International
Institute of Information
Technology Bangalore

NAME - SHASHANK BHATNAGAR

MACHINE LEARNING ASSIGNMENT SUBJECTIVE QUESTIONS

Assignment-based Subjective Questions

Q1 - From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Answer :

Please find below inferences based on the analysis of the categorical variables from the dataset

- Year wise : In year 2019 we have around 62.33 % of total Bike sharing i.e. Bike sharing is increasing with the years and the with the increase in the age of Bike Sharing Company

- Month wise : From may to oct we have observe a upward trend in bike sharing

- Season wise : Highest no of bike sharing will be in fall season and then in summer and then in winter and spring have least no of bike sharing.

- Weather-sit : Company Many Notice increase in Bike sharing count for below weather
    - 1: Clear, Few clouds, Partly cloudy, Partly cloudy
    - 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist

- Weather-sit : Bike sharing count will decrease with snow, rainfall and thunderstorm .

- Holiday : Bike sharing is more on non holidays so it seems people spend time with family during holidays or may be using private Vehicles.

- Working and non working days are almost uniformly distributed but bike sharing is slightly more on the working day

- With the clear weather (cloudy & Misty) during Fall & Summer season between May to Oct month for the year 2019 with moderate temp , low humidity and less windspeed we can experience more bike sharing.

Q2 - Why is it important to use drop_first=True during dummy variable creation?

Answer :

Dummy variables will be created with one-hot *encoding* and each attribute will have a value of either 0 or 1, representing the presence or absence of that attribute.

Dummy Variable Trap:

The Dummy variable trap is a scenario where there are attributes that are highly correlated (**Multi collinear**) and one variable predicts the value of others. When we use *one-hot encoding* for handling the categorical data, then one dummy variable (attribute) can be predicted with the help of other dummy variables. Hence, one dummy variable is highly correlated with other dummy variables. Using all dummy variables for regression models leads to a ***dummy variable trap***. So, the regression models should be designed to exclude one dummy variable that's why we us drop_first=True during dummy variable creation
For : E.g. Let's consider the case of gender having two values *male* (0 or 1) and *female* (1 or 0). Including both the dummy variable can cause redundancy because if a person is not male in such case that person is a female, hence, we don't need to use both the variables in regression models. This will protect us from the dummy variable trap.

Q3 - Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

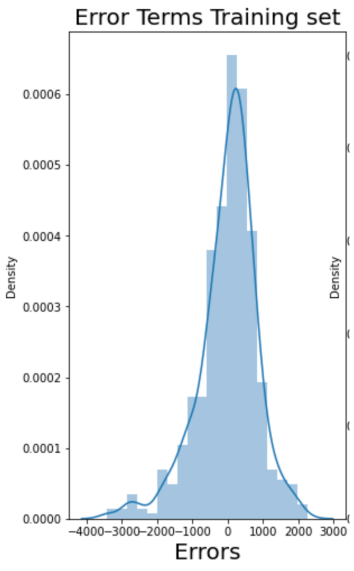
Answer:

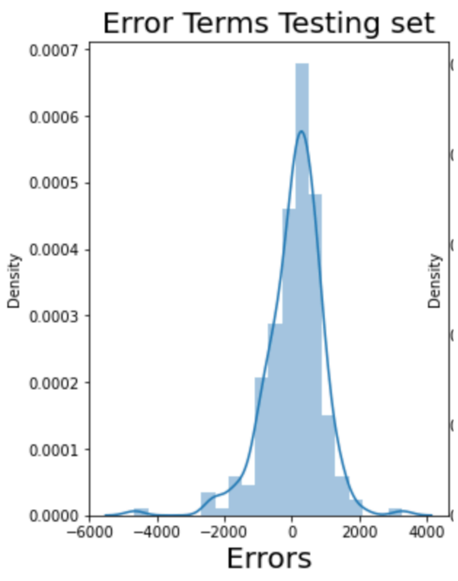
- temp & atemp have higher correlations with the cnt (i.e. target) variable

Q4 - How did you validate the assumptions of Linear Regression after building the model on the training set?

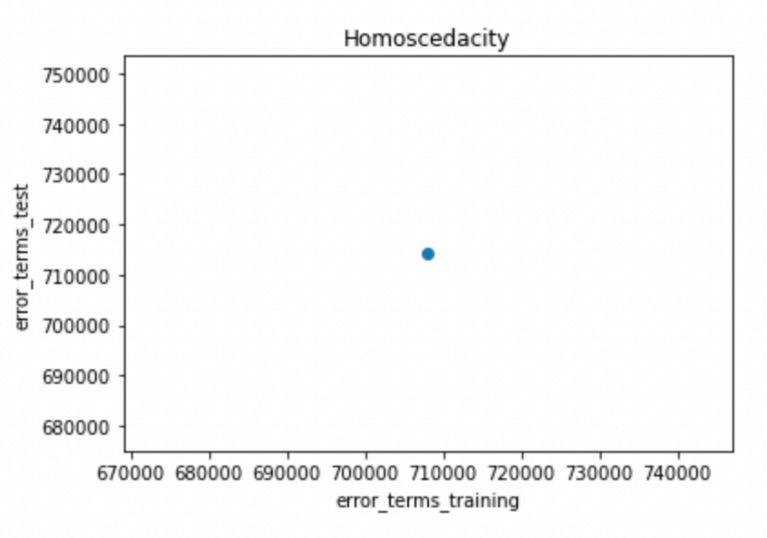
Answer:

- Using Residual Analysis and checked wether the **Error terms are normally distributed** with mean zero using distplot.

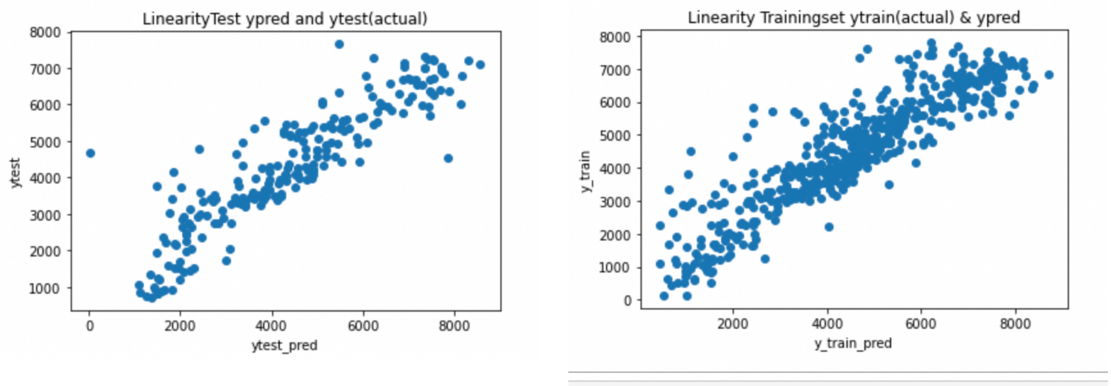

Error Terms Training set


Error Terms Testing set

- **Homoscedacity** : Calculated the **Error terms for training and testing set** and checked the variance error terms must have constant variance at different data points .



- **Linear Relationship : **Perform the Linearity check plotted actual and predicted values to test the linearity.

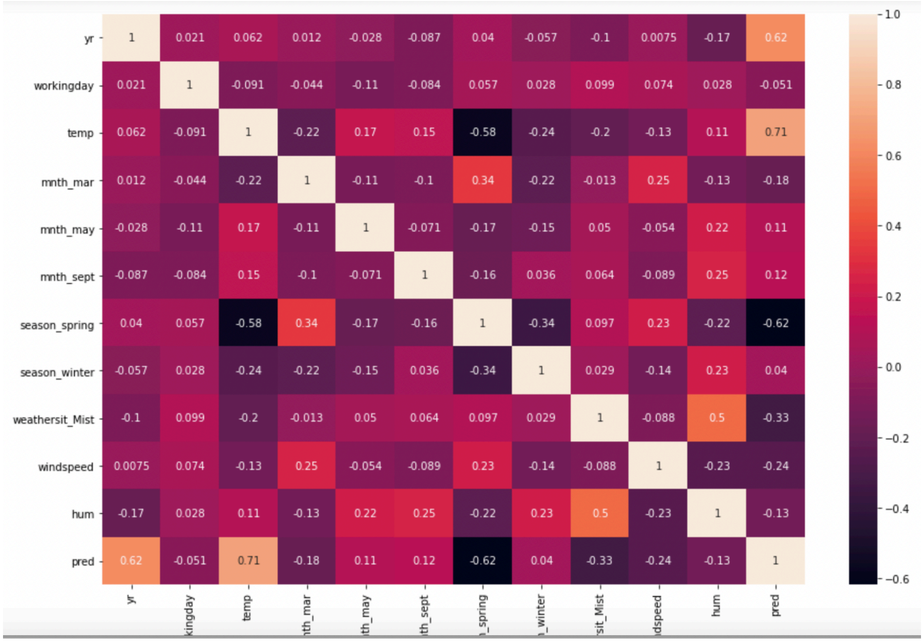


- **Error terms are independent** of each other training set error terms are independent of testing set and vice versa.

Q5 - Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Answer:

- **Yr, temp and season_spring** are the top 3 features contributing significantly towards explaining the demand of the shared bikes



General Subjective Questions

Q1. Explain the linear regression algorithm in detail.

Answer:

In Order to understand Linear Regression Algorithm we need to understand below concepts

- Regression
- Linear Regression
- Cost Function
- Gradient Descent
- Assumptions

Regression :

Regression analysis is nothing but a predictive modelling methodology that aims to investigate the relation that exists between independent variables or predictors and dependent variables or targets.

This is done by fitting a line or curve to different data points in a way that we can minimise the difference in data point distances from the line, or the curve.

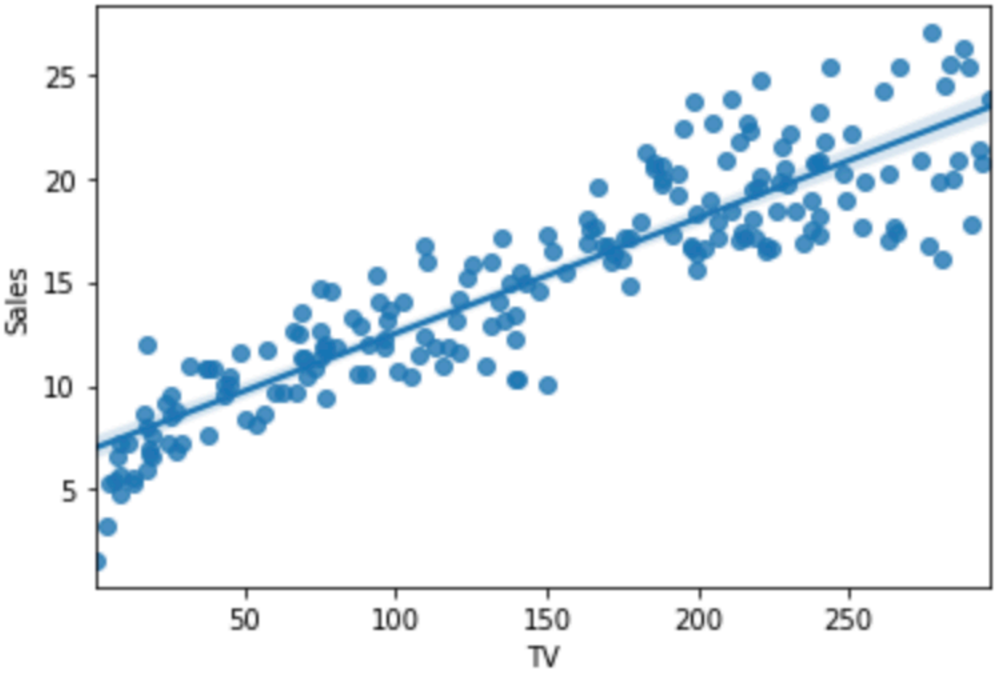
Regression shows a line or curve that passes through all the data points on a target-predictor graph in such a way that the vertical distance between the data points and the regression line is minimum.

Linear Regression :

Linear Regression shows the linear relationship between the independent variable x and dependent variable y , If we have only one independent variable then it will be a Simple Linear Regression problem and if we have multiple independent variables then it will be a Multiple Linear Regression Function

For eg : We can observe a linear relationship between TV and Sales equation will be like

Simple Linear Regression $Sales_y = B_0 + B_1 * TV$



Multiple Linear Regression $Sales_y = B_0 + B_1 * TV + B_2 * Radio$

Goal : Our goal is to find the best fit line by minimising the cost function

COST FUNCTION:

The cost function helps to figure out the best possible values for B_0 and B_1, which provides the best fit line for the data points.

Cost function optimises the regression coefficients or weights and measures how a linear regression model is performing. The cost function is used to find the accuracy of the **mapping function** that maps the input variable to the output variable.

$$CostFunction = \frac{1}{2m}\sum_{n=1}^{m}(y_{sales_{predicted}} - y_{sales_{actual}})^2$$

m= no of samples

$$y_{sales_{predicted}} = B_0 + B_1 * X_{TV}$$

We can minimise the cost function either by minimising the partial differential or by using Gradient Descent

Gradient Descent :

Gradient descent is a method of updating B_0 and B_1, to minimise the cost function (MSE). A regression model uses gradient descent to update the coefficients of the line by reducing the cost function by a random selection of coefficient values and then iteratively update the values to reach the minimum cost function.

Gradient descent is an iterative optimisation algorithm to find the minimum of a function

How Gradient Descent works :

The algorithm starts with some value of m and c (usually starts with m=0, c=0). We calculate MSE (cost function) at point m=0, c=0. Let say the MSE (cost) at m=0, c=0 is 100. Then we reduce the value of m and c by some amount (Learning Step). We will notice a decrease in MSE (cost). We will continue doing the same until our loss function is a very small value or ideally 0 (which means 0 error or 100% accuracy)

We need to make sure we should trap in local minima goal is to identify the Global Minima

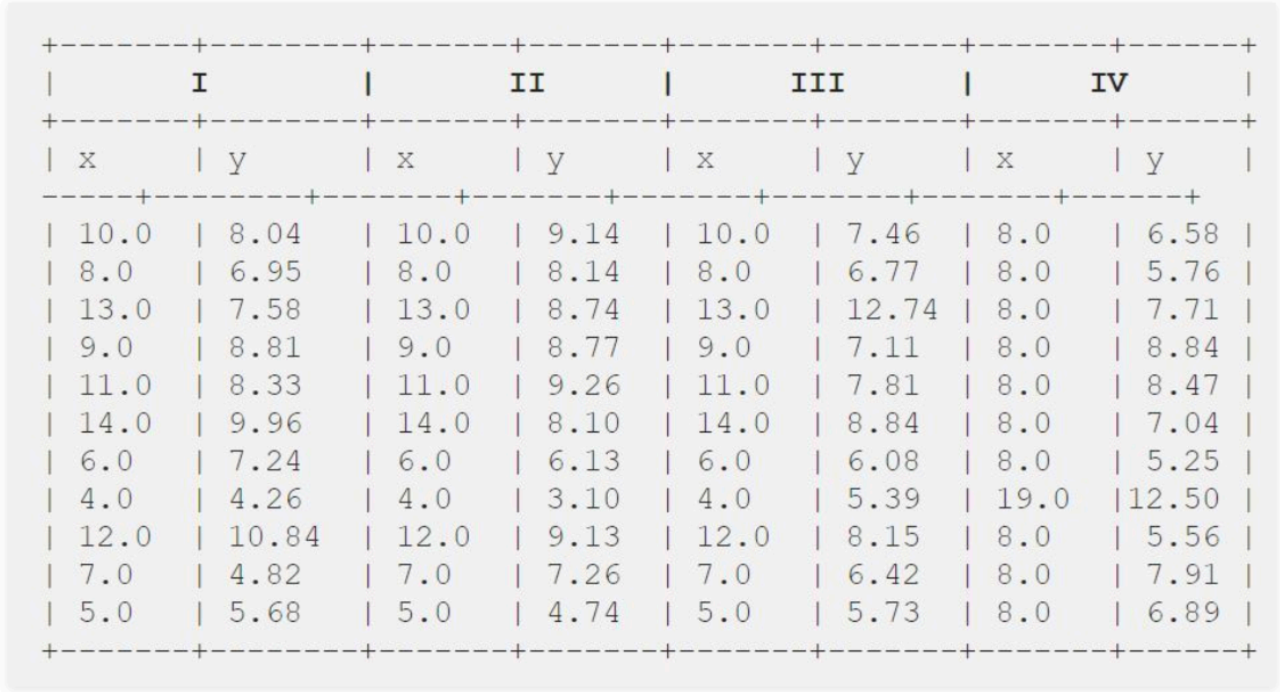
Assumptions:

The assumptions of simple linear regression were:
1. Linear relationship between X and Y
2. Error terms are normally distributed (notX,Y)
3. Error terms are independent of each other
4. Error terms have constant variance(homoscedasticity)

- No assumption on the distribution of X and Y, just that the error terms have to have a normal distribution.

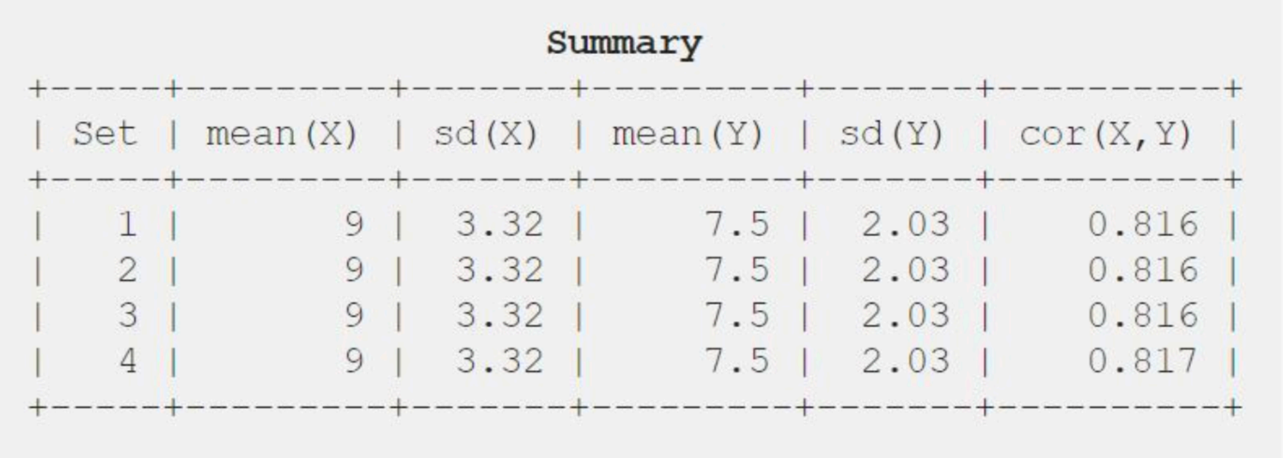
Q2. **Explain the Anscombe's quartet in detail.**

Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscomb'e to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.



I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

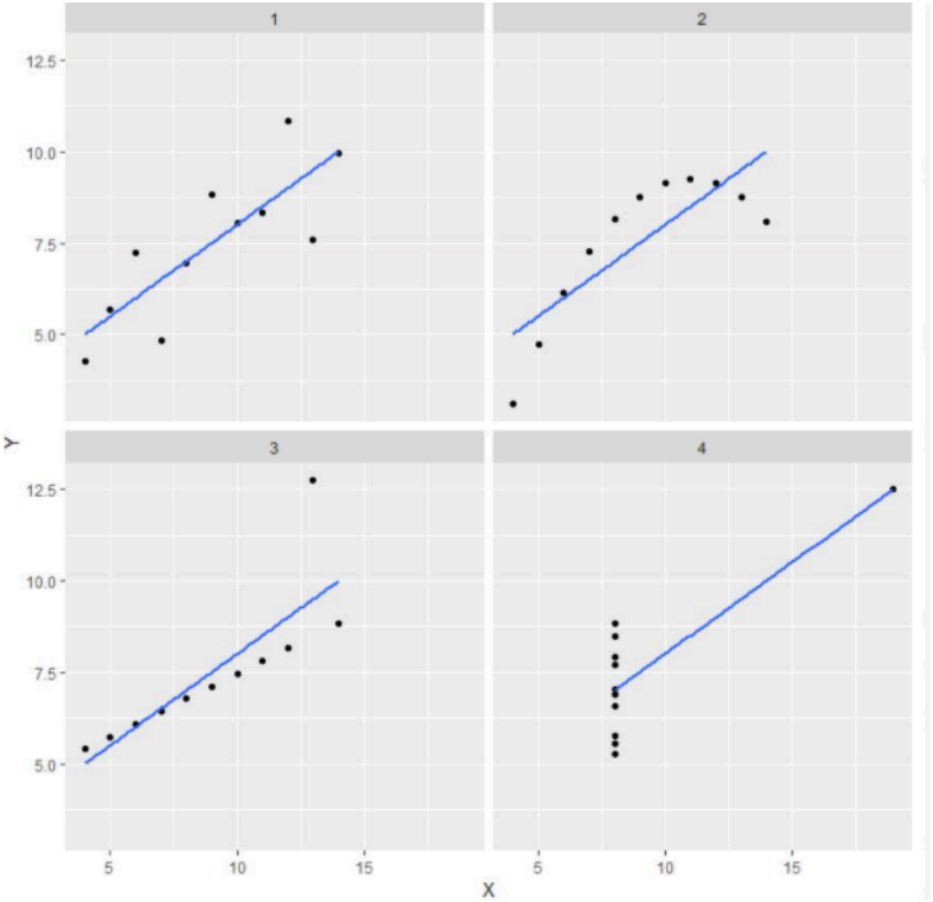
Using only descriptive statistics and found the mean, standard deviation, and correlation between x and y and are same



Summary

Set	mean(X)	sd(X)	mean(Y)	sd(Y)	cor(X,Y)
1	9	3.32	7.5	2.03	0.816
2	9	3.32	7.5	2.03	0.816
3	9	3.32	7.5	2.03	0.816
4	9	3.32	7.5	2.03	0.817

But on visualising It is mentioned in the definition that Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed.



- In the first one(top left) if you look at the scatter plot you will see that there seems to be a linear relationship between x and y.

- In the second one(top right) if you look at this figure you can conclude that there is a non-linear relationship between x and y.

- In the third one(bottom left) you can say when there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line.

- Finally, the fourth one(bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient.

Application:

The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

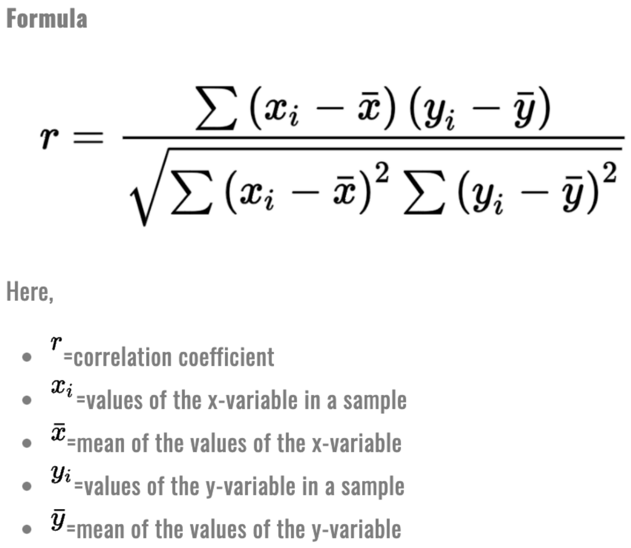
Q3. **What is Pearson’s R?**

In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations; thus it is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1.

The Pearson correlation coefficient has a high statistical significance. It looks at the relationship between two variables. It seeks to draw a line through the data of two variables to show their relationship. This linear relationship can be positive or negative

The Pearson's correlation coefficient varies between -1 and +1 where:

1. **Perfect:** If the value is near ± 1, then it said to be a perfect correlation: as one variable increases, the other variable tends to also increase (if positive) or decrease (if negative).
2. **High degree:** If the coefficient value lies between ± 0.50 and ± 1, then it is said to be a strong correlation.
3. **Moderate degree:** If the value lies between ± 0.30 and ± 0.49, then it is said to be a medium correlation.
4. **Low degree:** When the value lies below ± .29, then it is said to be a small correlation.
5. **No correlation:** When the value is zero.

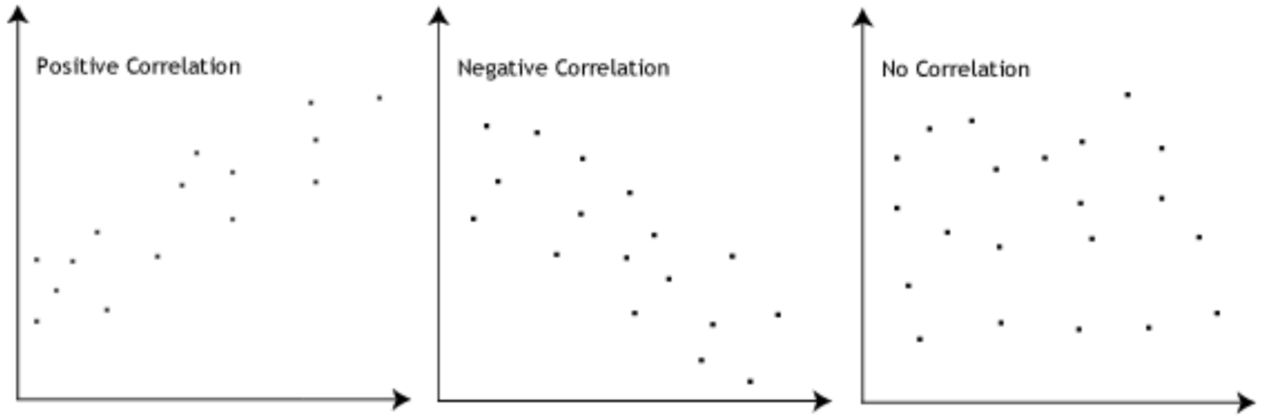


Formula

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

Here,

- 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



Q4. What is scaling? Why is scaling performed? What is the difference between normalised scaling and standardised scaling?

Scaling is a technique of bringing down the values of all the independent features of our dataset on the **same scale**. Feature selection helps to do calculations in algorithms very quickly. It is the important stage of data preprocessing.

If we didn't do feature scaling then the machine learning model gives higher weightage to higher values and lower weightage to lower values. Also, takes a lot of time for training the machine learning model.

Why Scaling

When you have a lot of independent variables in a model, a lot of them might be on very different scales which will lead a model with very weird coefficients that might be difficult to interpret. So we need to scale features because of two reasons:

1. Ease of interpretation
2. Faster convergence for gradient descent methods

If a feature in the dataset is **big in scale**(Salary) compared to **other(Age)** then in model it gives higher weightage to higher values and lower weightage to lower values. this big scaled feature becomes **dominating** and **needs to be normalised/standardised**.

Difference between normalised scaling and standardised scaling.

Standardising: The variables are scaled in such a way that their mean is zero and standard deviation is one.

$$x = \frac{x - mean(x)}{sd(x)}$$

Use-case of Standardiser

- In most of the Machine Learning models and it outperform MinMaxScaler(Normalization).

- Anywhere, where there is no need to scale features in the range 0 to 1.

- Since, it transforms the normal data distribution to standard normal distribution, which is the ideal & expected to have, most of the time it is the best to use in machine learning models

Normalisation : The variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data .

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Use-case of Normalisation:

- Every situation where the range of features should be between 0 to 1. For example, in Images data, there we have color pixels range from 0 to 255(256 colors in total), here Normaliser is the best one to use.

- There can be multiple scenarios where this range is expected, there it is optimal to use MinMaxScaler.

Q5 : You might have observed that sometimes the value of VIF is infinite. Why does this happen?

If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

Q6: What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions

Advantages:

a) It can be used with sample sizes also.

b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

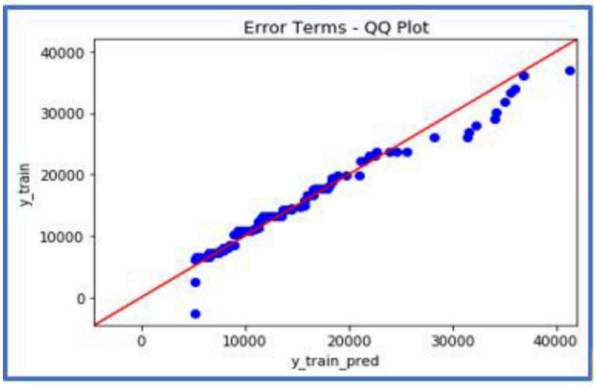
It is used to check following scenarios:

If two data sets —

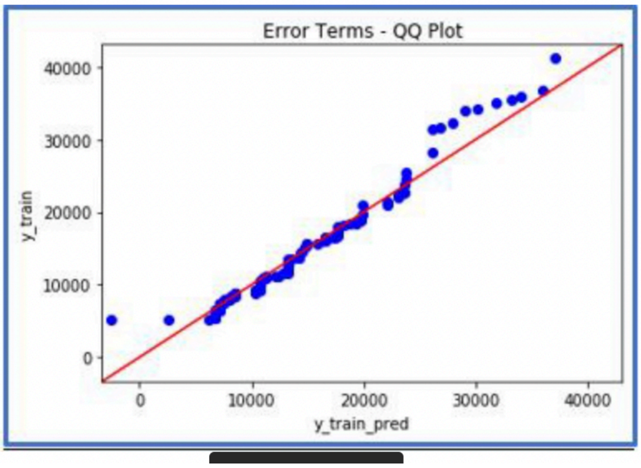
i. come from populations with a common distribution
ii. have common location and scale
iii. have similar distributional shapes
iv. have similar tail behavior

Interpretation

Below are the possible interpretations for two data sets.
a) **Similar distribution**: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis
b) **Y-values < X-values:** If y-quantiles are lower than the x-quantiles.



c) X-values < Y-values: If x-quantiles are lower than the y-quantiles.



Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis