Clustering Algorithms

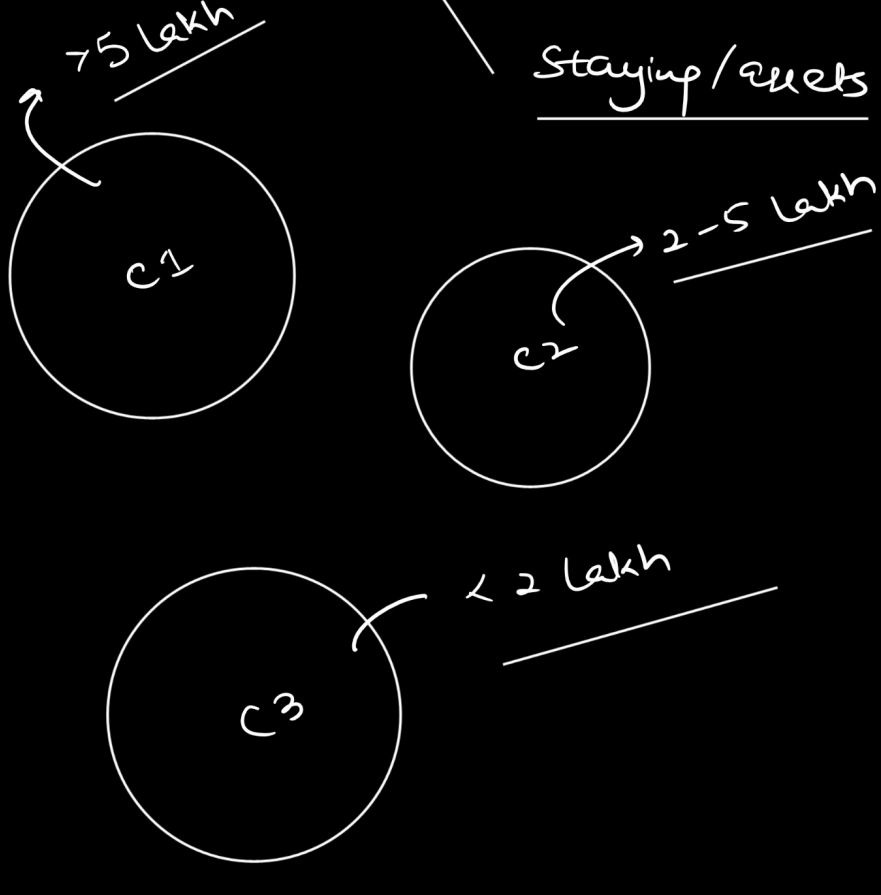
→ unsupervised machine learning tasks

unlabeled data (there is no target column available)

1. Customer Segmentation

Banking — monthly salary, CIBIL score

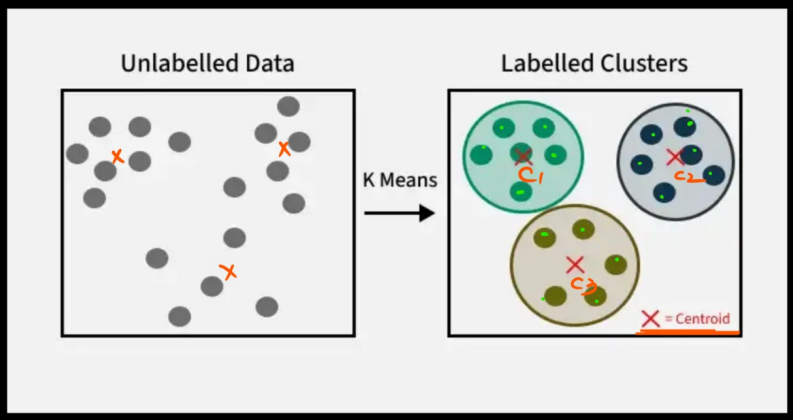
Staying / guests



- K-Means
- Hierarchial
- DBSCAN

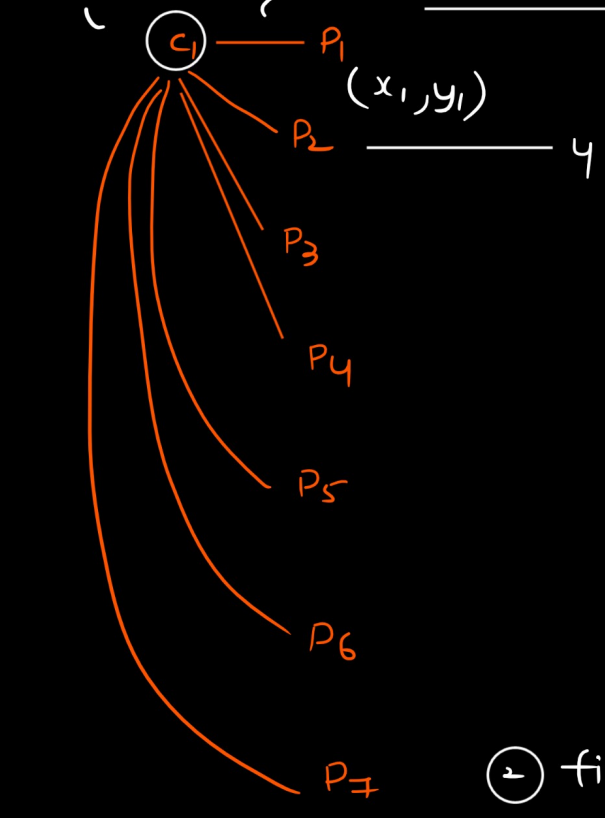
clusters = 3

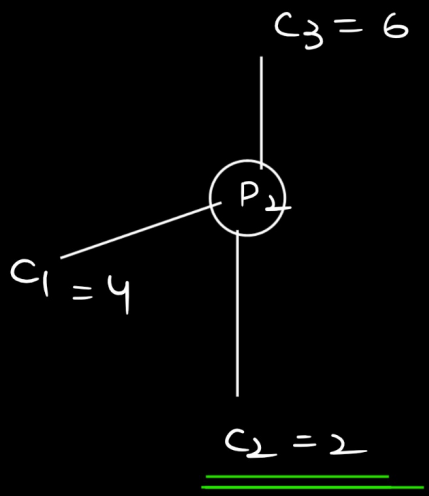
K-Means



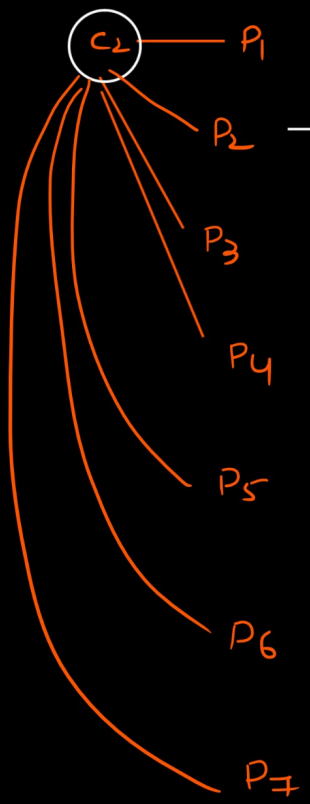
① Define the value of K (# clusters)

Euclidean Distance → $\sqrt{(x_1-x)^2+(y_1-y)^2}$





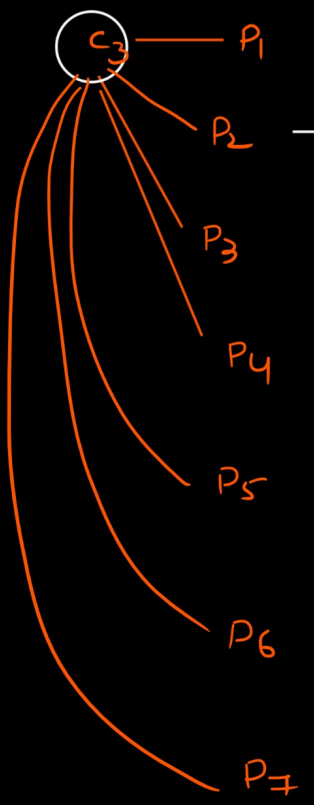
② finding the nearest points corresponding to each Centroid



2

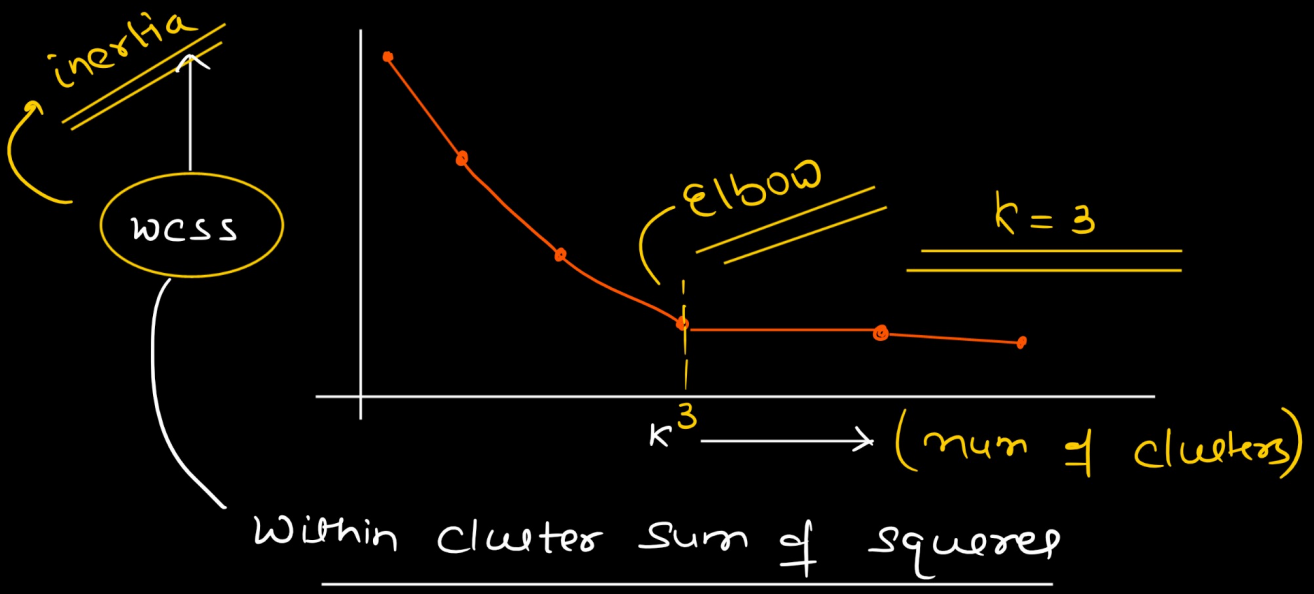
③ Average of all points available in each cluster to find the new centroids

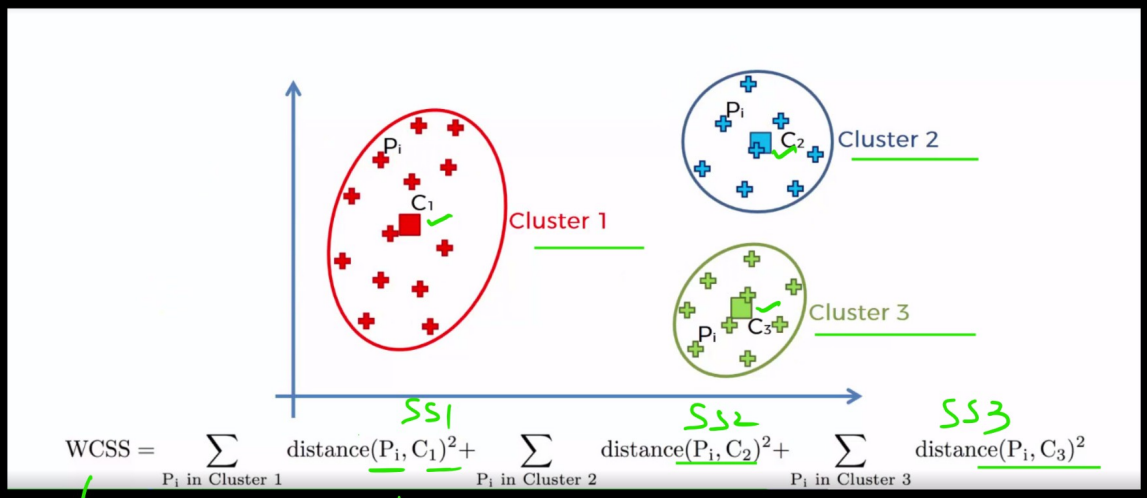
④ This process keeps on repeating until the model reaches to a point where it stops updating the centroids position.



6

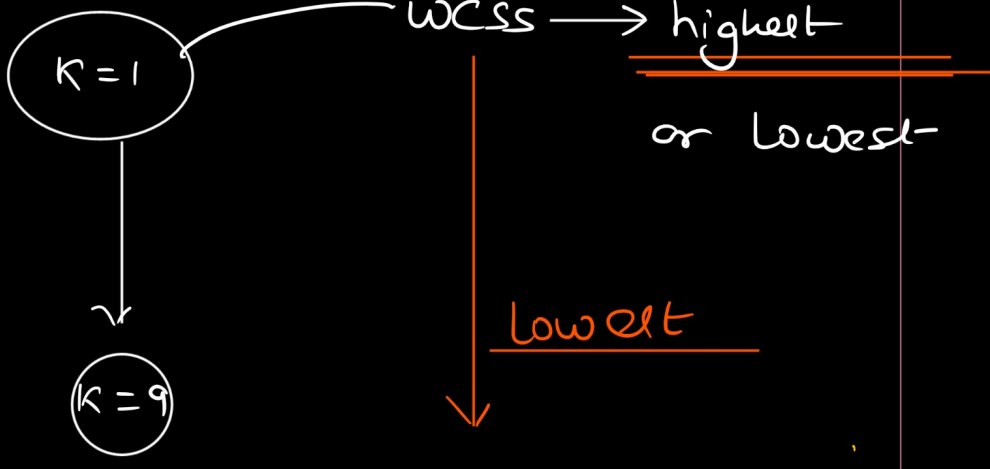
Elbow Method



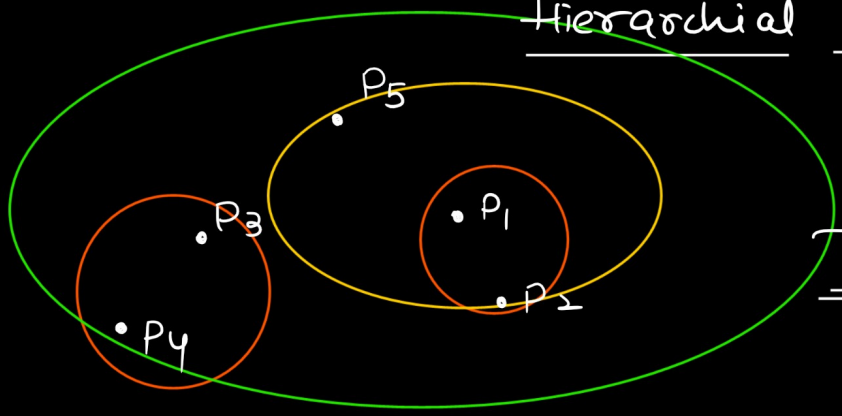


Within cluster sum of squared

Sum of squared

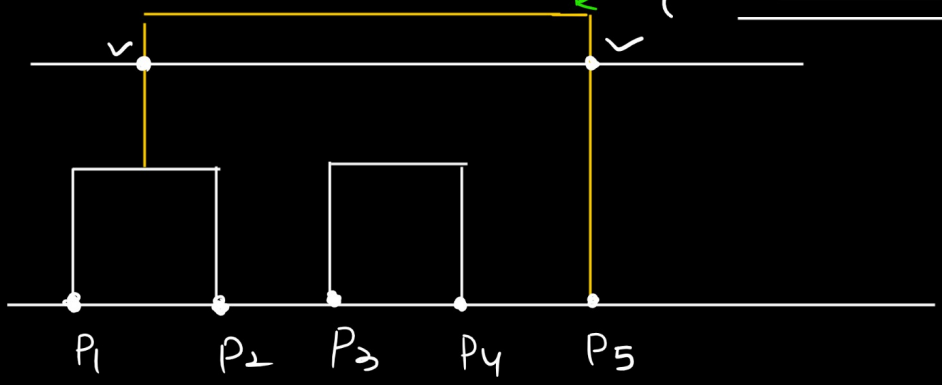


Hierarchical Clustering



Dendogram

num of clusters



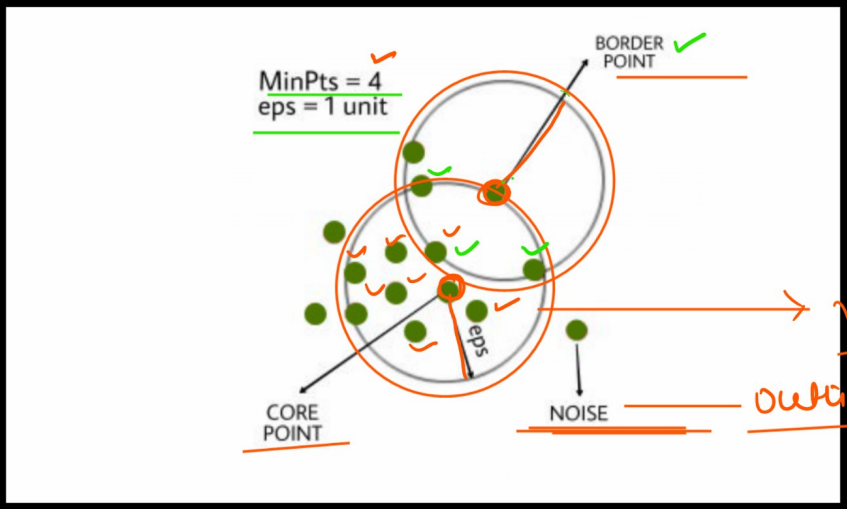
xx

DBSCAN → outliers

↳ eps → radius

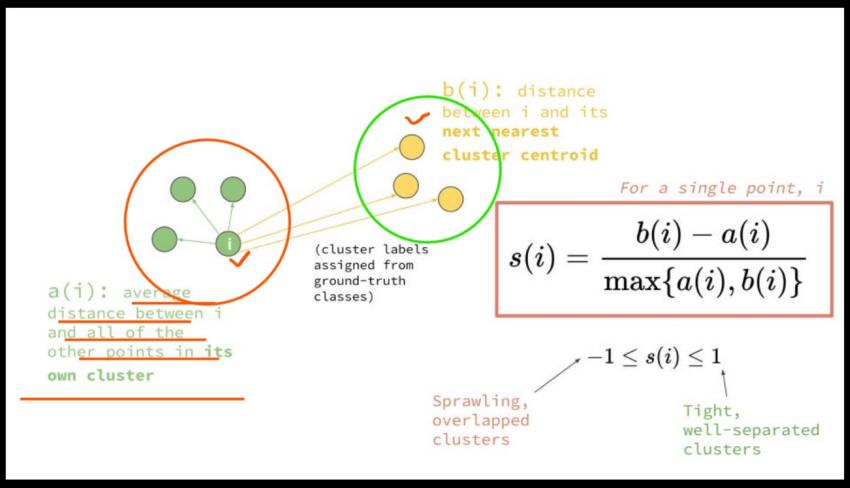
minpts

- Core_points
- border_points



→ valid cluster

outlier



→ b(i) → distance b/w i & next nearest cluster centroid

b(i) > a(i) ← ✓

or

b(i) < a(i) ← -ve

+ve

→ not having a good cluster