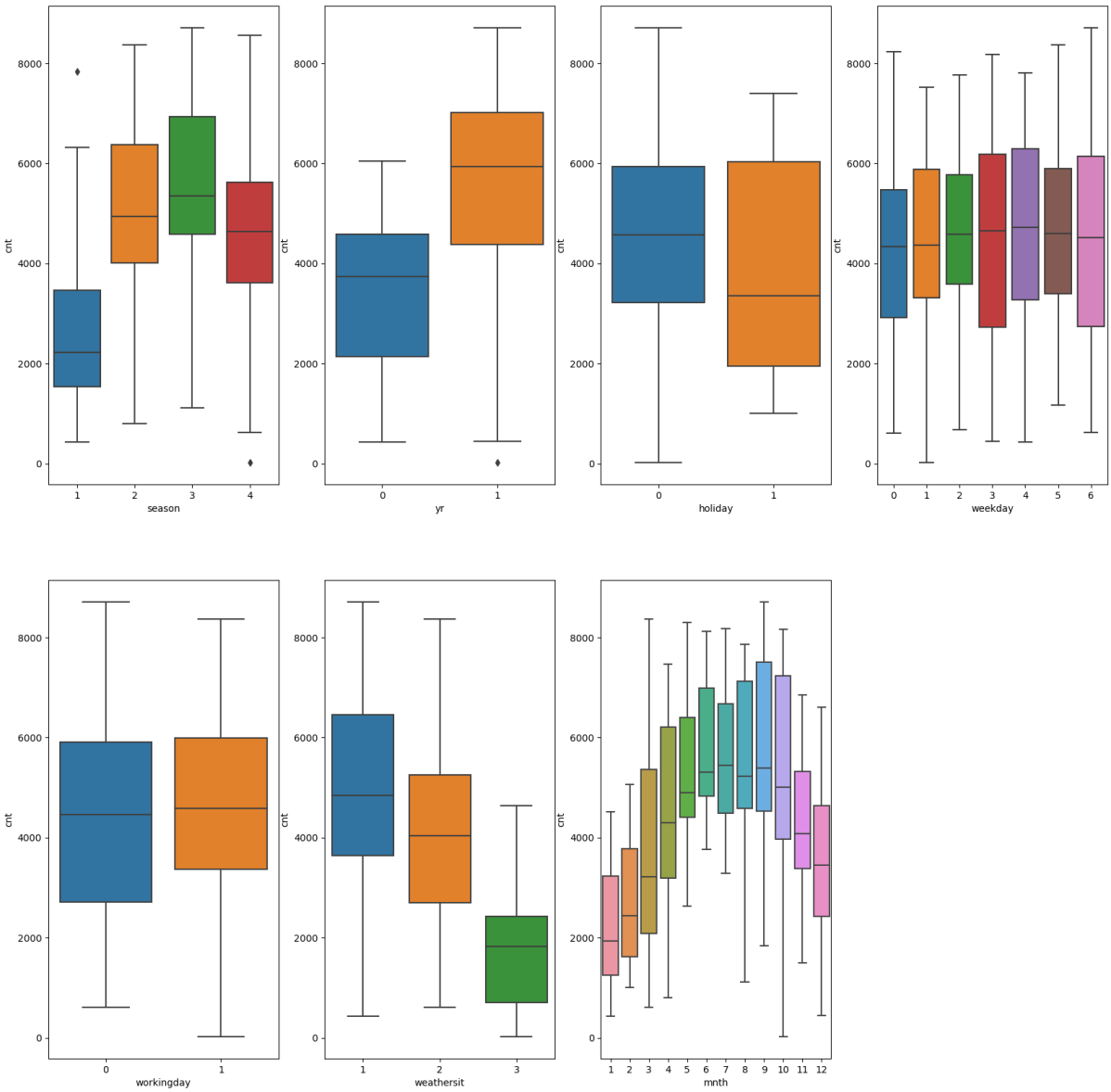
Assignment-based Subjective Questions

1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)**



The categorical variable in the dataset were season , yr , holiday, weekday ,workingday, and weathersit and mnth . These were visualized using a boxplot (Fig. attached) .
These variables had the following effect on our dependant variable:-

- Season - The boxplot showed that spring season had least value of cnt whereas fall had maximum value of cnt. Summer and winter had intermediate value of cnt.
- Weathersit - There are no users when there is heavy rain/ snow indicating that this weather is extremely unfavourable. Highest count was seen when the weathersit was' Clear, Partly Cloudy'.
- Yr - The number of rentals in 2019 was more than 2018
- Holiday - rentals reduced during holiday.

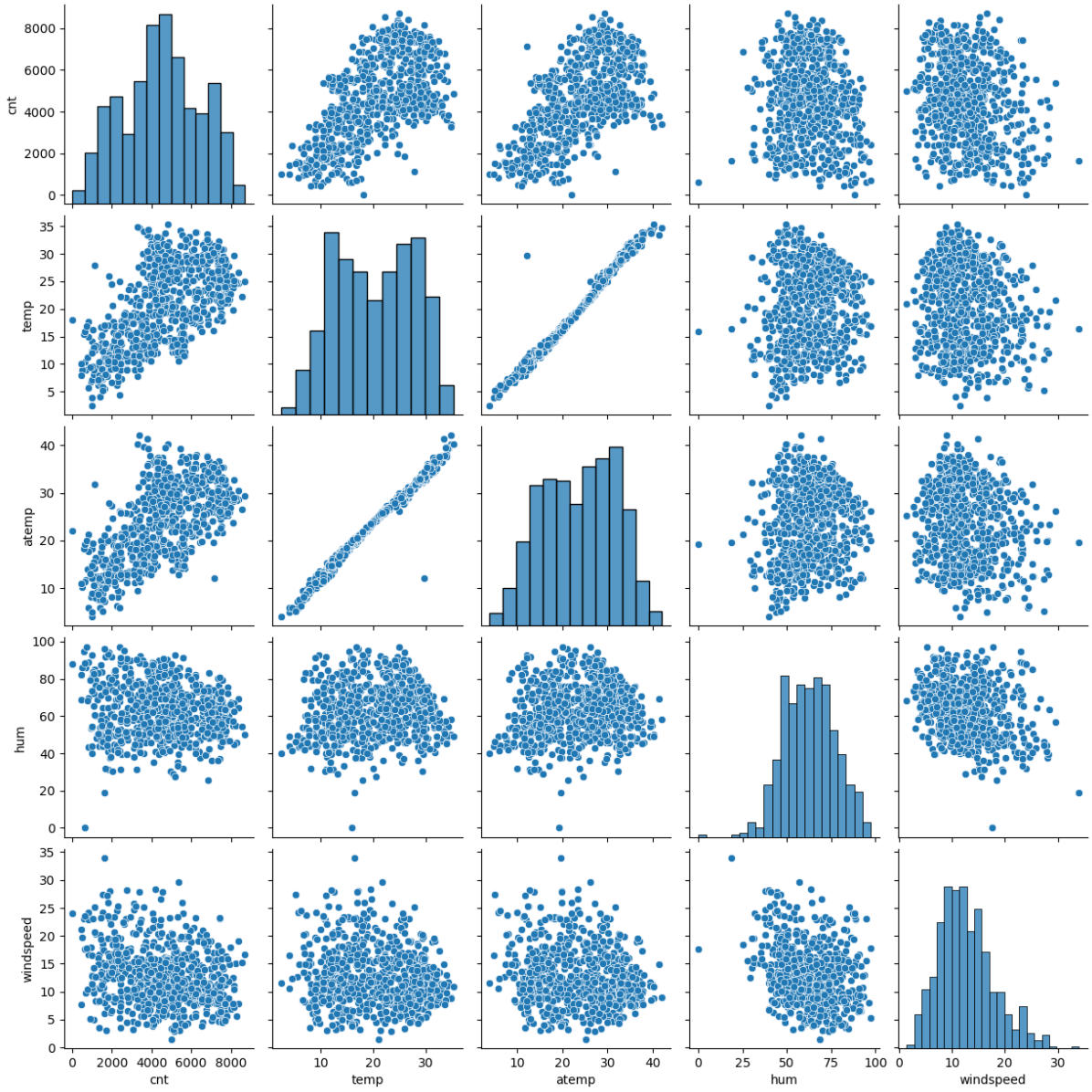
- Mnth - September saw highest no of rentals while December saw least. This observation is in accordance with the observation made in weathersit. The weather situation in December is usually heavy snow due to which the rentals might have dropped.
- Weekday - The count of rentals is almost even throughout the week
- Workingday – The median count of users is constant almost throughout the week.

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

Drop_first = True is important to use , as it helps in reducing the extra column created during dummy variable creation . Hence it reduces the correlations created among dummy variables .

In our boom bike case we will get one extra column for each categorical values like months, season, weathersit and weekday.

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

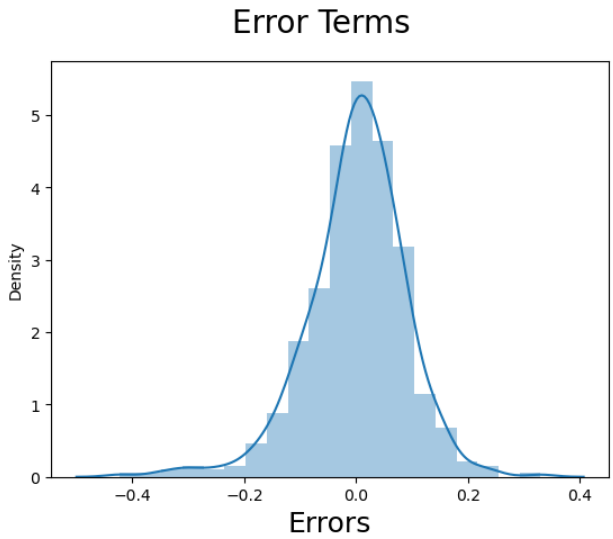


Using the below pairplot it can be seen that , “temp” and “atemp” are the two numerical variables which are highly correlated with the target variable (cnt)

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

First, linear regression needs the relationship between the independent and dependent variables to be linear. We visualised the numeric variables using a pairplot to see if the variables are linearly related or not. Refer to the notebook for more details.

Secondly, Residuals distribution should follow normal distribution and centred around 0 (mean = 0). We validated this assumption about residuals by plotting a distplot of residuals and saw if residuals are following normal distribution or not. The diagram below shows that the residuals are distributed about mean = 0.



Thirdly, linear regression assumes that there is little or no multicollinearity in the data. Multicollinearity occurs when the independent variables are too highly correlated with each other. We calculated the VIF (Variance Inflation Factor) to get the quantitative idea about how much the feature variables are correlated with each other in the new model. Refer to the notebook for more details.

	Features	VIF
12	temp	7.07
11	workingday	5.24
13	windspeed	4.67
0	season_Spring	3.08
1	season_Summer	2.33
9	yr	2.08
2	season_Winter	1.99
6	weekday_Saturday	1.97
3	mnth_Jan	1.62
4	mnth_Jul	1.59
8	weathersit_Mist & Cloudy	1.57
5	mnth_Sep	1.35
10	holiday	1.17
7	weathersit_Light Snow & Rain	1.09

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

The top 3 features are:

temp - coefficient : 0.472823

yr - coefficient : 0.234361

weathersit_Light Snow & Rain - coefficient : -0.291727

General Subjective Questions

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

Linear Regression is a type of supervised Machine Learning algorithm that is used for the prediction of numeric values. Linear Regression is the most basic form of regression analysis. Regression is the most commonly used predictive analysis model.

Linear regression is based on the popular equation “y = mx + c”.

It assumes that there is a linear relationship between the dependent variable(y) and the predictor(s)/independent variable(x).

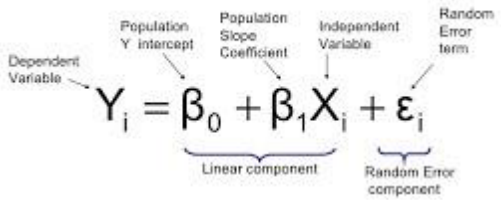
In regression, we calculate the best fit line which describes the relationship between the independent and dependent variable. Regression is performed when the dependent variable is of continuous data type and Predictors or independent variables could be of any data type like continuous, nominal/categorical etc. Regression method tries to find the best fit line which shows the relationship between the dependent variable and predictors with least error.

In regression, the output/dependent variable is the function of an independent variable and the coefficient and the error term.

Regression is broadly divided into simple linear regression and multiple linear regression.

1. Simple Linear Regression : SLR is used when the dependent variable is predicted using only one independent variable.

The equation for SLR will be:



2. Multiple Linear Regression :MLR is used when the dependent variable is predicted using multiple independent variables.

The equation for MLR will be:

$$observed\ data \rightarrow \quad y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p + \varepsilon$$

$$predicted\ data \rightarrow \quad y' = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p$$

$$error \rightarrow \quad \varepsilon = y - y'$$

B1 = coefficient for X1 variable

B2 = coefficient for X2 variable

B3 = coefficient for X3 variable and so on…

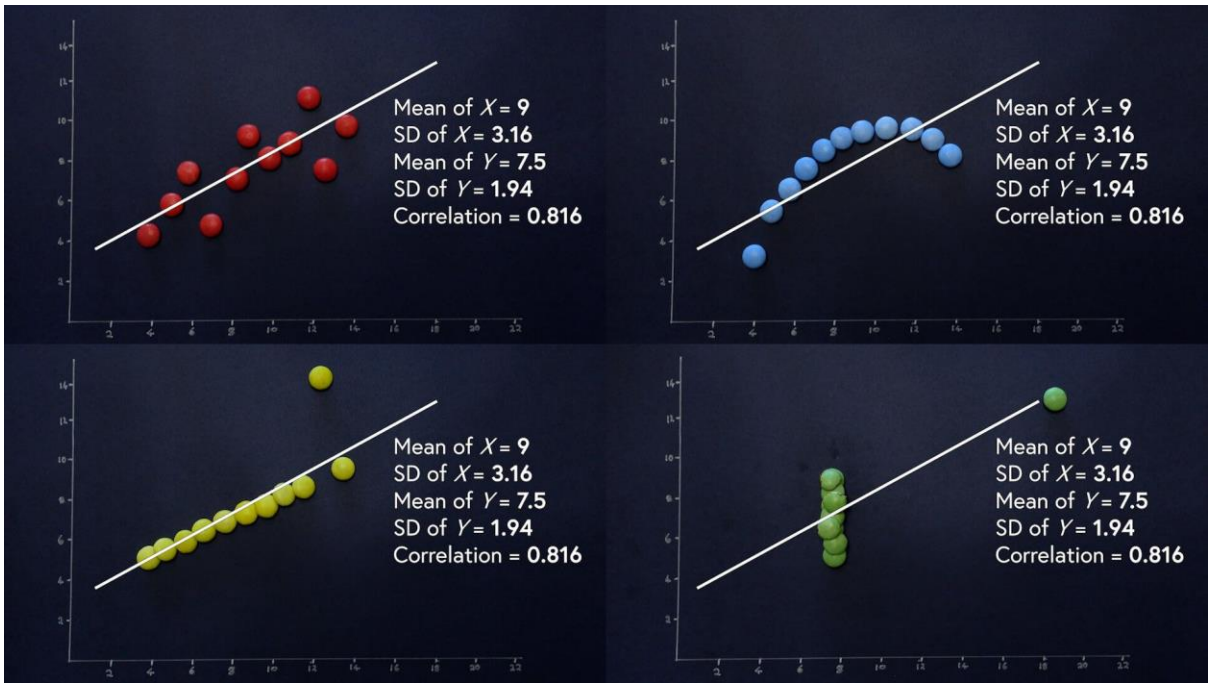
B0 is the intercept (constant term)

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

Anscombe's quartet is a group of datasets (x, y) that have the same mean, standard deviation, and regression line, but which are qualitatively different. It is often used to illustrate the importance of looking at a set of data graphically and not only relying on basic statistic properties.

Anscombe's quartet comprises a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph.

The four data sets comprising the Anscombe's quartet: all four sets have identical statistical parameters, but the graphs show them to be considerably different



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

The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables. When one variable changes, the other variable changes in the same direction.

Formula

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

r = correlation coefficient

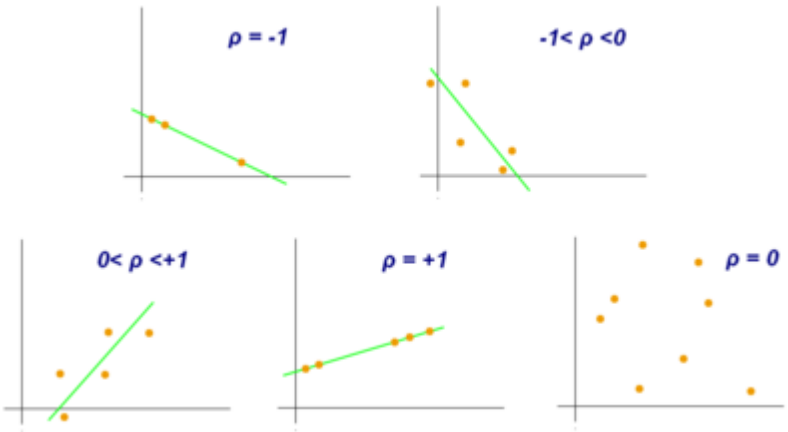
x_i = values of the x-variable in a sample

$\bar{x}$ = 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

As can be seen from the graph below,
r = 1 means the data is perfectly linear with a positive slope
r = -1 means the data is perfectly linear with a negative slope
r = 0 means there is no linear association



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

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.

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 have to do scaling to bring all the variables to the same level of magnitude.

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.

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).

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

- sklearn.preprocessing.scale helps to implement standardization in python.

- 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? (3 marks)

The VIF gives how much the variance of the coefficient estimate is being inflated by collinearity. If there is perfect correlation, then VIF = infinity. It gives a basic quantitative idea about how much the feature variables are correlated with each other. It is an extremely important parameter to test our linear model.

$$VIF = \frac{1}{1 - R^2}$$

Where R-1 is the R-square value of that independent variable which we want to check how well this independent variable is explained well by other independent variables. If that independent variable can be explained perfectly by other independent variables, then it will have perfect correlation and

it's R-squared value will be equal to 1. So, VIF = 1/(1-1) which gives VIF = 1/0 which results in "infinity" .

The numerical value for VIF tells you (in decimal form) what percentage the variance (i.e. the standard error squared) is inflated for each coefficient. For example, a VIF of 1.9 tells you that the variance of a particular coefficient is 90% bigger than what you would expect if there was no multicollinearity — if there was no correlation with other predictors.

A rule of thumb for interpreting the variance inflation factor:

- 1 = not correlated.
- Between 1 and 5 = moderately correlated.
- Greater than 5 = highly correlated.

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

Q-Q plots are also known as Quantile-Quantile plots. As the name suggests, they plot the quantiles of a sample distribution against quantiles of a theoretical distribution. Doing this helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential.

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:

a) It can be used with sample sizes also

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.

It is used to check following scenarios:

If two data sets —

i. come from populations with a common distribution

ii. have common location and scale

iii. have similar distributional shapes

iv. have similar tail behavior

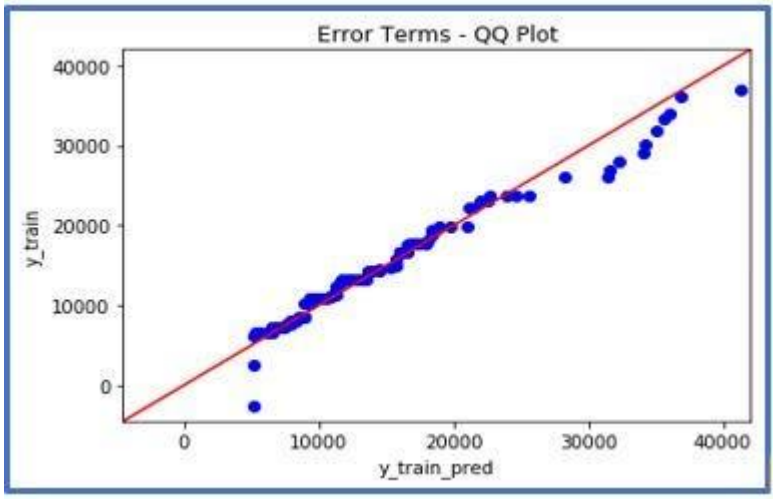
Interpretation:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

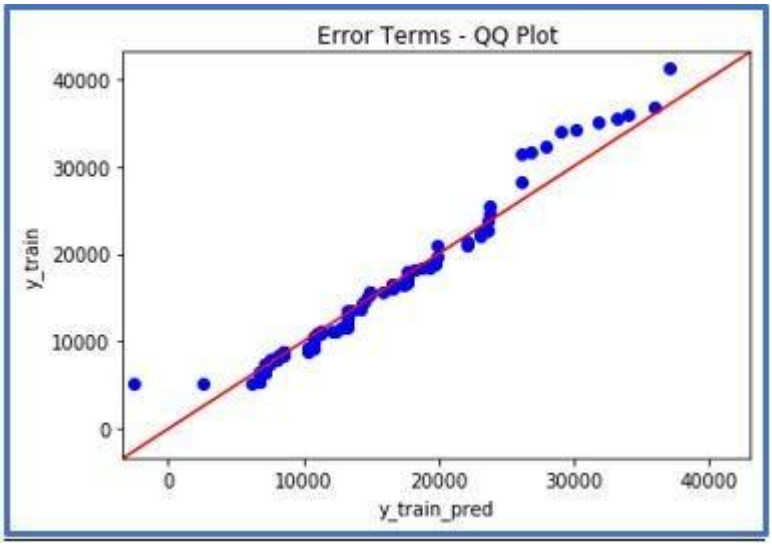
Below are the possible interpretations for two data sets.

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