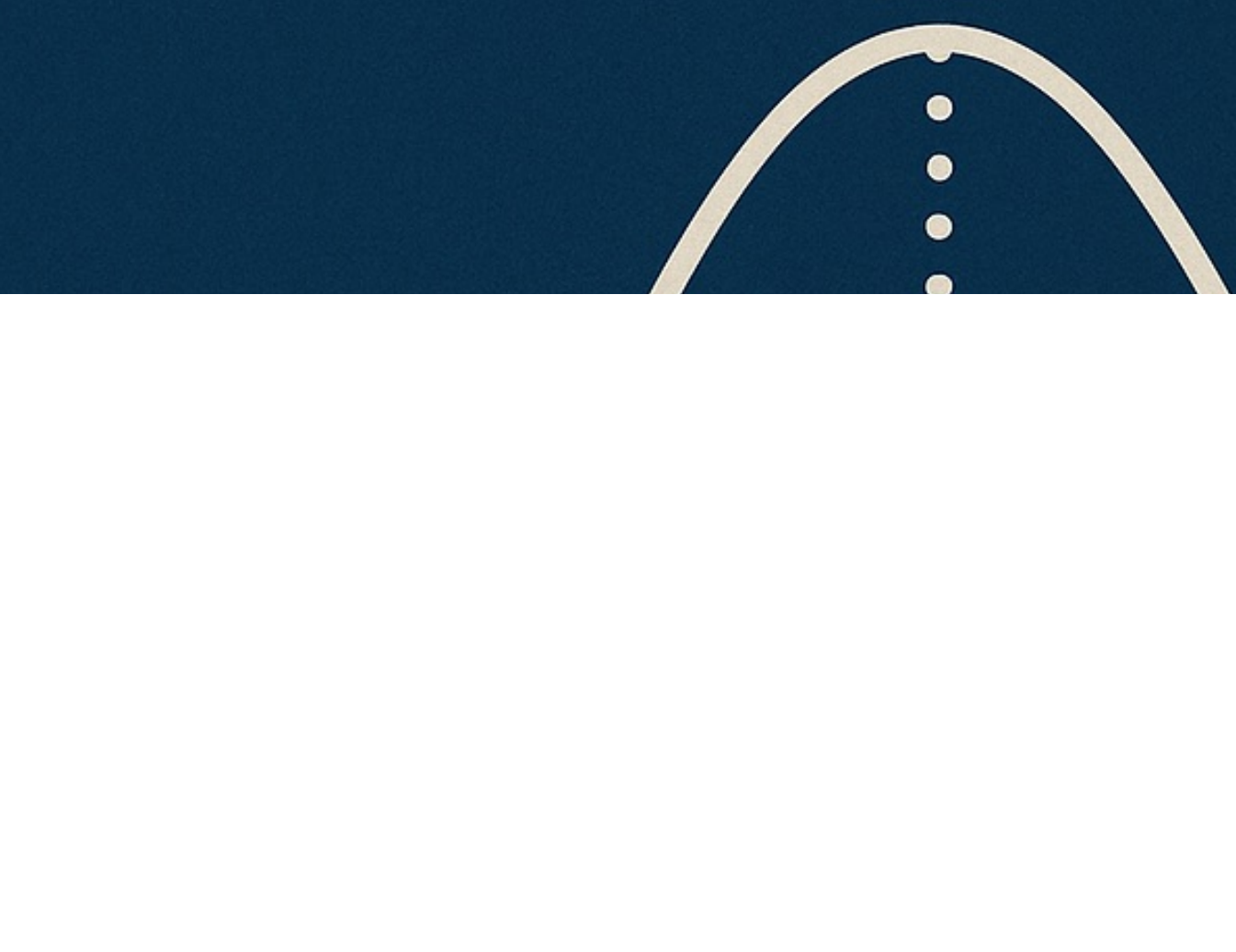
DAV-3

HYPOTHESIS TESTING

(Class starts @ 9:08 PM)



power of test

Act \ pred	0	1
0	TN	FN → β
1	FP → α	TP

β ↓ FN ↓ TP ↑ (1-β) ↑

TP ↑ → goal (power of Test)

(1 - β)

Lecture 8: Correlation

#Agenda

1. Parametric Vs Non parametric
2. Covariance & Correlation
3. Pearson Correlation
4. Spearmann Correlation
5. Vizualizing using Heatmap

Parametric

vs

Non-Parametric

Test

Parametric Vs Non-parametric Test

Parametric

Check the underlying distribution of data before Test

T-test, ANOVA

Non-Parametric

No need to check the underlying distribution of data before test

KS Test, KW Test

#Parametric HT

Assumptions:

- Parametric tests make specific assumptions about the population distribution from which the data is drawn.
- Common assumptions include normality (data follows a normal distribution) and variance is constant across groups or conditions.
- Parametric tests are typically used when the data reasonably follows the assumed distribution and other assumptions are met.
- Parametric tests tend to be more powerful (i.e., better at detecting true effects) than non-parametric tests when the assumptions are met.
- This is especially true when the sample size is large.

Non-Parametric HT

Assumptions:

- Non-parametric tests make fewer or no assumptions about the population distribution.
- They are distribution-free or rely on fewer assumptions, such as independence of observations.
- Non-parametric tests are useful when the assumptions for parametric tests are violated.
- They are also suitable for data types that don't fit well with parametric assumptions, such as ordinal or skewed data.
- Non-parametric tests are generally less powerful than parametric tests when data conforms to parametric assumptions.
- However, they can be more robust and appropriate when dealing with non-normally distributed data.

Parametric

Z test

paired T test

ANOVA

Levene's Test

Non-parametric

Chi Sq Test

KS Test

KW Test

Types of Test

So far

✓ Numeric Vs Categorical → 2 (T-test)
→ >2 (ANOVA)

✓ Categorical Vs Categorical → Chi square test

✓ Numeric Vs Numeric → Correlation Test

1. Pearson Correlation
2. Spearman Correlation

Covariance And Correlation

Class starts at 9:05 PM

Agenda

1. Covariance and Correlation
2. Pearson Correlation
3. Spearmann Correlation
4. Visualization using Heatmap.

Example : Height Weight

Statistician gathers Height & Weight data

↳ To find the relation b/w H & W.

Height	Weight

(Army Camp)

↳ Fit

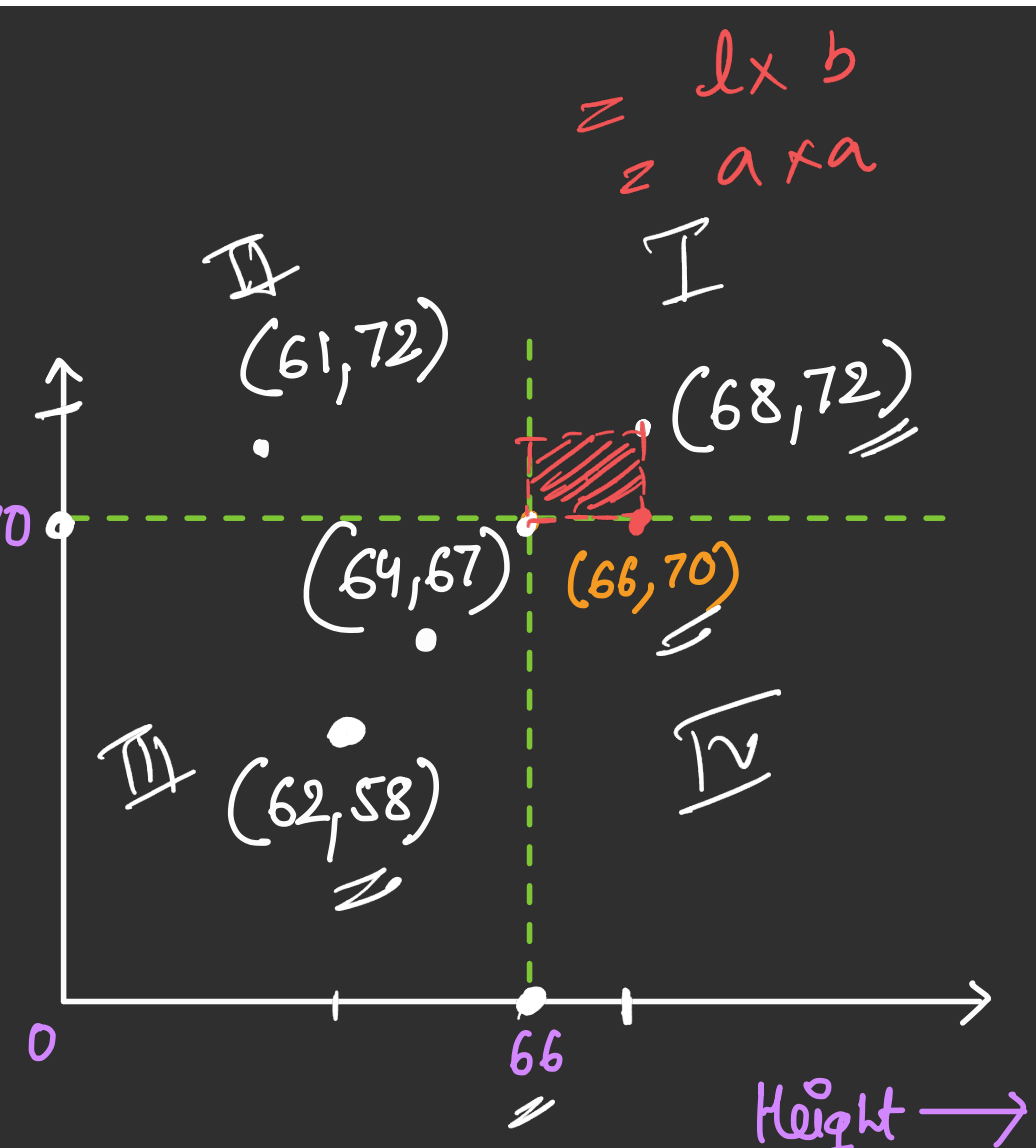
Height vs Weight → correlation

num num

Example 1 : Height and Weight

Height (inches)	Weight (kg)
68	72
62	58
64	67
61	72
70	79
66	61
61	68
65	64
71	80
72	79
$\bar{h} = 66$	$\bar{w} = 70$

mean mean

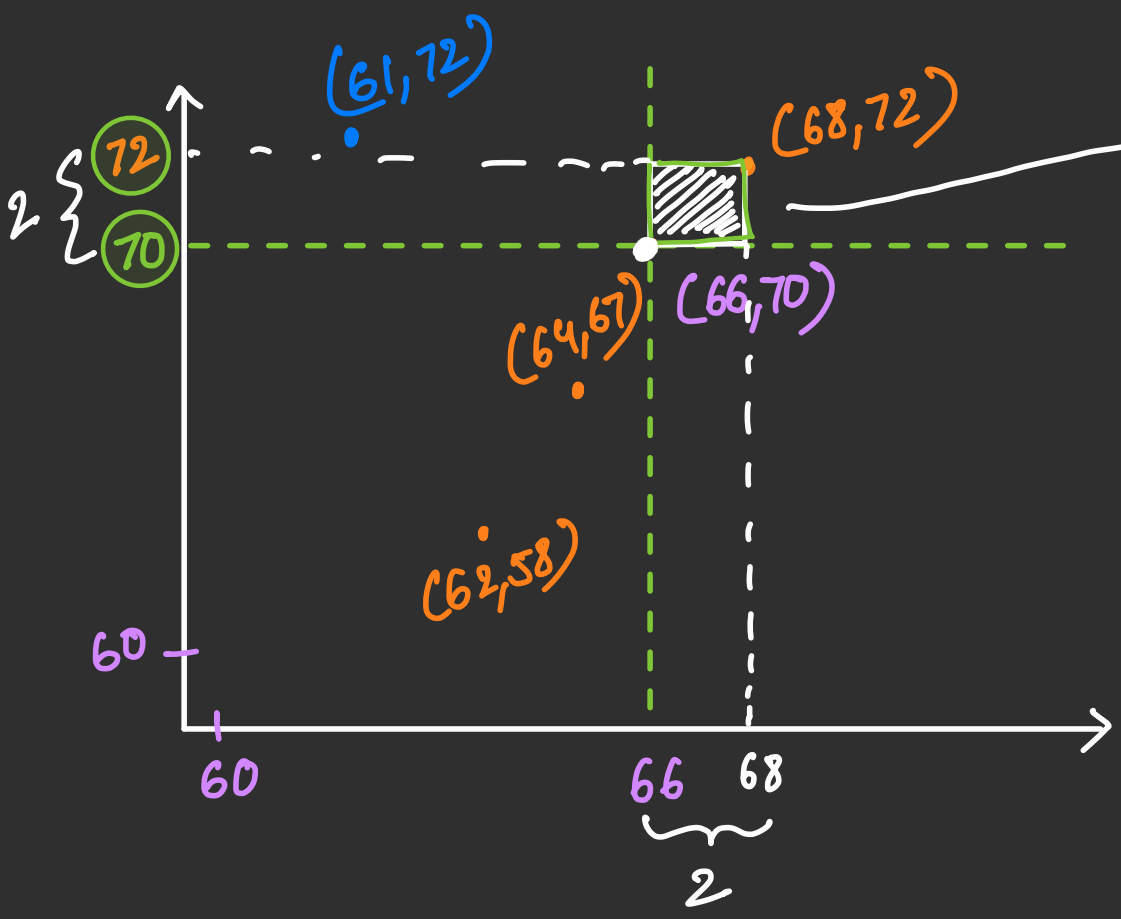


$= l \times b$

$= a \times a$

$(68-66) \times (72-70)$

$= 2 \times 2 = 4$



What will be the area?

$68 - 66 \Rightarrow 2$

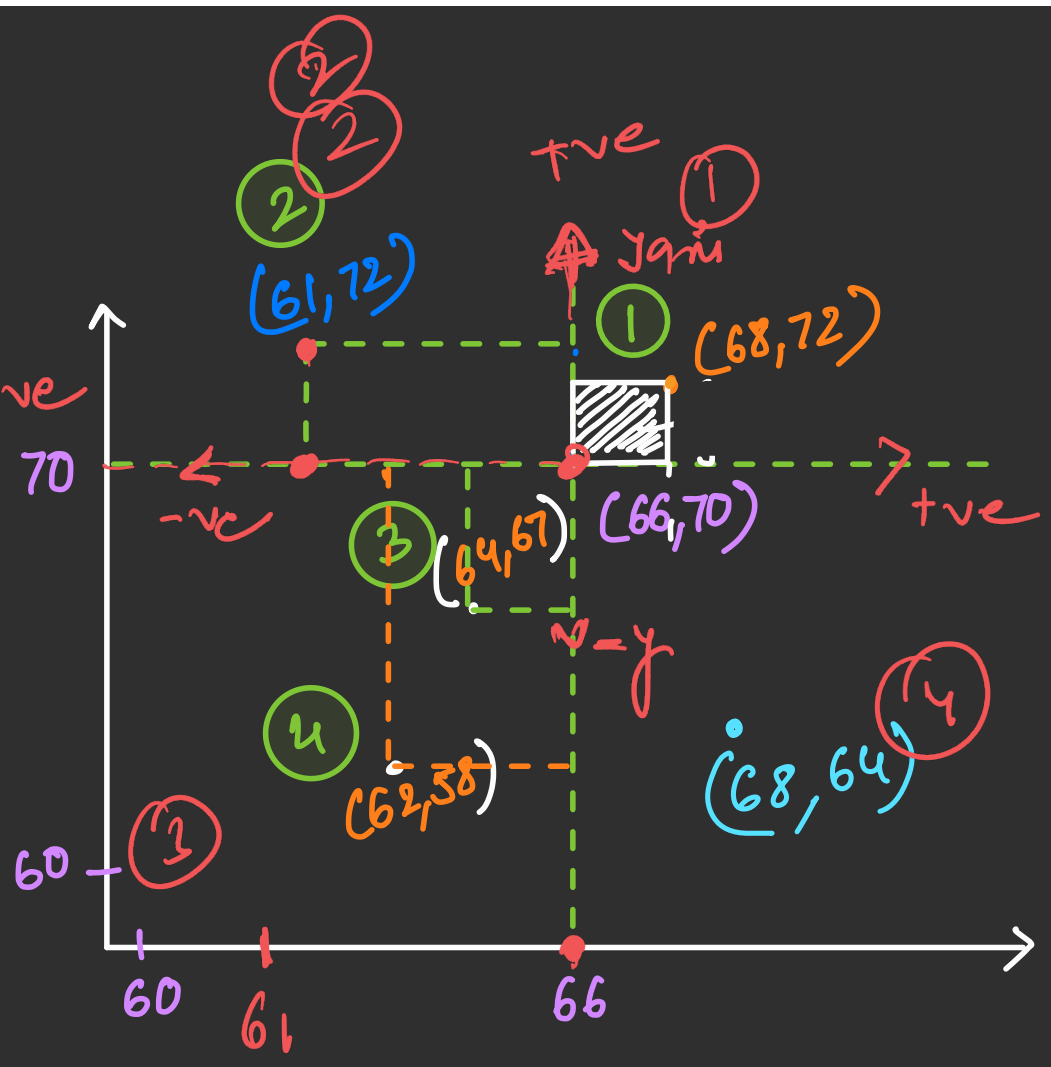
$72 - 70 \Rightarrow 2$

Area = $a \times a$

$\Rightarrow 2 \times 2$

$\Rightarrow$ 4 sq units

$$(68 - \underbrace{66}_{\text{mean}}) \times (72 - \underbrace{70}_{\text{mean}}) = 4$$



① $(68-66)\times(72-70) = 4$

② $(61-66)\times(72-70)$

$\hookrightarrow -10$

③ $(64-66)\times(67-70)$

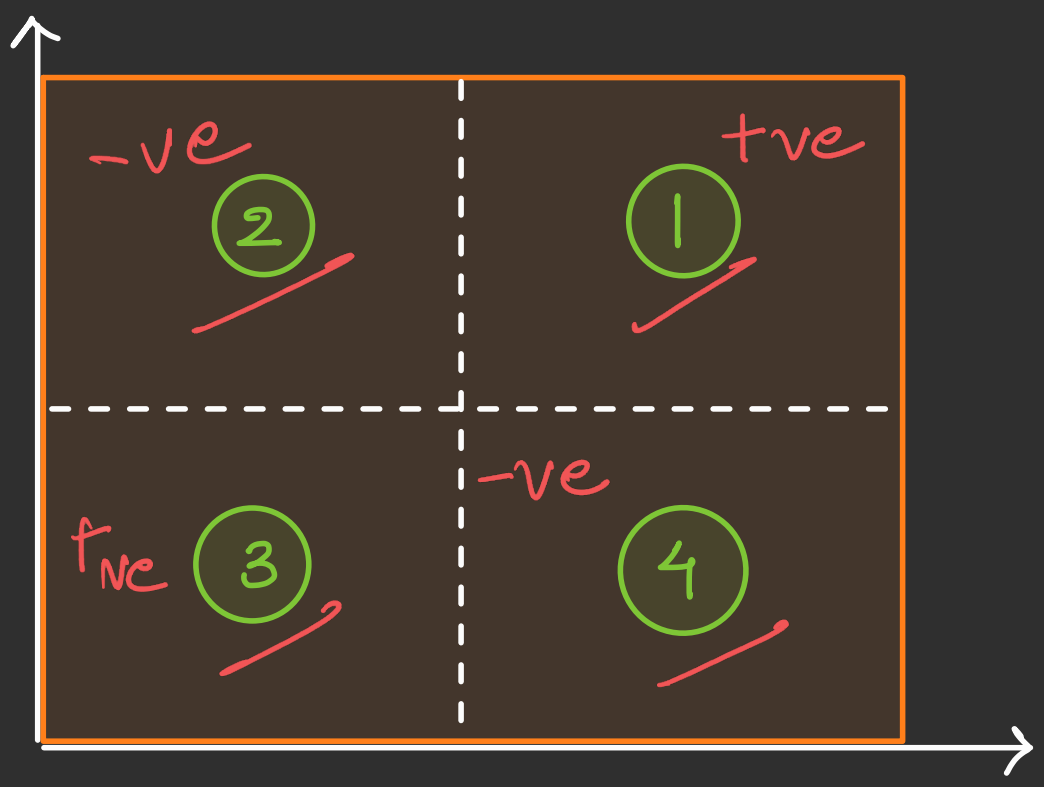
$\hookrightarrow 6$

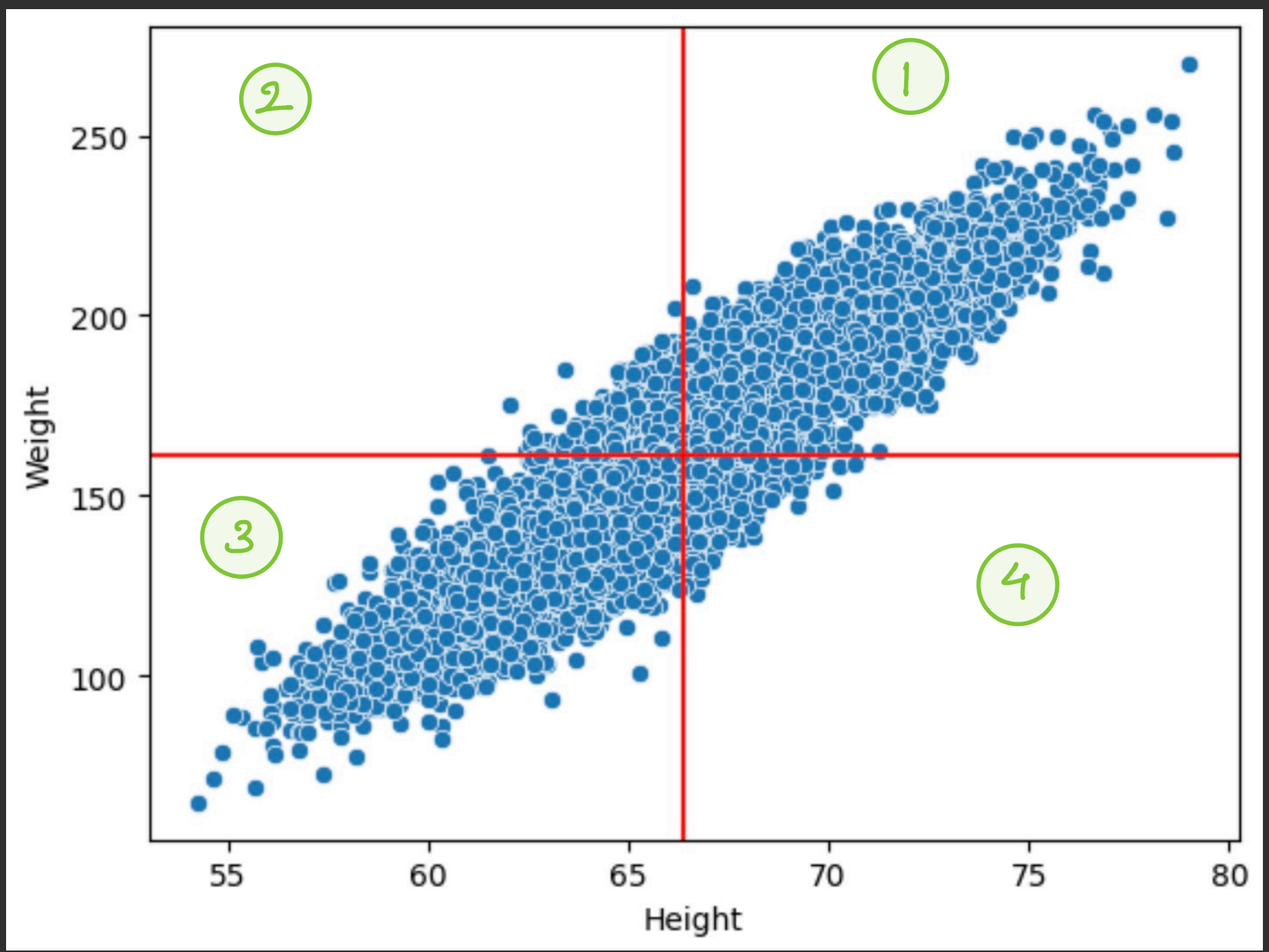
④ $(62-66)\times(58-70)$

$\hookrightarrow 48$

$(68-66)\times(64-70)$

$+ve(2) \times -6 = -12$


-ve
+ve
2
1
-ve
f_NC
3
4


2
1
250
200
150
100
3
4
55
60
65
70
75
80
Height

Average area for the points falling in area (1,3) will be positive

↳ Also known as "positively correlated"

Average area for the points falling in area (2,4) will be negative

↳ Also known as "negatively correlated"

1. $(68-66) \times (72-70) = 2 \times 2 = 4$
2. $(62-66) \times (58-70) = (-4) \times (-12) = 48$
3. $(64-66) \times (67-70) = (-2) \times (-3) = 6$
4. $(61-66) \times (72-70) = (-5) \times (2) = -10$

$[4, 48, 6, -10] \longrightarrow$ Avg (12)

↳ Avg [Covariance]

Formula for Covariance

$$cov(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}) \times (y_i - \bar{y})$$

x dtpt

$n \longrightarrow$ Total no of dtpts

mean

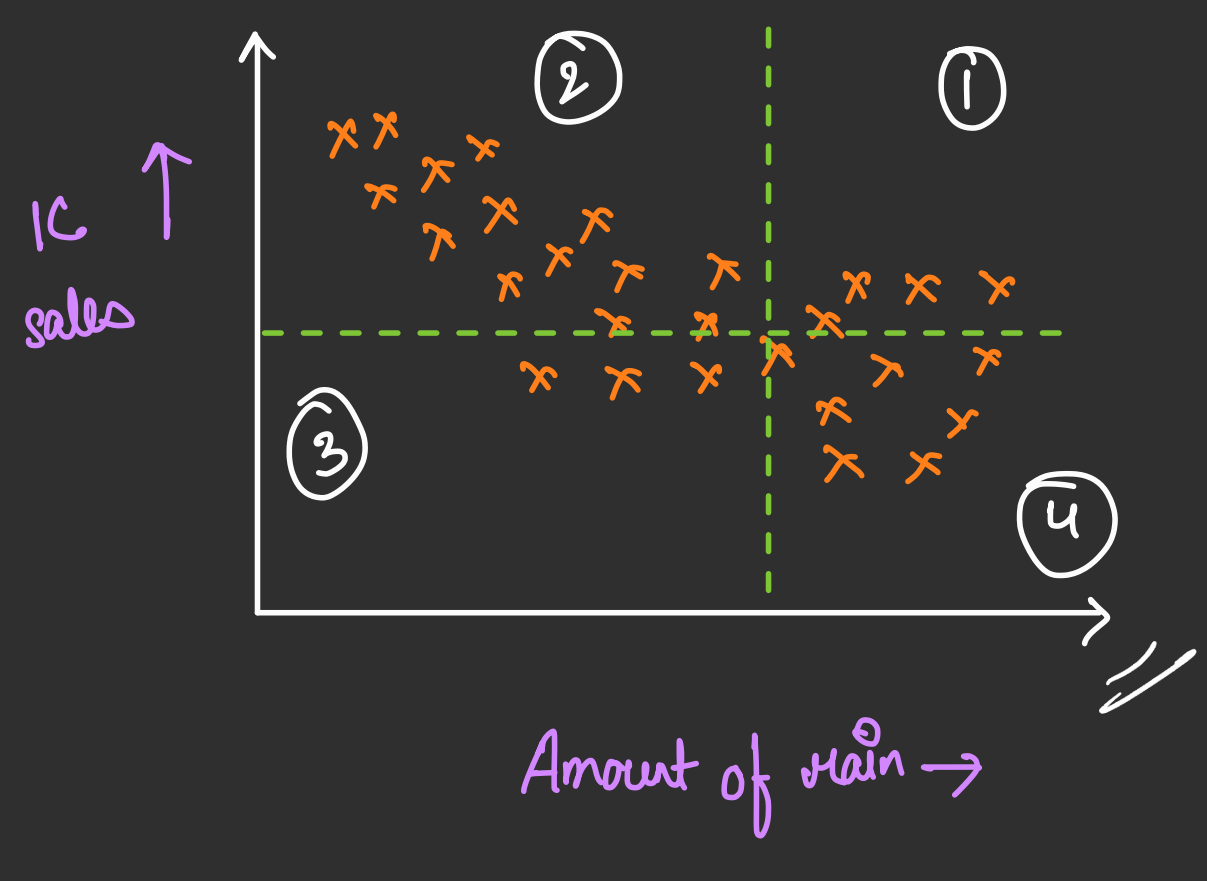
mean

y dtpt

Covariance is a mathematicaly entity to understand correlation

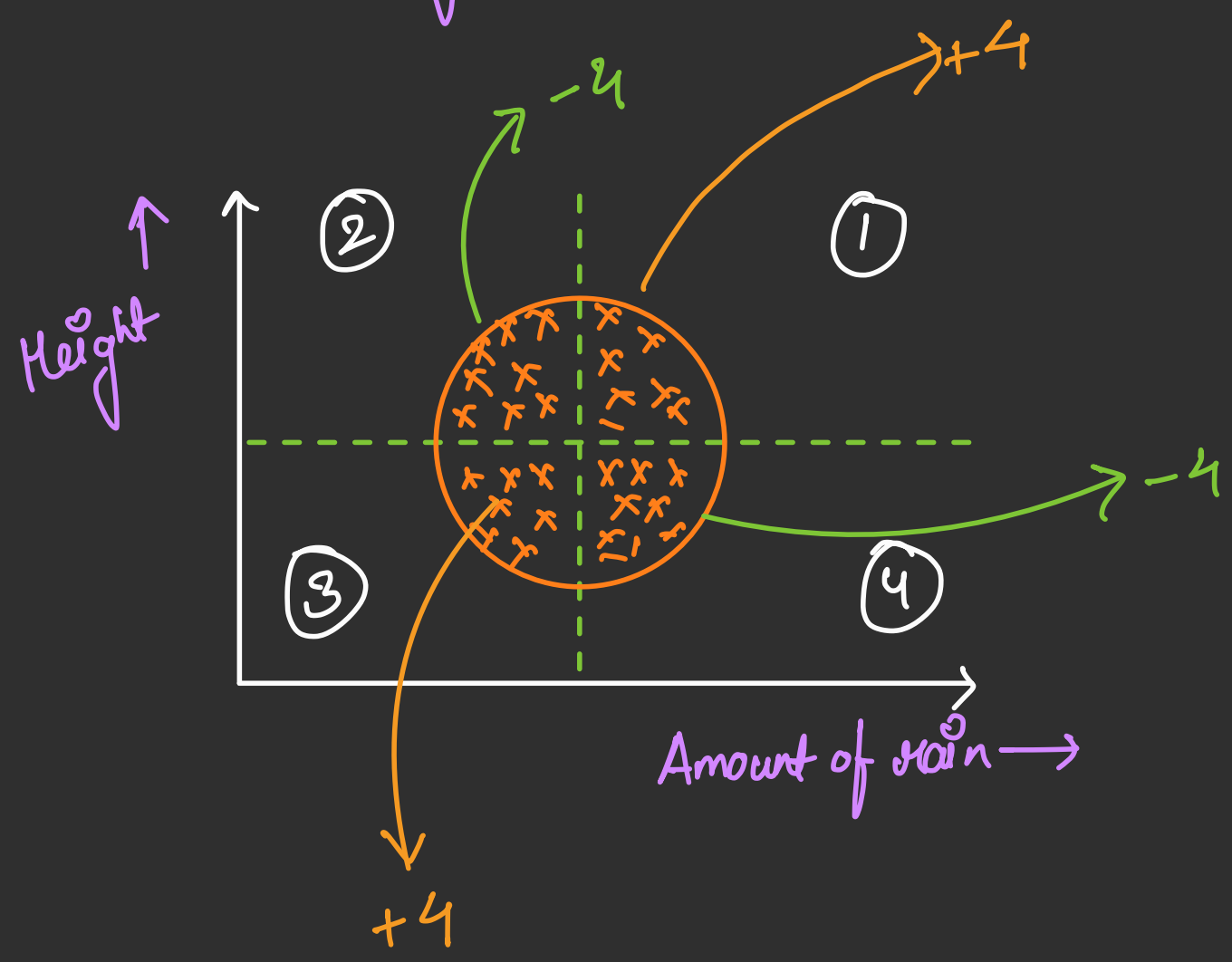
Co variance $\longrightarrow$ $x\uparrow$ $y\uparrow$

Example 2: Ice cream sales Vs Amount of Rain



$cov(x, y) \rightarrow$ -ve number

Example 3: Height Vs Rain



$Cov(x, y) \simeq 0$ ..

uncorrelated

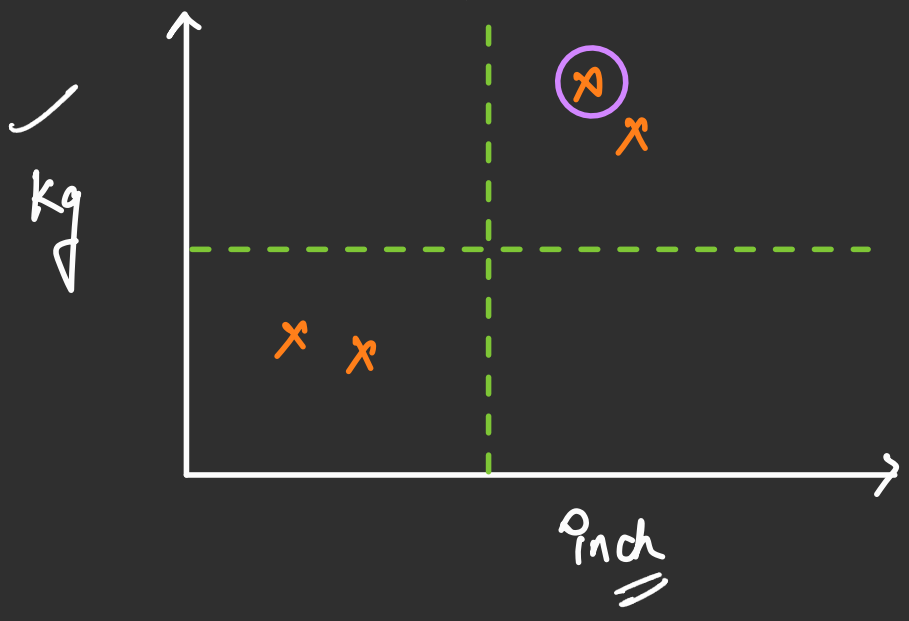
No correlation

Pearson Correlation

Pearson Correlation

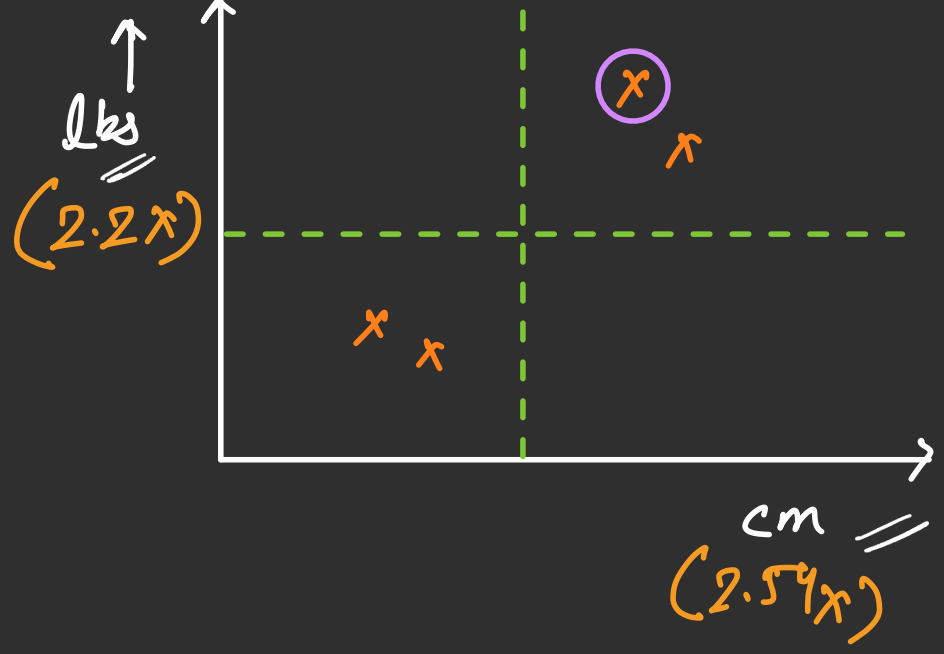
Lets take the example of Height & Weight

① Inch & kg (S_1)



(1 inch = 2.54 cm)

② Cm & pound (S_2)



(1 kg = 2.2 pounds)

$$\underset{s_1}{cov}(x,y) = (2.54)(2.2) \times Cov_{s_2}(x,y)$$

$X_{cm} = 2.54\ X_{inch}$

$Y_{lb} = 2.2\ Y_{kg}$

$Cov_{s_1}(x,y) = \underline{2.39}$

$Cov_{s_2} = (2.54) \times (2.2)$

$Cov_{s_2}(x,y) = \underline{48.39}$

$= 5.58 \times Cov_{s_1}$

S1

H (inch)	W Kg

men ne

inch kg

S2

H (cm)	W (pounds)

cm pound

Pearson Correlation

$$PC = \frac{cov(x,y)}{\sigma_x \sigma_y}$$

where σ_x & σ_y are std of (x) & (y) → Two cols

C (rho)

$$\underline{PC} = \frac{1}{n} \sum \left(\frac{x_i - \bar{x}}{\sigma_x} \right) \times \left(\frac{y_i - \bar{y}}{\sigma_y} \right)$$

Zscore

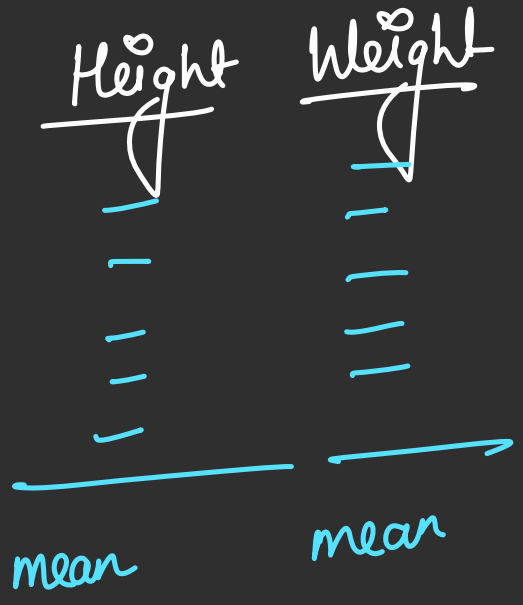
"Independent of scale"

→ Zscore

$$[-1 \leq \rho \leq 1]$$

np.cov

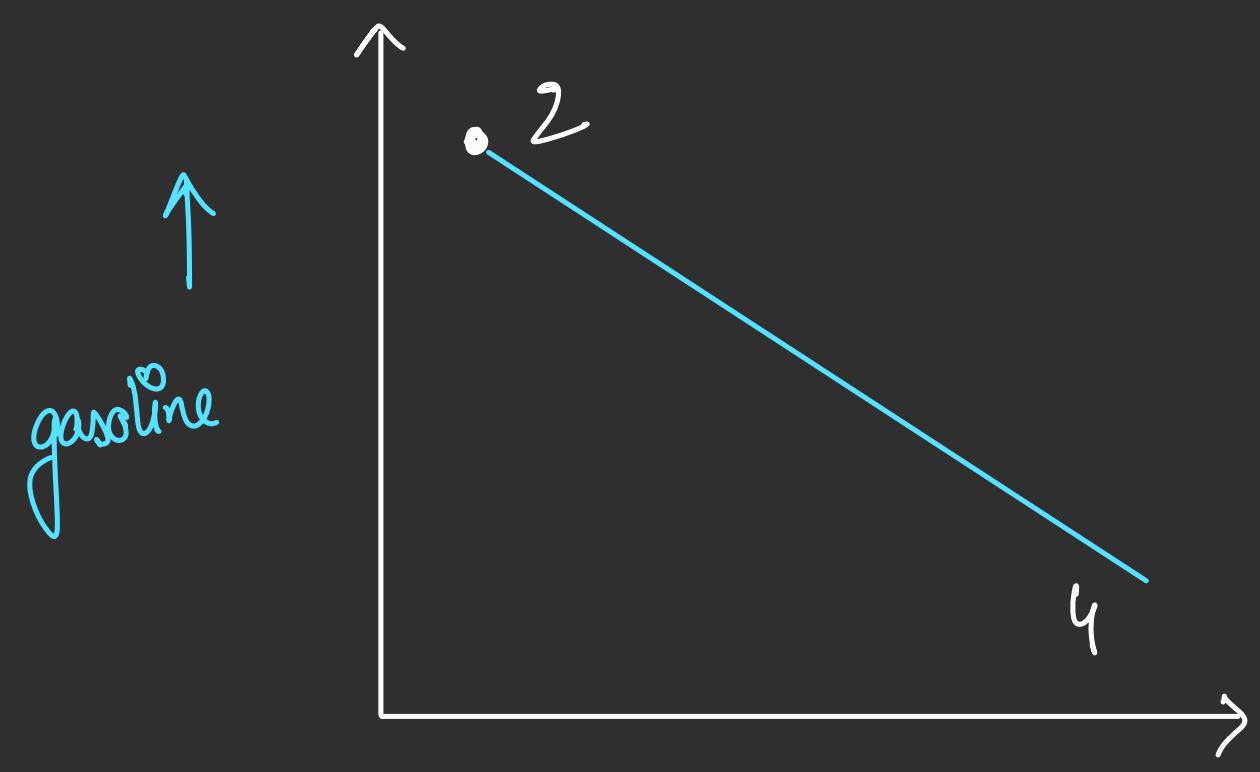
Two Col



$$cov(x,y) = \sum_{i=1}^{n} \frac{(x_i - \bar{x}) \times (y_i - \bar{y})}{n}$$

$$np.cov = \begin{bmatrix} var(x) & cov(y,x) \\ cov(x,y) & var(y) \end{bmatrix}$$

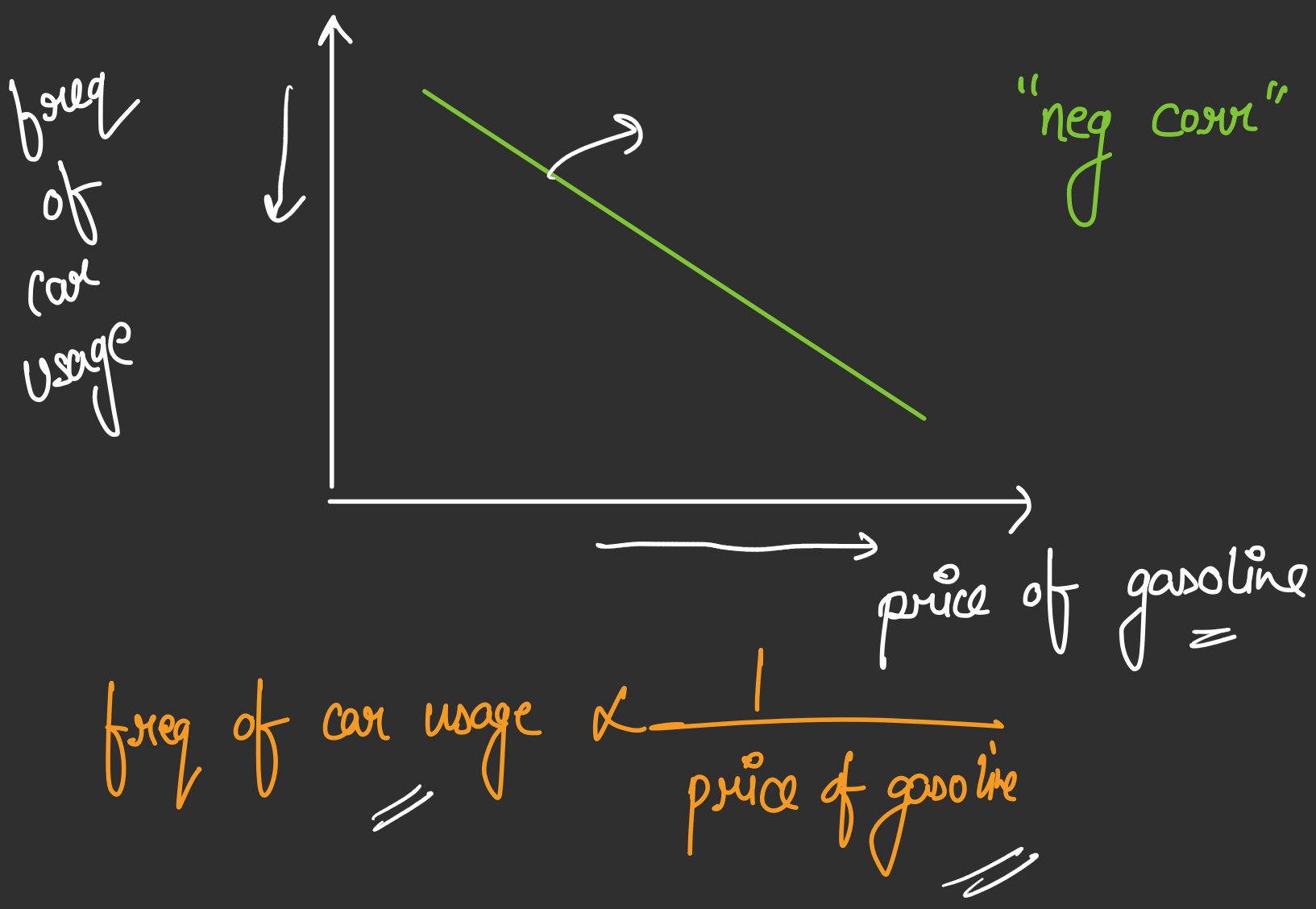
$$cov(x,y) = cov(y,x)$$



$$\text{cov}(x,y) = -\text{ve}$$

"negatively correlated"

Quiz



"neg corr"

freq of car usage $\propto \frac{1}{\text{price of gasoline}}$

X is the reason for change in Y

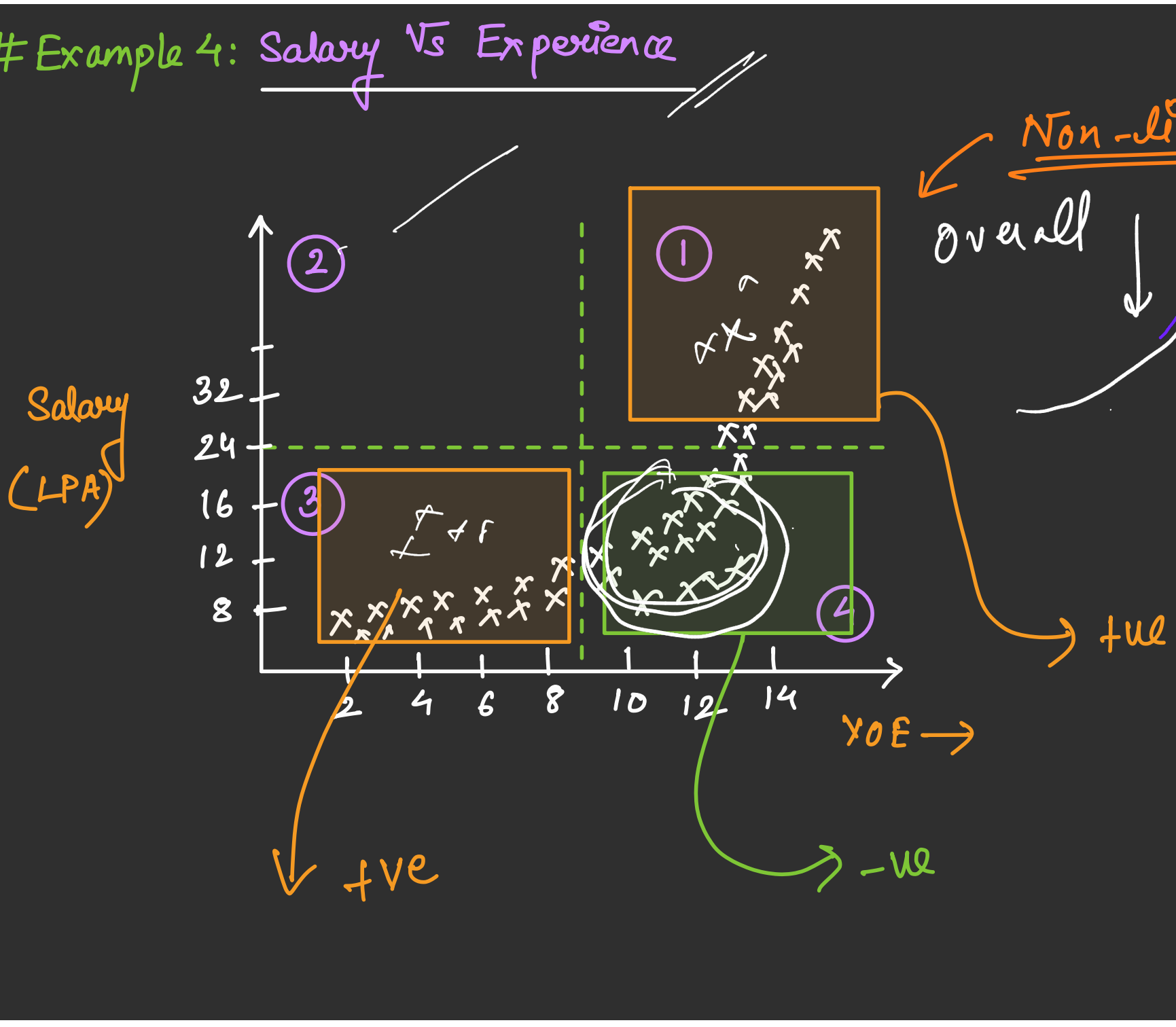
1) studying hours ↑ & marks ↑

2) Icecream sales ↑ & drowning ↑

⇒ This is not a causation

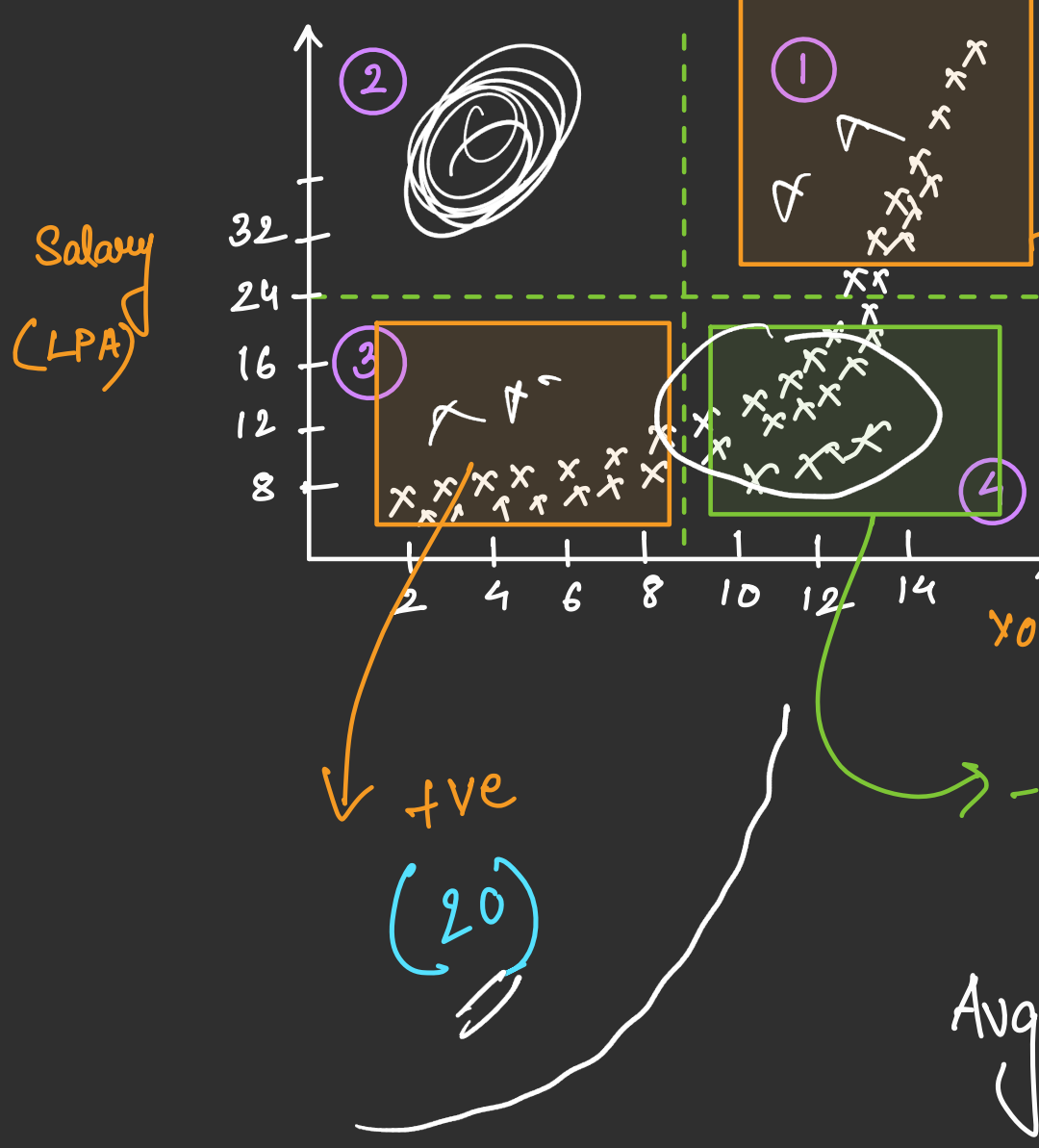
Spearman Correlation

Example 4: Salary Vs Experience

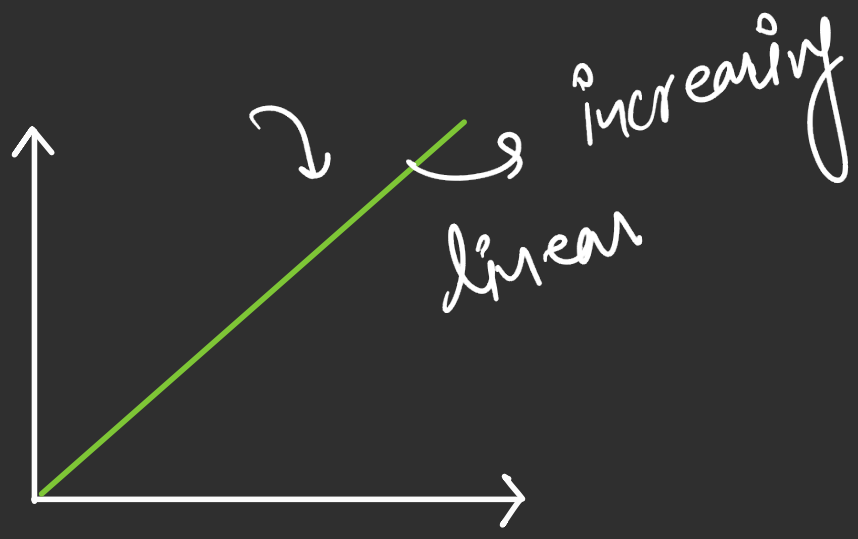


Non-linear

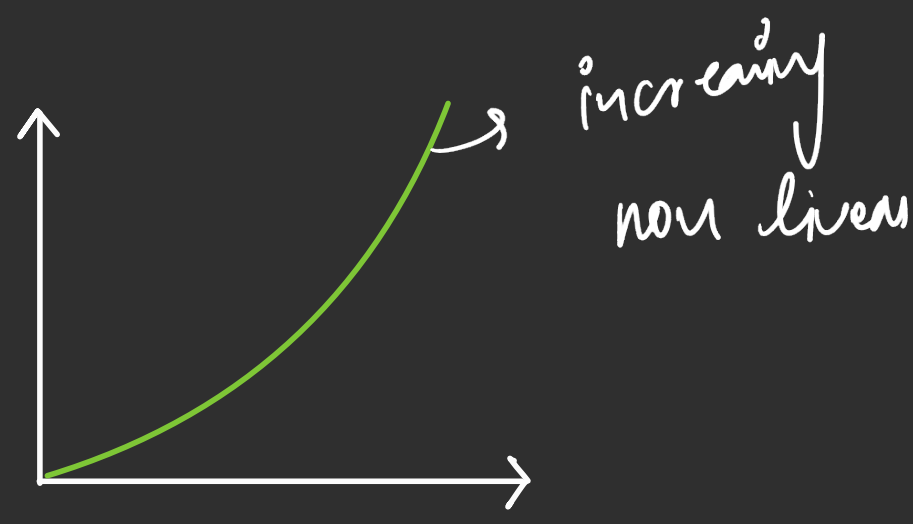
$$\rho = \frac{cov(x,y)}{\sigma_x \cdot \sigma_y}$$



Avg (10, 20, -15) ⇒ $\frac{+15}{3}$ (5)



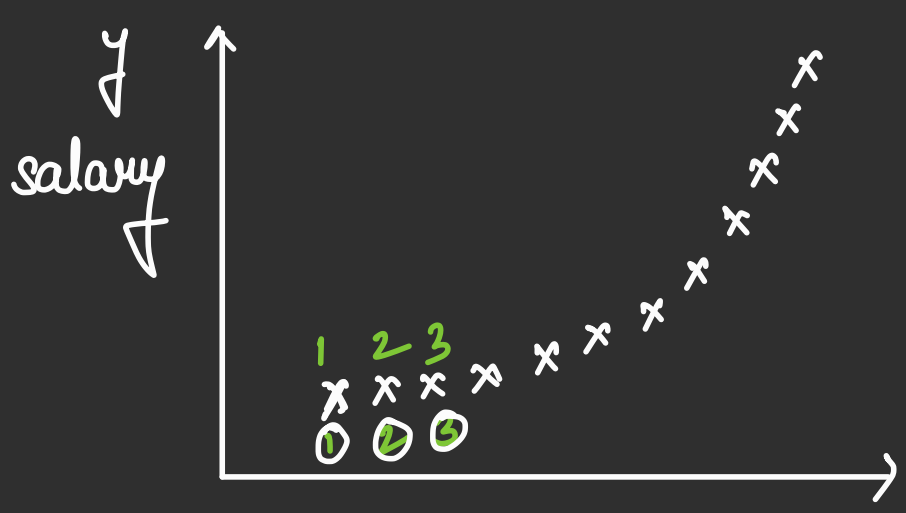
$y = x$ (linear)



$y = e^x$ (Non-linear)

Pearson Correlation ⟶ Cannot Handle Non-linear data

Spearman Correlation



x	y
32	20
42	36
100	48
46	31

Rank x	Rank y
1	1
2	2
3	3
4	4
⋮	⋮
10	10

linear

Example

Students	Maths	Science
A	35	24
B	20	35
C	49	39
D	44	48
E	30	45

difference in ranks

$$SC = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

Spearman Correlation

Students	Maths	Ranks M	Science	Ranks S	Difference d	$\|d\|^2$
A	35	3	24	5	−2	4
B	20	5	35	4	1	1
C	49	1	39	3	−2	4
D	44	2	48	1	1	1
E	30	4	45	2	2	4

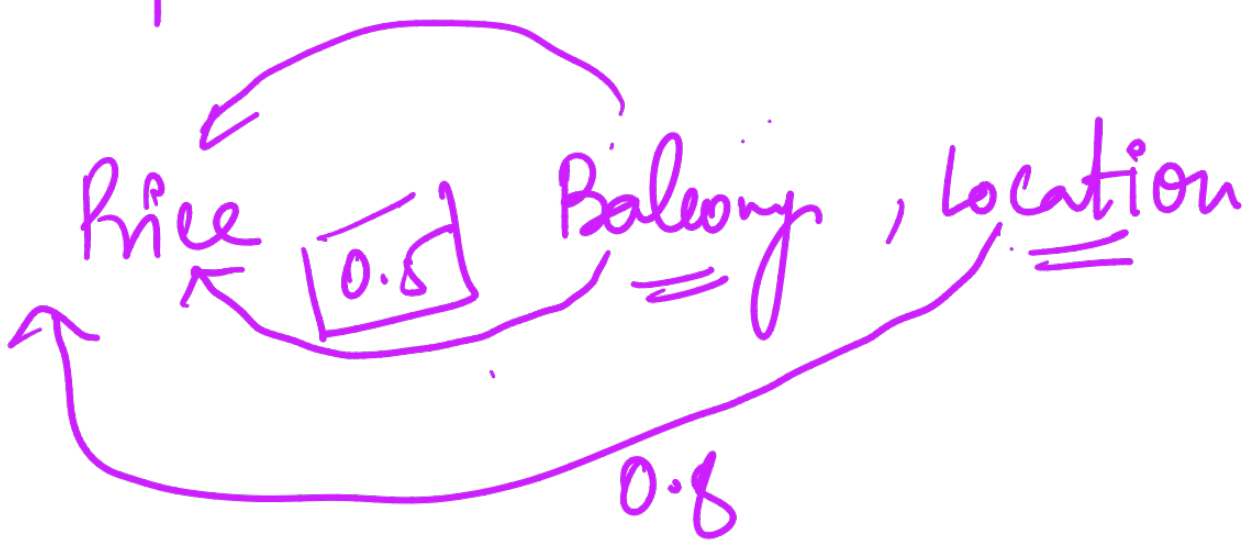
14

$$SC = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \Rightarrow 1 - \frac{6\times(14)}{5(25-1)} \Rightarrow 0.32$$

non linear data $\xrightarrow{\text{rank}}$ linear

↳ Pearson correlation

$$\text{Spearman}(X, Y) = \text{Pearson}(\text{rank}(X), \text{rank}(Y))$$



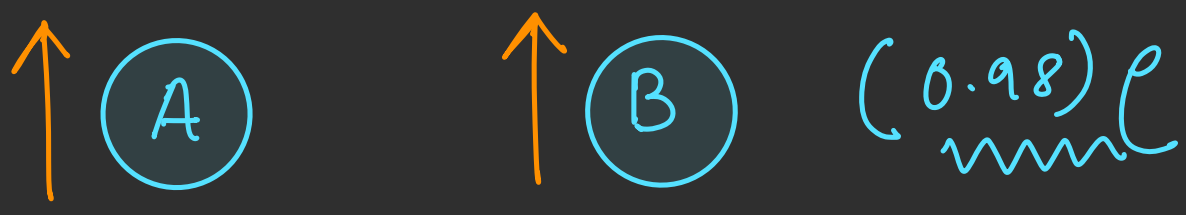
norm. cdf(z) = P(z ≤ z)

↳ left tail

Right tail 1 − norm.cdf(z)

→ norm.cdf(abs(z)) → p value

p = 2 × P(z > |z|) ↳ Two tailed tried


A
B
(0.98)c

pearson corr

1) Covariance

2) $C = \frac{Cov(x,y)}{\sigma_x \cdot \sigma_y}$

spear corr

1) $sc = 1 - \frac{6\sum d^2}{n(n^2-1)}$

(Non-linear data as well as linear data)

linear data

Nominal Data

35

24

36

22

Ordinal Data

2

3

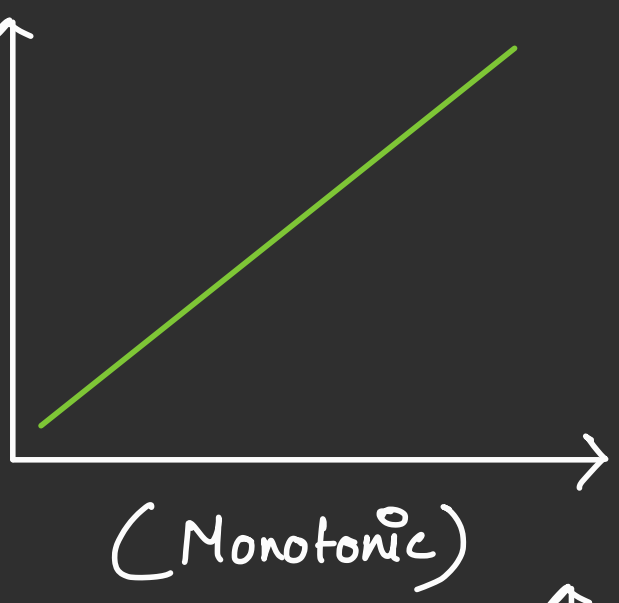
1

4

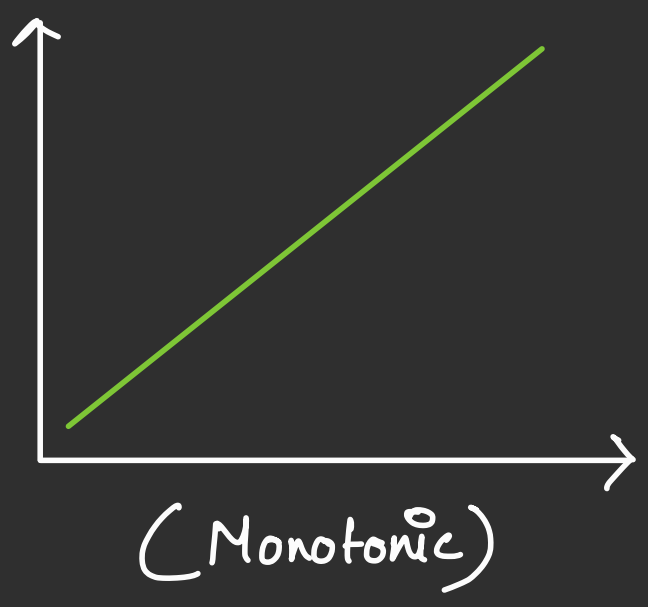
Rank

(Ordered data)

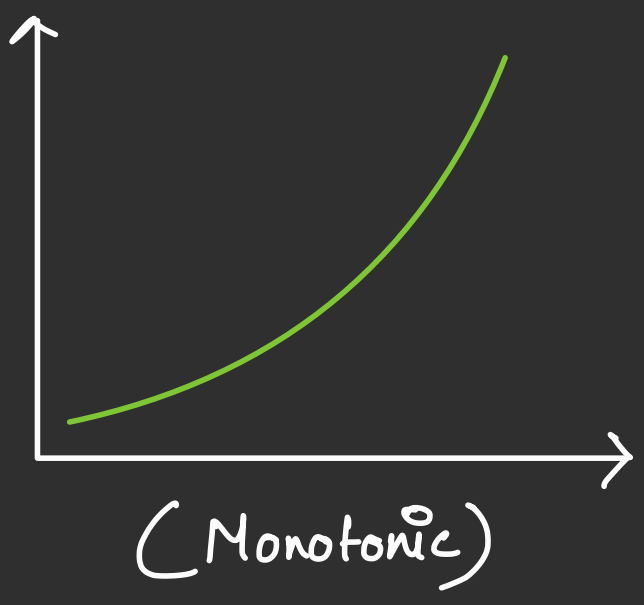
Monotonic Vs Non-Monotonic



(Monotonic)



(Monotonic)



(Monotonic)

(Any graph that is strictly increasing or Monotonic decreasing)

(Non Monotonic)