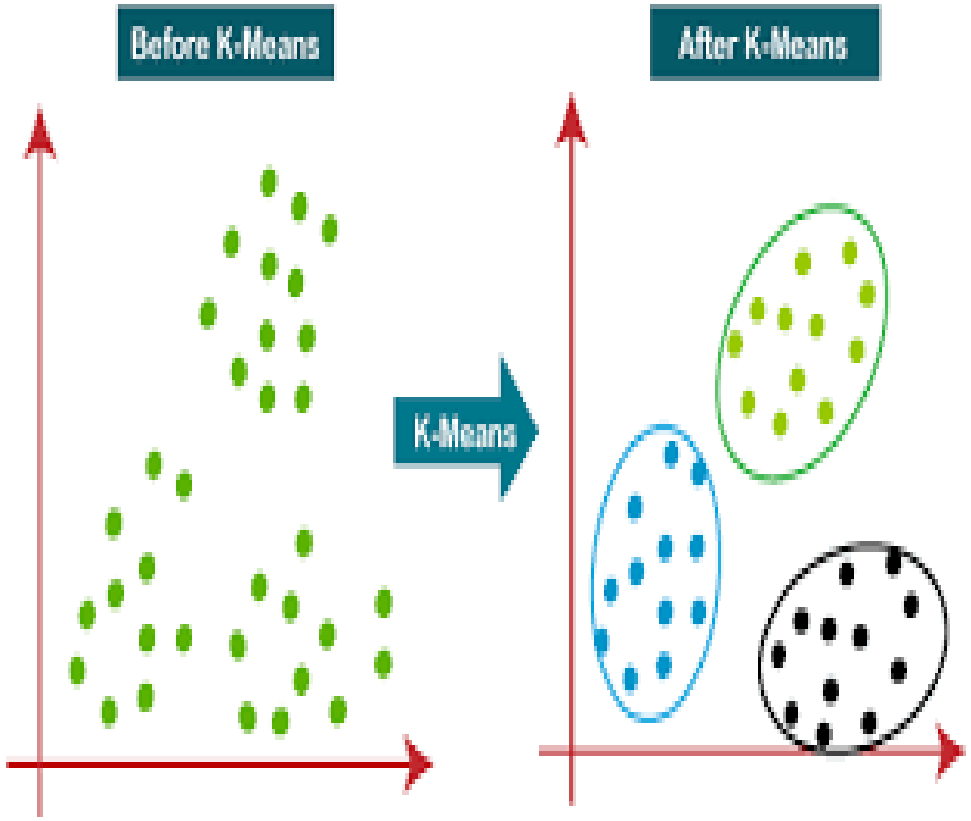
- **K-Means Clustering Algorithm**

Dr. Jagendra Singh

Machine Learning

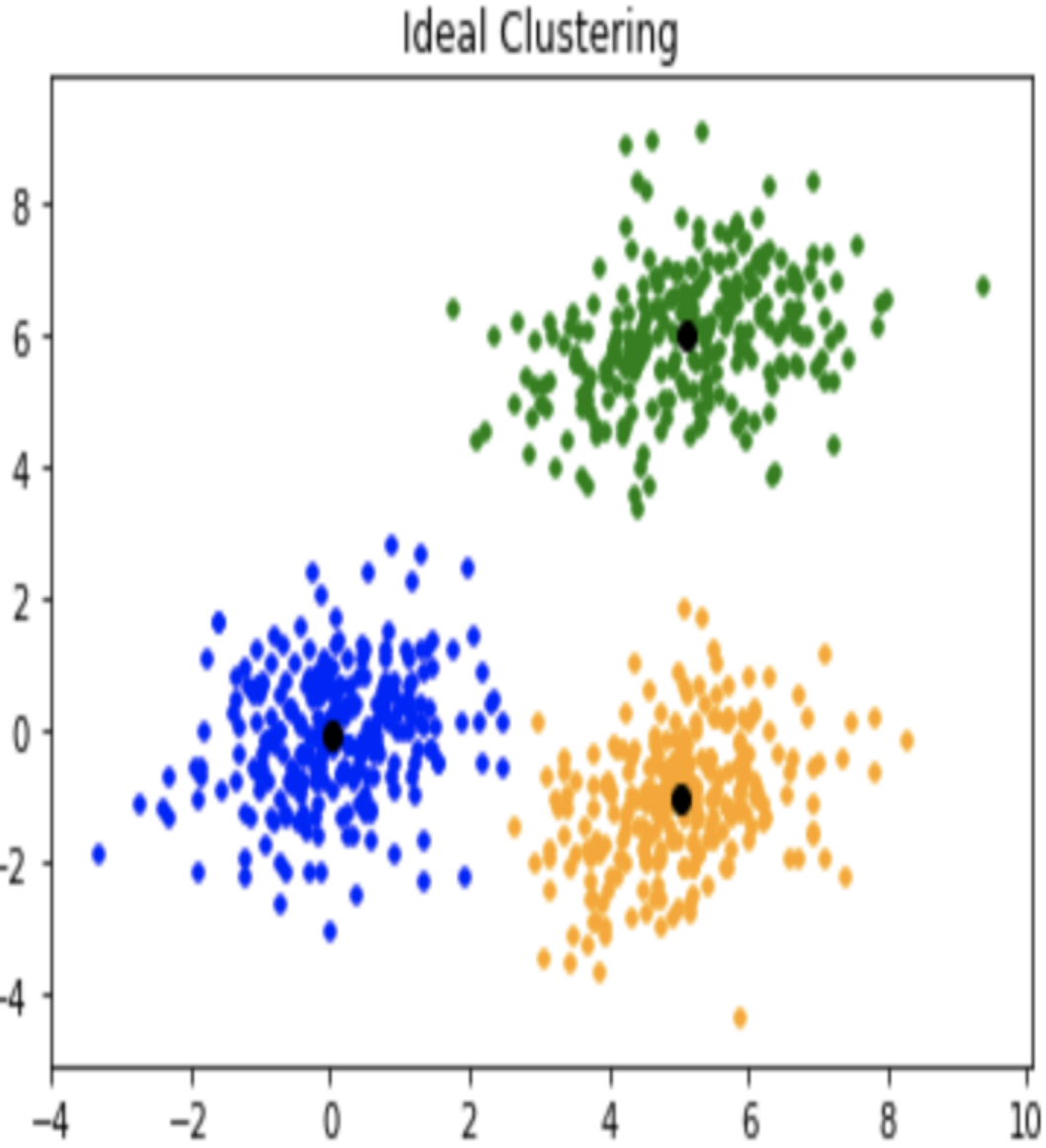


K-MEANS CLUSTERING ALGORITHM

- K-Means Clustering is an unsupervised learning algorithm that is used to solve the clustering problems in machine learning or data science.
- In this session, we will learn what is K-means clustering algorithm, how the algorithm works, along with the Python implementation of k-means clustering.

WHAT IS K-MEANS ALGORITHM

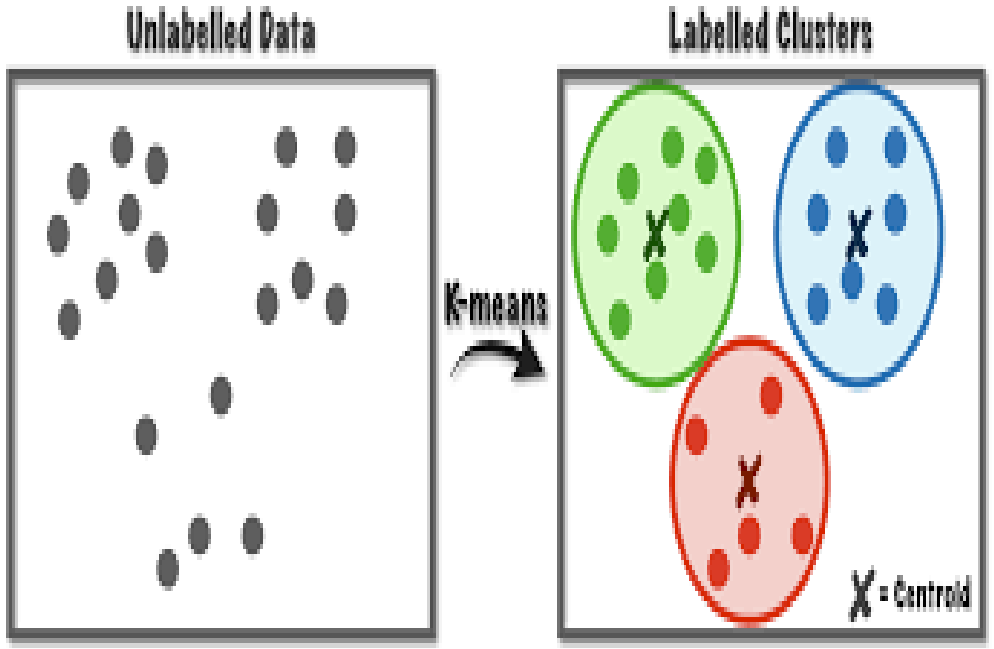
- K-Means Clustering algorithm groups the unlabeled dataset into different clusters.
- Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.
- It is an iterative algorithm that divides the unlabeled dataset into k different clusters in such a way that each dataset belongs only one group that has similar properties.



WHAT IS K-MEANS ALGORITHM

- It is a centroid-based algorithm, where each cluster is associated with a centroid.
- The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters.

K-MEANS ALGORITHM

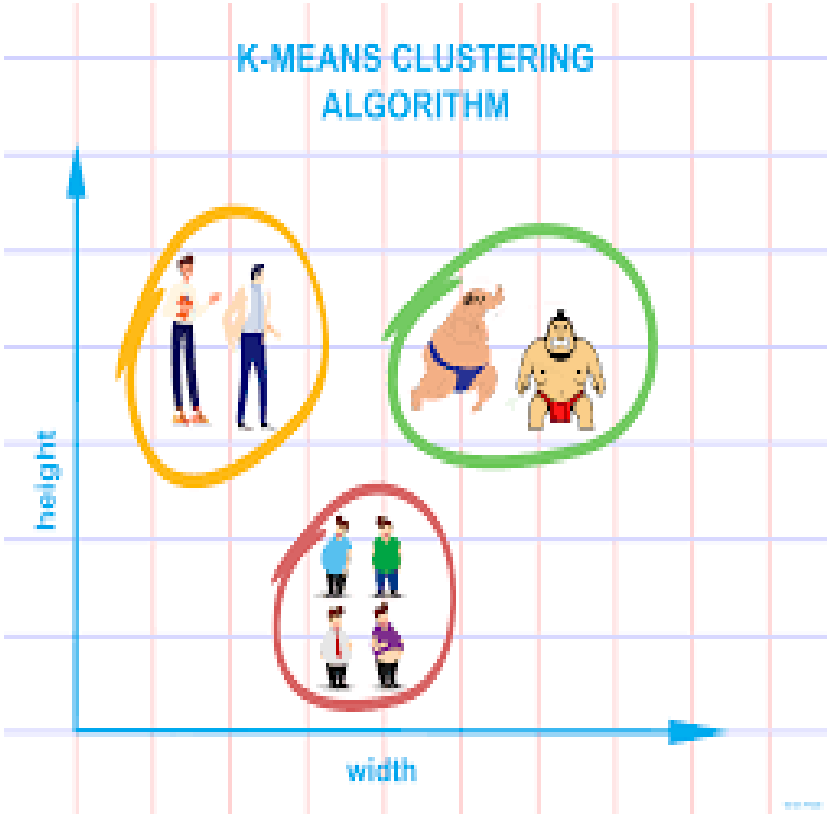


- The k-means clustering algorithm mainly performs two tasks:
  - Determines the best value for K center points or centroids by an iterative process.
  - Assigns each data point to its closest k-center. Those data points which are near to the particular k-center, create a cluster.
  - Hence each cluster has datapoints with some commonalities, and it is away from other clusters.
  - The below diagram explains the working of the K-means Clustering Algorithm:

HOW DOES THE K-MEANS ALGORITHM WORK?

The working of the K-Means algorithm is explained in the below steps:

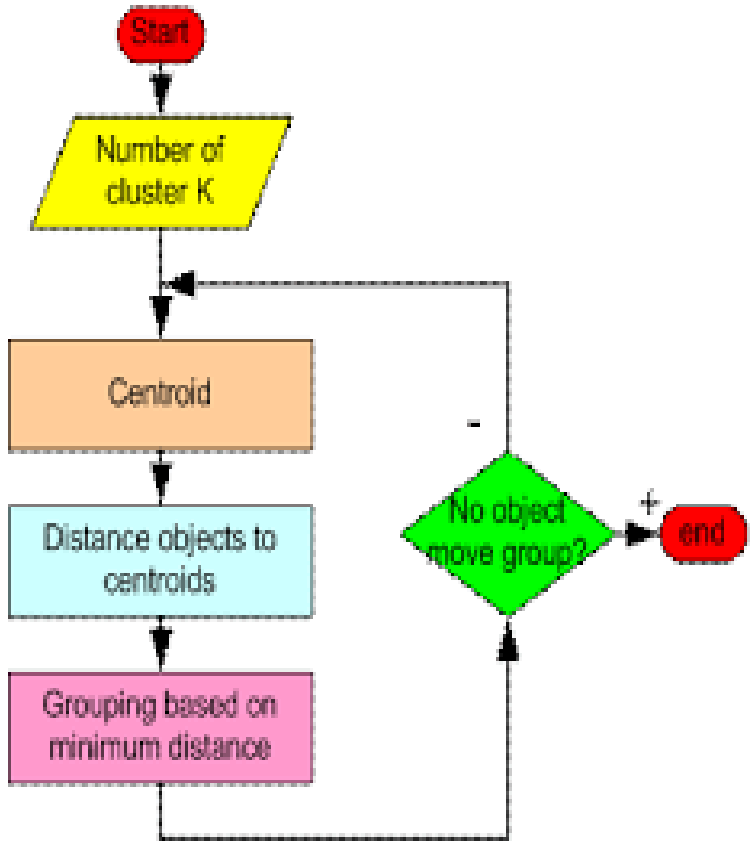
- **Step-1:** Select the number K to decide the number of clusters.
- **Step-2:** Select random K points or centroids. (It can be other from the input dataset).
- **Step-3:** Assign each data point to their closest centroid, which will form the predefined K clusters.
- **Step-4:** Calculate the variance and place a new centroid of each cluster.



HOW DOES THE K-MEANS ALGORITHM WORK?

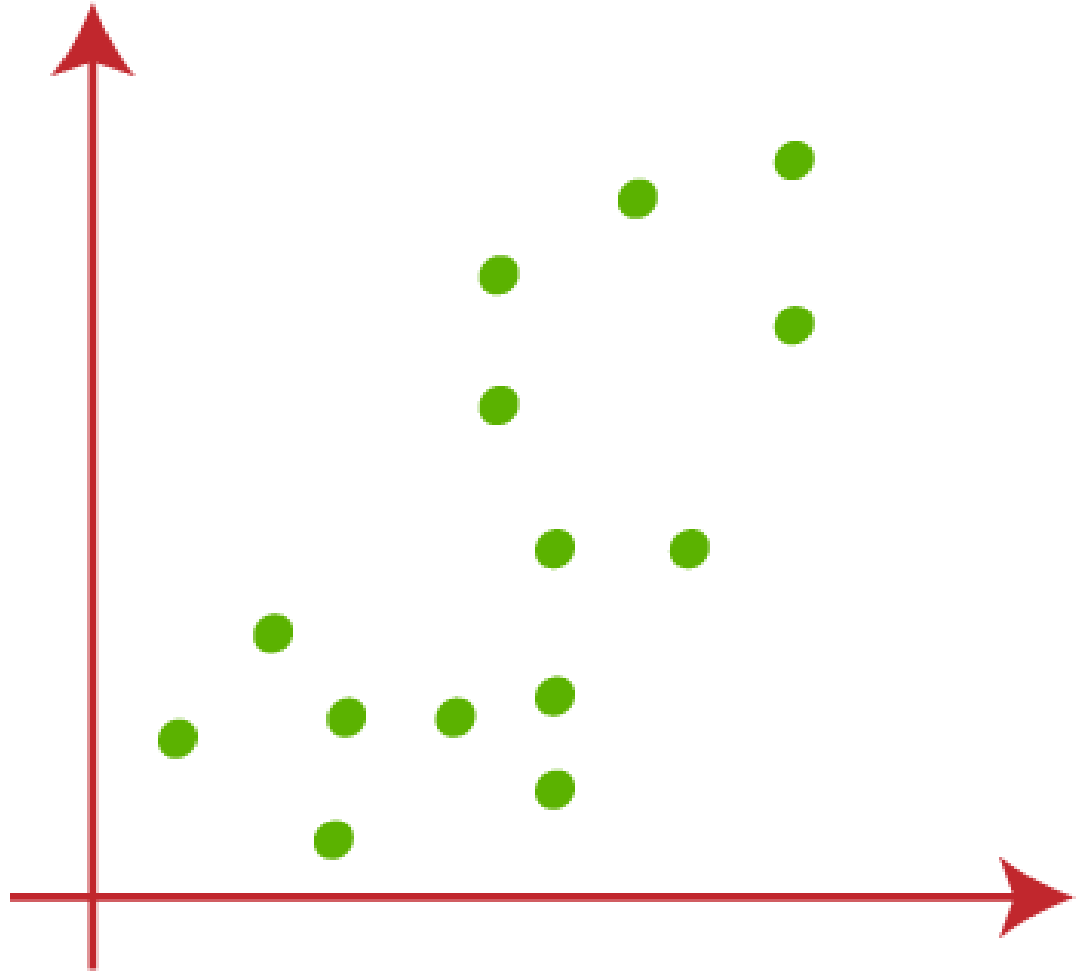
The working of the K-Means algorithm is explained in the below steps:

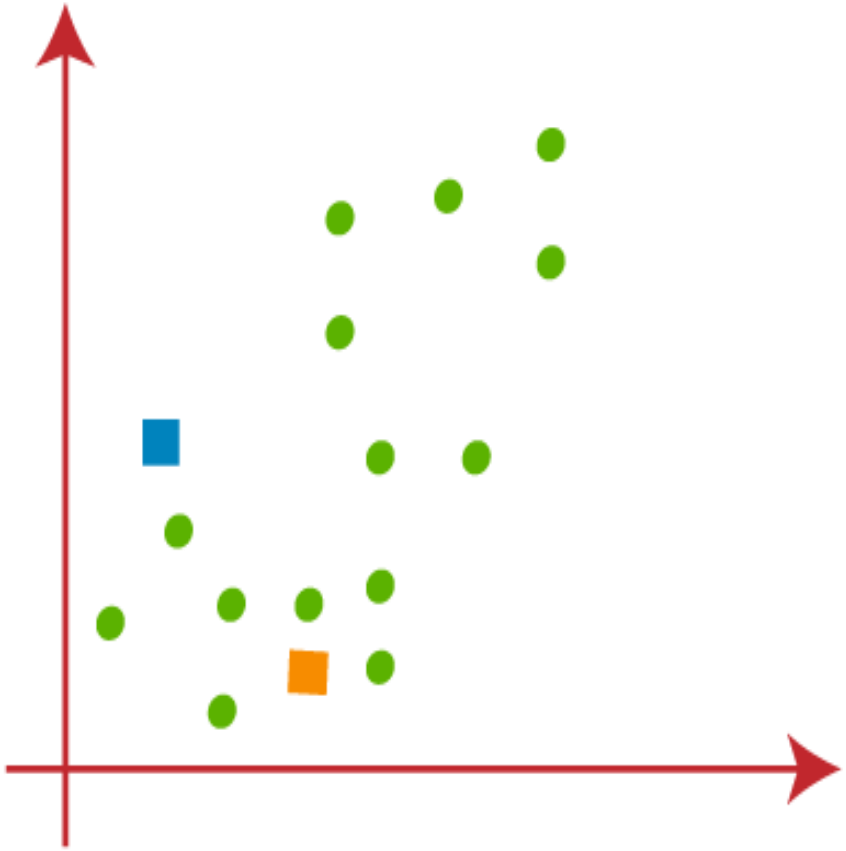
- **Step-5:** Repeat the third steps, which means reassign each datapoint to the new closest centroid of each cluster.
- **Step-6:** If any reassignment occurs, then go to step-4 else go to FINISH.
- **Step-7**: The model is ready.



HOW DOES THE K-MEANS ALGORITHM WORK?

- Let's understand the above steps by considering the visual plots:
- Suppose we have two variables M1 and M2. The x-y axis scatter plot of these two variables is given below:





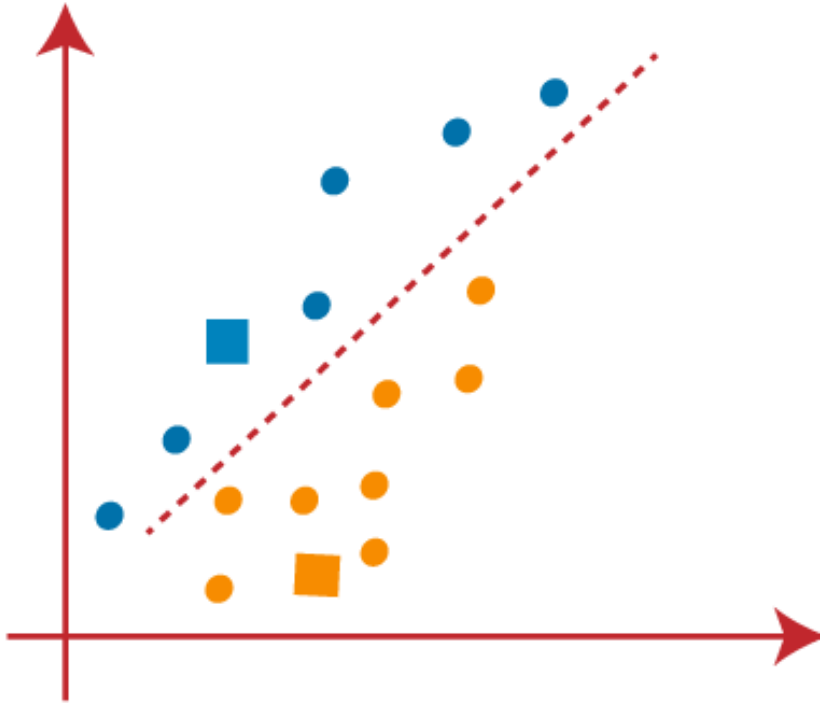
K-MEANS ALGORITHM WORK

- Let's take number k of clusters, i.e., K=2, to identify the dataset and to put them into different clusters.
- It means here we will try to group these datasets into two different clusters.

- We need to choose some random k points or centroid to form the cluster.
- These points can be either the points from the dataset or any other point.
- So, here we are selecting the below two points as k points, which are not the part of our dataset.
- Consider this image:

K-MEANS ALGORITHM WORK

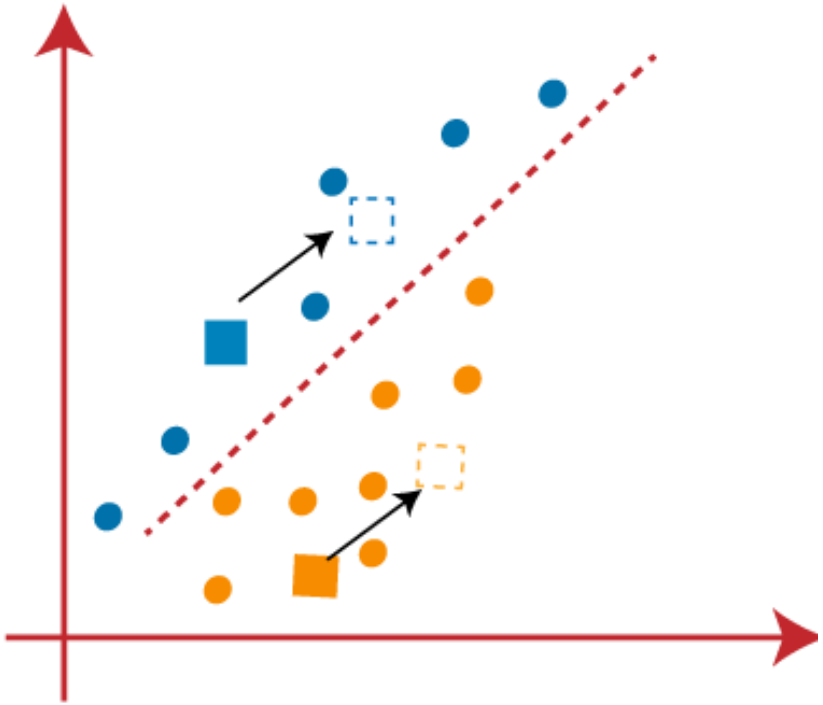
- From the above image, it is clear that points left side of the line is near to the K1 or blue centroid, and points to the right of the line are close to the yellow centroid.
- Let's color them as blue and yellow for clear visualization.



K-MEANS ALGORITHM WORK

- As we need to find the closest cluster, so we will repeat the process by choosing **a new centroid**.

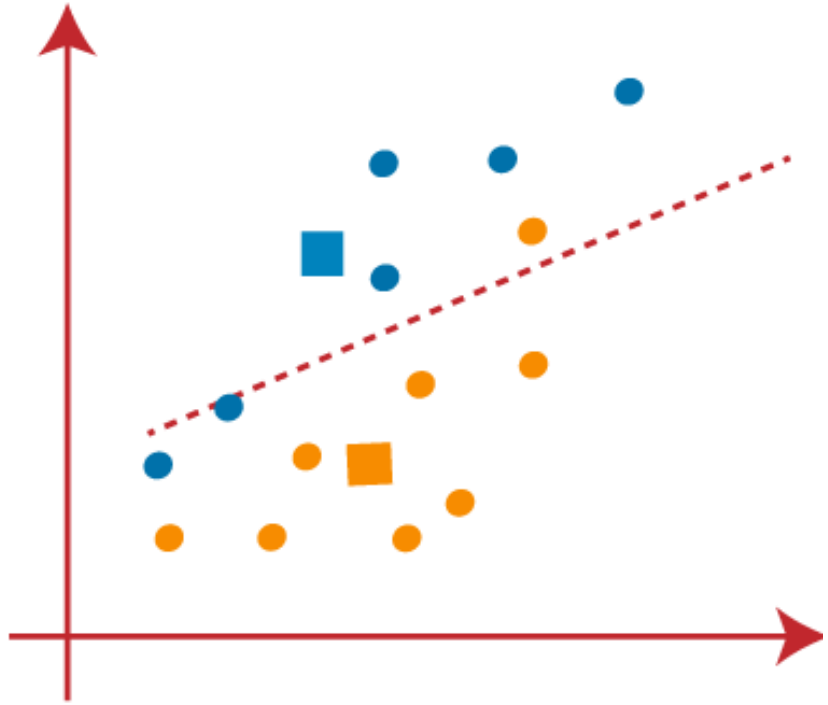
- To choose the new centroids, we will compute the center of gravity of these centroids, and will find new centroids as this figure:



K-MEANS ALGORITHM WORK

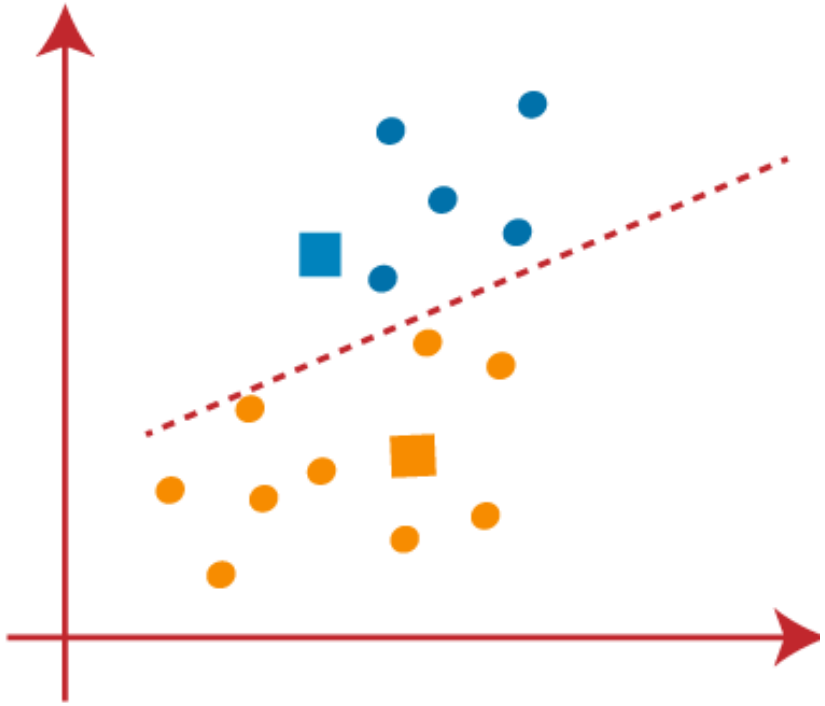
- Next, we will reassign each datapoint to the new centroid. For this, we will repeat the same process of finding a median line.

- The median will be like this image:



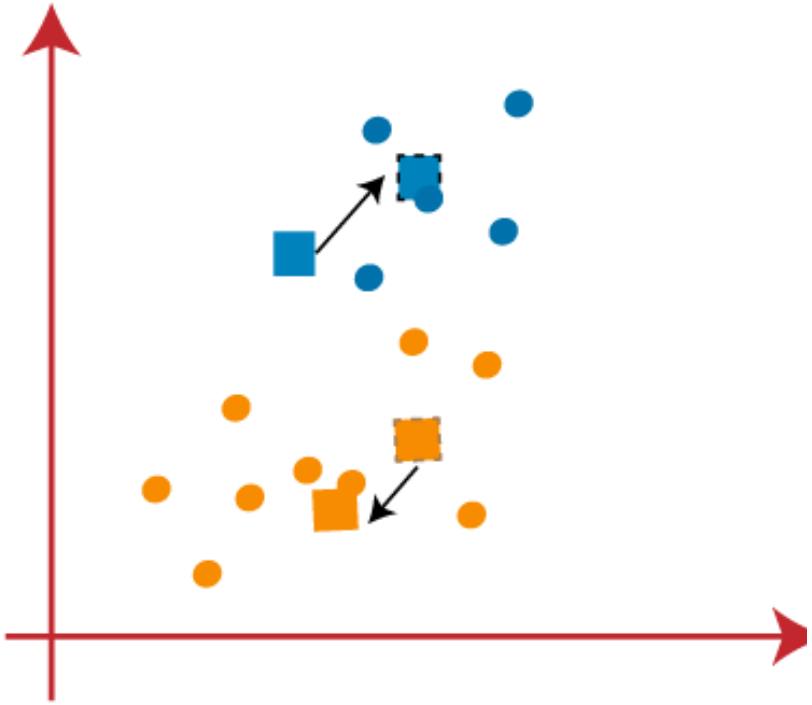
K-MEANS ALGORITHM WORK

- From the above image, we can see, one yellow point is on the left side of the line, and two blue points are right to the line.
- So, these three points will be assigned to new centroids.



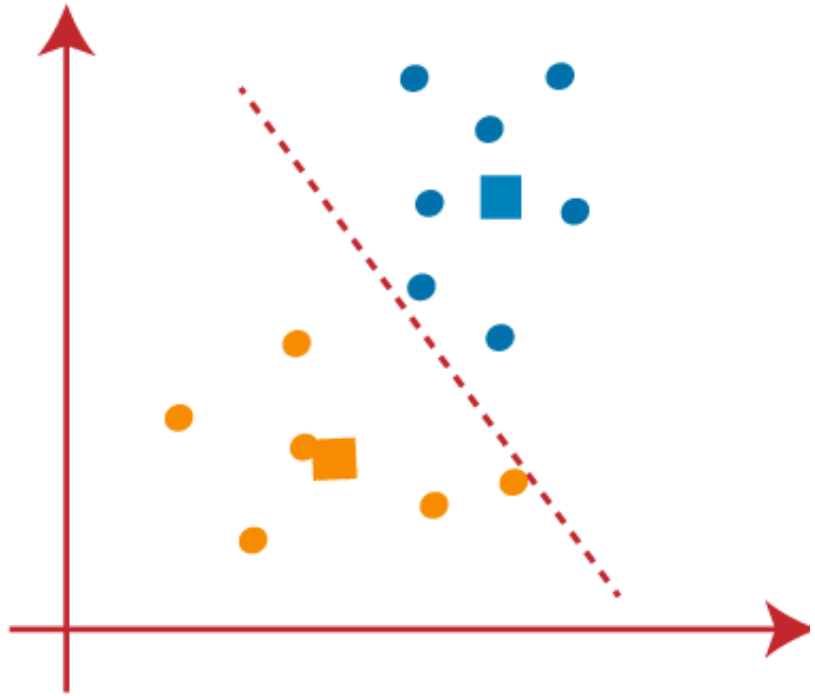
K-MEANS ALGORITHM WORK

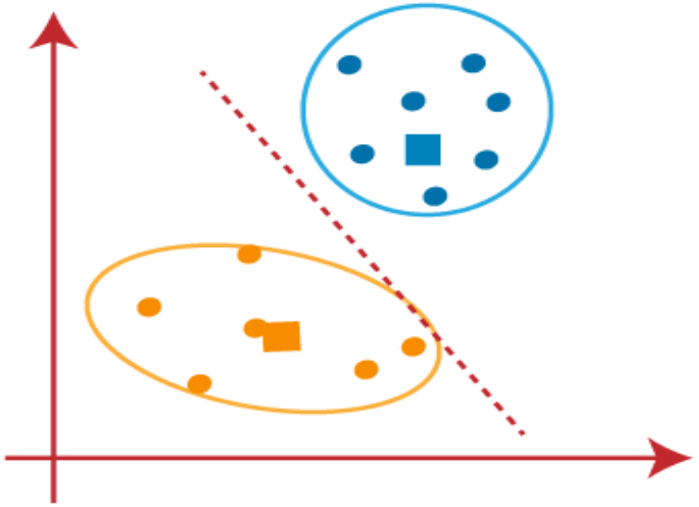
- As reassignment has taken place, so we will again go to the step-4, which is finding new centroids or K-points.
- We will repeat the process by finding the center of gravity of centroids, so the new centroids will be as shown in the this image:



K-MEANS ALGORITHM WORK

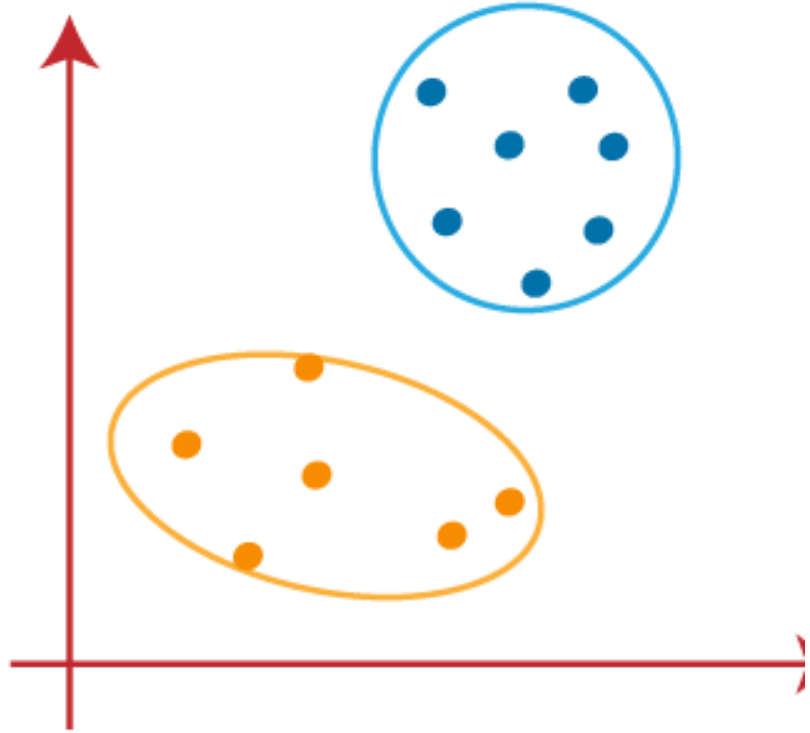
- As we got the new centroids so again will draw the median line and reassign the data points. So, the image will be:
- We can see in the above image; there are no dissimilar data points on either side of the line, which means our model is formed. Consider the below image:





K-MEANS ALGORITHM WORK

- As our model is ready, so we can now remove the assumed centroids, and the two final clusters will be as shown in the below image:



HOW TO CHOOSE THE VALUE OF K NUMBER OF CLUSTERS ?

- The performance of the K-means clustering algorithm depends upon highly efficient clusters that it forms.
- But choosing the optimal number of clusters is a big task.
- There are some different ways to find the optimal number of clusters.
- Here we are discussing the most appropriate method to find the number of clusters or value of K.
- The method is given below:

- Elbow Method

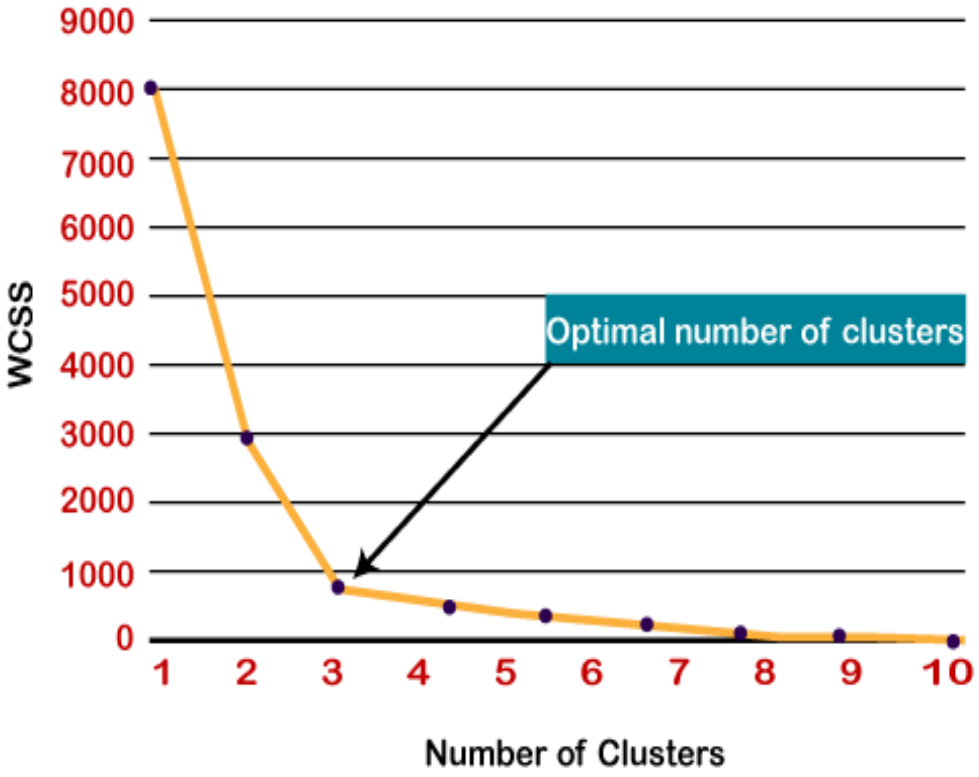
ELBOW METHOD

- Elbow Method
  - The Elbow method is one of the most popular ways to find the optimal number of clusters.
  - This method uses the concept of WCSS value.
  - **WCSS** stands for **Within Cluster Sum of Squares**, which defines the total variations within a cluster. The formula to calculate the value of WCSS (for 3 clusters) is given below:
- $WCSS = \sum_{Pi\ in\ Cluster1} distance(P_i\ C_1)^2 + \sum_{Pi\ in\ Cluster2} distance(P_i\ C_2)^2 + \sum_{Pi\ in\ CLuster3} distance(P_i\ C_3)^2$
  - In the above formula of WCSS,
  - $\sum_{Pi\ in\ Cluster1} distance(P_i\ C_1)^2$: It is the sum of the square of the distances between each data point and its centroid within a cluster1 and the same for the other two terms.

ELBOW METHOD

To measure the distance between data points and centroid, we can use any method such as Euclidean distance.

- 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 sharp point of bend or a point of the plot looks like an arm, then that point is considered as the best value of K.



IMPLEMENTATION OF K-MEANS CLUSTERING ALGORITHM

- So, we have a dataset of **Mall_Customers**, which is the data of customers who visit the mall and spend there.

- In the given dataset, we have **Customer_Id, Gender, Age, Annual Income ($), and Spending Score**
- **The steps to be followed for the implementation are given below:**
  - Data Pre-processing
  - Finding the optimal number of clusters using the elbow method
  - Training the K-means algorithm on the training dataset
  - Visualizing the clusters

THANK YOU