Linear Regression Subjective Questions

Pavani Gangula

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?

Answer: From the analysis of categorical variables from the data, we can infer that 'yr','workingday','weekday_Mon','mnth_Dec','mnth_July','mnth_Nov','mnth_Oct', and 'weathersit_Light Rain', 'casual', 'workingday', and 'windspeed' are the variables with VIF less than 5.

But 'weekday_Mon', and 'mnth_July' had high p-values. Hence they were discarded from the analysis.

At an overall level, 'mnth', 'weekday','yr', and 'weathersit' had been useful in determining the 'cnt' variable. Other categorical variables couldn't add value to the analysis.

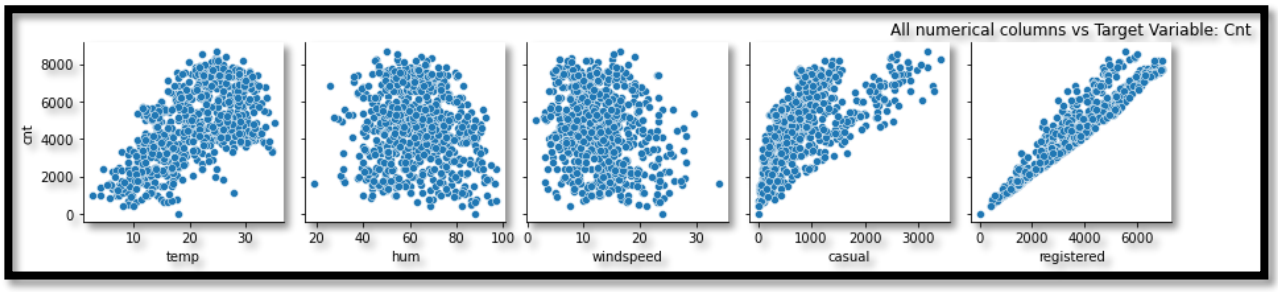
2. Why is it important to use drop_first=True during dummy variable creation?

Answer: To encode categorical data, one hot encoding is done, where a dummy variable is to be created for each discrete categorical variable for a feature. This can be done by using pandas.get_dummies() which will return dummy-coded data. Here we use parameter drop_first = True, this will drop the first dummy variable, thus it will give n-1 dummies out of n discrete categorical levels by removing the first level.
If we do not use drop_first = True, then n dummy variables will be created, and these predictors(n dummy variables) are themselves correlated which is known as multicollinearity and it, in turn, leads to Dummy Variable Trap.

Suppose, you have a column for gender that contains 4 variables- "Male", "Female", "Other", "Unknown". So a person is either "Male", or "Female", or "Other". If they are not either of these 3, their gender is "Unknown".

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

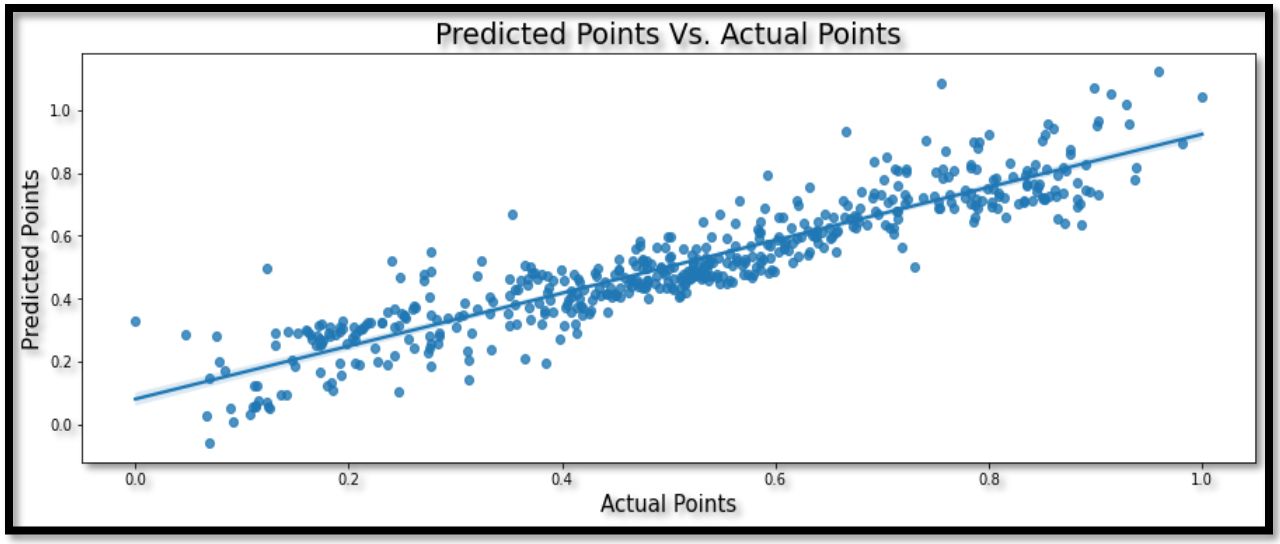
Answer: Out of all the numerical variables, 'registered' has the highest correlation with the target variable. Shown below are the pair plots and the correlation matrix.



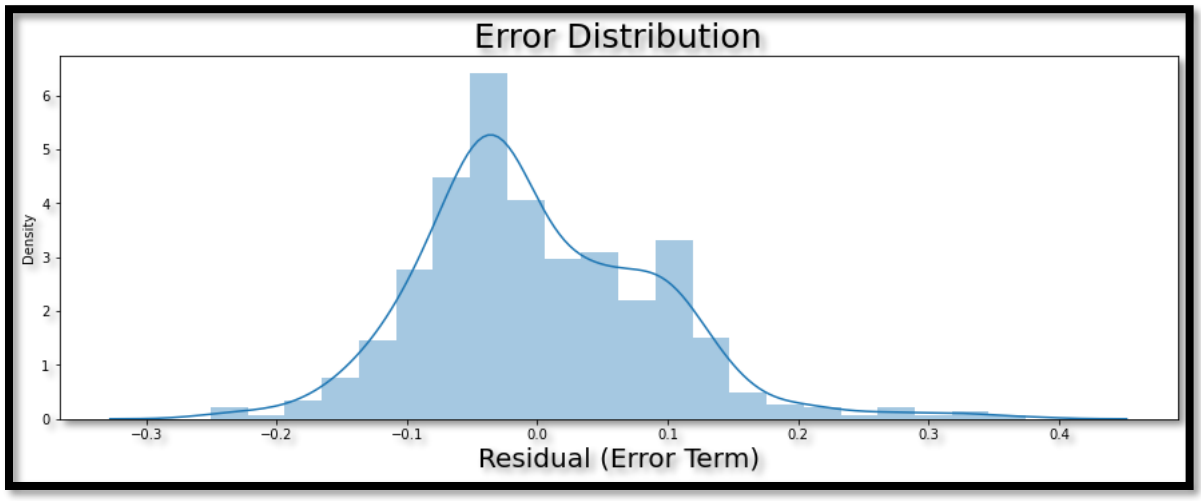
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Answer: Listed below are the assumptions of Linear Regression:

- Linear relationship between dependent and independent variables:
  Shown below is the distribution of predicted vs actual data points.



- <u>The residuals should follow a normal distribution:</u>
  Shown below is the distribution plot of residuals.

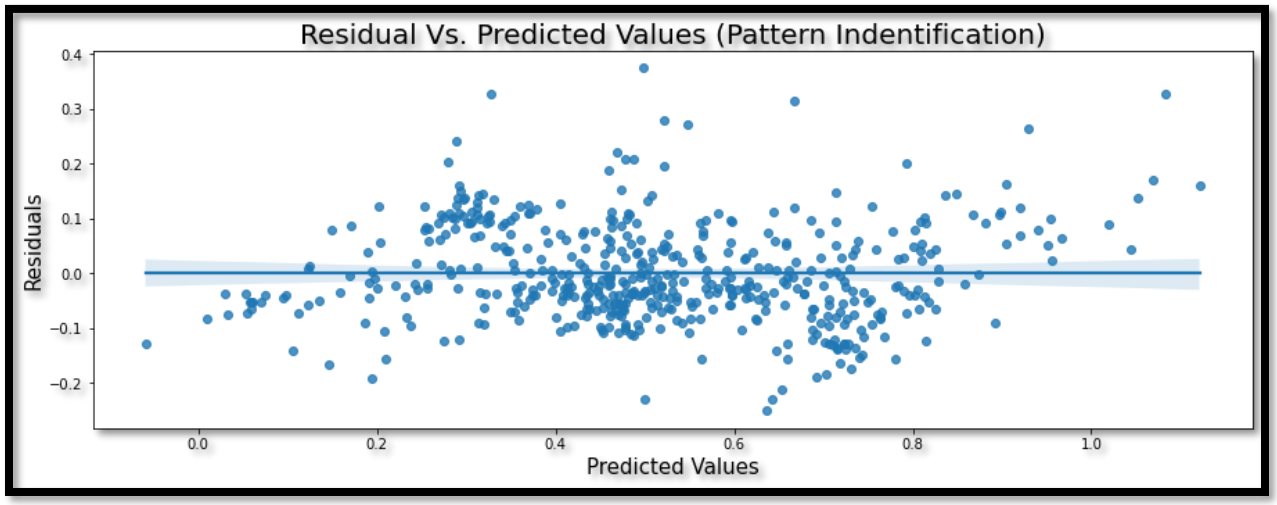


- <u>No Multicollinearity among predictors:</u>
  This can be verified by calculating VIF score. As a thumb rule, we consider predictors with VIF score <5 for the analysis. Shown below are the VIF scores for the champion model.

```
fetch_vif_df(X_train_rfe)
```

	Features	VIF
0	windspeed	2.91
1	yr	2.29
2	workingday	2.28
3	casual	2.22
4	mnth_Oct	1.14
5	mnth_Nov	1.09
6	weathersit_Light Rain	1.09
7	mnth_Dec	1.08

- Residuals should be independent of each other: Shown below is the distribution of residuals and the predicted values. We can observe that there is no pattern they follow and can be concluded that they are independent of each other.



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

Answer: Based on the champion model the top 3 features which contribute significantly in explaining the demand of shared bikes are 'workingday' with coefficient 0.2407, 'yr' with coefficient 0.1629 and 'weathersit_Light Rain' with coefficient -0.1189. 'workinhday' and 'yr' are positively correlated and 'weathersit_Light Rain' is negatively correlated.

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0571	0.017	3.363	0.001	0.024	0.090
yr	0.1629	0.008	19.187	0.000	0.146	0.180
workingday	0.2407	0.011	22.803	0.000	0.220	0.261
windspeed	-0.0901	0.025	-3.655	0.000	-0.139	-0.042
casual	0.0003	8.03e-06	34.024	0.000	0.000	0.000
mnth_Dec	0.0524	0.015	3.425	0.001	0.022	0.082
mnth_Nov	0.0693	0.015	4.728	0.000	0.041	0.098
mnth_Oct	0.0468	0.015	3.174	0.002	0.018	0.076
weathersit_Light Rain	-0.1189	0.024	-4.871	0.000	-0.167	-0.071

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer: Linear regression is a part of regression analysis. Regression analysis is a technique of predictive modeling that helps you to find out the relationship between Input and the target variable. In the case of linear regression as you can see the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.

Linear regression is used to predict a quantitative response 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:

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

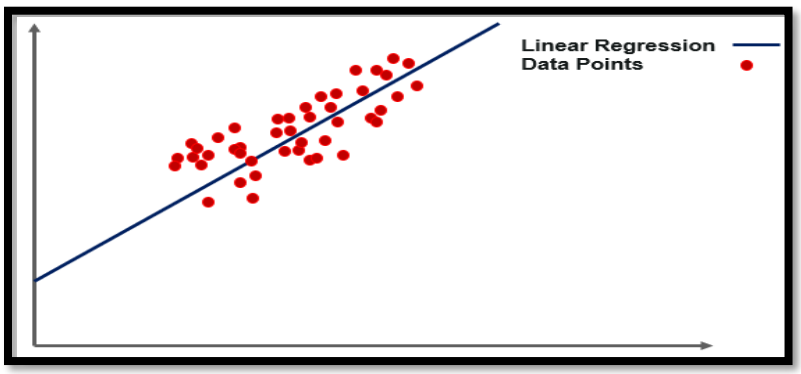
$$a\,(intercept) = \frac{n\sum y - b(\sum x)}{n}$$

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 from dataset; y = Dependent variable from dataset

Linear Regression algorithm finds a linear equation that best describes the correlation of the explanatory variables with the dependent variable. This is achieved by fitting a line to the data using least squares. The line tries to minimize the sum of the squares of the residuals. The residual is the distance between the line and the actual value of the explanatory variable. Finding the line of best fit is an iterative process.

Shown below is an example of data points and the best fit line of a Linear Regression Algorithm.



A linear regression model helps in predicting the value of a dependent variable, and it can also help explain how accurate the prediction is. This is denoted by the R-squared and p-value values. The R-squared value indicates how much of the variation in the dependent variable can be explained by the explanatory variable and the p-value explains how reliable that explanation is. The R-squared values range between 0 and 1. A value of 0.8 means that the explanatory variable can explain 80 percent of the variation in the observed values of the dependent variable. A value of 1 means that a perfect prediction can be made, which is rare in practice. A value of 0 means the explanatory variable doesn't help at all in predicting the dependent variable. Using a p-value, you can test whether the explanatory variable's effect on the dependent variable is significantly different from 0.

2. Explain the Anscombe's quartet in detail.

Answer: Anscombe's quartet tells us about the importance of visualizing data before applying various algorithms to build models. This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data

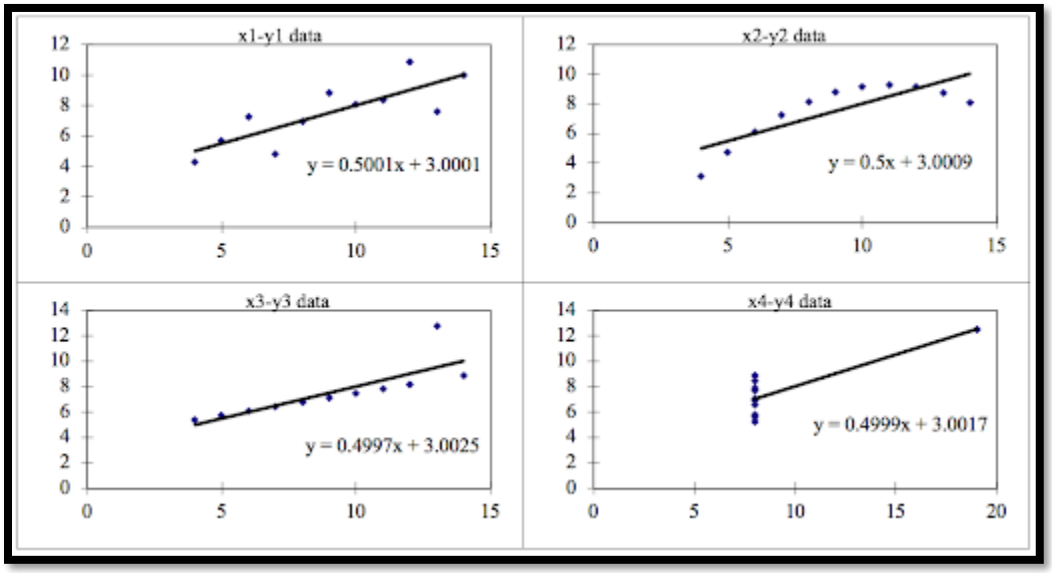
(outliers, diversity of the data, linear separability of the data, etc.). Moreover, the linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set. We can define these four plots as follows:

Anscombe's Data								
Observation	x1	y1	x2	y2	x3	y3	x4	y4
1	10	8.04	10	9.14	10	7.46	8	6.58
2	8	6.95	8	8.14	8	6.77	8	5.76
3	13	7.58	13	8.74	13	12.74	8	7.71
4	9	8.81	9	8.77	9	7.11	8	8.84
5	11	8.33	11	9.26	11	7.81	8	8.47
6	14	9.96	14	8.1	14	8.84	8	7.04
7	6	7.24	6	6.13	6	6.08	8	5.25
8	4	4.26	4	3.1	4	5.39	19	12.5
9	12	10.84	12	9.13	12	8.15	8	5.56
10	7	4.82	7	7.26	7	6.42	8	7.91
11	5	5.68	5	4.74	5	5.73	8	6.89

The statistical information for these four data sets are approximately similar. We can compute them as follows:

Anscombe's Data								
Observation	x1	y1	x2	y2	x3	y3	x4	y4
1	10	8.04	10	9.14	10	7.46	8	6.58
2	8	6.95	8	8.14	8	6.77	8	5.76
3	13	7.58	13	8.74	13	12.74	8	7.71
4	9	8.81	9	8.77	9	7.11	8	8.84
5	11	8.33	11	9.26	11	7.81	8	8.47
6	14	9.96	14	8.1	14	8.84	8	7.04
7	6	7.24	6	6.13	6	6.08	8	5.25
8	4	4.26	4	3.1	4	5.39	19	12.5
9	12	10.84	12	9.13	12	8.15	8	5.56
10	7	4.82	7	7.26	7	6.42	8	7.91
11	5	5.68	5	4.74	5	5.73	8	6.89
			Summary Statistics					
N	11	11	11	11	11	11	11	11
mean	9.00	7.50	9.00	7.500909	9.00	7.50	9.00	7.50
SD	3.16	1.94	3.16	1.94	3.16	1.94	3.16	1.94
r	0.82		0.82		0.82		0.82	

However, when these models are plotted on a scatter plot, each data set generates a different kind of plot that isn't interpretable by any regression algorithm, as you can see below:



We can describe the four data sets as:

ANSCOMBE'S QUARTET FOUR DATASETS

- Data Set 1: fits the linear regression model pretty well.
- Data Set 2: cannot fit the linear regression model because the data is non-linear.
- Data Set 3: shows the outliers involved in the data set, which cannot be handled by the linear regression model.

- Data Set 4: shows the outliers involved in the data set, which also cannot be handled by the linear regression model.

3. What is Pearson's R?

Answer: 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. However, it cannot capture nonlinear relationships between two variables and cannot differentiate between dependent and independent variables.

The Pearson Correlation Coefficient formula is as follows:

$$\mathbf{r} = \frac{\mathbf{n}(\sum \mathbf{xy}) - (\sum \mathbf{x})(\sum \mathbf{y})}{\sqrt{[\mathbf{n}\sum \mathbf{x^2} - (\sum \mathbf{x})^2][\mathbf{n}\sum \mathbf{y^2} - (\sum \mathbf{y})^2]}}$$

Where,

- r = Pearson Coefficient
- n= number of pairs of the stock
- ∑xy = sum of products of the paired stocks
- ∑x = sum of the x scores
- ∑y= sum of the y scores
- ∑x2 = sum of the squared x scores
- ∑y2 = sum of the squared y scores

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

Answer: Scaling is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Most of the time, the collected data set contains features highly varying in magnitudes, units and range. If scaling is not done, then algorithm only takes magnitude into account and not units hence leads to incorrect modelling. To solve this issue, we must do scaling to bring all the variables to the same level of magnitude.

Normalization/Min-Max Scaling:

It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

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

Standardization Scaling:

Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

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

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

Answer: If VIF is infinite, it 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.

Listed below are the various options to handle high VIF score:

- One approach is to review your independent variables and eliminate terms that are duplicates or not adding value to explain the variation in the model. For example, if your inputs are measuring the weight in kgs and lbs then just keep one of these variables in the model and drop the other one. Dropping the term with a large value of VIF will hopefully, fix the VIF for the remaining terms and now all the VIF factors are within the threshold limits. If dropping one term is not enough, then you may need to drop more terms as required.
- A second approach is to use principal component analysis and determine the optimal set of principal components that best describe your independent variables. Using this approach will get rid of your multicollinearity problem but it may be hard for you to interpret the meaning of these “new” independent variables.
- The third approach is to increase the sample size. By adding more data points to our model, hopefully, the confidence intervals for the model coefficients are narrower to overcome the problems associated with multicollinearity.
- The fourth approach is to transform the data to a different space like using a log transformation so that the independent variables are no longer correlated as strongly with each other.
- Finally, you can use a different type of model call ridge regression that better handles multicollinearity.

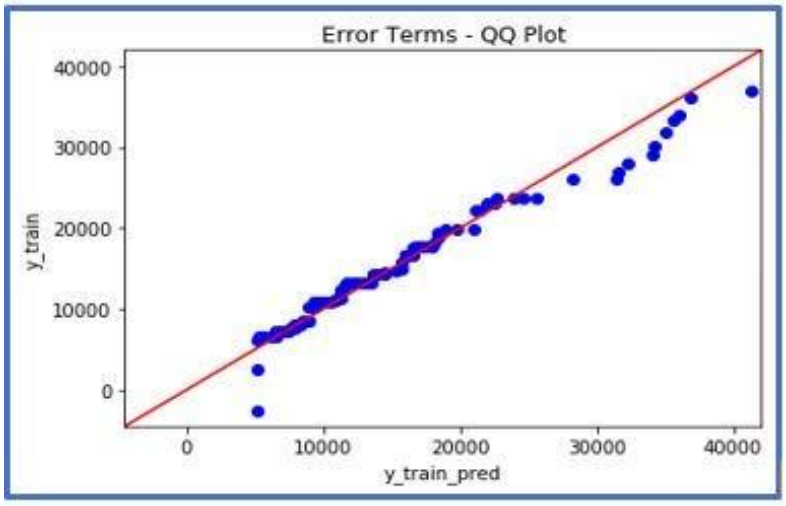
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Answer: Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data possibly 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.

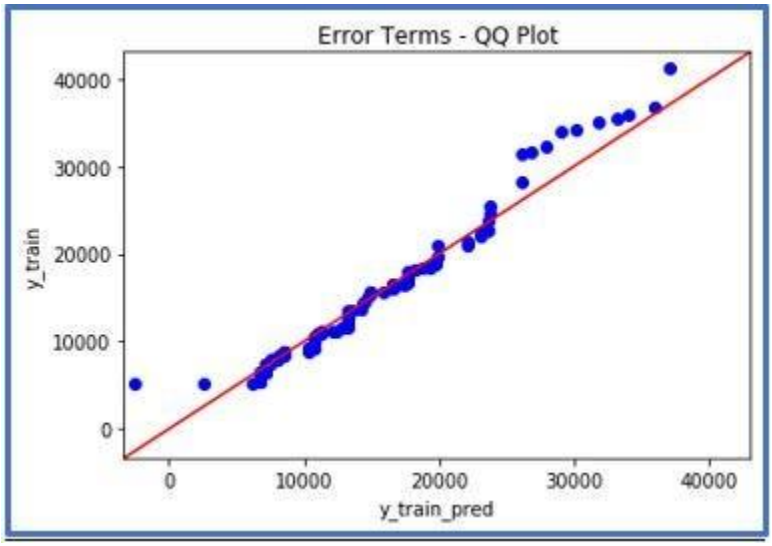
A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. Below are the possible interpretations for two data sets.

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