Correlation -

⇒ 1) Correlation
- Pearsom
- Spearman
- Viz of Correlation

⇒ Relationship b/w 2 Variables.

---

1) Num vs Categorial → T test (= 2 Cat)
→ Anova Test (> 2 Cat)

2) Cat vs Cat → Chi Square

3) Num vs Num → Correlation

⇒ T Test → IQ score of S1 to IQ score of S2

ANOVA

→ Not Multiple Numeric field

→ But multiple groups of same numeric field.

⇒ Num vs Num →

⇒ Height vs Weight →

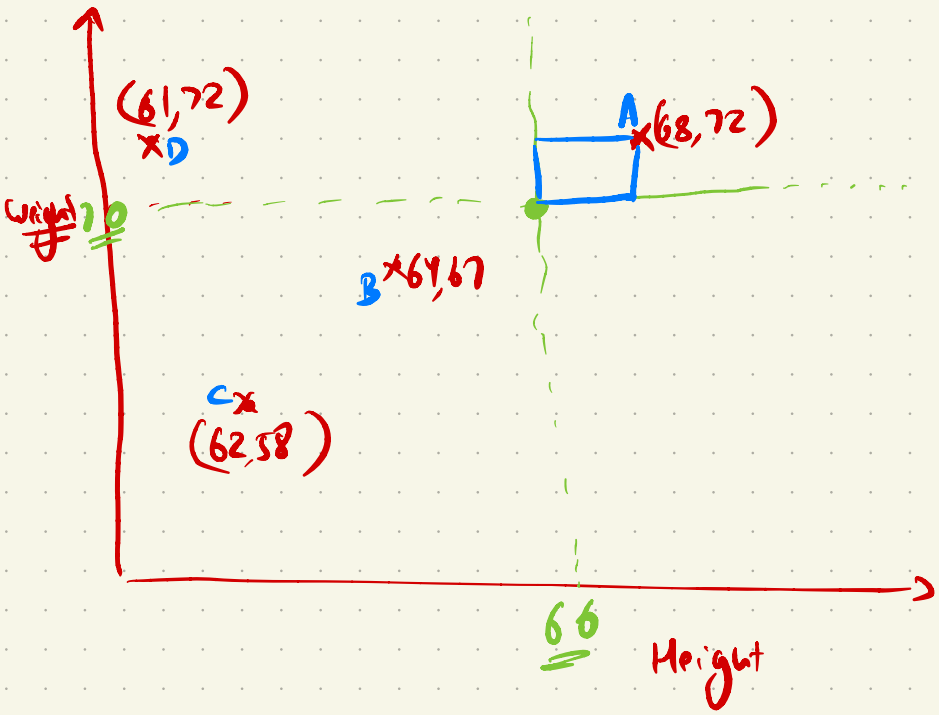
Height (inches)	Weight (kg)	
68	72	
62	58	
64	67	
61	72	
70	79	
66	61	
61	68	
65	64	
71	80	
72	79	
$\bar{h} = 66$	$\bar{w} = 70$	

Height & weight are related

Height → Numeric variable

weight → Num variable

⇒ Num vs Num relationship



A ⇒ $(68-66)(72-70) = 2 \times 2 = 4$

B ⇒ $(64-66)\ (67-70) = -2 \times -3 = 6$

C ⇒ $(62-66)\ (58-70) = -4 \times -12 = 48$

D ⇒ $(61-66)\ (72-70) = -5 \times 2 = -10$

Area → $(X-\mu_x)(Y-\mu_y)$ → +ve / -ve

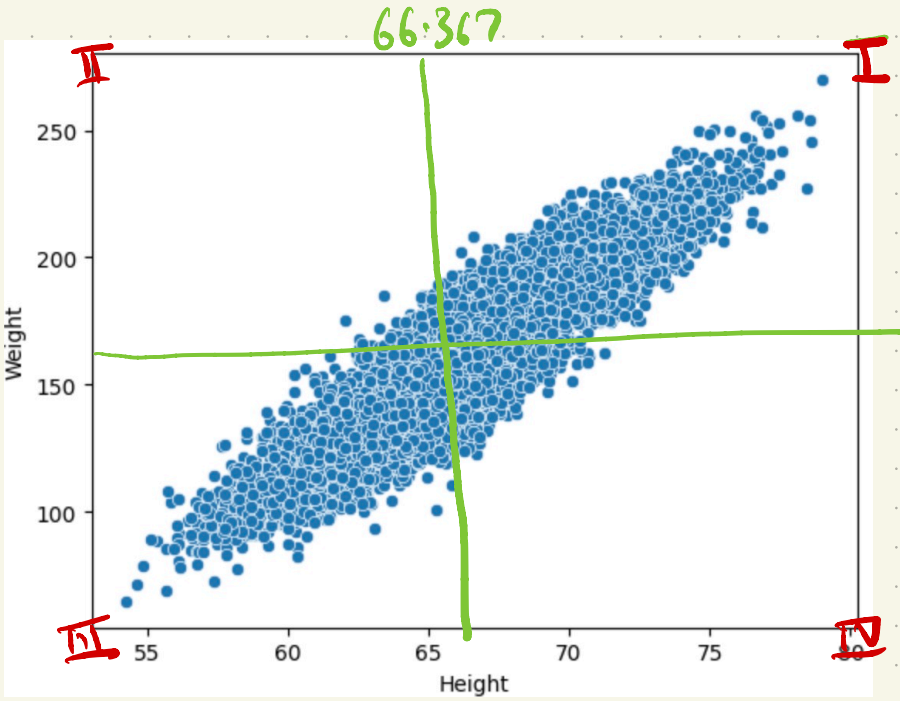
⇒ +ve area → +ve relationship

or Height increases when weight increases

-ve area → -ve relationship

or Height decreases when weight increases

⇒ Area can be added → slopes can't be added.



I → Area → +ve

III → Area → +ve

II → Area → -ve

IV → Area → -ve

O

# of I, III have >> # of points II & IV have

$$\Sigma (X-\mu_x)(Y-\mu_y)$$

Positive Correlation

Co variance

$$\frac{\sum_{i=1}^{n} (h_i - \bar{h})(\omega_i - \bar{\omega})}{n}$$

$\Rightarrow$ +ve number

$\Rightarrow$ +ve if more pts in Ist & IInd Quadrant

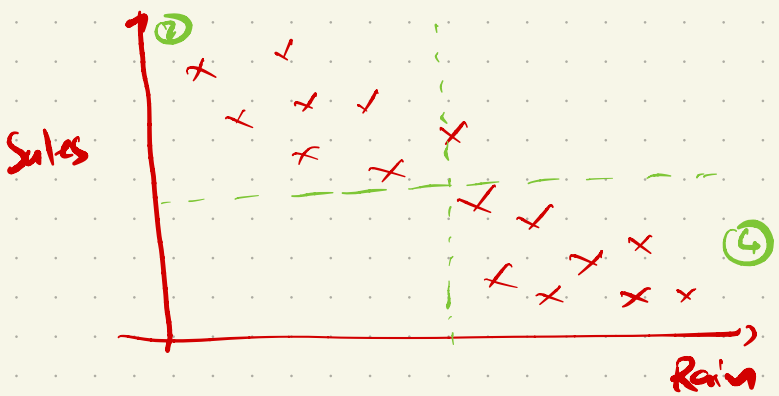
$$Cov(X, Y) = E\left[(X - E(X))(Y - E(Y))\right]$$

$E(X)$ = avg of all X pts

$E(Y)$ = avg of all Y pts

$$E\left([X - E(X)][Y - E(X)]\right) = \frac{\sum (x_i - \mu_x)(y_i - \mu_y)}{n}$$

$\Rightarrow$ Ice Cream Sales vs Amount of Rain



Negative Correlation

⇒ Most data will lie in II & IV Quardant →

So Covariance → ~~+ve~~

-ve ✓✓

Rainfall & Ice cream Sales are -ve related.

⇒ Height vs Rainfall

UNCorrelated

No Correlation

Covariance → Sum of areas

positive area ≈ negative area

Covariance ≈ 0

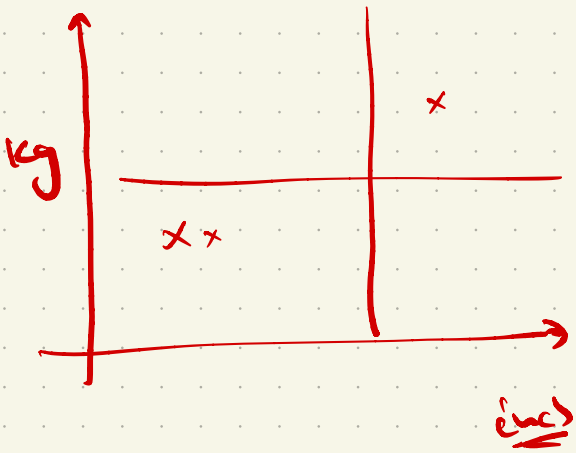
⇒

Pearson Correlation →

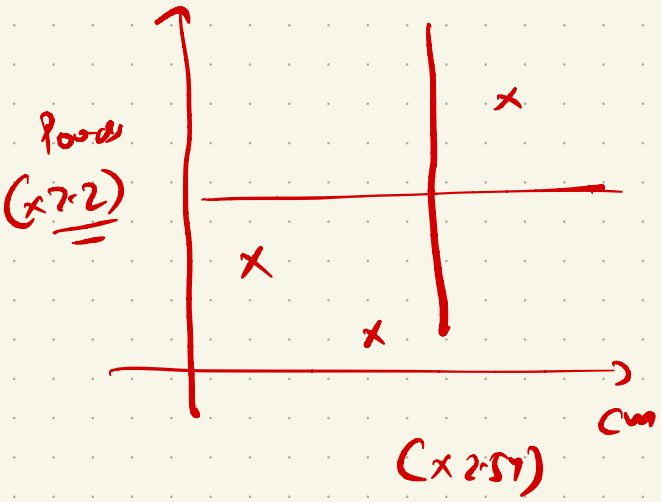
1 kg ≈ 2.2 pound
1 inch ≈ 2.54 cms

Covariance

① Inch vs kgs



② cms vs Pounds



$$(h - \bar{h})(\omega - \bar{\omega})$$

(3) (2)

h = 69 in
$\bar{h}$ = 66 in
ω = 72 kg
$\bar{\omega}$ = 70 kg

h = 69 × 2.54
$\bar{h}$ = 66 × 2.54
ω = 72 × 2.2
$\bar{\omega}$ = 70 × 2.2

(3 × 2.54)(2 × 2.2)

Covariance → dependent on the units →

Soln 1

$$\frac{(h-\bar{h})}{h}\frac{(\omega-\bar{\omega})}{\bar{\omega}}$$

$$\left(\frac{h}{\bar{h}}-1\right)\left(\frac{\omega}{\bar{\omega}}-1\right)$$

Soln 2

$$\frac{(h-\bar{h})}{\sigma_h}\quad\frac{(\omega-\bar{\omega})}{\sigma_\omega}$$

$$\boxed{Corr = \frac{Covariance(x,y)}{\sigma_x\ \sigma_y}}$$

$$\rho\ (rho) = \frac{1}{n}\sum\frac{(h_i-\bar{h})}{\sigma_h}\frac{(\omega_i-\bar{\omega})}{\sigma_\omega}$$

↳ Correlation Coffecient

Pearson's Correlation

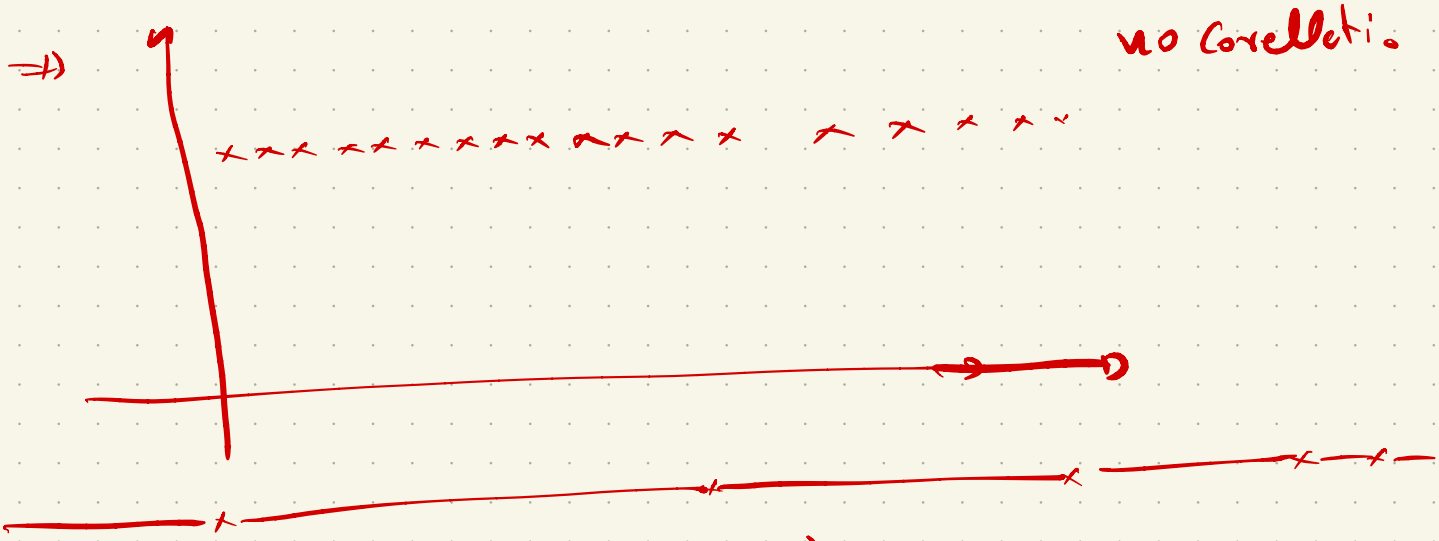
$$\boxed{-1 \le \rho \le 1}$$

$\rho \simeq -1 \Rightarrow$ High -ve Correlat

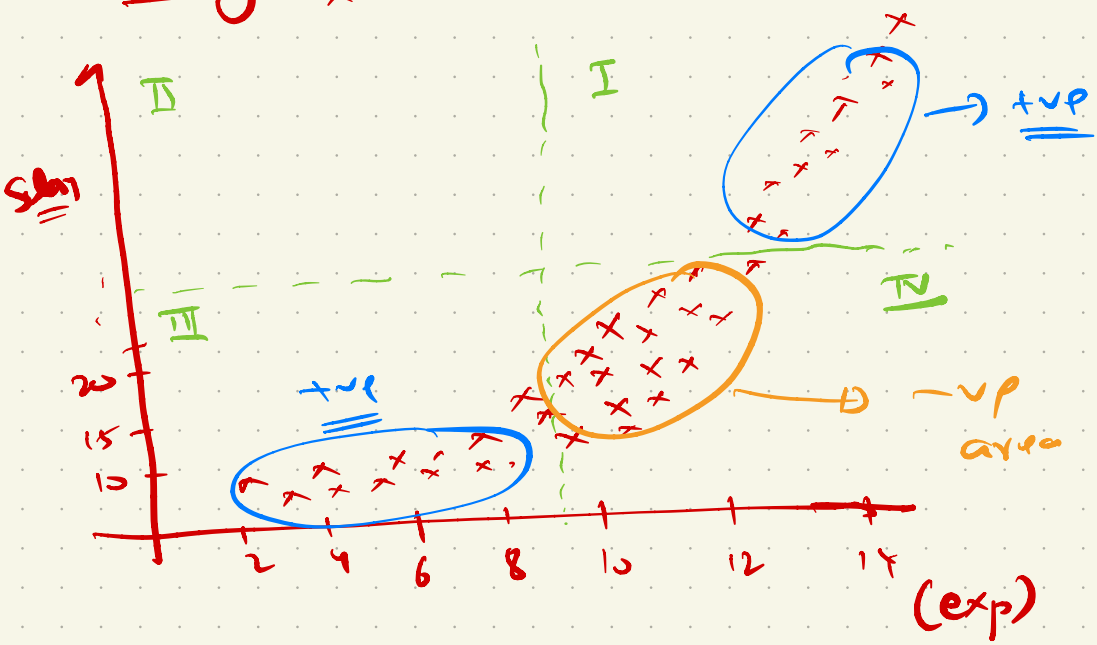
$\rho \simeq 1 \Rightarrow$ High +ve Correlat

$\rho \approx 0$ ⇒ no Correlation

Corr (ρ) is a unit free measurement →



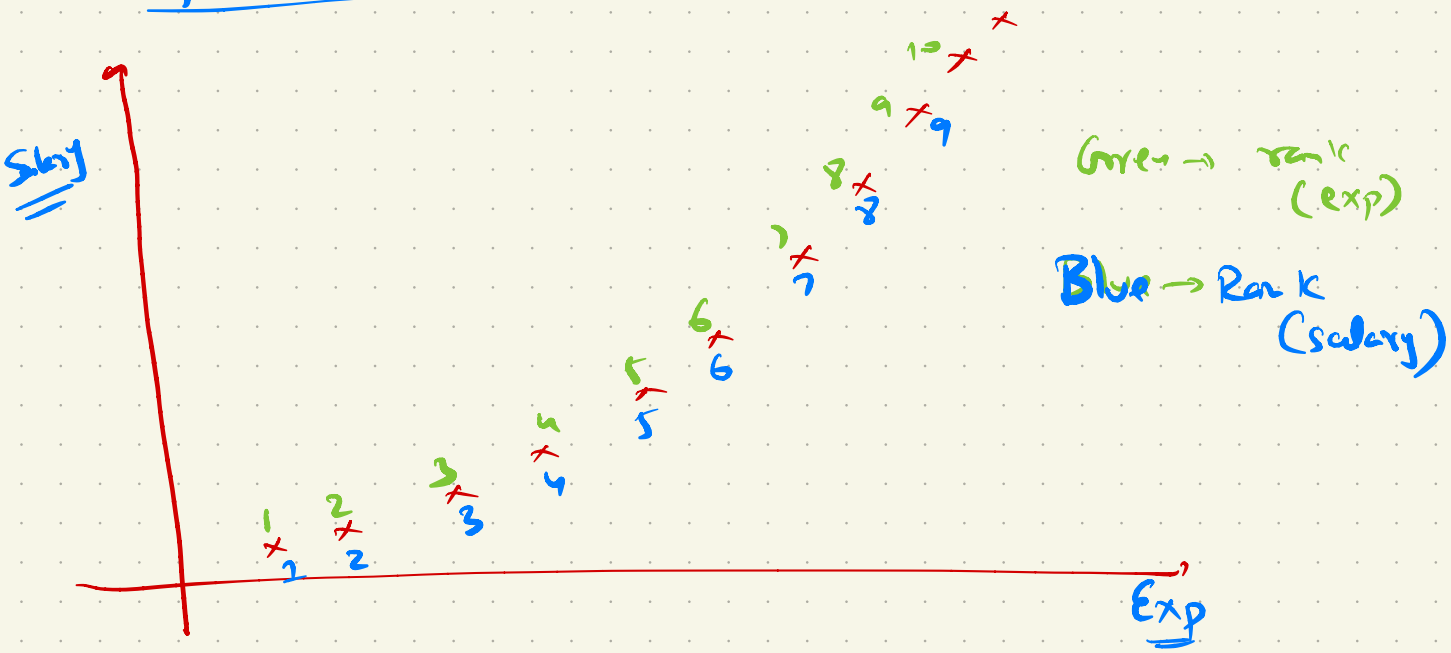
⇒ Salary vs Experience →



⇒ Total area may be ≈ 0 or -ve

⇒ Pearson's Correlation → works very well with linear dependencies

⇒ Spearman Correlation →



Rank(x)	Rank(y)
1	1
2	2
3	3
4	4
5	5
6	6

Spearman coefficient

$$\rho_{rank_x, rank_y}$$

$$\rho \simeq 1$$ →

⇒ Spearman would be high

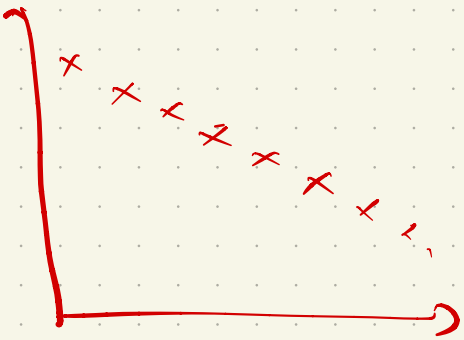
⇒ Pearson Could be low or even -ve.

$\rho \approx 0$

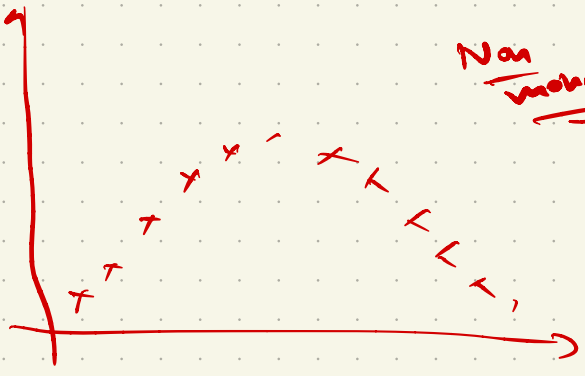
Both Pearson & Spearman $\rho \approx 0$

Spearmen Correlation ⟶ Monotonic (increasing / decreasing).

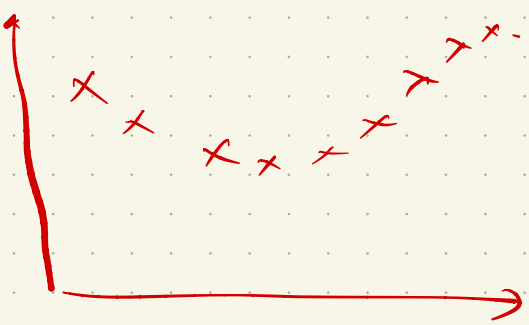
if x increases y should also increases →



Monotonic



Non monotonic



Ranks are +ve

Heights are +ve

Spearman → we have -ve corrd.

Rank x	Rank y
1	10
2	9
3	8
4	7
5	6

HW → Calculate Spearmen Correlation using basic python and .corr().

⇒ a way to get rank → ??

⇒ Hypothesis framework

H_0 → Fields are unrelated

H_a → one field is dependent on another.

Vizualize Correlation →

⇒ Parametric → Some params are assumed

Non parametric →

⇒ Pearson → Parametric

Correlation Test → Bivariate Normal distribution.

⇒ Spearman → Non parametric

→ on the basis of rank.

⇒ Both Height & weight follow a normal dist^n.

⇒ Bivariate → X, Y are Random variable

for every real no. a, b

such that

$$Z = aX + bY$$

Z → follows a normal distribution

p_value → tells that what are the chances that data is unrelated.

p_value ≈ 0 → data is Highly correlated

p_value < 0.05 → data is related

else

data is not related

Collab Link : https://colab.research.google.com/drive/1CTwBoTB7E11KoYiDGlugDfqlHFkxdZHw?usp=sharing