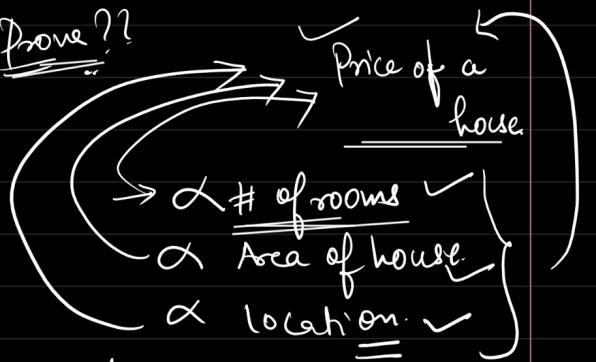
Data

↓

Descriptive Statistics



↓ ↓

transaction amt	transaction count
1000	1
2000	2
10000	3

Hardik Pandya → MI

→ whatever you are going to predict / price of a house, it should have some relation with the features

→ Price of house ∝ Area of house, No of rooms, locality

relationship

No of matches played	total score

* Measure the relationship

(1) Covariance

✓ (2) Correlation.

* Covariance ⟹ Variance:

⇓

Spread of data.

Co - variance

① Covariance

② Correlation

* Covariance → Co + Variance → relation of feature with itself

To determine relationship b/w two variable.

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

$$\frac{\sum (x_i - \bar{x})(x_i - \bar{x})}{n-1}$$

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

Co-variance

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

X	Y
2	3
3	5
6	6
1	8
$\bar{x} = 3$	$\bar{y} = 5.5$

$$\Rightarrow \frac{(2-3)(3-5.5) + \cancel{(3-3)(5-5.5)}^{0} + (6-3)(6-5.5) + (1-3)(8-5.5)}{3}$$

$$\Rightarrow \frac{(-1)(2.5) + 0 + 3 \times 0.5 + -2 \times 2.5}{3} \Rightarrow$$

$= -1/3 = -0.33$

→ The two features (X & Y) is negatively related.

X	Y
2	3
4	5
6	7
$\bar{x} = 4$	$\bar{y} = 5$

$$Cov(x,y) = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{n-1}$$

$$= \frac{(2-4)(3-5) + (\nearrow)^0 + (6-4)(7-5)}{3-1}$$

$\Rightarrow +4$

X & Y is +vely related

→ 0 → No relationship

magnitude

Cov(x y) – −0.33 → negative / inversely relation

Cov(x,y) – +4 → +ve / directly relation

Name of owner	f1 (sqft) (Area of house)	Wt of the person living	time.	Y
Ajay	1800	180	10 years	1.2 cr.

sqft (f1) ↑ ↗ Wt ↓

$$Cov = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{n-1} =$$

$Cov = sqft \cdot kg$

disadvantage of covariance

→ dimensional quantity — $Cov(x,y)$ → sqft kg

→ ranges between $-\infty$ to ∞

Correlation

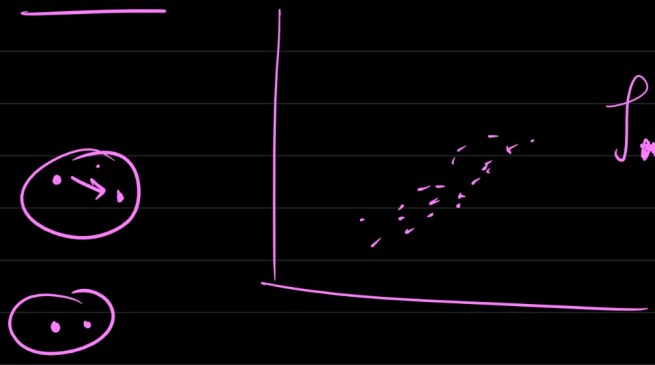
$$\frac{Cov(x,y)}{\sigma_x \times \sigma_y}$$ $\frac{kg \cdot kg}{kg \cdot kg}$ → dimensionless quantity

→ ranges b/w −1 and 1

−1 0 1

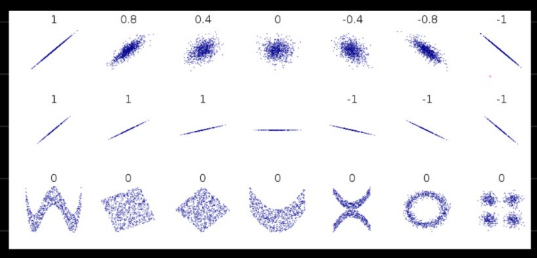
← −ve correlation → +ve correlation

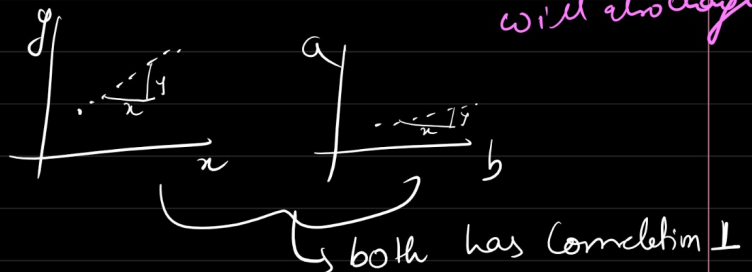
Interpretation



$\rho_{x,y} = 0.4$ → How much sure it is if there if 1 unit change in x, the other unit will also change

40%





* Correlation measures — linear relationship

* Spearman rank — Non linear relationship

$$r_s = \frac{Cov(R(x), R(y))}{\sigma(R_x) * \sigma_{R_y}}$$

x	y	R(x)	(R(y)
5	6	3	1
7	4	2	2
8	3	1	3
1	1	5	5
2	2	4	4

* Use case of correlation → To select the relevant feature.