30-08-2025

Agenda - FE-II.

2. Handling Outliers

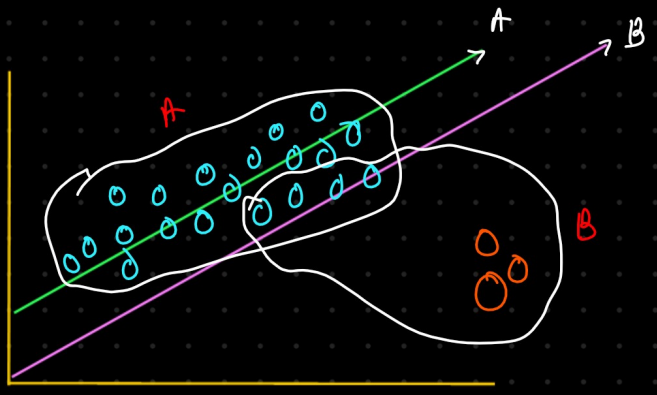
An outlier is a data point that is very different from the rest.

5m (250 cm)

150 ⟵ Students ⟶ 180

→ mean (distant mean)

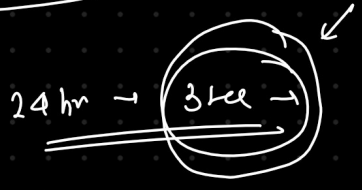
ML model → Linear Regression



Case A:

1 Dec - million

machine: NR → 1 million → zero



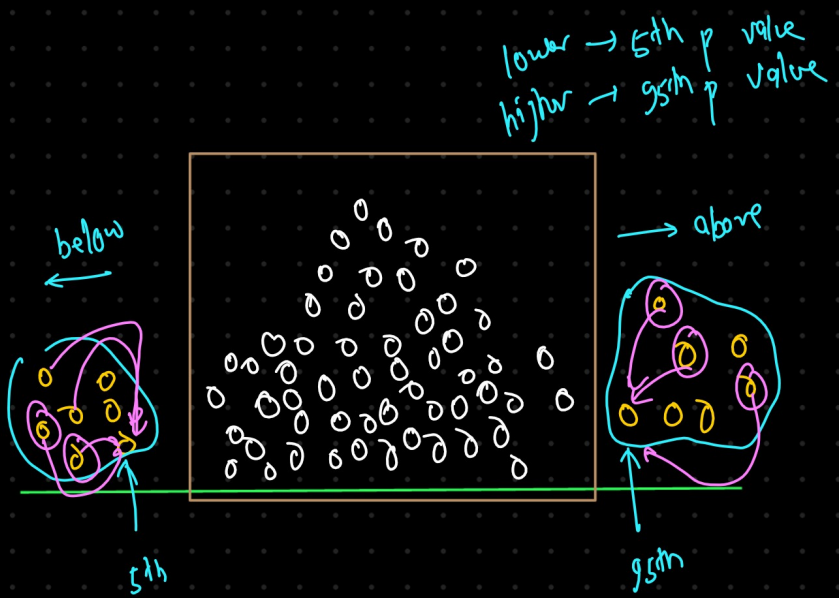
Detect & Handle outliers:

(1) Box plot (visualization)

(2) IQR

(3) Z-Score

(4) winsorization →



before winsorization:

Total : 100

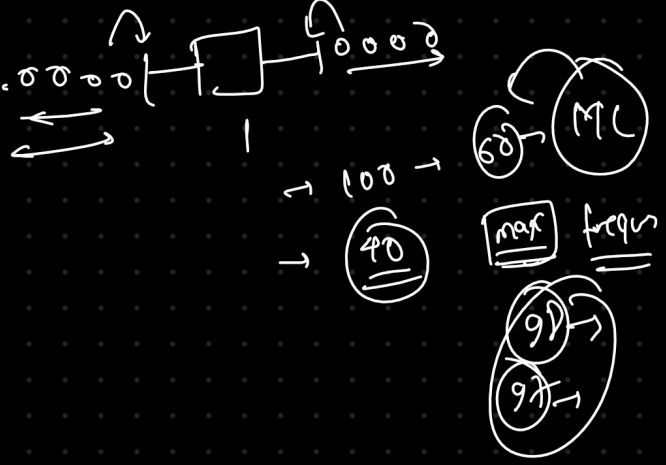
after outlier removal: 90

outlier removed: 10

winsorization:

Total : 100

after application: 100



(1) Finding missing values

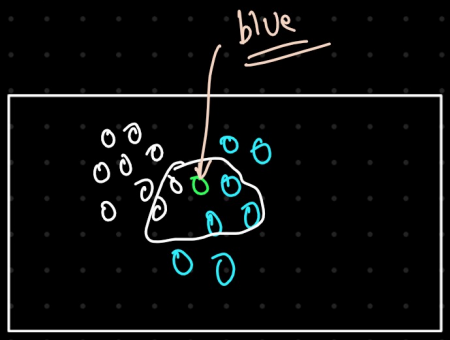
(2) Handling missing value

(3) Handling outliers

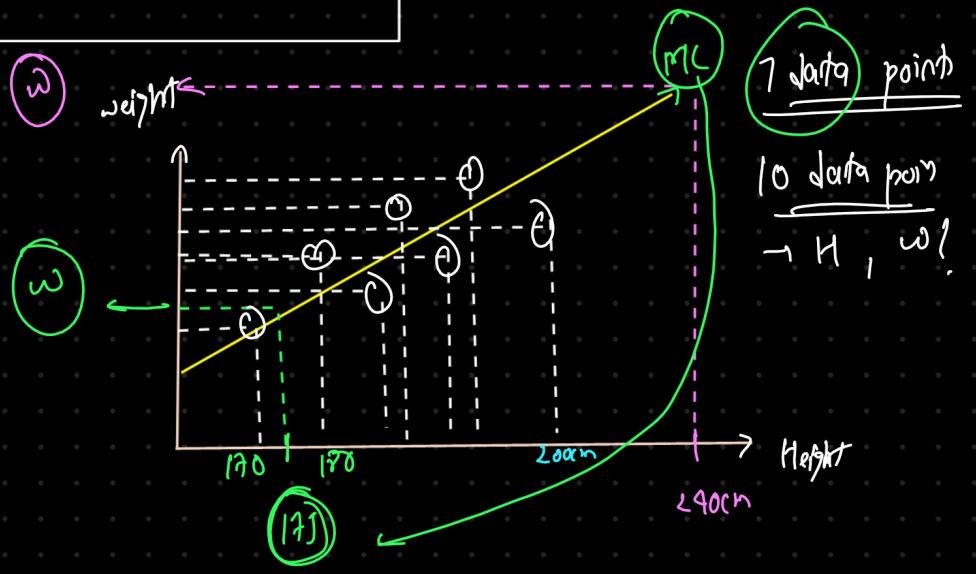
(4) Feature Scaling ← ML → calculation

(2) Handling Missing value

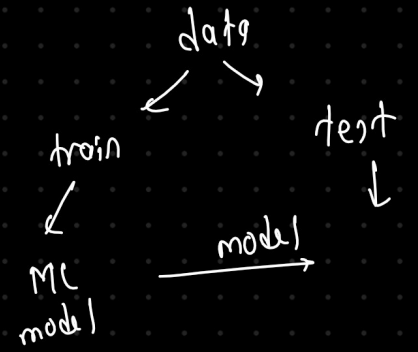
KNN
(ML algo)

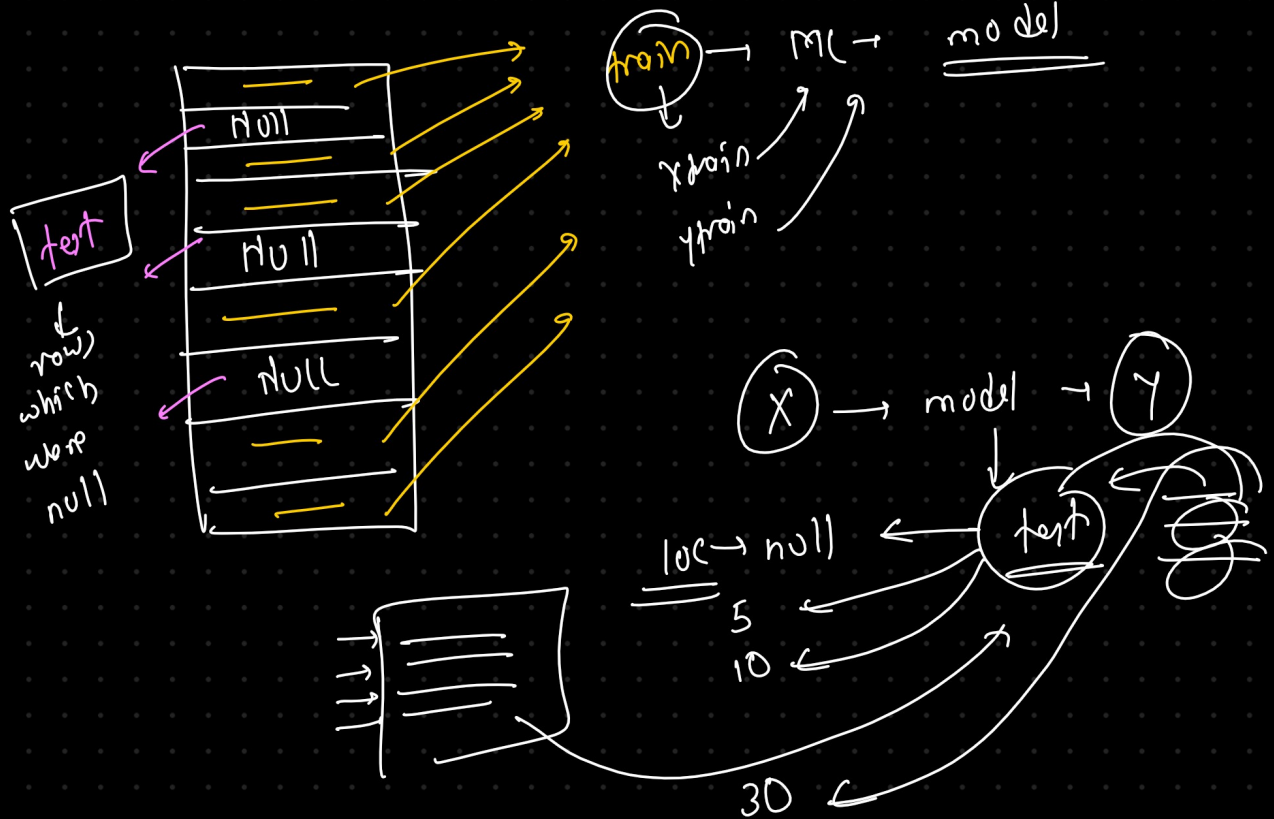


Linear Regression
(ML algo)



X → ML → Y





10
20
50
~~NULL~~
30
40
~~NULL~~
~~NULL~~
~~NULL~~
60
70
100

row-count : 12

NULL-rows : [4, 7, 8, 9]

features ← df.dropna(w)

target ←

all rows without null in w

df[df[w].isnull()] → [true, false ...]

boolean

→ rows where value is true

	Height	BMI	Age	weight
1	A_1	B_1	C_1	D_1
2	A_2	B_2	C_2	NULL
:				
12	A_3	B_3	C_3	NULL