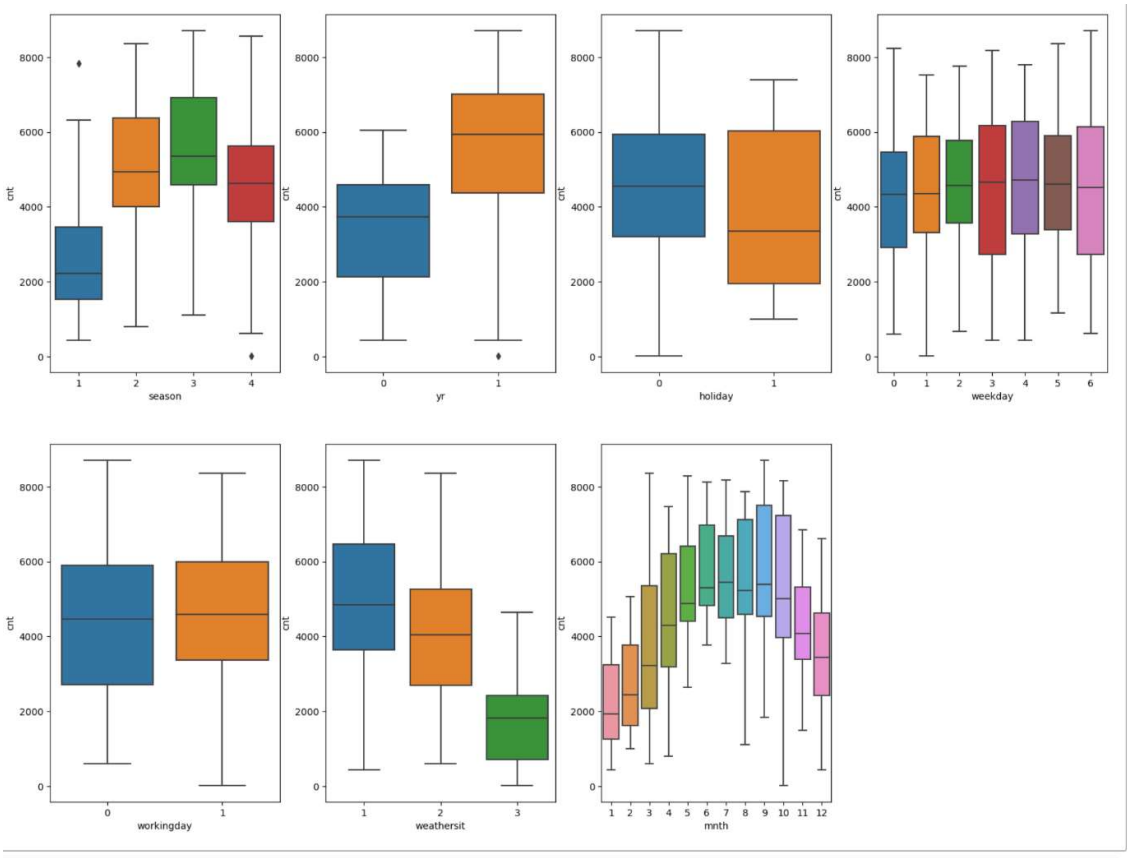
Assignment-based Subjective Questions

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

Answer

Let us review the Visualising Categorical Variables - Using a Boxplot which is plotted as part of the analysis.

The below graph is plotted for all the categorical variables in the dataset.



Below are the inferences from graphs.

- **Season -** The variable season's category 3 (Fall) has the highest median, which shows that the demand was high during this season. It is least for 1 (spring). So, the count of bike sharing is least for spring and high in fall.
- **Yr -** During the year 2019, number of bikes shared are higher as there is a high count of users as compared to the year 2018.
- **Weekday -** The bike demand is almost constant throughout the week.
- **Workingday -** The count of total users is in between 4000 to 6000 (~5500) during clear weather.
- **Mnth -** The count is highest in the month of September.
- **Holiday -** The count of users is less during the holidays.
- **Weathersit -** The count has zero values for weather situation with category 4 ('Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog')

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

Answer: When we have a categorical variable with say 'n' levels, the idea of dummy variable creation is to build 'n-1' variables, indicating the levels. But we can clearly see that there is no need of defining three different levels. If we drop a level, say 'XYZ', we would still be able to explain the three levels without this level.

If we don't drop the dummy variables, they will add to a correlation causing redundancy. This will affect the model adversely. Also, the effect is stronger when the cardinality is smaller.

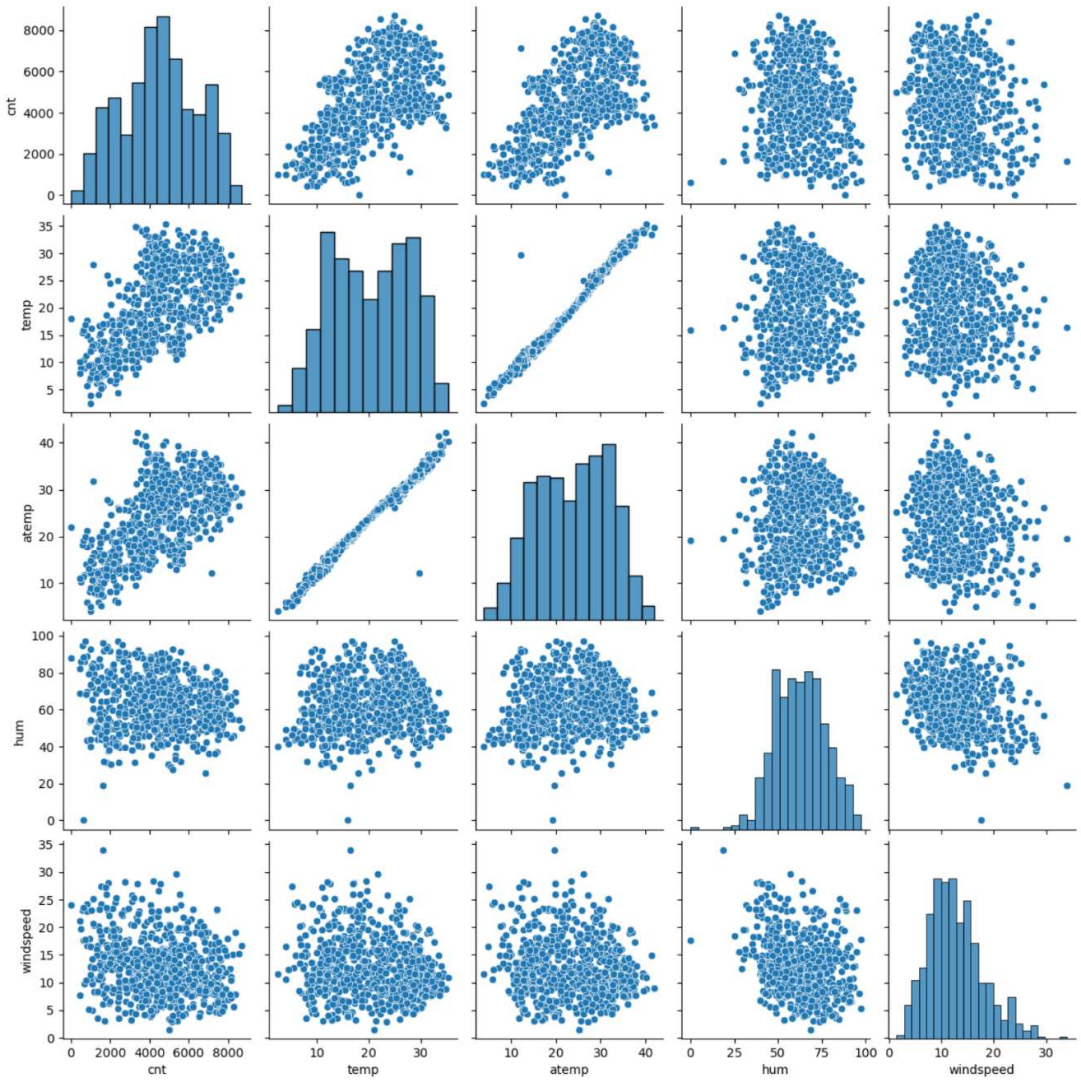
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation

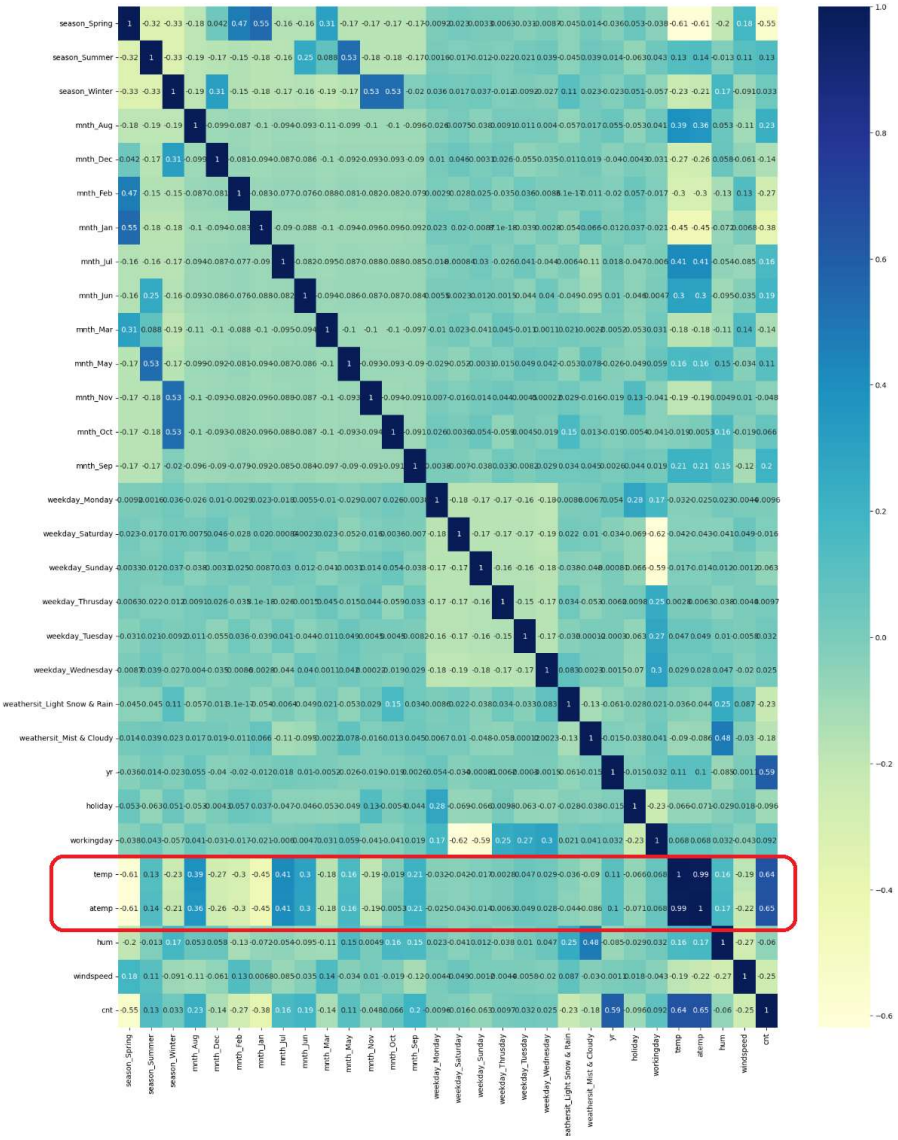
with the target variable? (1 mark)

Answer:

As can be seen from the map, atemp and temp seems to be correlated to the target variable cnt. Since, not much can be stated about the other independent variables, hence we will build a model using all the columns.

From the pair plot and correlation matrix, we can see that, **"temp"** and **"atemp"** are the two numerical variables which are highly correlated with the target variable **"Cnt".** Please check the pair plot as well as the correlation matrix explaining this.



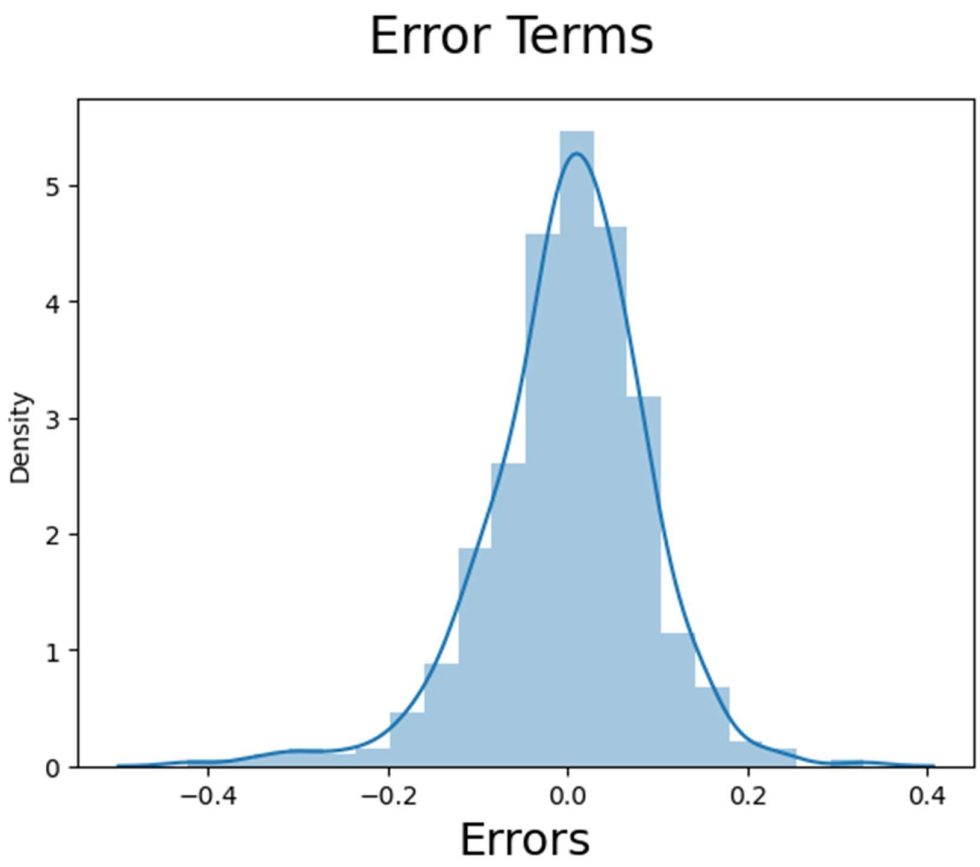


4. How did you validate the assumptions of Linear Regression after building the model on the

training set? (3 marks)

Answer:

As per one of the major assumptions of linear regression, the error items should be centred around zero and it should follow normal distribution pattern. Residual analysis is conducted to validate if residual distribution is following normal distribution pattern and if it's centred around 0(Mean = 0). This can be done by plotting the distplot of residuals. The below graph depicts the residuals are distributed with mean = 0.



5. Based on the final model, which are the top 3 features contributing significantly towards

explaining the demand of the shared bikes? (2 marks)

Answer: In the final model top three features contributing significantly towards explaining the demand are:

- **Temperature** - 0.472823 - Temp is the most significant with the largest coefficient and is followed by
- **weathersit** : Light Snow, Light Rain + Mist & Cloudy (-0.291727) and
- **year** (0.234361).
-

General Subjective Questions

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

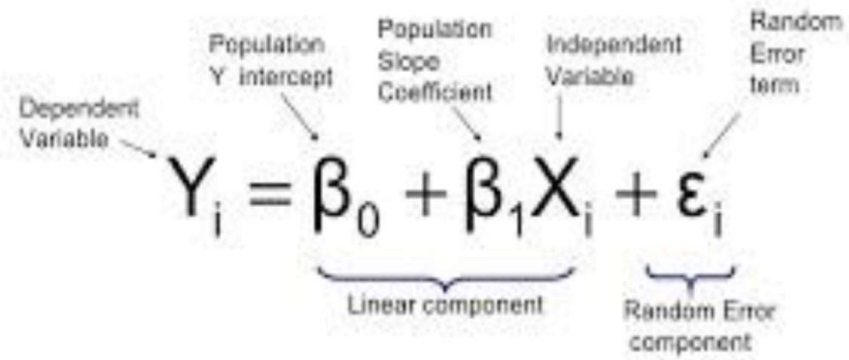
Linear Regression is a type of supervised Machine Learning algorithm that is used for the prediction of numeric values. Linear Regression is the most basic form of regression analysis. Regression is the most used predictive analysis model. Linear regression is based on the popular equation **"y = mx + 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 SLA is:



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

The equation for MLR is:

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

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

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

β1 = coefficient for X1 variable

β2 = coefficient for X2 variable

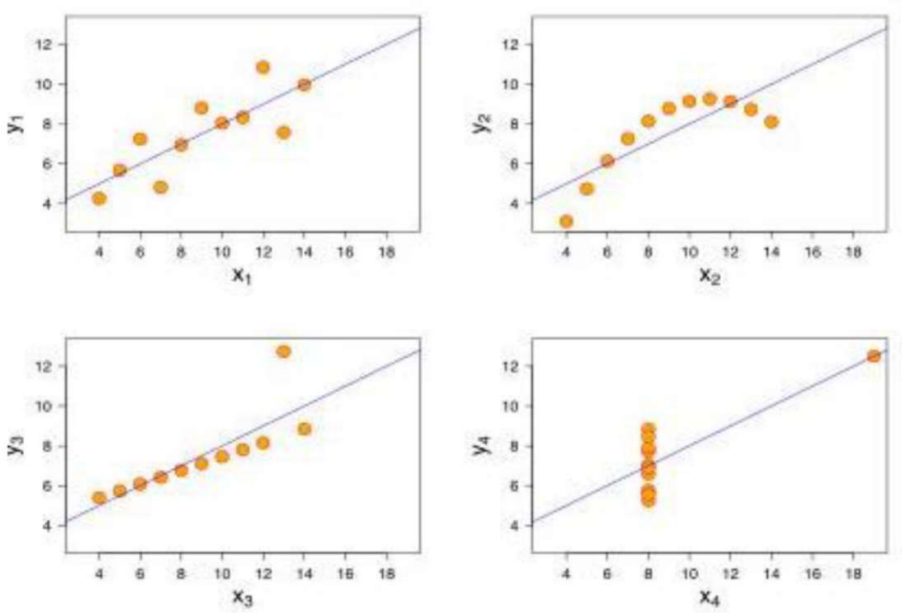
β3 = coefficient for X3 variable and so on…

β0 is the intercept (constant term).

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

Answer:

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 analysing 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.

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

Answer:

Pearson's r is a numerical summary of the strength of the linear association between the variables. It values ranges between -1 to +1. It shows the linear relationship between two sets of data.

In simple terms, it tells us can we draw a line graph to represent the data?

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

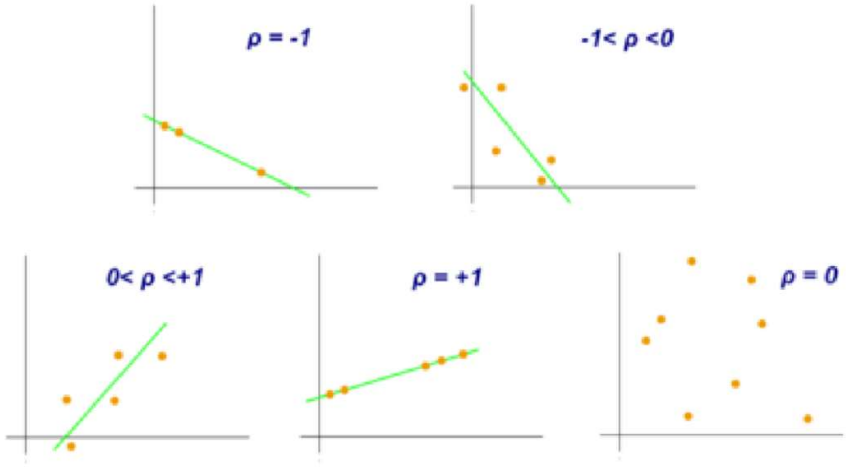
$\bar{y}$ = mean of the values of the y-variable

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)

Answer:

Feature scaling is a method used to normalize or standardize the range of independent variables or features of data. It is performed during the data pre-processing stage to deal with varying values in the dataset. 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, irrespective of the units of the values.

● **Normalization** is generally used when you know that the distribution of your data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbors and Neural Networks.

● **Standardization**, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalization, standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization.

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

(3 marks)

Answer:

The VIF gives how much the variance of the coefficient estimate is being inflated by collinearity. (VIF) =1/(1-R_1^2). 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. 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".

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

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)

Answer:

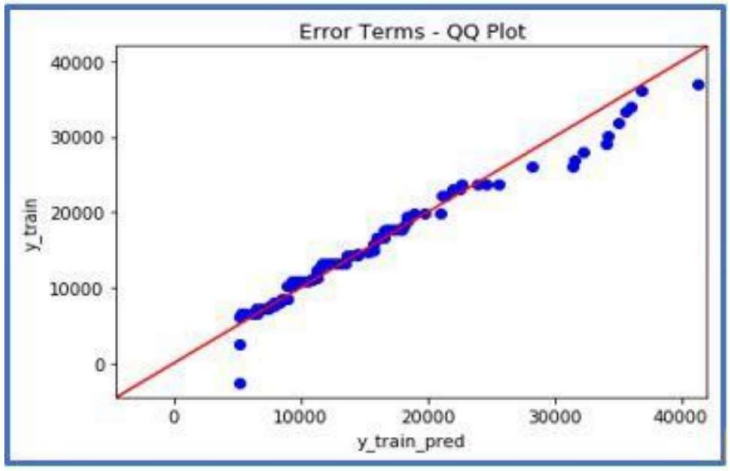
A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. It is used to compare the shapes of distributions. A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight. The q-q plot is used to answer the following questions:

- Do two data sets come from populations with a common distribution?
- Do two data sets have common location and scale?
- Do two data sets have similar distributional shapes?
- Do two data sets have similar tail behaviour?

Below are the possible interpretations for two data sets using a Q-Q plot:

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



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