K-Means Clustering.

Supervised Machine learning
- Linear Regression → Regression (Continous)
- Logistic Regression → Classificat^n (Categories)

H.W
→ Why Logistic Regression has regression in its name, but it actualy helps us to perform Classificat^n Task.

UnSupervised Machine Learning (e.g. Algorithm (K-Means))
- → DBScan, Heiracical Clustering.

K-Means Clustering

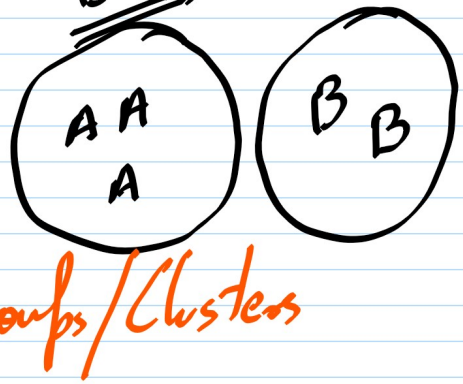
→ It is an unsupervised ML model used for cluster or group the data together.

→

e.g.

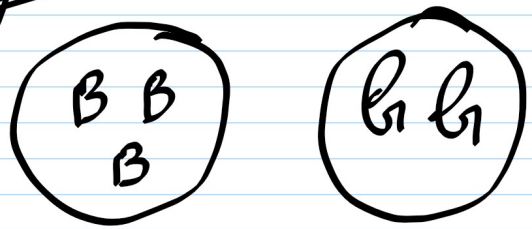
Age	Blood Group	Gender	Eye-color	Label
—	A	M	B	↓
—	B	M	B	Absent
—	A	F	Gr	
—	A	F	Gr	
—	B	M	B	

Based → BG



Groups/Clusters

Eye-Color

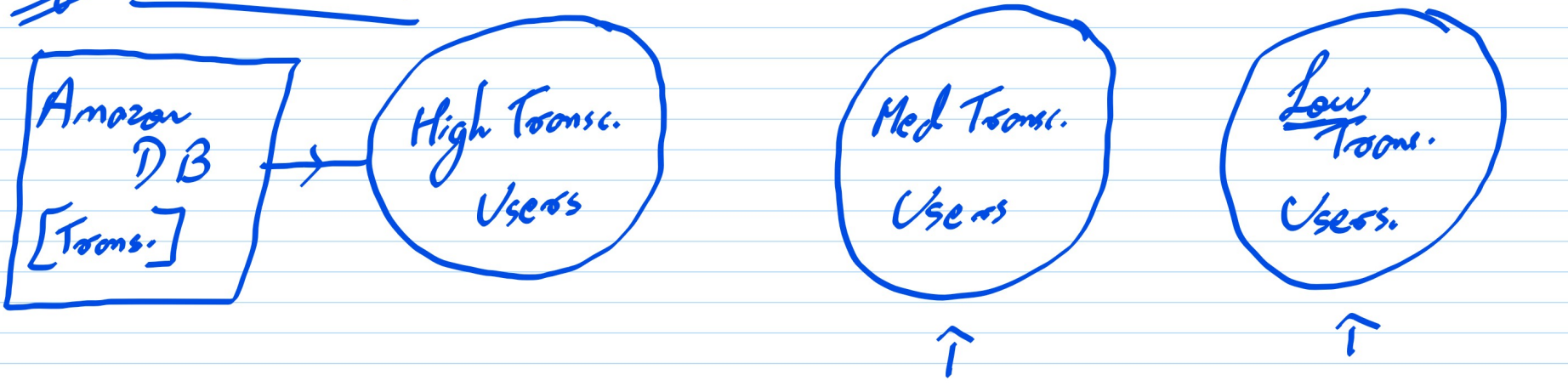


→ K-Means groups the datapoint together on the basis of certain common attributes.

Real World Application :=>

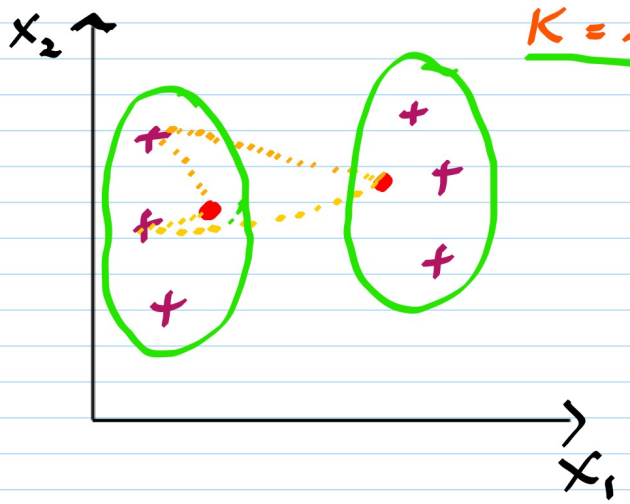
=> Customer Profilling => Grouping similar kind of Customers.

e.g. E-commerce.



If targets particular group of Customers.

Q) How K-Means Clustering Work?



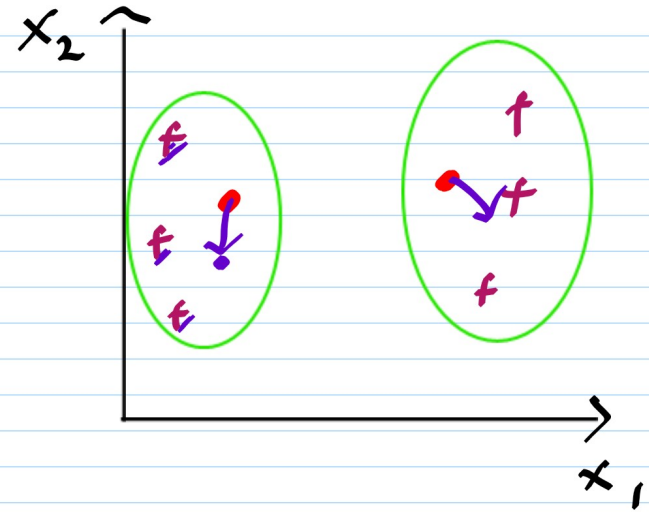
K-means Clustering
↳ Number of desired clusters.

→ Centroid (It will take two center points).

① Randomly chooses the centroid and assigning each data point to its nearest centroid to form a cluster.

Distance is calculated based on Euclidean Distance.

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



(2) Take the average value of the clusters and move the centroid to the average value.

(3) Again calculate the distance of all the data points from the centroid and assigns those point to nearest centroid.

→ Perform the steps again and again until the movement of the centroids stops.

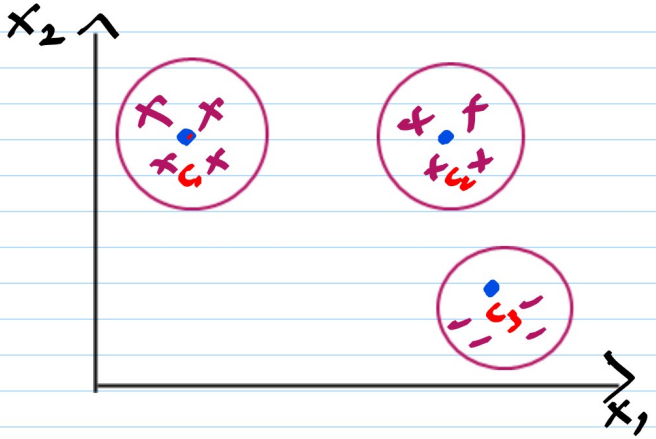
Q) What is the optimum number of clusters should we take?

K = 2, 3, 4, 5, 6, 7, ... n [1, 10]

↓

Clusters

WCSS / Inertia (Within Cluster Sum of Square).

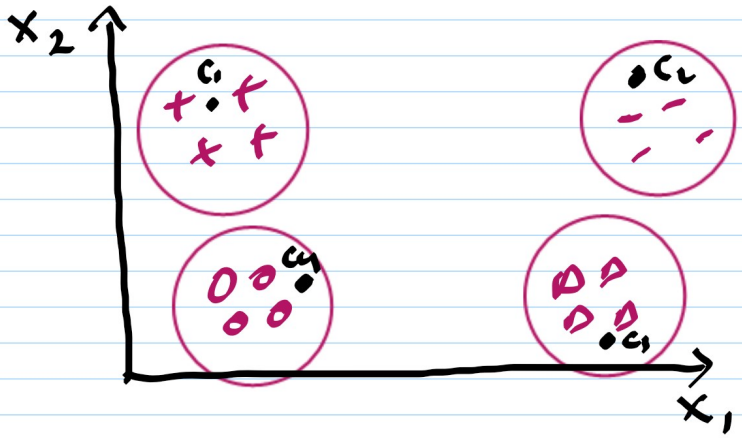


c = Centroid

x_i = Individual data-point.

$$WCSS = \sum\sum (x_i - c_i)^2$$

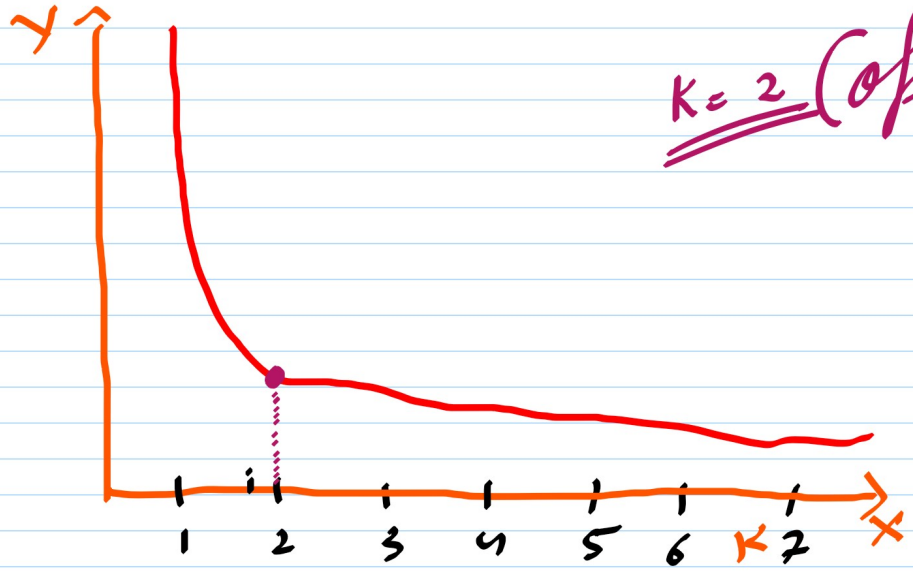
$$WCSS = \sum (x_i - c_1)^2 + \sum (x_i - c_2)^2 + \sum (x_i - c_3)^2$$



$$\Rightarrow \underline{WCSS} = \sum (x_i - C_1)^2 + \sum (x_i - C_2)^2 + \sum (x_i - C_3)^2 + \sum (x_i - C_4)^2$$

Increase the number of Clusters, WCSS ↓

Elbow Method



K = 2 (optimal k-value)