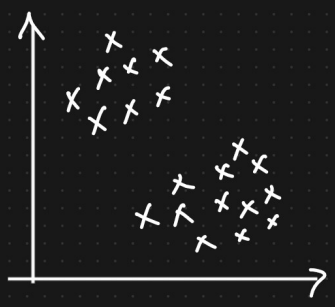
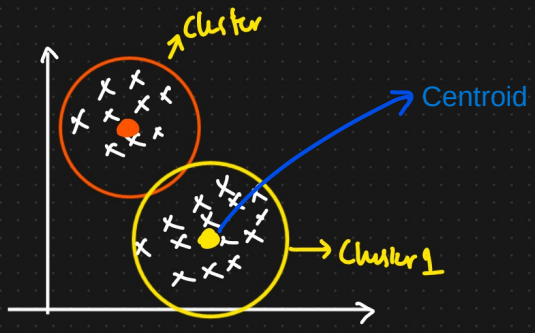
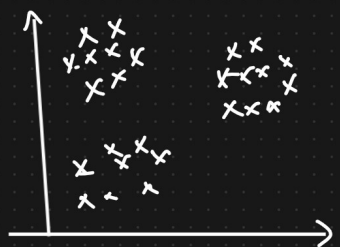
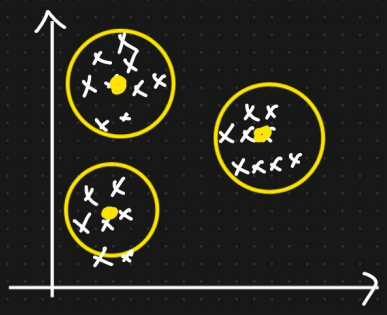
K Means Clustering Algorithm



K Means



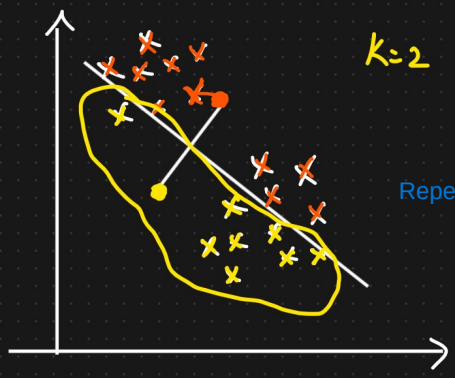




K=2

K Mean

Euclidian Distance
OR
MANHATTAN Distance

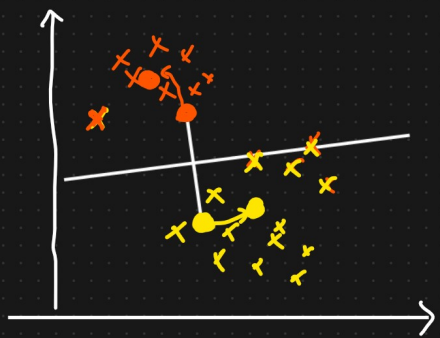


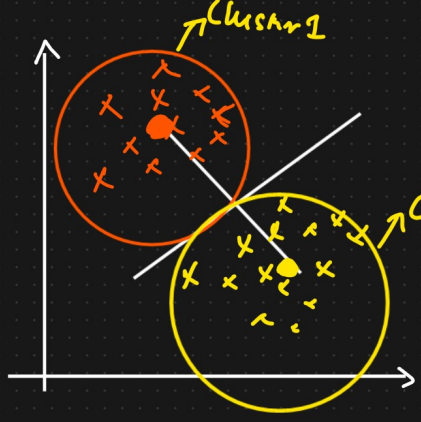
① Initialize Some K → centroids

② Points that are nearest to the centroids → Group

Repeat

③ Move the centroids → Average



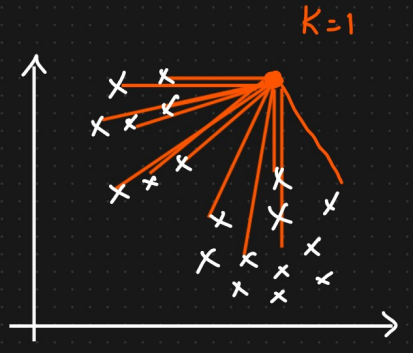


Here, I got two different cluster clearly so stop.

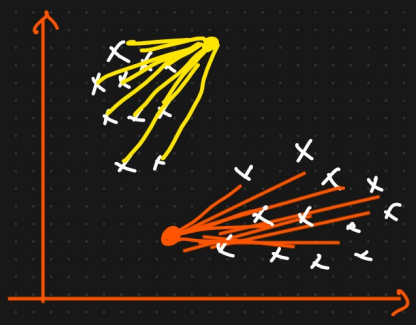
★ How do we select the K value?

WCSS = Within Cluster Sum of Squares

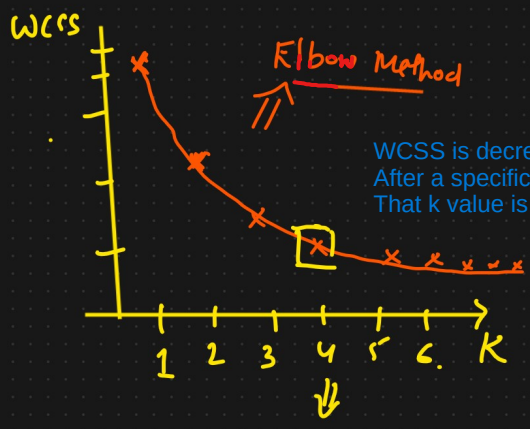
Initialize K=1 to 20



$$WCSS = \sum_{i=1}^{n} (\text{distance between points to nearest centroid})^2$$



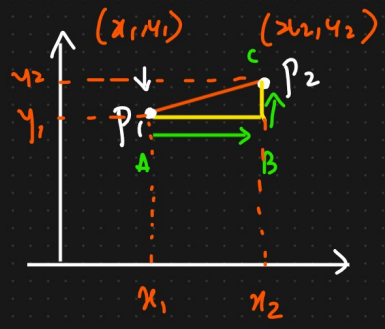
WCSS ↓



WCSS is decreasing.
After a specific point it is stabled.
That k value is best.

ⓕ Euclidian Distance

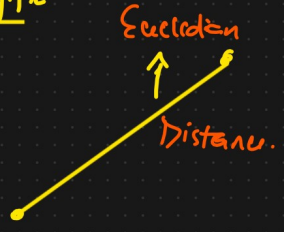
$$\text{Euclidian dist} = \sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$$



$$\text{Manhattan dist} = |x_2-x_1| + |y_2-y_1|$$

IRON MAN → U.S

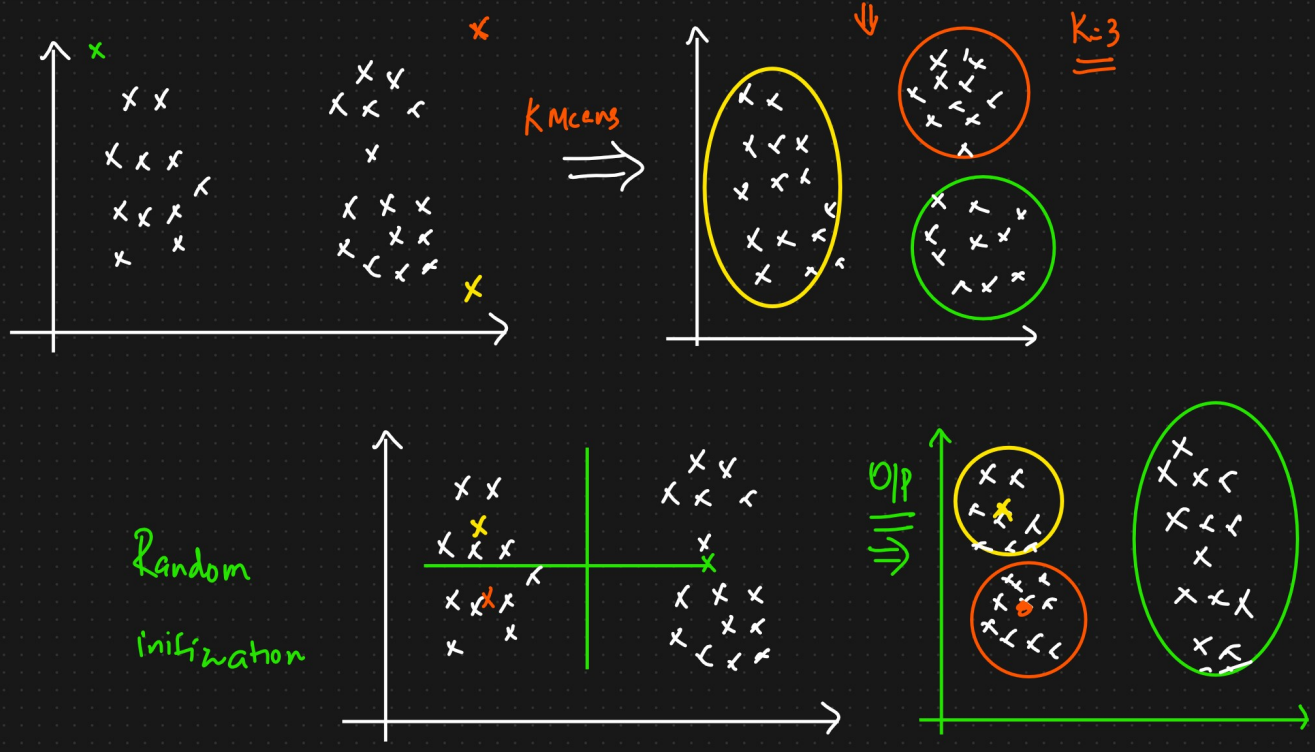
Air Traffic



⇒ Manhattan distance

When K-means starts, it randomly selects K initial centroids. If these centroids are not well chosen (e.g., too close to each other or far from actual cluster centers),
it can Cause inconsistent results across different runs

Random Initialization TRAP (Kmeans ++)



To avoid it :-

Use Kmeans ++ Initialization Technique

- Now in this technique, we initialize all the centroids in such a way that at least it should be at max distance that it can.
- Centroid are initialized far away to each other.