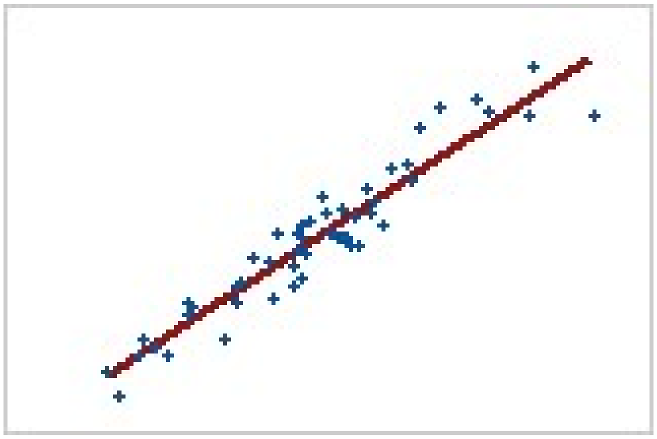
Mithilesh singh

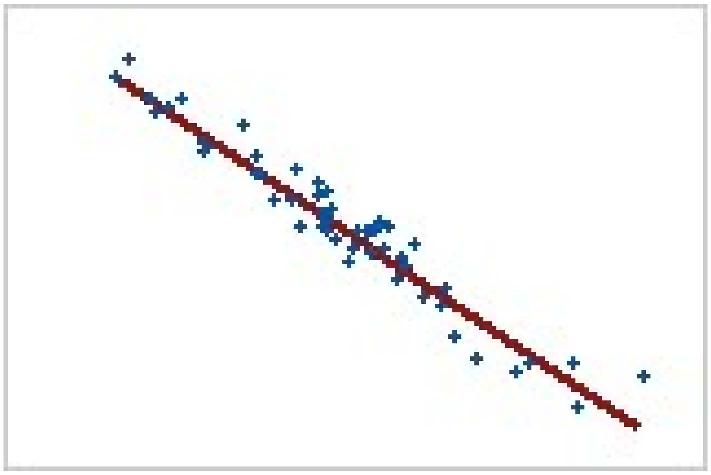
Linear, nonlinear, and monotonic relationships

When evaluating the relationship between two variables, it is important to determine how the variables are related. Linear relationships are most common, but variables can also have a nonlinear or monotonic relationship. It is also possible that there is no relationship between the variables. by creating a scatterplot of the variables to evaluate the relationship.

A linear relationship is a trend in the data that can be modeled by a straight line.



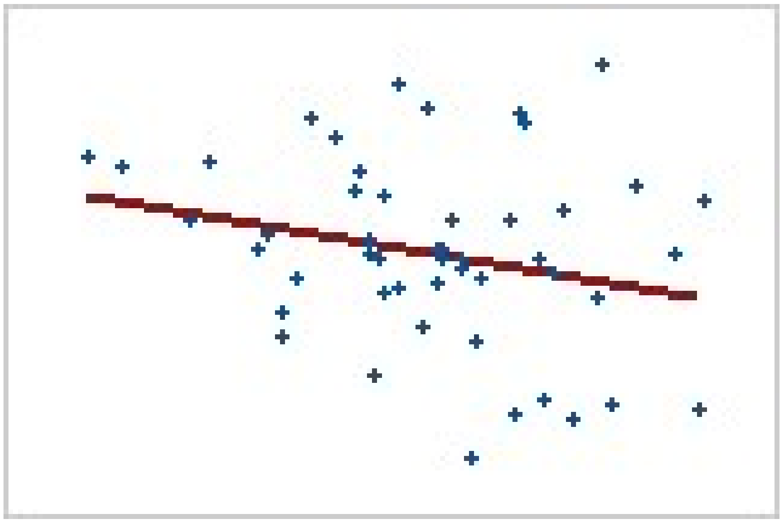
Plot 1: Strong positive linear relationship



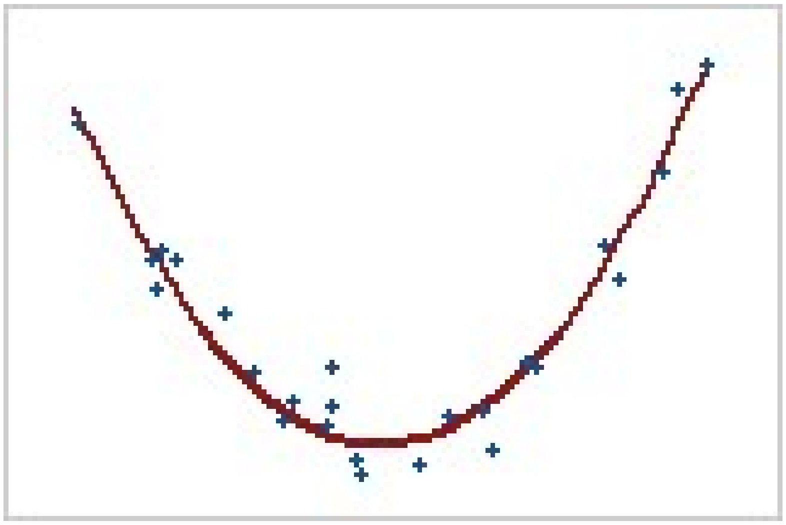
Plot 2: Strong negative linear relationship

When both variables increase or decrease concurrently and at a constant rate, a positive linear relationship exists. The points in Plot 1 follow the line closely, suggesting that the relationship between the variables is strong. The Pearson correlation coefficient for this relationship is +0.921.

When one variable increases while the other variable decreases, a negative linear relationship exists. The points in Plot 2 follow the line closely, suggesting that the relationship between the variables is strong. The Pearson correlation coefficient for this relationship is −0.968.



Plot 3: Weak linear relationship

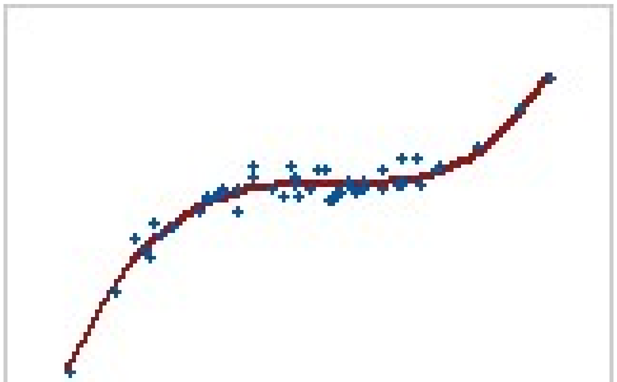


Plot 4: Nonlinear relationship

The data points in Plot 3 appear to be randomly distributed. They do not fall close to the line indicating a very weak relationship if one exists. The Pearson correlation coefficient for this relationship is −0.253.

If a relationship between two variables is not linear, the rate of increase or decrease can change as one variable changes, causing a "curved pattern" in the data. This curved trend might be better modeled by a nonlinear function, such as a quadratic or cubic function, or be transformed to make it linear.

Plot 4 shows a strong relationship between two variables. However, because the relationship is not linear, the Pearson correlation coefficient is only +0.244. This relationship illustrates why it is important to plot the data in order to explore any relationships that might exist.



Plot 5: Monotonic relationship

In a monotonic relationship, the variables tend to move in the same relative direction, but not necessarily at a constant rate. In a linear relationship, the variables move in the same direction at a constant rate.

Plot 5 shows both variables increasing concurrently, but not at the same rate. This relationship is monotonic, but not linear. The Pearson correlation coefficient for these data is 0.843, but the Spearman correlation is higher, 0.948.

Linear relationships are also monotonic. For example, the relationship shown in Plot 1 is both monotonic and linear.

Covariance & Correlation

Covariance and Correlation are two important concepts commonly used in statistics.

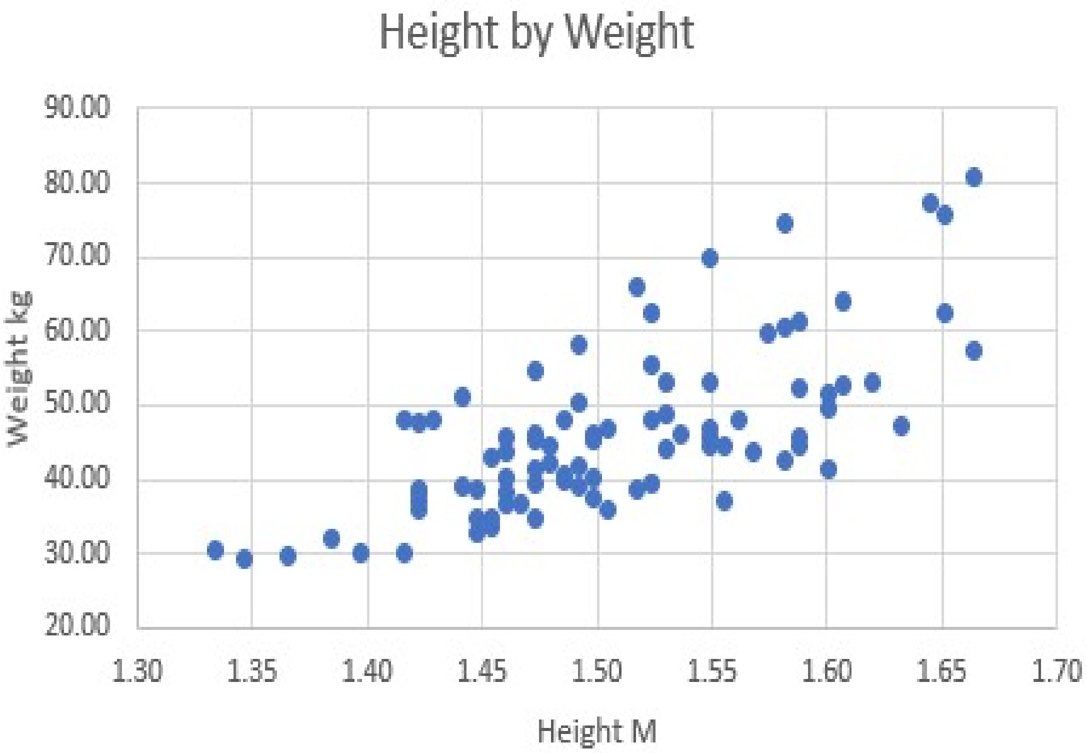
Covariance indicates the direction of the linear relationship between numerical variables.

If the value is positive, then the direction is upward.

If the value is negative, then the direction is downward.

For example, height and weight are related

scatterplot of the Height and Weight dataset.



The chart displays an upward slope.

person’s height increases, their weight also tends to increase.

Limitation of covariance: -

The covariance value won't help to compare the strength of the relationship between different variables, as the units of the variables may be the same or different.

This makes the covariance difficult to use to interpret the strength of direction.

For this reason, correlation is used, as it is a normalized version of covariance. And, the range of value is between - 1 to +1, which indicates both direction and strength of the linear relationship between 2 numerical variables.

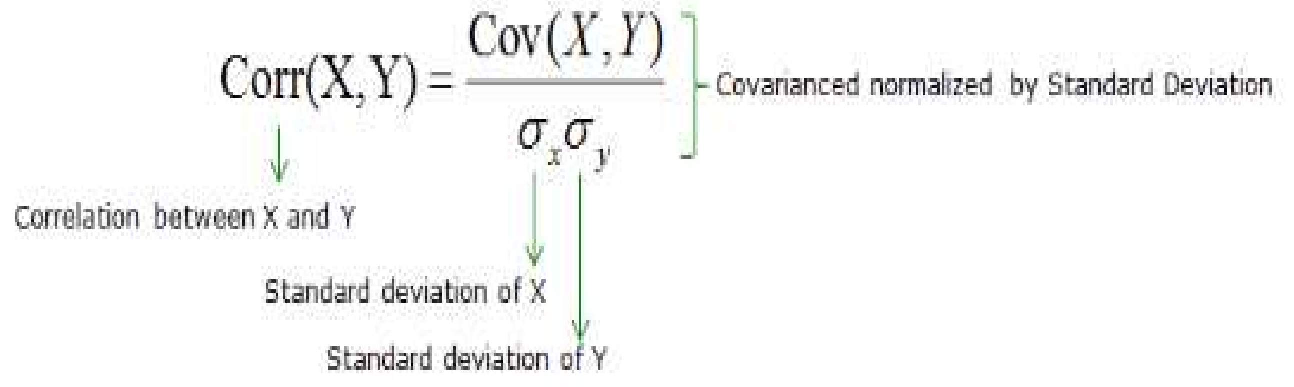
Correlation can be positive, negative and zero. If the correlation is

- positive: an increase in one of the variables results in an increase in the other
- negative: the variables are in opposite directions
- 0: then no relationship exists

Covariance and Correlation – Definition and Formula

A subset of the population is called a sample. Correlation and covariance are calculated on samples and not populations termed as sample covariance and correlation. Both terms define the relationship and dependency between the variables.

Correlation measures the association between the variables.



$$\mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_x \sigma_y}$$

Covarianced normalized by Standard Deviation
Correlation between X and Y
Standard deviation of X
Standard deviation of Y

Covariance explains the joint variability of the variables.

$$\mathbf{Cov(X,Y)=} \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{N-1}$$

Where

x_i = data value of x

y_i = data value of y

$\bar{x}$ = mean of x

$\bar{y}$ = mean of y

N = number of data values.

Applications: -

Evaluate the risk of particular stocks by comparing whether they move with or against each other.

For example, if the value of two stocks increases and decreases opposite one another, then they would be complementary, with minimal risk because they minimize financial loss by having one growing while the other shrinks.

You can also use covariance with correlation to determine if and how variables move together

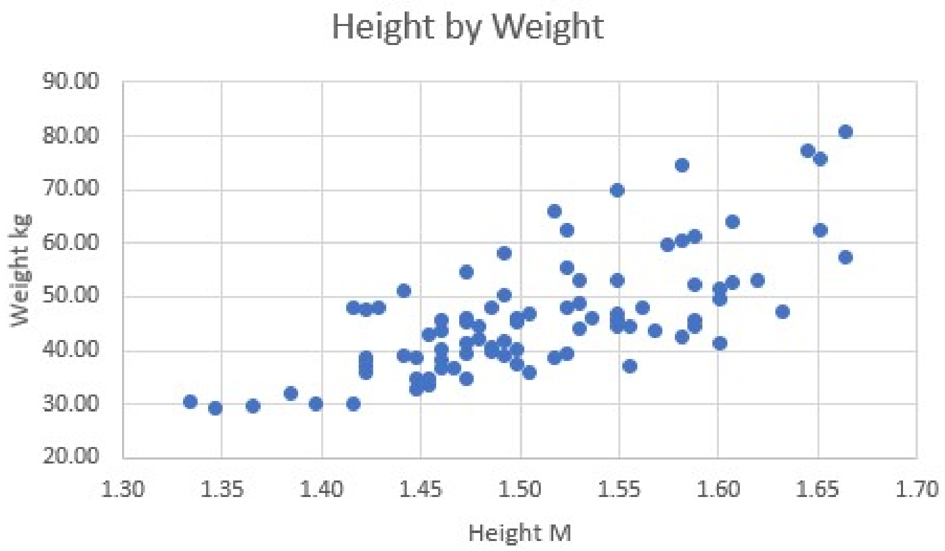
While covariance can tell you how two or more sets of the data move,

correlation can tell you what other factors influence that movement and if the two variables relate to each other.

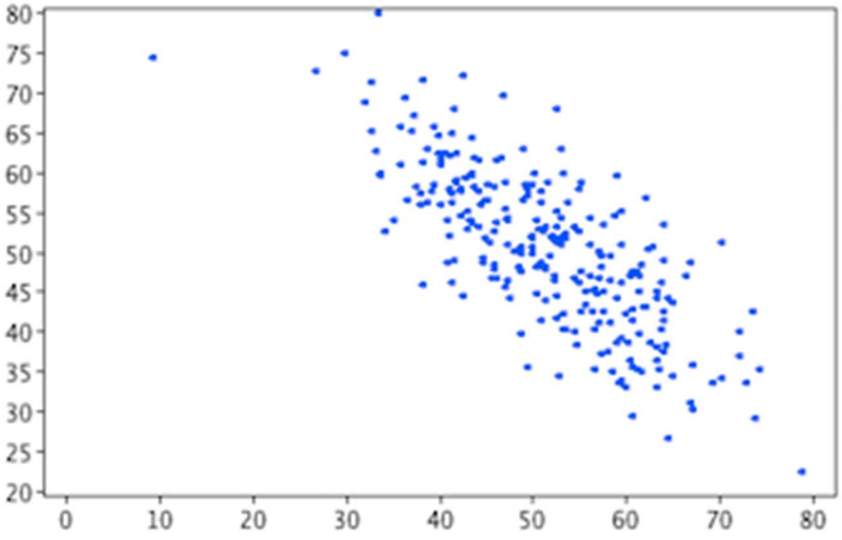
Types of Correlation

Correlation can be positive, negative, or no correlation.

Positive correlation means that as one data set increases, the other data set increases as well.

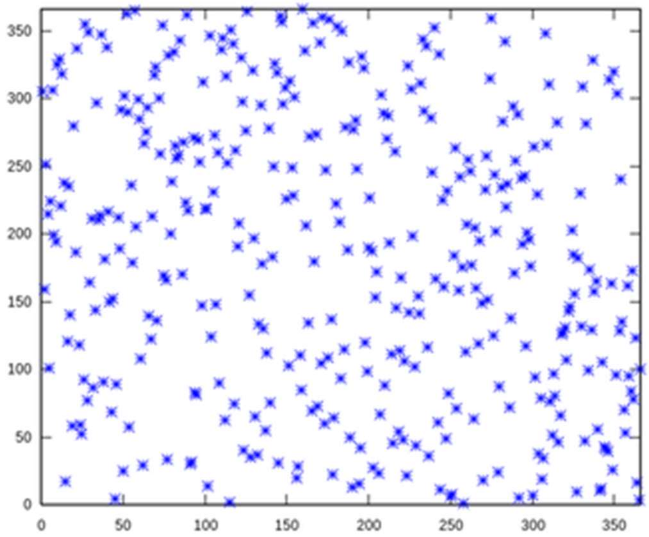


Negative correlation means that as one data set increases, the other decreases.



No correlation means that the two sets of data are not related at all.

means that one set of data does not increase or decrease with the other. No correlation is typically seen when the data points are very spread out .



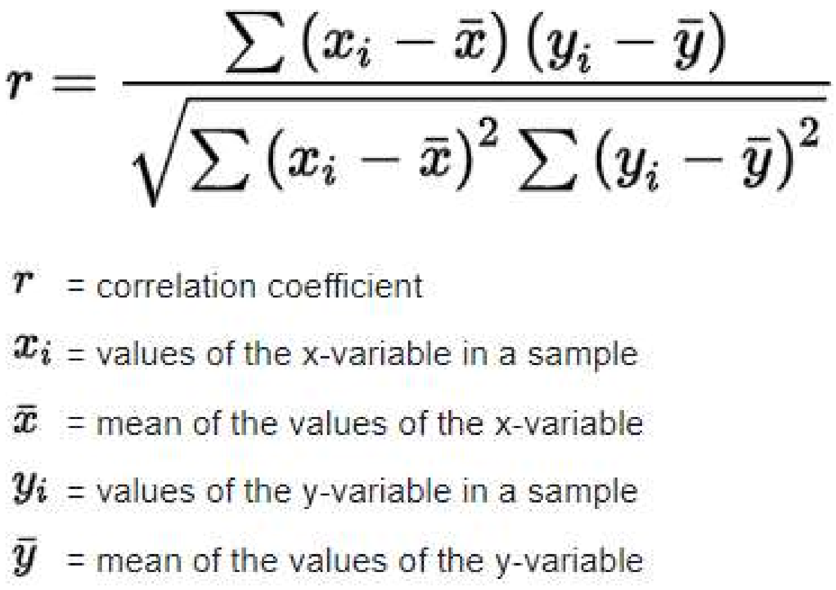
Positive and negative is not the only way to describe correlation; correlation can also be described by its strength.

Finding strength means: - Impact the change in one variable results in the impact in change in another variable.

We measure the correlation strength with the help of Correlation coefficient.

Correlation coefficient: - The Pearson’s correlation coefficient (r) is a measure that determines the degree to which the movement of two variables is associated. The value of correlation coefficient lies between -1 and +1.

Pearson’s co-correlation coefficient (r) Formula: -



$$r = \frac{\sum\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum\left(x_i - \bar{x}\right)^2 \sum\left(y_i - \bar{y}\right)^2}}$$

r = correlation coefficient

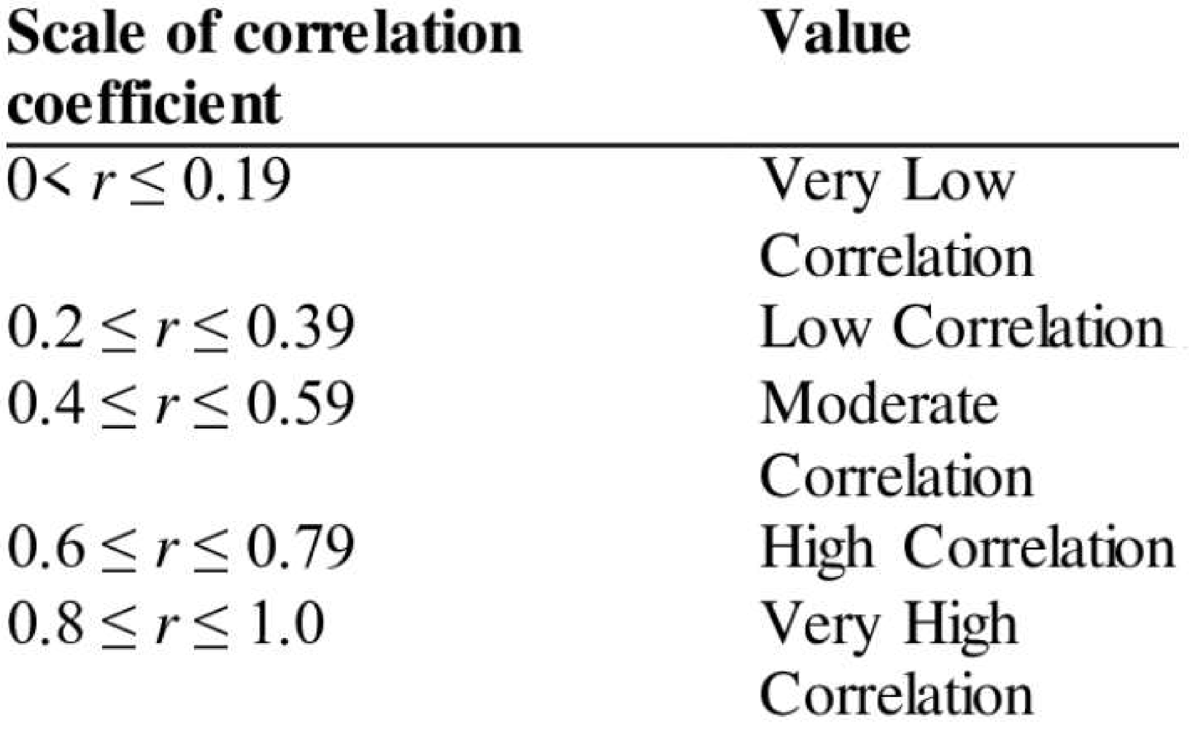
x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

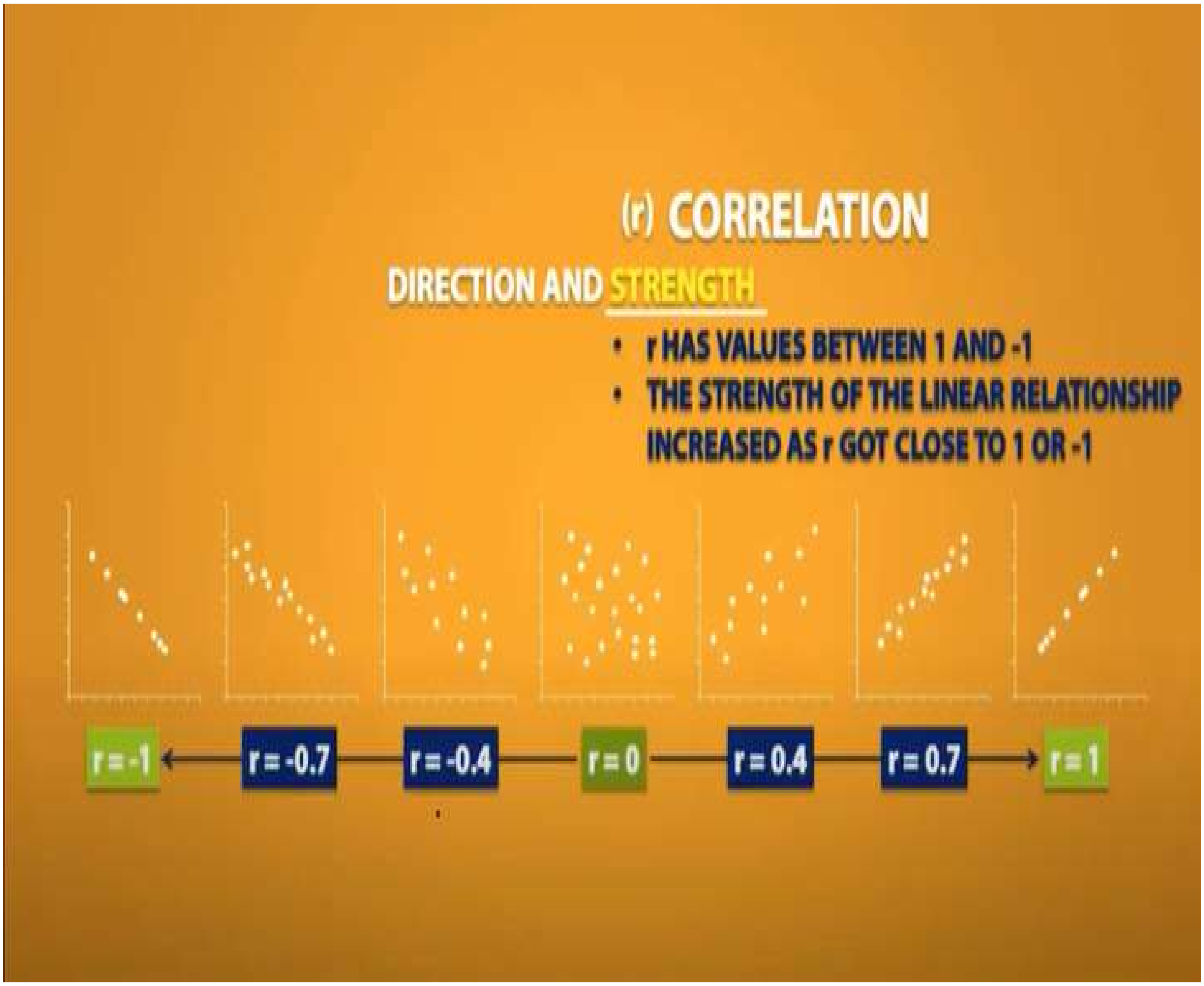
$\bar{y}$ = mean of the values of the y-variable

Pearson’s correlation coefficient is represented by the Greek letter rho (ρ) for the population parameter and r for a sample statistic. This coefficient is a single number that measures both the strength and direction of the linear relationship between two continuous variables



Scale of correlation coefficient	Value
$0< r \leq 0.19$	Very Low Correlation
$0.2 \leq r \leq 0.39$	Low Correlation
$0.4 \leq r \leq 0.59$	Moderate Correlation
$0.6 \leq r \leq 0.79$	High Correlation
$0.8 \leq r \leq 1.0$	Very High Correlation

The scale of Pearson's Correlation Coefficient



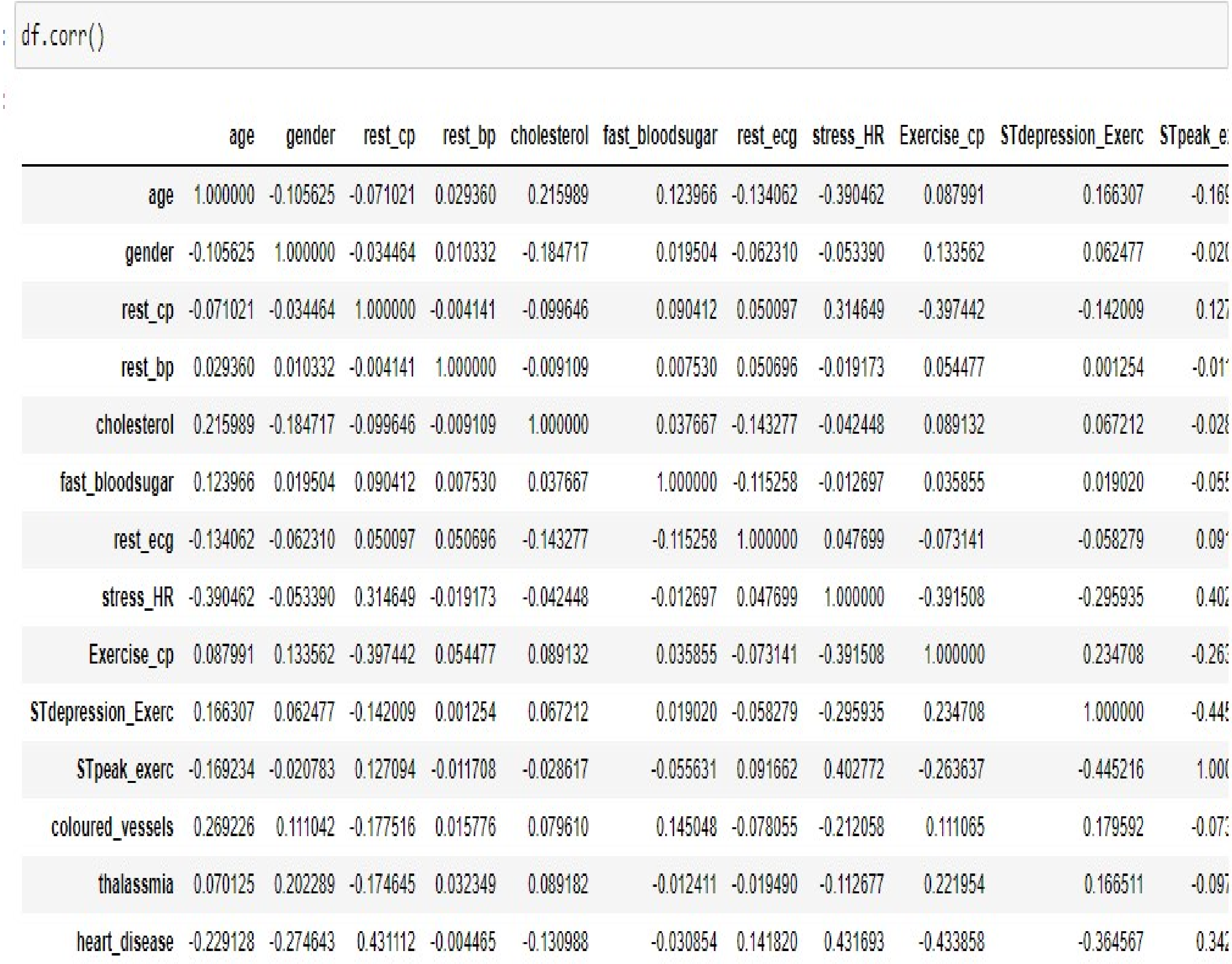
(r) CORRELATION

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

df.corr()

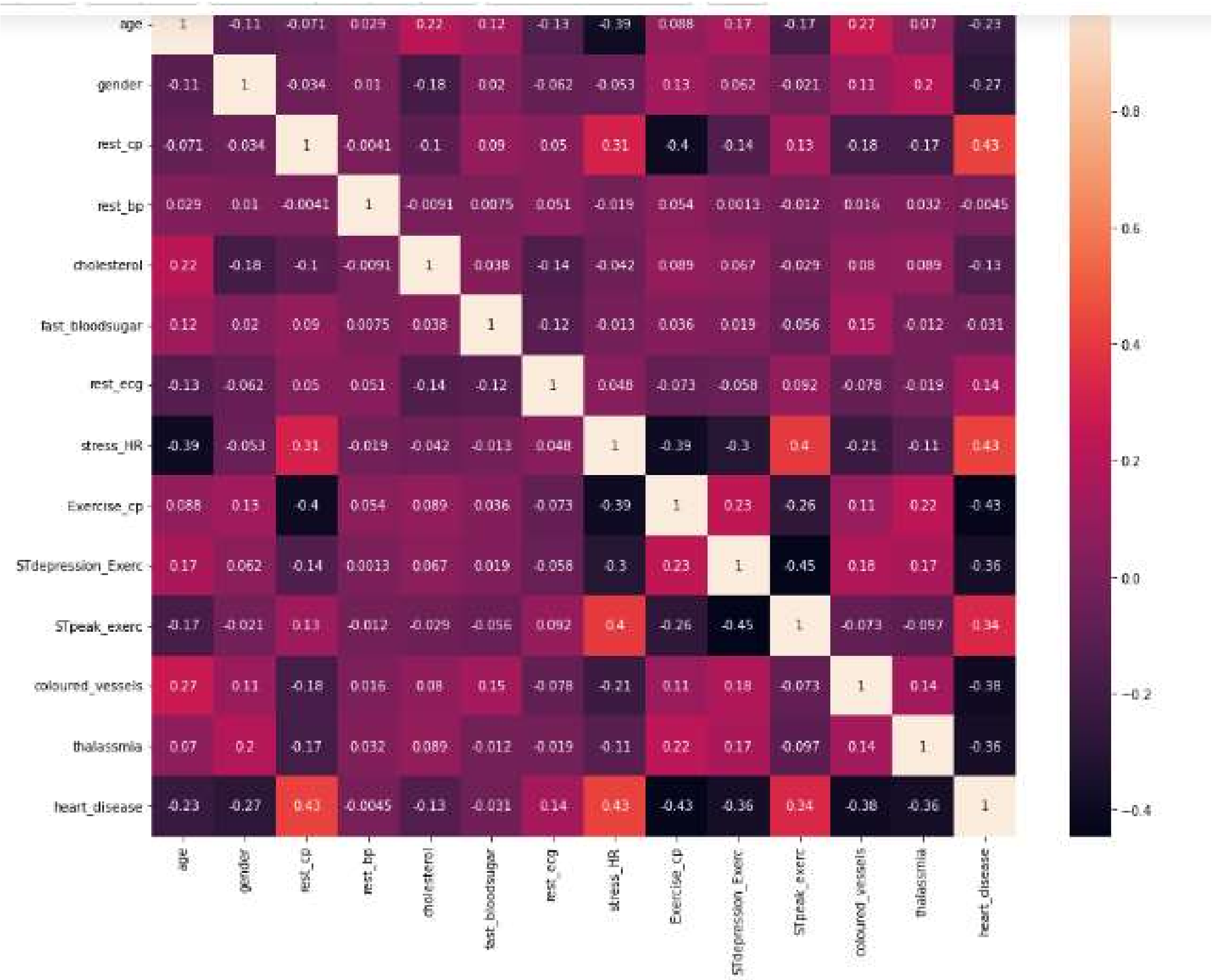
:- gives the Pearson’s co-correlation coefficient (r) all the features.



```
df.corr()
```

	age	gender	rest_cp	rest_bp	cholesterol	fast_bloodsugar	rest_ecg	stress_HR	Exercise_cp	STdepression_Exerc	STpeak_e
age	1.000000	-0.105625	-0.071021	0.029360	0.215989	0.123966	-0.134062	-0.390462	0.087991	0.166307	-0.16
gender	-0.105625	1.000000	-0.034464	0.010332	-0.184717	0.019504	-0.062310	-0.053390	0.133562	0.062477	-0.02
rest_cp	-0.071021	-0.034464	1.000000	-0.004141	-0.099646	0.090412	0.050097	0.314649	-0.397442	-0.142009	0.12
rest_bp	0.029360	0.010332	-0.004141	1.000000	-0.009109	0.007530	0.050696	-0.019173	0.054477	0.001254	-0.01
cholesterol	0.215989	-0.184717	-0.099646	-0.009109	1.000000	0.037667	-0.143277	-0.042448	0.089132	0.067212	-0.02
fast_bloodsugar	0.123966	0.019504	0.090412	0.007530	0.037667	1.000000	-0.115258	-0.012697	0.035855	0.019020	-0.05
rest_ecg	-0.134062	-0.062310	0.050097	0.050696	-0.143277	-0.115258	1.000000	0.047699	-0.073141	-0.058279	0.09
stress_HR	-0.390462	-0.053390	0.314649	-0.019173	-0.042448	-0.012697	0.047699	1.000000	-0.391508	-0.295935	0.40
Exercise_cp	0.087991	0.133562	-0.397442	0.054477	0.089132	0.035855	-0.073141	-0.391508	1.000000	0.234708	-0.26
STdepression_Exerc	0.166307	0.062477	-0.142009	0.001254	0.067212	0.019020	-0.058279	-0.295935	0.234708	1.000000	-0.44
STpeak_exerc	-0.169234	-0.020783	0.127094	-0.011708	-0.028617	-0.055631	0.091662	0.402772	-0.263637	-0.445216	1.00
coloured_vessels	0.269226	0.111042	-0.177516	0.015776	0.079610	0.145048	-0.078055	-0.212058	0.111065	0.179592	-0.07
thalassmia	0.070125	0.202289	-0.174645	0.032349	0.089182	-0.012411	-0.019490	-0.112677	0.221954	0.166511	-0.09
heart_disease	-0.229128	-0.274643	0.431112	-0.004465	-0.130988	-0.030854	0.141820	0.431693	-0.433858	-0.364567	0.34

Heat map option seaborn library :-



	age	gender	rest_cp	rest_bp	cholesterol	fast_bloodsugar	rest_ecg	stress_HR	Exercise_cp	STdepression_Exerc	STpeak_exerc	coloured_vessels	thalassmia	heart_disease
age	1	-0.11	-0.071	0.029	0.22	0.12	-0.13	-0.39	0.088	0.17	-0.17	0.27	0.07	-0.23
gender	-0.11	1	-0.034	0.01	-0.18	0.02	-0.062	-0.053	0.13	0.062	-0.021	0.11	0.2	-0.27
rest_cp	-0.071	-0.034	1	-0.0041	-0.1	0.09	0.05	0.31	-0.4	-0.14	0.13	-0.18	-0.17	0.43
rest_bp	0.029	0.01	-0.0041	1	-0.0091	0.0075	0.051	-0.019	0.054	0.0013	-0.012	0.016	0.032	-0.0045
cholesterol	0.22	-0.18	-0.1	-0.0091	1	0.038	-0.14	-0.042	0.089	0.067	-0.029	0.08	0.089	-0.13
fast_bloodsugar	0.12	0.02	0.09	0.0075	0.038	1	-0.12	-0.013	0.036	0.019	-0.056	0.15	-0.012	-0.031
rest_ecg	-0.13	-0.062	0.05	0.051	-0.14	-0.12	1	0.048	-0.073	-0.058	0.092	-0.078	-0.019	0.14
stress_HR	-0.39	-0.053	0.31	-0.019	-0.042	-0.013	0.048	1	-0.39	-0.3	0.4	-0.21	-0.11	0.43
Exercise_cp	0.088	0.13	-0.4	0.054	0.089	0.036	-0.073	-0.39	1	0.23	-0.26	0.11	0.22	-0.43
STdepression_Exerc	0.17	0.062	-0.14	0.0013	0.067	0.019	-0.058	-0.3	0.23	1	-0.45	0.18	0.17	-0.36
STpeak_exerc	-0.17	-0.021	0.13	-0.012	-0.029	-0.056	0.092	0.4	-0.26	-0.45	1	-0.073	-0.097	0.34
coloured_vessels	0.27	0.11	-0.18	0.016	0.08	0.15	-0.078	-0.21	0.11	0.18	-0.073	1	0.14	-0.38
thalassmia	0.07	0.2	-0.17	0.032	0.089	-0.012	-0.019	-0.11	0.22	0.17	-0.097	0.14	1	-0.36
heart_disease	-0.23	-0.27	0.43	-0.0045	-0.13	-0.031	0.14	0.43	-0.43	-0.36	0.34	-0.38	-0.36	1

Multicollinearity and Variance inflation factor (VIF)

- **Multicollinearity refers to high correlation in more than two independent variables.**
- Similarly, collinearity refers to a high correlation between two independent variables.

variance inflation factor (VIF), which measures the correlation and strength of correlation between the predictor variables in a regression model.

The value for VIF starts at 1 and has no upper limit. A general rule of thumb for interpreting VIFs is as follows:

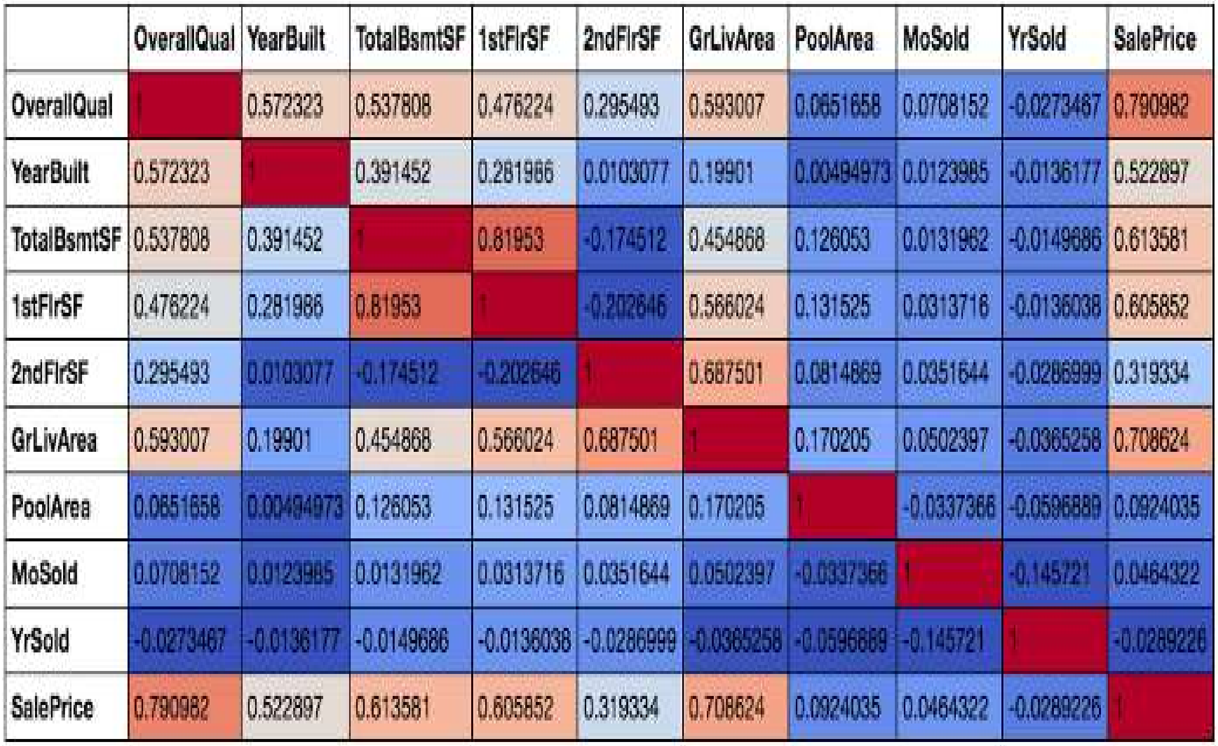
- A value of 1 indicates there is no correlation between a given predictor variable and any other predictor variables in the model.
- A value between 1 and 5 indicates moderate correlation between a given predictor variable and other predictor variables in the model, but this is often not severe enough to require attention.
- A value greater than 5 indicates potentially severe correlation between a given predictor variable and other predictor variables in the model. In this case, the coefficient estimates and p-values in the regression output are likely unreliable.

Multicollinearity in Regression

Multicollinearity happens when independent variables in the regression model are highly correlated to each other. It makes it hard to interpret of model and also creates an overfitting problem.

Why Multi-Collinearity is a problem?

When independent variables are highly correlated, change in one variable would cause change to another and so the model results fluctuate significantly.



	OverallQual	YearBuilt	TotalBsmtSF	1stFlrSF	2ndFlrSF	GrLivArea	PoolArea	MoSold	YrSold	SalePrice
OverallQual	1	0.572323	0.537808	0.476224	0.295493	0.593007	0.0651658	0.0708152	-0.0273467	0.790982
YearBuilt	0.572323	1	0.391452	0.281986	0.0103077	0.19901	0.00494973	0.0123985	-0.0136177	0.522897
TotalBsmtSF	0.537808	0.391452	1	0.81953	-0.174512	0.454868	0.126053	0.0131962	-0.0149686	0.613581
1stFlrSF	0.476224	0.281986	0.81953	1	-0.202646	0.566024	0.131525	0.0313716	-0.0136038	0.605852
2ndFlrSF	0.295493	0.0103077	-0.174512	-0.202646	1	0.687501	0.0814869	0.0351644	-0.0286999	0.319334
GrLivArea	0.593007	0.19901	0.454868	0.566024	0.687501	1	0.170205	0.0502397	-0.0365258	0.708624
PoolArea	0.0651658	0.00494973	0.126053	0.131525	0.0814869	0.170205	1	-0.0337366	-0.0596689	0.0924035
MoSold	0.0708152	0.0123985	0.0131962	0.0313716	0.0351644	0.0502397	-0.0337366	1	-0.145721	0.0464322
YrSold	-0.0273467	-0.0136177	-0.0149686	-0.0136038	-0.0286999	-0.0365258	-0.0596689	-0.145721	1	-0.0289226
SalePrice	0.790982	0.522897	0.613581	0.605852	0.319334	0.708624	0.0924035	0.0464322	-0.0289226	1

color scaled correlation matrix for housing data

. There is one pair of independent variables with more than 0.8 correlation which are total basement surface area and first-floor surface area. Houses with larger basement areas tend to have bigger first-floor areas as well and so a high correlation should be expected.