→ Pearson Correlation coefficient

1. Co-variance → $Cov(x, y)$

$$= \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu_x) * (y_i - \mu_y)$$

2. Pearson Correlation $= \rho_{x,y} = \frac{Cov(x,y)}{\sigma_x \sigma_y}$

3. Spearman, rank Correlation

1. Covariance.

Size	Price
1200sqm	$100K
1800 sqm	$200K
1500 sqm	$150K

Is there a relationship between Size and Price?

Quantifying the relationship

$$Cov(\underset{x}{Size}, \underset{y}{Price}) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu_x) * (y_i - \mu_y)$$

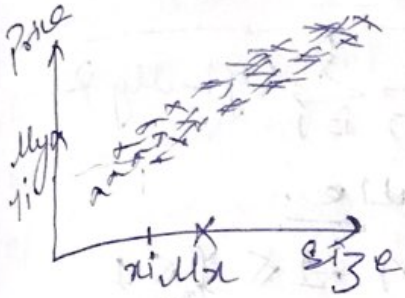
$$Var(x) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu_x)^2$$

$$= \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu_x) * (x_i - \mu_x)$$

If x increases and y increase or x increases and y decreases

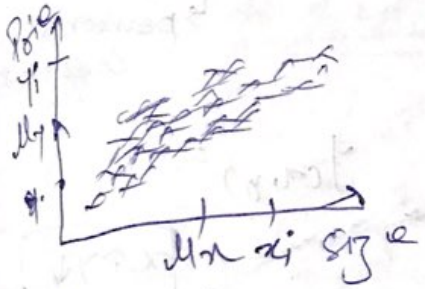
x ↑ and y ↑ = ▭ +ve

x ↑ and y ↓ = ▭ -ve



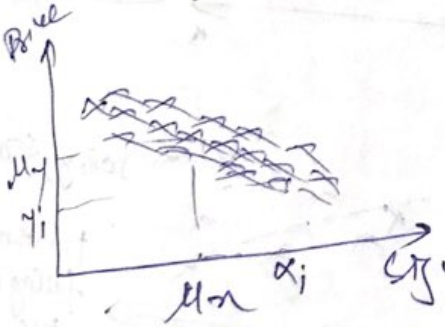
$$\frac{1}{n}(x_i - \mu_x)(y_i - \mu_y)$$

$$\frac{1}{n}(-ve) \times (-ve) = +ve$$



$$\frac{1}{n}(x_i - \mu_x) \times (y_i - \mu_y)$$

$$\frac{1}{n}(+ve) \times (+ve) = +ve$$



$$\frac{1}{n}(x_i - \mu_x) \times (y_i - \mu_y)$$

$$\frac{1}{n}(+ve) \times (-ve)$$

$$= -ve$$

So, covariance tells us about the direction of relationship between two random variable. But doesn't tell about the degree to which it is +vely related or -vely related.

For that we have correlation.

-> Pearson correlation coefficient - $\rho_{xy} = \frac{Cov(x,y)}{\sigma_x \sigma_y}$

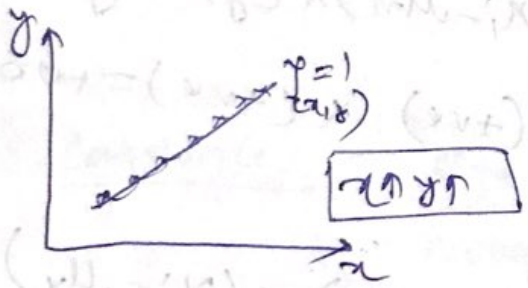
↳ std dev

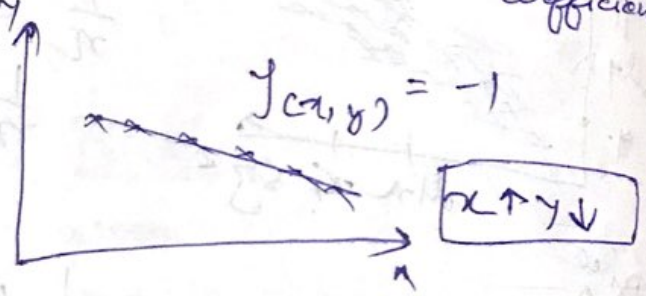
$$= \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \mu_y)^2}}$$

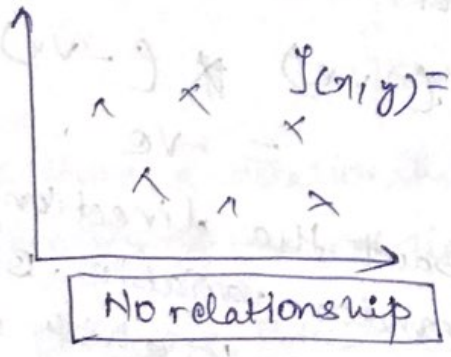
Tells the strength of two variable.

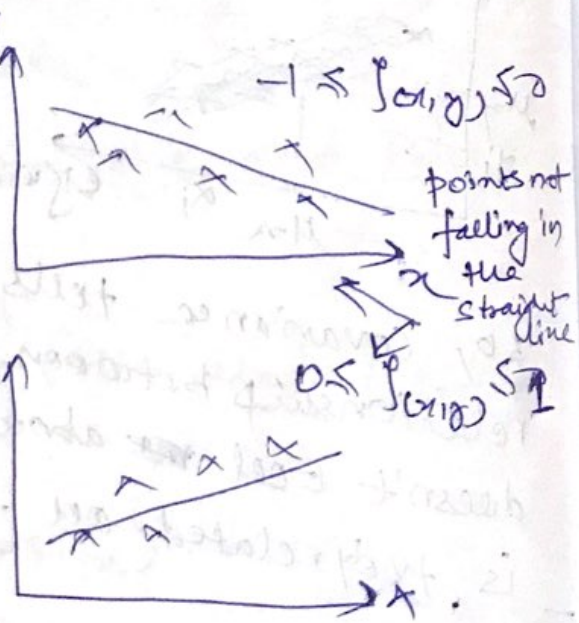
the value ranges between $-1 \leq \rho_{(x,y)} \leq 1$

↳ pearson's coefficient









→ Spearman rank correlation coefficient

$$r_{(x,y)} = \frac{cov(rg_x, rg_y)}{\sigma_{rg_x}\,\sigma_{rg_y}}$$

↳ we find the correlation bet^n rank of x and rank of y.

In this example, the raw data in the table below is used to calculate the correlation between the IQ of a person with the no. of hours spent in front of TV per week.

IQ x_i	Hours of TV/week y_i
106	7
86	2
86	2
101	50
99	28
103	29
97	20
113	12
112	6
110	7

1. sort the data by the 1st column (x_i). create a new column x_i and assign it the ranked values, 1, 2, 3, ... n.
2. create a 4th column and and sort the y_i value and place it accordingly against x_i value.
3. create a 5th col d_i to hold the differences between two rank columns (x_i and y_i)
4. create one final column d_i^2 to hold the value of column d_i^2.

IQ x_i	Hours TV/week y_i	rank x_i	rank y_i	d_i	d_i^2
86	0	1	1	0	0
97	20	2	6	-4	16
99	28	3	8	-5	25
100	27	4	7	-3	9
101	50	5	10	-5	25
103	29	6	9	-3	9
106	7	7	3	+4	16
110	17	8	5	3	9
112	6	9	2	7	49
113	12	10	4	6	36

$$\gamma = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \qquad \sum d_i^2 = 194$$

$$= 1 - \frac{6 \times 194}{10(10^2-1)} = -0.17575$$

→ Pearson focuses mainly on linear relationship, linear aspect of data.

→ where as spearman rank correlation focuses on the outliers also and gives the correlation value on non-linear relationships also

→ In python heatmap it uses this technique only.