Any doubts ?!

Stats,

Prob,

Sampling,

Hypothesis Test,

ML,

LR,

LR

Python Implementation!

Boxplot in Python?

Pandas?

function?

df. / Series

?

Heatmap?

Evaluation of Models?

Sklearn.metrics

StandardScaler?

sklearn.preprocessing

Prob. of Prediction! → which method of that ML model

model.predict_proba(features)

Xtrain

Xtest

→ Clustering → Create clusters / group / segments → Data Mining

↓

Set of Similar objects → rows

→ No Target required

features → Clustering ML Algo → Clusters

↓

Data

↓

No splitting required

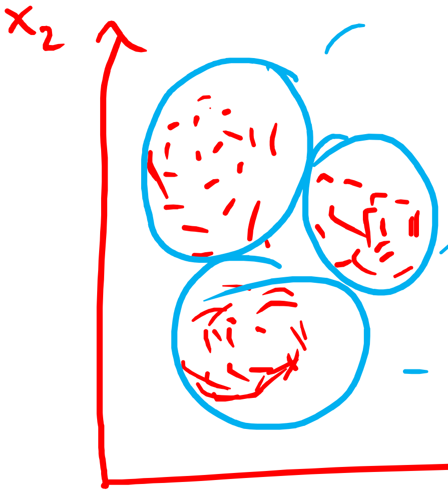
→ Quality ↑ Evaluation

2500, 7550

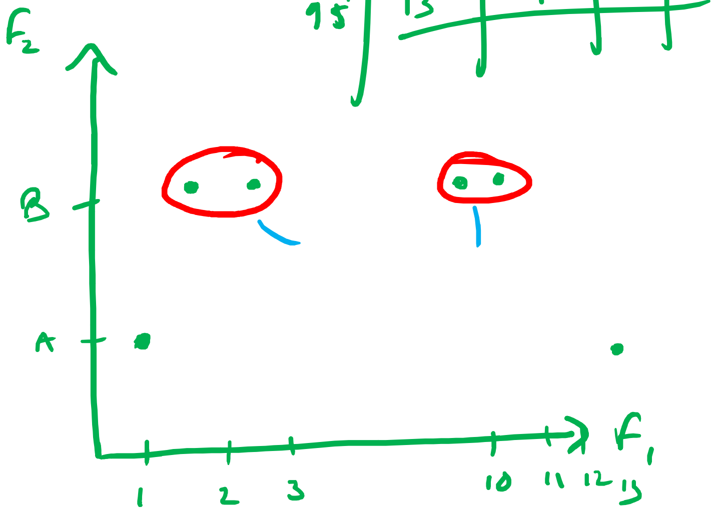
=) 100 Rows

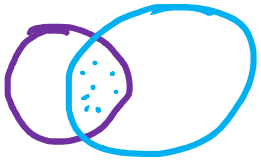
L

	F_1	F_2		
90	1	4		
91	10	B		
92	2	B		
93	1,	B		
94	3	B		
95	13	A		



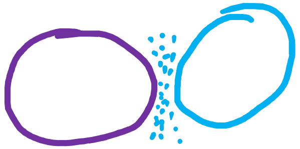
→ objective / o/p of clustering





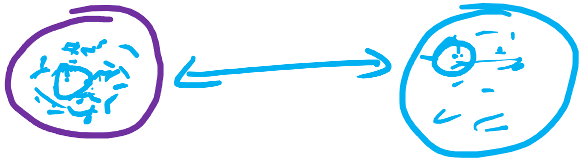
I

Incorrect clustering



II

Bad clustering



III

Good clustering

Clusters → collection of objects. , which are "Perceived" to be similar b/w them & dissimilar to objects belonging to diff cluster.

→ K - Means

↓

arbitrary Number

K ⇒ 1 , N → | K = {3 , 20}

↓

How many clusters!

Statistically a healthy cluster can have about 5% - 35% of the total data

20 ⟷ 3

1%

50% ✗

100 Rows → [illegible]

↓ 100

↓ 1000 clusters

1 Row in each. cluster

$$\frac{\cancel{100}^{25}}{\cancel{4}}$$

K=4

21%

5%

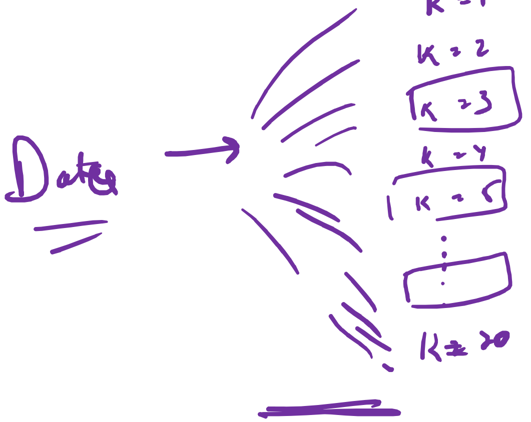
1

100

1M

On every data → 20 clusters →

↓



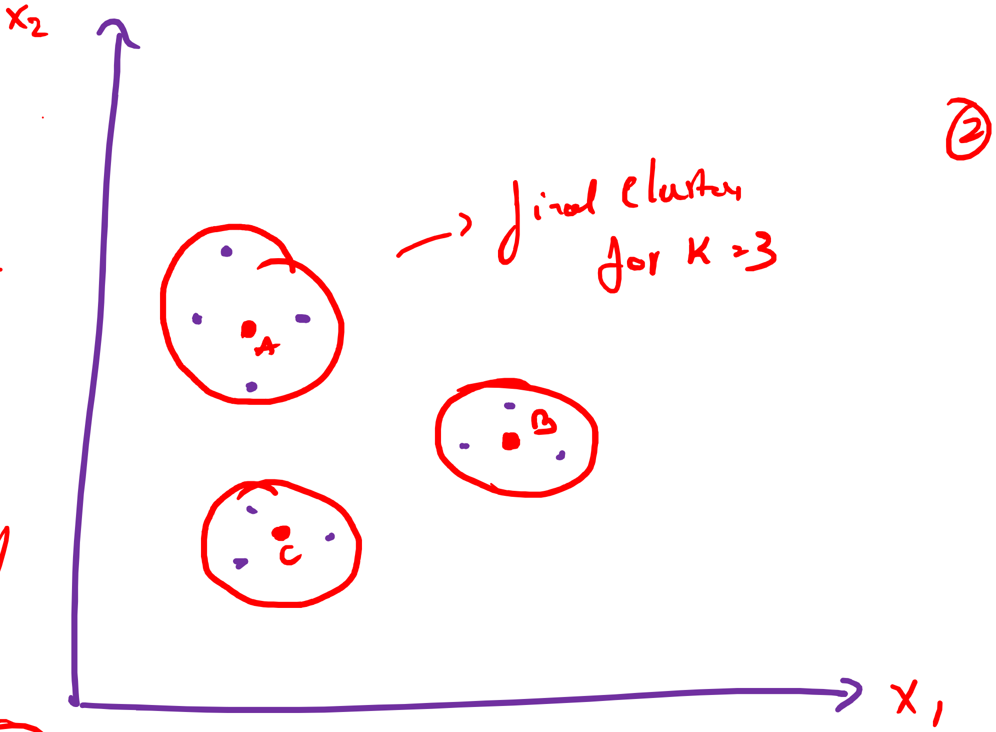
→ we will select the best value of K on that dataset

↓

final output

→ K-Mean Algorithm

K = 3



① will Add K Random points to your data
These are known as Centers/ Centroids/ Mean

② from each data point, it will calculate the distance to each added center

	A	B	C
→ 1	3	5	4
→ 2	4	2	3
→ 3	0 –	–	– –
	–	–	0

5 — – –

2 0 —

③ Group of shortest distance will be center

④ The centers will be recalculated/ shifted

⑤ Repeate from step ② Till in consecutive 2 Iterations you get same centers