Dr. Subramani Suresh

What is an Outlier ?

An outlier is a data point that is very different (much ⬆️ higher or ⬇️ lower) compared to the rest of the values in a dataset.

Why do outliers happen?

Natural Variation
Some values are just unusually high or low by nature.

📌 Example: Most students score between 40–70 marks, but one student scores 100.

Data Entry Errors
Mistakes while recording data.

📌 Example: Typing 500 instead of 50 for a student's marks.

Rare Events
Something unusual happens that shifts the data.

📌 Example: Example: A shop usually sells 20–30 items daily, but on a festival day it sells 200 items.

Why are outliers important?

👉They can twist the mean (average). 👉They may indicate errors or special cases worth studying. 👉They affect models like regression, clustering, etc.

How to find outliers in a dataset?

👉You can find outliers in multiple ways.
✅ The most common is using the IQR (Interquartile Range) method.

Step1 : Calculate Quartiles Q1,Q3 percentile : ex. Data = [-15,15,20,25,30,35,40,45,90]

A percentile shows the value below which a given percentage of data falls

IQR (Inter Quartile Range)

Q_0 Minimum — 0^{th} percentile

Q_1 — 25^{th} percentile

Q_2 Median — 50^{th} percentile

Q_3 — 75^{th} percentile

Q_4 Maximum — 100^{th} percentile

$$Q1 = \text{Value at position } P = \frac{25}{100} \times (n+1)$$

P=(25/100)x(9+1)=
2.5^{th} position= Q1 value =20

$$Q3 = \text{Value at position } P = \frac{75}{100} \times (n+1)$$

P=(75/100)x(9+1)=
7.5^{th} position= Q3 value =40

n = total number of data points
If P is an integer, Q = value at the P-th position
If P is not an integer, interpolate between the two closest data points.

Step 1: Q1 percentile = 20 , Q3 percentile = 40

Step 2: Compute IQR
IQR = Q3 − Q1
40 − 20 = 20

Step 3: Calculate bounds & Identify
👉 Lower Bound = Q1 − 1.5 × IQR
📌 20 − 1.5 × 20 = - 10

Any value < Lower Bound → Lower outlier (-15 Lower outlier)

👉 Upper Bound = Q3 + 1.5 × IQR
📌 40 − 1.5 × 20 = 70

Any value > Upper Bound → Upper outlier (90 Upper outlier)

💡 Reason for 1.5 × IQR rule
The IQR (Q3 − Q1) covers the middle 50% of the data.
💡 Multiplying by 1.5 extends this range to about 99% of a normal (bell- shaped) distribution.

Python

```
import numpy as np
# Example dataset
data = np.array([-15, 15, 20, 25, 30, 35, 40, 45, 90])
# Step 1: Calculate Q1, Q3
Q1 = np.percentile(data, 25)
Q3 = np.percentile(data, 75)
# Step 2: Compute IQR
IQR = Q3 - Q1
# Step 3: Calculate bounds
lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR
# Step 4: Find outliers
outliers = data[(data < lower_bound) | (data > upper_bound)]
print("Q1 (25%):", Q1)
print("Q3 (75%):", Q3)
print("IQR:", IQR)
print("Lower Bound:", lower_bound)
print("Upper Bound:", upper_bound)
print("Outliers:", outliers)
```

```
Q1 (25%): 20.0
Q3 (75%): 40.0
IQR: 20.0
Lower Bound: -10.0
Upper Bound: 70.0
Outliers: [-15  90]
```