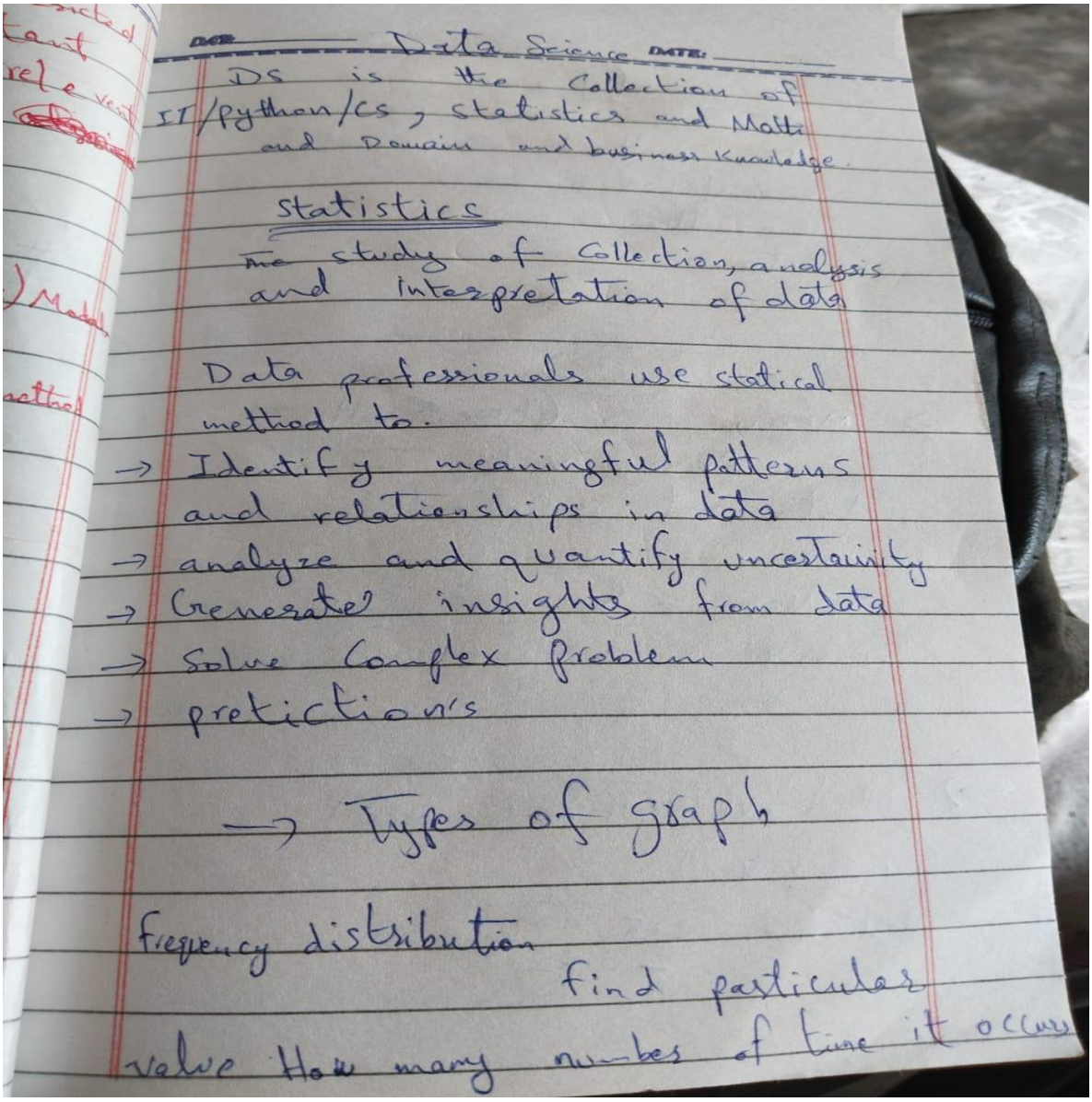
Introduction to Statistics

Some screenshot :



DATE:

Measure of frequency:

we use this method when outliers are found in dataset this method is use to clean data.

→ Mean (also called average) → (if use if outliers not occure)

Average value in data set is called mean.

$$\text{Mean} = \frac{\text{Sum of all values}}{\text{total \# of value}}$$

1, 5, 9

$$\text{Mean} = \frac{1+5+9}{3} = \frac{15}{3} = 5$$

→ Median (if use if out liers occure)

1, 2, 5 9

$$\text{Median} = \frac{2+5}{2} = \frac{7}{2}$$

Mod

when mod is useful

* The model is useful when Catagorical data is involved e.g Yes or no

Mod

Most frequently value

e.g

data = 1, 1, 9, 5, 3

In which 1 is most frequently value hence

Mode = 1

there should be more than one Mod. e.g

data = 5, 5, 5, 3, 1, 9, 9

In which 5 and 9 is most repeated value so Mod will be 5 and 9

$5\sqrt{2}$

Numpy

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

$$\text{mean} = \frac{\overbrace{1+2}^{5}+\overbrace{2+3+3+3+3}^{15}+\overbrace{3+4+4}^{8}+\overbrace{5+5}^{10}}{11}$$

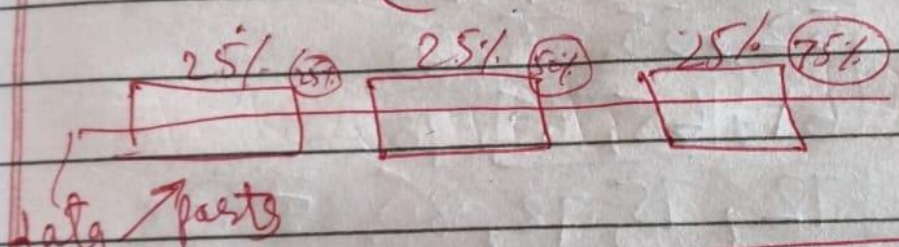
$$\text{mean} = \frac{38}{11} = \boxed{3.45}$$

DATE:

DATE:

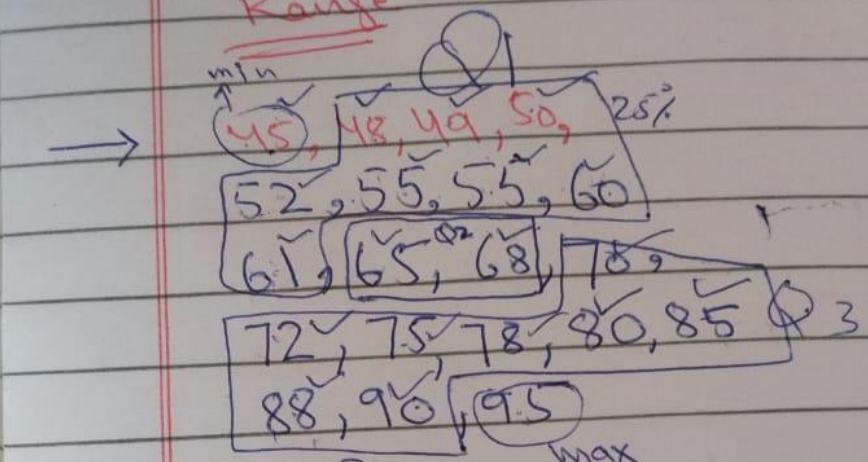
(1) Percentile → Measure of position finding below of total percentage (comparing values)

(2) Quartile → (Convert data into parts)



→ Range
→ standard deviation
→ 5 # Summary

Range



→ range = ?

range = 95 - 45

range = 50

→ 5 # Sumary = ?

min = 45

max = 95

$Q1 = \frac{48+49+50+52+55+55+60+61}{8}$

$Q1 = \frac{430}{8}$

Q1 = 53.75

DATE: DATE:

$$Q_3 = \frac{70+72+75+78+80+85+88+90}{8}$$

$$Q_3 = \frac{638}{8}$$

$$\boxed{Q_3 = 79.75}$$

Interquartile range $= Q_3 - Q_1$

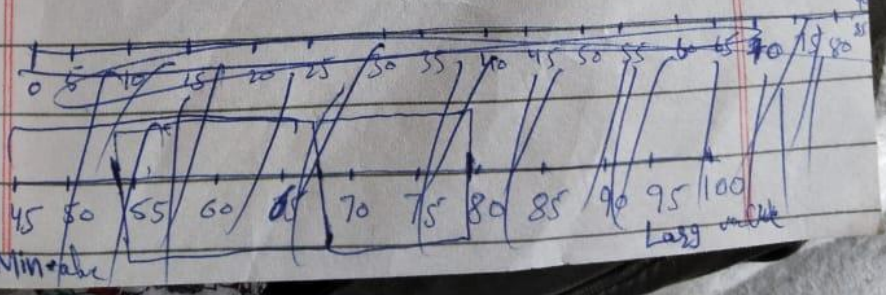
$IQ_{range} = 79.75 - 53.75$

$\boxed{IQ_{range} = 26}$

$$Q_2 = \frac{65+68}{2}$$

$$Q_2 = \frac{133}{2}$$

$$\boxed{Q_2 = 66.5}$$



Data Science

DS is the Collection of IT/Python/Cs, statistics and Math and Domain and business Knowledge.

Statistics

The study of Collection, analysis and interpretation of data

Data professionals use statical method to.

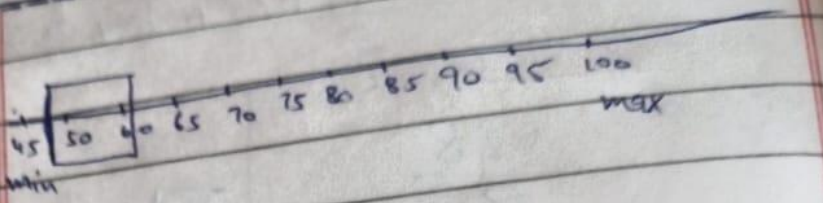
→ Identify meaningful patterns and relationships in data
→ analyze and quantify uncertainity
→ Generate insights from data
→ Solve Complex problem
→ pretictions

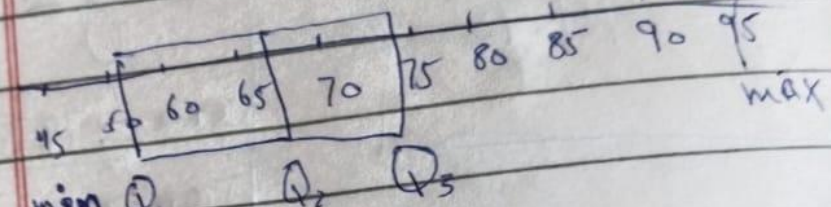
→ Types of graph

Frequency distribution

find particular value How many numbes of time it occurs

DATE:

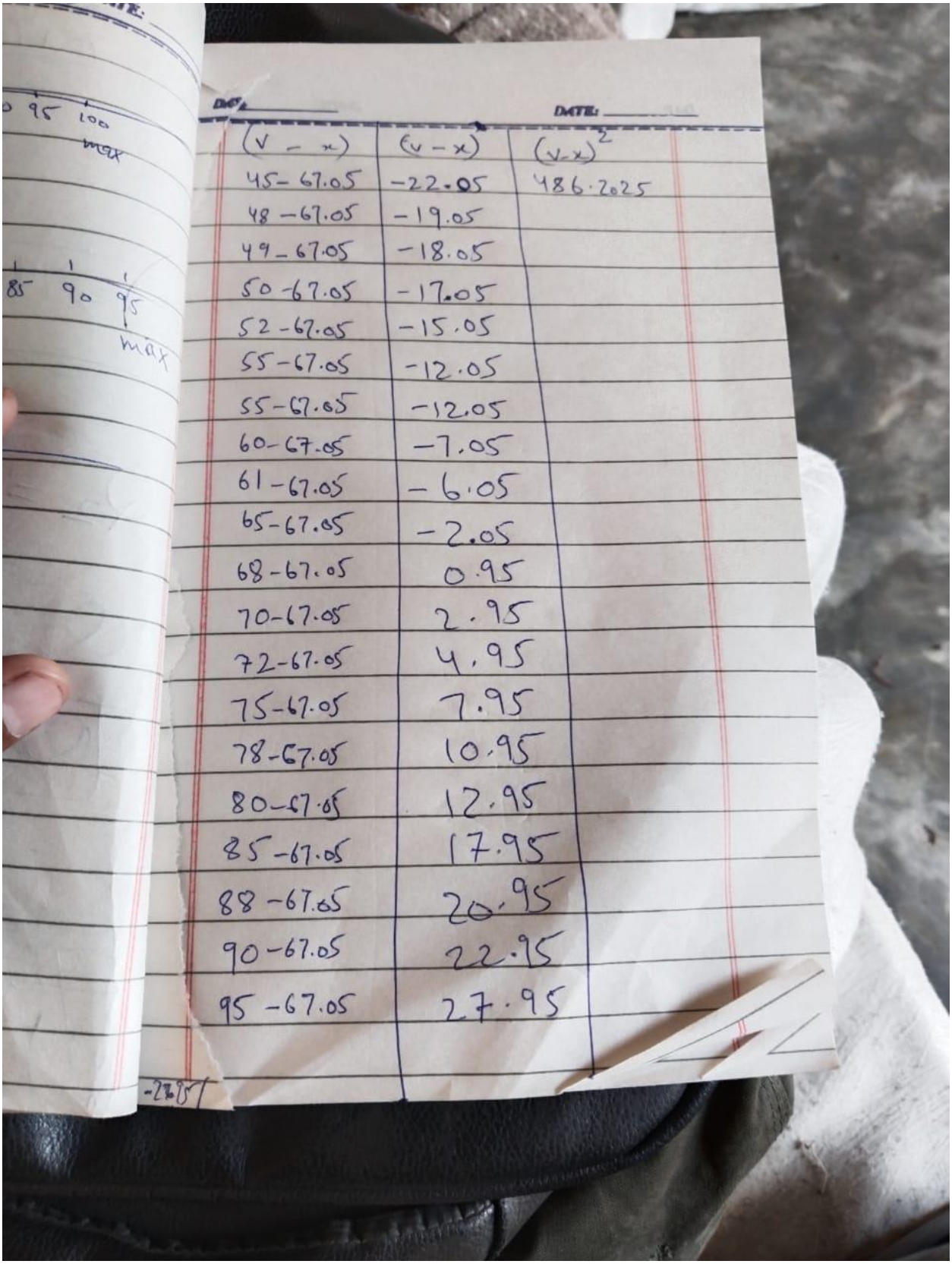




Standard deviation

~~standard deviation~~ $= \frac{1341}{20}$

$\sqrt{V} = 67.05$



95 100
max

85 90 95
max

DATE:

(v − x)	(v − x)	$(v-x)^2$
45 − 67.05	−22.05	486.2025
48 − 67.05	−19.05	
49 − 67.05	−18.05	
50 − 67.05	−17.05	
52 − 67.05	−15.05	
55 − 67.05	−12.05	
55 − 67.05	−12.05	
60 − 67.05	−7.05	
61 − 67.05	−6.05	
65 − 67.05	−2.05	
68 − 67.05	0.95	
70 − 67.05	2.95	
72 − 67.05	4.95	
75 − 67.05	7.95	
78 − 67.05	10.95	
80 − 67.05	12.95	
85 − 67.05	17.95	
88 − 67.05	20.95	
90 − 67.05	22.95	
95 − 67.05	27.95	

-2695