* PCA.

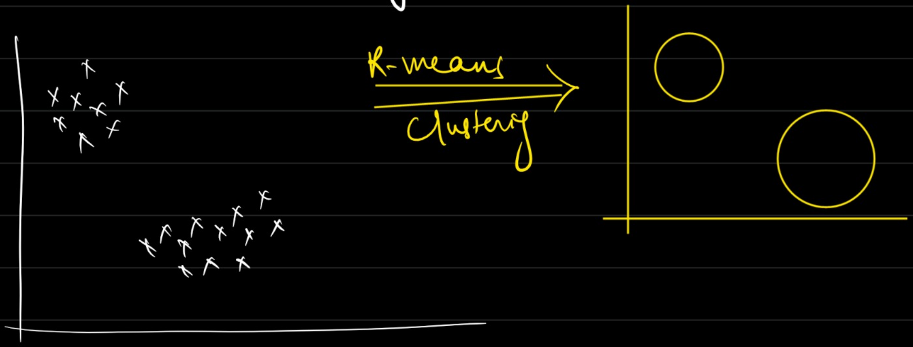
- * USL → Kmeans. / Hierarchical / DBSCAN
- * Anomoly detection → Isolation forest / DBSCAN / Local outlier fac
- * Time series. → Smoothing models / Autoregressive models.

USL

↳ y (target variabe is not there.

↳ finds groups | pattern in the data.

① K- means clustering



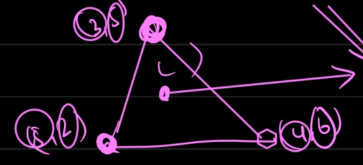
K- means

↓ ↓

K means (mean/Average)

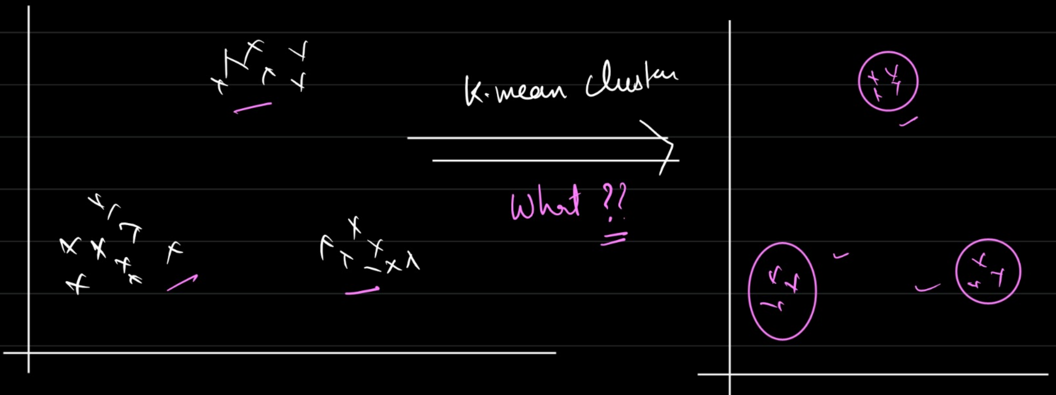
$$\begin{array}{ccc} f_1 & f_2 & f_3 \\ (5 & 3 & 2) \\ (2 & 4 & 3) \\ (6 & 1 & 5) \\ \hline 13/3 & 8/3 & \frac{10}{3} \end{array}$$

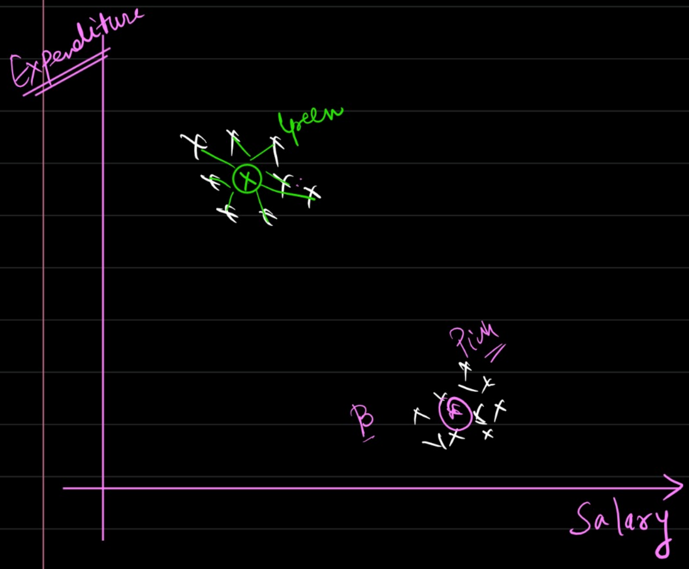
Average → Arithmetic center of the data. ⇒ Centroid.



$x_{cent} = \frac{5+3+4}{3}$

$y_{cent} = \frac{2+5+6}{3}$



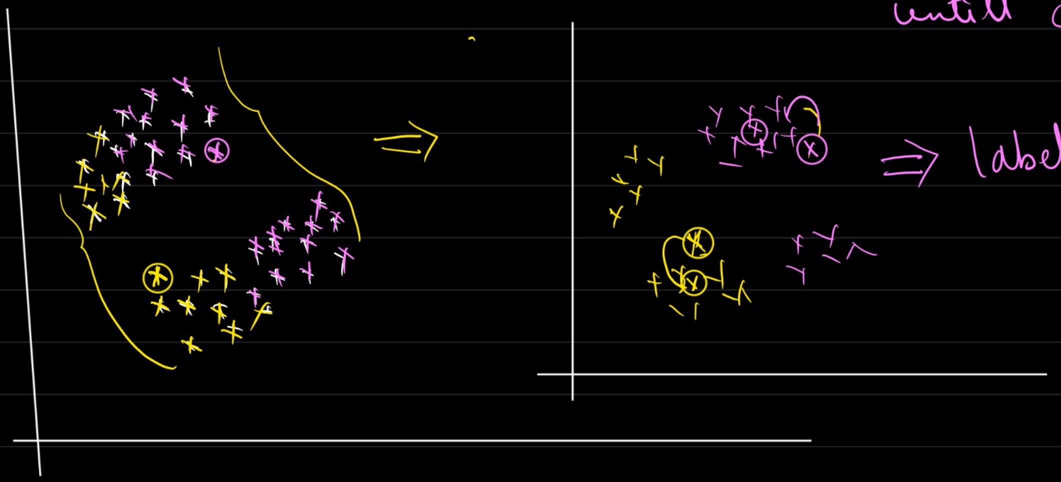


Steps of K-means clustering

(i) initialise centroid (K)
↳ Can be a random dp
↳ Centreid of all dPs
↳ A random new dP's

(ii) Poins nearer to the centroid will be labelled as that group.

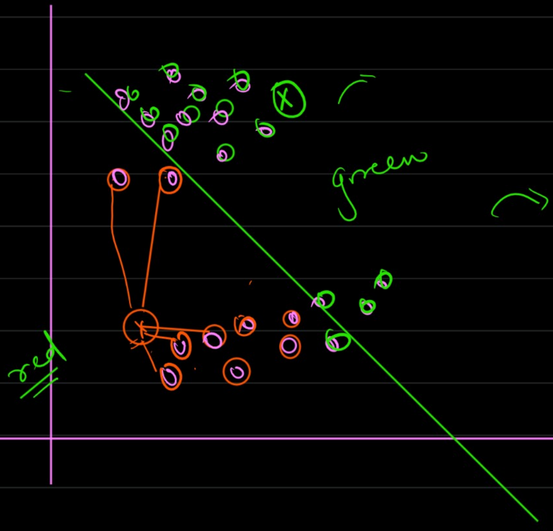
(iii) Recalculate the centroid
↳ step 2 will be repeated untill centroid doesnt change.

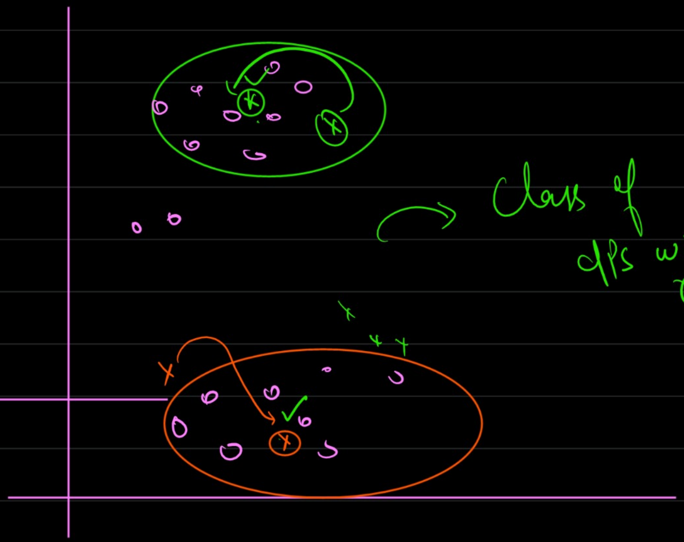


⇒ label will change

⇓

New centroid.





→ Class of d/p's will change

distance

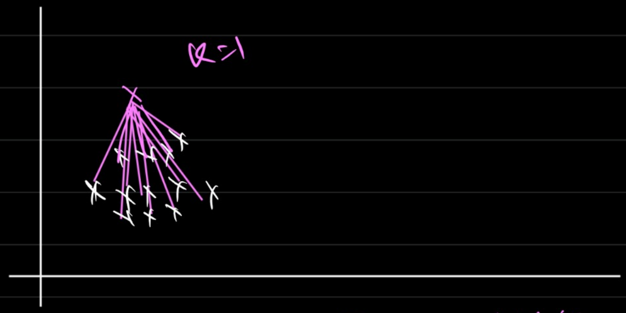
① Euclidean $\sqrt{(x_2-x_1)^2 + (y_2-y_1)^2}$

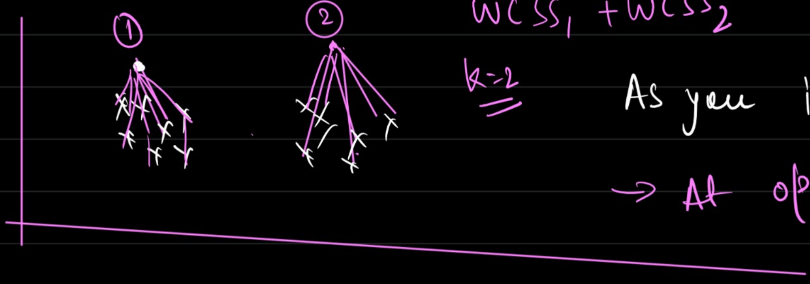
② Manhattan distance = $|x_2-x_1| + |y_2-y_1|$

* How to decide K.

WCSS → within Cluster Sum of Square distance

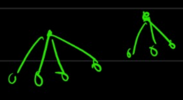
$$WCSS = \sum_{i=1}^{n} \left(\text{distance b/w points to nearest centroid} \right)^2$$

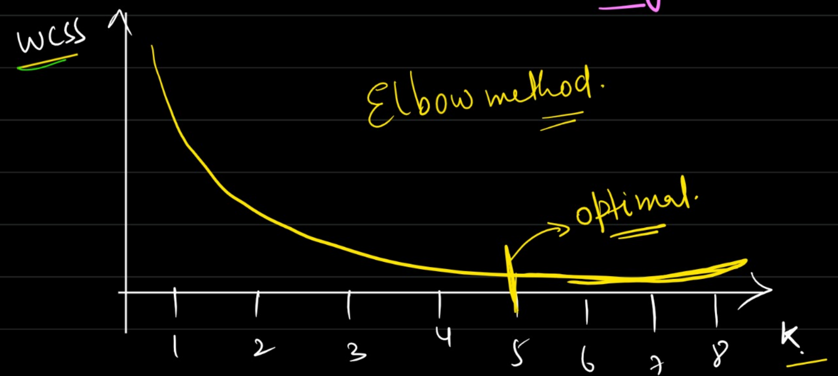




As you increase K, WCSS decreases.

→ At optimal K, WCSS will not change.





if all the dP are classified to the nearest Centroid class. ⟸ { if class label not Changing ⇓ Centroid not change }

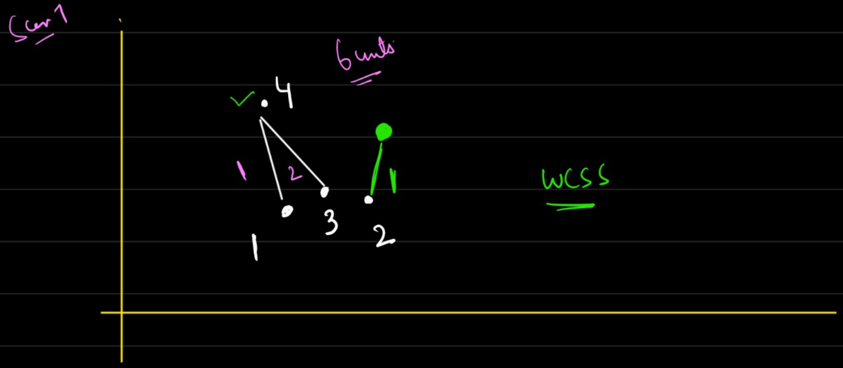
* if WCSS not Changing.

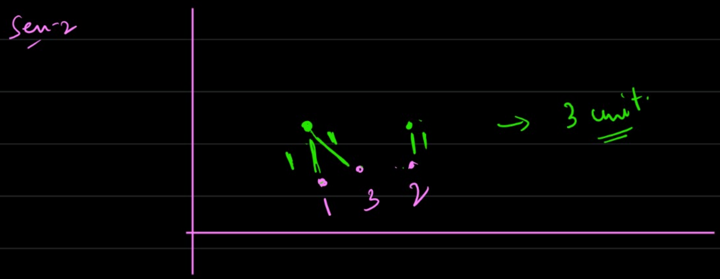
⇓

All the dp's are correctly marked to the nearest cluster class.

⇓

optimal k.



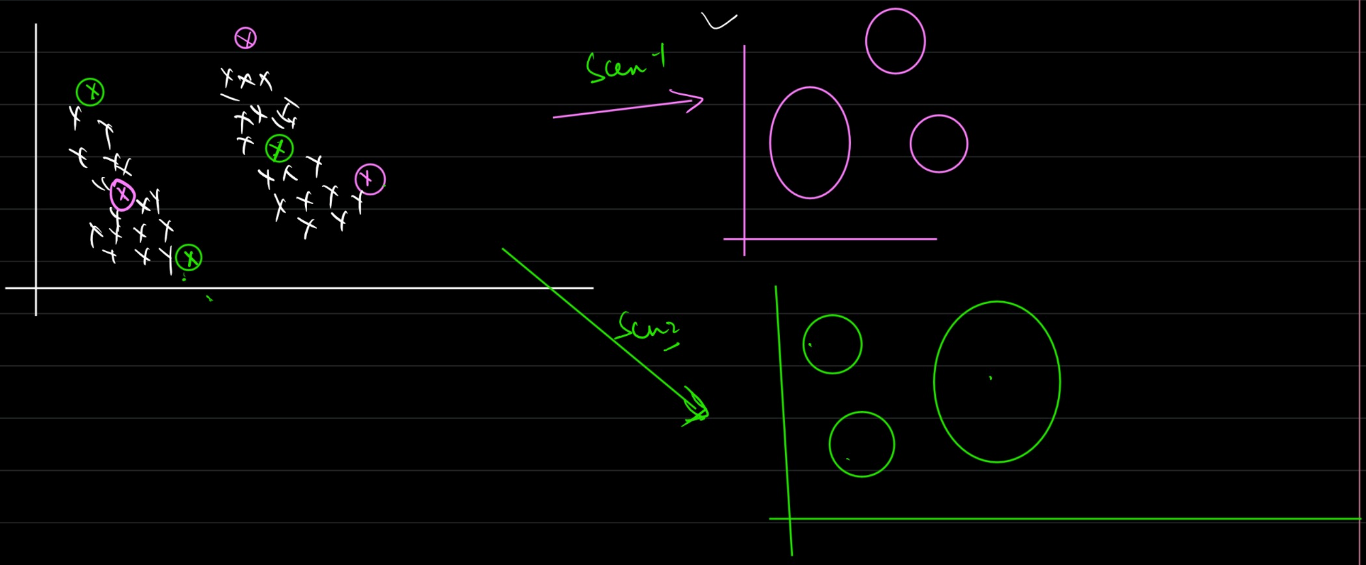


How k is decided :-

✓→ elbow method.

→ business team.

* Random initialization k-trap



* No of clusters depends on the Initialization of Centroid

* initialize the K's as much far as possible

⇓

Kmeans ++