**Correlation:**

Correlation explains how one or more variables are related to each other. These variables can be input data features which have been used to forecast our target variable.

Correlation, statistical technique which determines how one variables moves/changes in relation with the other variable. It gives us the idea about the degree of the relationship of the two variables. It’s a bi-variate analysis measure which describes the association between different variables. In most of the business it’s useful to express one subject in terms of its relationship with others.

For example: No of testing vs no of positive cases in Corona.

1. If two variables are closely correlated, then we can predict one variable from the other.

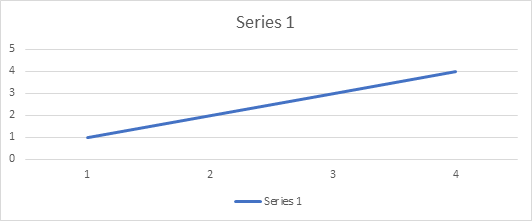
2. Correlation plays a vital role in locating the important variables on which other variables depend.

3. It’s used as the foundation for various modeling techniques.

4. Proper correlation analysis leads to better understanding of data.

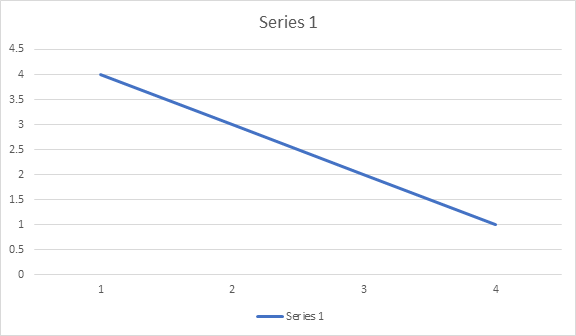
5. Correlation contribute towards the understanding of causal relationship (if any).

**Positive Correlation:**Two features (variables) can be positively correlated with each other. It means that when the value of one variable increase then the value of the other variable(s) also increases.



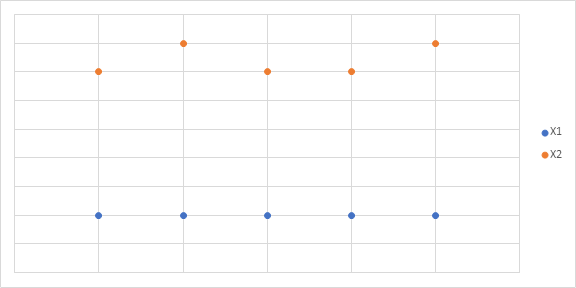
**Positive Correlation**

**Negative Correlation:**Two features (variables) can be negatively correlated with each other. It means that when the value of one variable increase then the value of the other variable(s) decreases.



**Negative Correlation**

**No Correlation:**Two features (variables) are not correlated with each other. It means that when the value of one variable increase or decrease then the value of the other variable(s) doesn’t increase or decreases.



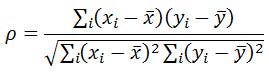
**No Correlation**

**Measures of Correlation**

**1. Pearson’s correlation of coefficient**

Pearson’s correlation coefficient is a measure of the strength of a linear association between two variables and is denoted by r. Basically, a Pearson’s correlation attempts to draw a line of best fit through two variables' data. The Pearson correlation coefficient, r, indicates how far away all these data points are to this line of best fit.

1. In Pearson’s correlation coefficient, variables can be measured in entirely different units. For example, we can correlate the height of a person with his weight. It is designed in such a way that unit of measurement can’t affect the study of covariation.
2. Pearson’s correlation coefficient(r) is a unitless measure of correlation and doesn’t change in the effect of origin or scale shift measurement.
3. It doesn’t take into consideration whether a variable has been classified as a dependent or independent variable. It treats all variables equally. We might want to find out whether basketball performance is correlated to a person’s height. But if we determine whether a person’s height was determined by their basketball performance (which makes no sense), the result will be the same.

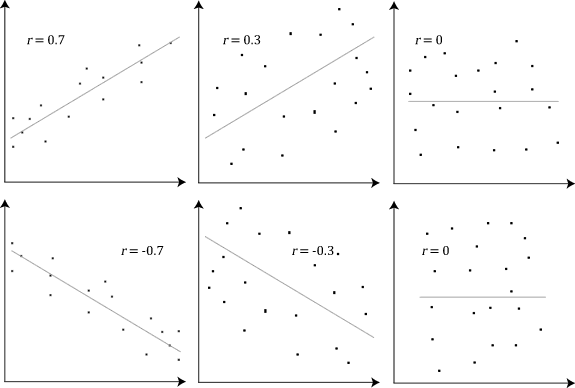


Pearson’s correlation coefficient formula

where xi, yi, are the variables and xbar, ybar, are the mean, respectively

**Properties:**

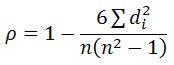
1. The range of r is between [-1,1].
2. The computation of r is independent of the change of origin and scale of measurement.
3. r = 1 (perfectly positive correlation), r =-1 (perfectly negative correction)  
   r = 0 (no correlation)



r with a linear relationship plot

**2. Spearman’s correlation of coefficient**

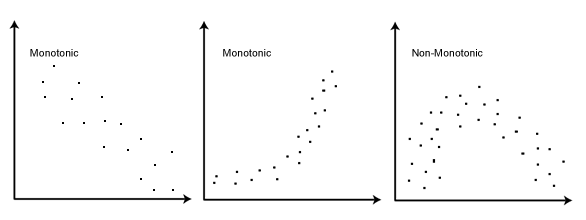
Spearman’s correlation coefficient is a non-parametric measure of the strength and direction of association that exists between two variables measured on at least an ordinal scale. The symbol rs or ρ denotes it. Such as:  
We may like to find out the correlation between ranks given by two Judges to candidates in an interview, marks secured by a group of students in five subjects, etc.



Spearman’s correlation coefficient formula

where n = total number of observations, di = (xi-Yi) where xi and yi are the observations

Spearman’s correlation determines the monotonic relationship's strength and direction between two variables rather than the strength and direction of the linear relationship between your two variables, which is what Pearson’s correlation determines.



ρ with monotonic plot

**Properties:**

1. The range of r is between [-1,1].
2. Preserves all properties of r.
3. As this is based on ordinal data, it doesn’t depend on any specific distribution(that is why we called non-parametric measure)

***Note: The Spearman correlation can be used when the assumptions of the Pearson correlation are markedly violated.***

**3. Kendall’s Tau correlation of coefficient**

Kendall’s Tau is a non-parametric measure of relationships between columns of ranked data. The Tau correlation coefficient returns a value of 0 to 1, where:

0 is no relationship,  
1 is a perfect relationship.

A quirk of this test can also produce negative values (from -1 to 0). Unlike a linear graph, a negative relationship doesn’t mean much with ranked columns (other than you perhaps switched the columns around), so remove the negative sign when you’re interpreting Tau.

[***Kendall’s Tau = (C — D / C + D)***](https://www.statisticshowto.com/kendalls-tau/)***,***where C is the number of concordant pairs and D is the number of discordant pairs.

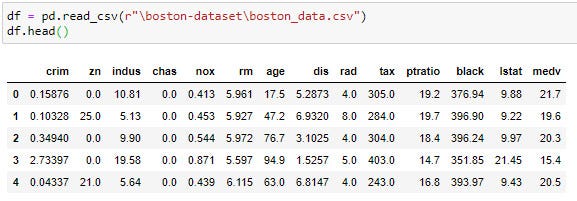
**Application in Machine Learning**

Correlation is a highly applied technique in machine learning during data analysis and data mining. It can extract key problems from a given set of features, which can later cause significant damage during the fitting model.  
Data having non-correlated features have many benefits. Such as:  
1. Learning of Algorithm will be faster  
2. Interpretability will be high  
3. Bias will be less

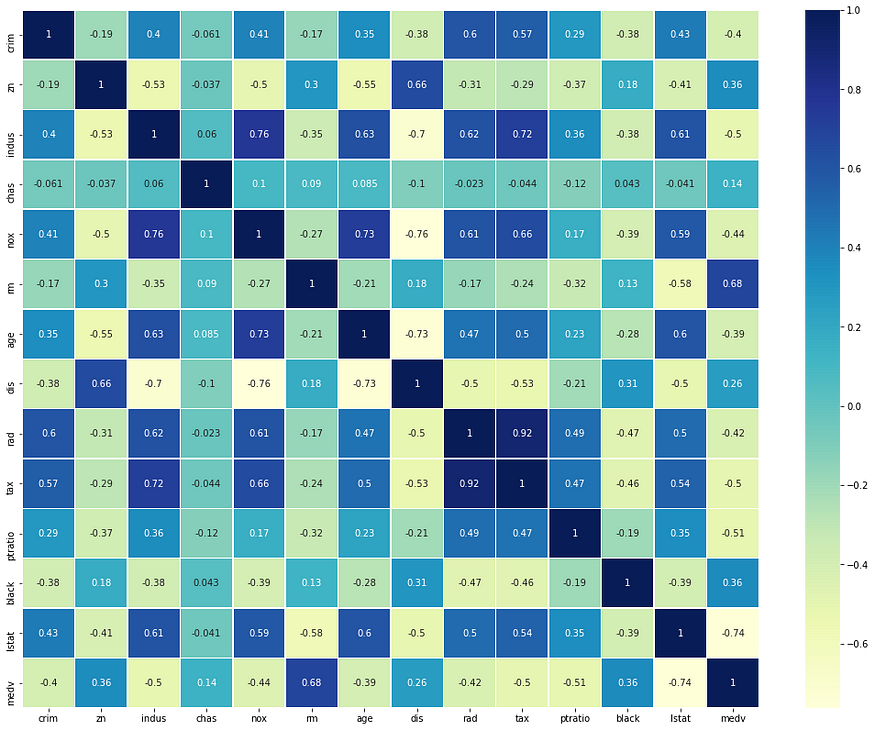
**How to Conduct Correlation Analysis**

To conduct a correlation analysis, you will need to follow these steps:

1. **Identify Variable:**Identify the two variables that we want to correlate. The variables should be quantitative, meaning that they can be represented by numbers.
2. **Collect data :**Collect data on the two variables. We can collect data from a variety of sources, such as surveys, experiments, or existing records.
3. **Choose the appropriate correlation coefficient.** The Pearson correlation coefficient is the most commonly used correlation coefficient, but there are other correlation coefficients that may be more appropriate for certain types of data.
4. **Calculate the correlation coefficient.** We can use a statistical software package to calculate the correlation coefficient, or you can use a formula.
5. **Interpret the correlation coefficient.** The correlation coefficient can be interpreted as a measure of the strength and direction of the linear relationship between the two variables.



13 numerical and one categorical(chas) feature is present

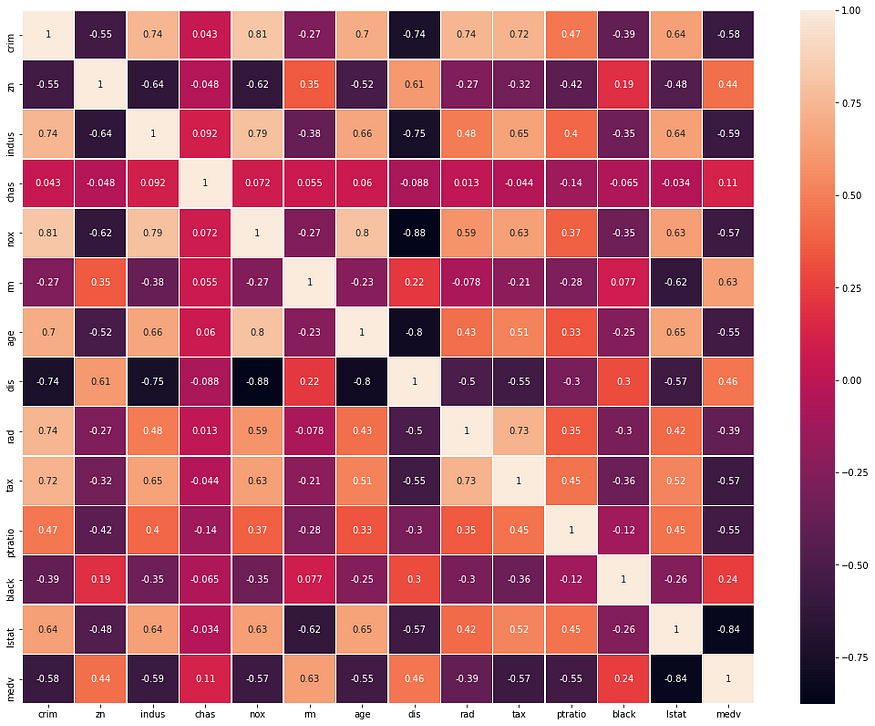


Pearson’s correlation coefficient heatmap

**Observations**

1. “tax” and “rad” columns are highly correlated with a value of 0.92(positive correlation).
2. Some of the features are negatively correlated, and their correlation value is negatively high. Such as “lstat” vs. “medv,” “dis,” vs. “indus,” “dis,” vs. “age.”

From the above observations, we can conclude that there is hidden covariation between the tax rate and the index of accessibility to radial highways. We can derive a relationship that if the house accessibility to highways is high, then the full-value property-tax rate will also be higher.  
This may sound like a genuine thing, because of the houses with high price generally nearer to market, good amenities, highways, etc.



Spearman’ correlation coefficient heatmap

**Observations**

1. “chas” is a categorical feature, and as we are taking Spearman’s correlation coefficient into account, It has also been included in the correlation.

**Effect of Multicollinearity**

A key goal in regression analysis in machine learning is to isolate each independent variable's relationship and the dependent variable. So change in one independent variable shouldn’t affect any other variables in the given data. However, when independent variables are correlated, it indicates that one variable's changes are associated with shifts in another variable. As the severity of the multicollinearity increases, so do these problematic effects.  
So during the fitting of the model, a small change in one variable can lead to a significant amount of swing in the model output. However, these issues affect only those independent variables that are correlated.

**Solution:**

1. The severity of the problems increases with the degree of multicollinearity. Therefore, if you have only moderate multicollinearity, you may not need to resolve it.
2. Multicollinearity affects only the specific independent variables that are correlated. Therefore, if multicollinearity is not present for the independent variables you are particularly interested in, you may not need to resolve it. Suppose your model contains the experimental variables of interest and some control variables.  
   If high multicollinearity exists for the control variables but not the experimental variables, you can interpret the experimental variables without problems.
3. If one of the collinear features has not much to contribute much to prediction or classification for distance-based ml algorithms, we can drop the analysis feature.

[What is Correlation Analysis? - GeeksforGeeks](https://www.geeksforgeeks.org/what-is-correlation-analysis/)

[Correlation: Meaning, Significance, Types and Degree of Correlation - GeeksforGeeks](https://www.geeksforgeeks.org/correlation-meaning-significance-types-and-degree-of-correlation/)