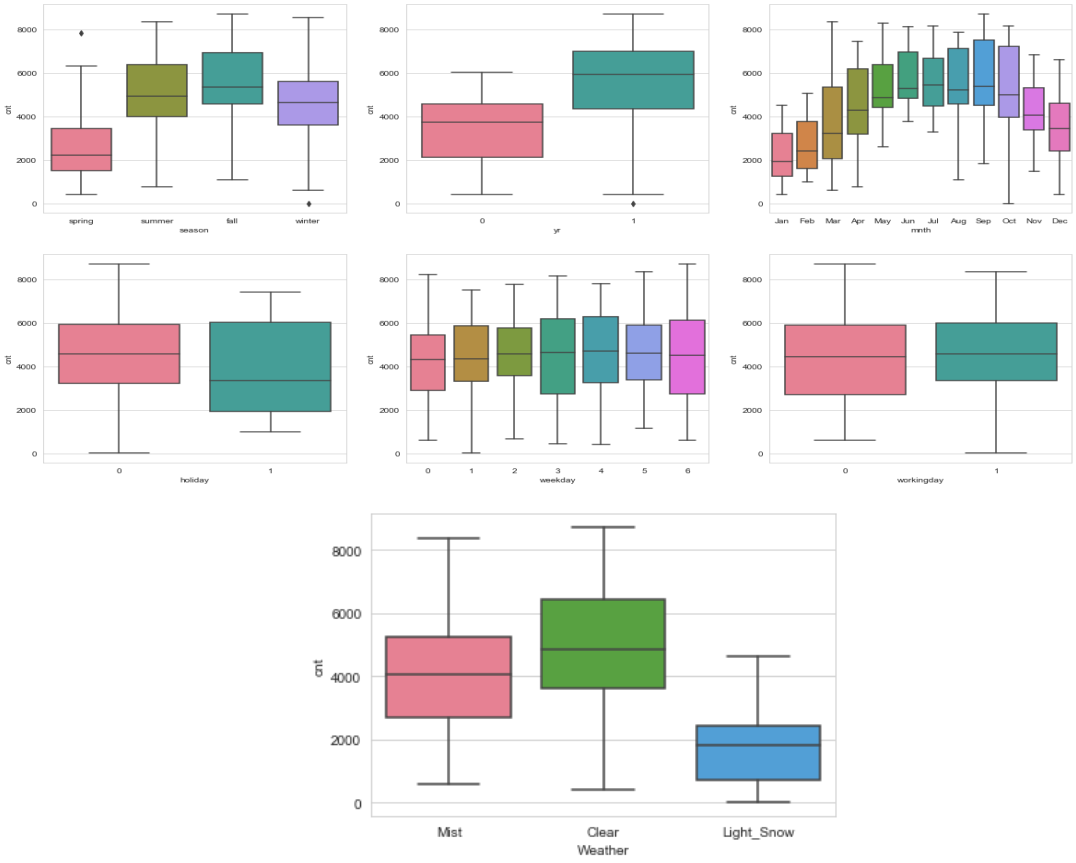
Submitted By- Koustav Chakraborty

Assignment-based Subjective Questions:-

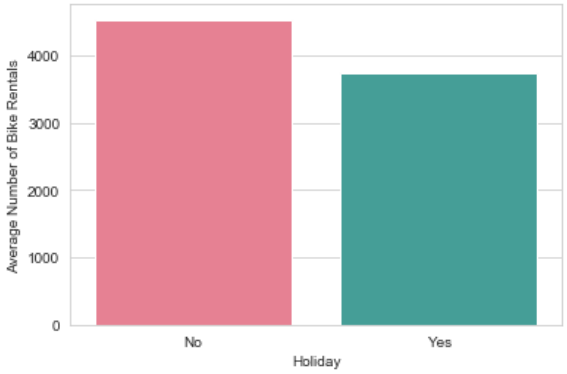
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

→

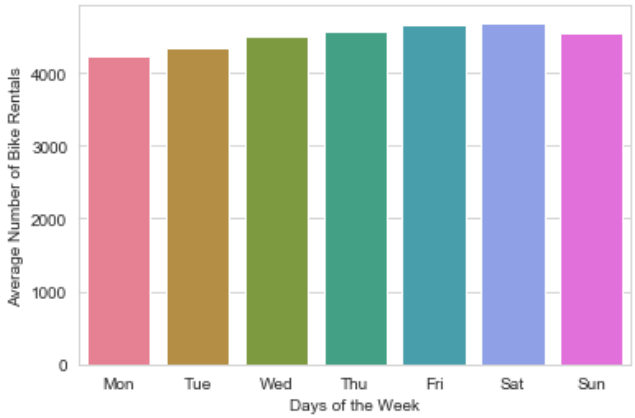


We can see that in summer months, the registration count goes up.

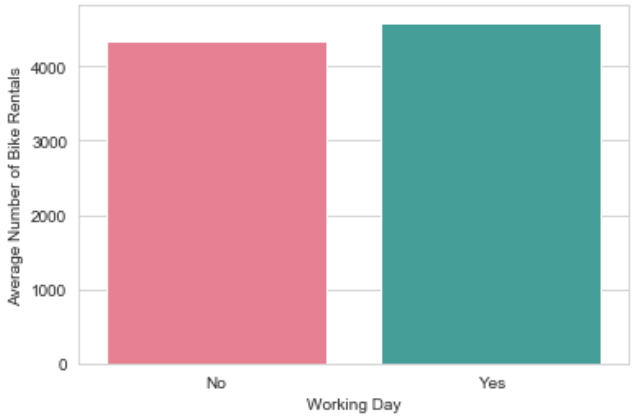
We can also see in clear weather the count goes up.



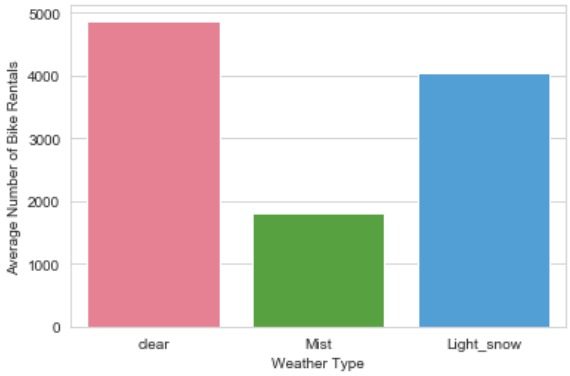
Average number is bike rentals is more on non-holidays



The average number of bike rentals is almost same throughout, but seems to be on the higher side as the weekend approaches.



The average number of Bike rentals is higher in working days.



We can also see in clear weather the count goes up.

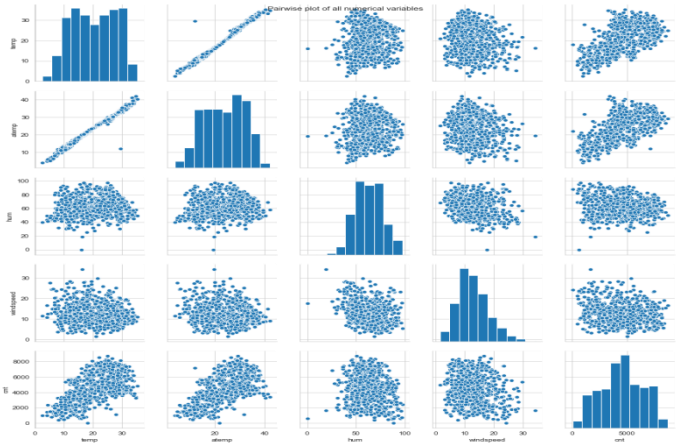
2. Why is it important to use drop_first=True during dummy variable creation?

➔ drop_first=True is important to use, because it helps to reduce the extra column created during dummy variable creation. As a result, it reduces the correlations created among dummy variables.

EX- Let's say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not furnished and semi_furnished, then It is obvious unfurnished. So we do not need 3rd variable to identify the unfurnished. Example

Hence if we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables.

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

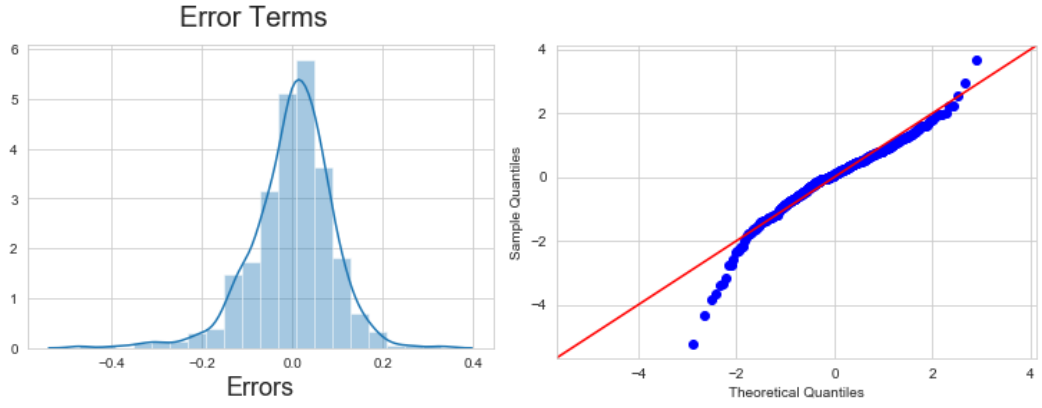


Temp and atemp has highest correlation here.

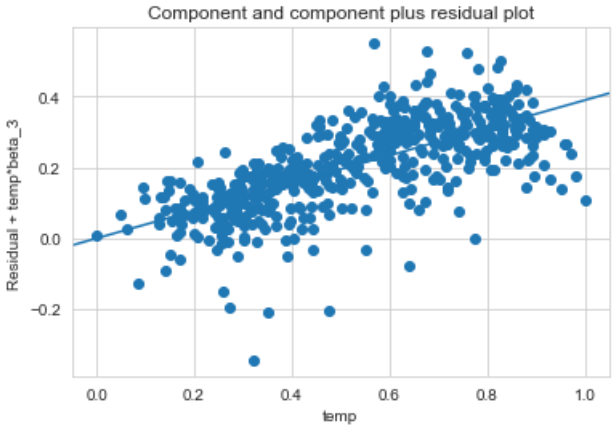
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

➔ The following assumptions are checked in the model-

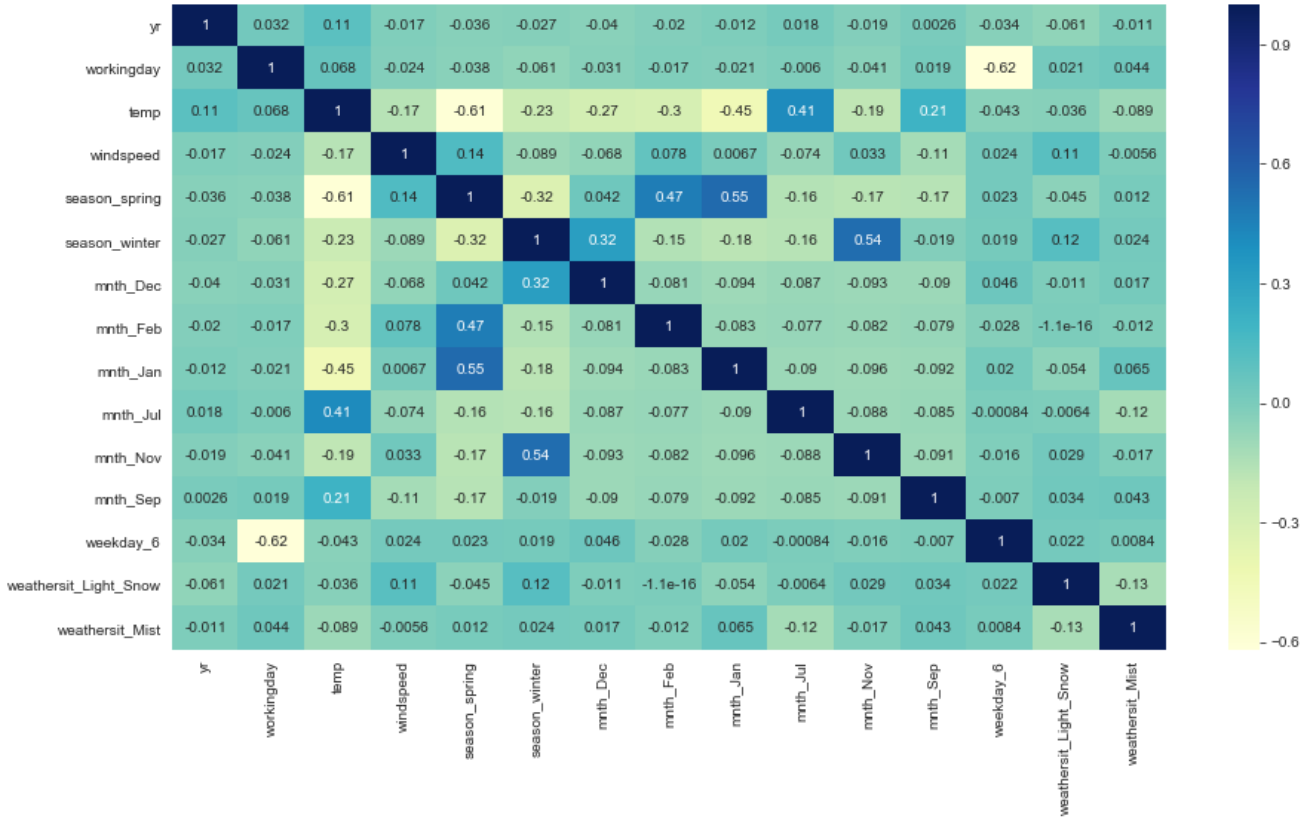
Residual Analysis of the train data- To check the error terms are normally distributed.



Linear Relationship- The partial residual plot represents the relationship between the predictor and the dependent variable while taking into account all the other variables. As we can see in the above graph, the linearity is well respected.



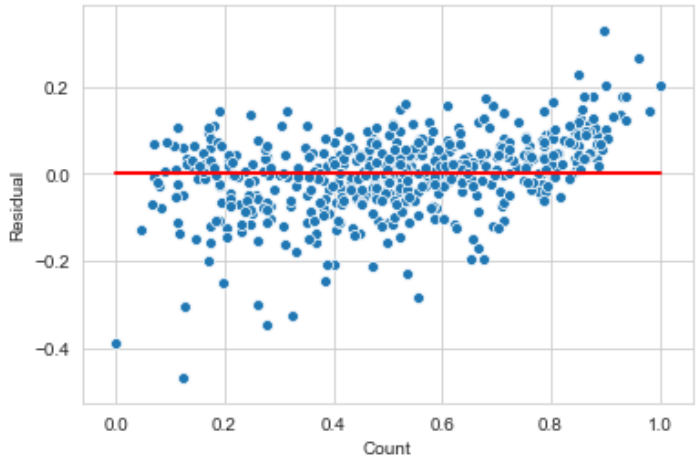
Absence of Multicollinearity-All variables have less than or equal to 0.55 correlation with each other.



	Features	VIF
2	temp	6.96
1	workingday	4.58
3	windspeed	4.49
4	season_spring	3.82
5	season_winter	2.61
8	mnth_Jan	2.22
0	yr	2.07
7	mnth_Feb	1.87
10	mnth_Nov	1.82
12	weekday_6	1.81
6	mnth_Dec	1.58
14	weathersit_Mist	1.57
9	mnth_Jul	1.37
11	mnth_Sep	1.21
13	weathersit_Light_Snow	1.10

Taking 5 as the maximum VIF permissible for this model but as temp has a good correlation and based on the business understanding I decided to keep it.

Homoscedasticity-As we can see from the plot above that the homoscedasticity is well and truely respected since the variance of the residuals are almost constant.



Independence of residuals (absence of auto-correlation)- Autocorrelation refers to the fact that observations' errors are correlated.
To verify that the observations are not auto-correlated, we can use the Durbin-Watson test.
The test will output values between 0 and 4. The closer it is to 2, the less auto-correlation there is between the various variables.
(0–2: positive auto-correlation, 2–4: negative auto-correlation)

The Durbin-Watson value for Model No.14 is 2.0027.

After checking all the assumption this model looks good.

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

➔ Workingday, temp and whethersit_Light_Snow are the three features that is significantly explaining the demand of the shared bikes.

General Subjective Questions:-

1. Explain the linear regression algorithm in detail.

➔ Linear regression algorithm is basically a machine learning algorithm of supervised learning. It is a technique by which we predict model that helps to find out the relationship between Input variable and the target variable.

Mainly this is used for three types of applications-

i) To find the effect of input variables on target variable.
ii) To find the change in target variable with respect to one or more input variable.
iii) To find the upcoming trends.

The types of Regression analysis are as follows-

Linear Regression

Multiple Linear Regression.

Logistic Regression

Polynomial Regression

Ex- Police department is running a campaign to reduce the number of robberies, so in this case we can expect the graph will be linearly downward.

Linear regression is used to predict a dependent variable Y from the predictor variable X.

Mathematically, we can write a linear regression equation as:

$$y = a + bx$$

Where a and b given by the formulas:

$$b\,(slope) = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

$$a\,(intercept) = \frac{n\sum y - b(\sum x)}{n}$$

Here, x and y are two variables on the regression line.

b = Slope of the line.

a = y-intercept of the line.

x = Independent variable

y = Dependent variable

The steps conducted to calculate regression analysis are as follows-

1) Data Preparation where we create dummy variables.
2) Splitting the data into training and test set where scaling is done.
3) Building the linear model.
4) Checking the assumptions of the linear model.
5) Making prediction using the final model.
6) Model evaluation.

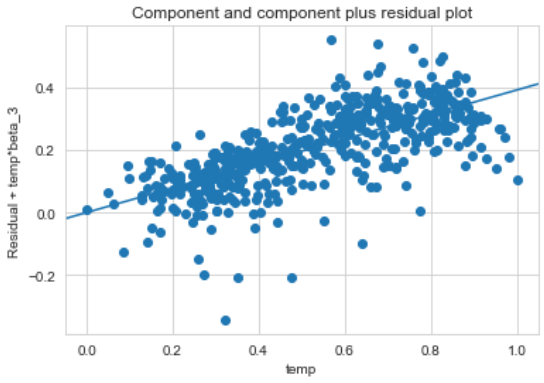
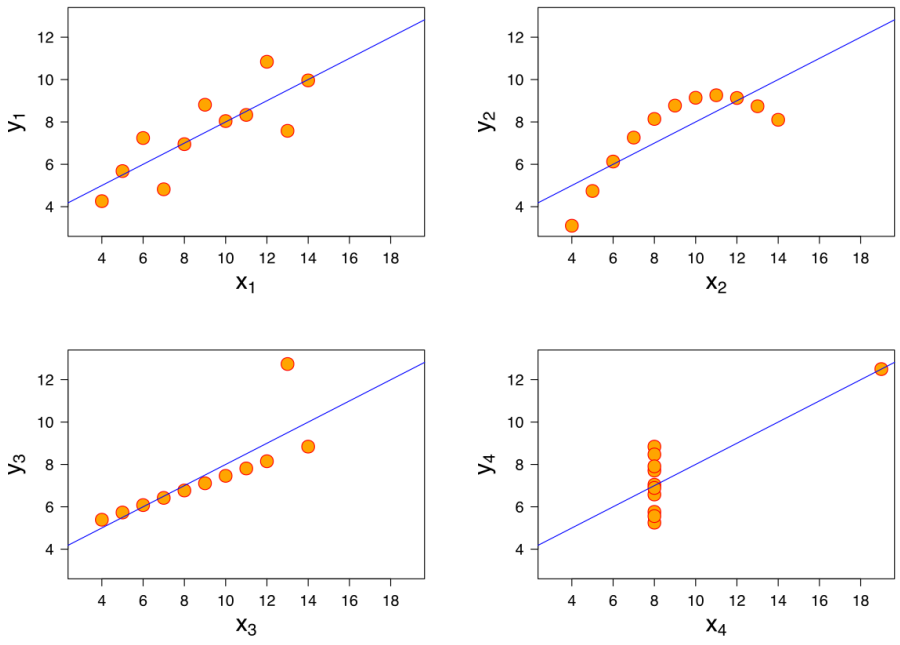


Image of regression line

2. Explain the Anscombe's quartet in detail.

➔ Anscombe's Quartet is basically a group of four data sets that are almost 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 generally they appear very differently when plotted on scatter plots.



So basically we can say that the first and 2nd order summary statistics don't say everything that we want to know about the data set, so usually it is plotted.

The bottom two panels show that these summary statistics are sensitive to outliers. If we generalize the bottom right panel that arises often in real life: then we might have to say that two noisy clouds corresponding to two groups.

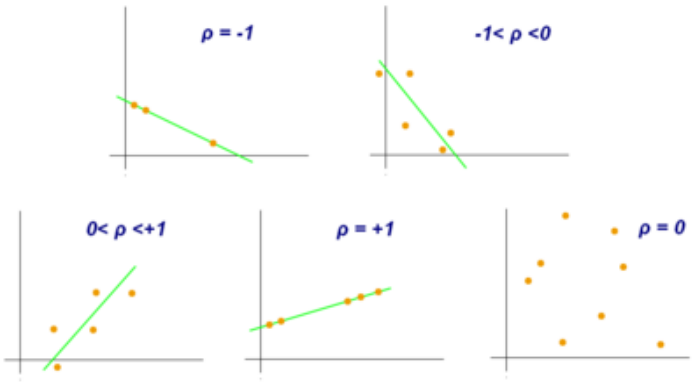
The top right panel shows that, even though we think about the correlation measures a linear association between two variables, we can have high correlations even when the relationship is nonlinear.

Application-

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

3. What is Pearson's R?

➔ The correlation between two variables that shows the degree to which the variables are related to each other. When correlation is computed in a sample, it is designated by the letter "r" and is known as "Pearson's r." Pearson's correlation shows the degree of linear relationship between two variables. It ranges from +1 to -1. So a correlation of +1 means that there is a perfect positive linear relationship between the two variables.



The scatterplot shown above states such a relationship. It is a positive relationship as we can see due to high scores on the X-axis is associated with high scores on the Y-axis.

So a correlation of -1 means that there is a total negative linear relationship between the two variables and r value of 0 indicates that there is no linear relationship between two variables.

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

➔ Feature Scaling is a technique to standardize the independent features present in the data in a fixed range to help in the modelling. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. This is performed after the data is spitted in train and test set.

Scaling only affects the coefficients and not any other parameters like F-Statistics, R squared, p-values.

Normalized Scaling- Also known as min-max scaling brings all the data in the range 0 and 1. **sklearn.preprocessing.MinMaxScaler** helps to calculate in python.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

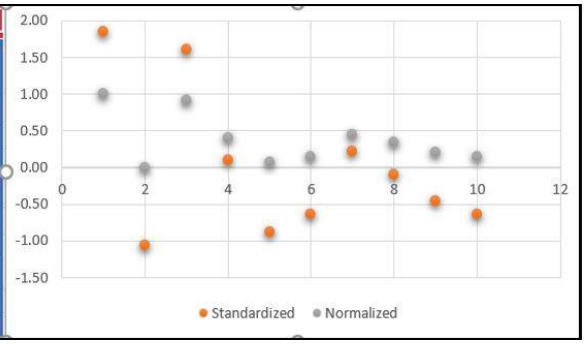
Standardized Scaling- It replaces the values by their z-score. It brings the data into standard normal distribution with mean=0 and standard deviation = 1.

sklearn.preprocessing.scale helps to calculate in python.

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

Disadvantage of normalization over standardization is it looses some information about data like outliers.

Original	Standardized	Normalized
59	1.85	1.000
11	-1.06	0.000
55	1.61	0.917
30	0.09	0.396
14	-0.88	0.063
18	-0.64	0.146
32	0.21	0.438
27	-0.09	0.333
21	-0.46	0.208
18	-0.64	0.146



Standardization and Normalization with original values .

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

➔ VIF is an index that basically shows the measure of how much the variance of an estimated regression coefficient increases due to collinearity. So to determine VIF, we need to fit a regression model between the independent variables.
Now, if there is a perfect correlation, then VIF = infinity. A large value of VIF shows that there is correlation between variables. If the independent variables are orthogonal then VIF is equal to 1. The standard error of the coefficient illustrates the confidence interval of the model coefficients. If the standard error is large, then the confidence intervals may also be large, and the model coefficient might come out to be non-significant due to the presence of multicollinearity.

Again a general rule of thumb is that if VIF>10 there is multicollinearity and also VIF<=5 is widely accepted to determine multicollinearity.

VIF	Conclusion
1	No multicollinearity
4 - 5	Moderate
10 or greater	Severe

.

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

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

This helps us 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.

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 during checking linear regression assumptions.

It is used to check following scenarios in the dataset-

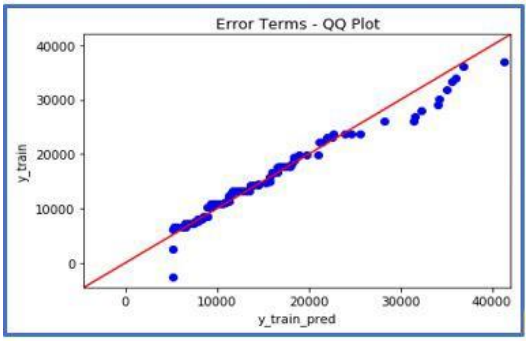
If two data sets —

Comes from population with a common distribution, have common location and scale, have similar distributional shapes, have similar tail behavior

Below are the possible interpretations for two data sets.

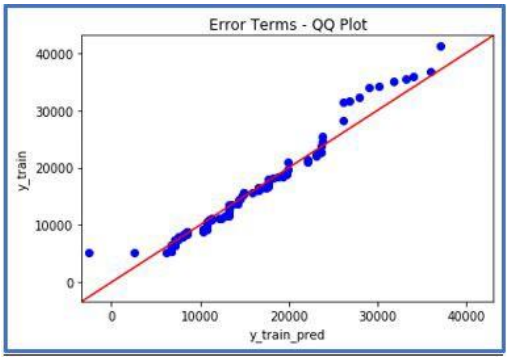
a) Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x –axis.

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.

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



This is the QQ-plot plotted in the assignment while checking the assumptions of linear regression model.