Covariance and Correlation:

Let's Say there are 2 variables

X	Y
2	3
4	5
6	7
8	9

Relationship between X & Y

X ↑ Y ↑

X ↑ Y ↓

X ↓ Y ↑

X ↓ Y ↓

Such types of relationships are very important in ML.

Example:

Size	location	Price

Size ↑ Price ↑

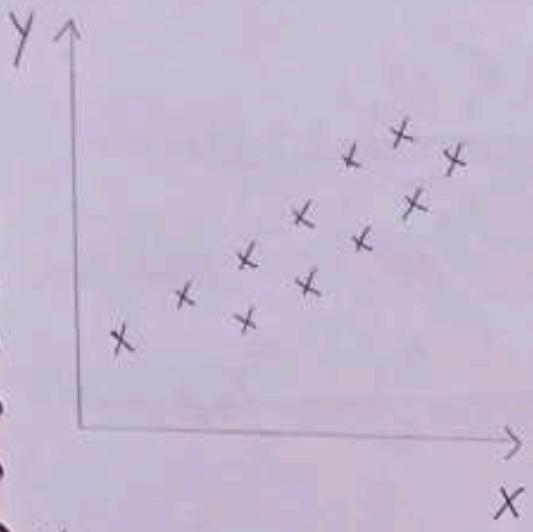
better location Price ↑

→ relationship

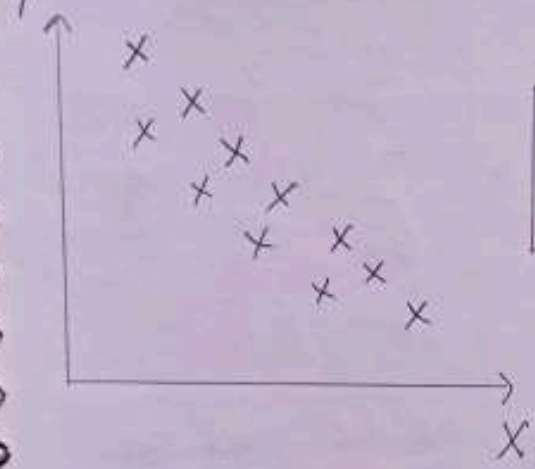
we can predict price, with help of Size

better location

Graphical understanding:



X↑	Y↑
X↓	Y↓



X↑	Y↓
X↓	Y↑

Covariance:

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

$$\text{Cov}(X,Y)$$

$$\text{Variance}(X) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$$

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

↓ Spread of data

Interview question

X↑	Y↑
X↓	Y↓

⊕ve Cov

X↑	Y↓
X↓	Y↑

⊖ve Cov

Example:

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

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

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

$$= 4 > 0$$

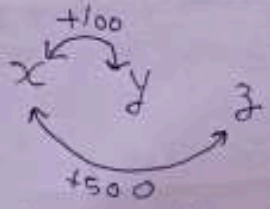
X, Y are having ⊕ve Covariance

Advantage of Covariance ↓

It Gives us relationship between X & Y whether ⊕ve or ⊖ve

Disadvantage of Covariance ↓

It doesn't a specific limit value.



There is no restriction on how much ⊕ve or ⊖ve.

Pearson Correlation Coefficient:

Its range is $[-1, 1]$

$$\rho_{x,y} = \frac{cov(X, Y)}{\sigma_x \sigma_y}$$

① More the value towards +1 more +ve correlated it is

② More the value towards -1 more -ve correlated it is

Application in data science projects:

It plays an important role in feature selection.

(Indep features)

Size ①	no. of rooms ②	location ③	no. of People staying ④	Price ⑤

features ①②③ are highly correlated with ⑤ (⑤ → O/P dep feature)

④ may not be related to ⑤

In feature Selection, if I get a value → 0 then it is not correlated.

We drop column Since redundant

Spearman Rank Correlation:

$$r_s = \frac{cov(R_x, R_y)}{\sigma_{R_x} \sigma_{R_y}}$$

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

SRC can also be used for feature selection.