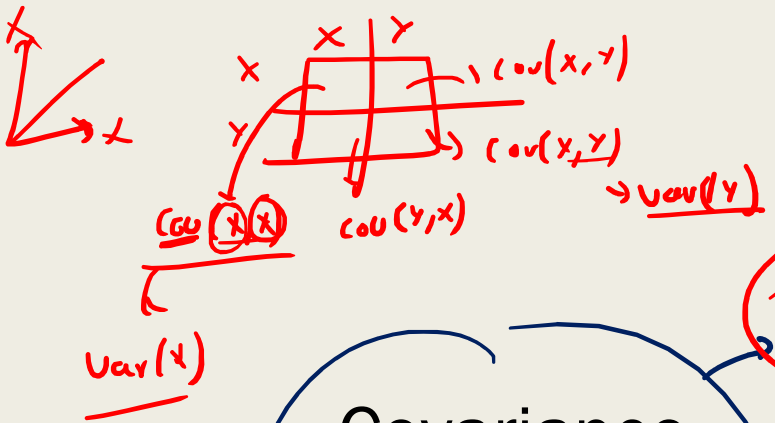
$\bar{X} \rightarrow \sigma \rightarrow cov(x, y)$

$x_1 \quad x_2 \quad x_3$

$x_1 x_2 \quad x_2 x_3 \quad x_3 x_1$ in pair always.

1 variable

Relationship b/w 2 variables



Covariance

dxn of the relationship

① $X \cdot N$ → +ve → $x\uparrow \; y\uparrow$ | $x\downarrow \; y\downarrow$

→ -ve → $x\uparrow \; y\downarrow$ | $x\downarrow \; y\uparrow$

→ 0 → No relationship / dxn

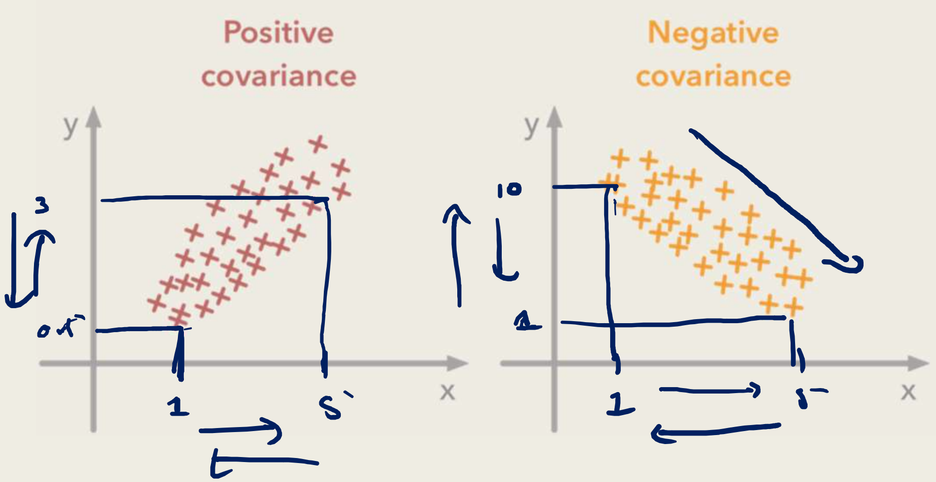
σ^2 (variance) $\Rightarrow \dfrac{\sum (x_i - \bar{x})^2}{N-1} \Rightarrow \dfrac{\sum (x_i - \bar{x})(x_i - \bar{x})}{N-1}$

Hack

$$\boxed{cov(x, y) \Rightarrow \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}}$$

Covariance

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

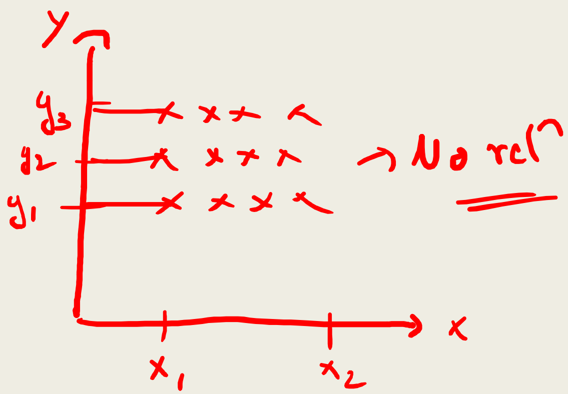


In Python: use *cov()* function

-2048

+ 1,10,111

magnitude does not matter!



$$Correlation = \frac{Cov(x,y)}{\sigma x * \sigma y}$$

Only dxn matters

cov(x,y) = 0 → No rel^n

cov(x,y) < 0 → -ve rel^n

cov(x,y) > 0 → +ve rel^n

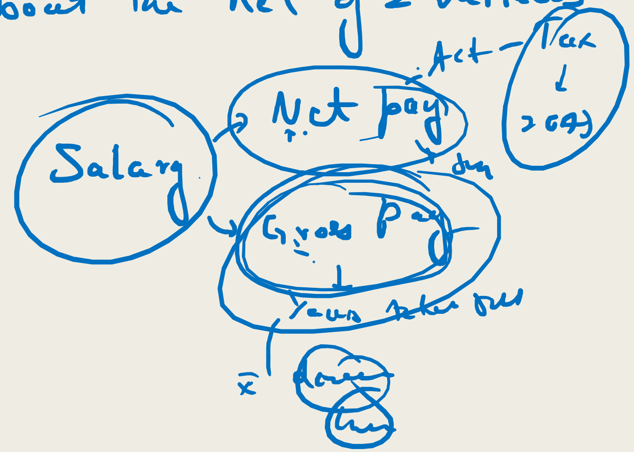
$\bar{x}$ → Central Tendency

σ^2 → spread

cov(x,y) → dxn of relⁿ

cor(x,y) → dxn + strength of relⁿ

Talks about the Relⁿ of 2 variables



dxn → sign (+, -)

[-1, +1]

Correlation

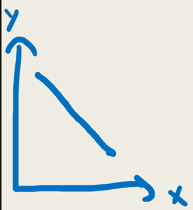
↳ strength → magnitude

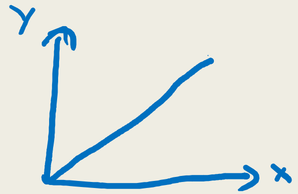
$$\text{Corr} = \frac{\text{cov}(x,y)}{\sigma_x \, \sigma_y} \Rightarrow \frac{\dfrac{\sum (x_i - \bar{x}) * (y_i - \bar{y})}{\cancel{n-1}}}{\sqrt{\dfrac{\sum (x_i - \bar{x})^2}{\cancel{n-1}}} * \sqrt{\dfrac{\sum (y_i - \bar{y})^2}{\cancel{n-1}}}}$$

$\sqrt{\sum x_i}$

$\sqrt{x_1 + x_2 + x_3}$

$\sqrt{x_1} + \sqrt{x_2} + \sqrt{x_3}$





strong -ve rel^n ← → No rel^n

→ strong +ve rel^n

-1 -0.75 -0.1 0 0.1 0.75 +1

weak -ve rel   weak +ve rel^n   -1 - 0 - 1

What is Correlation?

0.15
0.1
0.06
0.2

Correlation is a statistical technique to depict the relationship between 2 variables – strength and direction. We measure the correlation with the help of Correlation Coefficient.

Best Practice   decid share → 0.75

For example, height and weight are related; taller people tend to be heavier than shorter people.

0.5   06   0.72
50%   0.7
0.9

What is Correlation Coefficient?

→ linear Rel^n

The Pearson's correlation coefficient (*r*) is a measure that determines the degree to which the movement of two variables is associated. The value of Correlation Coefficient lies between -1 and 1.

Spearman coef of corr → Non-direct rel^n

Formula: (Pearson's Correlation Coefficient) - Standard Formula

In Python: *DataFrame.corr(method='pearson')*

↳ default

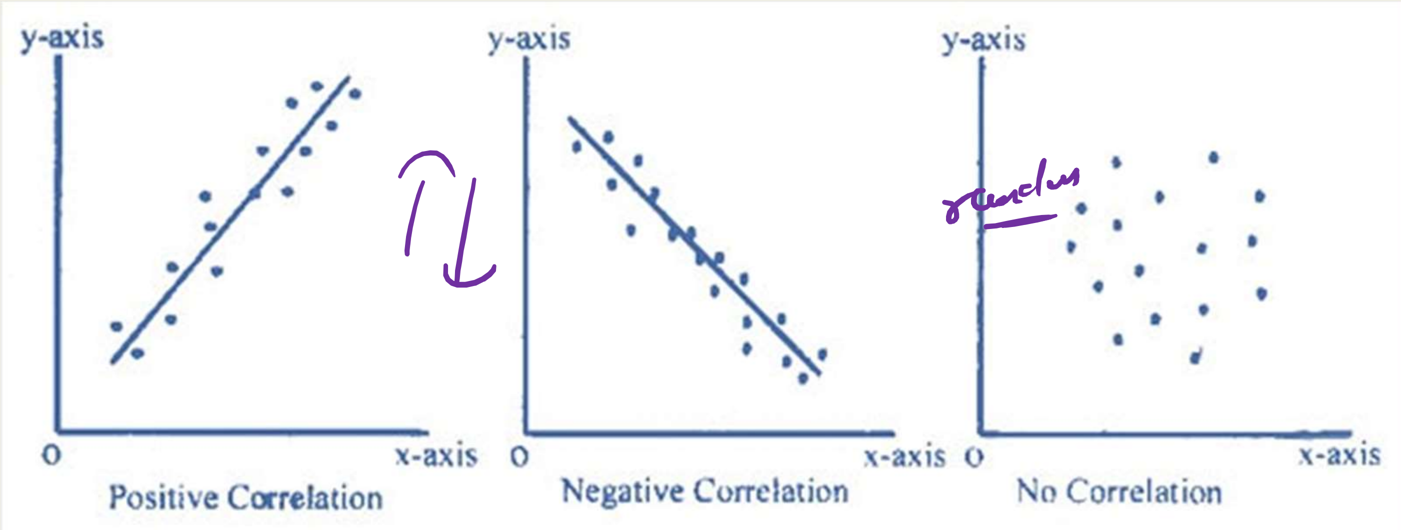
$$r = \frac{1}{(n-1)s_x s_y}\Sigma(x_i-\bar{x})(y_i-\bar{y})$$

(n = sample size, and Sx, Sy are the standard deviation of samples x and y. X-bar and y-bar are the respective means of x and y samples whereas Xi and Yi are sample points of X and Y respectively.)

Strength?

Positive and Negative Correlation:

1. Correlation Coefficient greater than zero indicates a *positive relationship*
2. while a value less than zero signifies a *negative relationship*
3. and a value of zero indicates no relationship between the two variables being compared.



Strong and Weak Correlation:

the closeness / packedness of the data points
↳ strength

Kind of correlation = depicted by sign of correlation coefficient
How Strong =. Value of Correlation Coefficient

Husband Wife

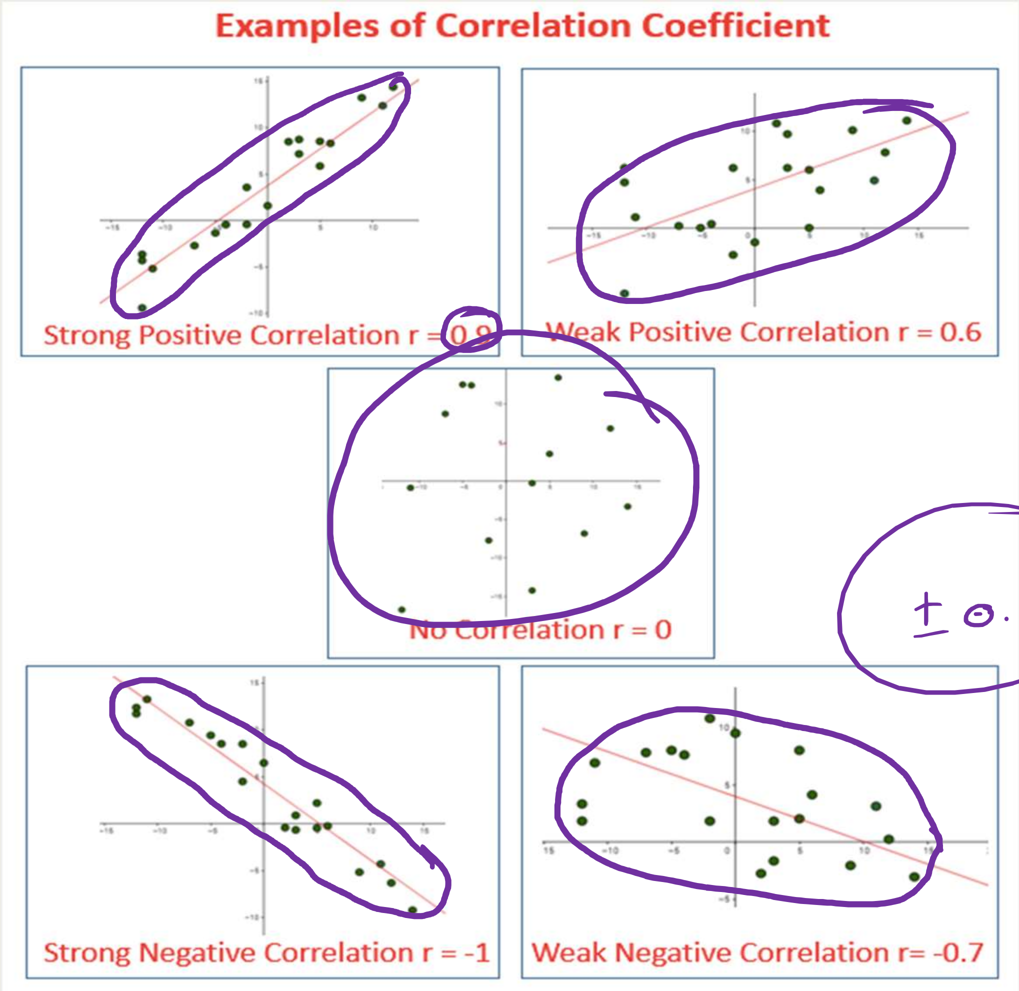
2 2 4
↳ 50%

3 3 4

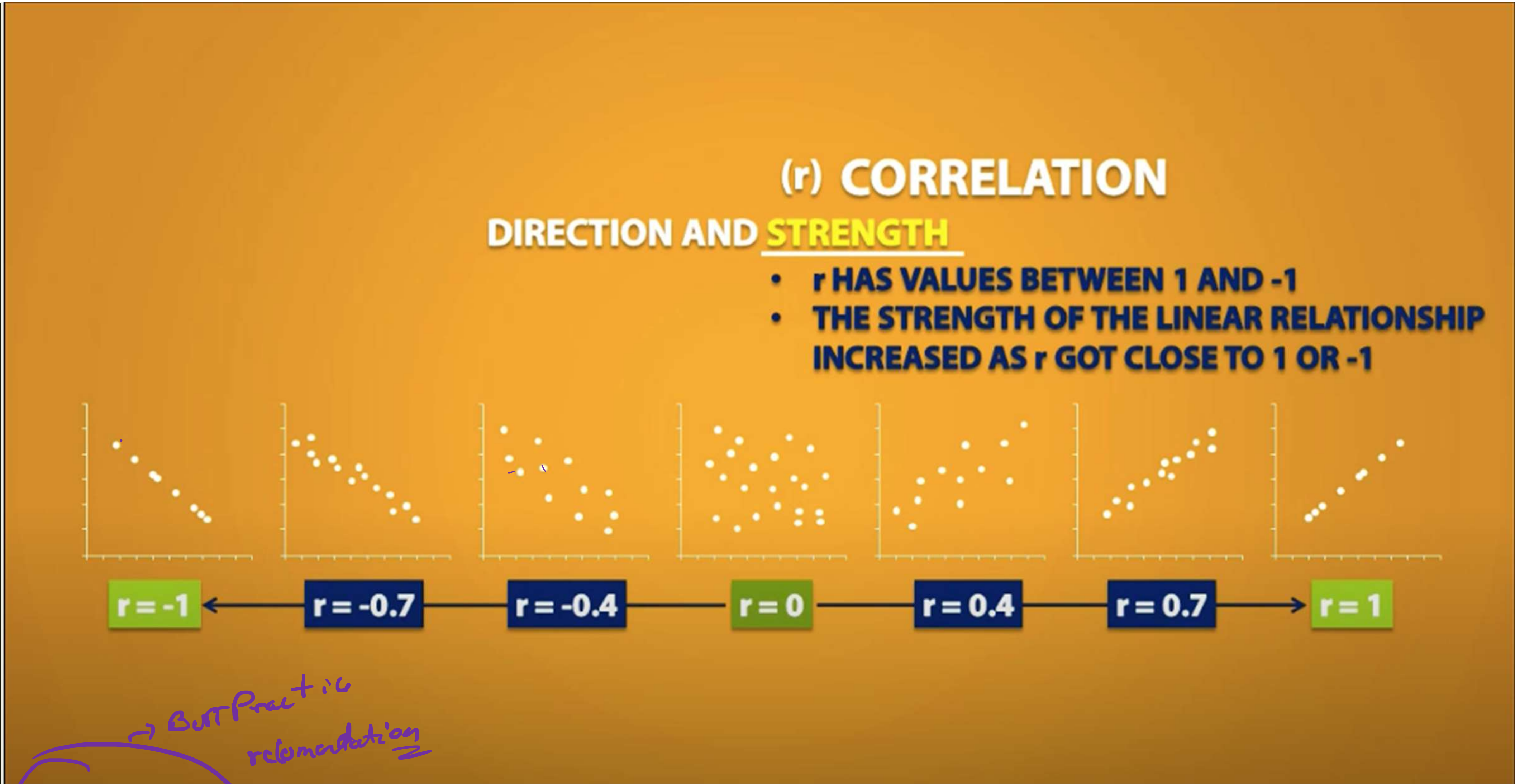
75%

Approach
Stats ≠ Maths
↳ 5≠6

5=6
↳ 2 weeks



±0.75



→ Best Practice recommendation

Rule of thumb: Any relationship with magnitude of r greater than 0.75 can be considered to be a strong correlation.
E.g.: -0.84 is a strong Negative correlation and 0.90 is a strong positive correlation.

A TEACHER WANTS TO DETERMINE THE CORRELATION BETWEEN THE NUMBER OF HOURS SPENT STUDYING AND TEST SCORES.

100

STUDENT NAME	hours x_i	score y_i
JOHN	13	53
ALLIE	15	69
MARK	7	92
SAMANTHA	3	10
JESSICA	10	85
JOSEPH	27	99

$r = \frac{821}{5 * 8.28 * 32.91} = 0.6$

↳ high +ve rel

x_i y_i

$r \Rightarrow \frac{cov(x,y)}{\sigma_x \sigma_y}$

$r = \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma (x_i - \bar{x})^2} * \sqrt{\Sigma (y_i - \bar{y})^2}}$

$$r = \frac{1}{(6-1)s_x s_y}\left[\ 821\ \right]$$

x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
13	53	0.5	−15	−7.5
15	69	2.5	1	2.5
7	92	−5.5	24	−132
3	10	−9.5	−58	551
10	85	−2.5	17	−42.5
27	99	14.5	31	449.5
$\bar{x} = 12.5$	$\bar{y} = 68$			SUM = 821
$s_x = 8.28$	$s_y = 32.91$			

Question: The local ice cream shop keeps track of how much ice cream they sell versus the temperature on that day, here are their figures for the last 12 days. Can you tell if Ice cream sales are correlated to that of temperature? Find out the nature and strength of correlation.

09:00 PM → 10 mins

Temperature	Ice Cream Sales
14.2°	$215
16.4°	$325
11.9°	$185
15.2°	$332
18.5°	$406
22.1°	$522
19.4°	$412
25.1°	$614
23.4°	$544
18.1°	$421
22.6°	$445
17.2°	$408

Youtube
Python

corr does not imply causation

np.corrcoef(,)

$r = 0.96$

Strong +ve corr

> 96%

Cause → further investigation

corr

$$\sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

$$\bar{x} = \frac{\sum x_i}{N}$$

$$cov(x,y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

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

14.2
+ 16.4
+ 11.9