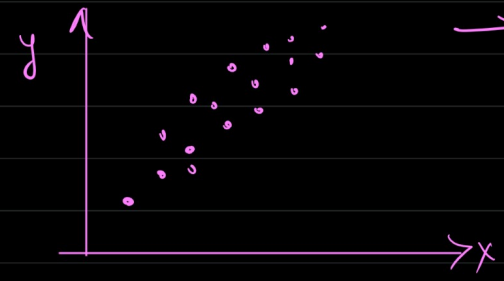
* Covariance and correlation

(X) transaction amount	(Y) transaction count
—	—
—	—
—	—
—	—

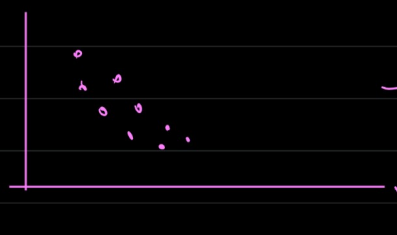
- Y↑ X↑
- Y↑ X↓
- Y↓ X↑
- Y↓ X↓

Understanding the relationship



→ relationship b/w X, Y is direct

X↑ Y↑
X↓ Y↓

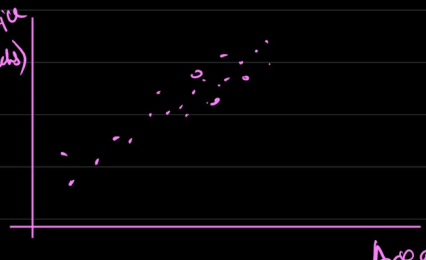


→ indirect relationship

X↓ Y↑
X↑ Y↓

* Example (direct)

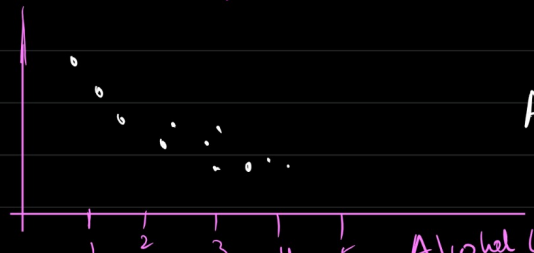
Predict price of house based on area of house



Area of house (sqft)	Price
1100	80
1200	85
—	—
—	—

AH↑ P↑
AH↓ P↓

* Example (Indirect



AC↑ YTL↓
AC↓ YTL↑

① Covariance

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

Co + Variance

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

↓

Variance was spread of data ⇒ relationship of a feature with itself

⇓

Cov mean, you are trying to understand the relationship of a feature with respect to other feature.

→ Covariance → relationship b/w two variable.

→

X ↑ Y ↑	or	X ↑ Y ↓
X ↓ Y ↓		X ↓ Y ↑
(+ve) Cov		(−ve Cov)

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

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

$$\frac{(2-3)(3-5.5) + (3-3)(5-5.5) + (6-3)(6-5.5) + (1-3)(8-5.5)}{4-1}$$

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

$$\Rightarrow \frac{2.5 + 0 + 1.5 - 5}{3} \Rightarrow \frac{-1}{3} = \boxed{-0.33}$$

→ The two features X and Y are negatively related.

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

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

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

$$\Rightarrow \frac{4+0+4}{2} = 4 = +ve$$

→ x & y are having a positive relation.

Advantage

→ Relationship b/w X, Y

+ve or -ve

* Disadvantage

①

X	Y
A	B

$Cov(x,y) = 50$

$Cov(A,B) = 100$

$\Downarrow$

$-\infty$ to ∞

→ No comparison of Strength of relationship in Covariance

→ No any standardized scale to interpret the strength.

② Covariance has dimension

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

x → height → ft

y → wt → kg.

Cov. ⇒ ft. kg

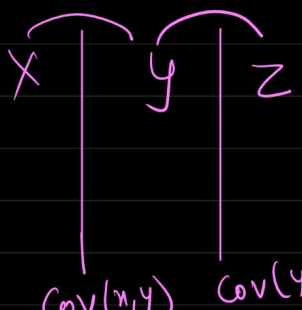
(X) (Y) (Z)

transaction Amount (Rs) | height (ft) | weight (kg)

$Cov(trAmt, ht) \Rightarrow Rs \cdot ft \Rightarrow 450\ Re \cdot ft$

$Cov(height, wt) \Rightarrow ft \cdot kg \Rightarrow 600\ ft \cdot kg$

→ We can not compare two different dimension



Not comparable → different dimension.

* Soln
  - → -1 to 1
  - → dimensionless Quantity

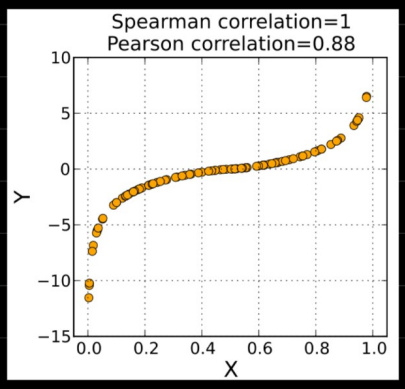
② Pearson Correlation Coefficient [-1 to 1]

$$\rho_{x,y} = \frac{Cov(x,y)}{\sigma_x \sigma_y} = [-1 \text{ to } 1]$$

$\rho_{x,y} = 0.4$ | $\rho_{A,B} = 0.8$

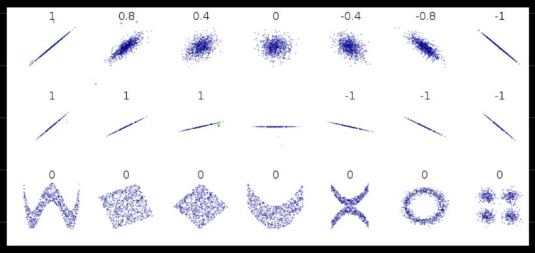
→ feature A,B is highly correlated as compared to x,y.

→ Pearson correlation coefficient always measures the linear relationship



Pearson correlation
↓
linear relationship

$x \uparrow y \uparrow$



What to do for Non linear relationship

Spearman Rank Correlation

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

$R(x)$ – Rank of x

$R(y)$ – Rank of 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

5th 4th 3rd 2nd 1st

$x \rightarrow 5, 7, 8, 1, 2 \longrightarrow$ Sort the value $\rightarrow 1, 2, 5, 7, 8$

Sort the no → Highest no will be rank 1.

Dataset → 1000 feature

o/p (y)

x_1 x_2 x_3 x_4 x_5 x_6 ---- x_{1000} Price

$x_1 - y$

$x_2 - y$

$x_3 - y$

$x_4 - y$

Corr ≈ 0