Outliers

Iron → Upinder ~~Doctor entry~~

0.53

53

What is outlier?

Spam calls → No spam call → Offence

A → ATM → cash

₹500 → Daily

₹5000 → Single incident

Out of Pattern/trend

An *outlier* is an observation that lies an abnormal distance from other values in a random sample from a population. In a sense, this definition leaves it up to the analyst to decide what will be considered abnormal.

1 → It deviates from the general pattern

2 ⇒ fraud/opportunity.

Friend → Daily → Meet.

med gun robbery

police.. Fir

Common Causes of Outliers

Dey Pok

Cen..

A mult?..

10 20

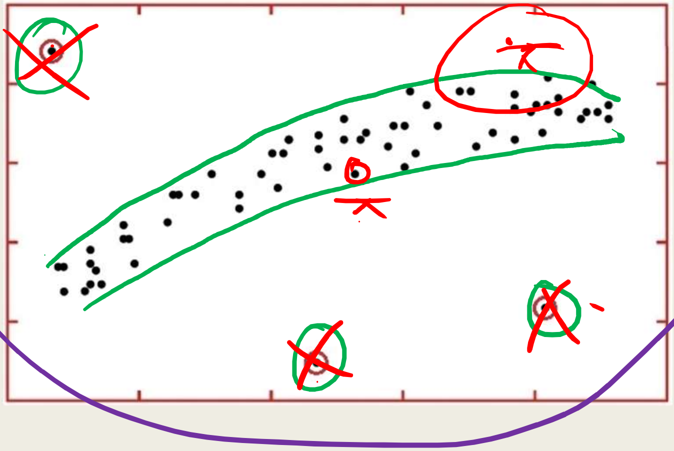
1. Data entry errors (human errors) 106.4
2. Measurement errors (instrument errors) Indi
3. Experimental errors (data extraction or experiment planning/executing errors) blue run..
4. Intentional (dummy outliers made to test detection methods)
5. Data processing errors (data manipulation or data set unintended mutations) Income → e^x → Inflation
6. Sampling errors (extracting or mixing data from wrong or various sources)
7. Natural (not an error, novelties in data)

Income

Afra Fan Safety Note Jour nidhi

£ slu

Extreme values



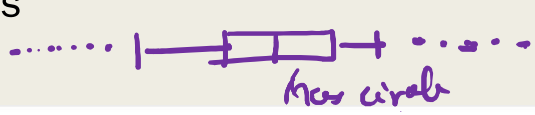
Common methods of determining an Outlier

500, 500, 500, 500, 5000

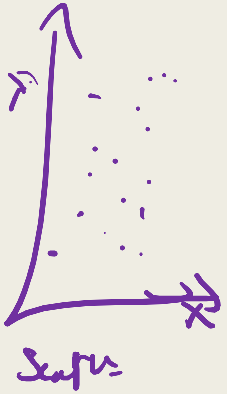
1. Sort the data and see for the extreme values
2. Plotting – Boxplot, Scatterplot
3. IQR Method

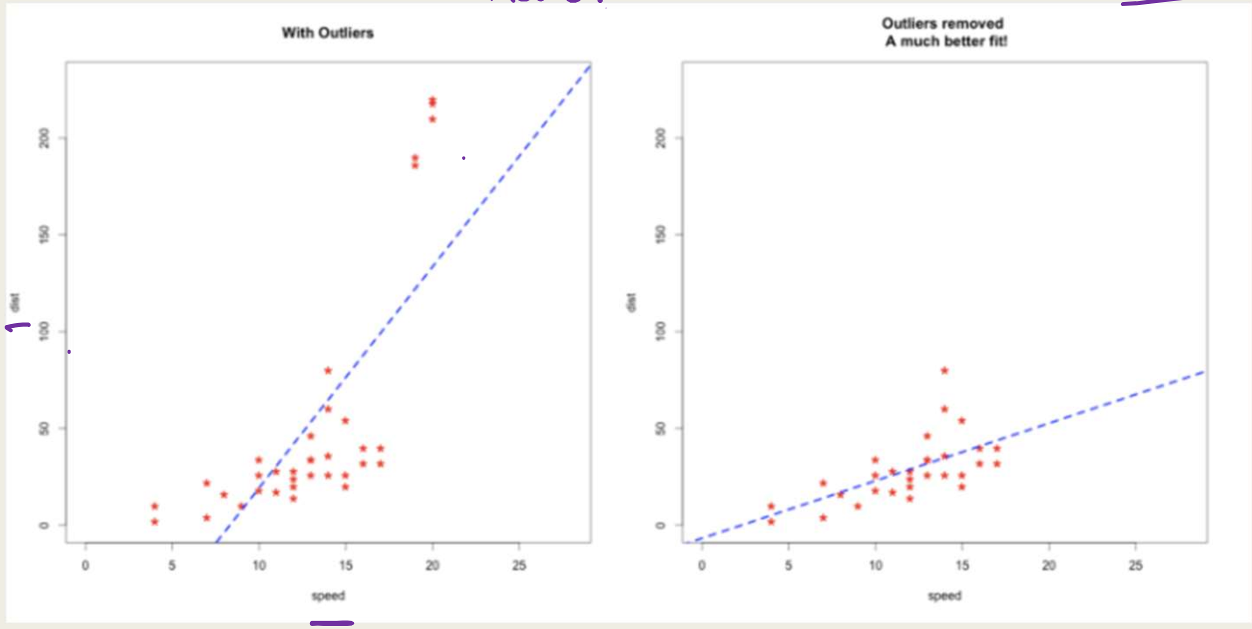
4. - Z - Score Method

N.D.


Has circle

No circle

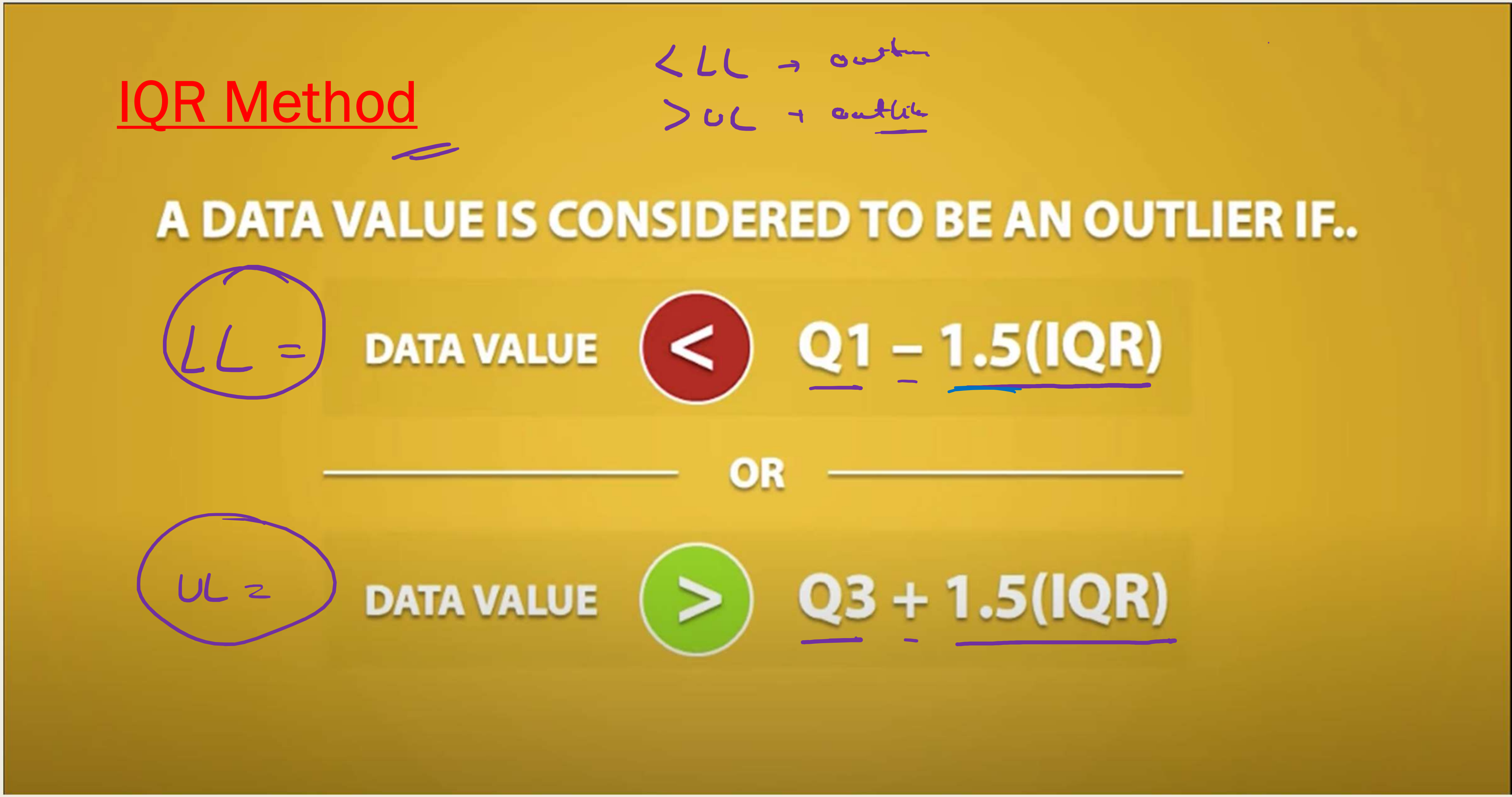




Why do we need to treat outliers?

Outliers can impact the results of our analysis and statistical modelling in a drastic way.

IQR Method



< LL → outlier

> UL → outlier

A DATA VALUE IS CONSIDERED TO BE AN OUTLIER IF..

LL = DATA VALUE < $\underline{Q1} \underline{-} \underline{1.5(IQR)}$

OR

UL = DATA VALUE > $\underline{Q3} \underline{+} \underline{1.5(IQR)}$

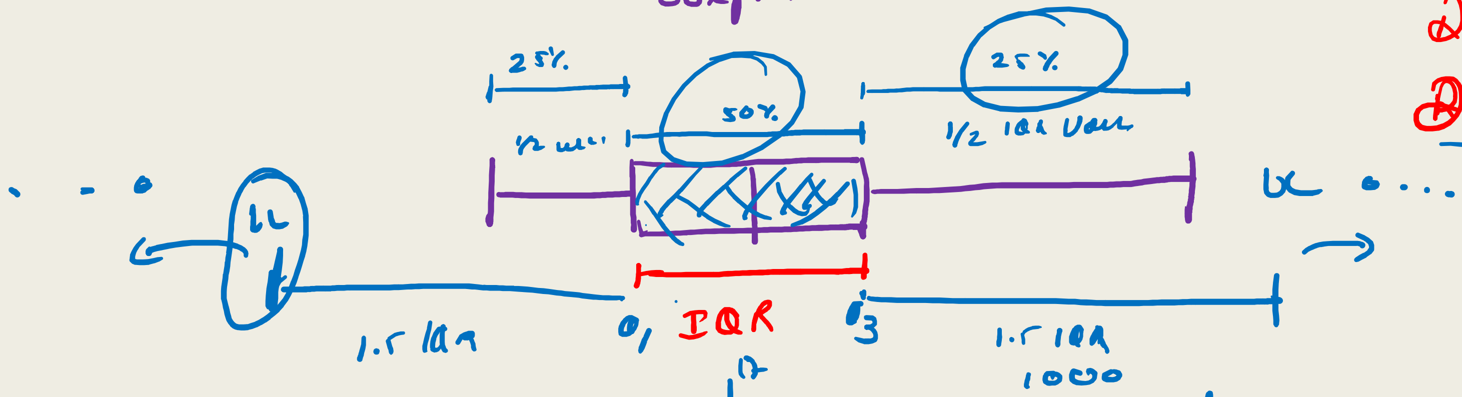
Boxplot

Data volume

Data Range

↓

IQR - 10



IQR →

1.5 IQR =

IQR
$Q_1 - Q_3$

data Volume
50%

data Range
100%

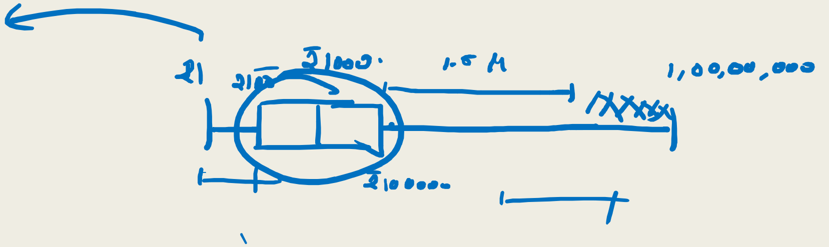
Min - Q_1 25%

Q_3 - Max 25%

50%

50%

1,00,000 - 100

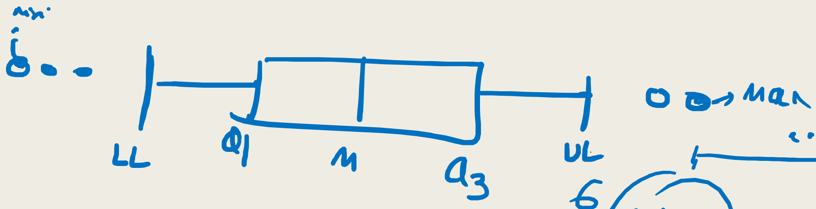


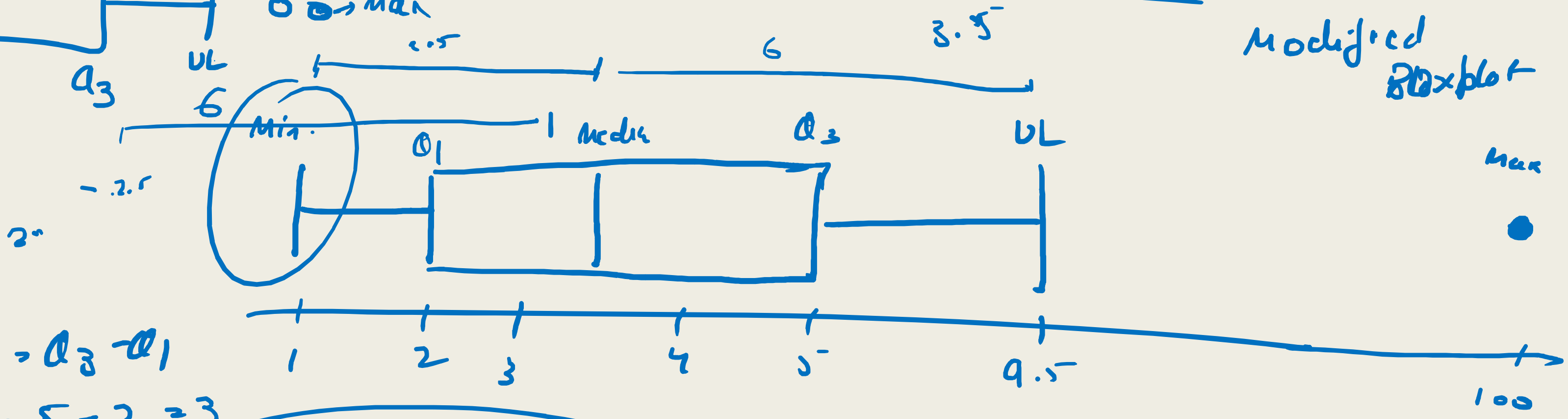
1/8 times of IQR

1 IQR 1/2 IQR

1 , ② , 3 , 4 , ⑤ , 100

Modified Boxplot





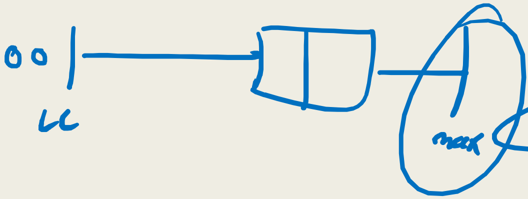
IQR = Q3 - Q1

= 5 - 2 = 3

$LL = Q_1 - 1.5\ IQR$

$= 2 - 1.5 * 3$

$= 2 - 4.5 = \boxed{-2.5}$



$UL = Q_3 + 1.5\ IQR$

$= 5 + 1.5(3) = 5 + 4.5$

$= \boxed{9.5}$

Min = 1

$Q_1 = 2$

med = 3.5

$Q_3 = 5$

max = 100

Q. Can you identify the outliers from the below dataset, using the IQR method?

-9 °C

26.0 °C , 15.0 °C , 20.5 °C , 31 °C , -350.0 °C , 31.0 °C , 30.5 °C

-350, 15, 20.5, 26, 30.5, 31, 31

7

Outlier

$1.5 * 16 = 24$

$UL = Q_3 + 1.5(IQR)$

$LL = Q_1 - 1.5(IQR)$

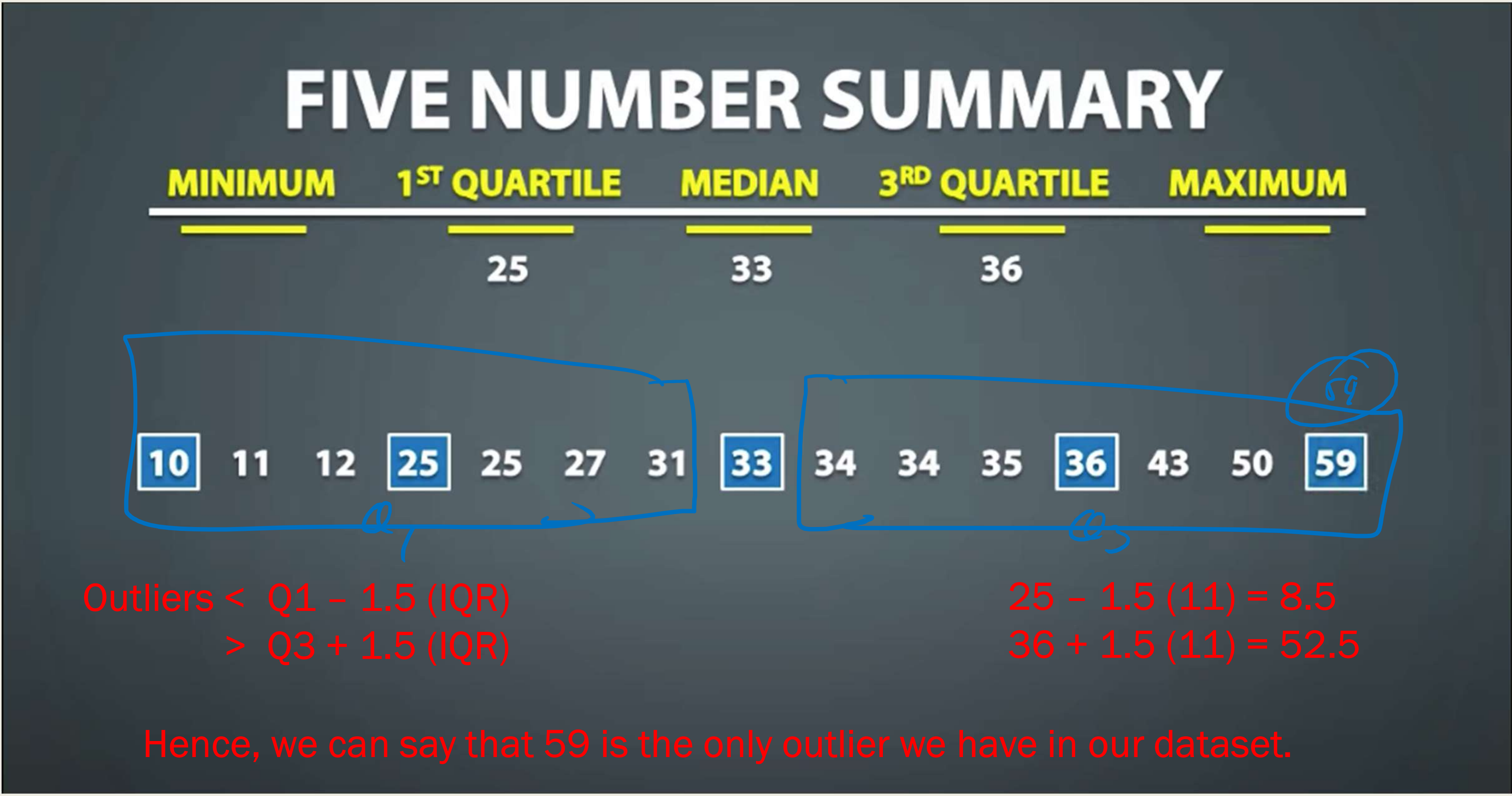
$IQR = Q_3 - Q_1$

$IQR = 31 - 15$

$= 16$

$UL = 31 + 24 = 55$

$LL = 15 - 24 = -9$



FIVE NUMBER SUMMARY

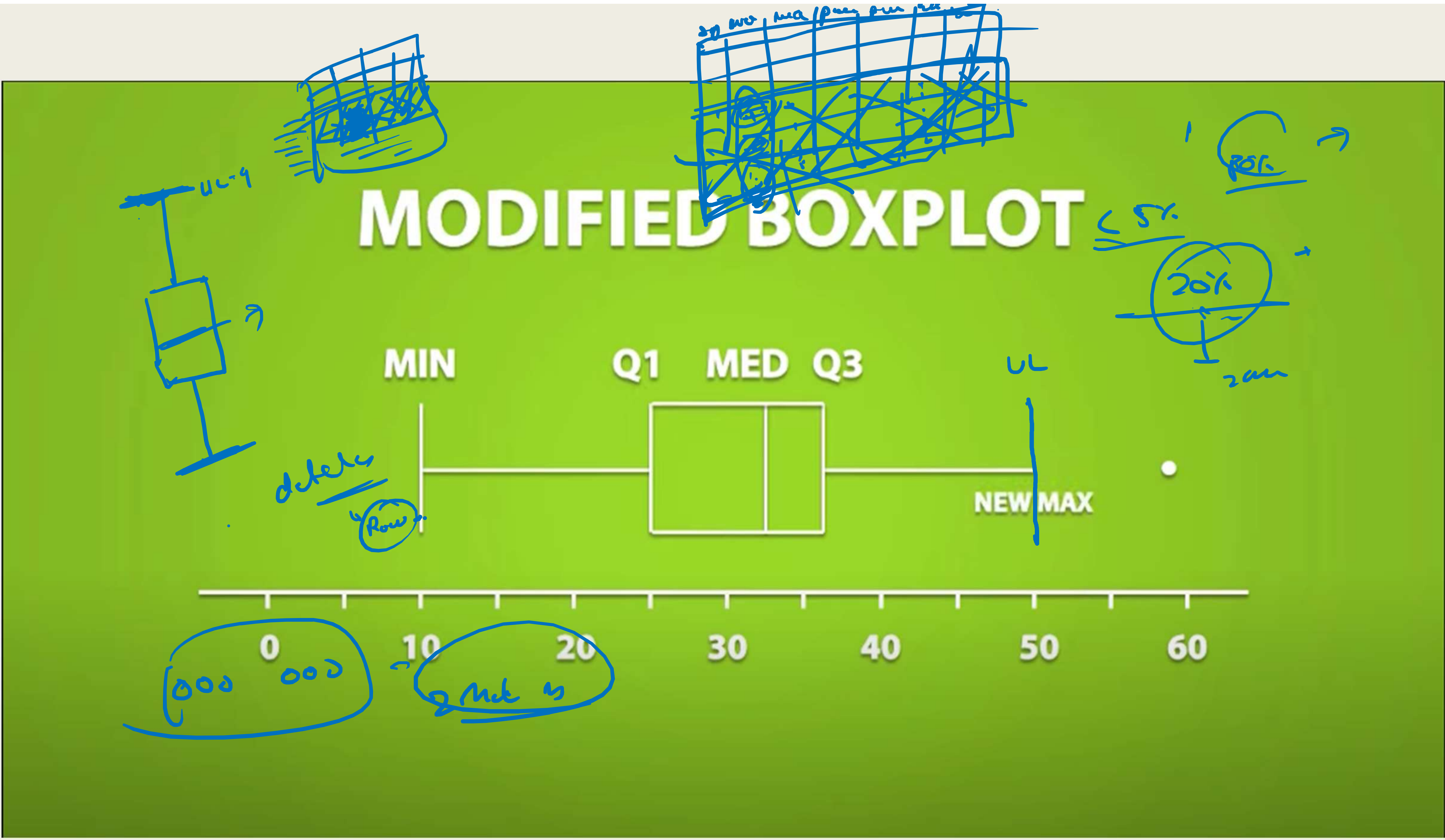
MINIMUM	1^{ST} QUARTILE	MEDIAN	3^{RD} QUARTILE	MAXIMUM
	25	33	36	

10 11 12 25 25 27 31 33 34 34 35 36 43 50 59

Outliers < Q1 – 1.5 (IQR)
> Q3 + 1.5 (IQR)

25 – 1.5 (11) = 8.5
36 + 1.5 (11) = 52.5

Hence, we can say that 59 is the only outlier we have in our dataset.



MODIFIED BOXPLOT