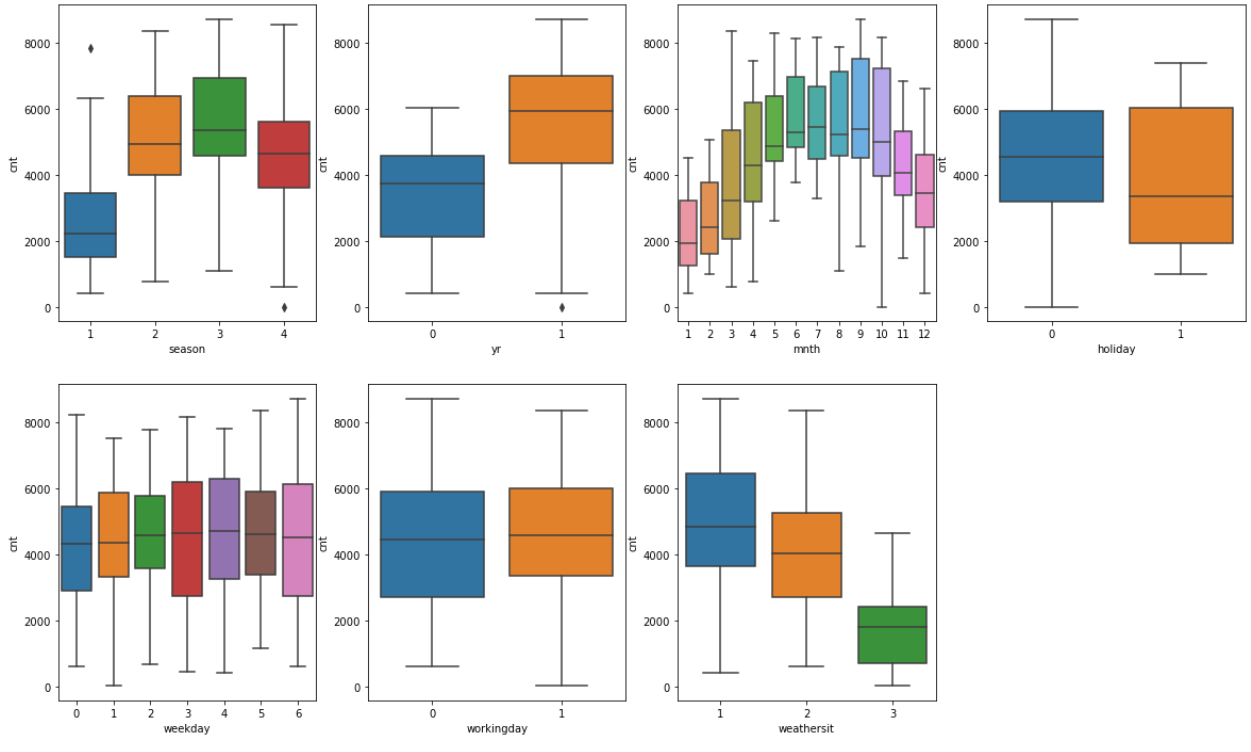
Assignment-based Subjective Questions

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

Answer: Categorical Variables: season, yr, mnth, holiday, weekday, workingday, weathersit

1. From the below box plots, it is clrealy visible that, for categories "season", "yr", "mnth", "holiday", "weathersit", there is significant variation in demand for different values of individual category. Hence it is worth considering these for model analysis and look for their individual co-efficients to see the impact of them on the end-model
2. For remaining two categorical variables "weekday" and "workingday", I do not see considerable variation in demand across all the categorical values. So for me it looks like, they may not identify themselves as affecting factors to our model. But I will not drop them in the beginning of our analysis but let our model building exercise decide the same



3. When I see season and mnth boxplots simultaneously, I see that season 1 which is Jan to Feb have similar cnt, season 2 which is Apr to Jun have similar cnt and so on. So there is a co-linearity observed between season and mnth variables. So in the final model, either a season or one of month from that season should appear but not all. This is the understanding I am getting while comparing these boxplots of categorical variables.

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

Answer: This has something to do with Multicollinearity in case if Multiple Linear Regression. Because, keeping k dummies for k levels of a categorical variable is good idea, but there is a redundancy of one level, which is here in separate column. This is not needed since one of the combinations will be uniquely representing this redundant column. Hence, it's better to drop one of the columns and just have k-1 dummies(columns) to represent k levels.

This Overall approach reduces Multicollinearity in the dataset, which is one of the prime Assumption of Multiple Linear Regression.

Let us take example of categorical variable "season". We will replace column values with below four values (1: spring, 2: summer, 3: fall, 4: winter)

The mean of count for every season is:

Fall = 5644.3
Spring = 2608.4
Summer = 4992.3
Winter = 4728.2

Dummy Coding:

Level Of Season	Season 1 vs 2	Season 1 vs 3	Season 1 vs 4
Fall	0	0	0
Spring	1	0	0
Summer	0	1	0
Winter	0	0	1

	coef	P>|t|
const	5700.3657	0.000
spring	-3102.5915	0.000
summer	-754.1737	0.000
winter	-1103.4523	0.000

The parameter estimate for the first season compares the mean of the dependent variable, cnt, for levels 1 and 2 yielding (2648-5644=3036) which is close to co-eff of spring i.e., -3102. The results of the second season, comparing the mean of cnt for levels 1 and 3. The expected difference in variable write between group 1 and 3 is -652, and so on. Notice that the intercept corresponds to the cell mean for season = fall i.e., close to 5644.

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

Answer:

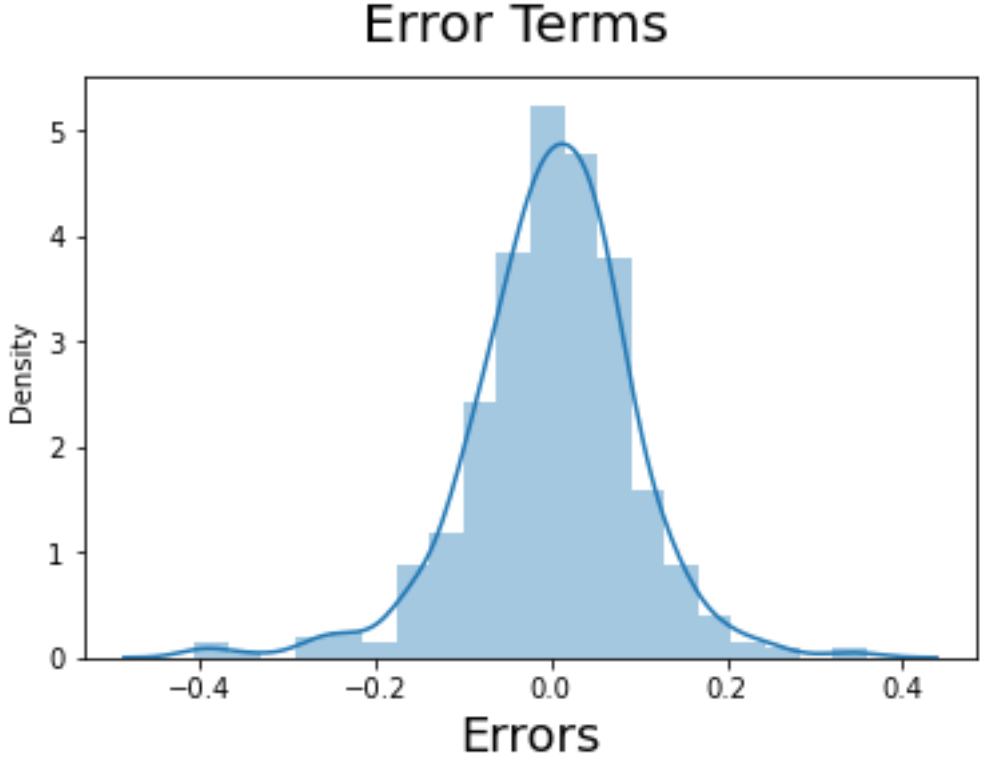
1. If we consider all the data without removing any column, we observe that column 'registered' has the highest correlation of 0.95 with the target variable.
2. But if we remove unnecessary columns like instant, registered, casual, temp then, with target variable 'cnt' we get highest correlation of 0.63 from column 'atemp'

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

Answer:

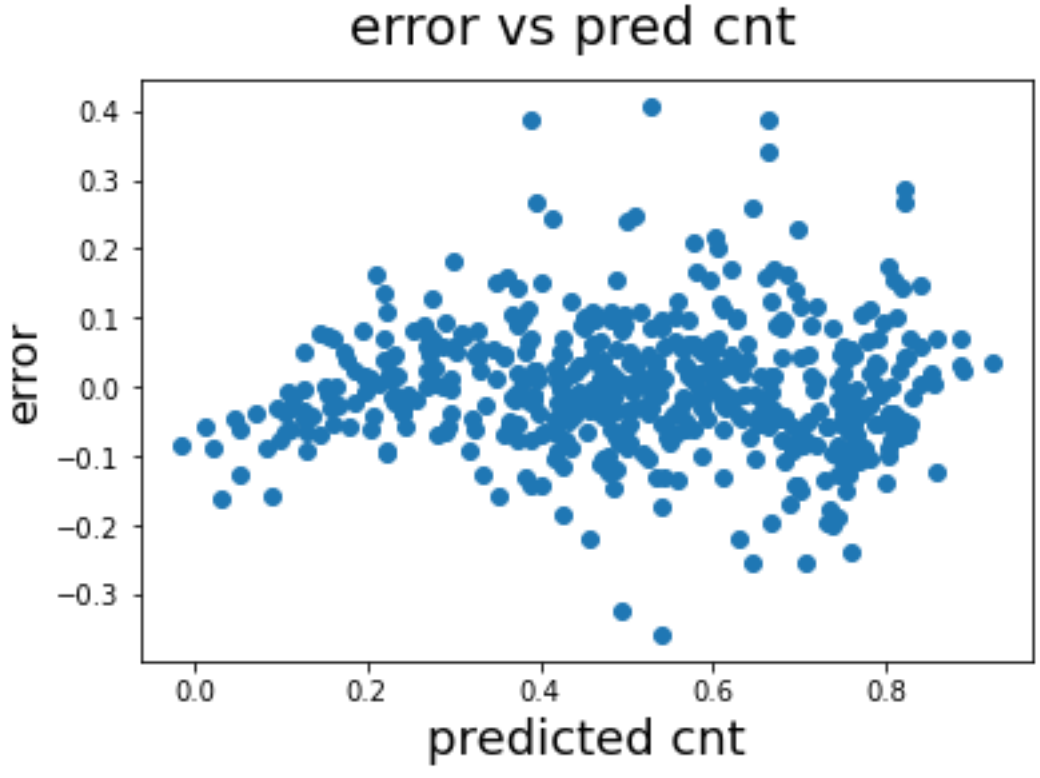
1. **linearity assumption**: Multiple variables show some kind of linear relationship with target variables "cnt".
2. **Residuals**:
    a. *Normality assumption:* It is assumed that the error terms, are normally distributed.
    b. *Zero mean assumption:* It is assumed that the residuals have a mean value of zero, i.e., the error terms are normally distributed around zero.

Both these assumptions can be validated from the below plot which is created from our final model



c. *Constant variance assumption:* It is assumed that the residual terms have the same (but unknown) variance, σ2. This assumption is also known as the assumption of homogeneity or homoscedasticity.
d. Independent error assumption: It is assumed that the residual terms are independent of each other, i.e., their pair-wise covariance is zero.

Both these assumptions can be validated from the below plots which is created from our final model



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

Answer: Top 3 features contributing significantly are:

1. atemp (feeling temperature in Celsius)
2. weathersit with value 3 (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds)
3. yr (especially the year 2019, which is the second year after introducing this service)

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer: When the data follows a straight-line trend, we think of linear regression. Linear regression attempts to model the relationship between two variables by fitting a linear equation to the observed data. One variable is considered to be a target/dependent variable, and the other(s) is/are considered to be a dependent variable(s).

What linear regression does is simply tell us the value of the dependent variable for an arbitrary independent/target variable.

From a machine learning context, it is the simplest model one can try out on the data. If we have a hunch that the data follows a straight-line trend, linear regression can give us quick and reasonably accurate results.

Simple predictions are all cases of linear regression. We first observe the trend and then predict based on the trend.

Hence it is important to understand that even though linear regression can be the first attempt at understanding the data it may not always be ideal.

Here's how we do linear regression:

1. We plot our dependent variable (y-axis) against the independent variable (x-axis)
2. We try to plot a straight line and measure correlation
3. We keep changing the direction of our straight line until we get the best correlation
4. We extrapolate from this line to find new values on y-axis

Characteristics of Linear Regression:

1. Linear Regression provides you with a straight line that lets you infer the dependent variables
2. Linear regression at its core is a method to find values for parameters that represent a line.
3. Linear regression is a form of supervised learning. Supervised learning involves those set of problems where we use existing data to train our machine.
4. Linear regression can involve multiple independent variables. but in its simplest form it involves 1 independent variable.

Here is how it works:

1. In its generic form it is written as
   $$Y = \beta 0Xo + \beta 1X1 + \beta 2X2 + \ldots.$$
   where all the betas are coefficients that our machine learning algorithm must figure out. The x's are known because they are independent. We can set them anything. What we need to find is Y.
2. For a single independent variable, the equation is reduced to
   $$Y = \beta 0Xo + \beta 1X1$$
   For simplification X0 is set to be equal to 1 and β0 is given the name c. x1 is called x and β1 = m. It reduces to:
   $Y = mX + c$

   a. To figure out m and c we draw a line, using an initial guess of m and c through the set of points that we already have.

b. We calculate the distance of this line from each of these points.
c. We take square-root of the sum of the squares of these distances (Cost Function) .
d. We keep changing m and c in small steps to see if this Cost Function decreases.
e. When the cost stops decreasing, we fix that m and c as our final result.
f. The resulting line is our best linear fit through the data.
g. Now for any new x we can figure out the y using this line.

3. It means that we keep *redrawing* the line until it seems to fit the data best.

What is clear is that linear regression is a simple approach to predict based on a data that follows a linear trend. It follows that we will fail rather drastically if we were to fit a sine curve or a circular data set.

Finally, linear regression is always a good first step if the data is visually linear.

2. Explain the Anscombe's quartet in detail.

Answer: 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 helps 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 provide same statistical information that involves variance, and mean of all x, y points in all four datasets.

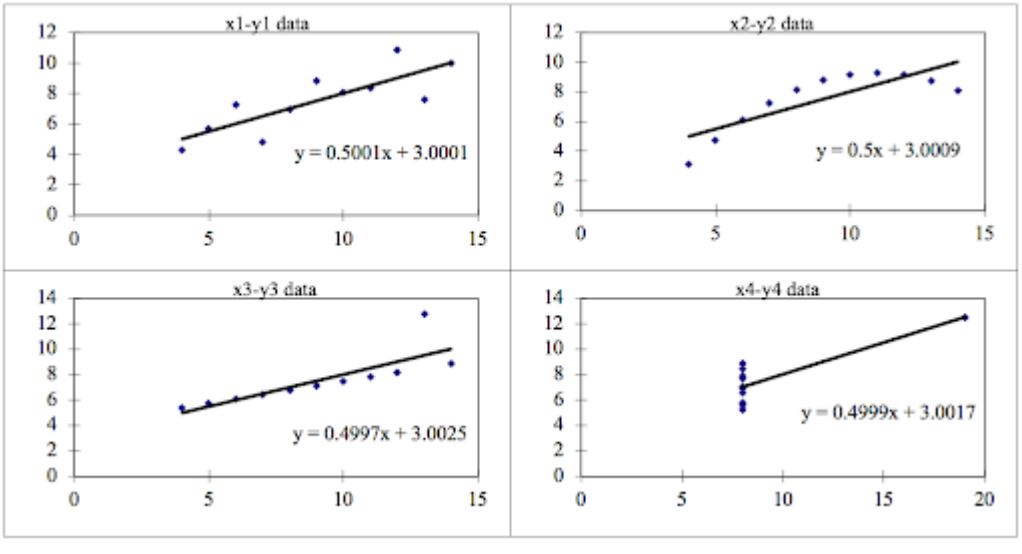
This tells us about the importance of visualizing the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc.

Also, the Linear Regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets.

These four plots can be defined as follows. The statistical information for all these four datasets are approximately similar.

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	

When these models are plotted on a scatter plot, all datasets generate a different kind of plot that is not interpretable by any regression algorithm which is fooled by these peculiarities and can be seen as follows:



The four datasets can be described as:

- Dataset 1: this fits the linear regression model well.
- Dataset 2: this could not fit linear regression model on the data quite well as the data is non-linear.
- Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression model.
- Dataset 4: shows the outliers involved in the dataset which cannot be handled by linear regression model.

We have described the four datasets that were intentionally created to describe the importance of data visualization and how any regression algorithm can be fooled by the same. Hence, all the important features in the dataset must be visualized before implementing any machine learning algorithm on them which will help to make a good fit model.

3. What is Pearson's R?

Answer: The Pearson's Correlation Coefficient is also referred to as Pearson's R or bivariate correlation. Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations.

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.

Using the formula proposed by Karl Pearson, we can calculate a linear relationship between the two given variables. For example, a child's height increases with his increasing age (different factors affect this biological change). So, we can calculate the relationship between these two variables by obtaining the value of Pearson's Correlation Coefficient R.

There are certain requirements for Pearson's Correlation Coefficient:

- Scale of measurement should be interval or ratio
- Variables should be approximately normally distributed
- The association should be linear
- There should be no outliers in the data

The formula for Pearson's R is given by:

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

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

Interpretations from Pearson's R:

- The positive value of Pearson's correlation coefficient implies that if we change either of these variables, there will be a positive effect on the other.
- The more inclined the value of the Pearson correlation coefficient to -1 and 1, the stronger the association between the two variables.

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

Answer:

Scaling: It 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

Need: 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 modelling. To solve this issue, we must do scaling to bring all the variables to the same level of magnitude.

Normalized Scaling: It brings all the data in the range of 0 and 1

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

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

$$x_{\text{stand}} = \frac{x - \text{mean}(x)}{\text{standard deviation } (x)}$$

- It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.
- One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

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

Answer: If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables.

In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity.

Multicollinearity can also occur when we use dummy variables to handle discrete independent variables and if we include the default values as well in the model.

To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

Multicollinearity refers to the problem when the independent variables are collinear. Collinearity refers to a linear relationship between two explanatory variables. Two variables are perfectly collinear if there is an exact relationship between the two variables. If the independent variables are perfectly collinear, then our model becomes singular, and it would not be possible to uniquely identify the model coefficients mathematically

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 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.

Few advantages:

1. It can be used with sample sizes also
2. 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.

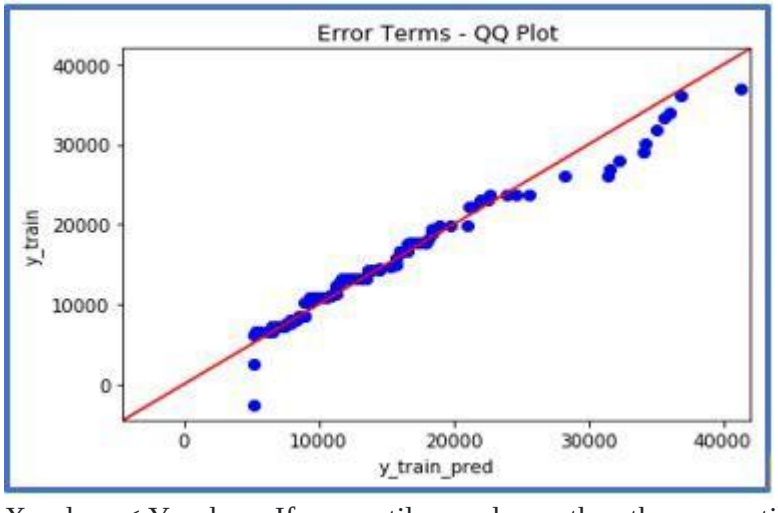
It is used to check following scenarios: If two data sets —

1. come from populations with a common distribution
2. have common location and scale
3. have similar distributional shapes
4. have similar tail behavior

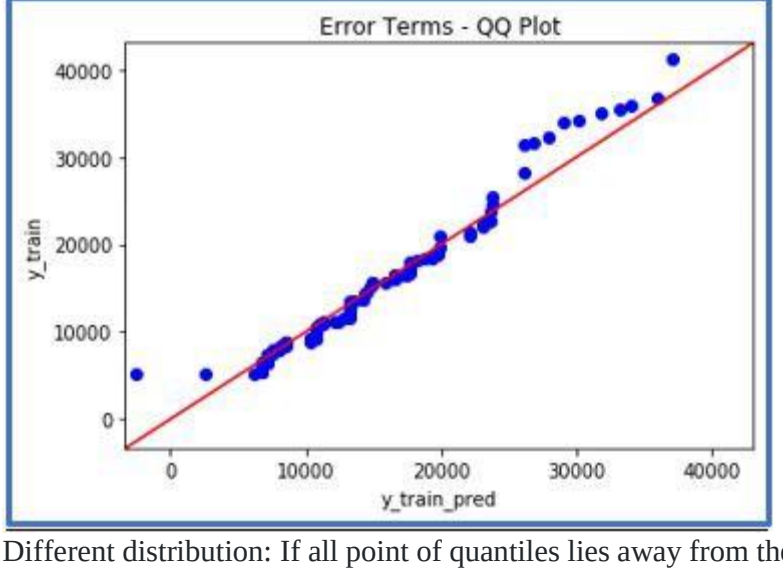
Interpretation:

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.

1. Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis
2. Y-values < X-values: If y-quantiles are lower than the x-quantiles.



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



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