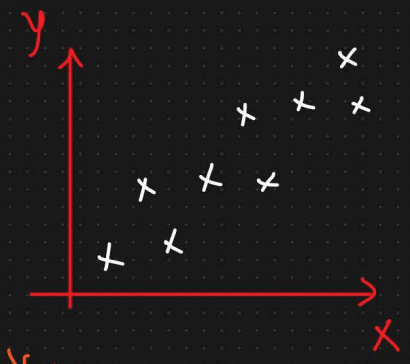
Covariance And Correlation

[Relationship between X and Y]

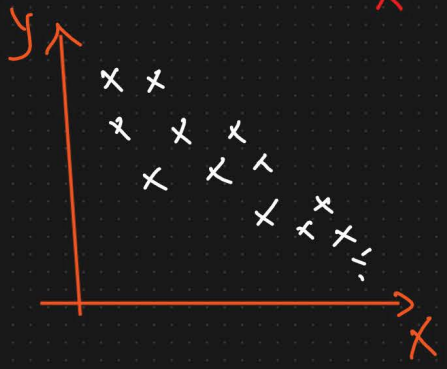
X	y
2	3
4	5
6	7
8	9

- $X\uparrow \; y\uparrow$
- $X\downarrow \; y\uparrow$
- $X\downarrow \; y\downarrow$
- $X\uparrow \; y\downarrow$

↓↑ Size of House → Price of House ↑↓ ↓o/p



- $X\uparrow \; y\uparrow$
- $X\downarrow \; y\downarrow$



- $X\uparrow \; y\downarrow$
- $X\downarrow \; y\uparrow$

Covariance

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

$$Var(x) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})}{n-1} = Cov(x,x) \Rightarrow \text{Spread}$$

$x_i \rightarrow$ Datapoints of x

$\bar{x} \rightarrow$ Sample mean of x

$y_i \rightarrow$ Datapoints of y

$\bar{y} \rightarrow$ Sample mean of y

Cov(x,y)

x↑	y↑
x↓	y↓

+ve Covariance

x↑	y↓
x↓	y↑

-ve Covariance

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

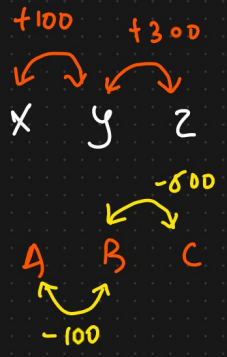
$$Cov(x,y) = \frac{\sum_{i=1}^{n} (x_i-\bar{x})(y_i-\bar{y})}{n-1}$$

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

$$= \frac{4+0+4}{2} = \frac{8}{2} = 4 \text{ tve value}$$

⇓

Positive Covariance



X & y are having a positive Covariance

Advantages

① Relationship between X and y

+ve or -ve value

Disadvantages

① Covariance does not have a specific limit value

② Pearson Correlation Coefficient [-1 to 1]

$$\rho_{x,y} = \frac{Cov(x,y)}{\sigma_x \cdot \sigma_y}$$

① The more the value towards +1 the more +ve Correlated it is (x, y)

② The more the value towards -1 the more -ve Correlated it is (x, y)

③ Spearman Rank Correlation [-1 to 1]

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

X	y	R(x)	R(y)
1	2	5	5
3	4	4	4
5	6	3	3
7	8	2	1
0	7	6	2
8	1	1	6

Feature Selection

+ve — Size of House ↑

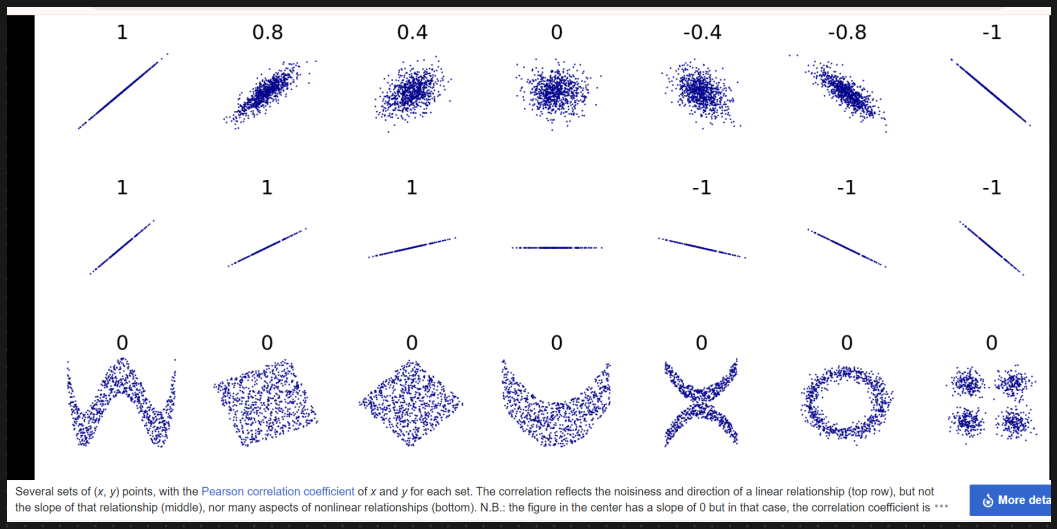
+ve — No. of Rooms ↑

+ve — Location ↑

≈0 — ~~No. of people Staying~~

-ve — Haunted

Price ↑



Several sets of (*x*, *y*) points, with the Pearson correlation coefficient of *x* and *y* for each set. The correlation reflects the noisiness and direction of a linear relationship (top row), but not the slope of that relationship (middle), nor many aspects of nonlinear relationships (bottom). N.B.: the figure in the center has a slope of 0 but in that case, the correlation coefficient is ***

More deta

	Covariance	Correlation
Aim	Both indeces are used to get an idea of how two random variables change together.	
Domain	Defined between -∞ and +∞	Defined between -1 and 1
Measurement Unit	Depends on the measurement units of the random variables	Unitless number
Scale change/Normalization	*Affects* covariance	*Does not affect* correlation

Key Differences

Key	Covariance	Correlation
Meaning	Covariance is a measure indicating the extent to which two random variables change in tandem.	Correlation is a statistical measure that indicates how strongly two variables are related.
What is it?	Measure of correlation	Scaled version of covariance
Values	Lie between -∞ and +∞	Lie between -1 and +1
Change in scale	Affects covariance	Does not affects correlation
Unit free measure	No	Yes