* Descriptive Statistics

↳ Summarising / organising the data / population without modifying it.

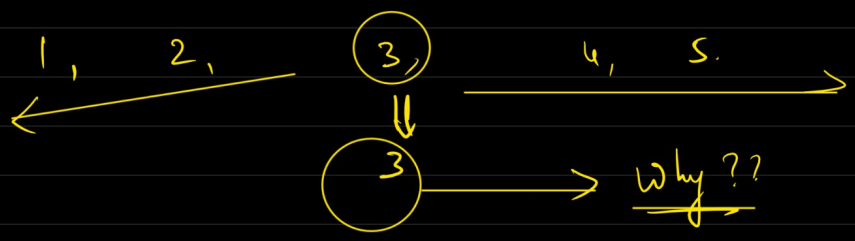
① Measures of CT
② Measures of Dispersion
③ Measures of Symmetricity.

* Measures of Central Tendency

⇓

Central → one value around which data is revolving.

What represents your village?
⇒ Mukhiya / Panchayat.



* CT → represents the center of data.

① Mean
② Median
③ Mode

① Mean (Arithmetic mid value of data)

Population → $x = \{1, 2, 3, 4, 5\}$

$\mu = \sum_{i=1}^{n} \frac{x_i}{N}$

Σ → Summation

$\sum_{i=1}^{N} x_i \Rightarrow \frac{x_1+x_2+x_3+x_4+x_5}{5}$

$$\frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$$

mu

μ (population mean)

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$$

n → Total no of elements in sample

$\bar{x}$ → Sample mean

→ Summing of all obs and dividing by no of observation.

② **Median** (Physical mid point of data)

6, 5, 2, 3, 1, 2

→ Sort the data, → 1, 2, 2, 3, 5, 6

→ Count is even

median = avg (two middle most ele)

$\frac{2+3}{2} = 2.5$

→ 4, 5, 2, 3, 1

→ Sort the data

→ Count = 5

→ middle most element of the data.

1, 2, 3, 4, 5

median = 3

Scenario 1 — Reliance Board of directors.

HR

	Salary
→ Sandeep	40
→ Abdullah	42
→ Kapish	41
→ Rahul	44
→ Deep	38

$$\frac{40 + 42 + 44 + 38}{4} = \frac{(164)}{4} \Rightarrow 41$$

Scenario 2

Sandeep → 40

Abdullah — 42

Kapish —

Rahul — 44

Deep — 38

Mukesh Ambani → 500

Mean ?

Why ??

Outlier

Mean

Scen 1

1, 2, 3, 4, 5

Avg value ⇒ 3

Scen 2

1, 2, 3, 4, 100

Avg ⇒ 22

Scen 1

Median 1, 2, 3, 4, 5

median = 3

Scen - 2

1, 2, 3, 4, 100

median = 3

* Mean is inflated by outliers.

Application → ①

Subu → 6 kids of 13 years
→ 5 babies of 1 years

fun activity

birthday → mean ⇒ $\frac{13+13+13+13+13+1+1+1+1+1}{11} = 7.5$

Avg age of children → 7.5 years.

Jumpy Castle

App-2

Age	Gender	weight	Salary
25	M	50	50
26	M	—	31
—	M	55	60
	M	60	70
23	M	61	30
25	M		
	F	miss	
	F	150	
	M		
	M		

Avg

Median

Mean

Scen-1

0, 1, 2, 3, 4

→ mean = 2

Scen-2

0, 1, 2, 3, 1000

mean = 169

* Mean can change / inflated due to presence of outlier

→ Arithmetic mid.

Median

Scen-1

0, 1, 2, 3, 4

median = 2

Scen-2

0, 1, 2, 3, 1000

medi = 2

⇒ Median will not be changed due to presence of outlier.

* Mode → maximum frequency.

Mode is the term appearing maximum times in the data.

mean / median ⟹ Numerical

Mode ⟹ Categorical ⇒ frequency

Gender

M - 1

0.4

Will it make sensen ??

0 - 1

0.4

2
2
2

M.C.T

↳ mean
↳ medi
↳ mode

1, 2, 3, 4, 5, 6, 7, 8, 9, 10 ..... 999 999

mode → 999

What would be correct way ??

1, 2, 3, 4, 5

→ mode → no mode value.

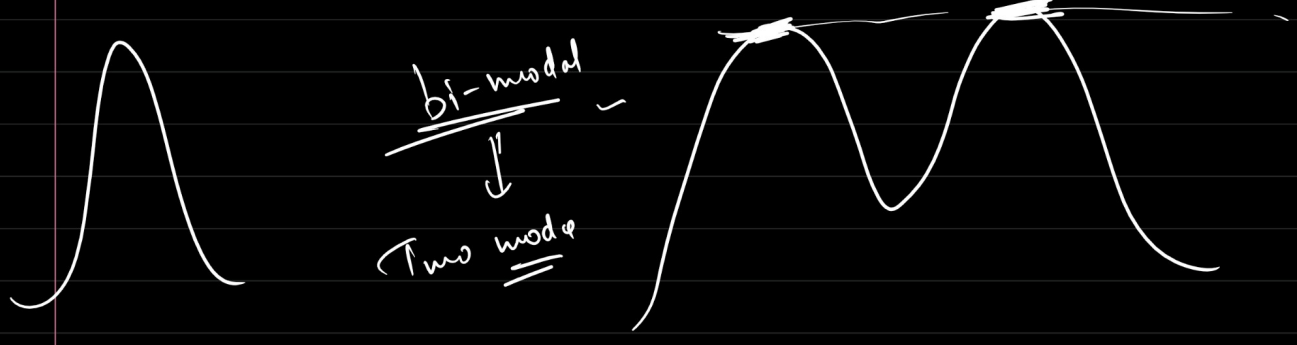
⇒ 1, 2, 2, 3, 3, 4, 5 → 2 → mode

2 3

bi-modal

↓

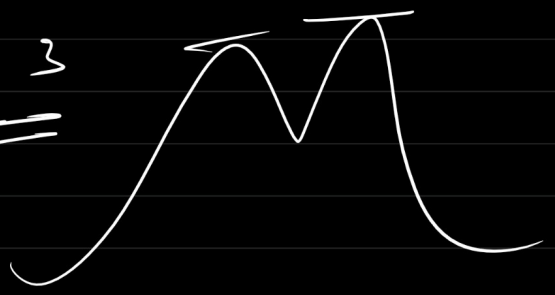
Two mode



2, 2, 3, 3, 4, 5, 6

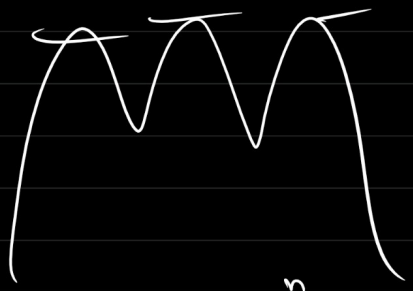
mode 2 & 3

bi-modal



2, 2, 3, 3, 4, 4, 5, 6, 7, ...

→ 2, 3, 4 (multi-modal)



Mean → Arithmatic mean → $\sum_{i=1}^{n} \frac{x_i}{n}$

O Geometric mean → multiplies the no together and does a square root / cube root or how many nos are there.

2 & 10

$\sqrt[2]{2 \times 10} = (2 \times 10)^{1/2}$

$= 4.4$

2, 2, 5

→ $\sqrt[3]{2 \times 2 \times 5} = (2 \times 2 \times 5)^{1/3}$

↳ 2, 2, 5, 1, 8 ⇒ $\sqrt[5]{2 \times 2 \times 5 \times 1 \times 8} = (2 \times 2 \times 5 \times 1 \times 8)^{1/5}$

Defn

for n number, multiply them all together and then take n^{th} root

$$\sqrt[n]{\quad}$$

$a_1, a_2, a_3 \ldots a_n.$

$$\Rightarrow \sqrt[n]{a_1 a_2 a \ldots \ldots a_n.}$$

Q - Camera 1 : Zoom 200 Review : 8

- Camera 2 : Zoom 250 Review : 6

Camera 1 $\frac{200+8}{2} = 104$

Camera 2 $\frac{250+6}{2} = 128$

→ larger diff in magnitude

→

Cam 1 $= \sqrt{200 \times 8} = 40.$

Camera 2 $\sqrt{250 \times 6} = 38.$

* The geometric is used when we want to compare two different properties.

* Harmonic mean

$$HM = \frac{n}{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \ldots \frac{1}{n}}$$

$2, 4, 5, 100 \Rightarrow \frac{4}{\frac{1}{2} + \frac{1}{4} + \frac{1}{5} + \frac{1}{100}} = \frac{4}{0.5 + 0.25 + 0.20 + 0.001}$

→ The reciprocal of avg of all reciprocals.

1, 2, 3

$\frac{\frac{1}{1}+\frac{1}{2}+\frac{1}{3}}{3}$ reciprocal $= \frac{3}{\frac{1}{1}+\frac{1}{2}+\frac{1}{3}}$

ML

① Confusion matrix ⇒ F score

② We travelled 10 km at 60 km/hr, then another 10 km at 20 km/hr. What is avg speed?

$$\frac{2}{\frac{1}{60}+\frac{1}{20}} \qquad \frac{n}{\frac{1}{a}+\frac{1}{b}}$$

frequency mean

f	fx
2	1×2
5	2×5
3	3×3
6	6×4
5	5×5
2	6×2
(20)	(150)

Score x	Frequency (f)
1	2
2	5
3	3
4	6
5	5
6	2

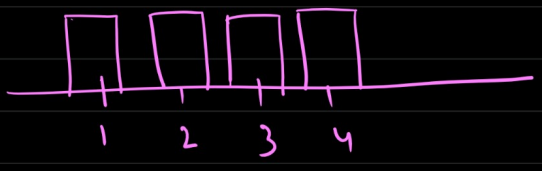
$$\frac{\Sigma fx}{\Sigma f} = \frac{150}{20}$$

1 + 2 + 2 + 2 + 2 + 2 + 3 + 3 + 3 + 4 + 4 + 4 + 4 + 4 + 4 + 4 + 4 + 4 + 4 + 5 + 5 + 5 + 5 + 5 + 6 + 6

* Weighted mean

Mean = $\frac{1+2+3+4}{4}$

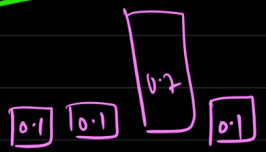
All the values have equal weights



$\Rightarrow 1\times\frac{1}{4} + 2\times\frac{1}{4} + 3\times\frac{1}{4} + 4\times\frac{1}{4}$

$\Rightarrow$ 1 2 3 4 $\quad \sum w = 1$

w_i = 0.1 0.1 0.7 0.1



Calculate mean $\Rightarrow$ $1\times0.1 + 2\times0.1 + 3\times0.7 + 4\times0.1$

$\Rightarrow 2.8$

Weighted mean can be used where something are important as compared to other.

Q. Awanish

- Image Quality : 50%
- Battery life : 30%
- Zoom range : 20%

Sony → IQ 8, BL 6, ZR 7

Sony = $0.5\times8 + 0.3\times6 + 0.2\times7$

Canon

IQ	BL	ZR
9	4	6

Canon = $0.5\times9 + 0.3\times4 + 0.2\times6$

= 7.2

= 6.9

What if weights don't add to 1

↳ divide the sum by weights

Dinesh → 7 times a week, 1, 2 or 5 lunches.

On 2 weeks → only one lunch for whole week.

" 14 weeks = two " " "

" 8 weeks = 5 " " "

" 32 " = 7 " " "

What is the mean no of lunch that Dinesh have every week??

Weeks × lunch = 2×1 + 14×2 + 5×8 + 32×7

total weeks = 2 + 14 + 8 + 32 ⇒ 56

Descriptive Stats

↳ M.CT

S_1

↳ 1, 2, 3, 4, 5

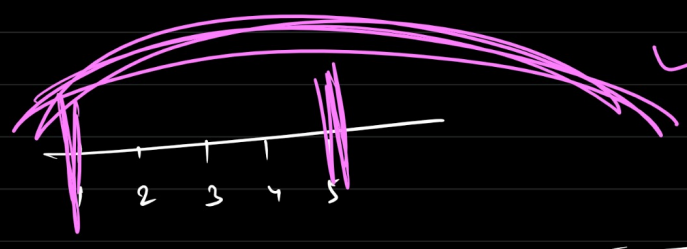
M.CT → 3, 3.

S_2

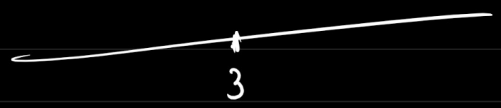
↳ 3, 3, 3, 3, 3.

3, 3

S_1



S_2



M. CT alone can not say everything about the data

② Measures of Spread | dispersion

- Range
- Percentage | percentile
- Quartile (IQR, Box plot)
- variance
- standard dev^n

① **Range** → difference b/w maximum and min value

1, 2, 3, 4, 5

range = 5 − 1 = 4

⇒ 1, 2, 3, 4, 1000 → range = 1000 − 1 = 999

* outlier affects the range.

② **Percentage & Percentile**

1, 2, 3, 4, 5.

No's are odd → $\frac{3}{5} \times 100 = 60\%$

* Percentile — A percentile is a value below which a certain percentage of observation lie.

1, 2, 3, 4, 4, 6, 7, 7, 8, 10

What is the percentile rank of 3?

percentile rank of a no = $\frac{\text{No of values below that no.}}{\text{Total no}} \times 100$

$= \frac{2}{10} \times 100 = 20^{th}$ percentile

What value exist at 75^{th} percentile

$$\text{Value} = \frac{\text{percentile} \times (n+1)}{100}, \quad n \rightarrow \text{total no.}$$

$$= \frac{75}{100} \times (10+1)$$

$$= \frac{75}{100} \times 11 = 8.25^{th} \text{ no}$$

$\Downarrow$

8^{th} no.

The 8^{th} no that is 7 is my 75^{th} percentile.

Quartile

↳ Quartiles are values that divides a list of no into Quarters.

↓

4 equal parts

eg 6, 8, 5, 5, 7, 3, 9

Order → 3, 5, 5, 6, 7, 8, 9.

1st, 2nd, 3rd, 4th part.

Quartile-1 Quartile-2 Quartile-3

$Q_1 - 5$

$Q_2 - 6$

$Q_3 - 8$

Quarters = 4

3, 5, 5, 6, 7, 8, 9 10

Quarter 1 Quarter-2 Quarter Quarter.

Quarter-1 →

Jan Feb March April May June July Aug Sep Oct Nov Dec

Quarter 1 Quarter-2 Quarter-3 Quarter 4

eg 1, 1, 1, 1, 2, 2, 2, 3, 3, 3, 4

↑ Q_1 ↑ Q_2 ↑ Q_3

Order

total count = 11.

df.describe()

Count - odd

$Q_1 = 1 \times \left(\frac{n+1}{4}\right)^{th} = \left(\frac{11+1}{4}\right)$ obs. $= 3^{rd}$ obs

$Q_2 = \left(\frac{n+1}{2}\right)^{th} = \left(\frac{11+1}{2}\right)^{th} = \left(\frac{12}{2}\right)^{th} = 6^{th}$ obs

$Q_3 = 3\left(\frac{n+1}{4}\right)^{th} = 3 \times \left(\frac{11+1}{4}\right)^{th} \Rightarrow 9^{th}$ obs

Count = even

$Q_1 = \frac{n}{4}^{th}$ obser

$Q_3 = 3 \times \frac{n}{4}^{th}$ obser

$Q_2 = \frac{\left(\frac{n}{2}\right)^{th} + \left(\frac{n}{2}+1\right)^{th}}{2}$

$Q_1 = 1.5$

1, 2, 3, 3, 4, 4 — Q_3

Count: n = 6.

$Q_1 = \frac{6}{4}^{th} = \frac{3}{2}^{th} = 1.5^{th}$ obse

$Q_3 = \frac{3n^{th}}{4} = \frac{3 \times 6}{4} = 4.5^{th}$ term $= 4$

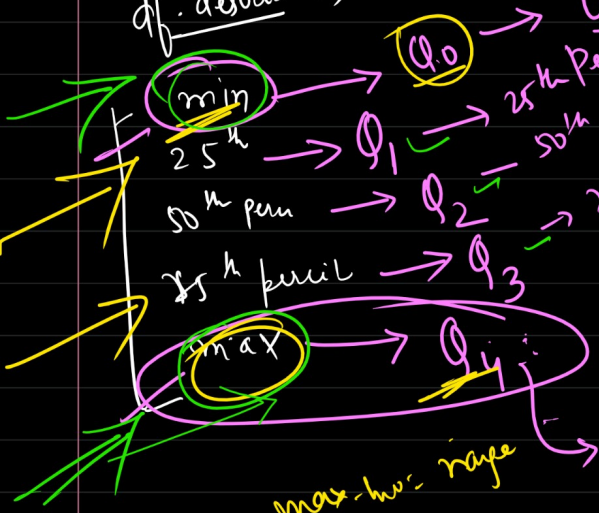
$Q_2 = \text{avg}\left(\frac{n}{2}^{th}, \frac{n}{2}+1^{th}\right)$ term $= \frac{3+4}{2} = 3.5$

$= \text{avg}(3^{rd}\ 4^{th})$

$\frac{3+3}{2} = 3$

FIVE POINT SUMMARY

df.describe() → FIVE POINT SUMMARY



$Q_1 - 1000$ (The 25% of population do the transaction of 2000 Rs or below 2000)

$Q_2 - 2000$

$Q_3 - 3100$

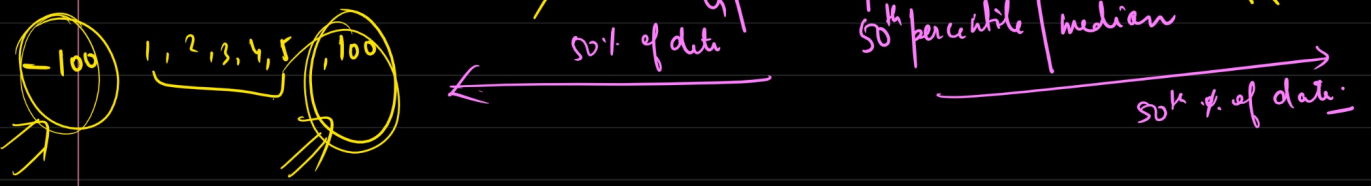
transactia amount
506
106
600
700
.
.
.

Outlier

↓

extreme value

much-much higher
much-much lower



Outliers will lie either is Q_1 & Q_4

$\text{inter Quartile range} = Q_3 - Q_1$

$\frac{3+1}{2} = 2$

2, 3, 3, 3, 3, 4, 4, 5, 5, 5, 6, 6, 6, 7, 8, 99

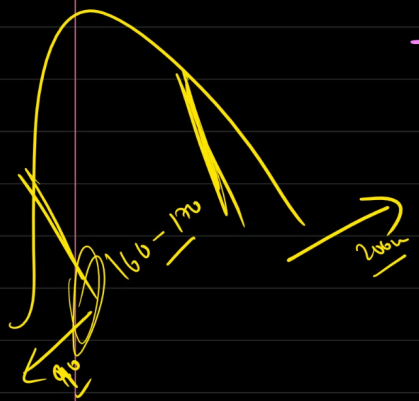
$\frac{n+1}{2}$ th

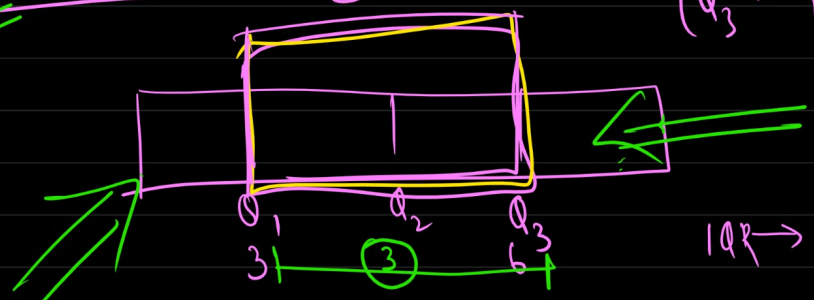
$Q_1 = 25^{th}$ percentile $= \frac{25}{100} \times 16 = \frac{1}{4} \times 16 \Rightarrow 4^{th}$ no $= Q_1 = 3$

$Q_3 = 75^{th}$ percent $= \frac{75}{100} \times 16 = \frac{3}{4} \times 16^{4} = 12^{th}$ no $= Q_3 = 6$

$$IQR = Q_3 - Q_1 = 6 - 3 = 3$$

$(Q_3 - Q_1)$





IQR → $Q_3 - Q_1 = 6 - 3 = 3$

↳ Most of elements lie here 68% — 1 IQR

✓ Lower fence $= Q_1 - (1.5) \times IQR$

✓ Upper fence $= Q_3 + (1.5) \times IQR$

$Q_1 - 3$

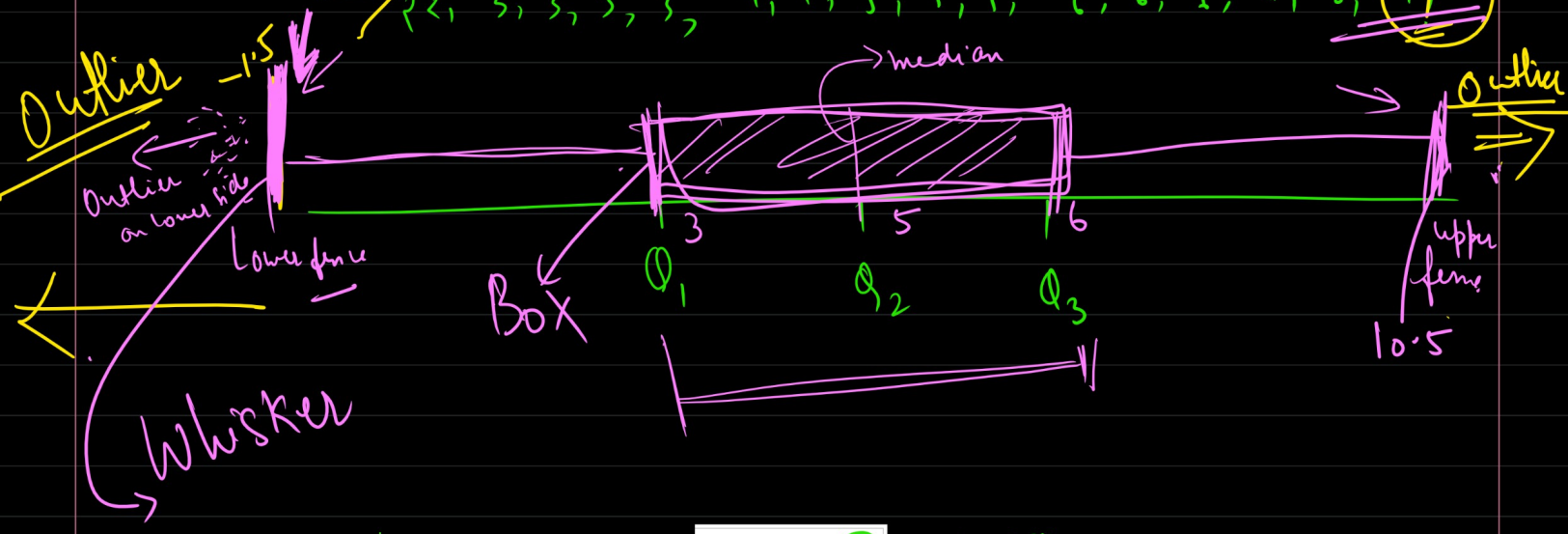
$Q_3 - 6$

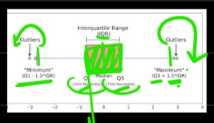
LF $= 3 - 1.5 \times 3 = -1.5$

UF $= 6 + 1.5 \times 3 = 10.5$

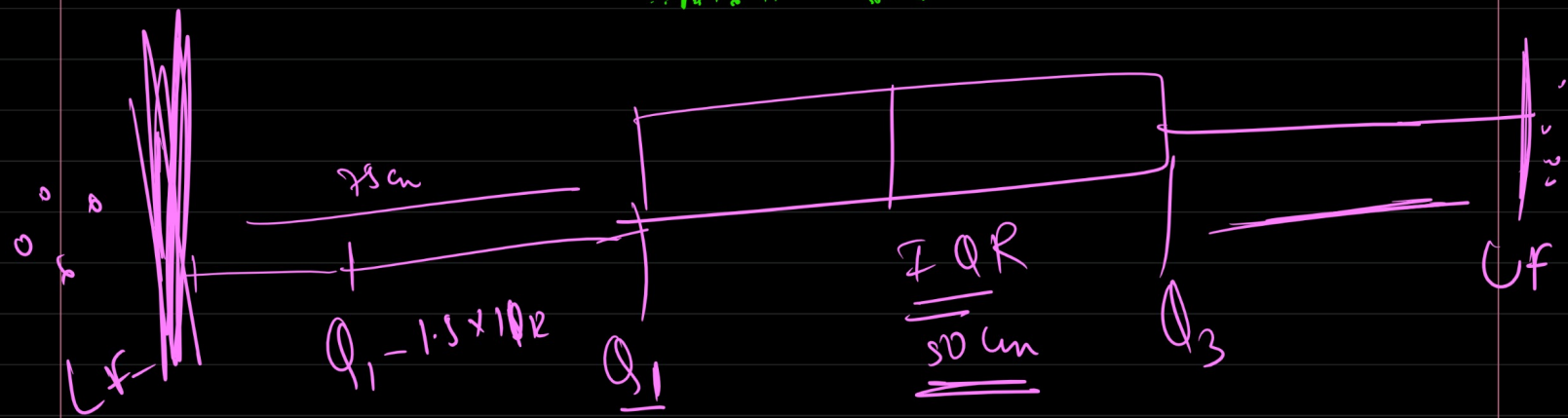
{2, 3, 3, 3, 3, 4, 4, 5, 5, 5, 6, 6, 7, 8, 99}

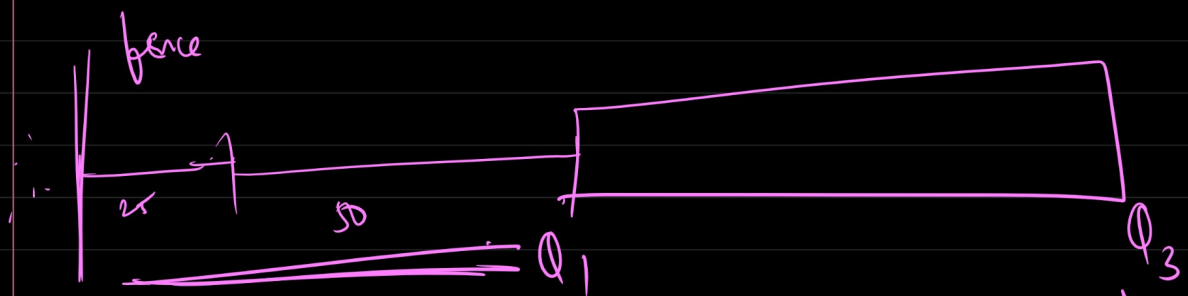
Outlier





LF $= Q_1 - 1.5 \times IQR$





$Q_1 - 1.5 \times IQR$

Box Plot → outlier → are extreme values

$$LF = Q_1 - 1.5 \times IQR$$

$$UF = Q_3 + 1.5 \times IQR$$