Module :Machine Learning -1

Linear Regression Assignment

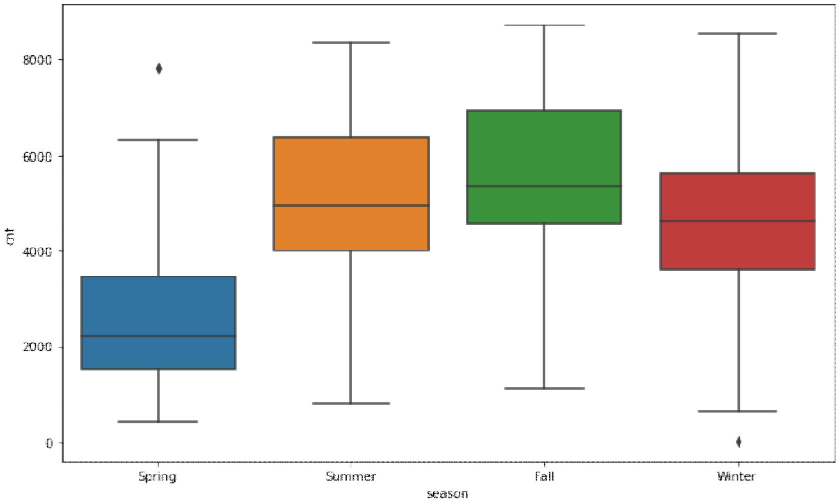
1. Assignment-Based Subjective Questions
2. General Subjective Questions

Submitted By:

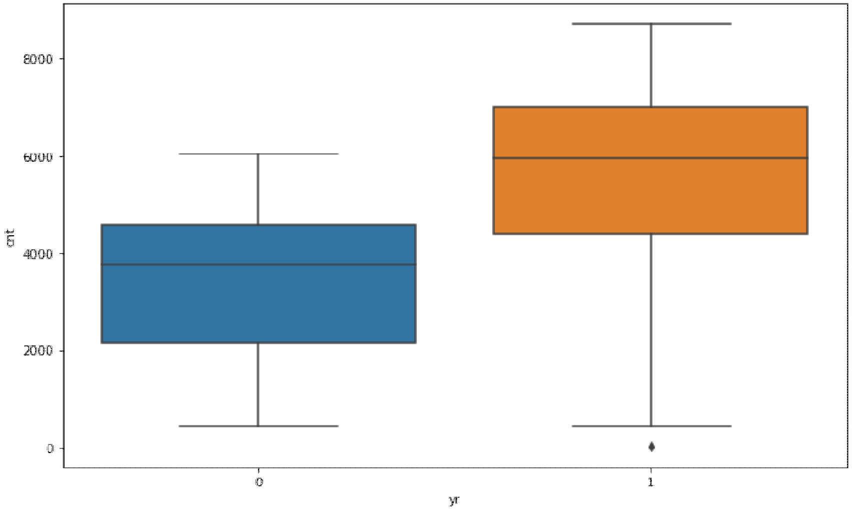
Md Nasiruzzaman

Assignment-Based Subjective Questions

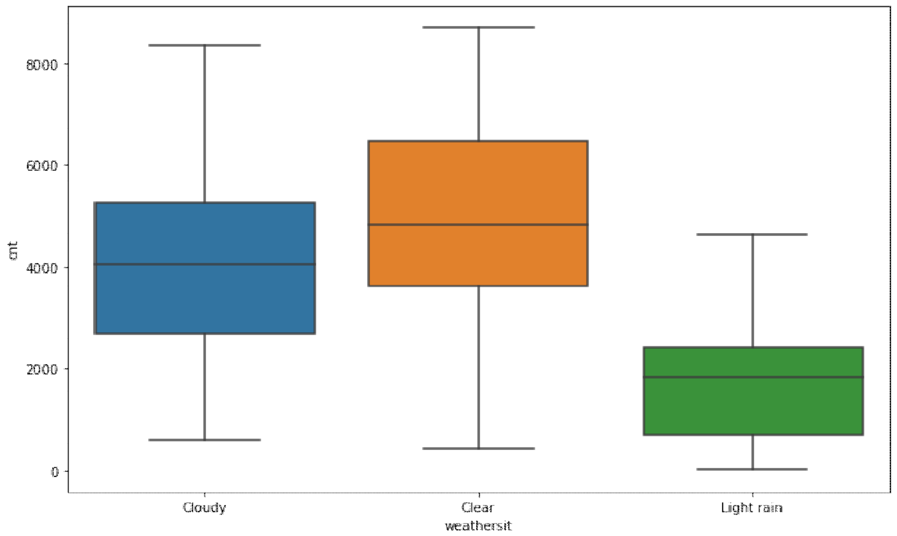
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?



It is observed from the season v/s count that the bike usage is high during fall and summer season and low during spring season.



Here 0 denotes 2018 and 1 denotes 2019 .
Box plot implies that the number of users are increasing over year.

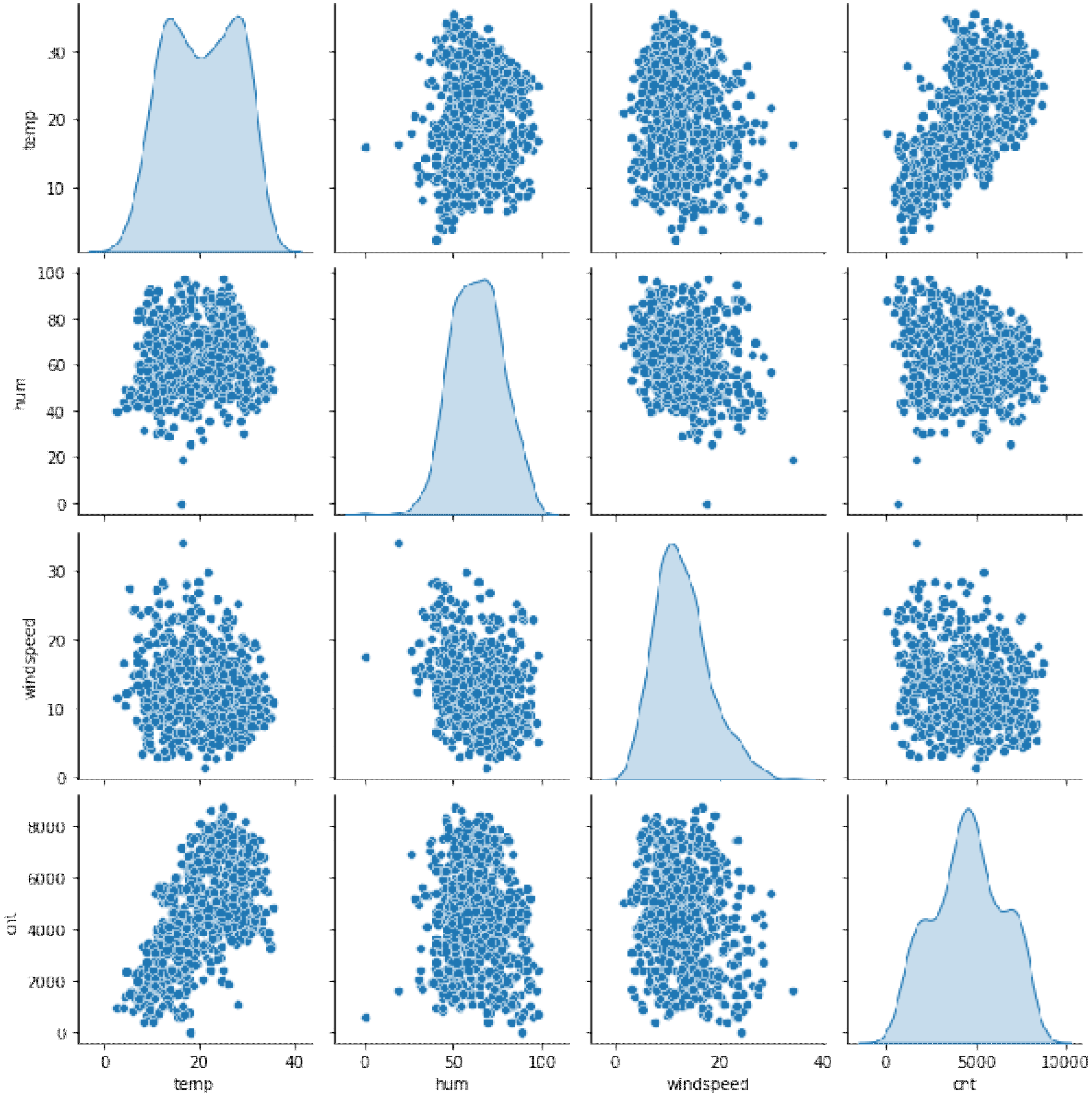


Bikes are generally used when weather is sky or bit cloudy. Hence weather plays a vital role over the use of the bikes for rental

2. Why is it important to use drop_first=True during dummy variable creation?

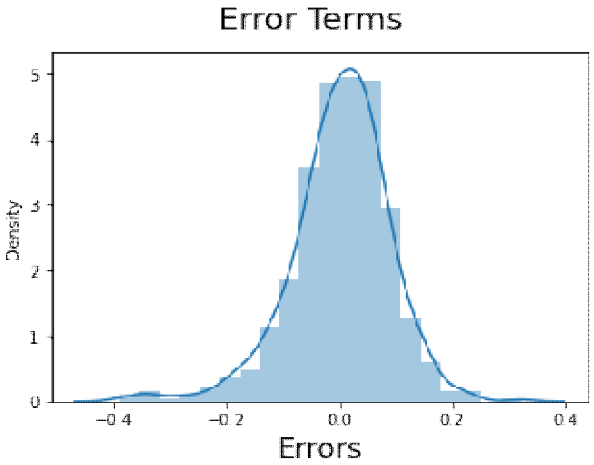
During dummy variable creation drop_first=True is important as it helps in reducing the extra column created .
Also if we have k categorical levels in categorical variable we can have (k-1) dummy column to represent the dummy variable.
It reduce the collinearity between dummy variable.

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



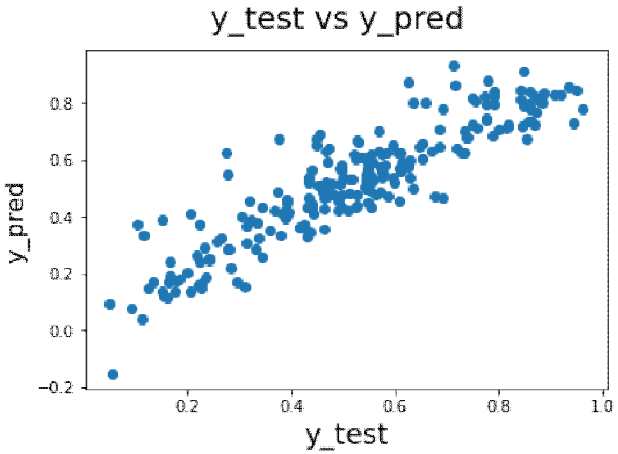
From the pair plot temperature holds highest correlation with target variable count.

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



Assumption : Residuals must be normally distributed.

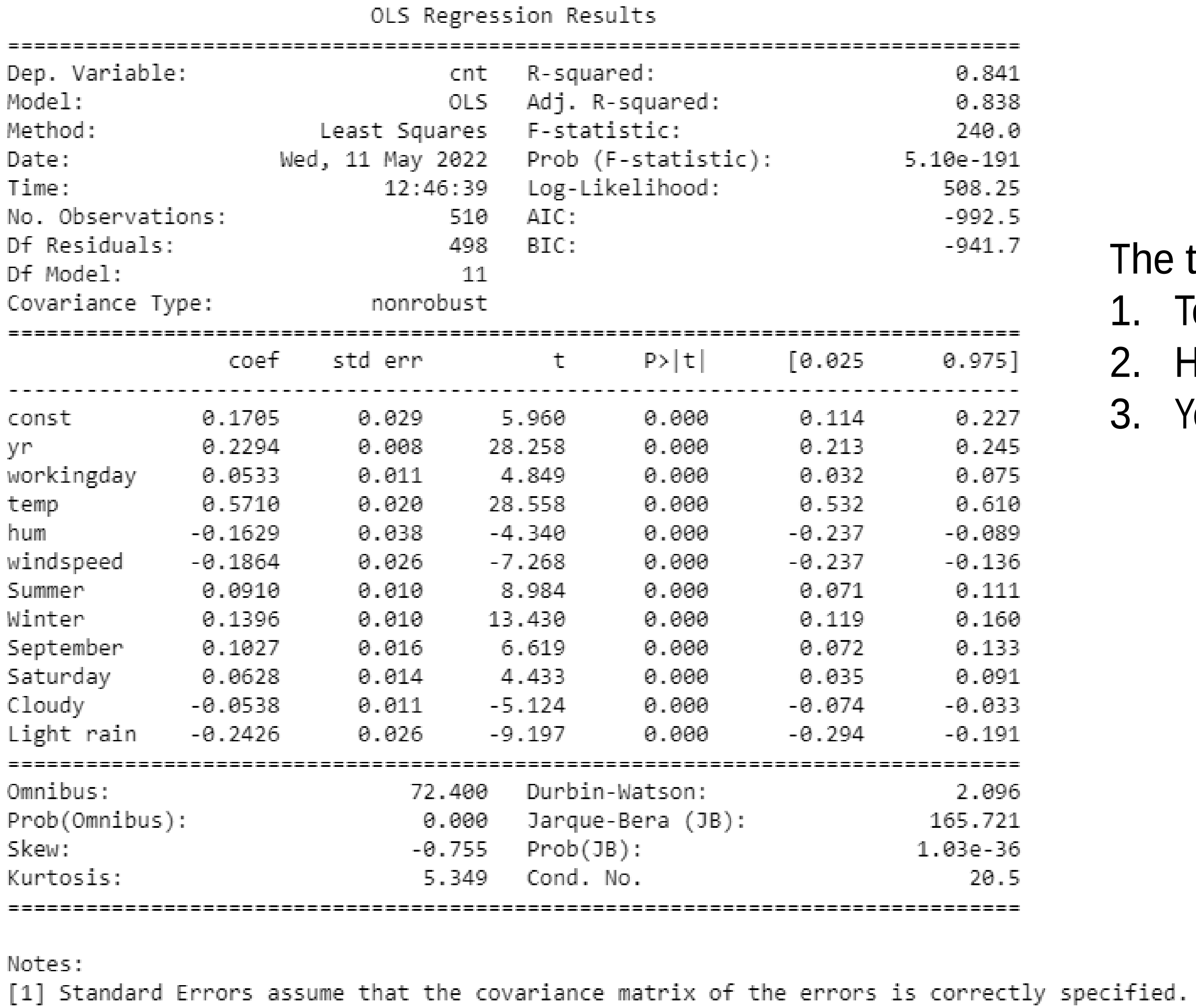
We can find the residuals to be normally distributed.



Assumption : The Dependent variable and Independent variable must have a linear relationship.

We checked the y_prediction v/s y_test and its shows linear relation as shown in the figure

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



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    cnt   R-squared:                       0.841
Model:                            OLS   Adj. R-squared:                  0.838
Method:                 Least Squares   F-statistic:                     240.0
Date:                Wed, 11 May 2022   Prob (F-statistic):          5.10e-191
Time:                        12:46:39   Log-Likelihood:                 508.25
No. Observations:                 510   AIC:                            -992.5
Df Residuals:                     498   BIC:                            -941.7
Df Model:                          11
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.1705      0.029      5.960      0.000       0.114       0.227
yr             0.2294      0.008     28.258      0.000       0.213       0.245
workingday     0.0533      0.011      4.849      0.000       0.032       0.075
temp           0.5710      0.020     28.558      0.000       0.532       0.610
hum           -0.1629      0.038     -4.340      0.000      -0.237      -0.089
windspeed     -0.1864      0.026     -7.268      0.000      -0.237      -0.136
Summer         0.0910      0.010      8.984      0.000       0.071       0.111
Winter         0.1396      0.010     13.430      0.000       0.119       0.160
September      0.1027      0.016      6.619      0.000       0.072       0.133
Saturday       0.0628      0.014      4.433      0.000       0.035       0.091
Cloudy        -0.0538      0.011     -5.124      0.000      -0.074      -0.033
Light rain    -0.2426      0.026     -9.197      0.000      -0.294      -0.191
==============================================================================
Omnibus:                       72.400   Durbin-Watson:                   2.096
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              165.721
Skew:                          -0.755   Prob(JB):                     1.03e-36
Kurtosis:                       5.349   Cond. No.                         20.5
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

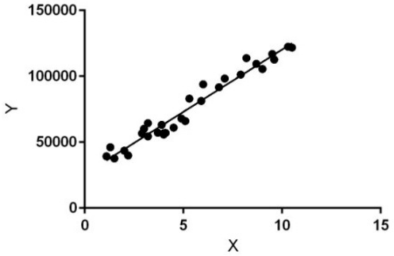
The top three features are:

1. Temperature
2. Humidity
3. Year

General Subjective Questions

1.Explain the linear regression algorithm in detail.

- Regression estimates the relationship among variables for prediction.
- Regression analysis helps to understand how the dependent variable changes when some of the independent variables are varied, while the other independent variables are held fixed.
- Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.
- This form of analysis estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable. Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values. There are simple linear regression calculators that use a "least squares" method to discover the best-fit line for a set of paired data. You then estimate the value of X (dependent variable) from Y (independent variable).
- While training the model we are given :
  x: input training data (univariate – one input variable(parameter))
  y: labels to data (supervised learning)
- When training the model – it fits the best line to predict the value of y for a given value of x. The model gets the best regression fit line by finding the best m and c values.



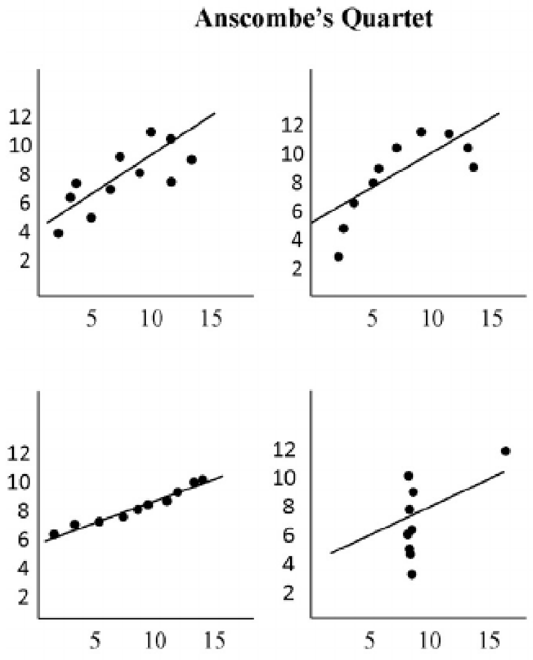
Y= mx + c

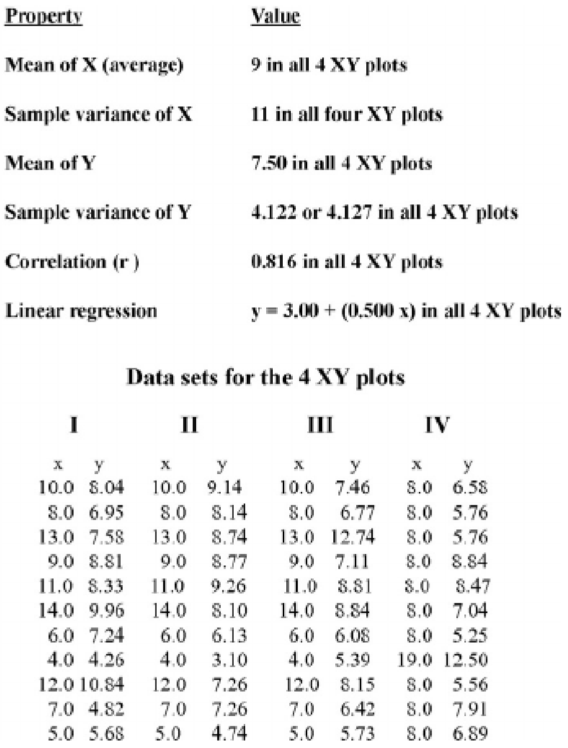
where m= coefficient of x

c= intercept

2. Explain the Anscombe’s quartet in detail.

- Anscombe’s Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.
- It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties.There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x,y points in all four datasets.





Property	Value
Mean of X (average)	9 in all 4 XY plots
Sample variance of X	11 in all four XY plots
Mean of Y	7.50 in all 4 XY plots
Sample variance of Y	4.122 or 4.127 in all 4 XY plots
Correlation (r)	0.816 in all 4 XY plots
Linear regression	y = 3.00 + (0.500 x) in all 4 XY plots

Data sets for the 4 XY plots

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	5.76
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	8.81	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	7.26	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

2. What is Pearson’s R?

- In Statistics, the Pearson's Correlation Coefficient is also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation. It is a statistic that measures the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0.
- Whenever we discuss correlation in statistics, it is generally Pearson's correlation coefficient. However, it cannot capture nonlinear relationships between two variables and cannot differentiate between dependent and independent variables.
- Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations.
- The formula given as:

$$r = \frac{N\Sigma xy-(\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2- (\Sigma x)^2][N\Sigma y^2 - (\Sigma y)^2]}}$$

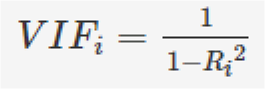
Where,
N = the number of pairs of scores
Σxy = the sum of the products of paired scores
Σx = the sum of x scores
Σy = the sum of y scores
Σx2 = the sum of squared x scores
Σy2 = the sum of squared y scores

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

- Scaling is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range and also helps in speeding up the calculations in an algorithm.
- Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modeling.
- **Normalization or Min-Max Scaling** is used to transform features to be on a similar scale. The new point is calculated as:
- X_new = (X - X_min)/(X_max - X_min)
- This scales the range to [0, 1] or sometimes [-1, 1].
- Normalization is useful when there are no outliers.
- **Standardization or Z-Score Normalization** is the transformation of features by subtracting from mean and dividing by standard deviation. This is often called as Z-score.
- X_new = (X - mean)/Std
- Standardization can be helpful in cases where the data follows a Gaussian distribution.

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

- The variance inflation factor (VIF) quantifies the extent of correlation between one predictor and the other predictors in a model. It is used for diagnosing collinearity/multicollinearity. Higher values signify that it is difficult to impossible to assess accurately the contribution of predictors to a model.
- The value of VIF is calculated by the below formula:

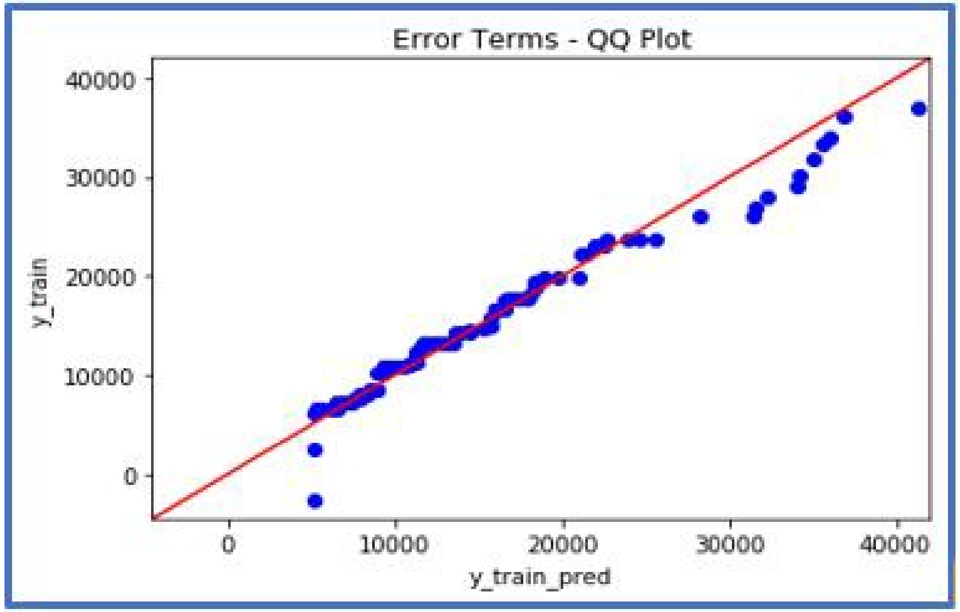


$$VIF_i = \frac{1}{1-R_i{}^2}$$

- Where, ‘i’ refers to the ith variable.
- If R-squared value is equal to 1 then the denominator of the above formula become 0 and the overall value become infinite. It denotes perfect correlation in variables.

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

- QQ plot can also be used to determine whether or not two distribution are similar or not. If they are quite similar you can expect the QQ plot to be more linear.
- A quantile-quantile plot, or Q-Q plot, is a plot of the sorted quantiles of one data set against the sorted quantiles of another data set. It is used to visually inspect the similarity between the underlying distributions of 2 data sets. Each point (x, y) is a plot of a quantile of one distribution along the vertical axis (y-axis) against the corresponding quantile of the other distribution along the horizontal axis (x-axis). If the 2 distributions are similar, then the points would lie close to the identity line, y = x.


Error Terms - QQ Plot
y_train
y_train_pred