K-Means Clustering

Sunday, 1 June 2025 10:00

Unsupervised Learning

↓

Data with input cols only

Amazon Customer's Data

↓

	Name	Age	City	Gender	Purchases	Amount spent ↓
0	—	—	—	—	—	—
1						
2						
3						
4						
5						

↓

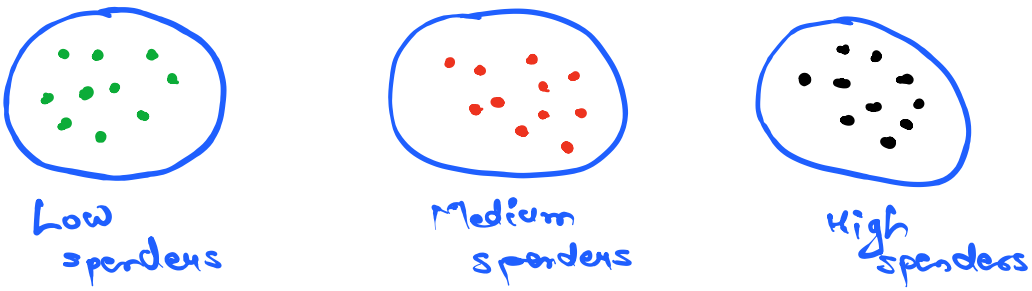
Unsupervised Algorithm

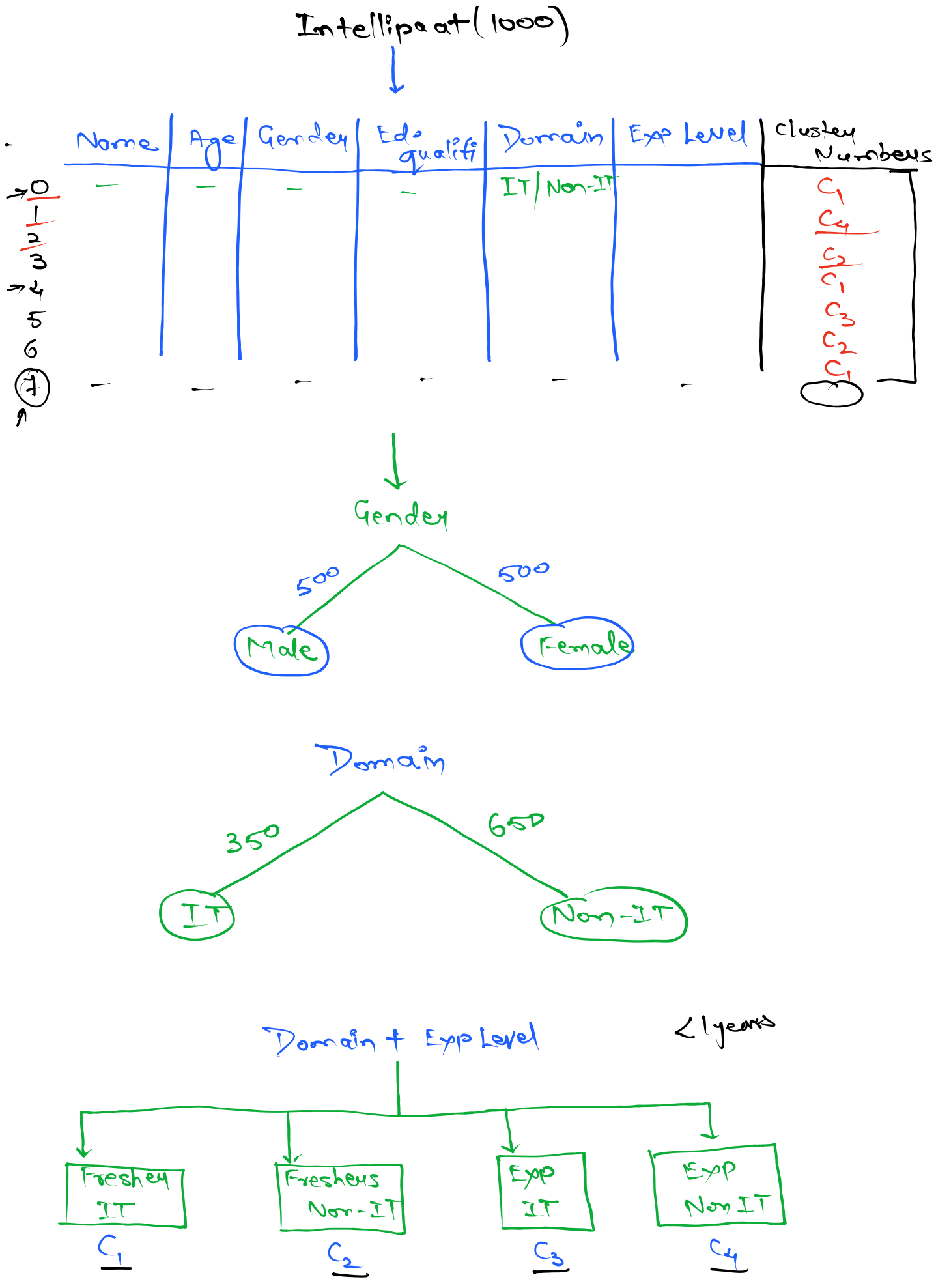
↓

They will learn two things :

→ How many groups / clusters exist in our data ?

→ Which customer / datapoint belongs to which cluster.




Intellipaat (1000)
Name
Age
Gender
Edu qualifi
Domain
Exp Level
Cluster Numbers
IT/Non-IT
C1
C4
C2
C1
C3
C2
C1
Gender
500
500
Male
Female
Domain
350
650
IT
Non-IT
Domain + Exp Level
< 1 years
Fresher IT
Freshers Non-IT
Exp IT
Exp Non IT
C1
C2
C3
C4

1. Targeted Marketing
2. Recommendations
3. offers, coupons, discounts.

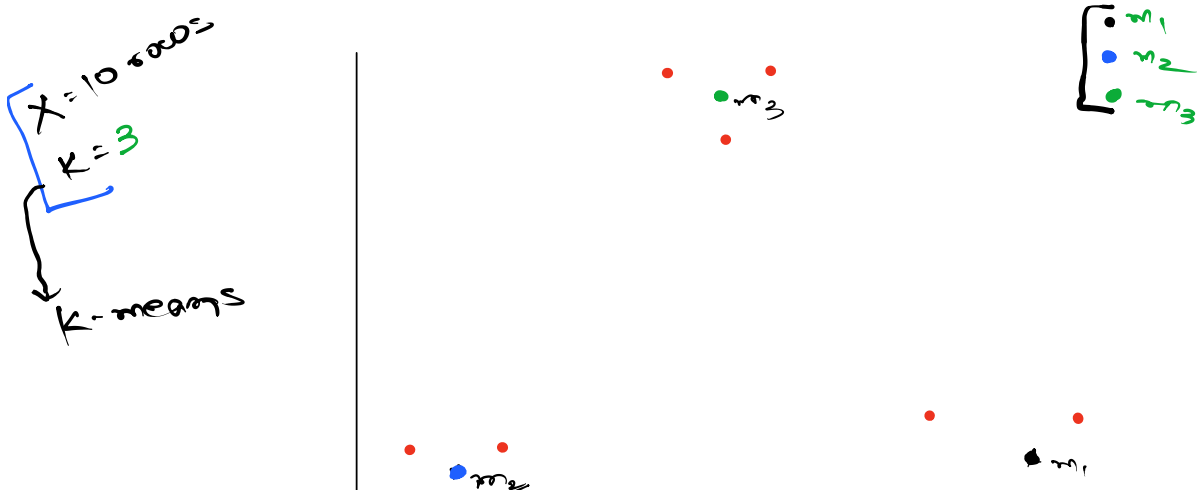
Unsupervised Learning Algorithms →

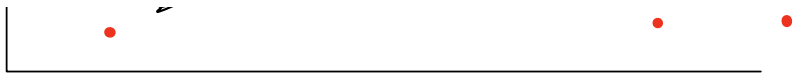
1. K-Means Clustering.
2. Hierarchical Clustering.
3. DBScan Clustering.
4. Gaussian Mixture Models (GMMs)

K-Means Clustering

Two requirements of k-means :

1. Data (input cols) → X
2. k-value : No of clusters to be identified.





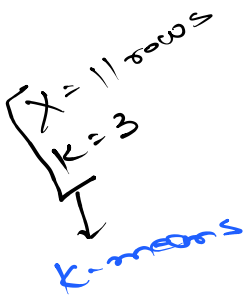
① Create centroids/cluster means on the data randomly.

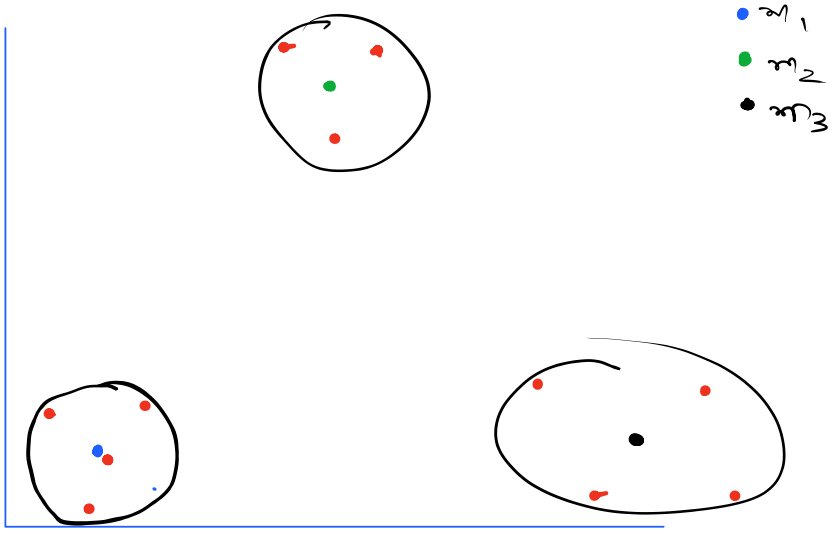
② Calculate the distance b/w each datapoint to each centroid.

③ Put together the datapoints with their closest centroids in clusters.

④

⑤





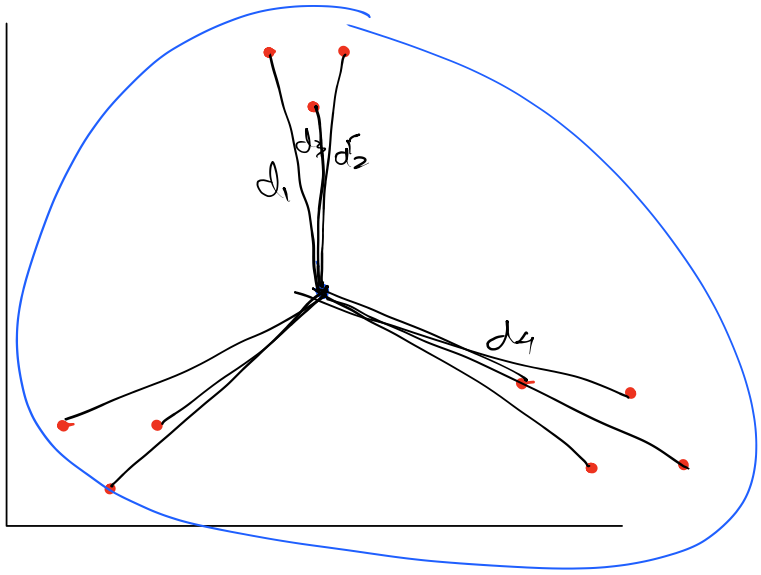
Finding optimal k-value using wcss/elbow method:

WCSS: Within Cluster Sum of squared distances.

or

SSD

K=1
K=2
K=3
⋮

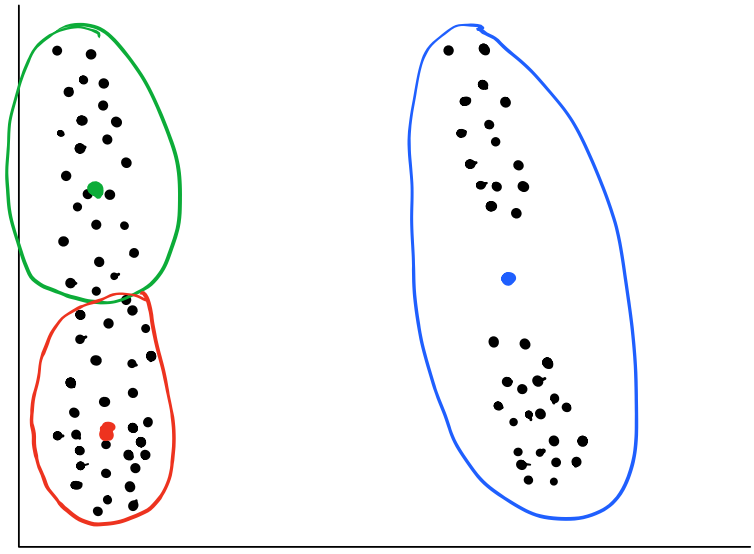


$$WCSS(K=1) = d_1^2 + d_2^2 + d_3^2 + \text{----} + d_{10}^2$$

$$= 2023$$

Initialization Trap :

$K = 2$



To overcome this trap, we now have updated version of K-means algorithm which is known as 'K-means++'.

K-means++