Covariance and Correlation

Covariance and Correlation are very helpful in understanding the relationship between two continuous variables.

Covariance tells whether both variables vary in same direction (positive covariance) or in opposite direction (negative covariance). There is no significance of covariance numerical value only sign is useful.

Whereas Correlation explains about the change in one variable leads how much proportion change in second variable. Correlation varies between -1 to +1.

Correlation tells you about if there exist any linear relationship between two continuous variables if yes then in what proportion two continuous variables are linearly corelated.
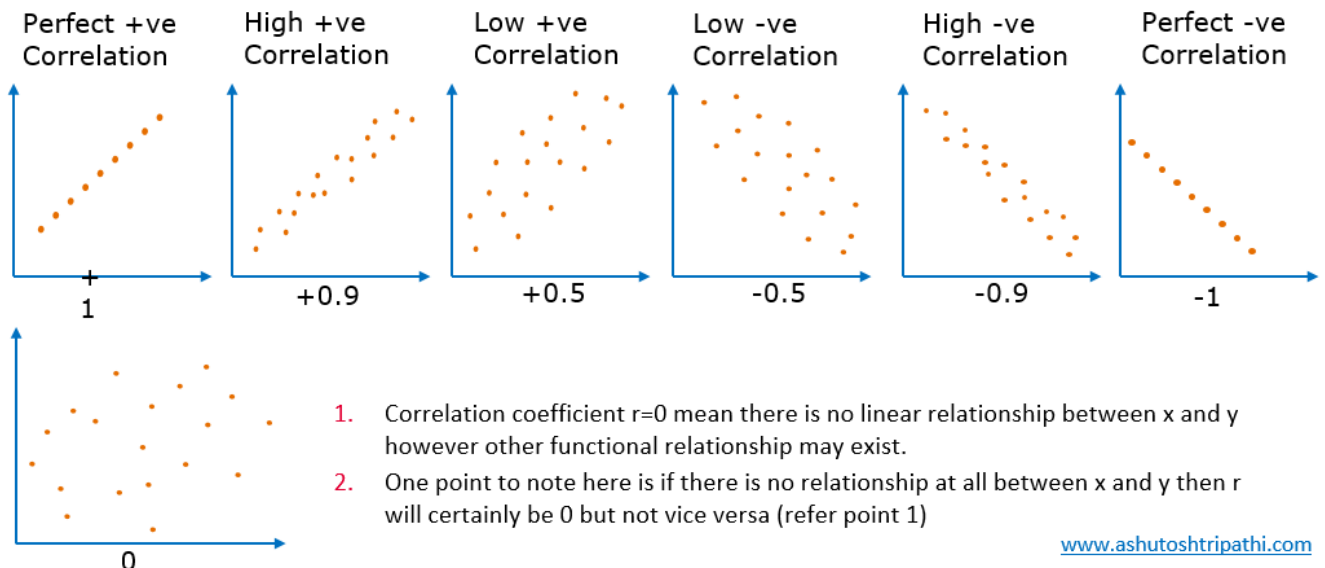
If correlation value is 0 then it means there is no Linear Relationship between variables however other functional relationship may exist.

Correlation coefficient r is number between -1 to +1 and tells us how well a regression line fits the data and defined by

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

where,

- s_{xy} is the covariance between x and y
- s_x and s_y are the standard deviations of x and y respectively.

Perfect +ve Correlation	High +ve Correlation	Low +ve Correlation	Low -ve Correlation	High -ve Correlation	Perfect -ve Correlation
+1	+0.9	+0.5	-0.5	-0.9	-1

1. Correlation coefficient r=0 mean there is no linear relationship between x and y however other functional relationship may exist.
2. One point to note here is if there is no relationship at all between x and y then r will certainly be 0 but not vice versa (refer point 1)

Let's understand these terms in detail:

Covariance:

In the study of covariance only sign matters. Positive value shows that both variables vary in same direction and negative value shows that they vary in opposite direction.

Covariance between two variables x and y can be calculated as following:

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

Where:

- x̄ is sample mean of variable x
- ȳ is sample mean of target variable y
- x_i and y_i are the values of x and y for ith record in sample.
- n is the no of records in sample

Significance of the formula:

- Numerator: Quantity of variance in x multiplied by quantity of variance in y.
- Unit of covariance: Unit of x multiplied by unit of y
- Hence if we change the unit of variables, covariance will have new value however sign will remain same.
- Therefore, numerical value of covariance does not have any significance as it is unit dependent and hence only sign matter. If it is positive then both variables vary in same direction else if it is negative then they vary in opposite direction.

Correlation:

As covariance only tells about the direction which is not enough to understand the relationship completely, we divide the covariance with standard deviation of x and y respectively and get correlation coefficient which varies between -1 to +1.

- -1 and +1 tells that both variables have perfect linear relationship.
- Negative means they are inversely proportional to each other with the factor of correlation coefficient value.
- Positive means they are directly proportional to each other mean vary in same direction with the factor of correlation coefficient value.
- if correlation coefficient is 0 then it means there is no linear relationship between variables however there could exist other functional relationship.

- if there is no relationship at all between two variables then correlation coefficient will certainly be 0 however if it is 0 then we can only say that there is no linear relationship but there could exist other functional relationship.

Correlation between x and y can be calculated as following:

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

Where:

- S_{xy} is the covariance between x and y.
- S_x and S_y are the standard deviation of x and y respectively.
- r_{xy} is correlation coefficient.
- Correlation coefficient is dimensionless quantity. Hence if we change the unit of x and y then also coefficient value will remain same.
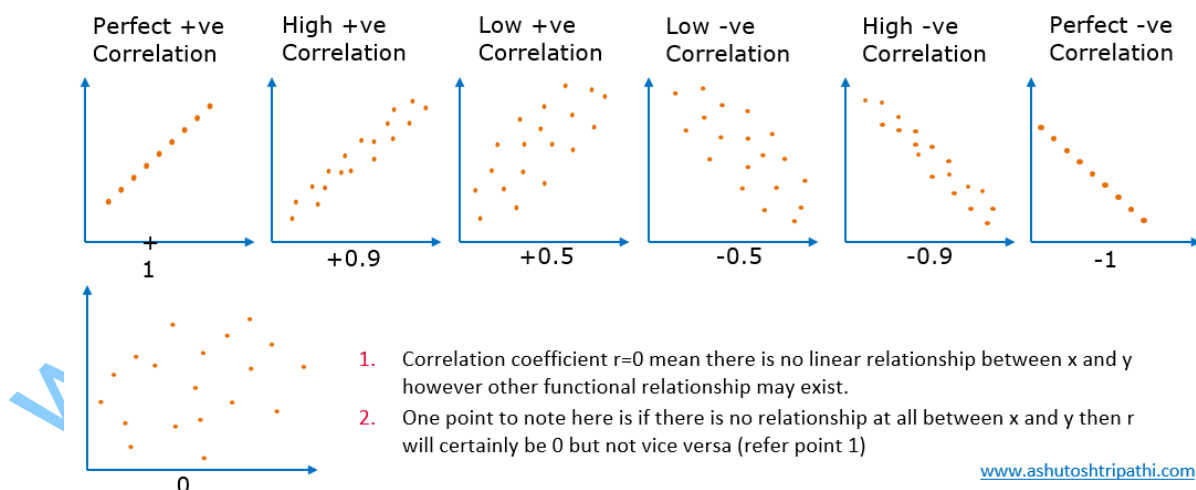
Let's understand what is the significance of correlation coefficient with the help of below infographics:

Correlation coefficient r is number between -1 to +1 and tells us how well a regression line fits the data and defined by

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

where,

- s_{xy} is the covariance between x and y
- s_x and s_y are the standard deviations of x and y respectively.

Perfect +ve Correlation	High +ve Correlation	Low +ve Correlation	Low -ve Correlation	High -ve Correlation	Perfect -ve Correlation
+1	+0.9	+0.5	-0.5	-0.9	-1

0

1. Correlation coefficient r=0 mean there is no linear relationship between x and y however other functional relationship may exist.
2. One point to note here is if there is no relationship at all between x and y then r will certainly be 0 but not vice versa (refer point 1)

www.ashutoshtripathi.com

Thank You. Happy Reading. Follow me on LinkedIn where I share easy to understand Data Science topics.

LinkedIn: https://www.linkedin.com/in/ashutoshtripathi1/

ML Blog: https://ashutoshtripathi.com