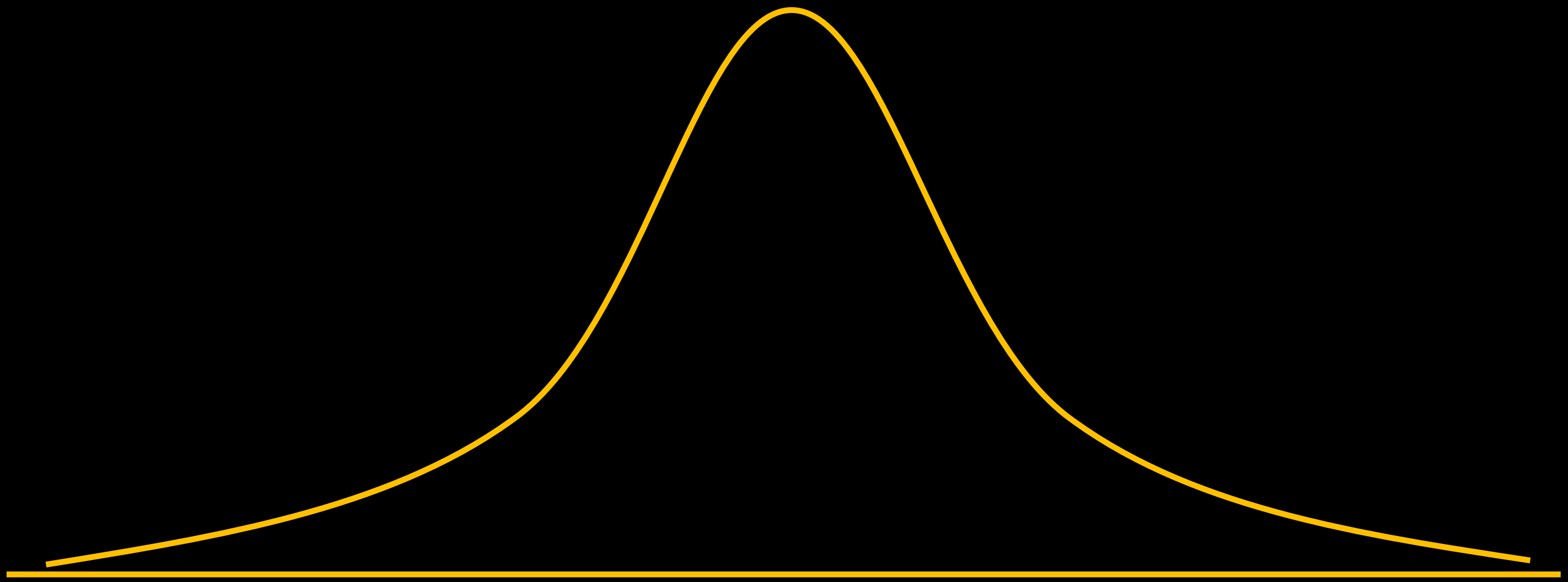
Identify Outliers: Using Standard Deviation

Method best for:

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



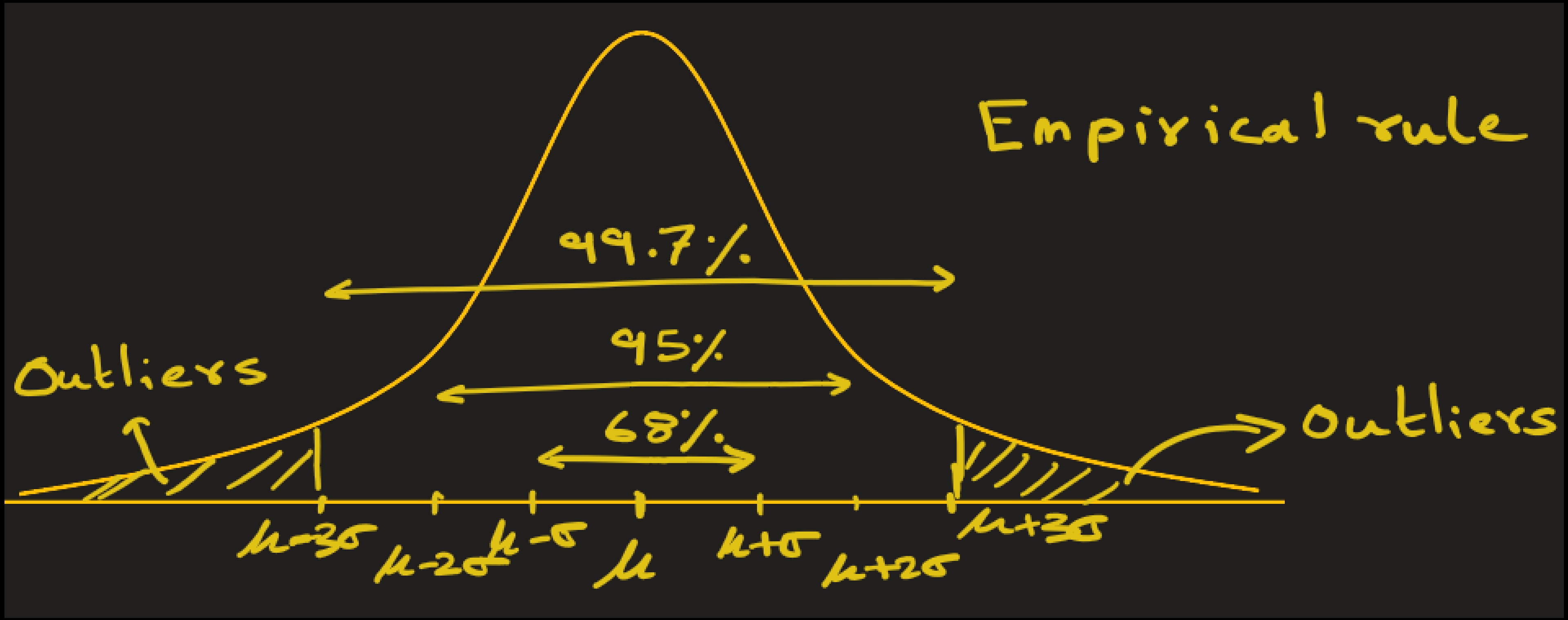
Identify Outliers: Using Standard Deviation

Idea:

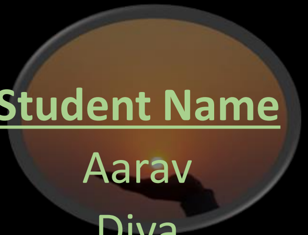
Since most of data should lie around the mean value, so, any data **far away from mean** is an outlier.

These data points lie towards extreme left or right of distribution curve.

Here you typically choose standard deviation threshold of 3. If a data is outside the threshold, that data is an outlier.



Identify Outliers: Using Standard Deviation

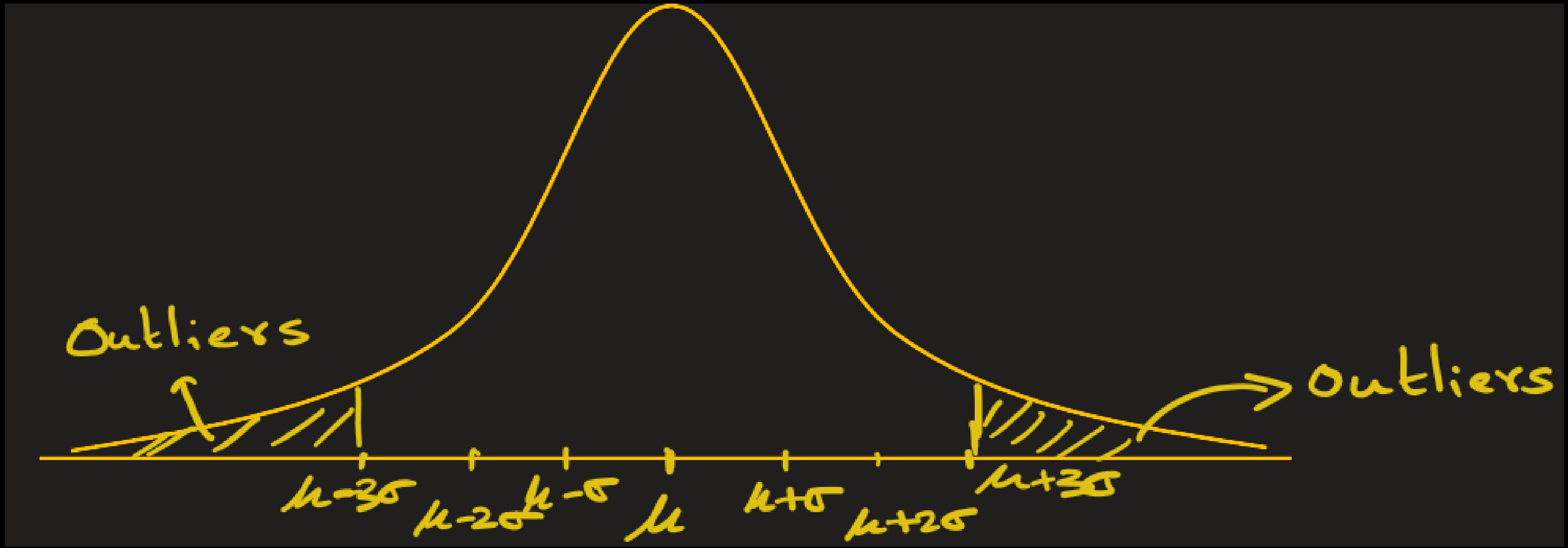


Student Name	Score/age/income
Aarav	52
Diya	48
Rohan	45
Ananya	44
Kabir	49
Isha	55
Vivaan	40
Meera	46
Arjun	44
Neha	53
Aditya	47
Pooja	47
Rahul	48
Sneha	52
Kunal	48
Priya	34
Siddharth	52
Nisha	52
Manish	48
Kavya	44
Amit	51
Ritu	51
Varun	47
Shreya	48
Nikhil	45
Tanvi	49
Suresh	42
Pallavi	52
Mohit	45
Ayesha	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 Z-score

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

Here we find mean μ = 48.28 and standard deviation σ = 8.53
If threshold is 3σ, then values outside the range of (μ - 3σ, μ + 3σ) are outliers
-> Values outside of (48.28 - 3x8.53, 48.28 + 3x8.53) = (22.69, 73.87) are 74 and 75

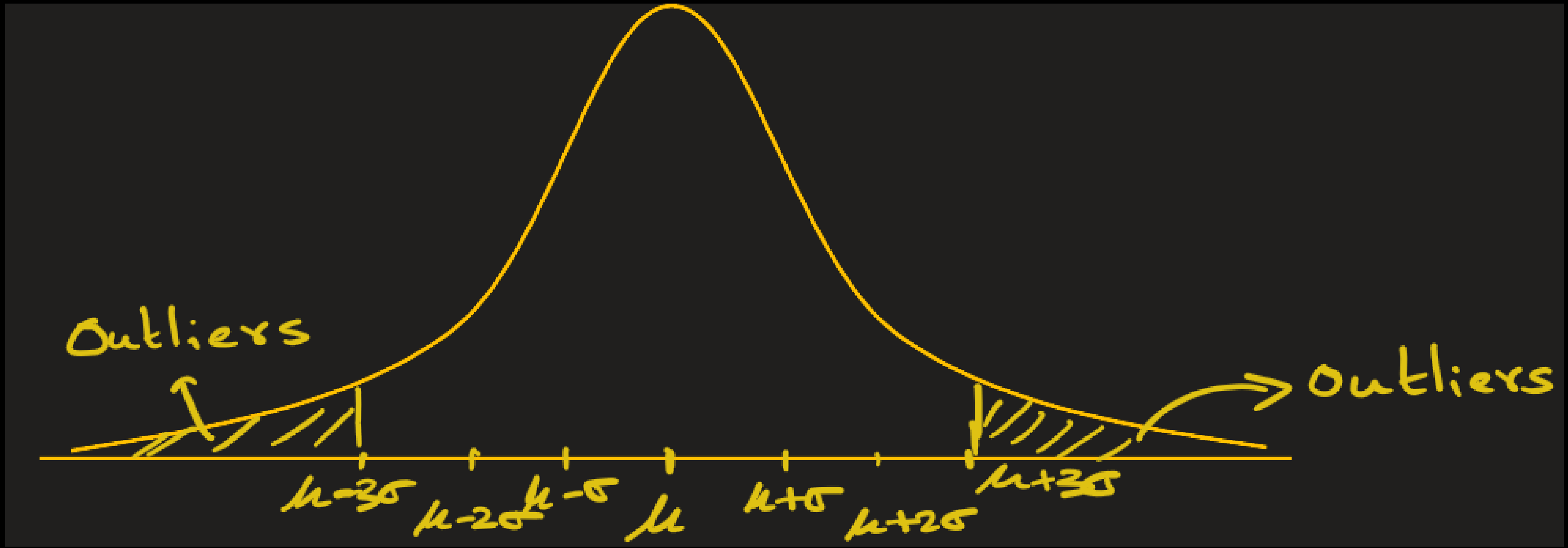


Identify Outliers: Using Z-score

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

If threshold is 2.3σ, then values outside the range of (μ - 2.3σ, μ + 2.3σ) are outliers

-> Values outside of (48.28 - 2.3x8.53, 48.28 + 2.3x8.53) = (28.66, 67.89) are 28, 31, 68, 74, 75



Identify Outliers: Using Z-score

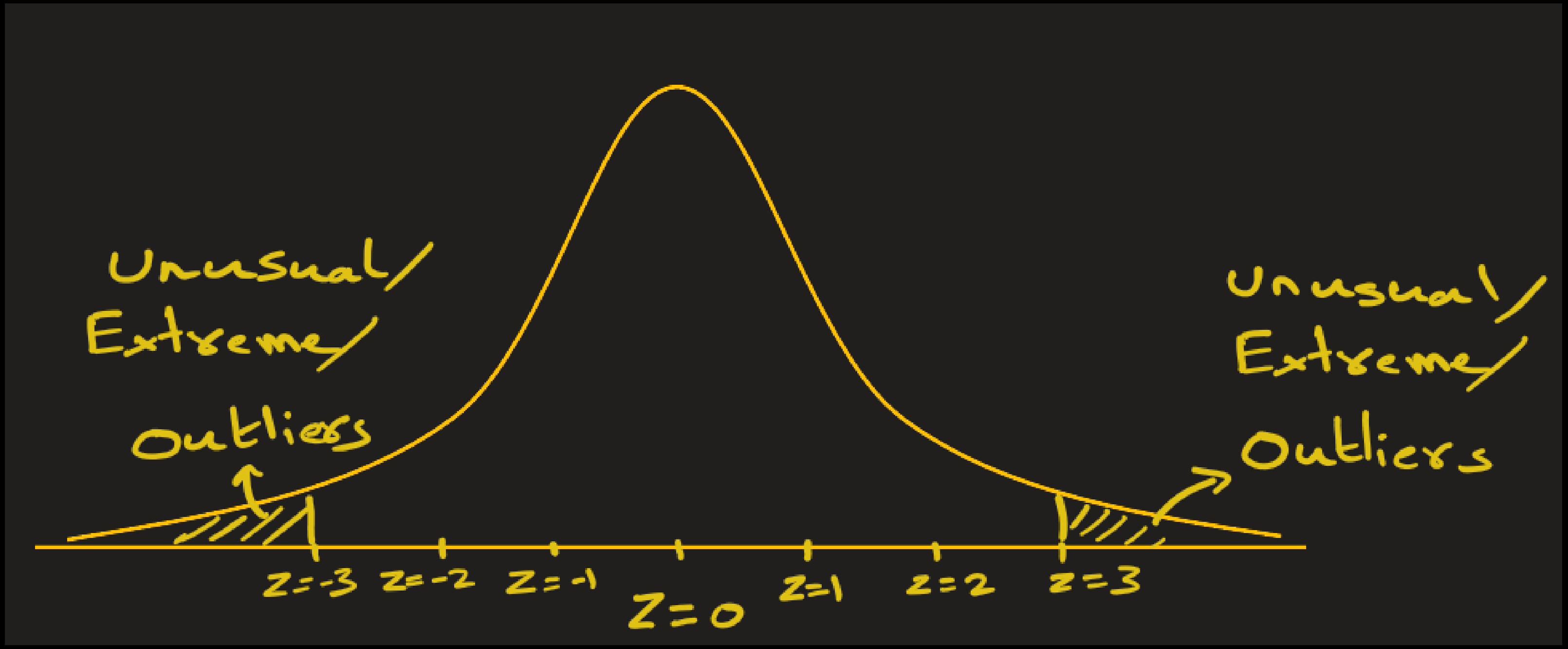
$$Z = \frac{X - \mu}{\sigma}$$

2 Step Process

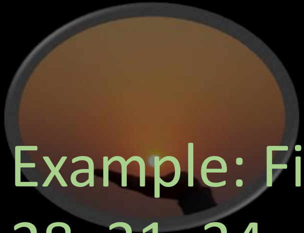
Step 1: Convert your data into z-score.

The z-score measures how many standard deviations a data point is from the mean.

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



Identify Outliers: Using Z-score



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: Convert your data into z-score.

Here we find mean μ = 48.28 and standard deviation σ = 8.53

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

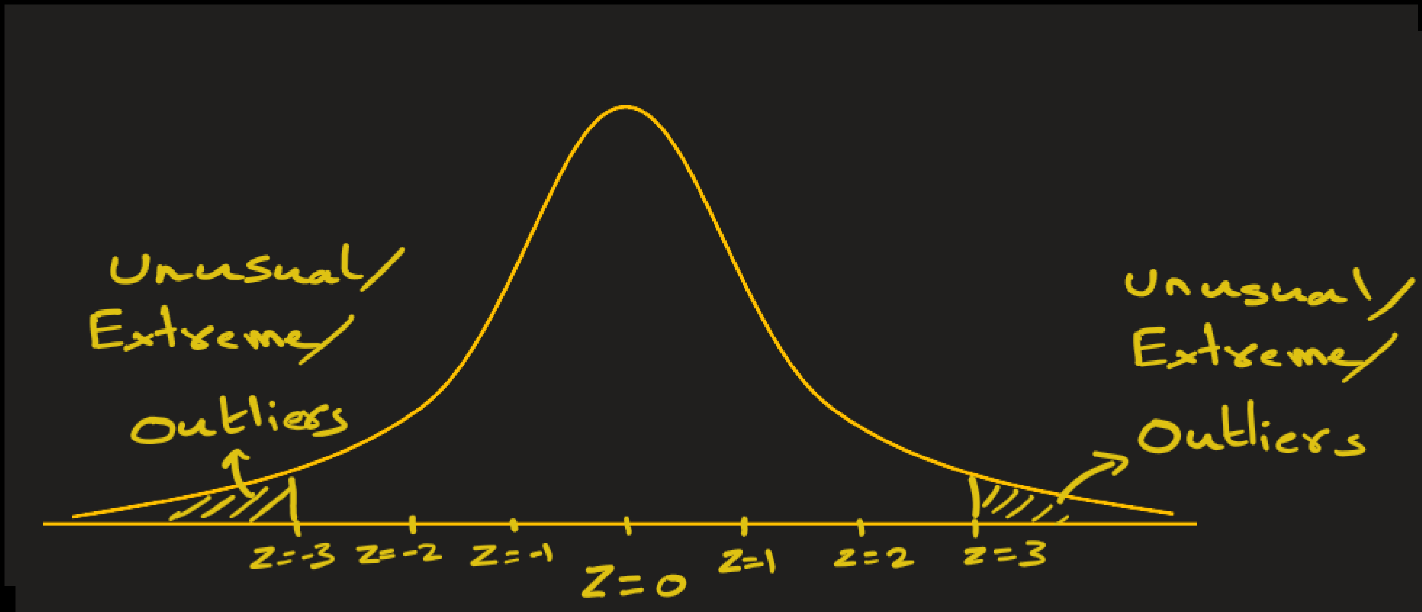
<u>Data</u> -> <u>z-score</u>

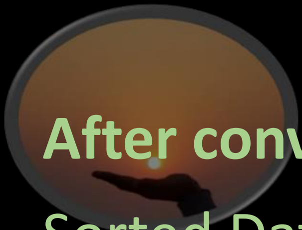
28 -> (28 – 48.28) / 8.53 = -2.38

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

34 -> (34 - 48.28) / 8.53 = -1.67

and so on.





Identify Outliers: Using Z-score

After conversion

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:

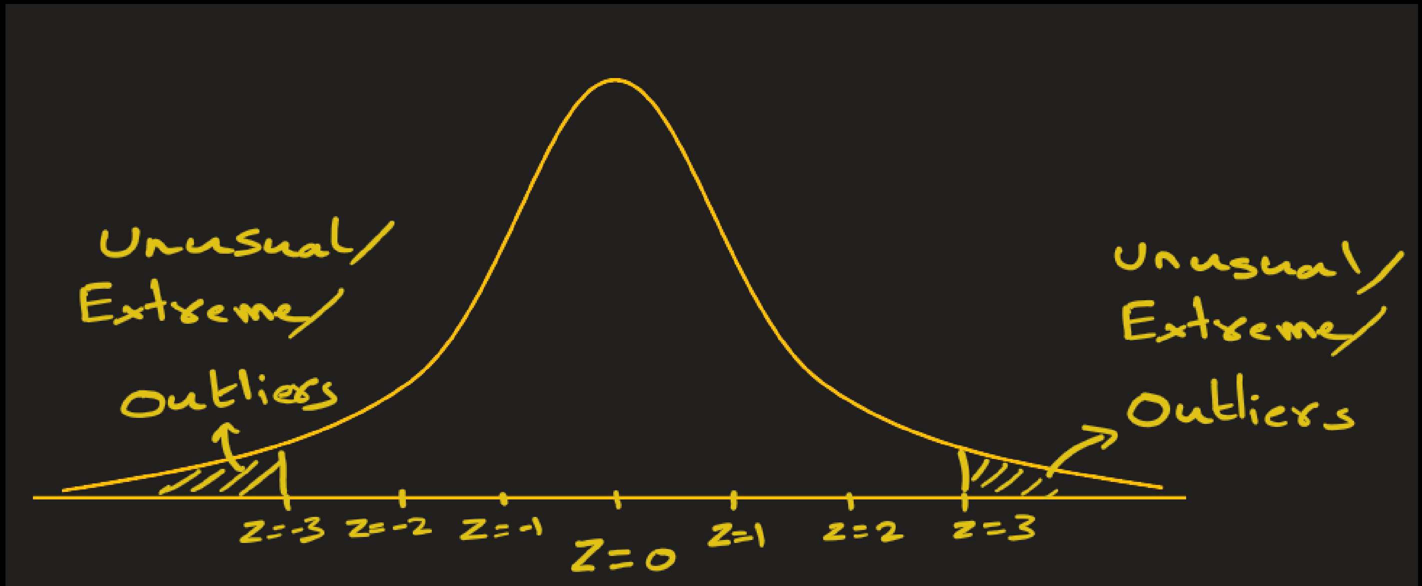
-2.38, -2.03, -1.67, -0.97, -0.85, -0.85, -0.74, -0.74, -0.62, -0.5, -0.5, -0.5, -0.39, -0.39, -0.39, -0.27, -0.15, -0.15, -0.15, -0.15, -0.03, -0.03, -0.03, -0.03, -0.03, -0.03, 0.08, 0.08, 0.08, 0.2, 0.32, 0.32, 0.32, 0.43, 0.43, 0.43, 0.43, 0.43, 0.55, 0.55, 0.67, 0.79, 2.31, 3.01, 3.13

Step2: Let's define threshold = 3

Detected Outliers (z < **-3** or z > **+3**):

3.01 -> 74

3.13 -> 75



Identify Outliers: Using Z-score

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

If I change threshold = 2.3

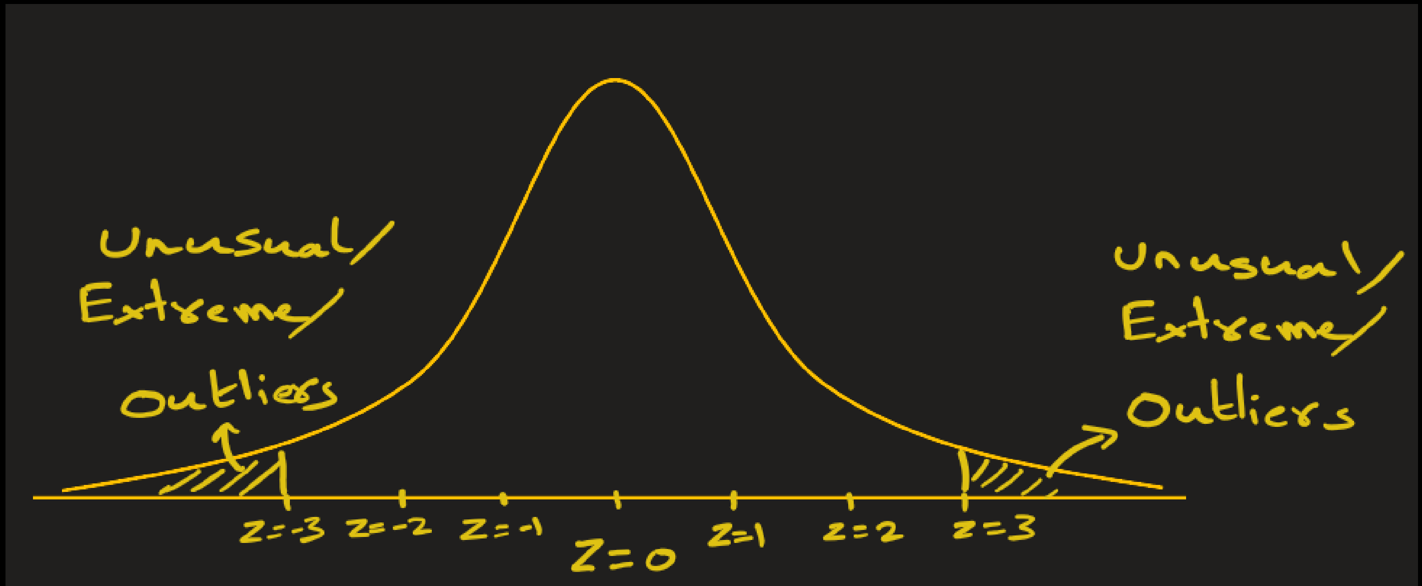
Detected Outliers (z < **-2.3** or z > **+2.3**):

-2.38 -> 28

2.31 -> 68

3.01 -> 74

3.13 -> 75



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, ….)

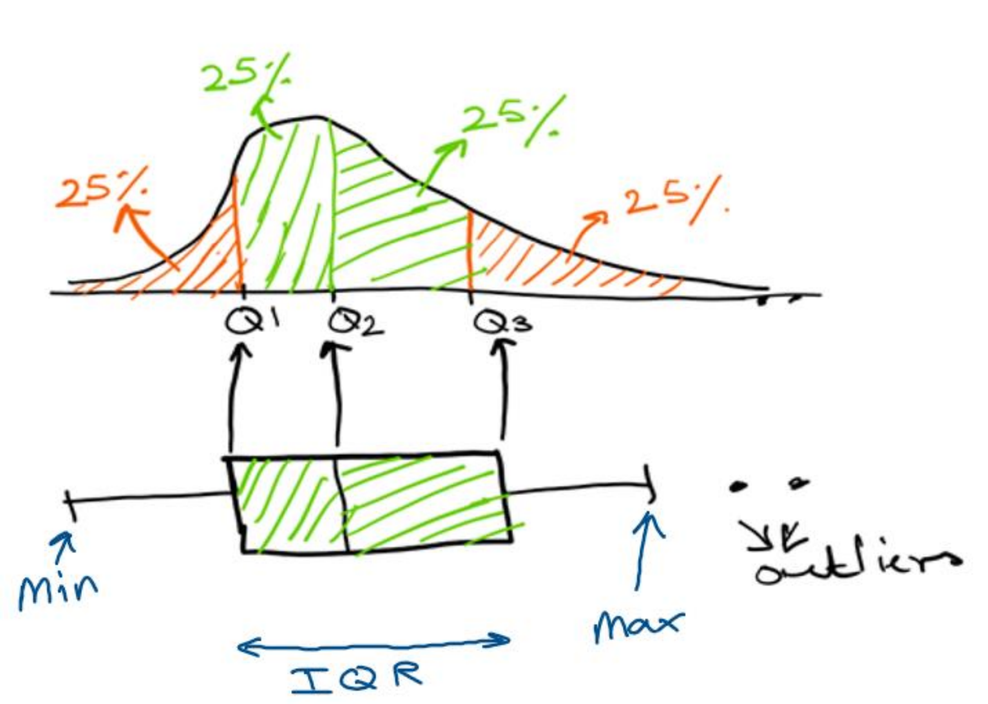
STOP

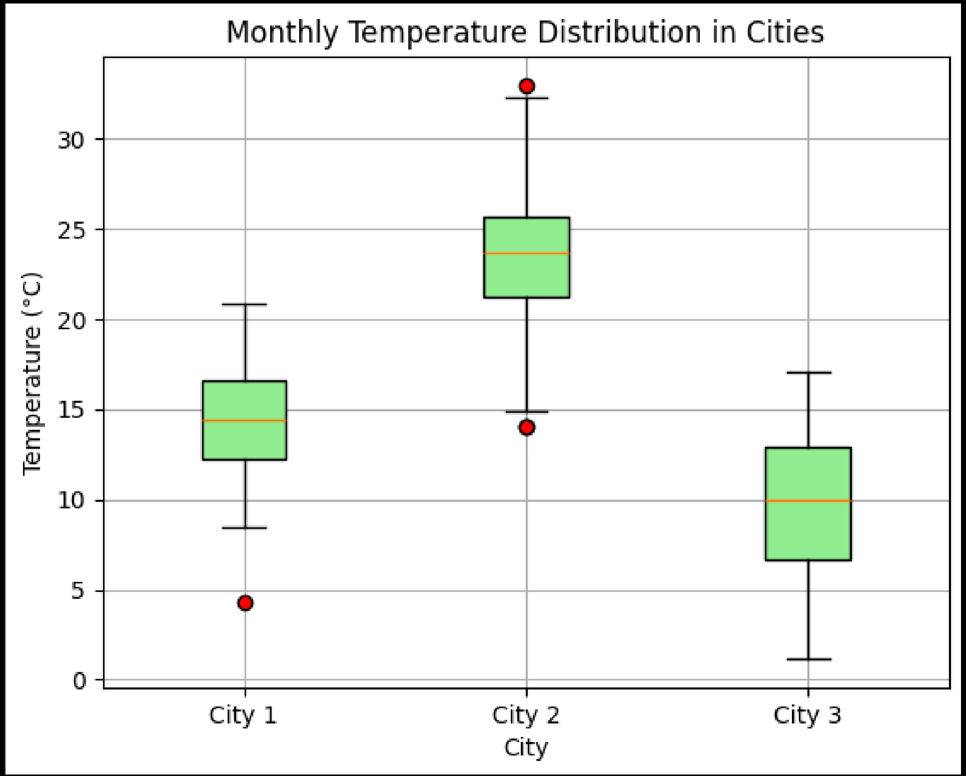
Box plots

- 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.