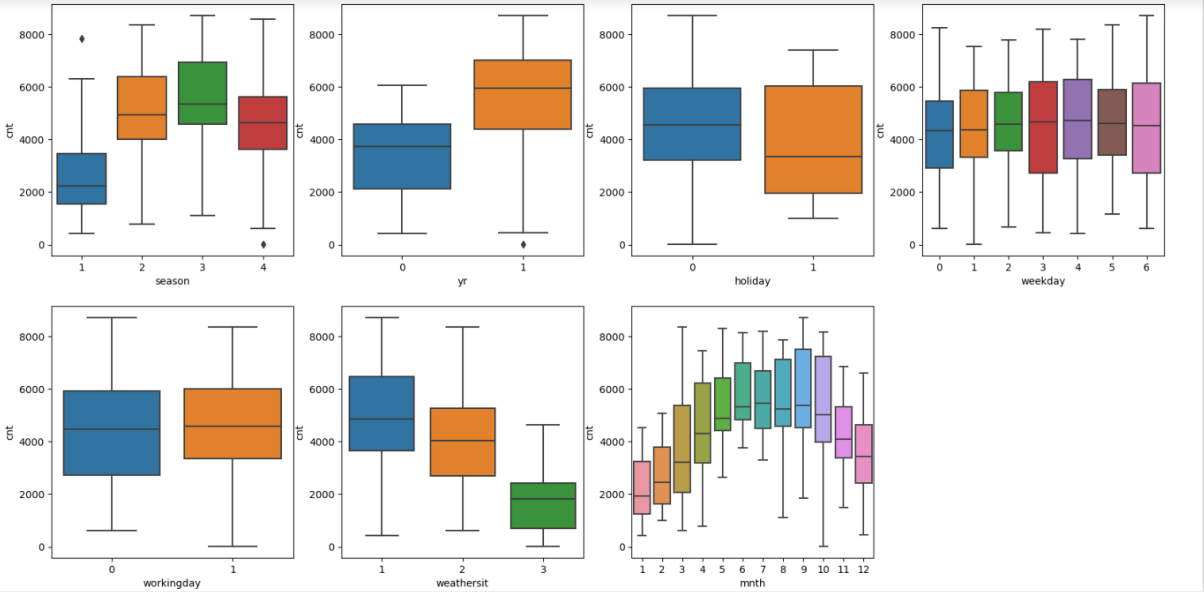
Assignment-based Subjective Questions

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



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.

Also, from our analysis of the categorical variables from the dataset we can predict the formula for the best fit line equation:

$$cnt = 0.12155 + (0.2342 * year) - (0.0973 * holiday) + (0.4728 * temp) - (0.1549 * windspeed) + (0.0762 * \text{mnth\_}sep) + (-0.2864 * \text{weathersit\_Light Snow \& Rain}) + (-0.0807 \times \text{weathersit\_Mist \& Cloudy}) + (-0.0615 \times \text{season\_}Spring) + (0.0422 \times \text{season\_}Summer) + (0.0771 \times \text{season\_}Winter)$$

- **YR** - Coefficient of yr indicates that a unit increase in yr variable, will **increase** bike hiring by **0.2342** values.
- **HOLIDAY** - Coefficient of holiday indicates that a unit increase in holiday variable, will **decrease** the bike hiring by **0.0973** values.

- **TEMP** - Coefficient of temp indicates that a unit increase in temp variable, will **increase** the bike hiring by **0.4728** values.
- **WINDSPEED** - Coefficient of mnth_3 indicates that w.r.t. to mnth_1, a unit increase in the mnth_9, will **increase** the bike hiring by **0.0806** values.
- **MNTH_*sep*** - Coefficient of mnth_sep indicates that w.r.t. to Season_1, a unit increase in the season_3, will **increase** the bike hiring by **0.0762** values.
- **WEATHERSIT_LIGHT SNOW & RAIN** - Coefficient of weathersit_Light Snow & Rain indicates that a unit increase in the weathersit_Light Snow & Rain will **decrease** the bike hiring by **0.2864** values
- **WEATHERSIT_LIGHT MIST & CLOUDY** - Coefficient of weathersit_Light Mist & Cloudy indicates that a unit increase in the of weathersit_Light Mist & Cloudy will **decrease** the bike hiring by **0.0807** values
- **SEASON_SPRING**- Coefficient of season_spring indicates that a unit increase in the season_spring, will **decrease** the bike hiring by **0.0615** values.
- **SEASON_SUMMER** - Coefficient of season_summer indicates that, a unit increase in the mnth_9, will **increase** the bike hiring by **0.0422** values.
- **SEASON_WINTER** - Coefficient of season_winter indicates that a unit increase in the season_winter, will **increase** the bike hiring by **0.0771** values.

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

   If we don't drop the first column then the dummy variables will be correlated (redundant). This may affect some models adversely and the effect is stronger when the cardinality is smaller. For example iterative models may have trouble converging and lists of variable importance may be distorted. In the figure below , check the number of columns, instead of 13 we got 12 columns. It removes the first column of the get_dummies dataframe. The first column for the “Body Color” column is Beige. If there is a beige car, all columns are 0. When all columns are 0, the model knows it’s a beige car. More columns mean less performance and more training time. Imagine we have 20 columns that are not numerical. If we use ‘drop_first’, we get 20 columns less. So it is useful to use the drop_first = True parameter for model performance.

```
1 pd.get_dummies(df['Body Color'])
```

	Beige	Black	Blue	Bronze	Brown	Green	Grey	Orange	Red	Silver	Violet	White	Yellow
0	0	1	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	1	0	0	0	0
2	0	1	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	1	0	0	0	0	0	0	0	0
4	0	1	0	0	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
4795	0	0	0	0	0	0	0	0	0	0	0	1	0
4796	0	0	0	0	0	0	0	0	0	0	0	1	0
4797	0	0	0	0	0	0	0	0	0	1	0	0	0
4798	0	0	0	0	0	0	0	0	0	1	0	0	0
4799	0	0	0	0	0	0	0	0	0	1	0	0	0

4800 rows × 13 columns

```
1 pd.get_dummies(df['Body Color'], drop_first = True)
```

	Black	Blue	Bronze	Brown	Green	Grey	Orange	Red	Silver	Violet	White	Yellow
0	1	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	1	0	0	0	0
2	1	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	1	0	0	0	0	0	0	0	0
4	1	0	0	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...
4795	0	0	0	0	0	0	0	0	0	0	1	0
4796	0	0	0	0	0	0	0	0	0	0	1	0
4797	0	0	0	0	0	0	0	0	1	0	0	0
4798	0	0	0	0	0	0	0	0	1	0	0	0
4799	0	0	0	0	0	0	0	0	1	0	0	0

4800 rows × 12 columns

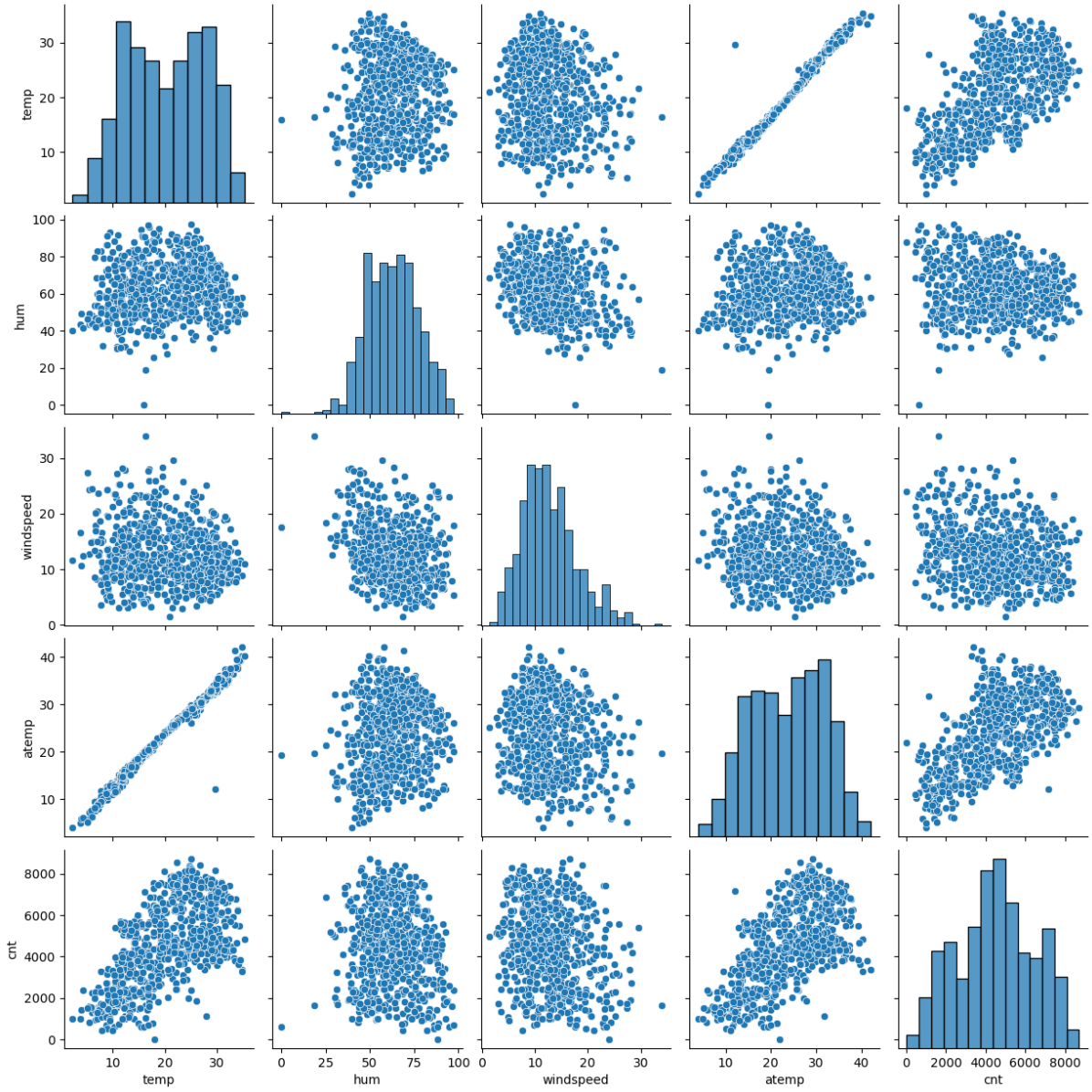
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

   By looking at the pair-plot, **'TEMP'** has the highest correlation among the other numerical variables with the '**CNT'** as the target variable.

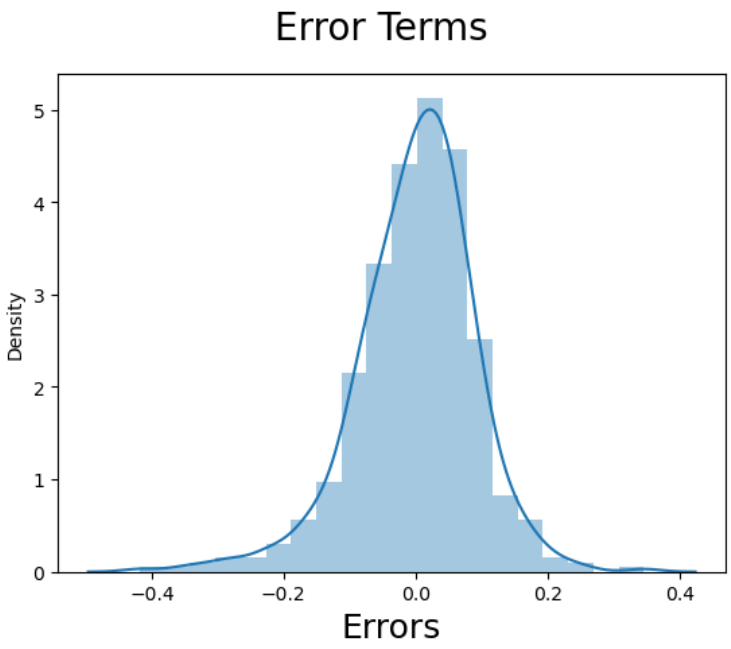
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

The following tests were done to validate the assumptions of linear regression:

1. 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 below figure for more details. We could see there is a linear relation between temp and atemp variable with the predictor 'cnt'.



2. 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.



3. 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.
4. 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.
5. Independent error assumption: It is assumed that the residual terms are independent of each other, i.e. their pair-wise covariance is zero.

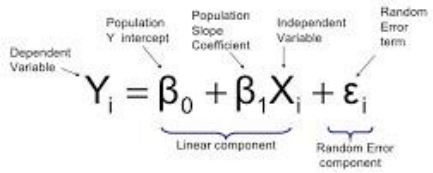
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
   The top 3 features are:
   1. Temperature (0.4728)
   2. year (0.234287)
   3. weathersit_Light Snow & Rain (−0.2864)

General Subjective Questions:

1. Explain the linear regression algorithm in detail.

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:

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

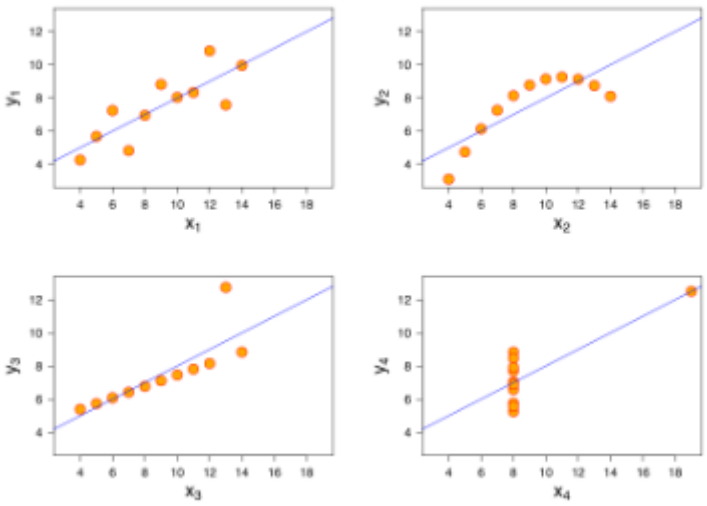
$$predicted\ data \rightarrow \quad y' = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p$$

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

b1 = coefficient for x1 variable, b2 = coefficient for x2 variable, b3 = coefficient for x3 variable and so on… ,b0 is the intercept (constant term)

3. Explain the Anscombe's quartet in detail.

Anscombe's Quartet was developed by statistician Francis Anscombe. It includes four data sets that have almost identical statistical features, but they have a very different distribution and look totally different when plotted on a graph. It was developed to emphasize both the importance of graphing data before analyzing it and the effect of outliers and other influential observations on statistical properties



- The first scatter plot (top left) appears to be a simple linear relationship.
- The second graph (top right) is not distributed normally; while there is a relation between them,it's not linear.
- In the third graph (bottom left), the distribution is linear, but should have a different regression line The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

4. What is Pearson's R?

Correlation coefficients are used to measure how strong a relationship is between two variables. There are several types of correlation coefficient, but the most popular is Pearson's. Pearson's correlation (also called Pearson's *R*) is a correlation coefficient commonly used in linear regression. If you're starting out in statistics, you'll probably learn about Pearson's *R* first. In fact, when anyone refers to the correlation coefficient, they are usually talking about Pearson's.

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

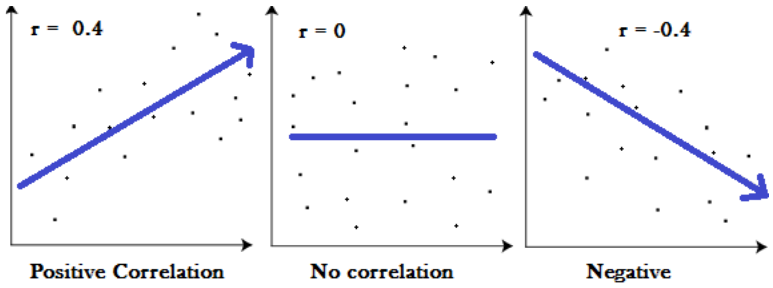
x_i = values of the x-variable in a sample

$\bar{x}$ = 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



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

Scaling: Feature Scaling is a technique to standardize the independent features present in the data in a fixed range.
It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

Example: If an algorithm is not using feature scaling method then it can consider the value 3000 meter to be greater than 5 km but that's actually not true and in this case, the algorithm will give wrong predictions. So, we use Feature Scaling to bring all values to same magnitudes and thus, tackle this issue B3 = coefficient for X3 variable and so on…
B0 is the intercept (constant term)

Difference between Normalization and Standardization:

a. **Min-Max Normalization**: This technique re-scales a feature or observation value with distribution value between 0 and 1.

$$X_{new} = \frac{X_i - \min(X)}{\max(x) - \min(X)}$$

b. **Standardization:** It is a very effective technique which re-scales a feature value so that it has distribution with 0 mean value and variance equals to 1.

$$X_{new} = \frac{X_i - X_{mean}}{\text{Standard Deviation}}$$

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

If there is perfect correlation, then VIF (Variance Inflation Factor) = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which leads 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.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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.

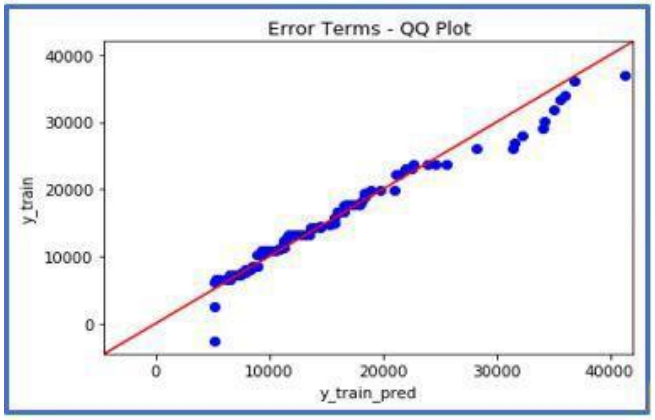
7. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression

A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight.

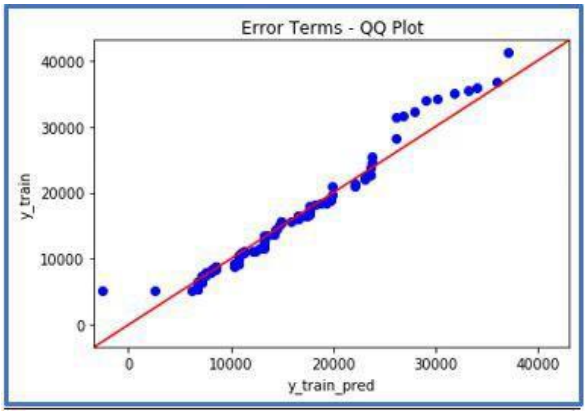
This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

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.
If two data sets —
• Come from populations with a common distribution.
• Have common location and scale.
• Have similar distributional shapes.
• Have similar tail behavior.

Below are the possible regressions 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.