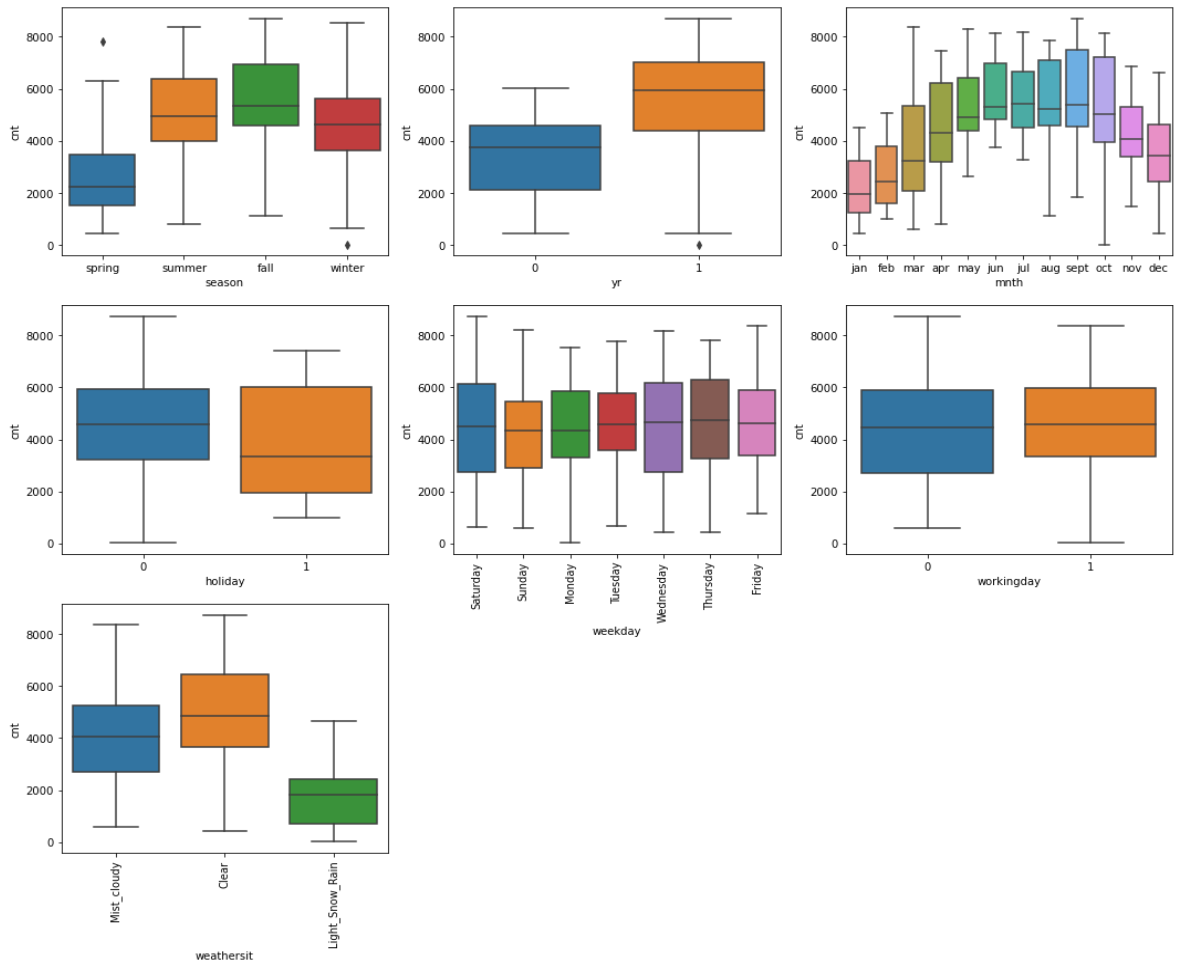
Linear Regression Assignment

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? (3 marks)

Sol:



From above boxplots for all the categorical variables, we can derive following:

1. There is less demand for bikes in spring and winter, high in summer and fall.
2. There is less demand in the year 2018 and high in 2019.
3. Less demand in jan, feb, nov and dec, and more in other months in a year.
4. More demand on holidays than other days.
5. Almost same demand on all working days, but on Saturday and Wednesday there will be more demand.
6. Non-working days has more demand than working days.
7. Less demand in Light_Snow_Rain weather situation, and high in Clear weather situation.

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

Sol:

Dummy variables are created to convert categorical values into numerical values.

- So, while we convert ‘n’ categories into dummy variables it creates ‘n’ variables.
- Instead of creating ‘n’ variables to represent ‘n’ categories, we can drop one variable after dummy creation. Because ‘n-1’ variables can represent ‘n’ categories, this is the reason why we drop one variable using drop_first = True.
- Below is an example:
    - Before doing drop_first = True.

yr_2018	yr_2019	
1	0	Indicates year = 2018
0	1	Indicates year = 2019

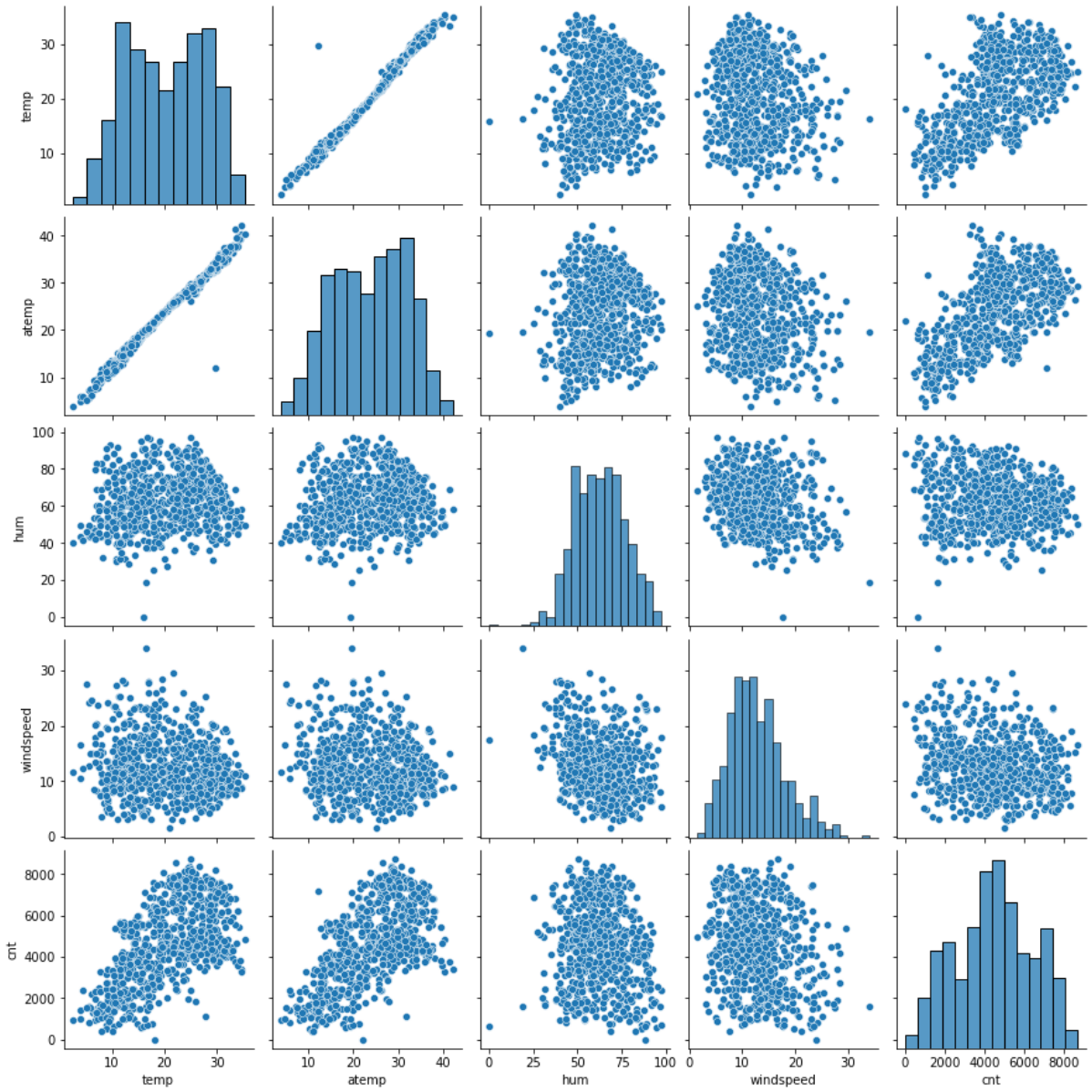
    - After doing drop_first = True.

yr_2019	
0	Indicates year = 2018
1	Indicates year = 2019

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Sol:

By looking at the below pairplot, we can say that temp/atemp has the highest correlation with the target variable.



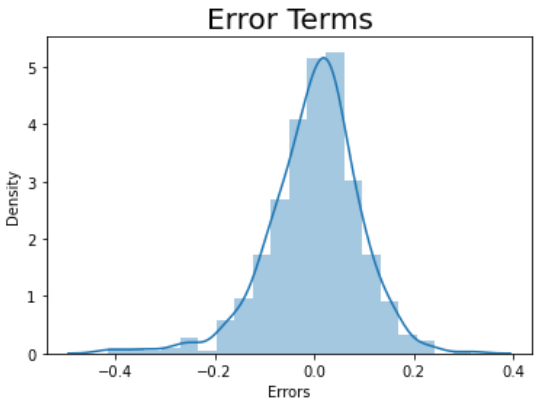
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Sol:

There are 5 assumptions for Linear Regression:

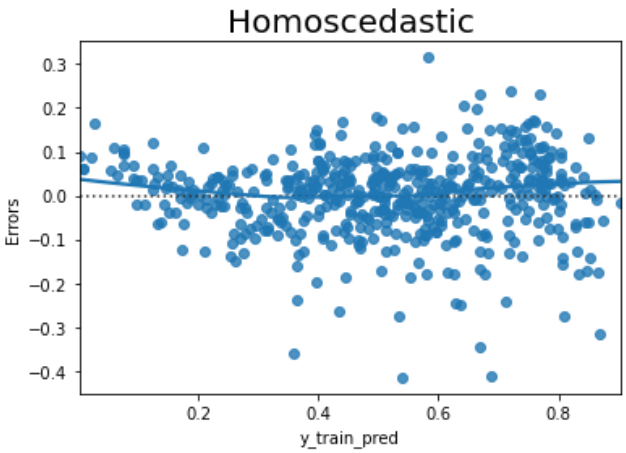
- Error terms with normal distribution
- Homoscedasticity
- Linear relationship
- Independence of variables using Autocorrelation
- Multicollinearity

Error terms with normal distribution:



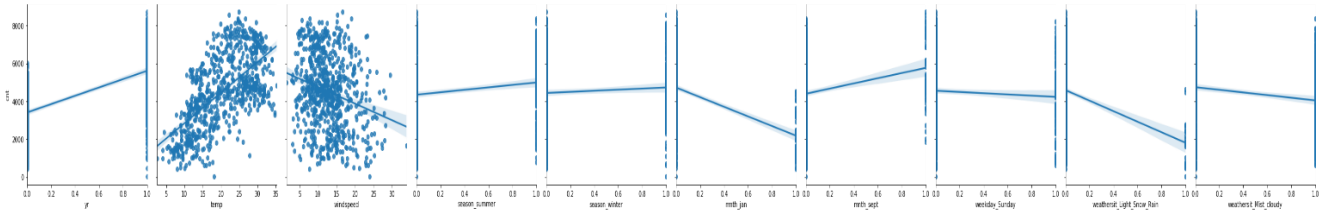
- We can observe the normal distribution of the data from above plot.

Homoscedasticity:



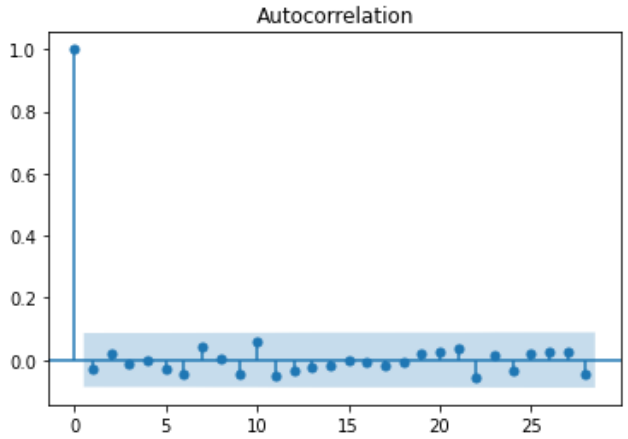
- We can observe any patterns in the data and how far all data-points from 0-axis.

Linear relationship:



- We can observe how linearly the data is distributed, it can be positive or negative.

Independence of variables using Autocorrelation:



- We can observe the variance in the datapoints from the 0-axis.

Multicollinearity:

- We can observe how many variables are closely correlated to each other.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Sol:

After final model building, we can draw the top 3 highly correlated variables with target variable as below:

1. temp (0.521582)

2. yr (0.234075)

3. season_winter (0.117836)

General Subjective Questions

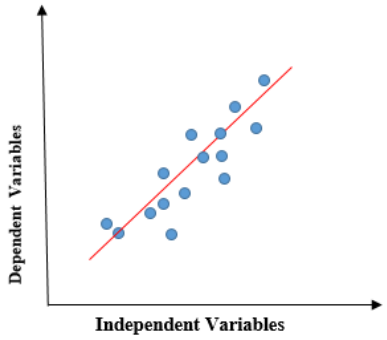
1. Explain the linear regression algorithm in detail. (4 marks)

Sol:

- Linear regression is a simple statistical regression method used for predictive analysis and shows the relationship between the continuous variables. Linear regression shows the linear relationship between the independent variables and the dependent variable, consequently called linear regression.

- If there is a single input variable, such linear regression is called **simple linear regression**.

- If there is more than one input variable, such linear regression is called **multiple linear regression**. The linear regression model gives a sloped straight line describing the relationship within the variables.



To calculate best-fit line linear regression uses a traditional slope-intercept form.

$$y = mx+b \quad \Longrightarrow \quad y = a_0+a_1x$$

y = Dependent Variable.
x = Independent Variable.
a0 = intercept of the line.
a1 = Linear regression coefficient.

As mentioned above, Linear regression estimates the relationship between a dependent variable and an independent variable.

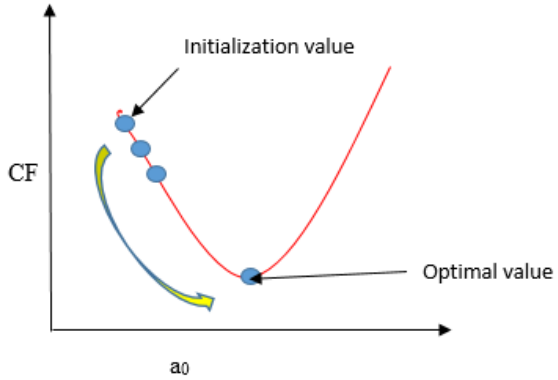
Cost function

- The cost function helps to figure out the best possible values for a0 and a1, which provides the best fit line for the data points.
- In Linear Regression, Mean Squared Error (MSE) cost function is used, which is the average of squared error that occurred between the predicted values and actual values.
- *Let's y = actual values, y_i = predicted values*

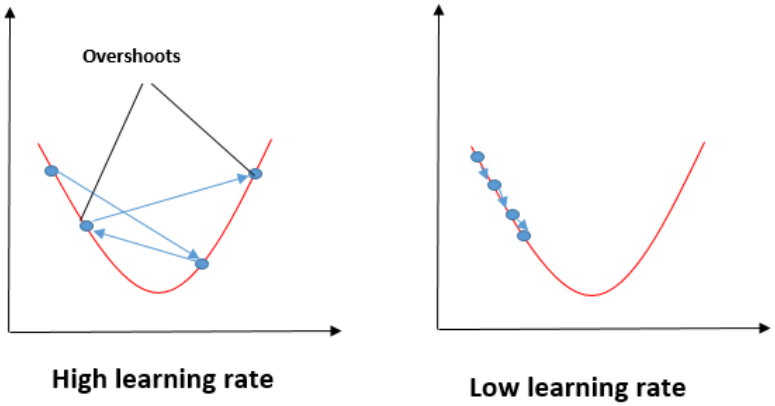
$$MSE = \frac{1}{N}\sum_{i=1}^{n}(y_i - (mx_i + b))^2$$

Gradient descent

- Gradient descent is a method of updating a0 and a1 to minimize 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.



- In the gradient descent algorithm, the number of steps you take is the learning rate, and this decides how fast the algorithm converges to the minima.



2. Explain the Anscombe's quartet in detail. (3 marks)

Sol:

- *Anscombe's Quartet* is the method to demonstrate the importance of data visualization which was developed by the statistician *Francis Anscombe* in 1973 to signify both the importance of plotting data before analyzing it with statistical properties.

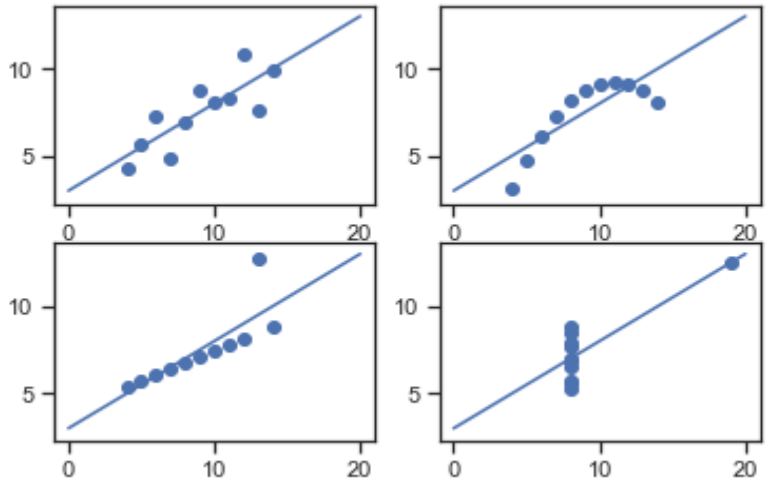
- It comprises of four data-set and each data-set consists of eleven (x,y) points.

- The basic thing to analyze about these data-sets is that they all share the same descriptive statistics(mean, variance, standard deviation etc) but different graphical representation.

- Each graph plot shows the different behavior irrespective of statistical analysis.

x1	y1	x2	y2	x3	y3	x4	y4
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

- However, the statistical analysis of these four data-sets are pretty much similar. But when we plot these four data-sets across the x & y coordinate plane, we get the following results & each pictorial view represent the different behavior.

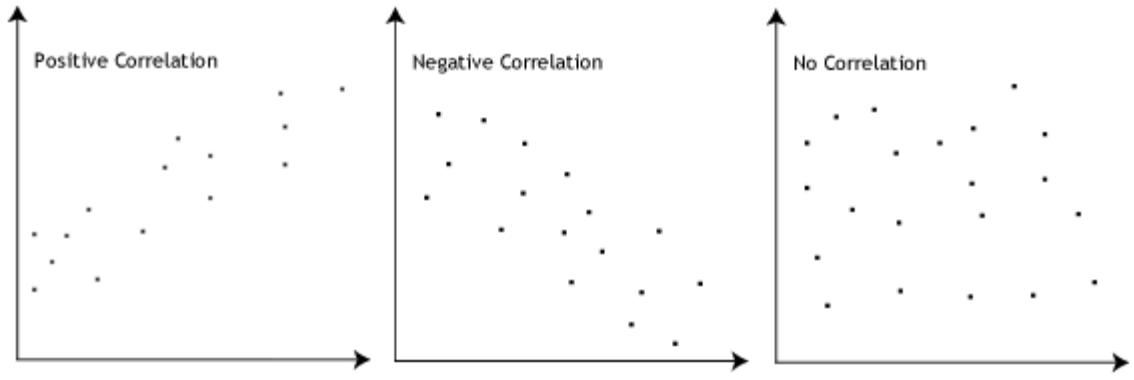


- Data-set I - consists of a set of (x,y) points that represent a linear relationship with some variance.
- Data-set II - shows a curve shape but doesn't show a linear relationship (might be quadratic?).
- Data-set III - looks like a tight linear relationship between *x* and *y*, except for one large outlier.
- Data-set IV - looks like the value of *x* remains constant, except for one outlier as well.

3. What is Pearson's R? (3 marks)

Sol:

- In Statistics, the Pearson's Correlation Coefficient is also referred to as Pearson's r. 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.



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

- r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
- r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
- r = 0 means there is no linear association
- r > 0 < 5 means there is a weak association
- r > 5 < 8 means there is a moderate association
- r > 8 means there is a strong association

Pearson r Formula

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

Where,

- 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

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Sol:

- **Feature Scaling** is a technique to standardize the independent features present in the data in a fixed range.

- **Reason**: It is performed during the data pre-processing to handle highly varying magnitudes or values or units.

- Machine learning algorithm just sees number — if there is a vast difference in the range say few ranging in thousands and few ranging in the tens, and it makes the underlying assumption that higher ranging numbers have superiority of some sort. So these more significant number starts playing a more decisive role while training the model.

Normalization	Standardization
Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
It is really affected by outliers.	It is much less affected by outliers.
Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.
This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.
It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
It is a often called as Scaling Normalization	It is a often called as Z-Score Normalization.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Sol:

- VIF is a measure of calculating correlation between 2 or more variables.

- VIF = infinity, indicates the 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.

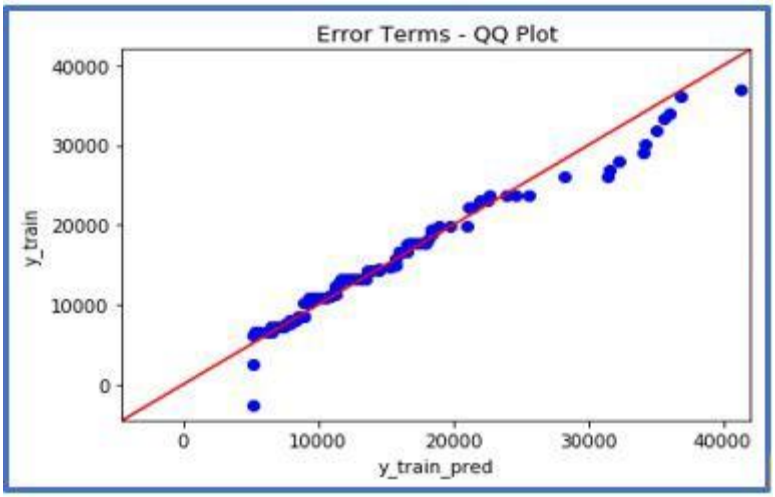
- VIF = 1, indicates there is no correlation between the variables.

- VIF >1 or very high, indicates a high correlation between the variables.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Sol:

- The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.



The advantages of the q-q plot are:

1. The sample sizes do not need to be equal.

2. Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot. For example, if the two data sets come from populations whose distributions differ only by a shift in location, the points should lie along a straight line that is displaced either up or down from the 45-degree reference line.