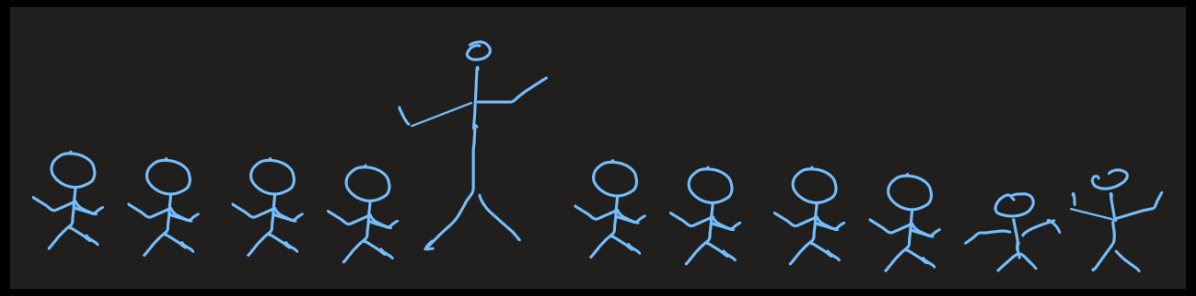
Can you tell which numbers are **statistically extreme points (AKA outliers)** ?

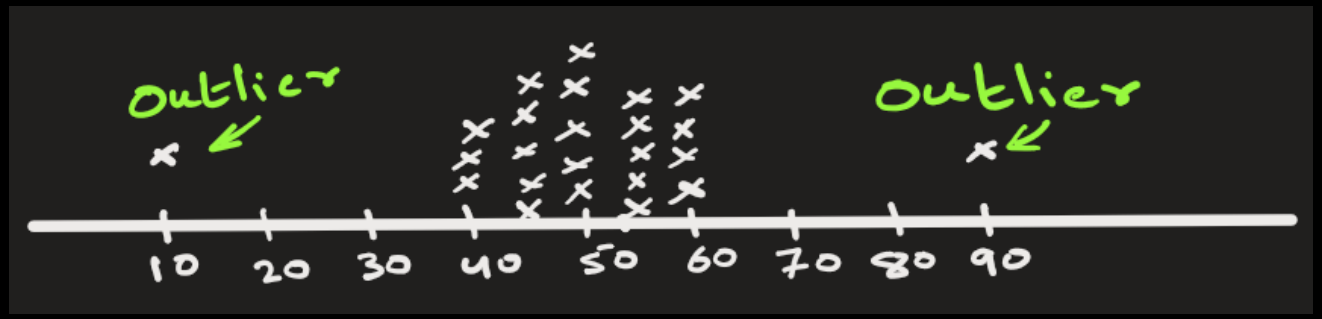
52, 48, 45, 44, 49, 55, 40, 46, 44, 53, 47, 47, 48, 52, 48, 34, 52, 52, 48, 44, 51, 51, 47, 48, 45, 49, 42, 52, 45, 51, 42, 47, 49, 41, 50, 48, 41, 43, 53, 54, 28, 31, 68, 75, 74

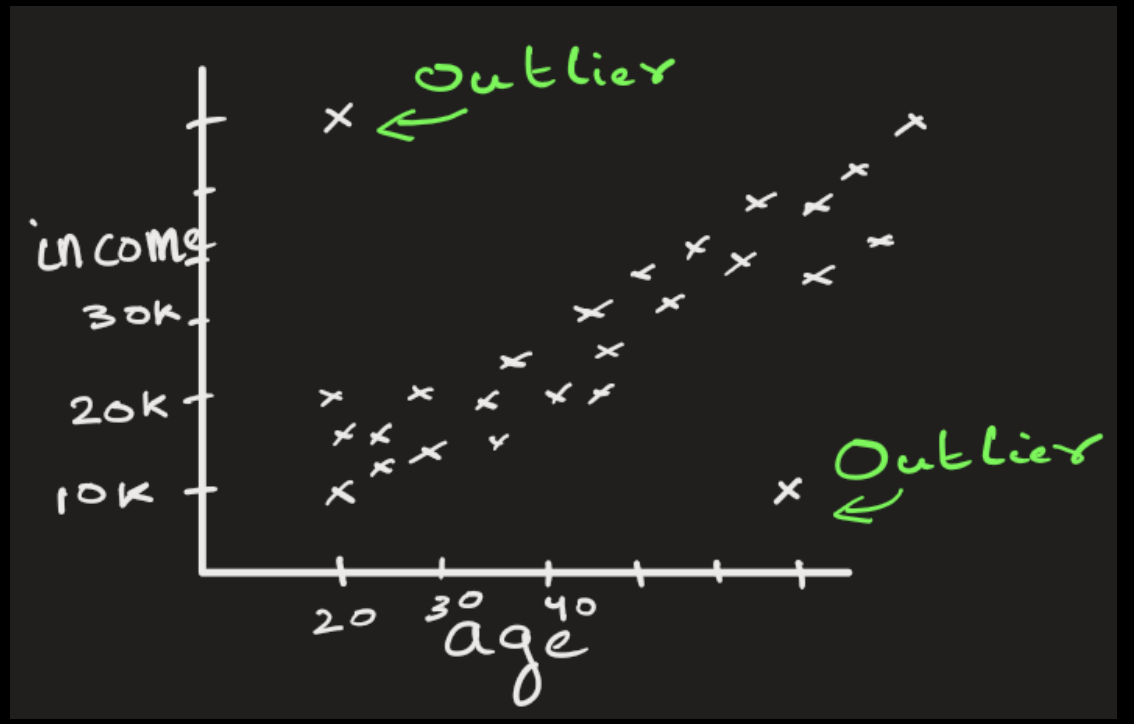


OUTLIERS

What are outliers ?

Outliers are data points that **significantly deviate from the average or typical values** within a dataset.



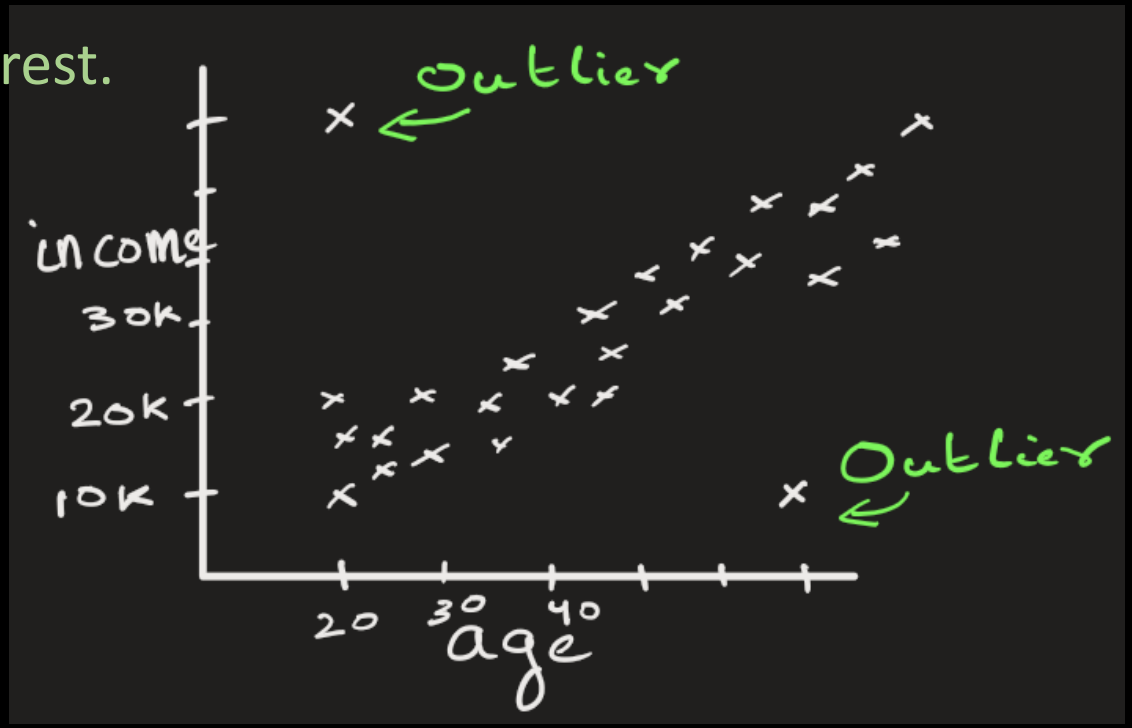


OUTLIERS

Why do they happen?

They can arise due to **measurement errors, experimental anomalies, or truly exceptional observations.**

Example of truly exceptional observations:
Maybe the young kid has higher education than the rest.
Maybe the older guy is not experienced.




Problems with outliers

- Mean:

3, 3, 4, 5, 5, 7 => mean = (3 + 3 + 4 + 5 + 5 + 7) / 6 = 4.5

3, 3, 4, 5, 5, 70 => mean = (3 + 3 + 4 + 5 + 5 + 70) / 6 = 15

Presence of outliers changed the mean value drastically

- Median are not affected by presence of outliers:

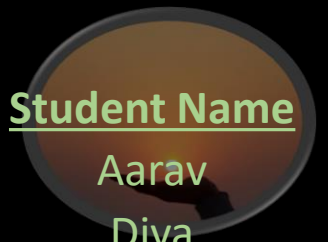
3, 3, 4, 5, 5, 7 => median = (4 + 5) / 2 = 4.5

3, 3, 4, 5, 5, 70 => median = (4 + 5) / 2 = 4.5

Identify outliers: Using IQR

IQR
Inter Quartile Range

Identify outliers: Using IQR



Student Name	Score/age/income
Aarav	52
Diya	48
Rohan	45
Ananya	44
Kabir	49
Isha	55
Vivaan	40
Meera	46
Arjun	44
Neha	53
Aditya	47
Pooja	47
Rahul	48
Sneha	52
Kunal	48
Priya	34
Siddharth	52
Nisha	52
Manish	48
Kavya	44
Amit	51
Ritu	51
Varun	47
Shreya	48
Nikhil	45
Tanvi	49
Suresh	42
Pallavi	52
Mohit	45
Ayesha	51
Rakesh	42
Simran	47
Deepak	49
Ankit	41
Bhavya	50
Reena	48
Vikas	41
Tina	43
Gaurav	53
Sonali	54
Harsh	28
Riya	31
Naveen	68
Pankaj	75
Lokesh	74

Identify outliers: Using IQR

- Step1: Order the dataset:

Actual dataset:

52, 48, 45, 44, 49, 55, 40, 46, 44, 53, 47, 47, 48, 52, 48, 34, 52, 52, 48, 44, 51, 51, 47, 48, 45, 49, 42, 52, 45, 51, 42, 47, 49, 41, 50, 48, 41, 43, 53, 54, 28, 31, 68, 75, 74

Ordered dataset:

28, 31, 34, 40, 41, 41, 42, 42, 43, 44, 44, 44, 45, 45, 45, 46, 47, 47, 47, 47, 48, 48, 48, 48, 48, 48, 49, 49, 49, 50, 51, 51, 51, 52, 52, 52, 52, 52, 53, 53, 54, 55, 68, 74, 75

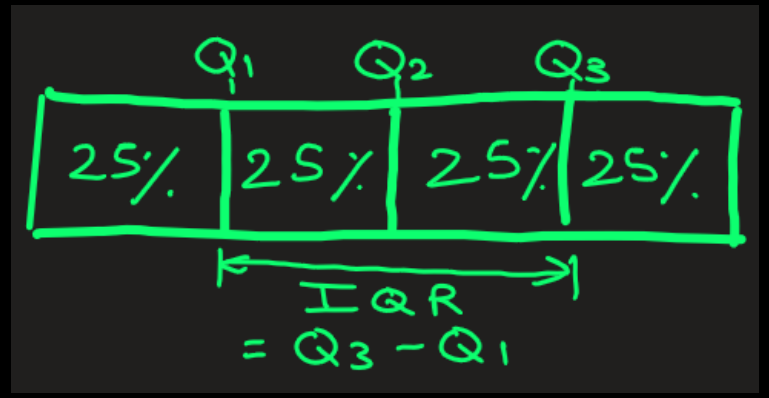
Identify outliers: Using IQR

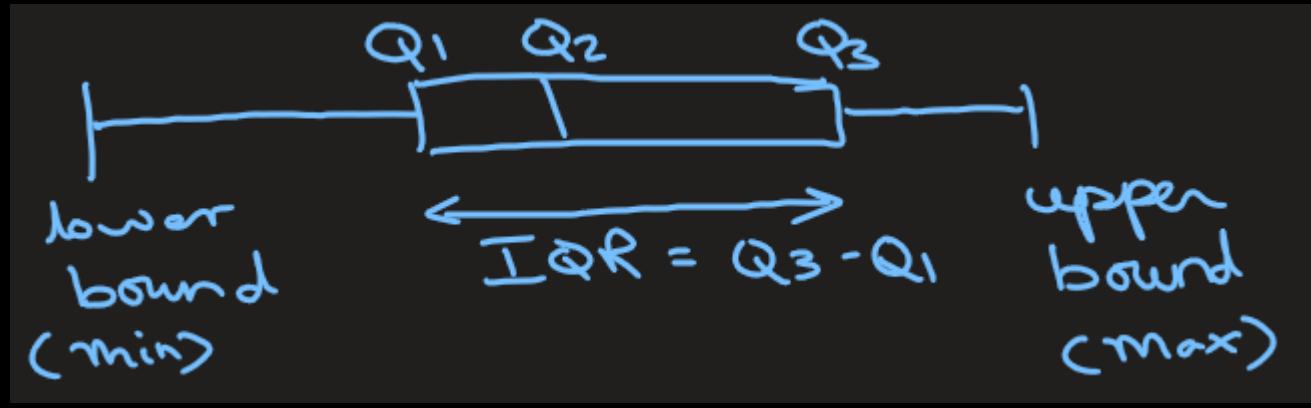
- Step2: Find Q1, Q3 and IQR

28, 31, 34, 40, 41, 41, 42, 42, 43, 44, **44**, 44, 45, 45, 45, 46, 47, 47, 47, 47, 48, 48, **48**,
48, 48, 48, 49, 49, 49, 50, 51, 51, 51, 52, 52, 52, 52, 52, 53, 53, 54, 55, 68, 74, 75

Here Q1 = 44, and Q3 = 52.

IQR = Q3 – Q1 = 52 – 44 = 8





Identify outliers: Using IQR

- Step3: Identify Lower and upper bound

Lower bound = Q1 – 1.5 × IQR = 44 – 1.5 x 8 = 32

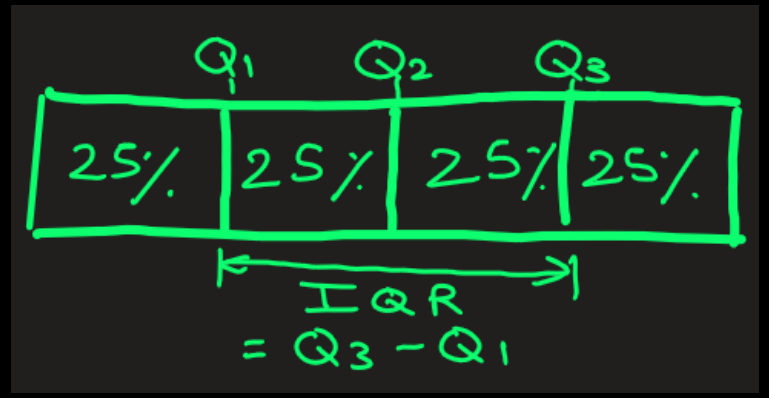
Upper bound = Q3 + 1.5 × IQR = 52 + 1.5 x 8 = 64

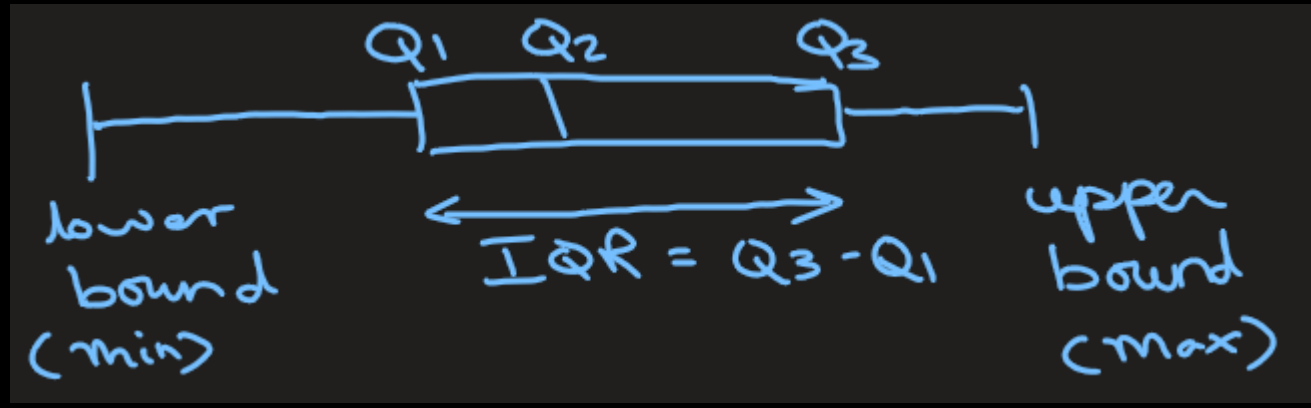
- Step4: Points outside the lower and upper bound are outliers.

Data = 28, 31, 34, 40, 41, 41, 42, 42, 43, 44, 44, 44, 45, 45, 45, 46, 47, 47, 47, 47, 48, 48, 48, 48, 48, 48, 49, 49, 49, 50, 51, 51, 51, 52, 52, 52, 52, 52, 53, 53, 54, 55, 68, 74, 75

Points 28, 31, 68, 74, 75 are outside the range of lower and upper bound.

So, outliers = 28, 31, 68, 74, 75. These values represent **statistically extreme points** in the dataset.



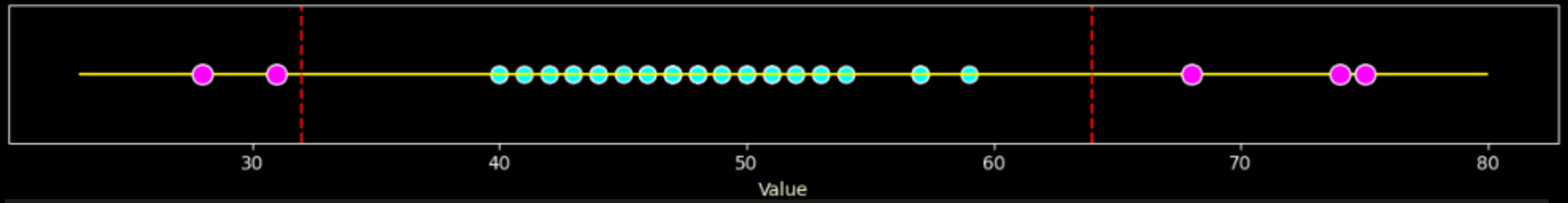


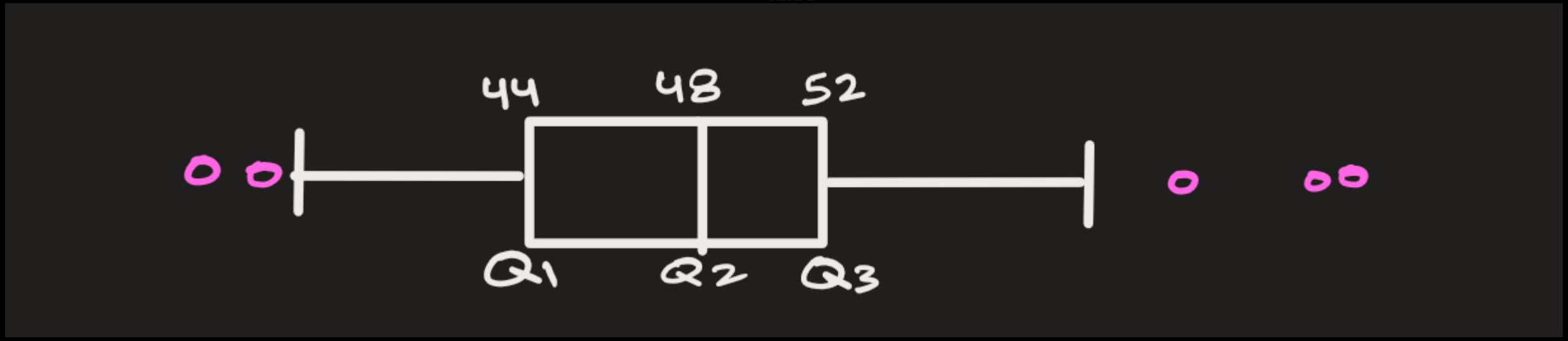
Identify outliers: Using IQR

Final Result

Ordered dataset:

28, 31, 34, 40, 41, 41, 42, 42, 43, 44, 44, 44, 45, 45, 45, 46, 47, 47, 47, 47, 48, 48, 48, 48, 48, 48, 49, 49, 49, 50, 51, 51, 51, 52, 52, 52, 52, 52, 53, 53, 54, 55, 68, 74, 75



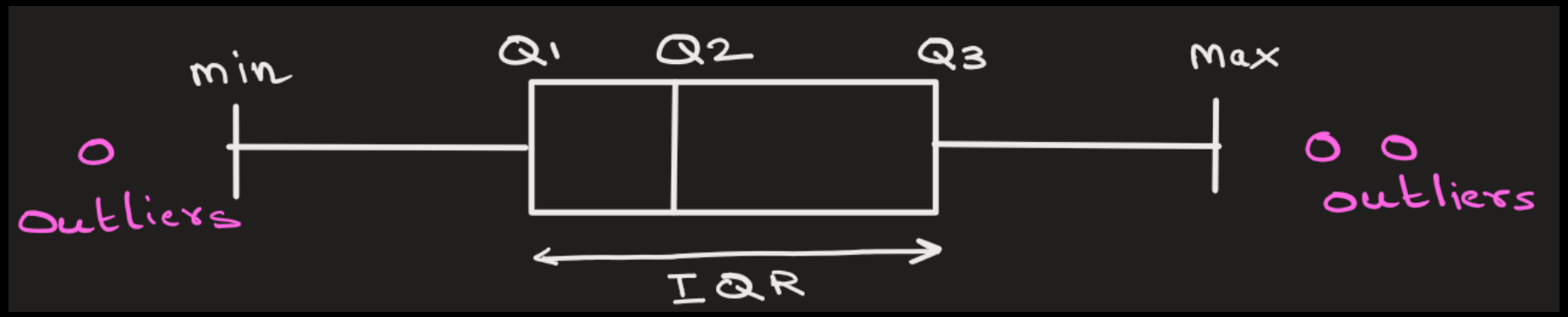


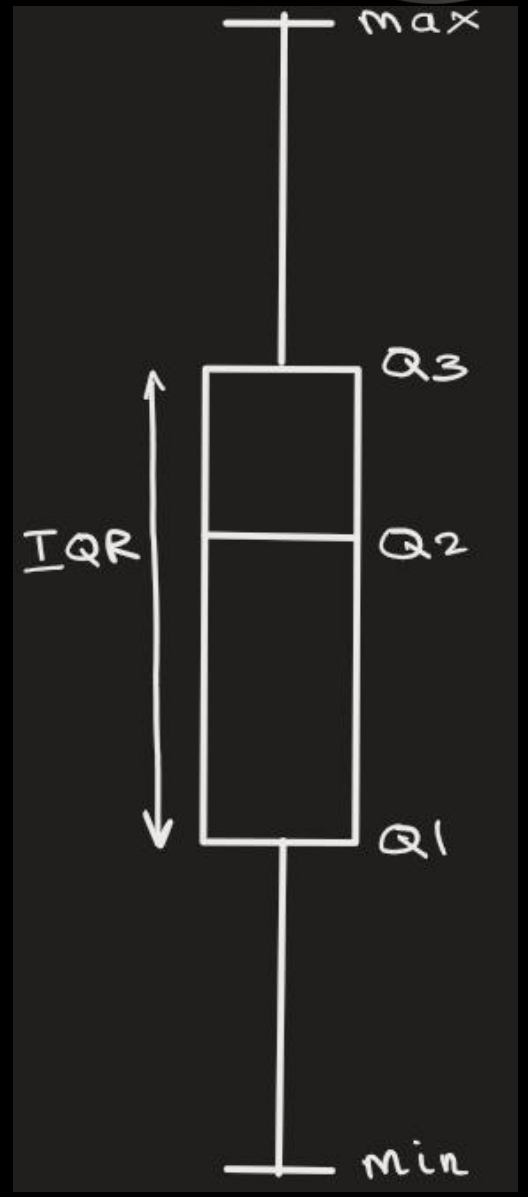
Identify outliers: Using Box Plot

Box Plot:

Horizontal

Vertical



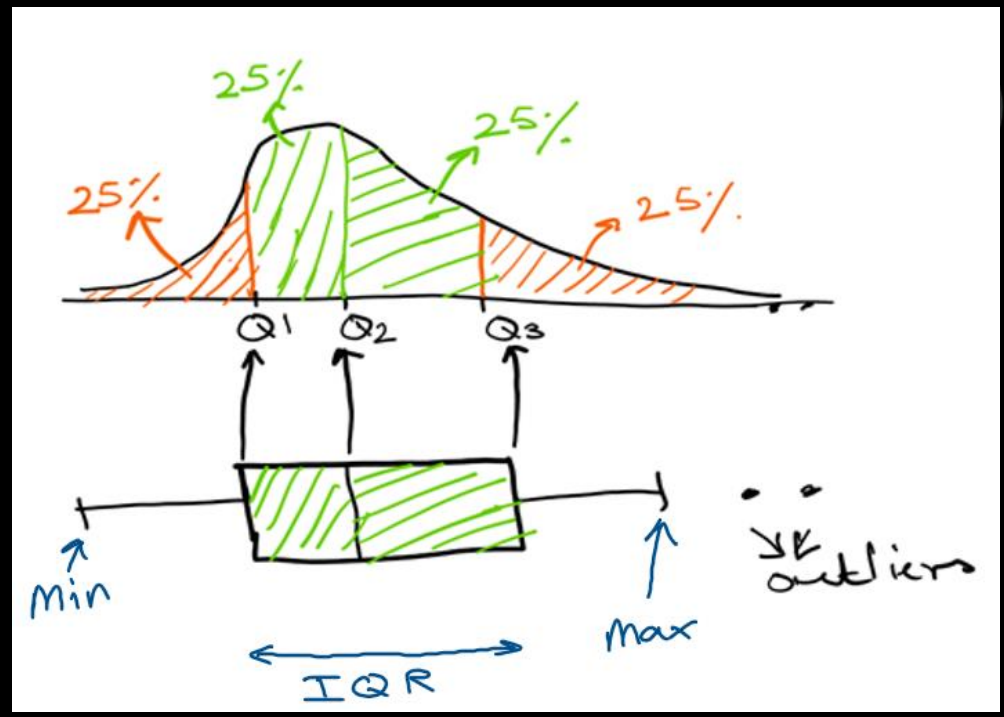


STOP

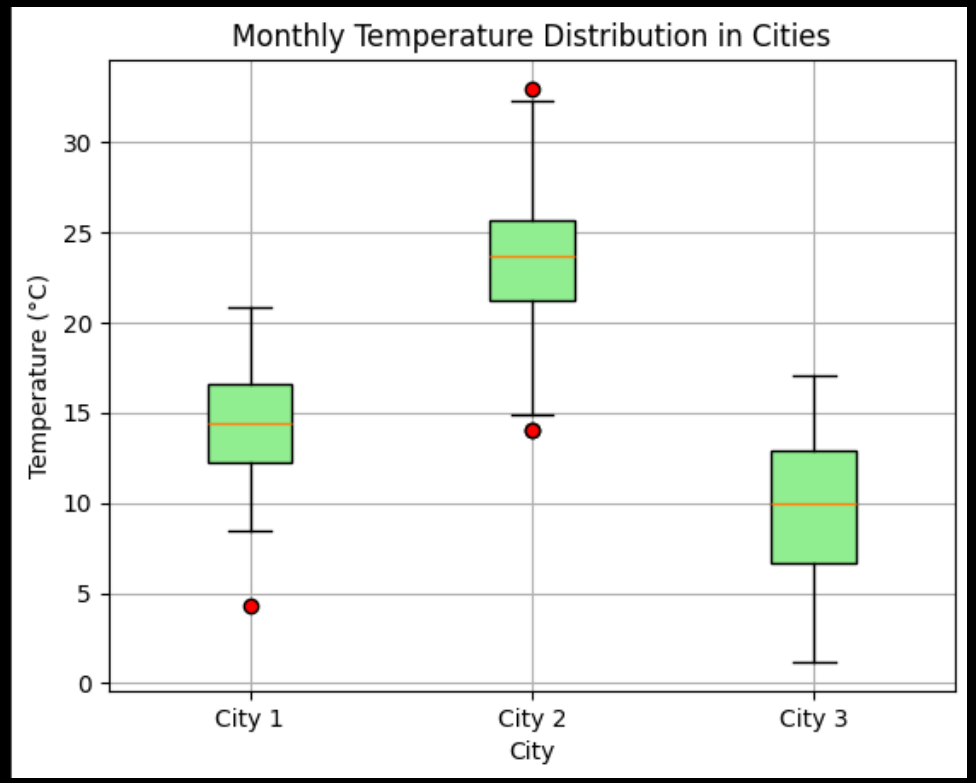
Box plots

- Show **distributions of numeric data** values, especially when you want to compare them **between multiple groups**.
- Provide visuals on **data's symmetry, skew, variance, and outliers**.

4, 3, 5, 2, 4, 3, 6, 7, 8, 3, 5, 2, 3, 4, **78**, 3, 2,-**30**, 3, 4, 5, 3, 2: here -30 and 78 seem outliers



- Easy to see where the main bulk of the data is, and make that comparison between different groups.
- 25% of data falls below Q1 (quartiles)
- 50% of data falls below Q2
- 75% of data falls below Q3



For city1:

- most of temp is between 13 to 16. There is one outlier, temp = 4
- Q1 = 13. So, 25% of temp data falls below 13.
- Q2 =14. So, 50% of temp data falls below 14
- Q3 = 17.

STOP