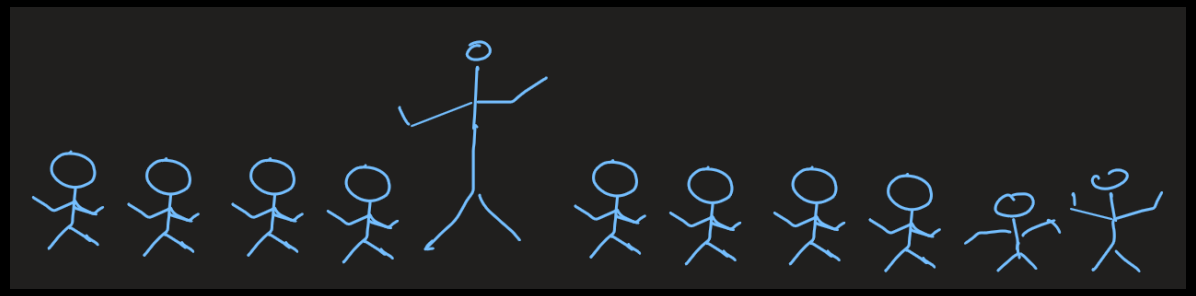
Can you tell which numbers are **statistically extreme points (AKA outliers)** ?

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




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 method

- 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 method

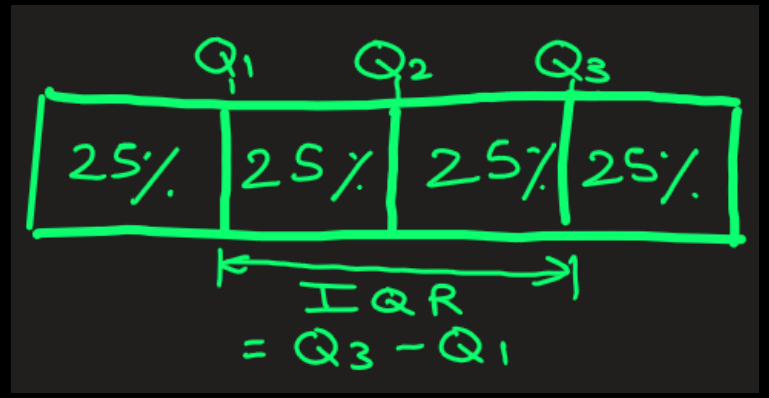
- Step2: Find Q1, Q3 and IQR

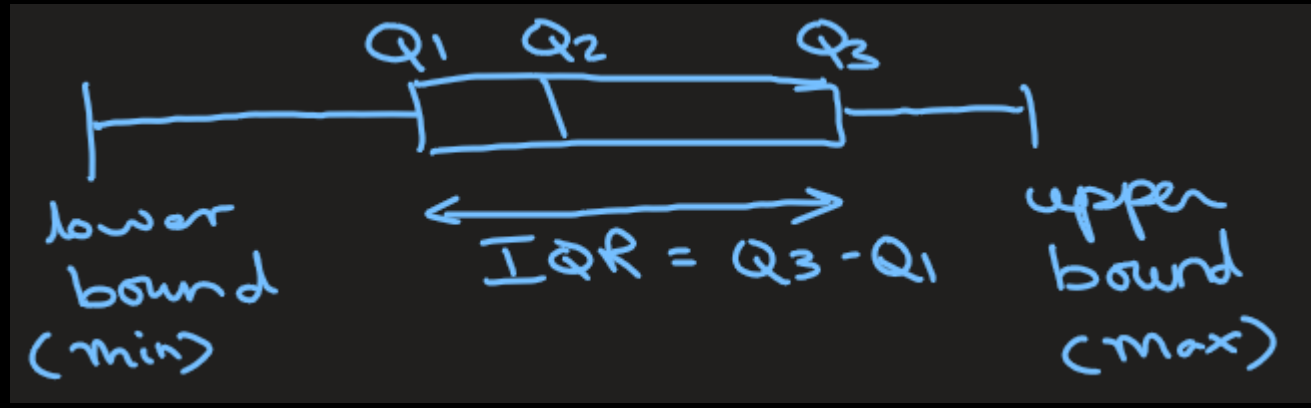
28, 31, 34, 40, 41, 41, 42, 42, 43, 44, **44**, 44, 45, 45, 45, 46, 47, 47, 47, 47, 48, 48, **48**,

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 method

- Step3: Identify Lower and upper bound

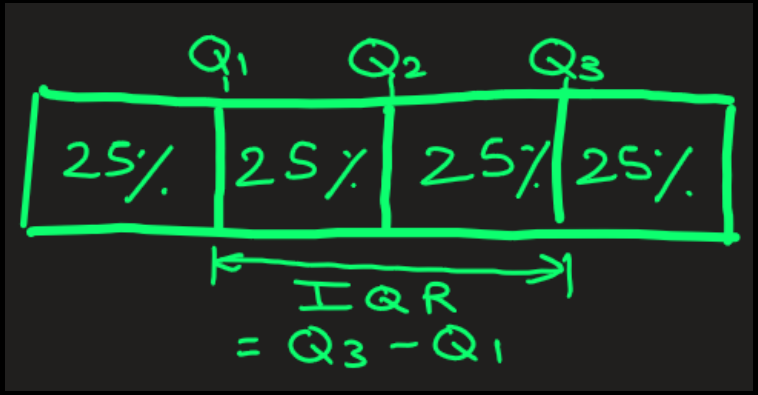
Lower bound = Q1 − 1.5 × IQR = 44 – 1.5 x 8 = 32

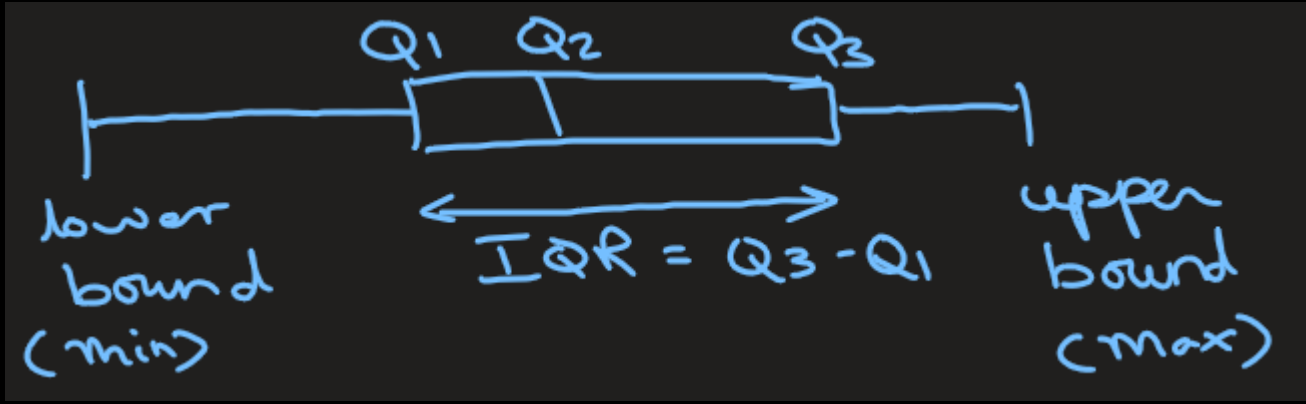
Upper bound = Q3 + 1.5 × IQR = 52 + 1.5 x 8 = 64

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

Outliers = 28, 31, 68, 74, 75. These values represent

statistically extreme points in the dataset.

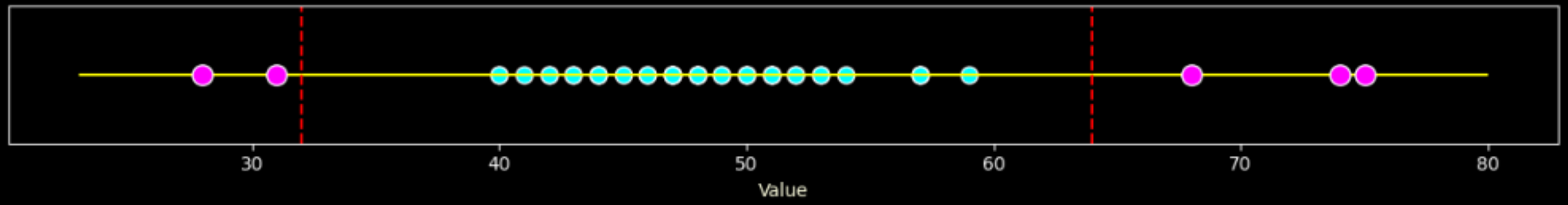




Identify outliers: Final Result

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



STOP

Identify outliers: Using z-score method

2 step process:

Step 1: Convert your data into z-score. The z-score measures how many standard deviations a data point is from the mean.

$$Z = \frac{X - \mu}{\sigma}$$

Step 2: Set up your threshold (commonly 3), and if |Z| > threshold, it's considered an outlier.

Example: Find outliers in following 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

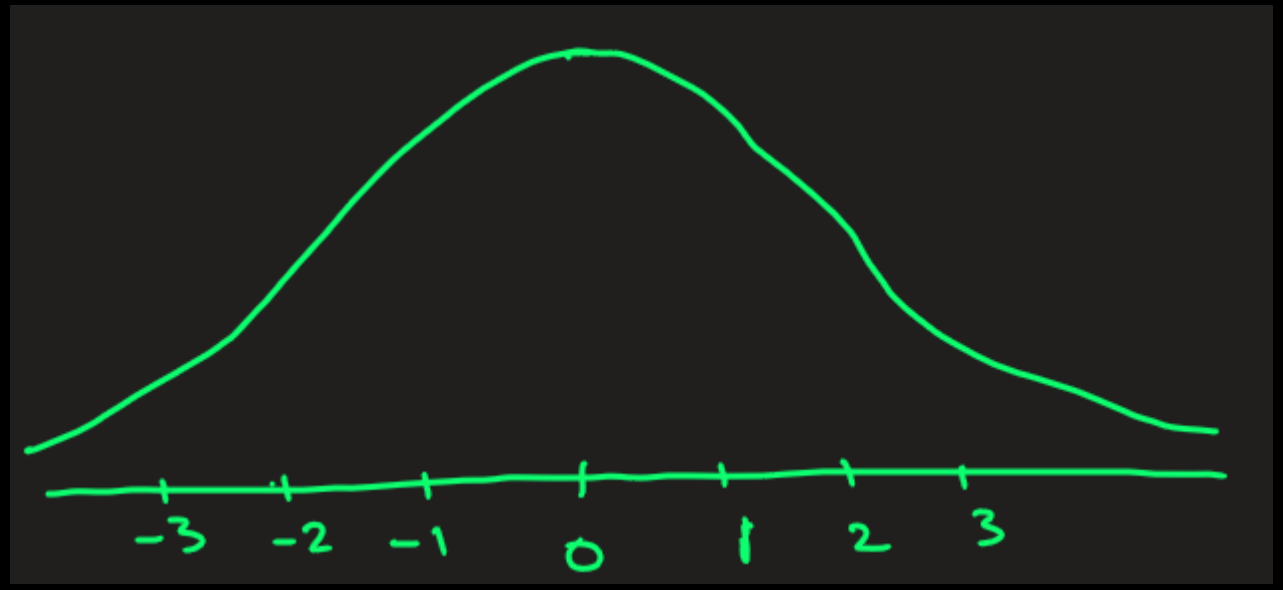
Step1: After calculation, mean μ = 48.28 and
std. dev. σ = 8.53

<u>Data</u> -> <u>z-score</u>

28 -> (28 – 48.28) / 8.53 = -2.38

31 -> (31 - 48.28) / 8.53 = -2.03

34 -> (34 - 48.28) / 8.53 = -1.67 and so on.



Identify outliers: Using z-score method

Sorted 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

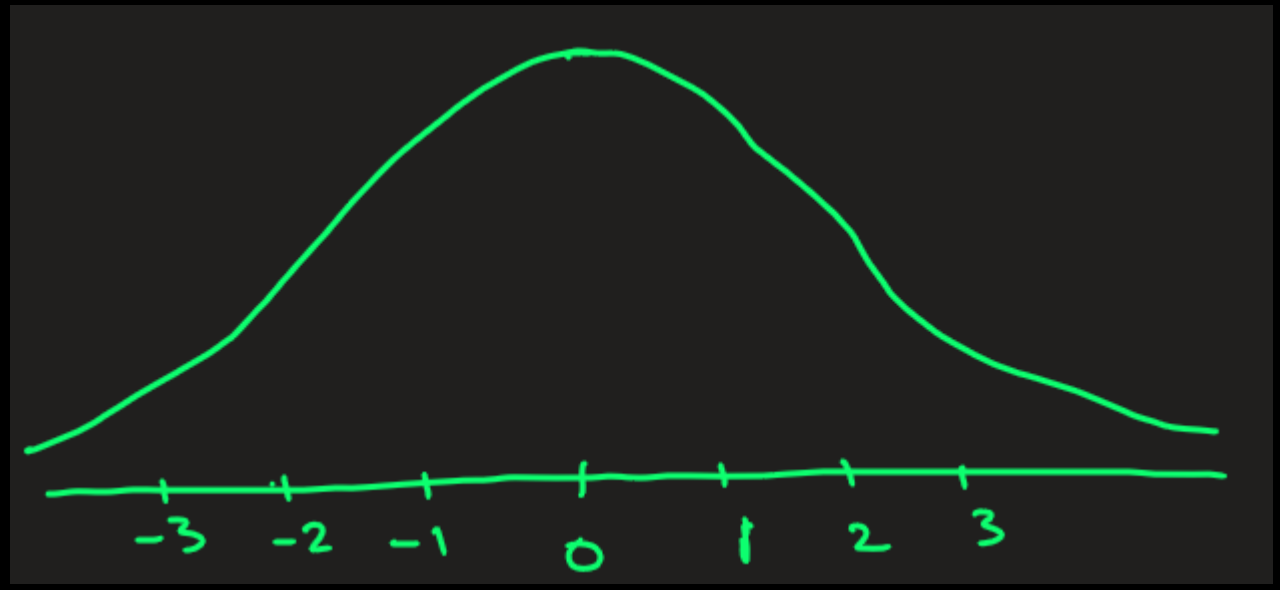
Z-scores:

-2.38, -2.03, -1.67, -0.97, -0.85, -0.85, -0.74, -0.74, -0.62, -0.5, -0.5, -0.5, -0.39, -0.39, -0.39, -0.27, -0.15, -0.15, -0.15, -0.15, -0.03, -0.03, -0.03, -0.03, -0.03, -0.03, 0.08, 0.08, 0.08, 0.2, 0.32, 0.32, 0.32, 0.43, 0.43, 0.43, 0.43, 0.43, 0.55, 0.55, 0.67, 0.79, 2.31, 3.01, 3.13

Step2: Let's define threshold = 2.3

Detected Outliers (z < **-2.3** or z > **+2.3**):

28, 68, 74, 75



STOP

IQR method or z-score method ?

IQR method is best for:

- **Non-normal (skewed)** or **small datasets** (e.g. 3, 2, 4, 1, 3, 4, 93)
- **Data with unknown distribution**
- **Ordinal or not strictly continuous data** (e.g. 3, 2, 4, 1,3, 4, 93)

Z-score method best for:

- **Normally distributed (bell-shaped)** data
- **Continuous and large datasets (**e.g. 1.12, 2.34, 3.12, 1.94, ….)

STOP

How to remove outliers ? Few basic techniques

A) Trim them: - 40, 1, 1, 2, 2, 2, 2, 2, 3, 4, 4, 5, 56, 89

1, 1, 2, 2, 2, 2, 2, 3, 4, 4, 5

If outliers are due to data entry errors or measurement mistakes, you can safely remove them.

Best for: clear, obvious outliers not representing true behavior.

B) Cap or Winsorize them: - 40, 1, 1, 2, 2, 2, 2, 2, 3, 4, 4, 5, 56, 89

1, 1, 1, 2, 2, 2, 2, 2, 3, 4, 4, 5, 5, 5

Instead of removing, you limit extreme values to a threshold.

Best for: preserving dataset size and avoiding bias.

C) Replace with mean/median (Imputation): - 40, 1, 1, 2, 2, 2, 2, 2, 3, 4, 4, 5, 56, 89

M, 1, 1, 2, 2, 2, 2, 2, 3, 4, 4, 5, M, M

You can replace extreme values with more **typical** ones.

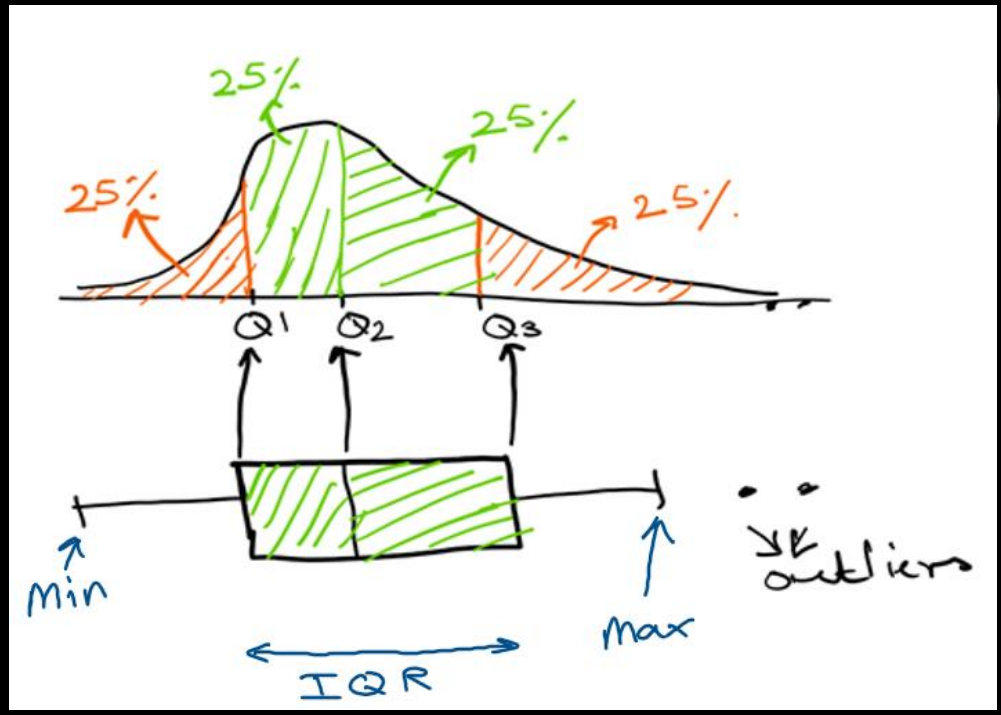
Best for: small datasets where every record matters.

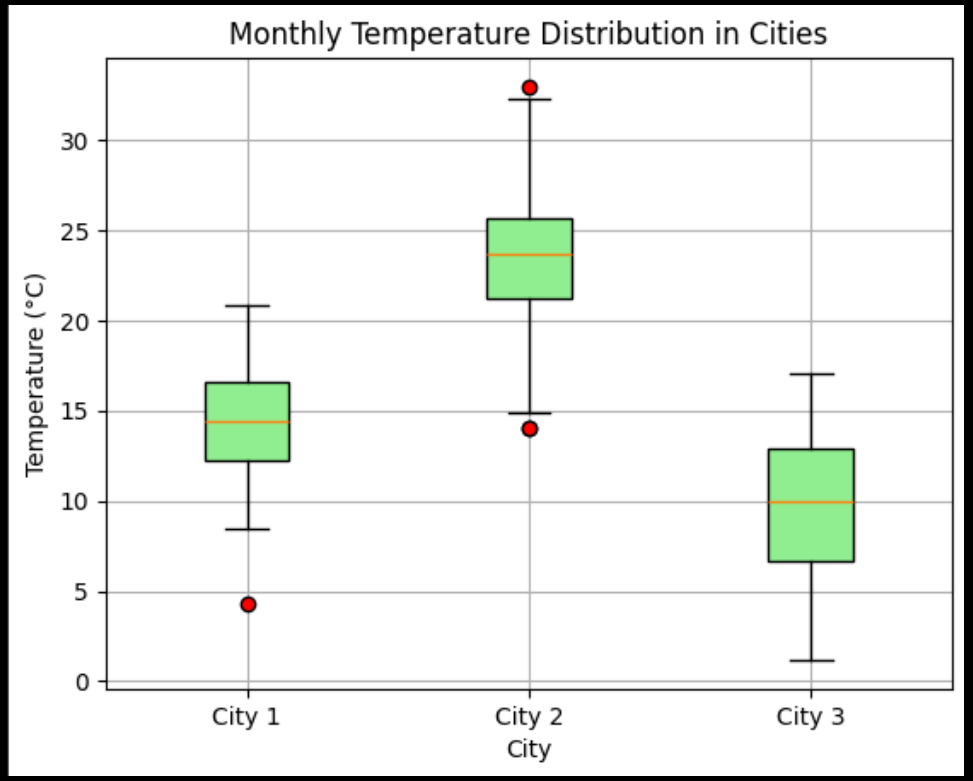
Box plots

- 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