Covariance

Let's have a look at the things we are going to discuss in this article:

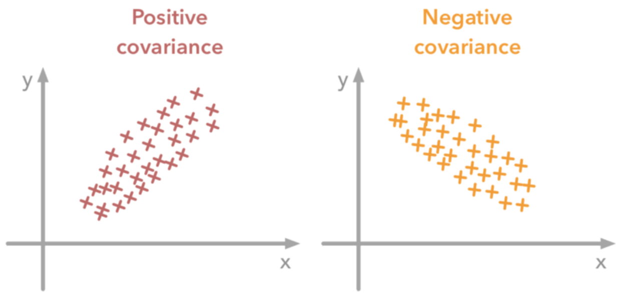
- What is covariance?
- What is correlation?
- Difference between covariance and correlation?

Covariance and correlation are both mathematical notions that are employed in statistics and probability theory. The most beneficial in terms of comprehending variables. Generally, the data science field is used for comparing data samples from different populations, and covariance is used to determine how much two random variables are related to each other, and correlation is used to evaluate whether a change in one variable affects another variable.
Both terms refer to the linear relationship between variables. In other words, a positive correlation exists when one variable moves in the same direction as another variable. When both variables are pointing in the opposite direction, this is referred to as negative correlation. When there is no relationship, there are no changes. Correlation describes how a change in one variable affects the amount of change in the second variable.

We calculate covariance and correlation on samples rather than the complete population.

Covariance

Covariance is only affected by sign. A positive value indicates that both variables are moving in the same direction. The same as A negative score indicates that they are moving in opposite directions. Covariance is a metric for determining how much a variable fluctuates at random. The covariance is a product of the two variables' units. Covariance has a value between -∞ and +∞. The covariance of two variables (x and y) can be represented by cov(x,y). E[x] is the expected value or also called as a means of sample ‘x’.



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

Where,

- $\bar{x}$ = sample mean of x
- $\bar{y}$ = sample mean of y
- x_i and y_i=the values of x and y

for i th record in the sample.

- N = is the no of records in the sample

In Python: use cov() function

$x > \mu_{x,}\ y > \mu_y$	+	+
$x < \mu_{x,}\ y < \mu_y$	-	-
$y < \mu_y\ x > \mu_x$	+	-
$y > \mu_y\ x < \mu_{x,}$	-	+

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

The formula's significance

The numerator represents the amount of variance in x multiplied by the amount of variance in y. The unit of covariance is defined as a unit of x multiplied by a unit of y.

As a result, if we alter the unit of variables, covariance will have a new value but the sign will remain unchanged.

If it is positive, both variables will fluctuate in the same direction; if it is negative, they will vary in the opposite direction.

Correlations

Correlation is defined as the correlation of two variables, which is a normalised version of covariance. Correlation coefficients are always between -1 and 1. The correlation coefficient is also known as Pearson’s correlation coefficient. When you read about Covariance, you will only learn about the direction, which is insufficient to fully comprehend the relationship. As a result, we divide the covariance by the standard deviations of x and y.

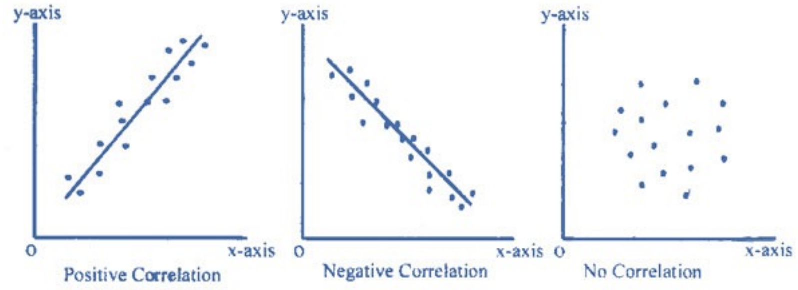
To get the correlation coefficient between the random variables X and Y, divide the sample covariance of X and Y by the product of X and Y's sample sat.deviation.
For example, height and weight are related; taller people tend to be heavier than shorter people.

$$r = \frac{1}{(n-1)s_x s_y} \Sigma(x_i - \bar{x})(y_i - \bar{y})$$

(n = sample size, and Sx, Sy are the standard deviations of samples x and y. X-bar and y-bar are the respective means of x and y samples whereas Xi and Yi are sample points of X and Y respectively.)

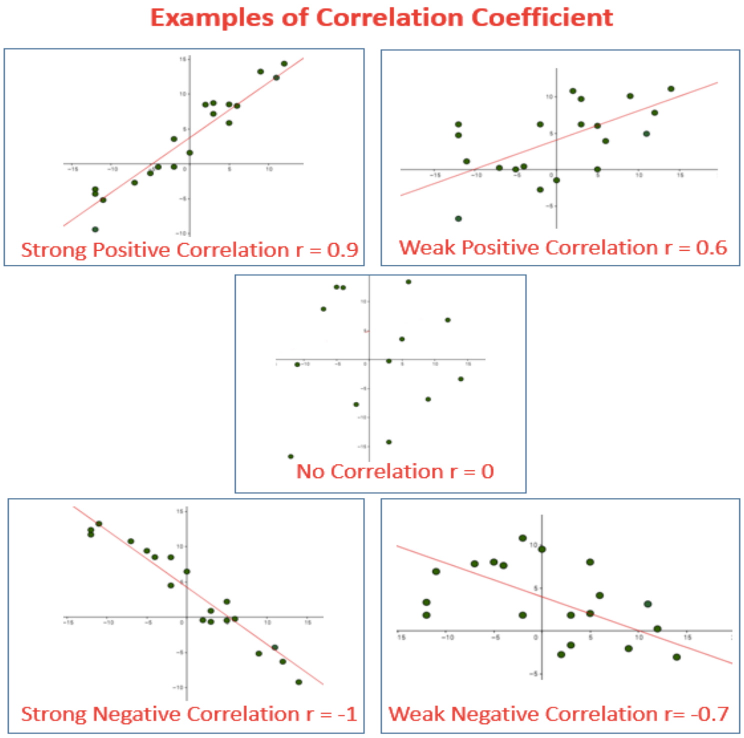
Positive and Negative Correlation:

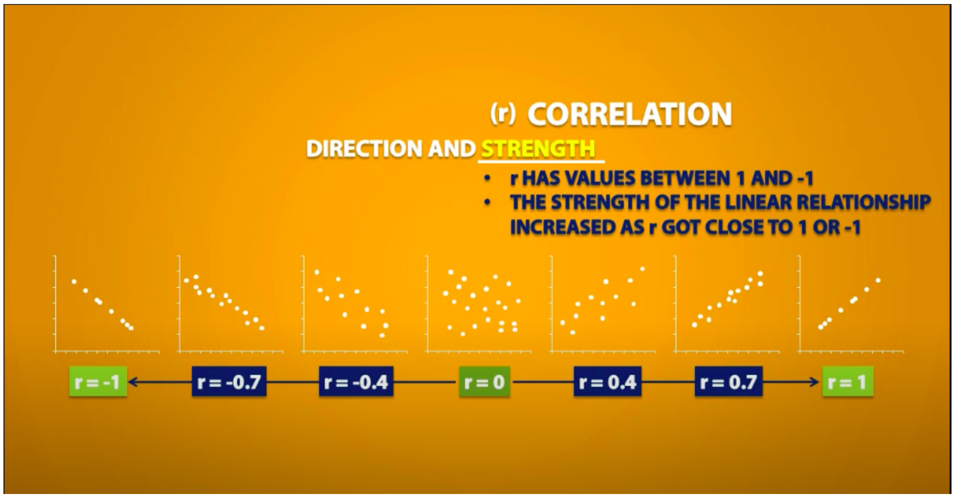
1. Correlation Coefficient greater than zero indicates a positive relationship.
2. while a value less than zero signifies a negative relationship.
3. and a value of zero indicates no relationship between the two variables being compared.



Strong and Weak Correlation:

Kind of correlation = depicted by sign of correlation coefficient
How Strong = Value of Correlation Coefficient



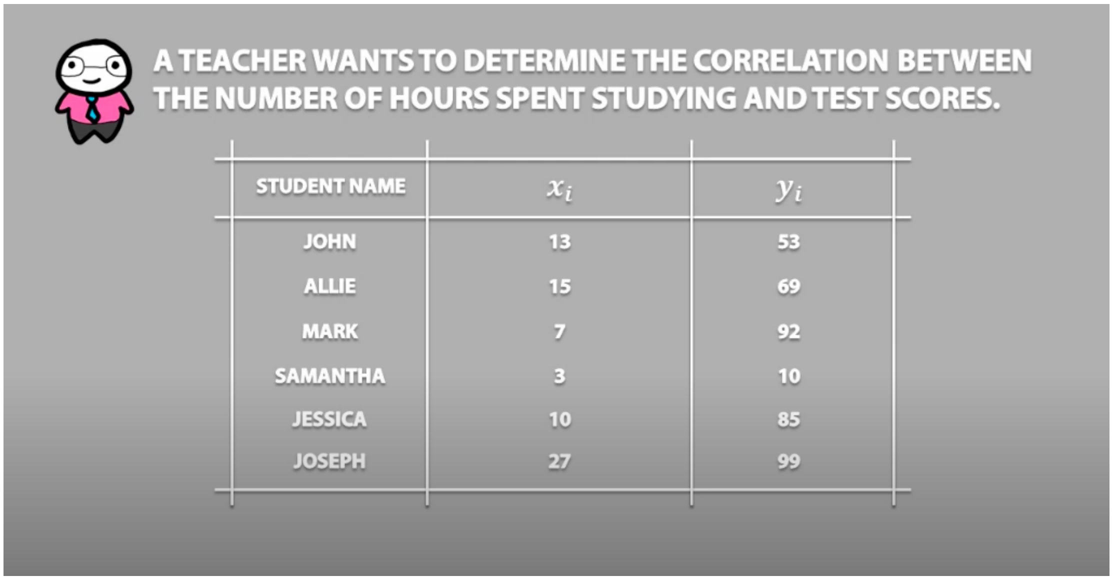


(r) CORRELATION

DIRECTION AND STRENGTH

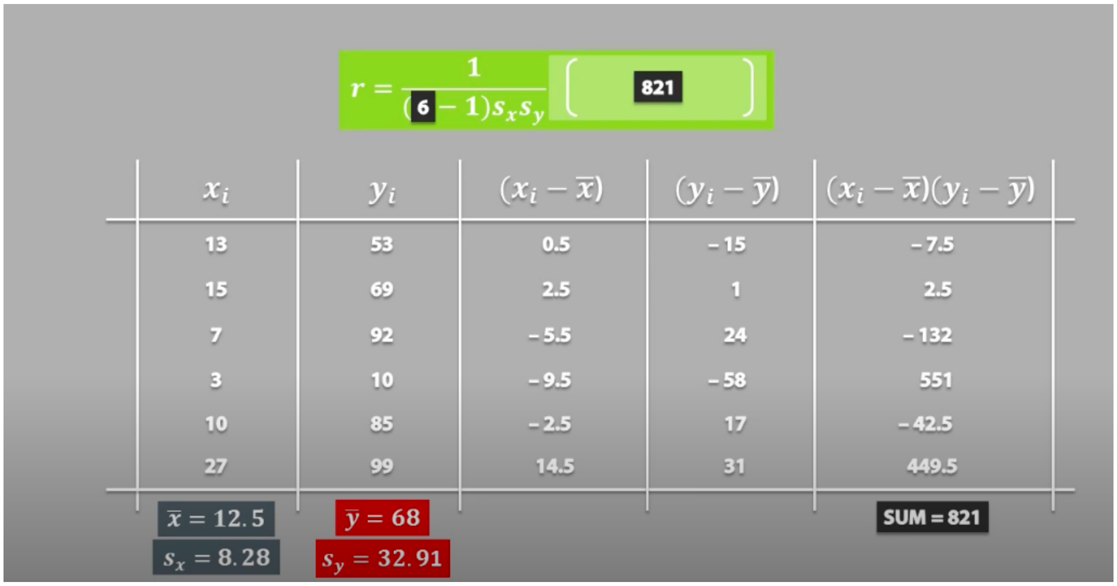
- r HAS VALUES BETWEEN 1 AND -1
- THE STRENGTH OF THE LINEAR RELATIONSHIP INCREASED AS r GOT CLOSE TO 1 OR -1

Rule of thumb: Any relationship with magnitude of r greater than 0.75 can be considered to be a strong correlation.
E.g.: -0.84 is a strong Negative correlation and 0.90 is a strong positive correlation.



A TEACHER WANTS TO DETERMINE THE CORRELATION BETWEEN THE NUMBER OF HOURS SPENT STUDYING AND TEST SCORES.

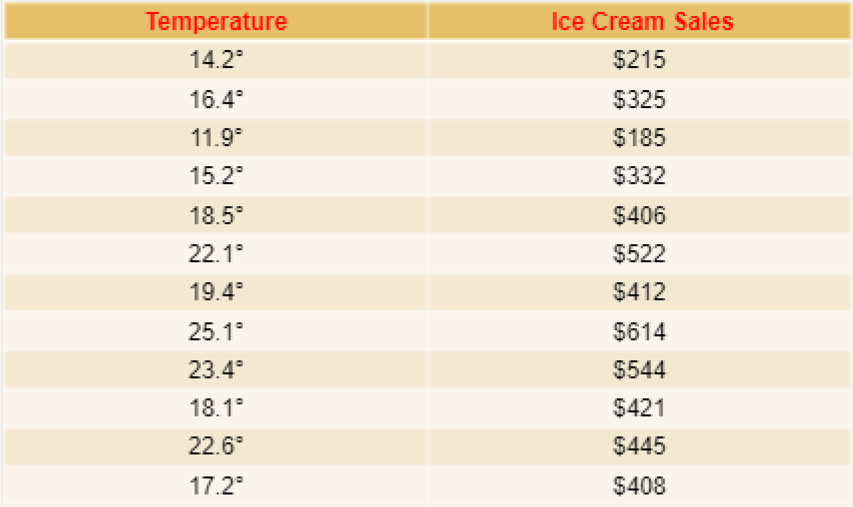
STUDENT NAME	x_i	y_i
JOHN	13	53
ALLIE	15	69
MARK	7	92
SAMANTHA	3	10
JESSICA	10	85
JOSEPH	27	99



$$r = \frac{1}{(6-1)s_x s_y}\left[\ 821\ \right]$$

x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
13	53	0.5	– 15	– 7.5
15	69	2.5	1	2.5
7	92	– 5.5	24	– 132
3	10	– 9.5	– 58	551
10	85	– 2.5	17	– 42.5
27	99	14.5	31	449.5
$\bar{x} = 12.5$ $s_x = 8.28$	$\bar{y} = 68$ $s_y = 32.91$			SUM = 821

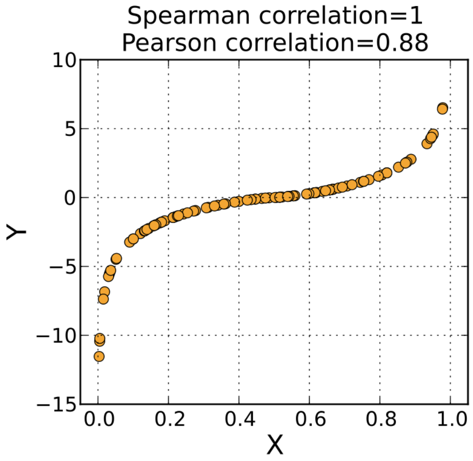
Question: The local ice cream shop keeps track of how much ice cream they sell versus the temperature on that day, here are their figures for the last 12 days. Can you tell if Ice cream sales are correlated to that of temperature? Find out the nature and strength of correlation.



Temperature	Ice Cream Sales
14.2°	$215
16.4°	$325
11.9°	$185
15.2°	$332
18.5°	$406
22.1°	$522
19.4°	$412
25.1°	$614
23.4°	$544
18.1°	$421
22.6°	$445
17.2°	$408

Spearman Rank Correlation

- Used for Non-Linear Variables
- Spearman Corr Coeff = Pearson Corr coeff (rank variables)
- In Python: DataFrame.corr(method='spearman')
- Denoted by rho.



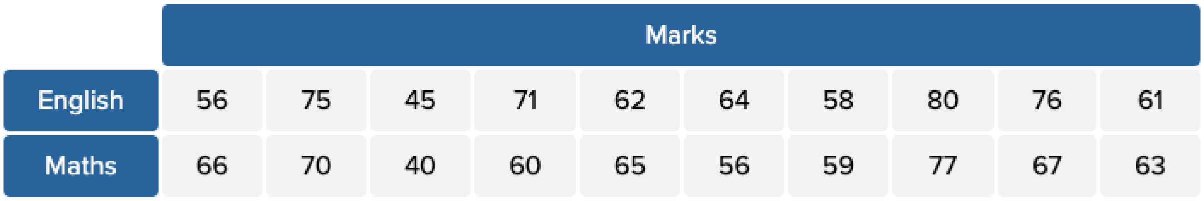
$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

Steps for Spearman Correlation Coefficient

1. Create a new column for rank(x) and assign the rank of each variable.
2. Assign the rank of 2nd variable in a new column rank(y).
3. Calculate the difference in rank of both the variables = d.
4. Calculate the d-squared.
5. Add up d-squared score.
6. Put in the formula provided:

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

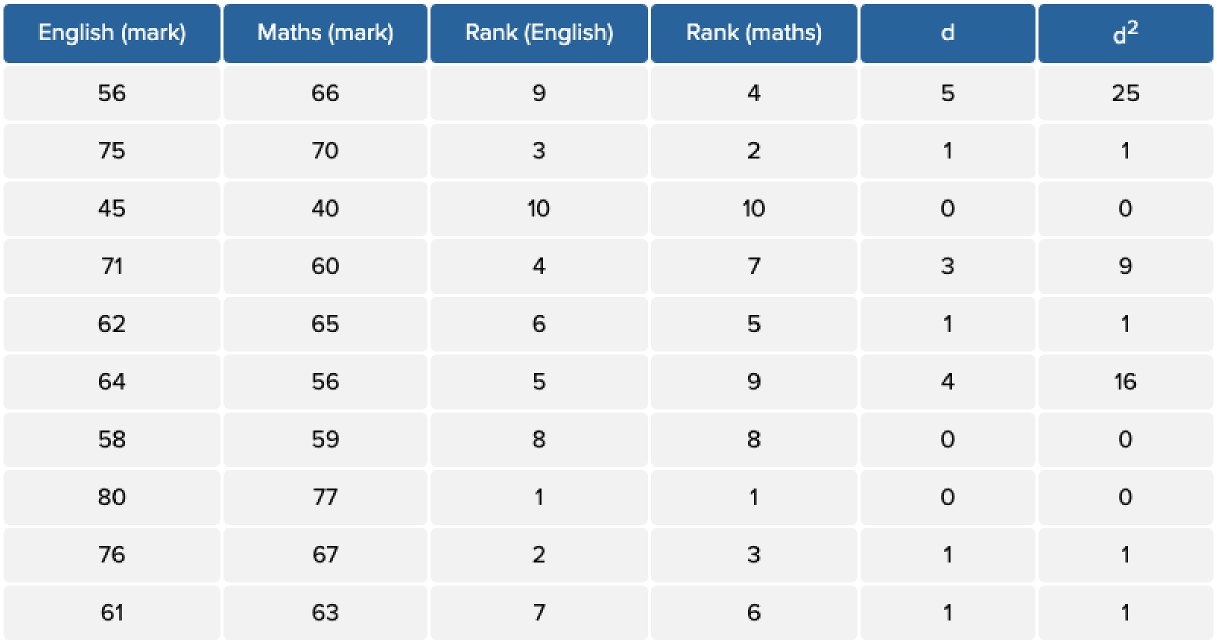
Question: The scores for 10 students in English and Maths are as follows:



	Marks									
English	56	75	45	71	62	64	58	80	76	61
Maths	66	70	40	60	65	56	59	77	67	63

Compute the Spearman rank correlation.
Solution:

Step 1,2,3 and 4:



English (mark)	Maths (mark)	Rank (English)	Rank (maths)	d	d^2
56	66	9	4	5	25
75	70	3	2	1	1
45	40	10	10	0	0
71	60	4	7	3	9
62	65	6	5	1	1
64	56	5	9	4	16
58	59	8	8	0	0
80	77	1	1	0	0
76	67	2	3	1	1
61	63	7	6	1	1

Step 5: $\sum d_i^2 = 25 + 1 + 9 + 1 + 16 + 1 + 1 = 54$

Step 6:

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

$$\rho = 1 - \frac{6 \times 54}{10(10^2 - 1)}$$

$$\rho = 1 - \frac{324}{990}$$

$$\rho = 1 - 0.33$$

$$\rho = 0.67$$

Hence, the Spearman Rank Coefficient is 0.67.

Difference Between Covariance and Correlation

Correlation is simply a normalized form of covariance. It is obviously important to be precise with language when discussing the two, but conceptually they are almost identical.

The value of the correlation coefficient ranges from [-1 – 1]. -1 is indicated for a negative relationship. 1 means a positive relationship. 0 means no relationship.