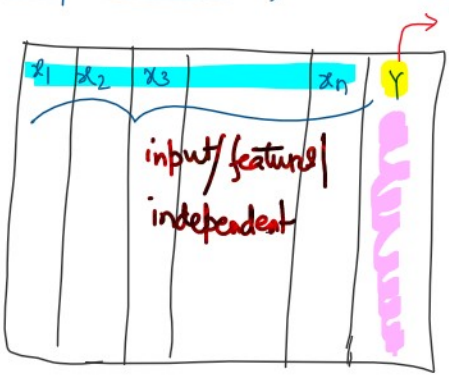
K-Means Clustering

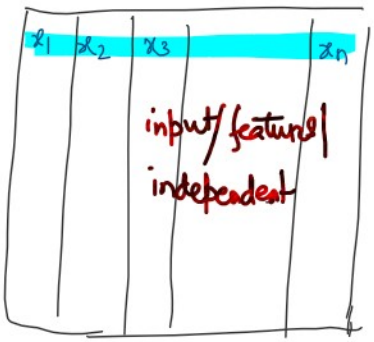
07 April 2024 19:59

Supervised learning



→ Target/Response/Dependent.

Unsupervised learning



→ No target
↓
No labels
↓
Unlabelled data.

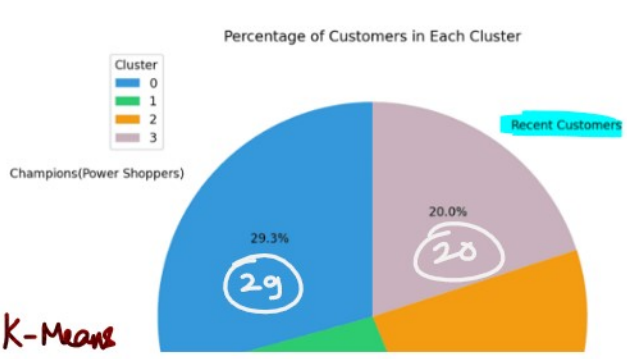
In unsupervised learning, there is no output variable to guide the learning process, and data is explored by algorithms to find the patterns.

relationships between x's (features only)

Given that the data has no labels, the algorithm identifies similarities on data points and groups them into clusters/segments.

Customer Segmentation

K-Means

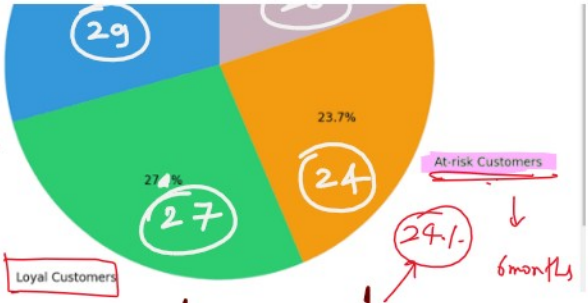


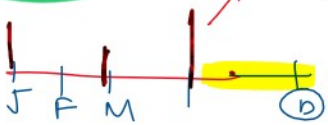
K-Means clustering

100

Variables
Name
1. Age
2. Gender
3. Buying pattern
4. Products range
5. Value
6. Mobile
7. location
and so on so forth

[source: credit history]

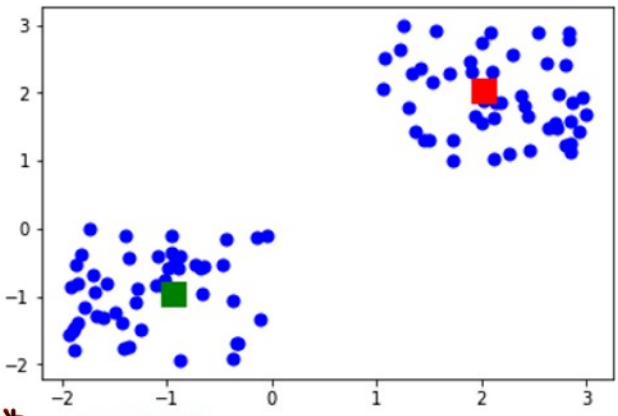




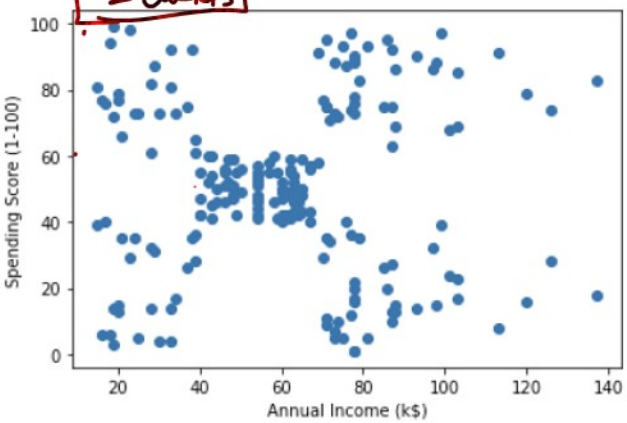
— they call you and take feedback

Royal Enfield → service → [stop using their service]

* K-Means Clustering



* 2 clusters



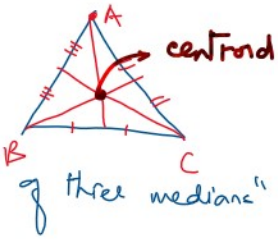
Is it easy ??

No

In K-Means clustering technique, each cluster is represented by its center (called a centroid), which corresponds to arithmetic mean of the

Q: What is centroid ?

Geometry

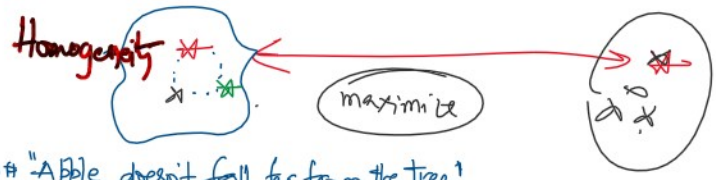


"Intersection of three medians"

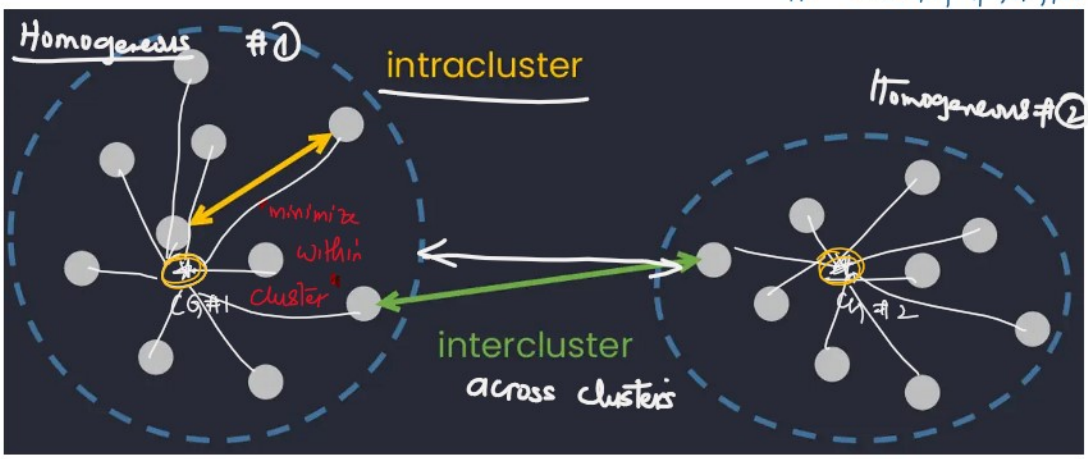
(called a centroid), which corresponds to arithmetic mean of the data points assigned to the cluster.

"Intersection of three medians"

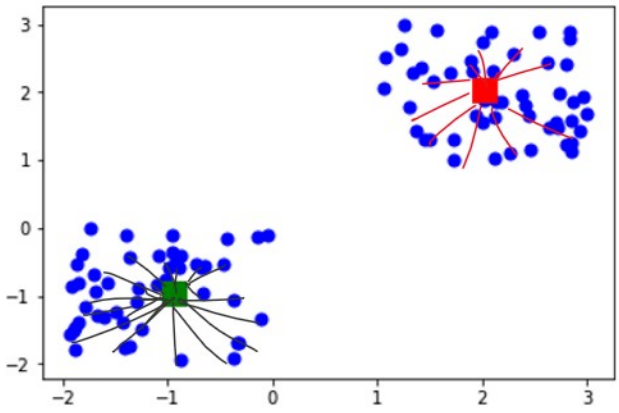
A centroid is a single data point within the cluster which represents the center and might not be necessarily be a member of the dataset



Idiom # "Apple doesn't fall far from the tree"



K-Means clustering algorithm tries to minimize the distance of the points in a cluster with their respective centroids.



Centroid #1
centroid # 2

K-Means Algorithm steps

1. Specify the no. of clusters 'K' — User input

Modeller — Business

"Scree plot" aka

Modeller ↓ Business ↓

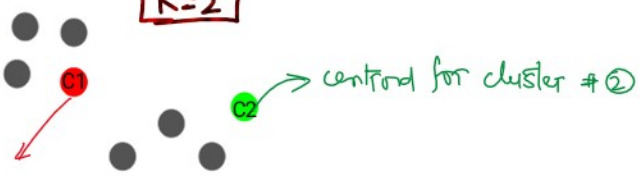
K=3 → ML based approach → "Scree Plot" aka (Elbow Chart) → "optimal no. of clusters"

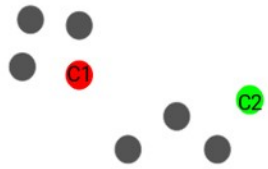
2. Initialize centroids by first shuffling the dataset and then randomly select 'K' data points for the centroids without replacement.

100 000 rows (4)

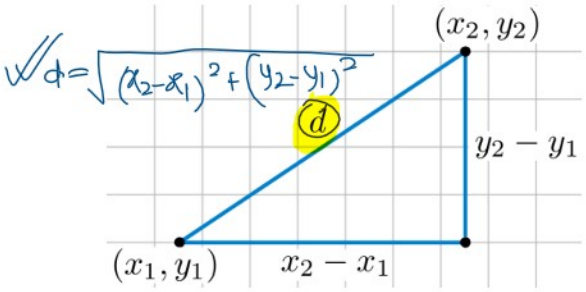
Total #8

K=2

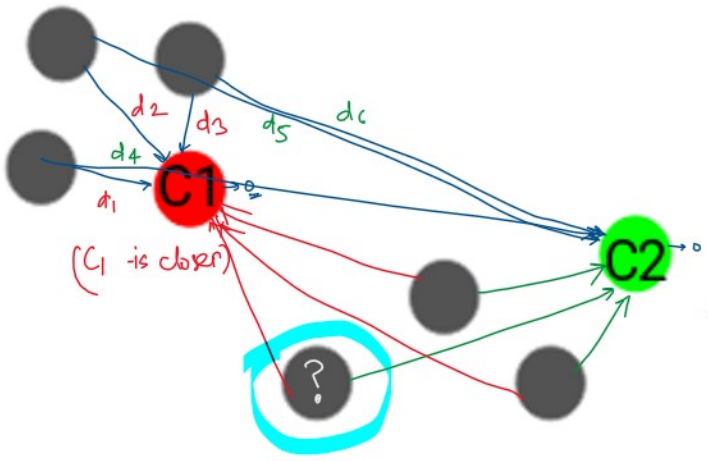


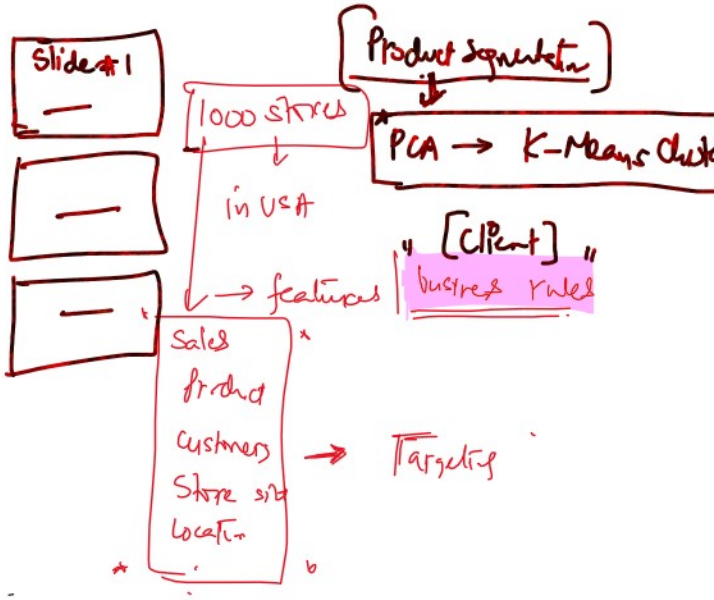


Euclidean Distance:

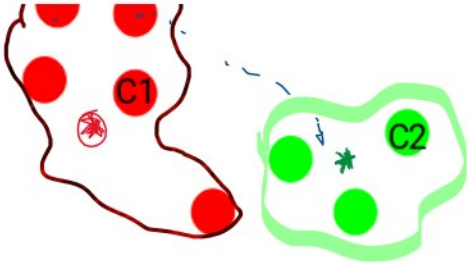


$d = \sqrt{(x_2-x_1)^2 + (y_2-y_1)^2}$ (d)



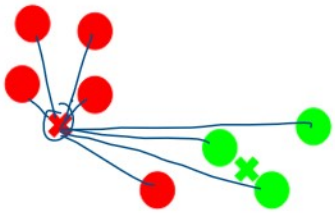


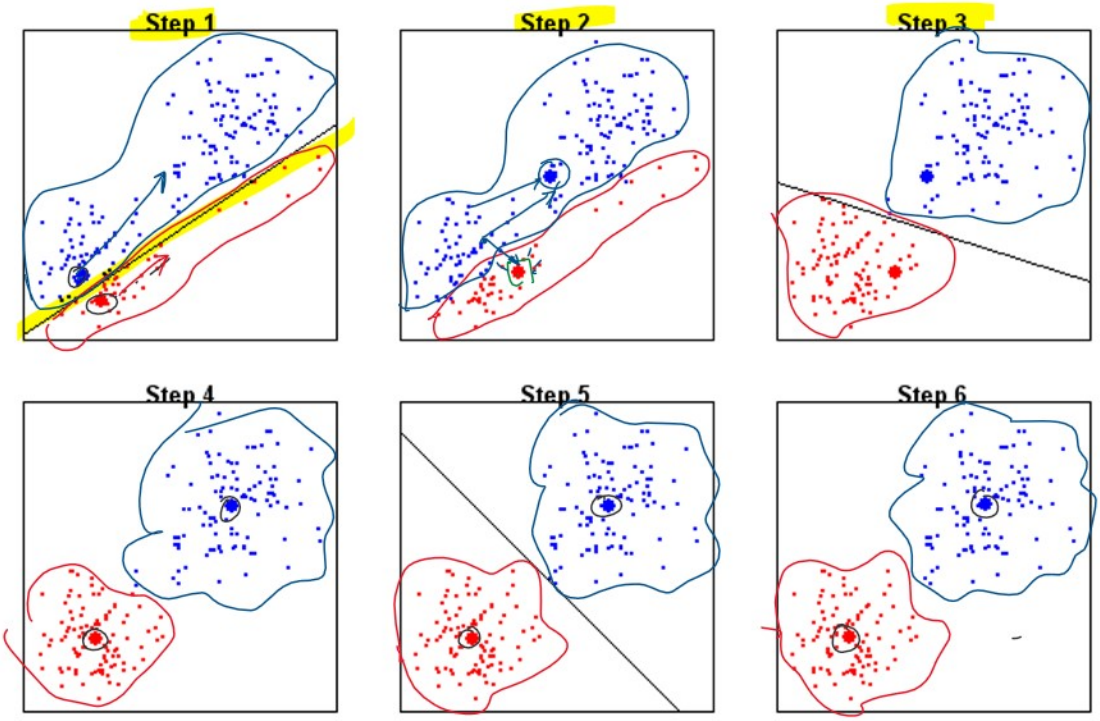
3. Assign all the points to the closest cluster centroid.



10:55

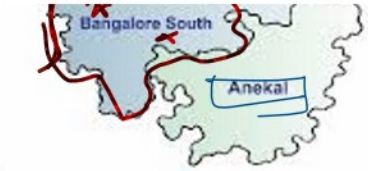
4. Recompute the centroids of newly formed clusters.





→ 2 clusters

Zara steps

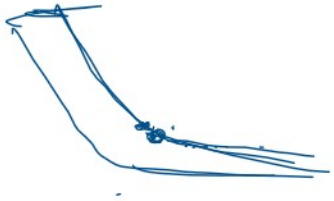


.2 → K

4 ≤ K ≤ 6 → Business + Elbow chart

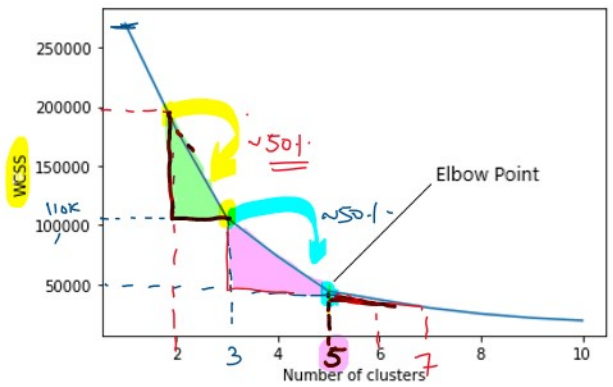
3 .

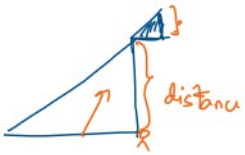
#Elbow chart



* Elbow method is a heuristic method to determine the optimal number of clusters (K) in K-Means clustering

* It involves plotting the within-cluster sum of squared distances (WCSS) again the no. of clusters and identifying the "elbow" point where the rate of decrease in WCSS slows down.



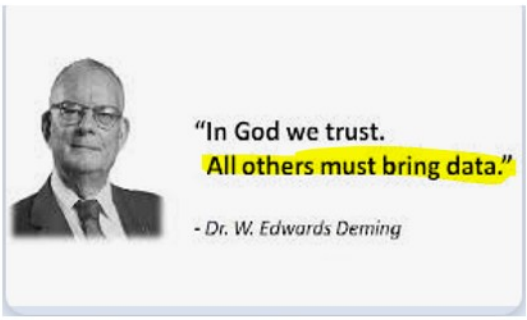


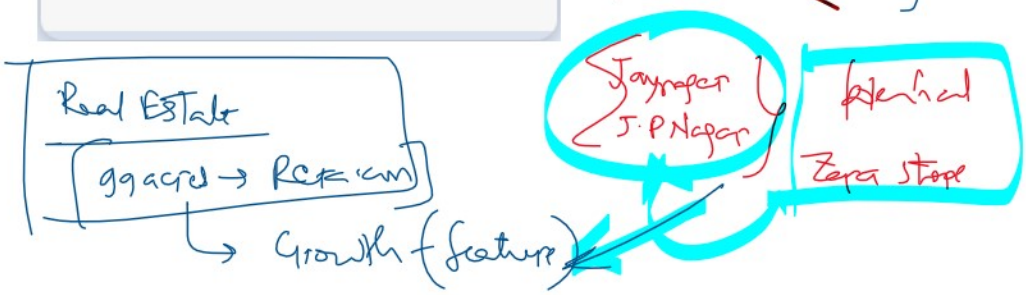
$\frac{d\,WCSS}{dK}$ ↓↓↓ —

K=5 → no of optimal clusters

Testing different no. of clusters and measuring the resulting WCSS and choosing the 'K' value at which an increase in 'K' (say K=5 to K=6) will cause a very small decrease in WCSS.

one more cluster.





Task # Hierarchical Clustering
- Agglomerative
- Divisive.