④ Outliers

→ What is outlier ?

• An outlier is a value that is unusually far from most of the other values in your dataset.

Eg ex:- Every one earns 25k - 40k per month...
and one person earns 5 cr.

→ That 5cr guy? outlier.

Eg:- Ages of students :- 18, 19, 20, 21, 22, 99.

→ 90 is clearly an outlier.

→ Why outlier is matter in ML ?

• outliers can

① Skew mean
② Break linear regression
③ Confuse distance based models (KNN, K-means)
④ Cause poor generalization.

→ Not all outliers are bad.
Sometimes they are valuable signals.
(fraud detection, anomly detection)

→ Types of outliers :-

① Global outliers :- Far away from all values.

Ex :- Income = 20k, 25k, 25k, 24k, [1 CR]

② Contextual outliers :- outliers depends on the context.

Ex :- Temperature - Summer → 45°c (Normal)
Winter → 45°c (outlier)

③ Collective outliers :- A group of abnormal points.

Ex :- Sudden Spike in website trafic at midnight.

→ How to detect outlier ?

① visual :- Box plot, Scatter plot, Histogram.

② IQR method :- lower Bound = $Q_1 - 1.5(IQR)$
upper bound = $Q_3 + 1.5(IQR)$

If the Value < [Lower bound]
If the Value > [upper bound] → outlier

IQR Exam

10., 12

Q_1

→ $Q_1 = 12$
$Q_3 = 2$
IQR =
=

$0., 100

IQR Example :-

10, 12, 14, 15, 18, 20, 100

Q_1 → 12 ; Q_3 → 20

$\Rightarrow Q_1 = 12$

$Q_3 = 20$

$IQR = 20 - 12$

$= 8.$

① Upper Bound $= 20 + 1.5(8)$

$= 32.$

② Lower Bound $= 12 - 1.5(8)$

$= 0.$

So., 100 is a outlier.