MEASURES OF DISPERSION
Range, IQR

- Arsh

WHAT IS DISPERSION ?

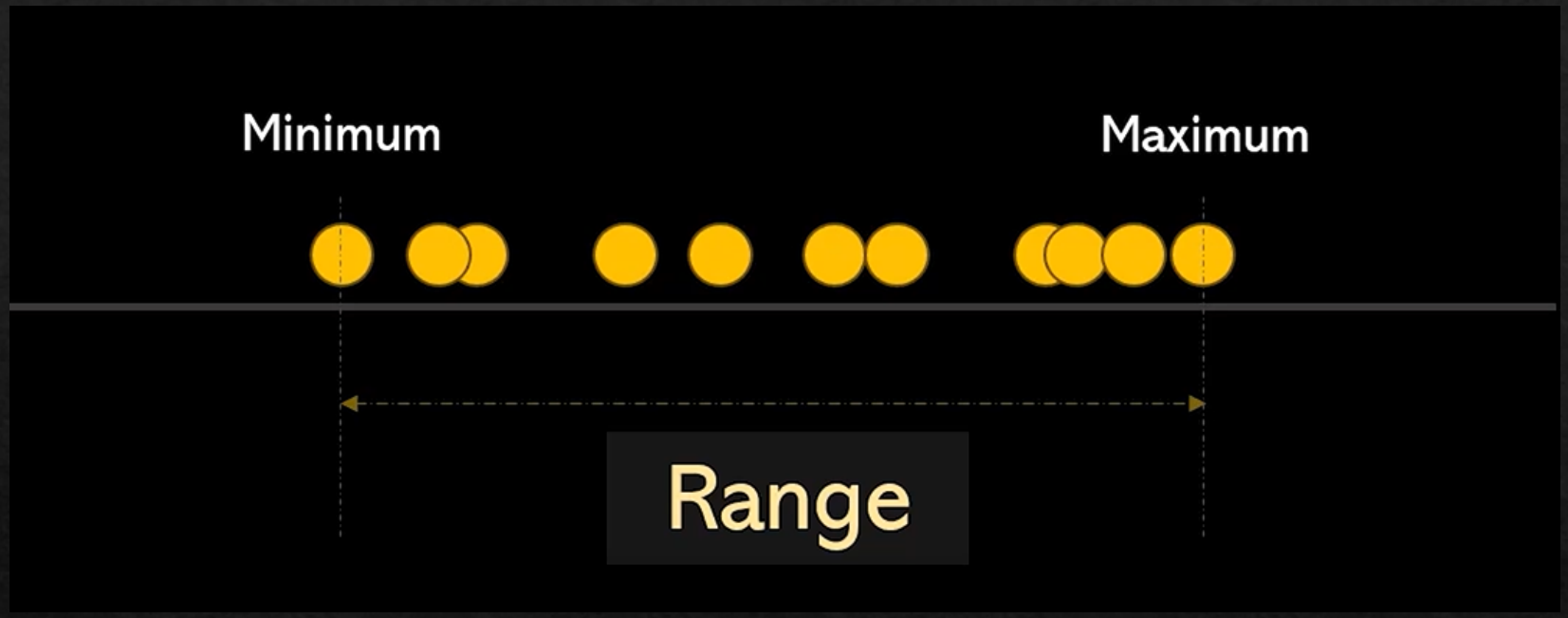
Dispersion in statistics refers to the measure of how spread out or varied the data is from the mean or median value. It gives an idea of how much the individual data points differ from each other.

STUDENT	SCORE
1	80
2	70
3	90
4	60
5	85
6	75
7	95
8	65
9	88
10	72

The mean score is 78.4. However, the scores are spread out, with some students scoring much higher or lower than the mean. This indicates a high level of dispersion in the data.

EMPLOYEE	SALARY
1	50K
2	60K
3	70K
4	80K
5	90K

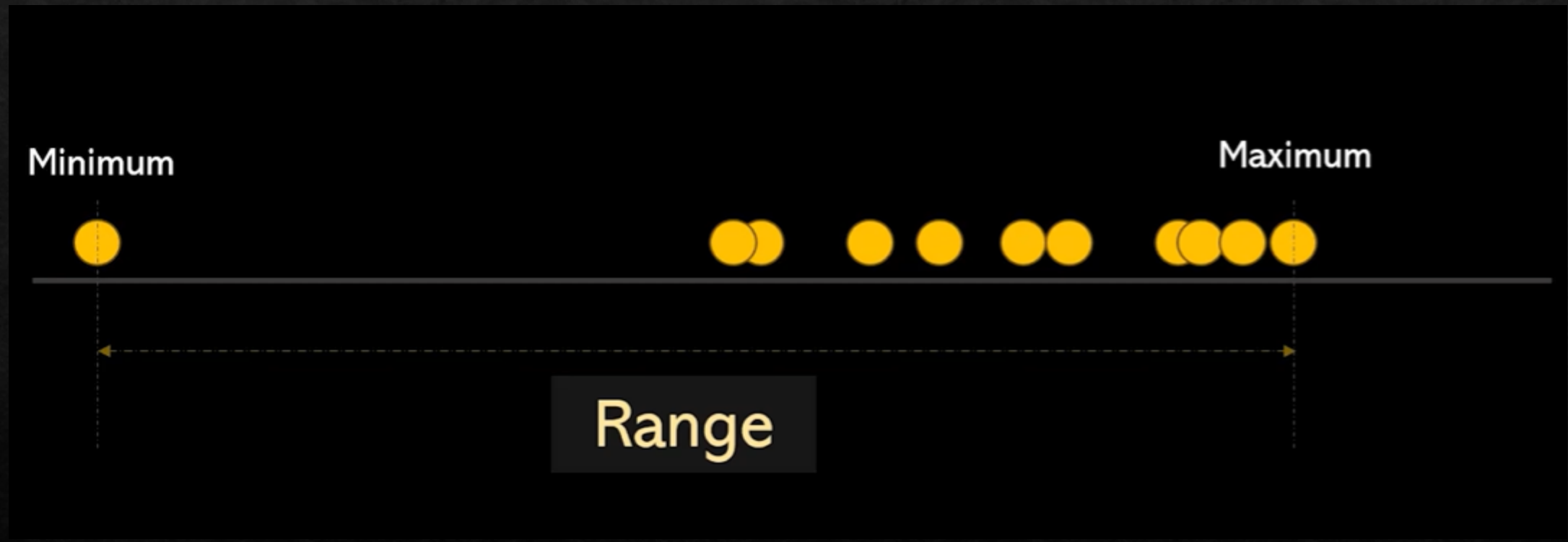
The mean salary is 65000. However, the salaries are all very close to each other, with only a small range of values. This indicates a low level of dispersion in the data.



RANGE

The range is the difference between the maximum and minimum values in the data. It reflects the data spread.

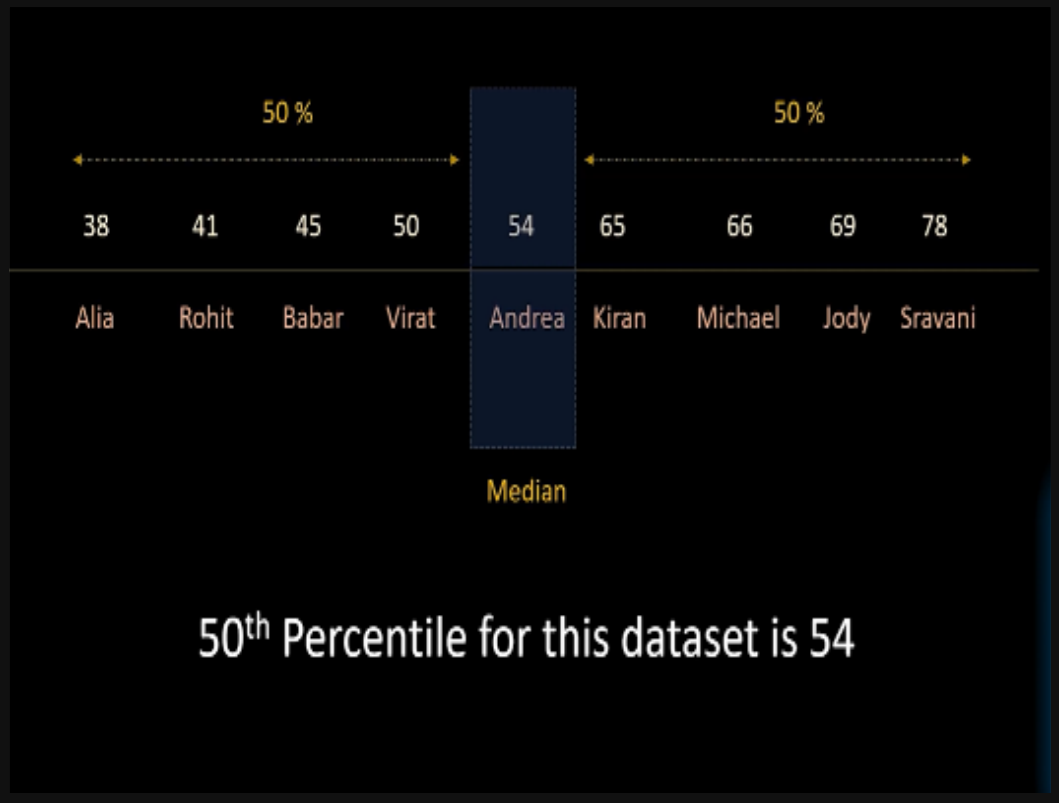
SENSITIVE TO OUTLIERS….



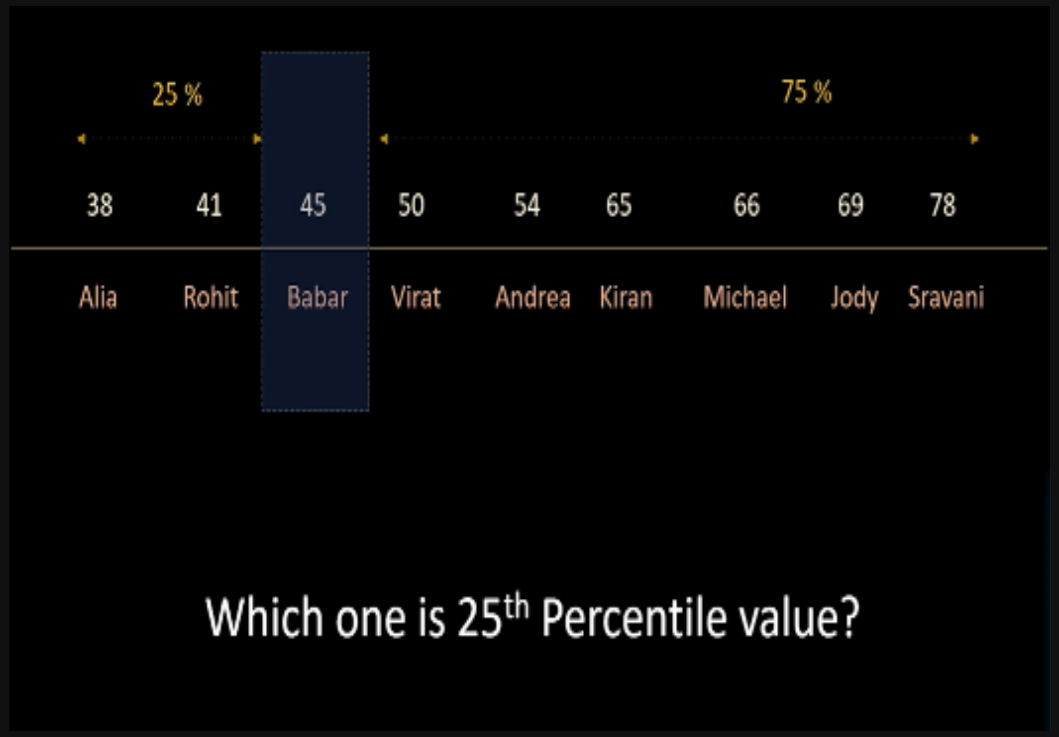
IQR – Inter Quartile Range

Unlike Range, IQR (Inter Quartile Range) is less influenced by outliers making it a robust measure.

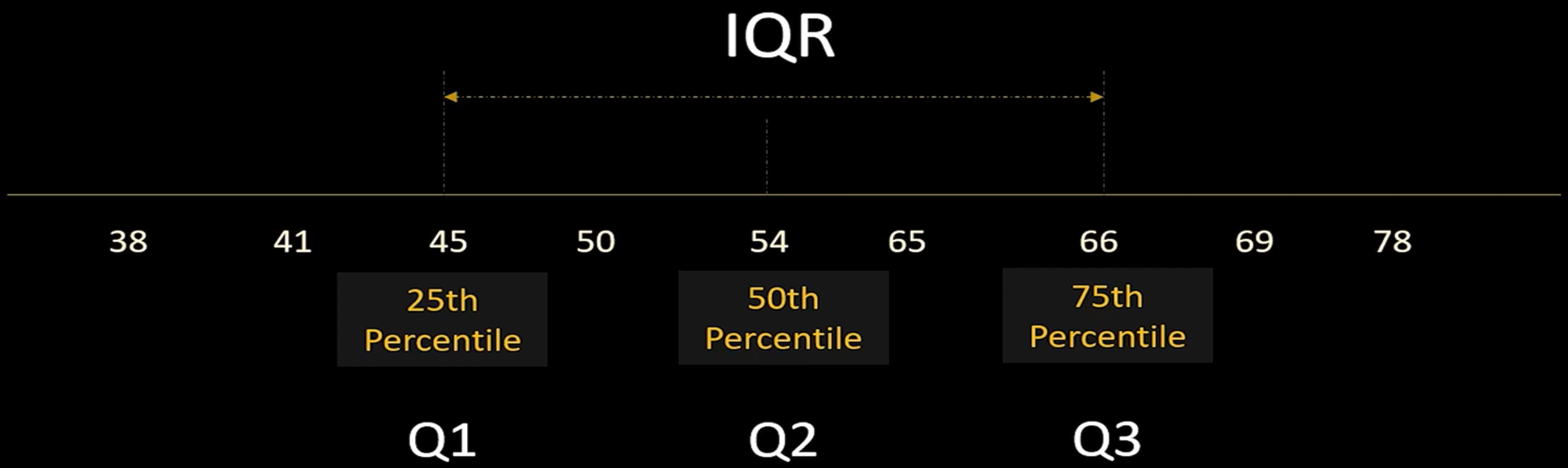
The IQR is the difference between the 75^{th} percentile (Q3) and the 25^{th} percentile (Q1) in the data, showing the spread of middle 50 % of the data.



50th Percentile for this dataset is 54



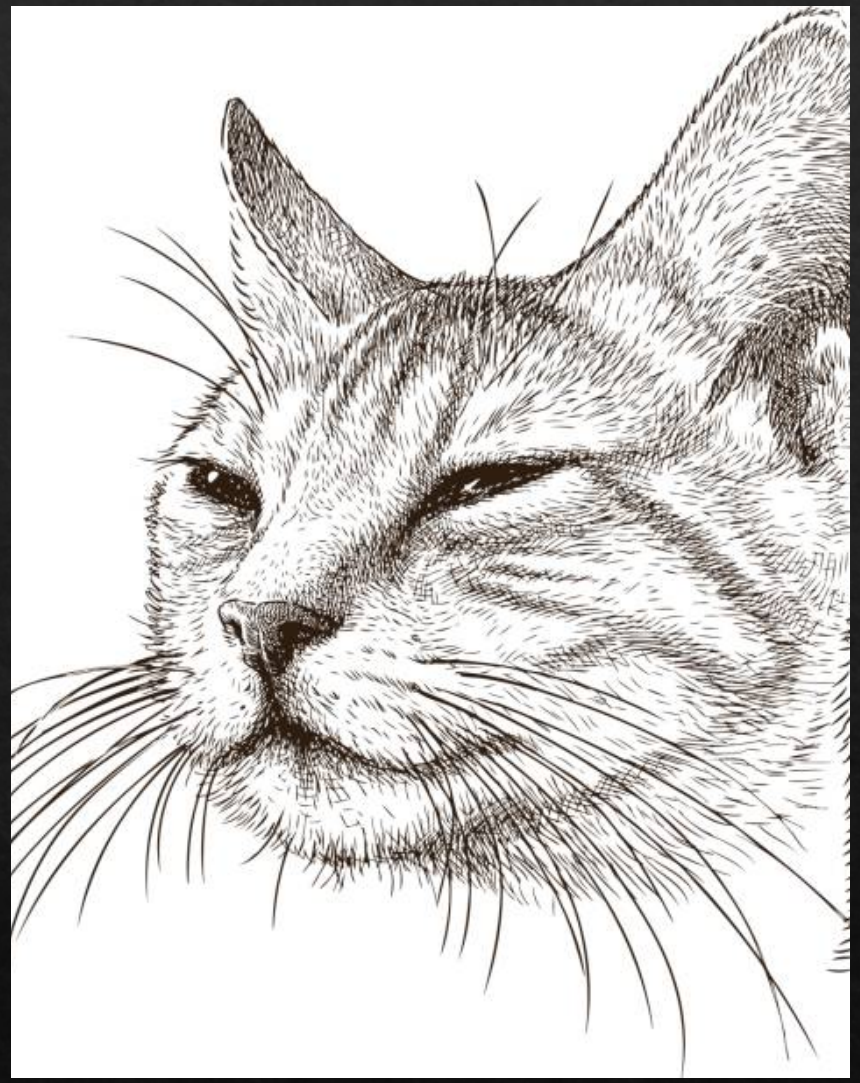
Which one is 25th Percentile value?



$$IQR = Q3 - Q1$$
$$= 66 - 45$$
$$= 21$$

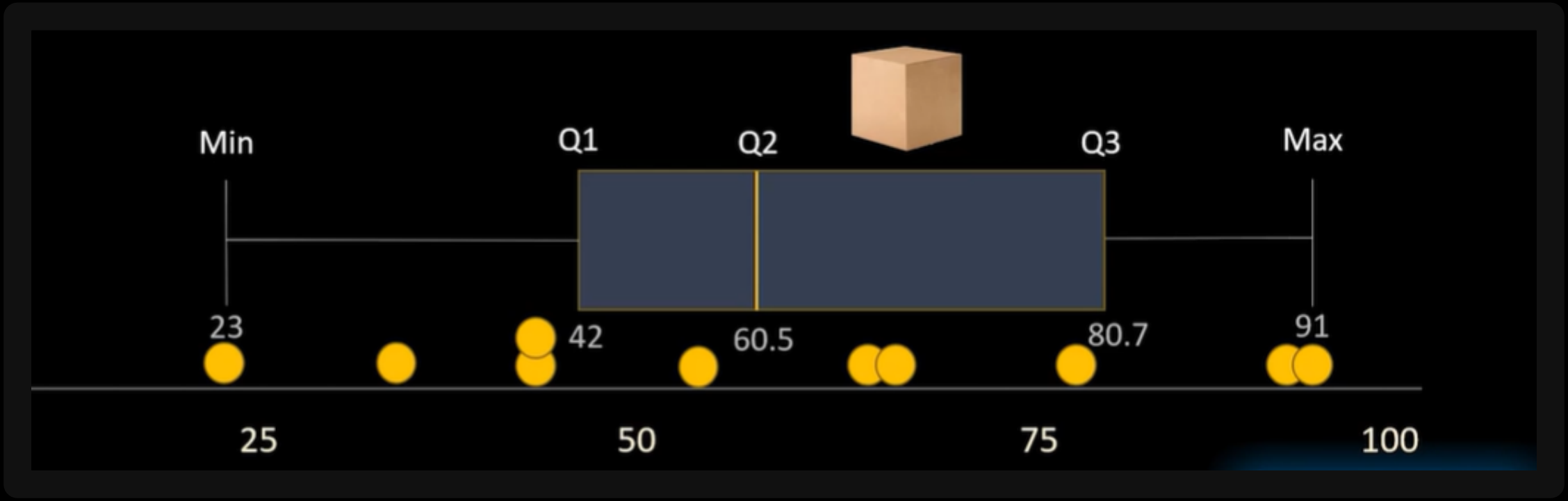
BOX OR WHISKER PLOT

A Box Plot or a Whisker Plot provides a visual summary of the central tendency (mean, median, mode), spread and presence of outliers in a dataset.

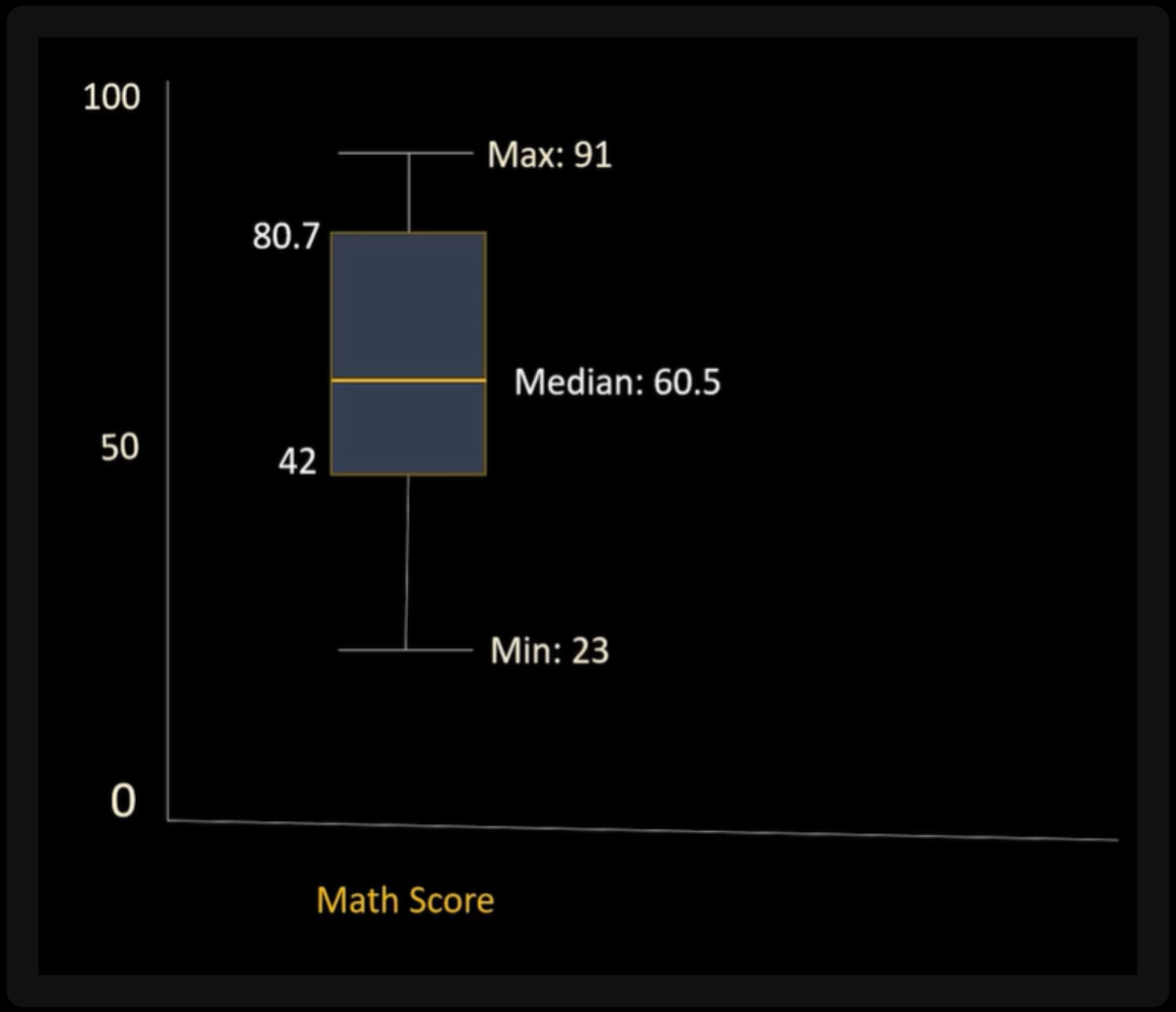



Min
Q1
Q2
Q3
Max
23
42
60.5
80.7
91
25
50
75
100

The ‘box’ in the box plot shows the 50% of the data, with the line inside representing the median



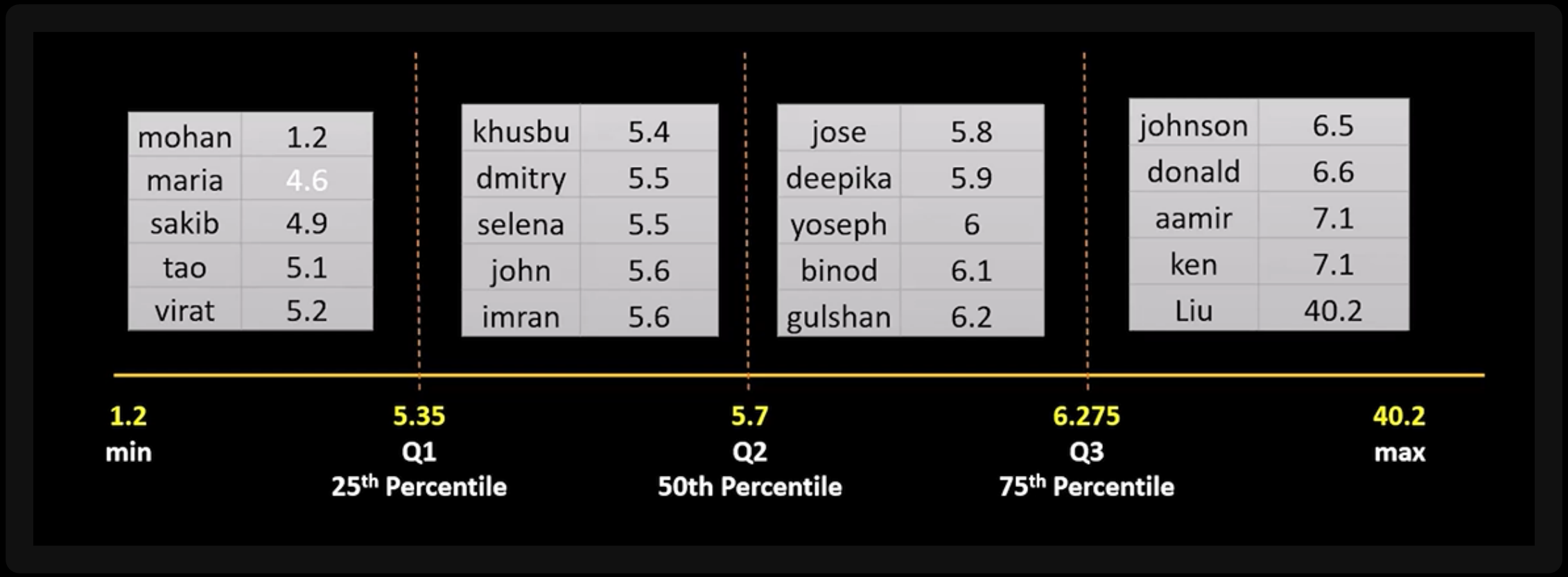
The ‘whiskers’ in box plot shows the minimum and maximum values with a specific range


100
50
0
Max: 91
80.7
Median: 60.5
42
Min: 23
Math Score

OUTLIER TREATMENT USING IQR AND BOX PLOT

Outliers can greatly effect the statistical measures such as Mean, Median and Standard Deviation. This can lead to inaccurate conclusions. It can also affect the performance of machine learning models and statistical models.

name	height
mohan	1.2
maria	4.6
sakib	4.9
tao	5.1
virat	5.2
khusbu	5.4
dmitry	5.5
selena	5.5
john	5.6
imran	5.6
jose	5.8
deepika	5.9
yoseph	6
binod	6.1
gulshan	6.2
johnson	6.5
donald	6.6
aamir	7.1
ken	7.1
Liu	40.2



IQR = Q3 – Q1
IQR = 6.275 – 5.35
IQR = 0.925

Lower_Limit = Q1 – 1.5*IQR = 3.96

Upper_Limit = Q3 + 1.5*IQR = 7.66

WHY DO WE USE 25^{TH} AND 75^{TH} PERCENTILE

1. They capture the middle 50% of the data,
2. They reduce sensitivity to extreme values while keeping most valid data.
3. Using 10^{th} and 90^{th} percentiles would remove too many data points

WHY MULTIPLY BY 1.5 ?

1.5 × IQR → The standard rule, used in most cases.
1.0 × IQR → More aggressive, detects more points as outliers.
3.0 × IQR → Only flags extreme outliers.