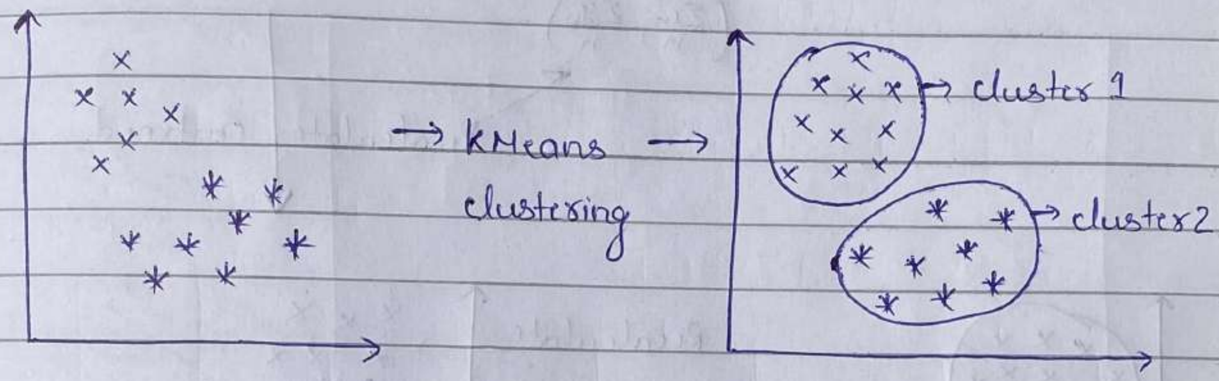
K Means Clustering

→ It is an unsupervised machine learning algorithm which is used to solve clustering problem by grouping the unlabeled dataset into different clusters.



* Mathematical Intuition :

Steps :

(i) Initialize some k value.

(ii) Randomly initialize k centroids

(iii) Assign data points to nearest centroid

(iv) Recalculate the centroid value by mean of datapoints.

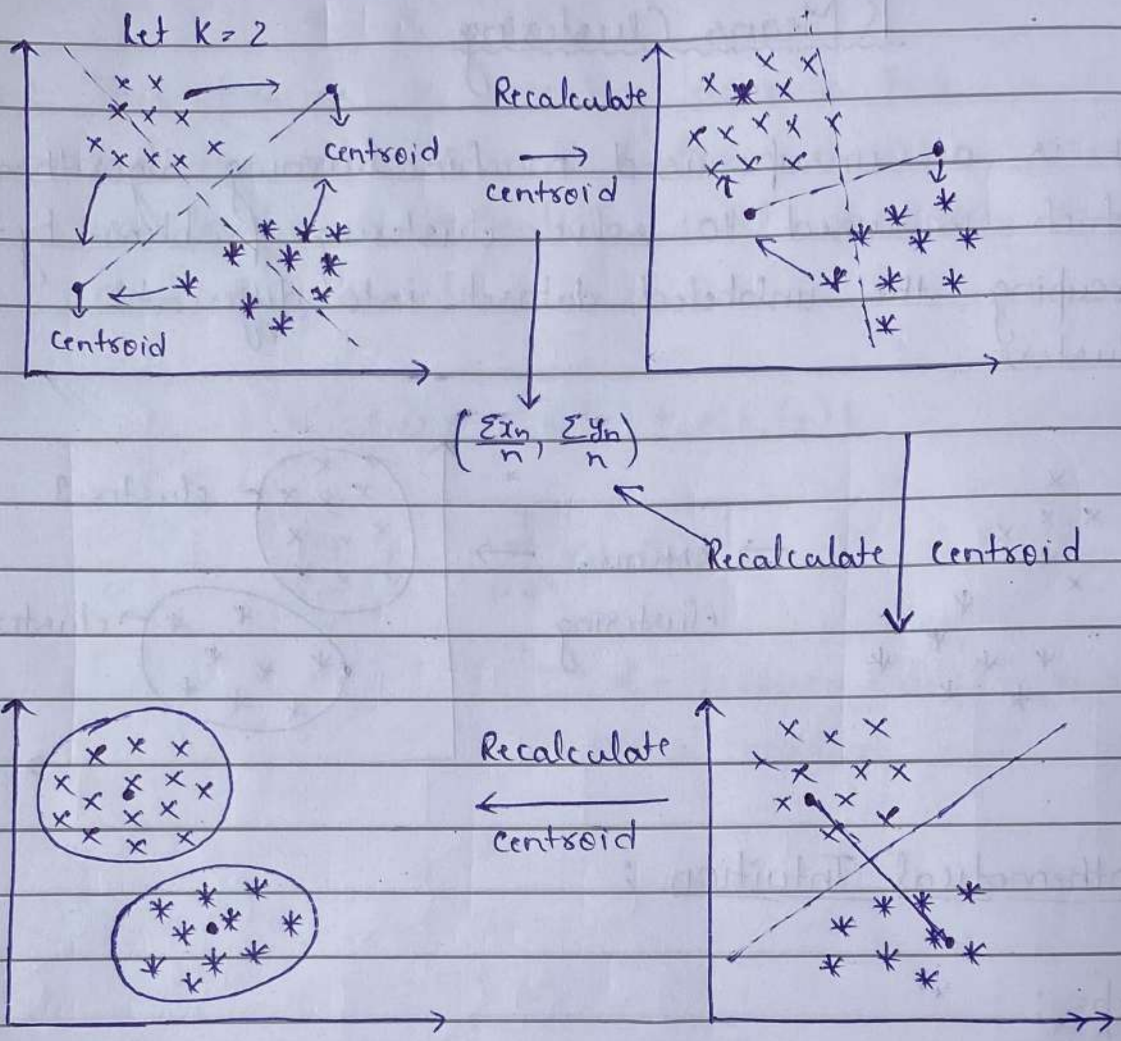
(v) Repeat 3rd and 4th step until centroid of current iteration become same ~~of~~ its previous iteration.

- centroid calculation :

$$C = \left(\frac{\Sigma x_n}{n} , \frac{\Sigma y_n}{n} \right)$$

- Distance calculation : → Eulidean or Manhattan

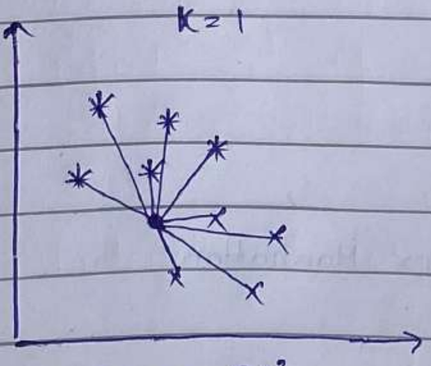
$$ED = \sqrt{(x_2-x_1)^2 + (y_2-y_1)^2}$$



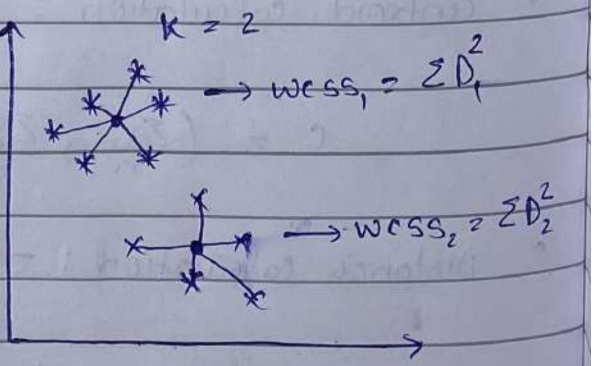
Q How we can select the k value ?

- WCSS → within cluster sum of square

$$\boxed{WCSS = \sum_{j=1}^{K} \left[\text{Distance between points to nearest centroid} \right]^2}$$

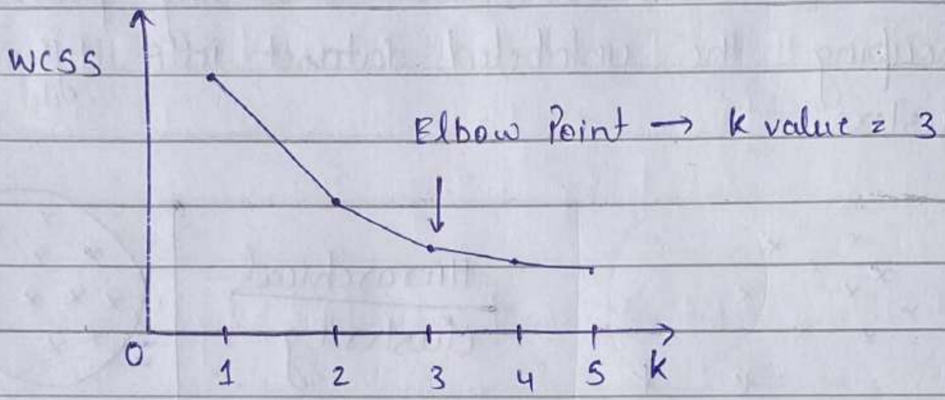


$WCSS = \Sigma D^2$

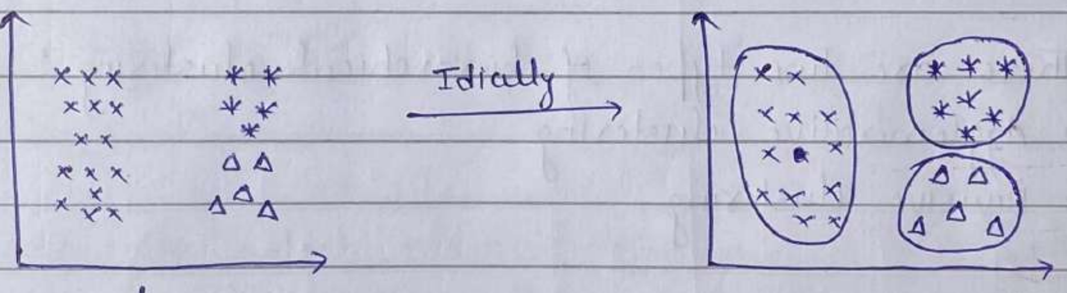


$WCSS = WCSS_1 + WCSS_2$

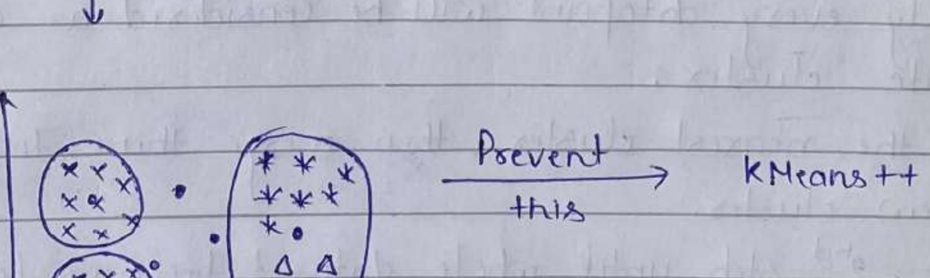
Note: As the value of k increases, the value of wcss decreases.



* Random Initialization Trap :



But sometimes leads to this when centroid initializes very close to each other



Note: KMeans++ is an initialization technique which ensures that the centroids should initialize far from each other.