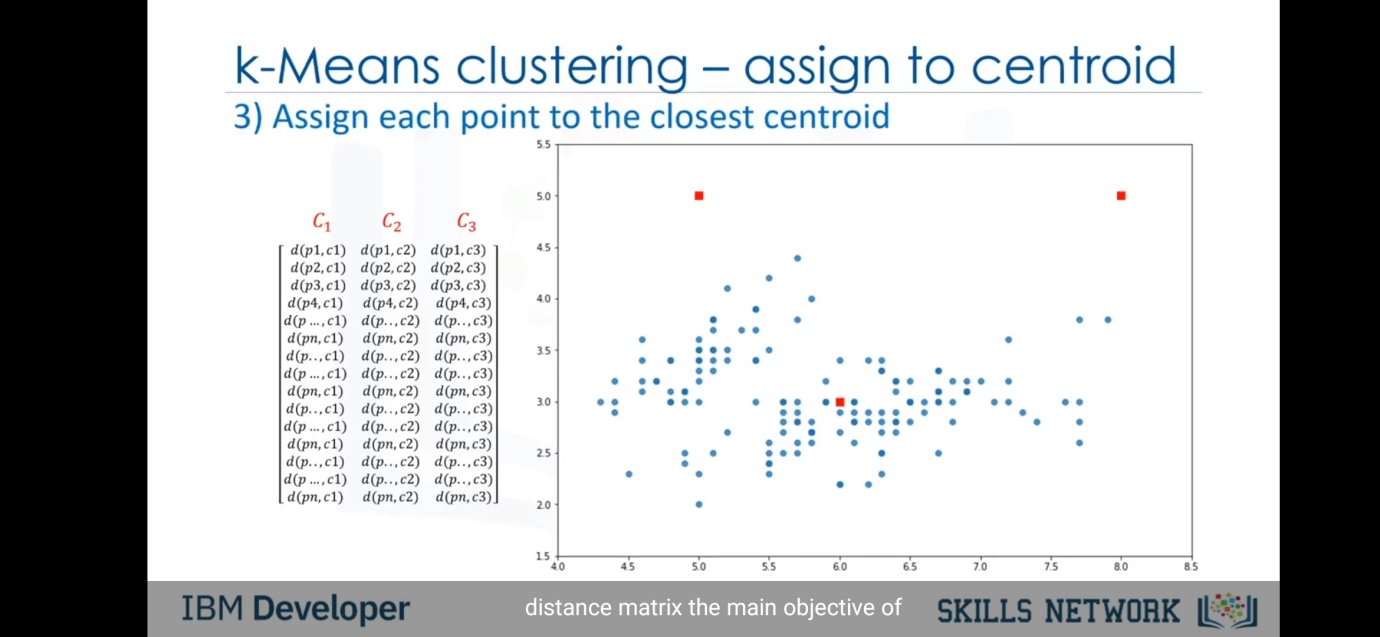
Step 2) Distance calculation

Now we will calculate the distance between the centroid points and each points of data displayed. The distance is calculated as above example we encountered, by Euclidian distances. After calculating the distance of each point with all k (here 3) centroids , we store it in a matrix technically called as distance matrix.



Step 3) Assign the closest centroid.

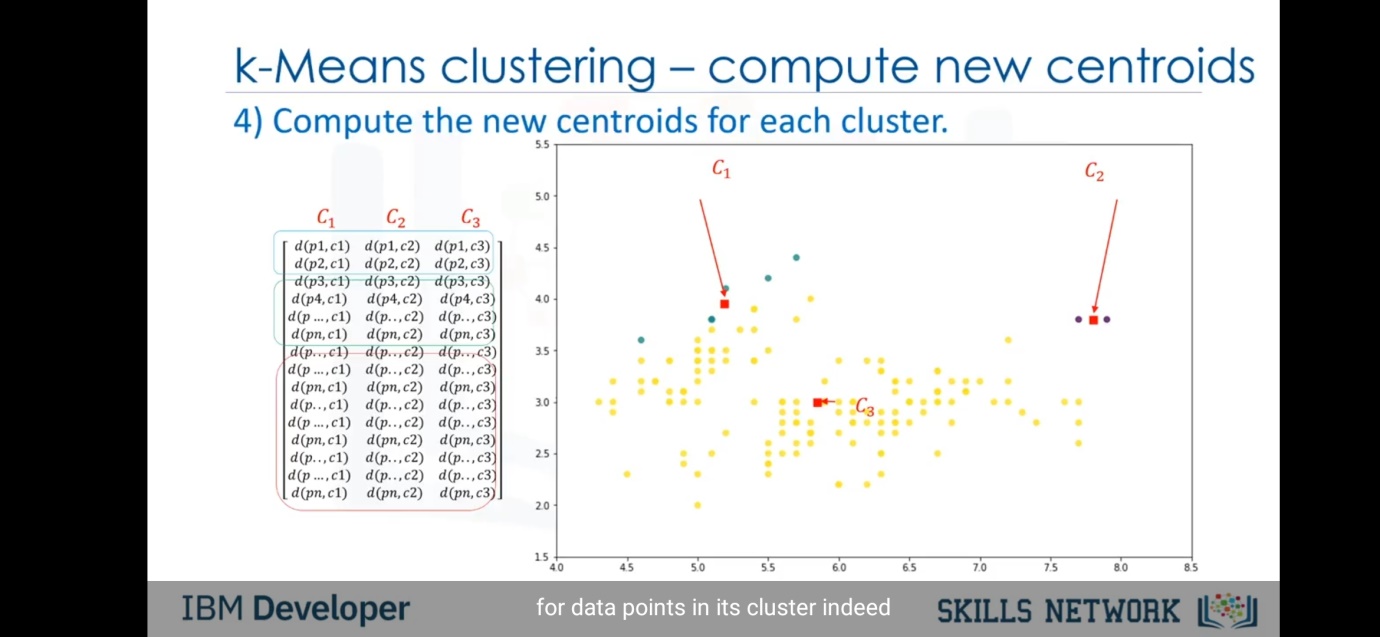
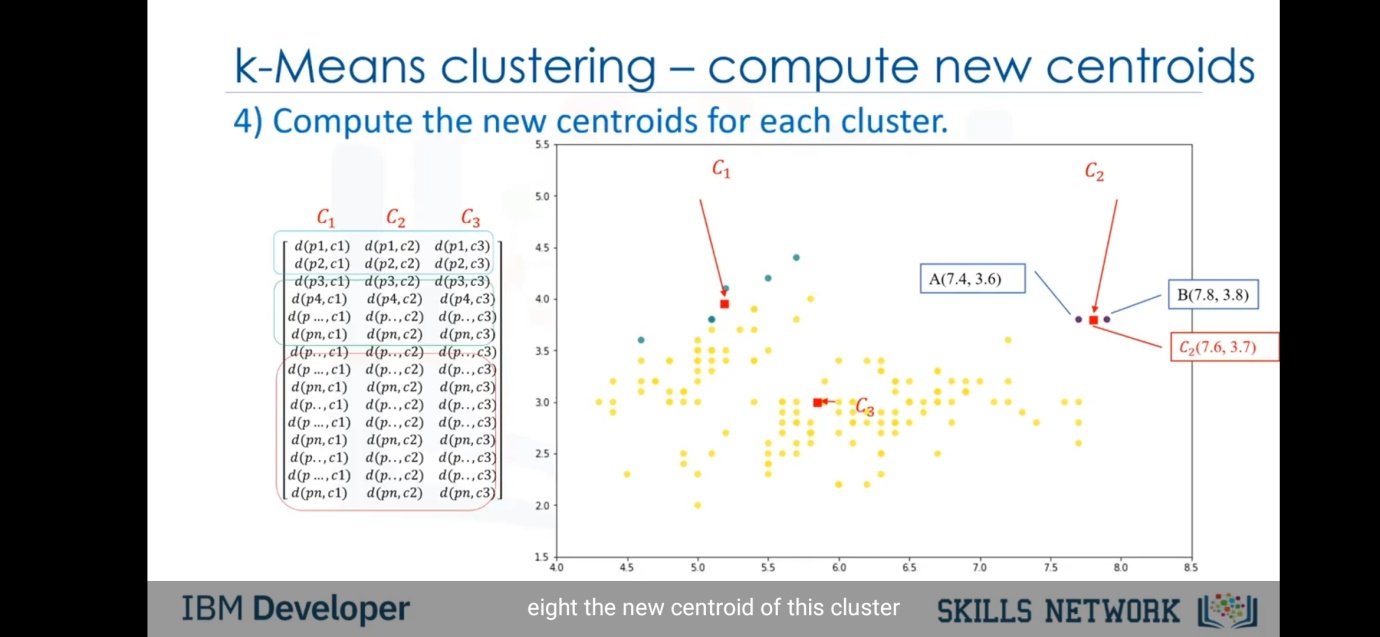
After we got the distance matrix , which has data of distance between the points and 3 centroids, we will assign the points to the closest centroid. So for each point in the distance matrix, it will consist of 3 values i.e. the distance between the point and 1st centroid, distance between point and 2nd centroid and so on. So from that the one having less distance from that 3 , we will assign the respective centroid to respective point.



But as seen above in the figure the cluster formed are not properly distributed . The reason is because we have selected the centroid initial points randomly. So this triggers the error which is generally called SSE(sum of squares error) whose formula is given in the figure.Its just the distance between the point and centroids. Now for accurate model we require to reduce the error.

Step 4) Compute new centroids for each cluster.

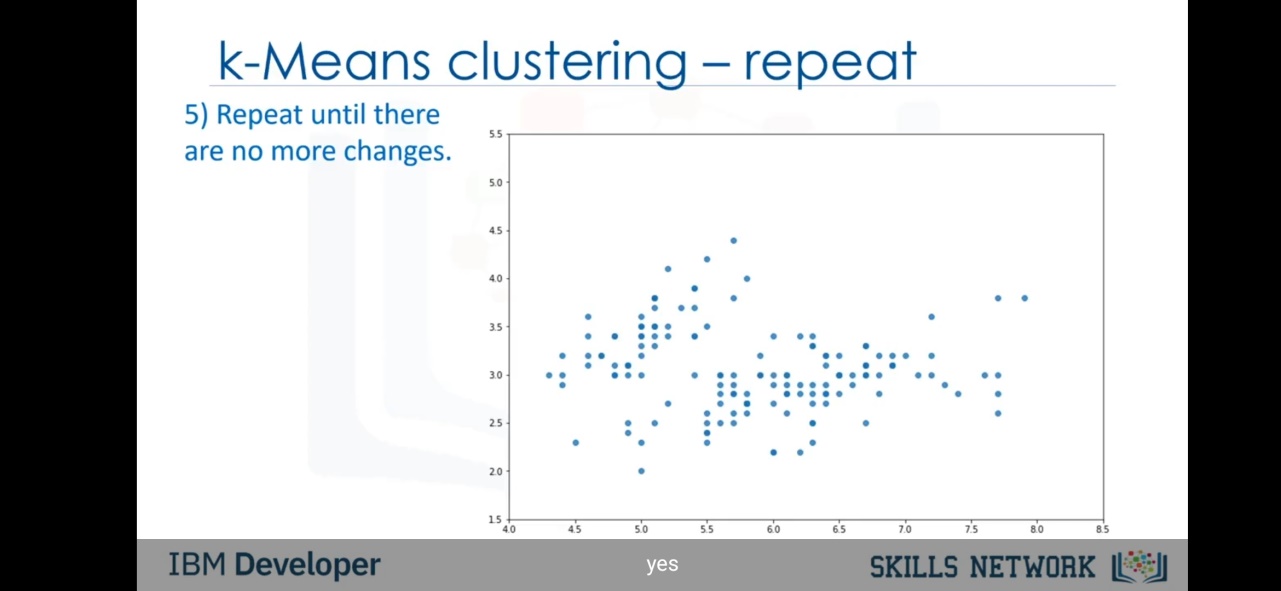
Now we got the SSE , we have to reduce it. To rescue form the error we will compute new centroids for cluster. The meaning is we will change the position of the centroids from the initially random ones. But then to which position we should keep the centroids? So we change the position such that new position of centroid is the average of the points assign to the centroid.

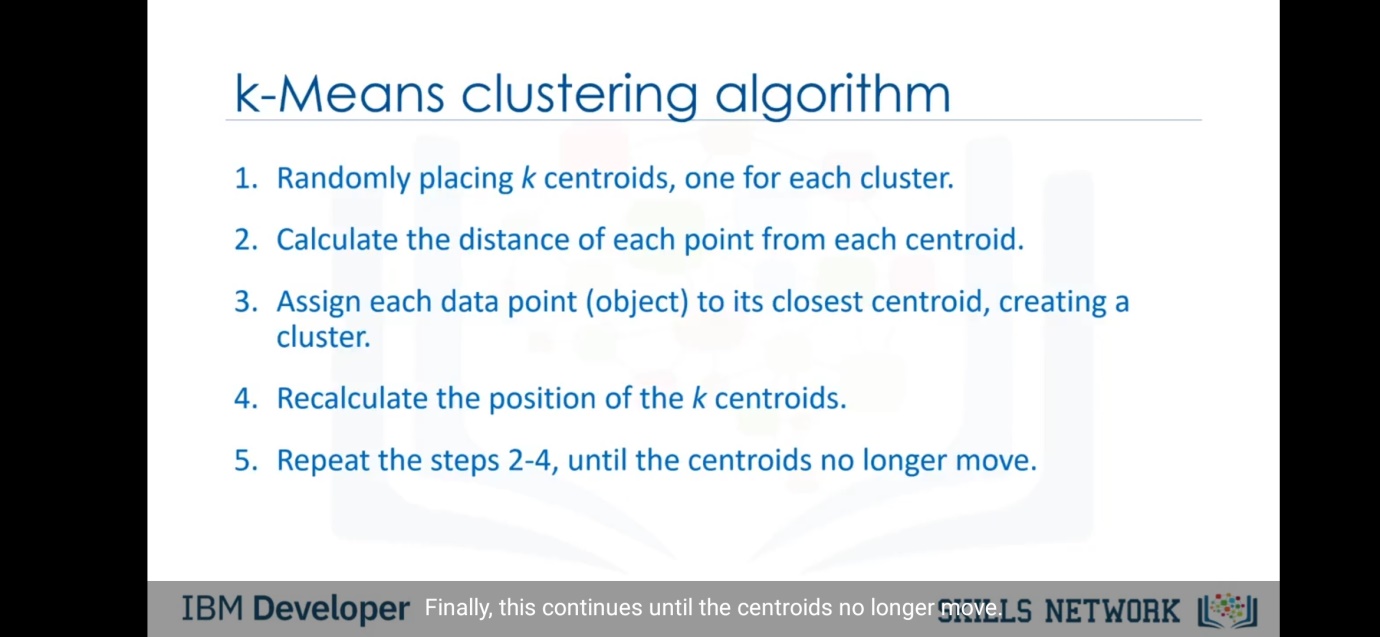
 Suppose as shown in example C2 has two points at (7.4,3.6) and (7.8,3.8) . So average of it is the new position of C2(7.6,3.6). Similiar to C1 and C3.

Now as the position of centroid is changed , the distance between the centroids and points are changed and so the distance matrix needs to change.

Step 5) Repeat until no more changes

Now we will calculate the distance matrix again, again we will assign the nearest centroid, and then change the position of centroids with respect to average of other points position. So we will repeat Step 2 to 4 repeatedly till there is minimum error and data is properly distributed in different clusters.

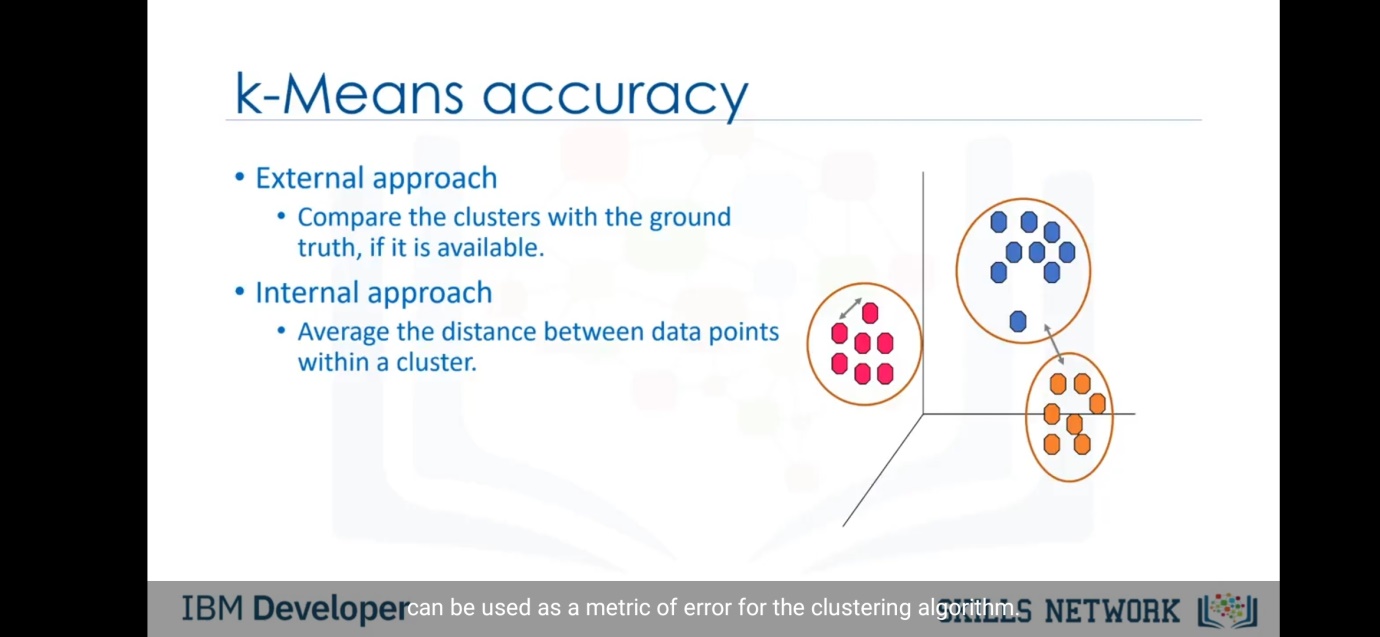


Hence the overview steps are as follows:-

Now we have created the clusters , but we have to cross check that the model is accurate or not. So to check the accuracy of this method we got two approaches.

1) External Approach :- If available , compare the results obtained with the ground truth i.e. actual results obtained by any specialist.

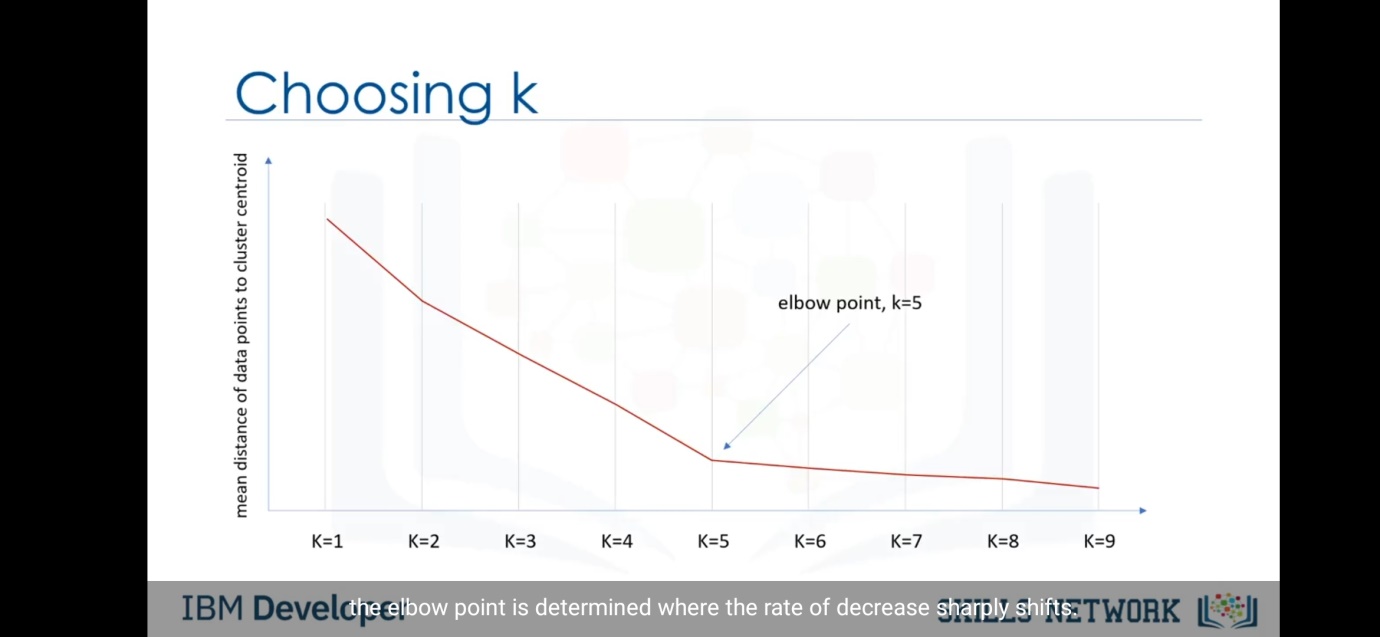
2) Internal approach :- Average between the data points distance within the cluster.



Now at the kickstart of this method we discussed that we will take the value of k as random to make our life easy. But the value of k is much important as it can change the accuracy score of the model. So we can select the value of k as follows :-

We will take k=1, and then will form clusters from the steps we discussed above. By that we will get one cluster. Now we will find the mean distance of data points to cluster centroid. Now we will increment the value of k by 1 i.e. k=2 and then repeat the same steps . Now if we keep this date in graph of mean distance v/s k we get such graph.



So we should select the value of k such that we get elbow point in this graph at that value of k.This method is called elbow method to select k. 



Now after discussing the portioning clustering , we shall proceed to the second type of clustering.

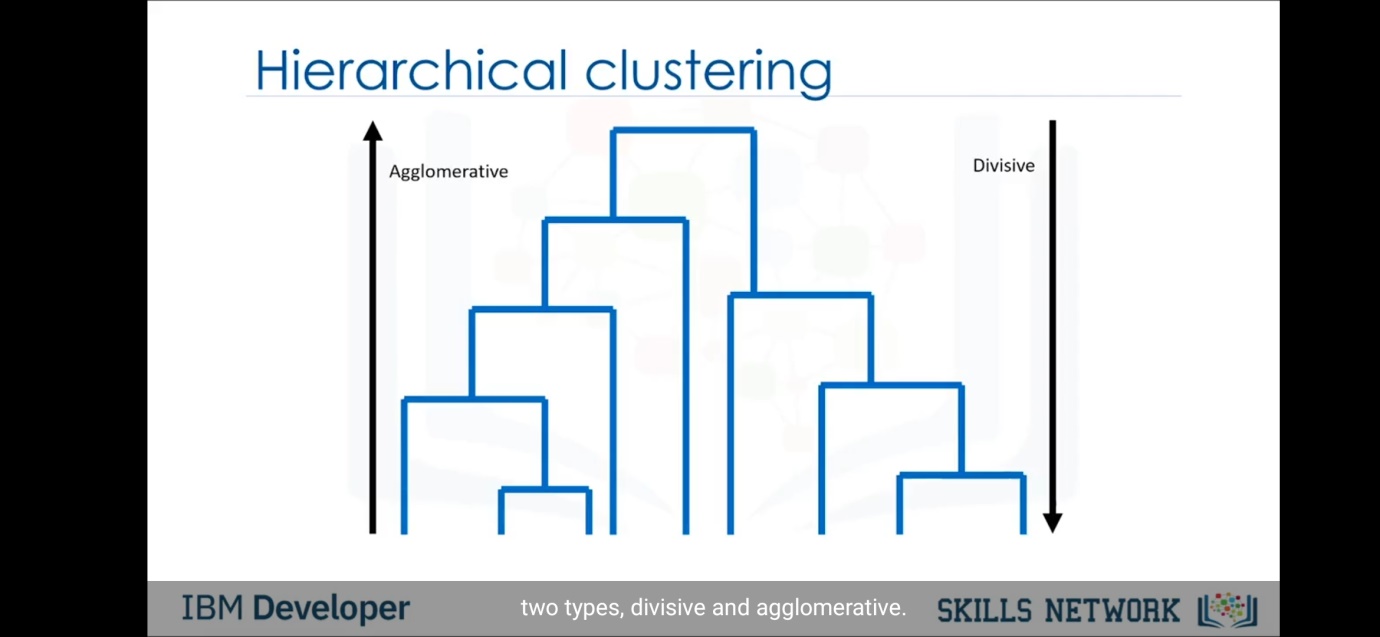
**HIERACHICAL CLUSTERING:-**

The basic techy definition of hierarchical clustering is as follows:-

“Hierarchical clustering algorithms build a hierarchy of clusters where each node is a cluster consists of the clusters of its daughter nodes . ”

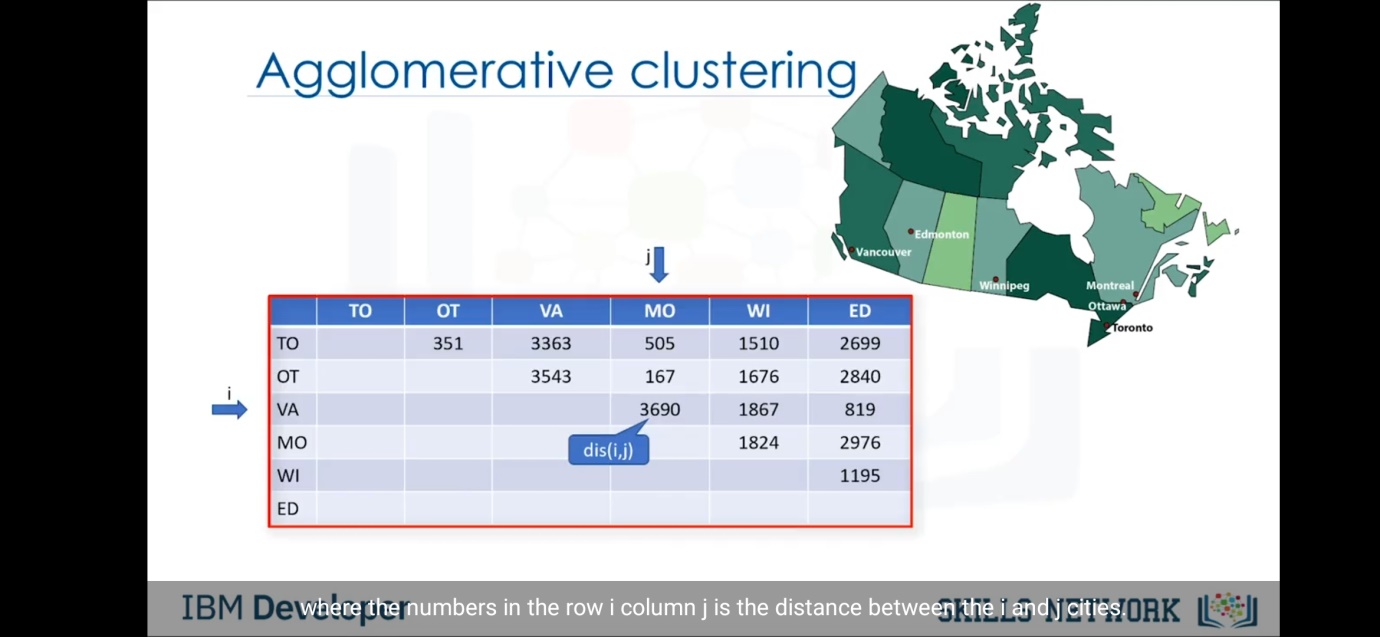
The below shown figure is example of hierarchical clustering. Its generally have a outlook as the trees. Also we have two types of hierarchical clustering :- Agglomerative and divisive. Agglomerative clustering is the bottom to up model and divisive is the one with top to bottom.

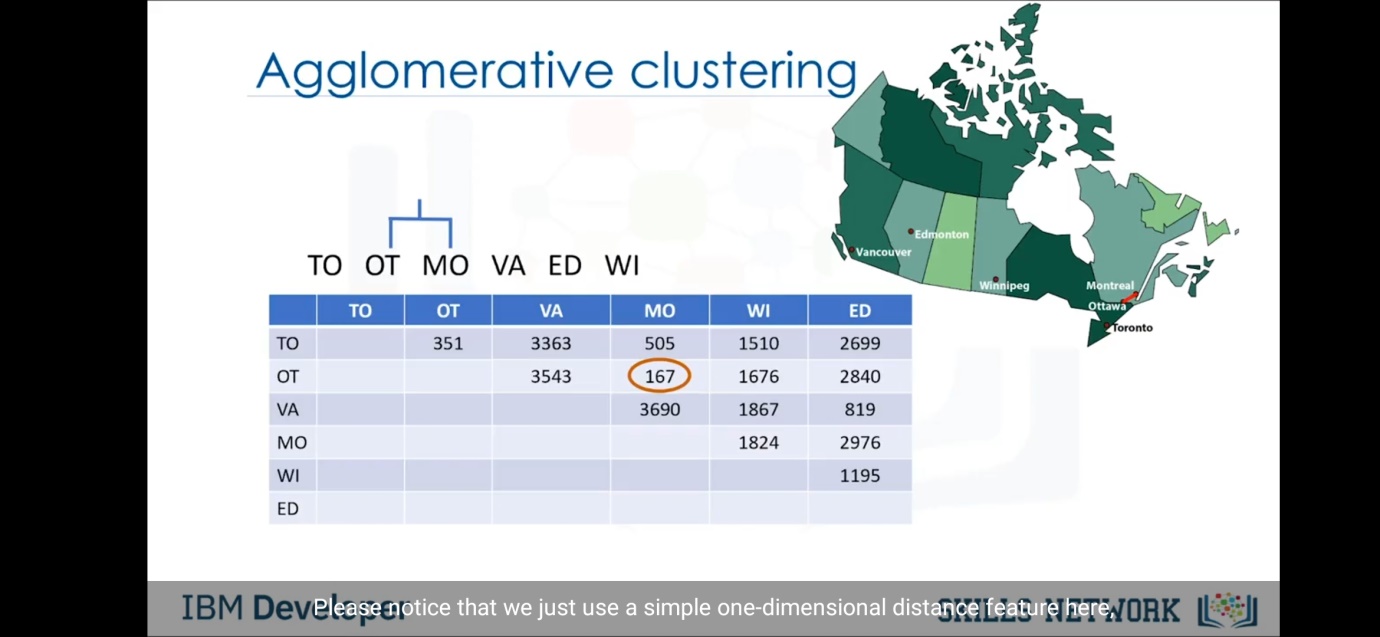
So in agglomerative we take the nodes as an individual nodes and with their similarity we cluster it. But in divisive we have a cluster already, in which we further divide them into similar parts till the nodes have the individual value.

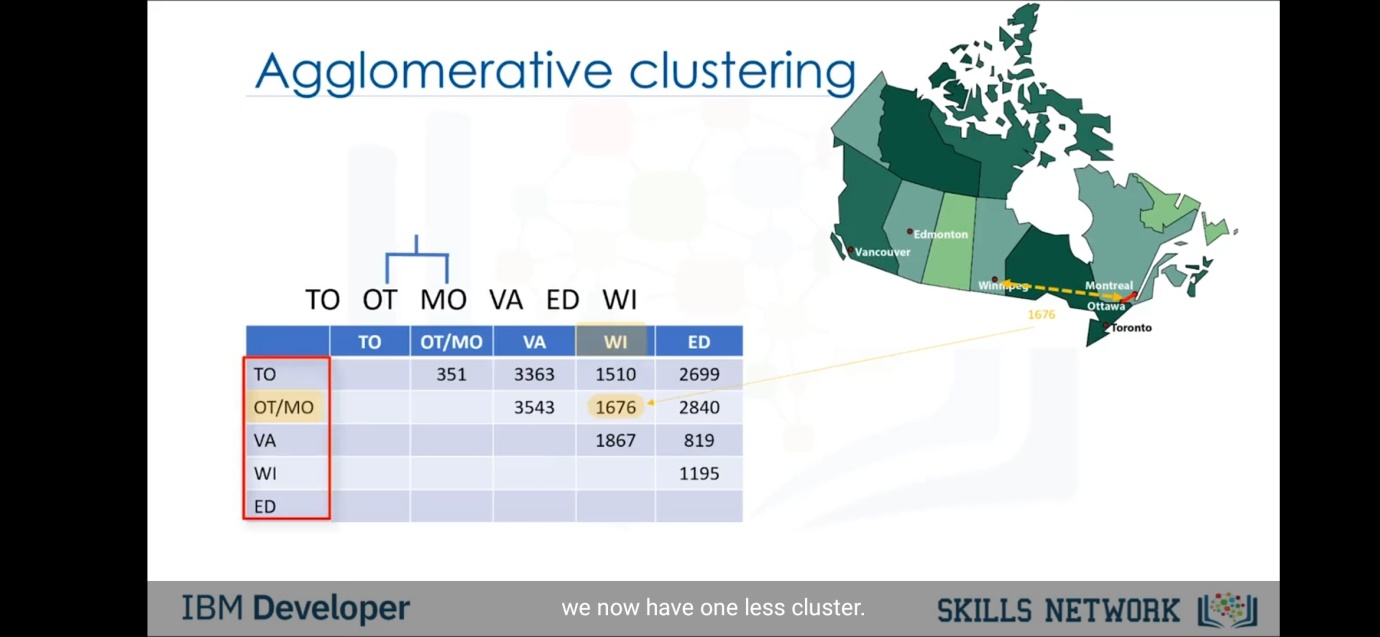


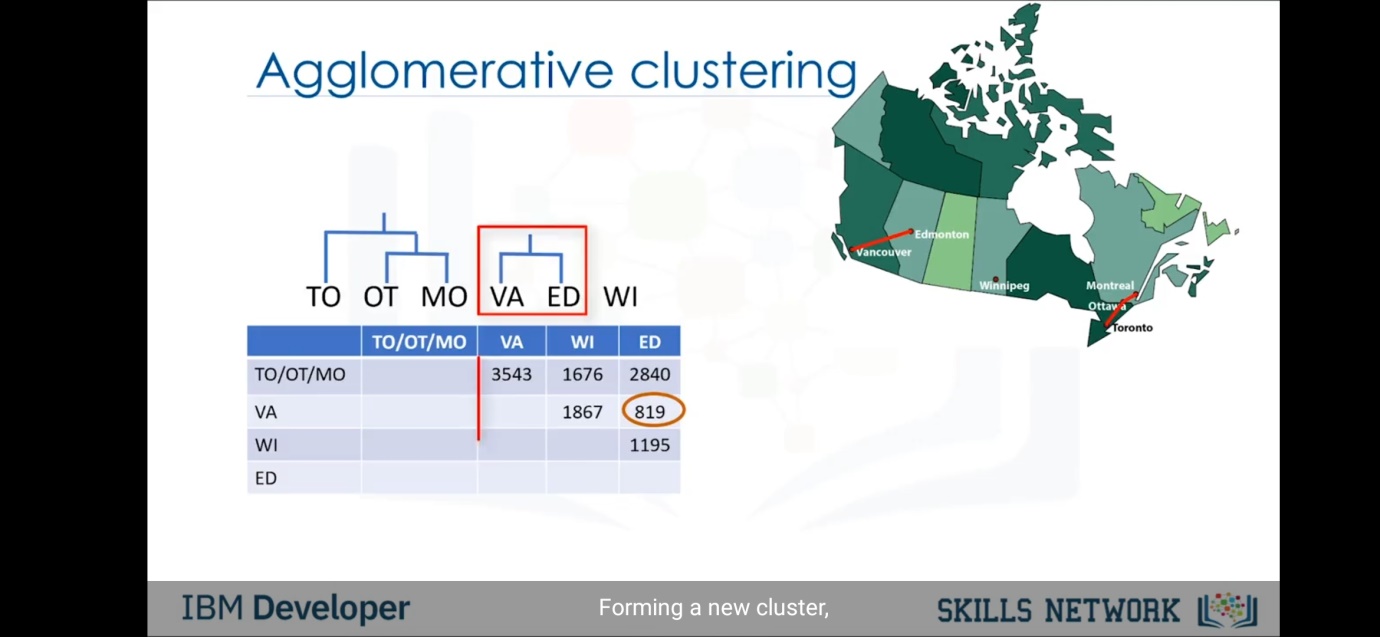
But in most of the case the agglomerative clustering is used. So we shall proceed for the example of agglomerative clustering .

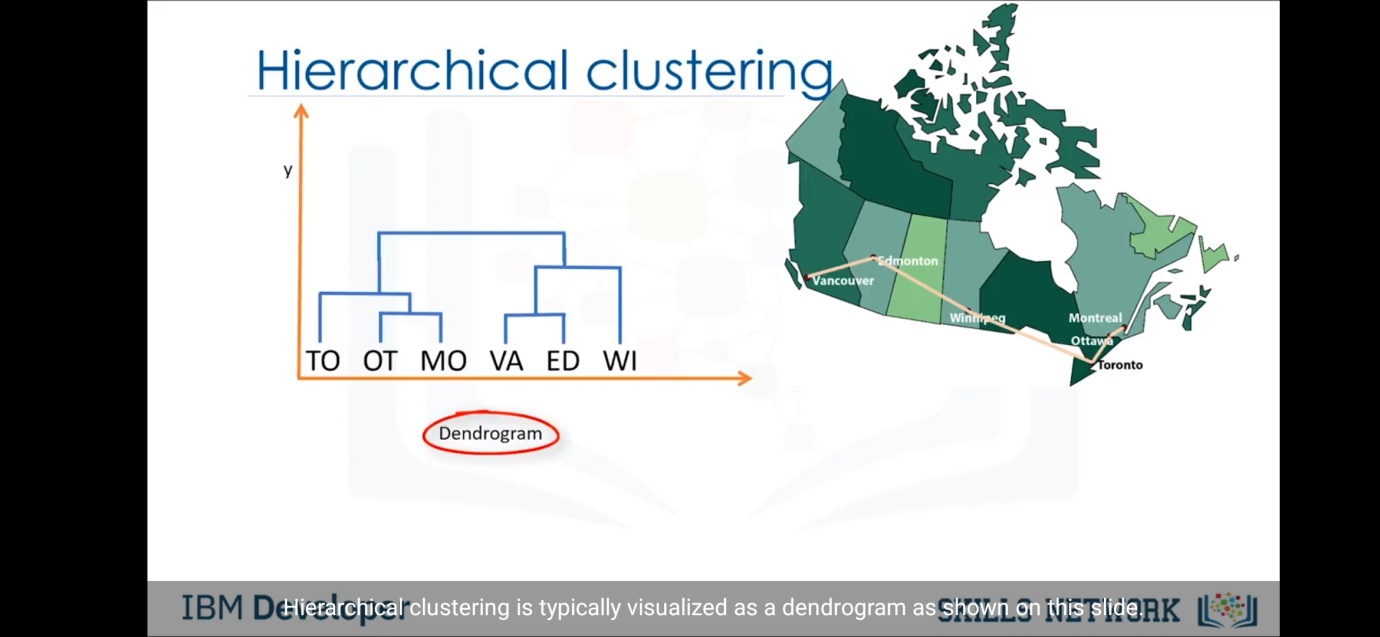
As shown in the figure , we have taken the example of the six states and we have to cluster them with their distance.

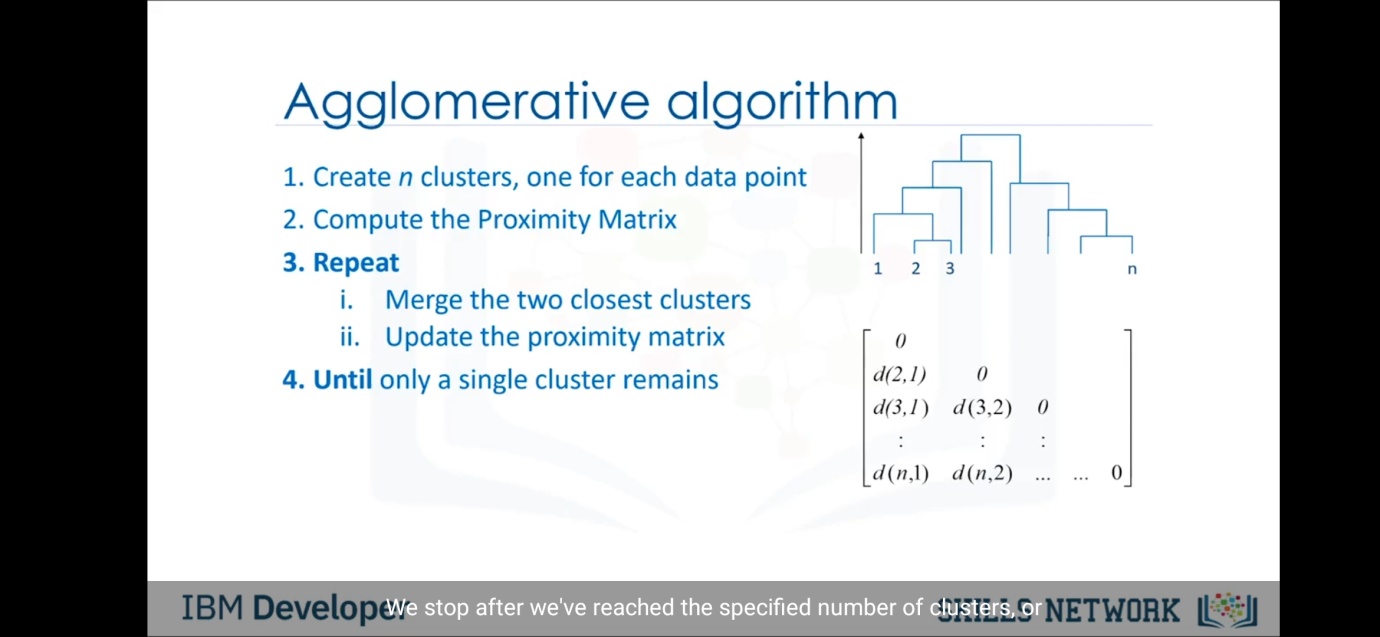
Now to cluster them up we will first make a matrix of n\*n where (i,j) is the distance between i and j respectively. 

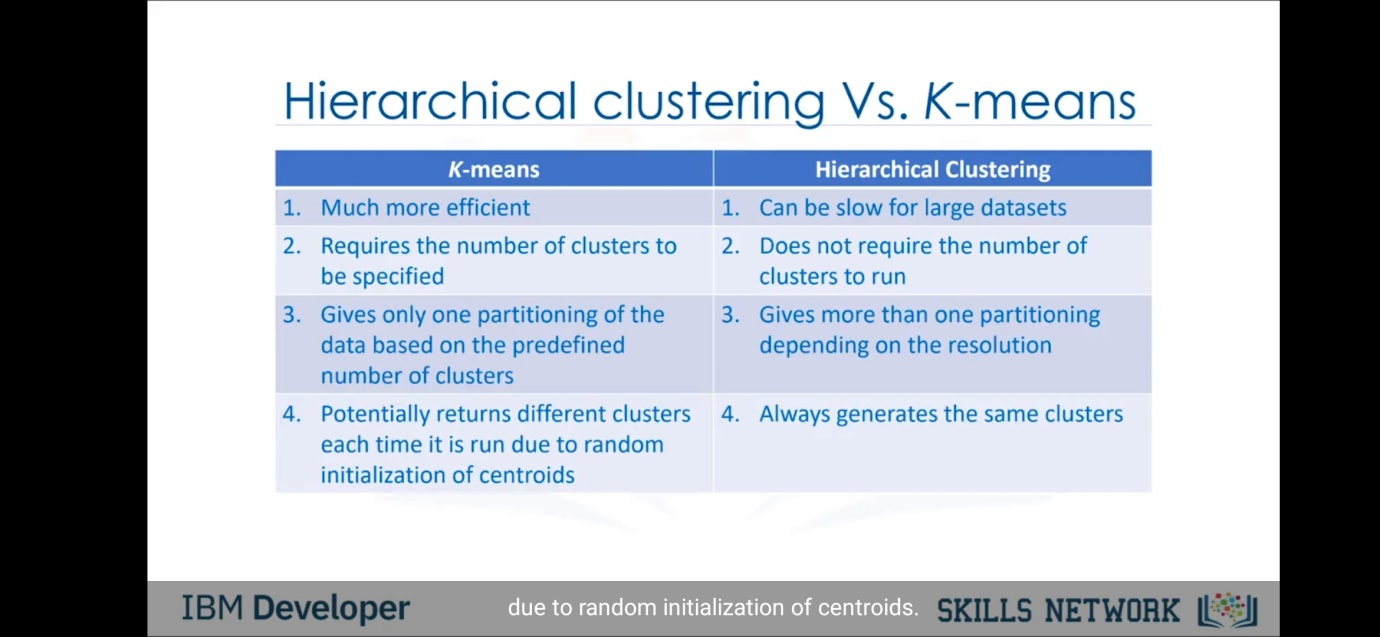
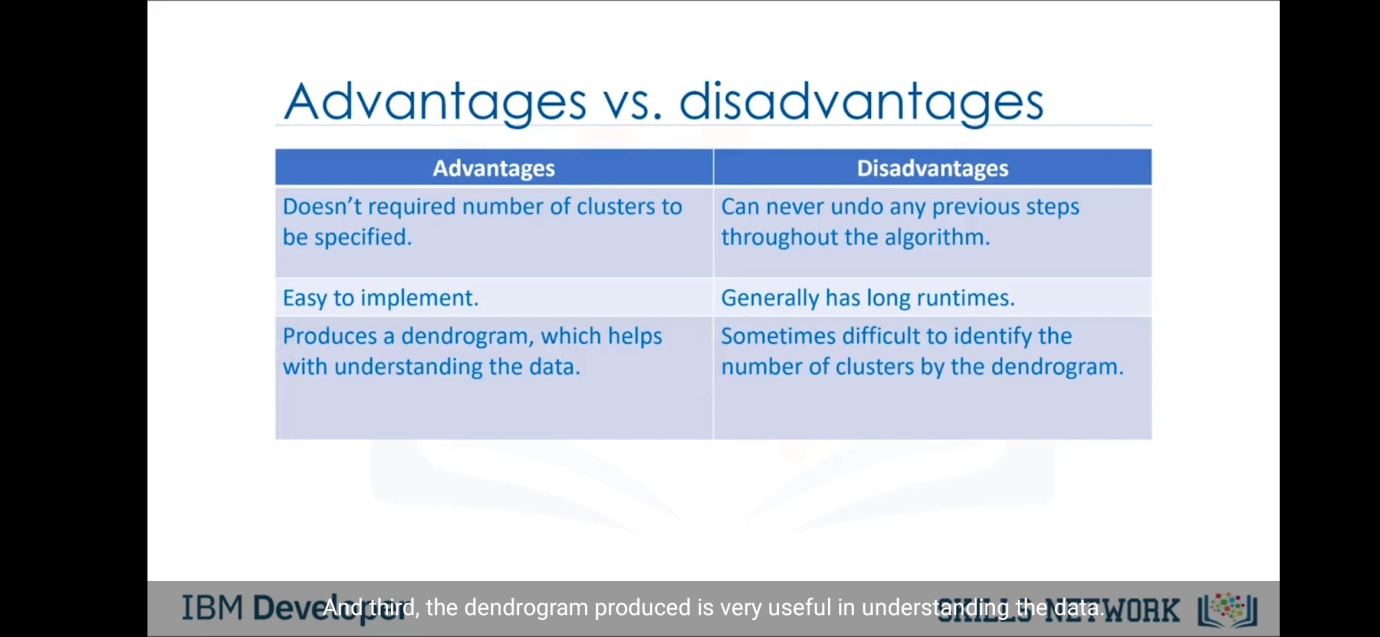
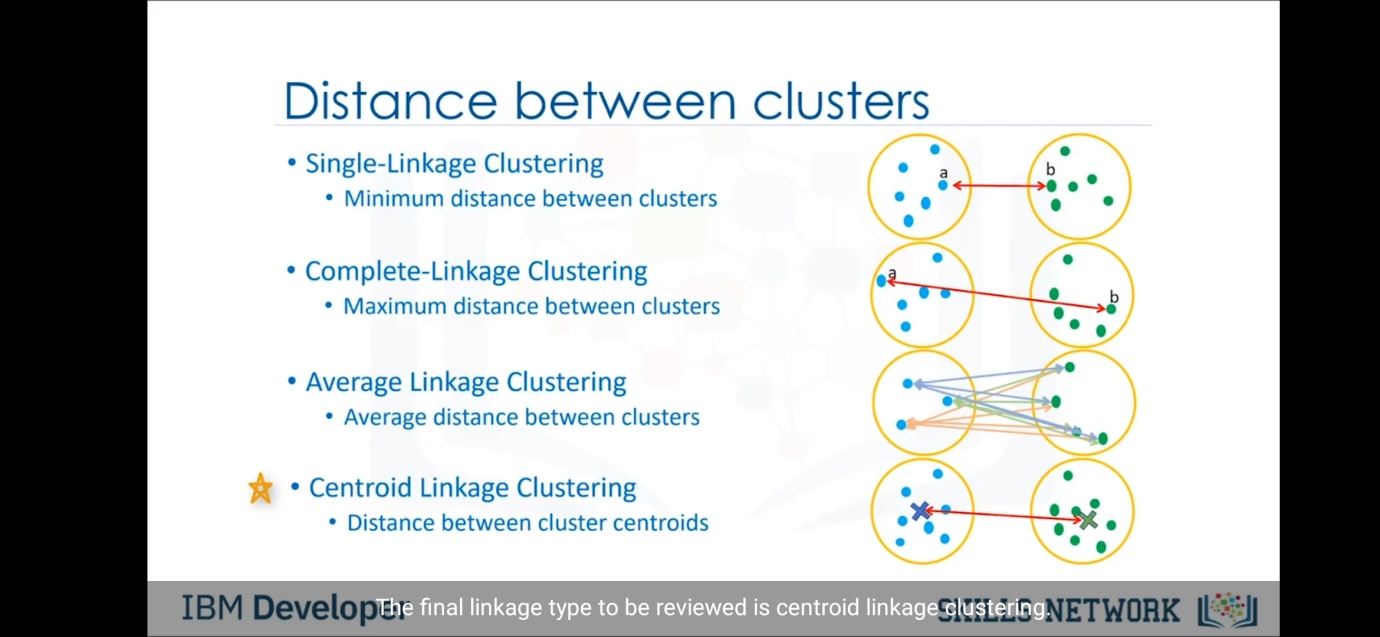
Also we will write all names as the nodes of the tree to make the clusters. After doing so we find the smallest distance. Here it is 167 of MO/OT. Here distance signifies the similarity ,the data varies with different case of examples. So now we will form a cluster of MO and OT as shown in figure.

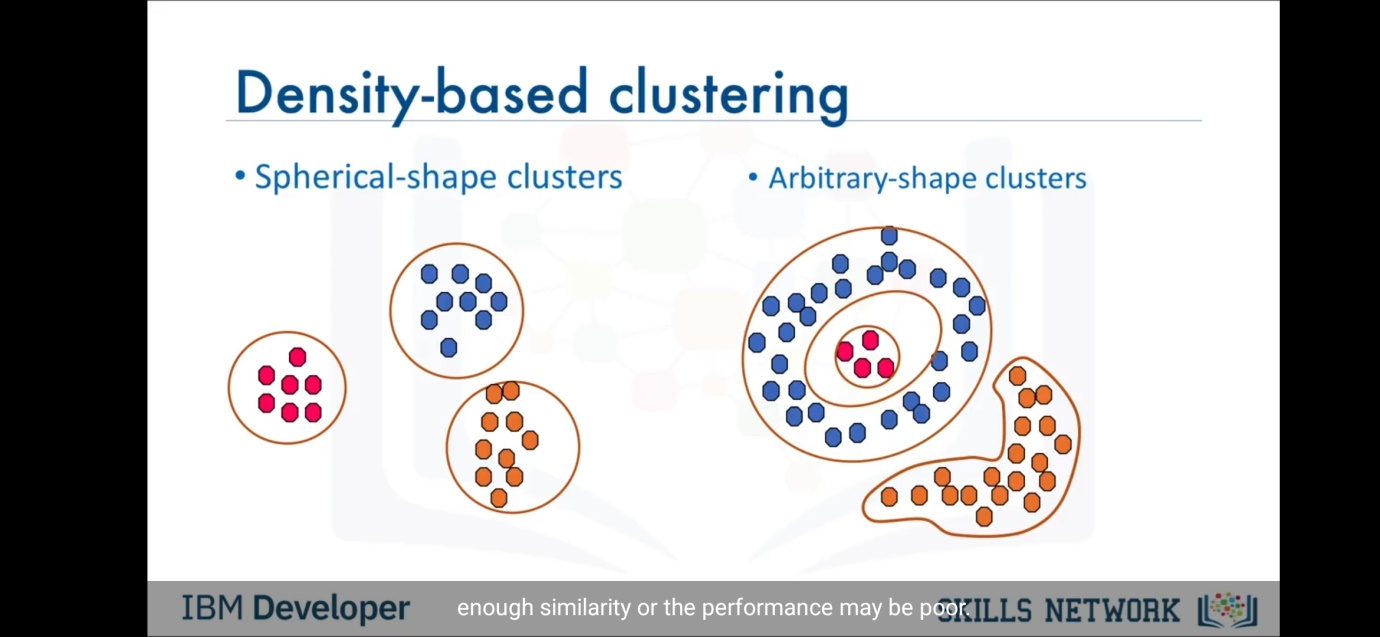
After forming the cluster , we will change the matrix and will remove the columns and rows of OT and MO and change with the OT/MO as they are marked for making cluster with each other. So the matrix is changed by calculating the distance again. But how to measure the distance of any city with OT/MO cluster. So generally machine learning practitioners take the mid point between OT/MO and then calculate the distance between other city and OT/MO.

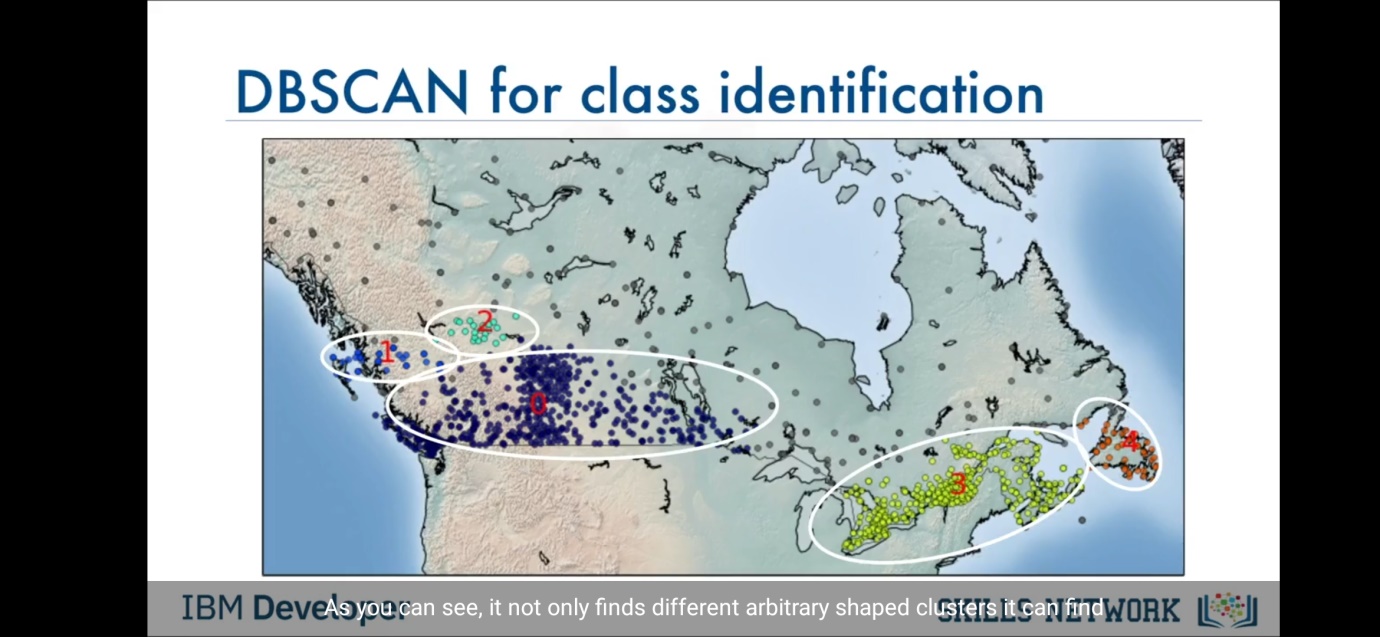
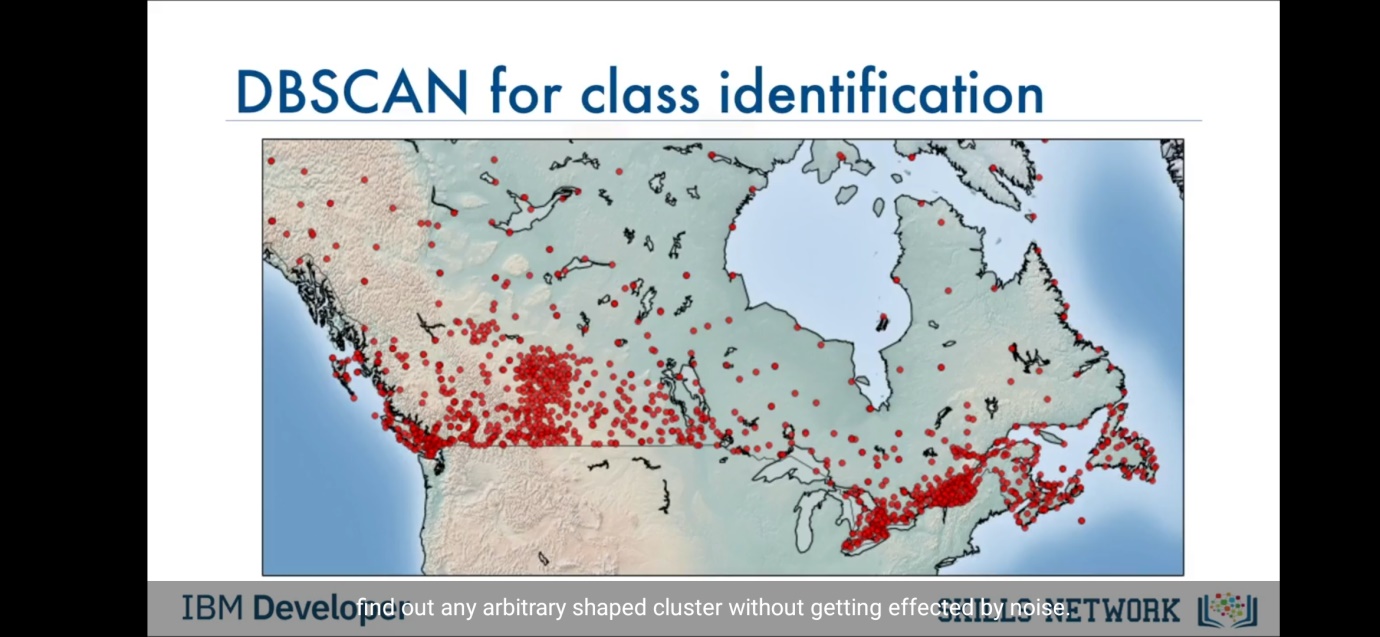
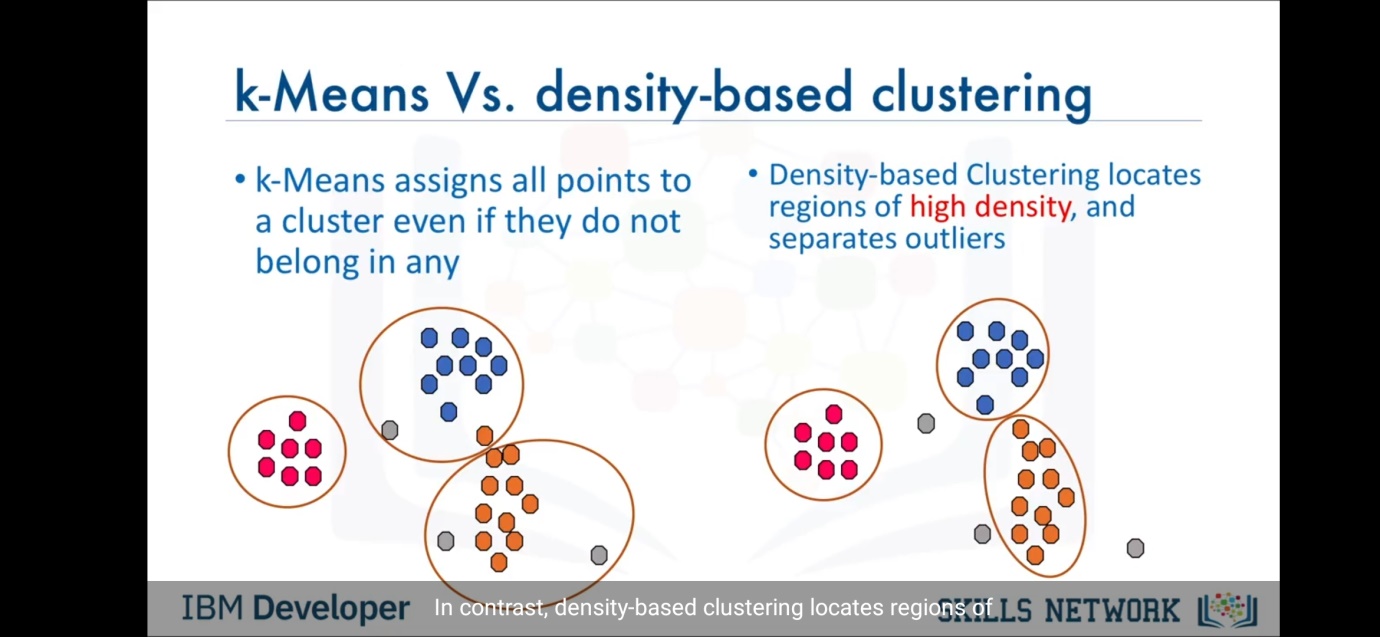
Now after changing the matrix, we will repeat the process again. Again find the smallest distance (here it is 819) and form the cluster of two cities with that distance. And repeat changing the matrix.

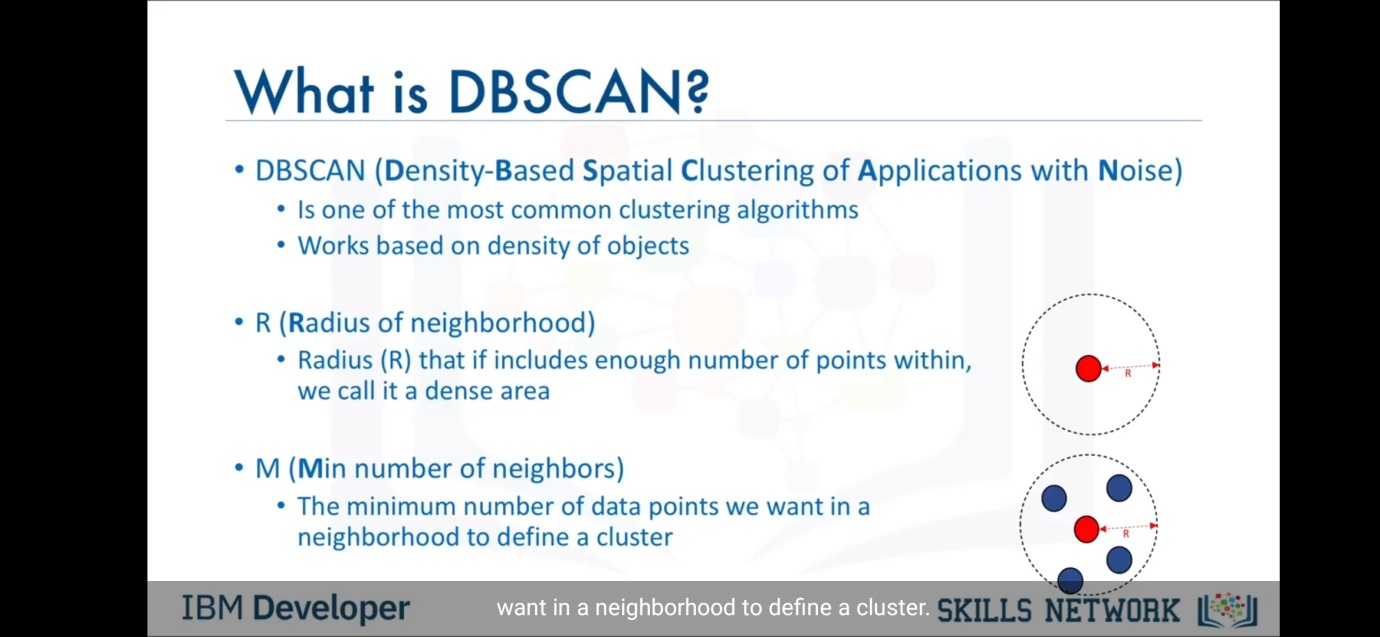
So after repeating the steps , we will reach to the stage when the different cities are converted into one whole cluster as shown in the figure. This whole cluster has a term also coined as dendogram. So we have generated a dendogram from different city data. 

The overview of the steps of agglomerative algorithm are as follows:-

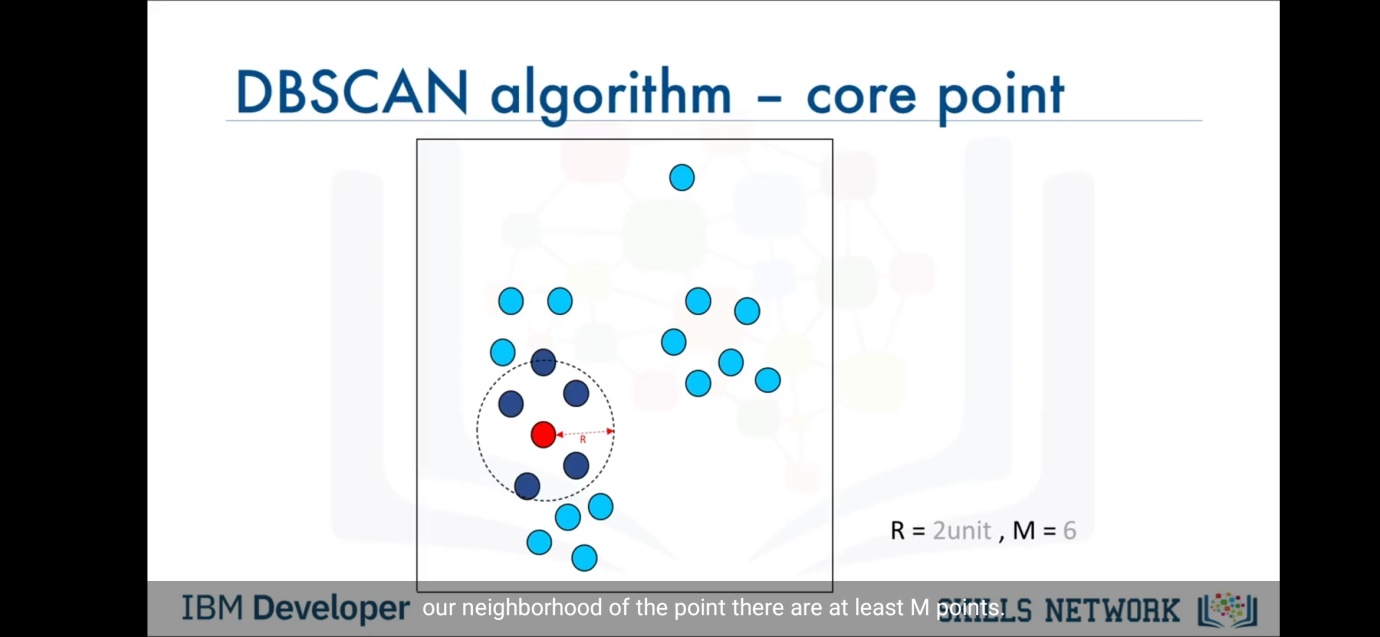
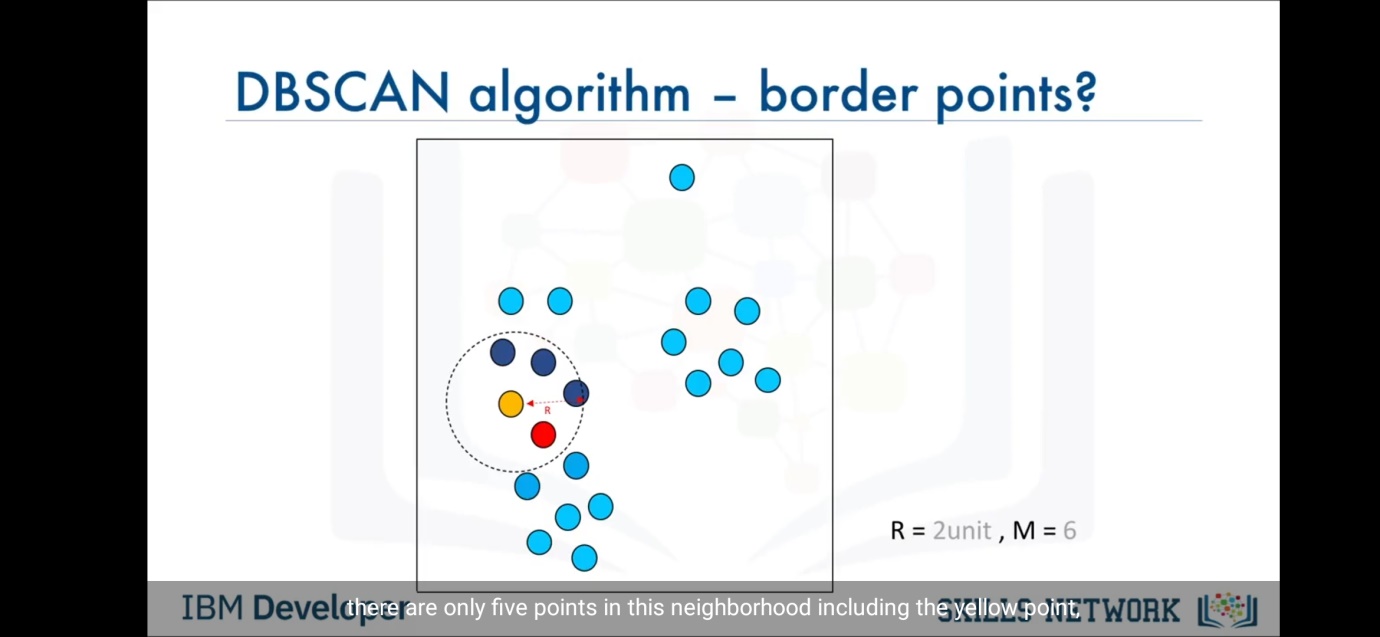
EXTRA DATA:- 

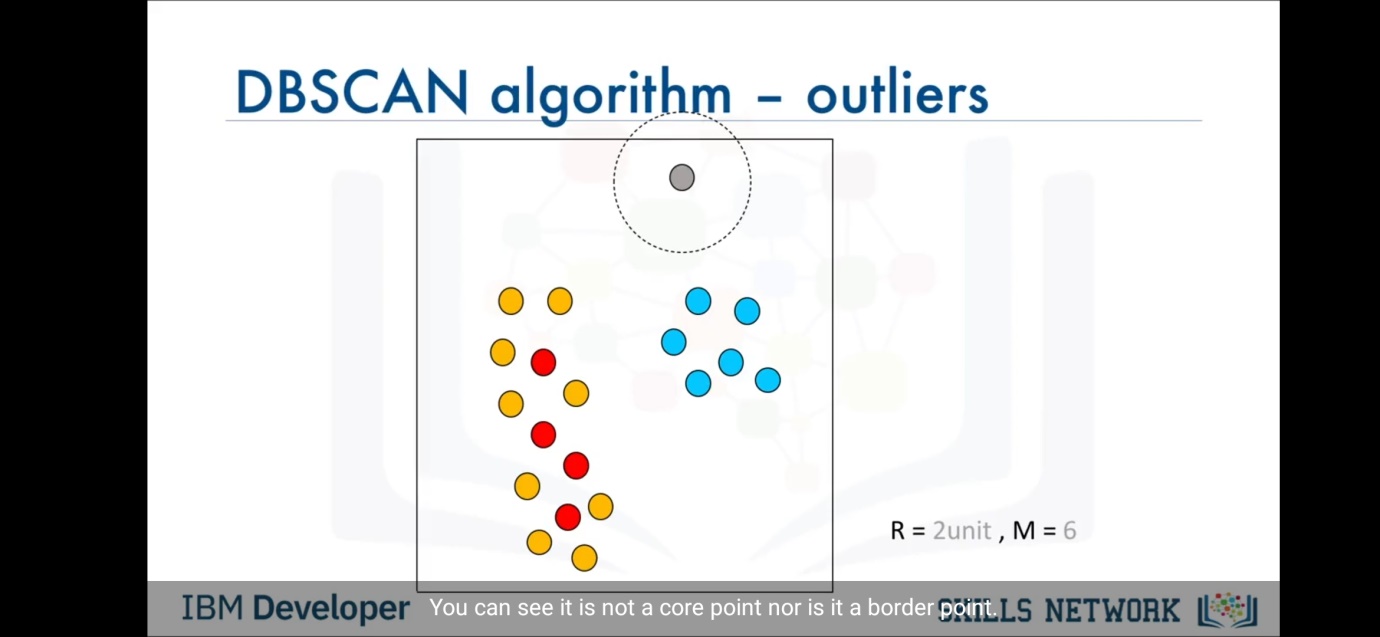
Now we will study the last type of clustering that is density based clustering. In our previous studies we have learned the k-means algorithm in which the shape of the cluster is spherical. But in this method the cluster can form arbitary shape. But how this change the cluster or affects the clustering? Let’s dive into this topic to find the answer. 

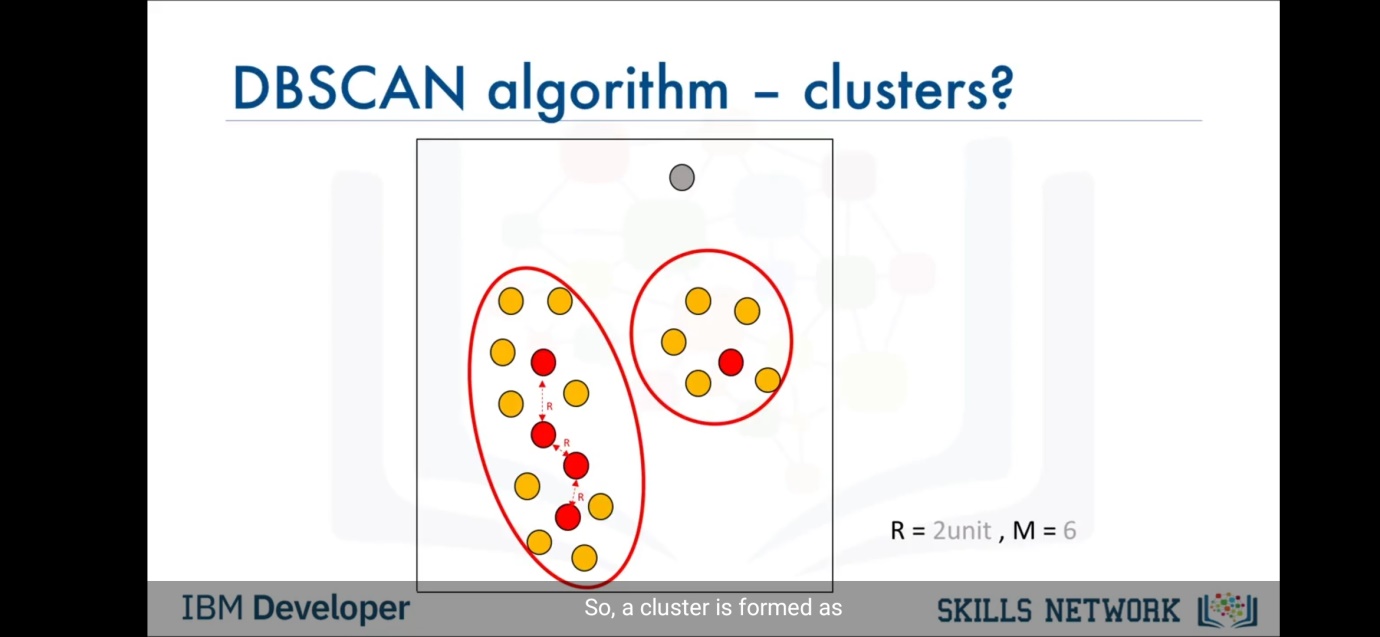
In k-means algorithm we have specific spherical shape clusters, so it gets bound to take the points which comes in the radius of sphere to be the part of cluster even if its characteristic is not similar to the data points of that respective cluster. So this generates the error in the model. This points are generally called the outlier among the data points. So we get bounded by the specific spherical shape to include the outlier in clusters. Hence it can be good if we form cluster of arbitary shape which doesn’t include outlier and hence can reduce the error or noise. So we here take the density as subject of clustering, we cluster the points in such cluster that the density of cluster is high. 

General information of DBSCAN:-

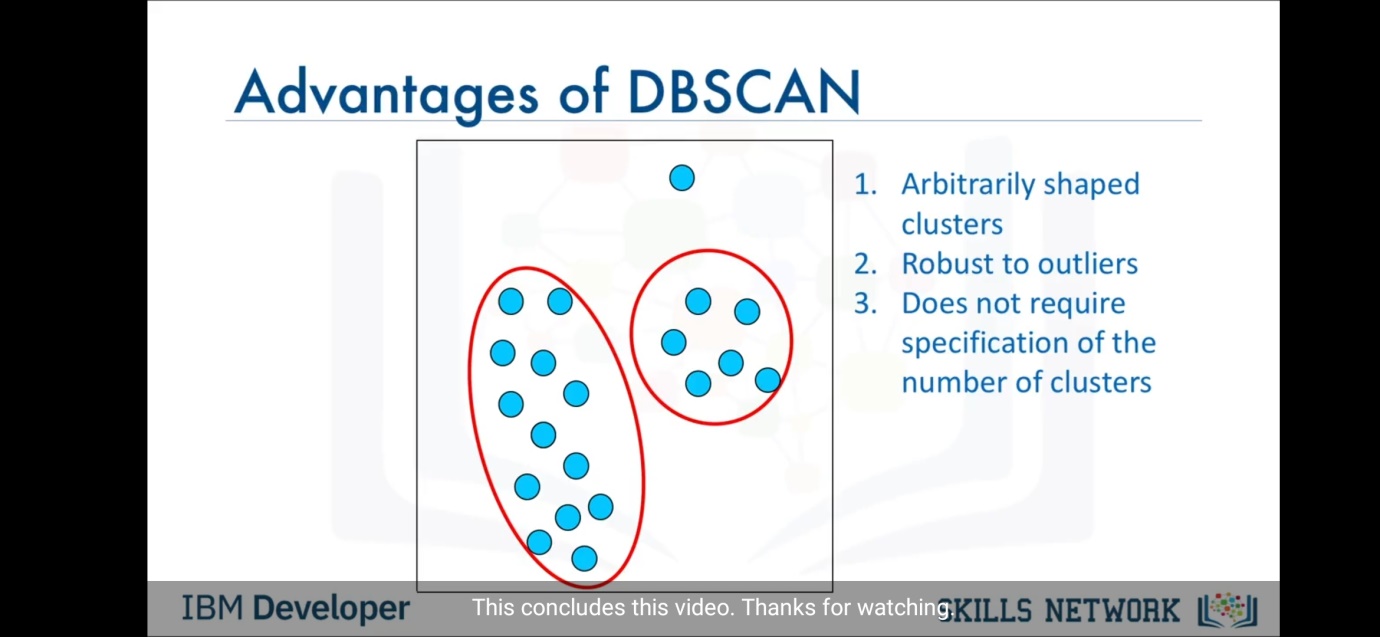
Now as shown in figure , we take a example and cluster it using DBSCAN.

So for that , we initially have given the value of radius and M. R=2 and M=6. Now we will select a random point, suppose the red point shown in the figure is the point we selected initially. Now we will draw a circle of given radius R(here given 2). Now if this circle consist of M (here 6) points inside it, then this point is called core point. Here this point is core point because threquirements are satisfied. Now we will select another point. Here shown yellow point is the one we selected. Now we check again by drawing circle of R radius,but here does not satisfy the requirement of M points in the circle.So this cannot be a core point . But it has a point which is core point inside the circle, so this point is called as border point. 

Now suppose we select a point again. Suppose we select the grey point as shown in figure and again draw a circle of radius 2. But here we can see that its niether core nor border point. So this point is called outlier point. Hence this point should not be added in any of the clusters.

Now we will repeat the algorithm on each point and we will get figure as shown. So now we can form the cluster such that the distance between two core points is less than or equal to R. If its so then we add that two core points and there respective border points into a cluster . So as shown we will get this type of two cluster. Here we can observe that the core point at the right side has the distance from any core point in data with more than R distance . So it forms different cluster. And left core points are having distance less than R among themselves so the form one clusters.

Here the best thing of this method is that we do not have to get or select a random value k and start and then again change the value of k and check accuracy of model. Also the outlier are not added because of the arbitary shape allowance.

So the advantages of DBSCAN are as follows:- 

This concludes the clustering concept. Thanks for reading. Happy Clustering ☺ .