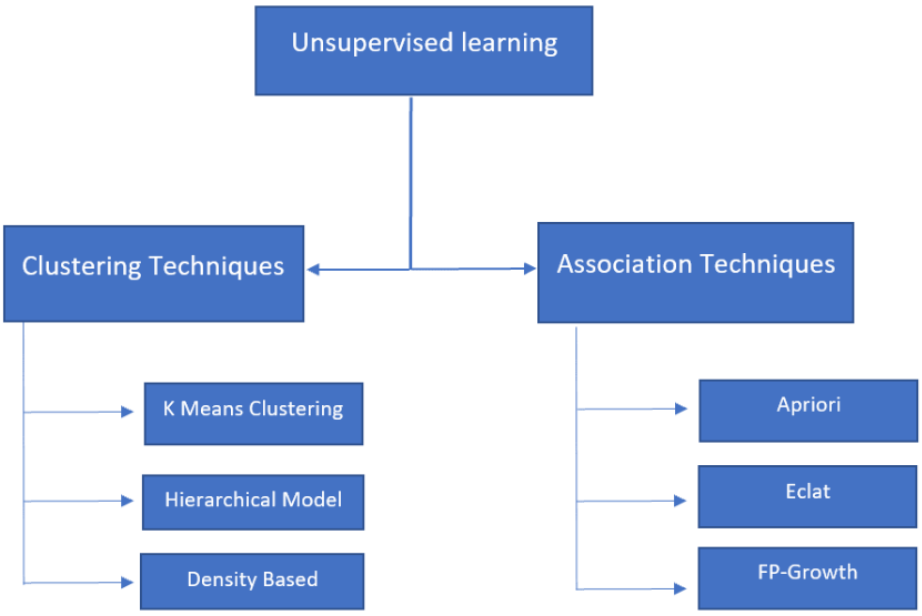
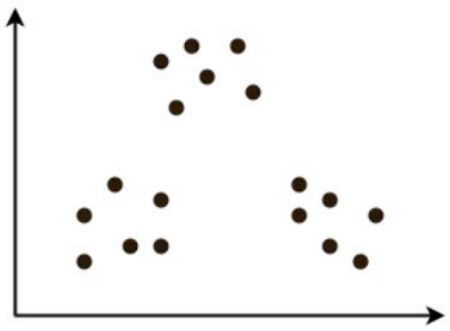
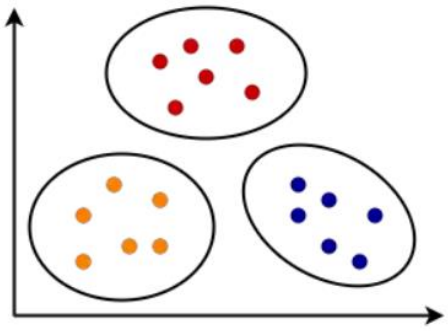


K-Means clustering

- K-Means clustering is an unsupervised iterative clustering technique.
- It partitions the given data set into k predefined distinct clusters.
- A cluster is defined as a collection of data points exhibiting certain similarities.



Before K-Means



After K-Means

How did it do that ?

STEP 1: Choose the number K of clusters

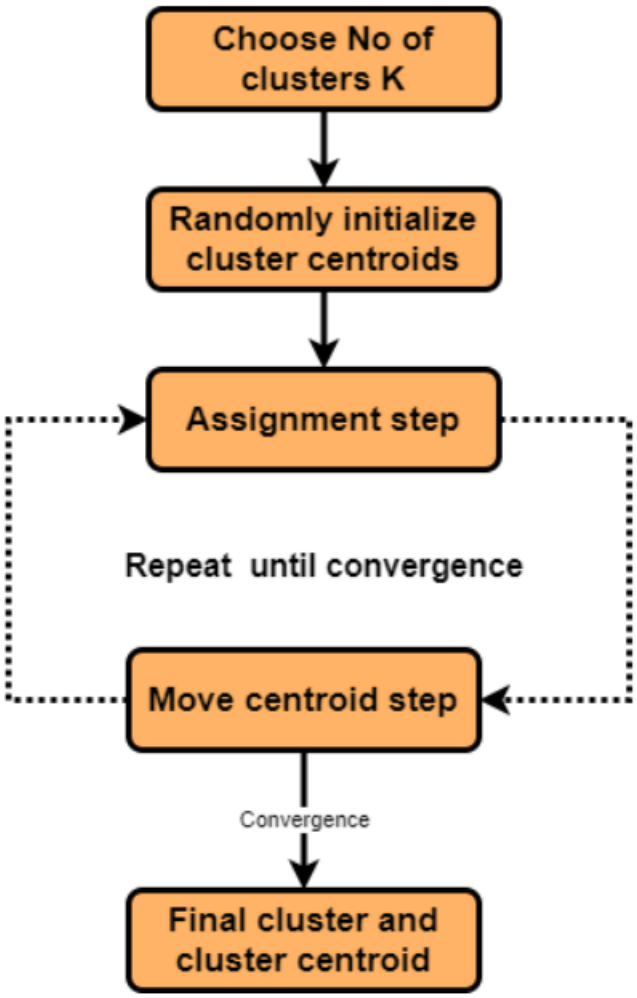
STEP 2: Select at random K points, the centroids (not necessarily from your dataset)

STEP 3: Assign each data point to the closest centroid → That forms K clusters

STEP 4: Compute and place the new centroid of each cluster

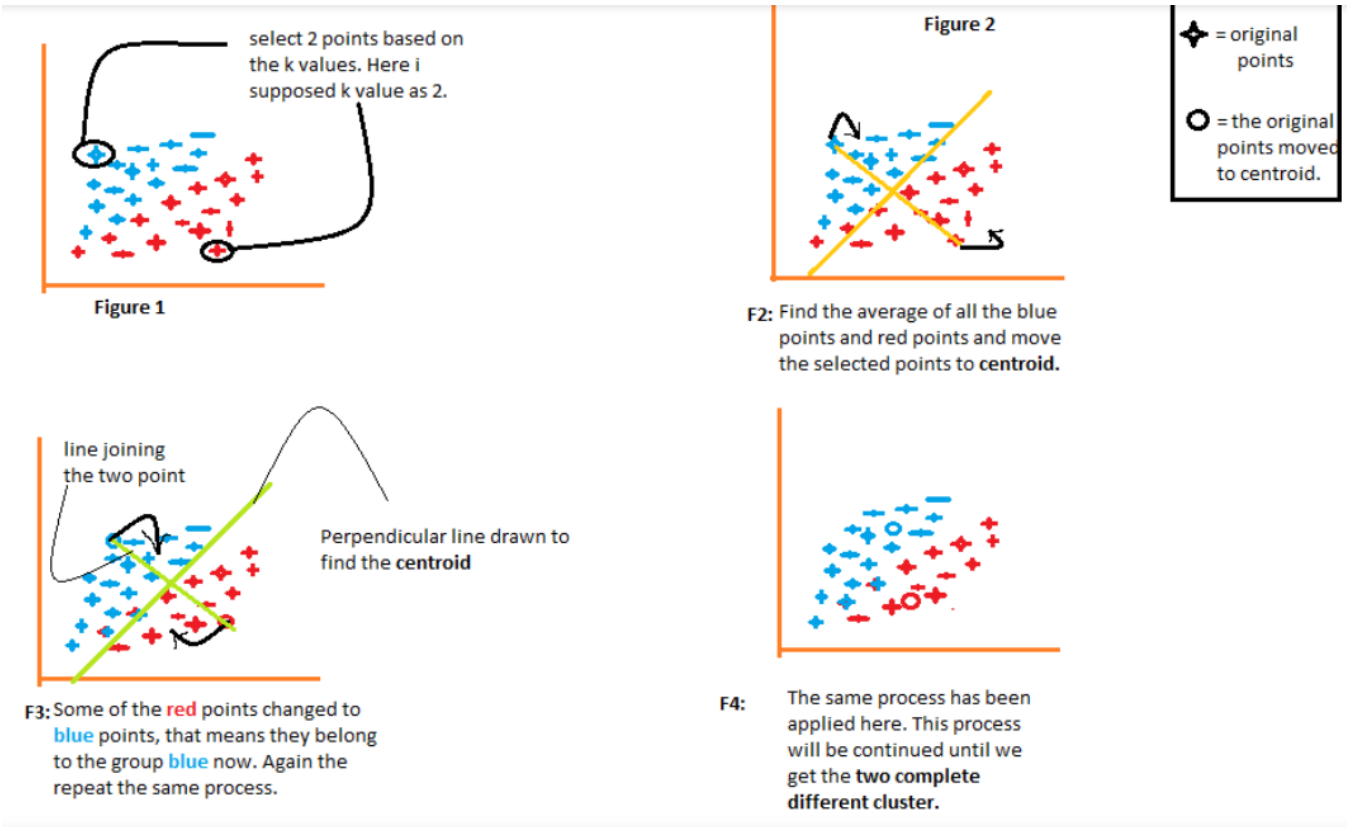
STEP 5: Reassign each data point to the new closest centroid.
If any reassignment took place, go to STEP 4, otherwise go to FIN.

Your Model is Ready



three steps.

1. Select the k values.
2. Initialize the centroids.
3. Select the group and find the average.



How to choose the value of K?

1. Elbow Method.
2. Silhouette Method.

Elbow Method for K-Means Clustering:

Purpose: Determine the optimal number of clusters (k) for K-Means algorithm.

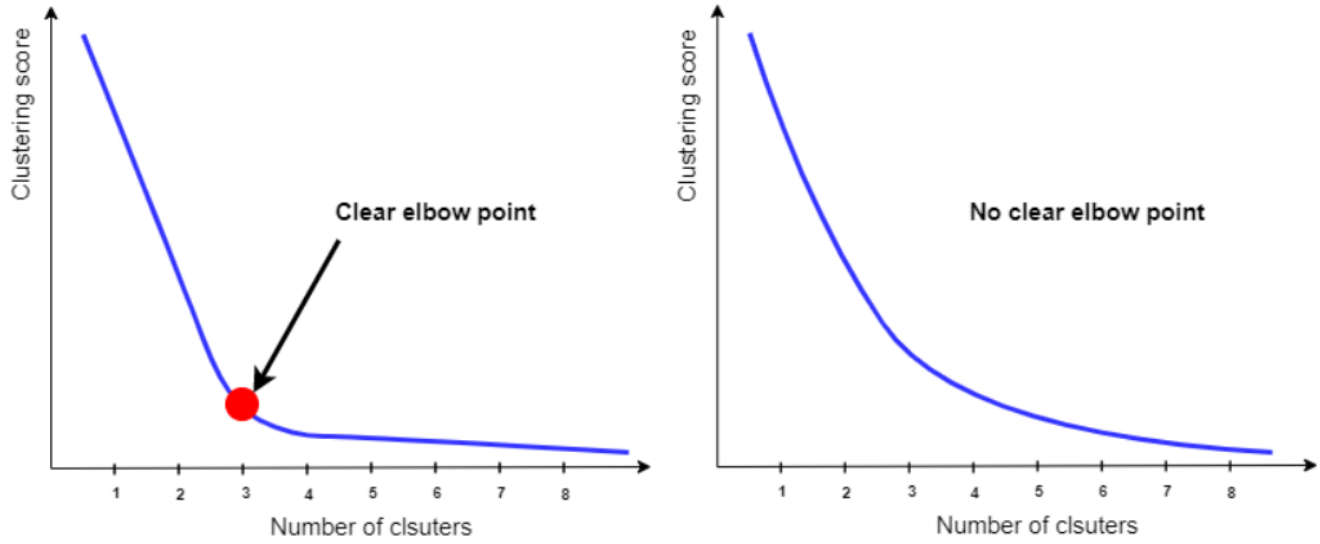
Method:

- Run K-Means multiple times with increasing k values (e.g., 1 to 10).

- ❖ Calculate the clustering score (e.g., within-cluster sum of squares) for each k.
- ❖ Plot clustering score vs. number of clusters.
- ❖ Look for the "elbow" point where the score decrease slows down significantly.

Challenge: Not always a clear elbow point exists.

Difficulties without a clear elbow: Choosing the optimal k becomes subjective and requires further analysis.



Use of K-Mean Clustering

- **Search engine**: Search engine, groups results together using clustering algorithm
- **Customer segmentation**: K-mean clustering can be used to create customer clusters based on demographic information, geographical information and behavioral data.
- **Social network analysis**: To find groups of people with specific interest to direct the personalized ads.
- **Data center**: To organize the computer clusters in data center.
- **Inventory management**: Create inventory clusters based on sales number and manufacturing capacity

Advantages of K-means

1. It is very simple to implement.
2. It is scalable to a huge data set and also faster to large datasets.
3. it adapts the new examples very frequently.
4. Generalization of clusters for different shapes and sizes.

Disadvantages of K-means

1. It is sensitive to the outliers.
2. Choosing the k values manually is a tough job.
3. As the number of dimensions increases its scalability decreases.