K-means Clustering:

It is an unsupervised ML Algorithm.
It groups / clusters the data into 'k' clusters.

Steps:

1) Determine 'K' clusters.
2) scatter the 'k' centroids randomly.
   Why K?
   Every cluster has a centroid.
3) Assign the data points to the closet centroid
   This can be done by
   1) Euclidean
   2) Manhattan
4) move the centroid
   (calculate the mean of data points within the cluster)
5) Repeat 3 & 4 until:
   - centroids do not change
   - max. iteration is reached.

join the centroids & draw perpendiculars.

How to find the 'k' value?

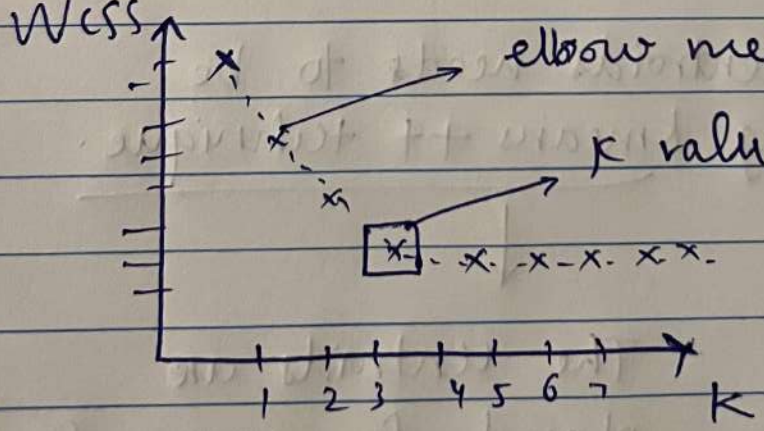
'k' means the number of clusters.

- Initialize k = 1 to some value.
  Since k = 1, there will only be 1 cluster and ∴ only 1 centroid.

  Compute the distance of every data point to the centroid.

WCSS (within cluster sum of squares)

$$\sum_{i=1}^{n} (\text{distance between data pt \& centroid})^2$$



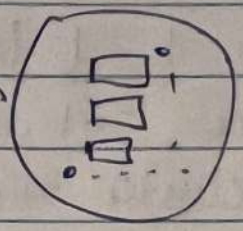
The K value ↑, WCSS ↓
We have to find a threshold or an arbitrary point from which the WCSS it decreases slowly.

- Euclidian distance:

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

- Manhattan distance:

$$|x_2-x_1|+|y_2-y_1|.$$



~~The~~ If the centroids are initialized randomly, then there are chances that we may get a different/ output. wrong.

If the centroids are close together, it result into random initialization trap.

Therefore, the centroids needs to be initialized using kmeans ++ technique.

↓

- The centroids are placed far away from each other.