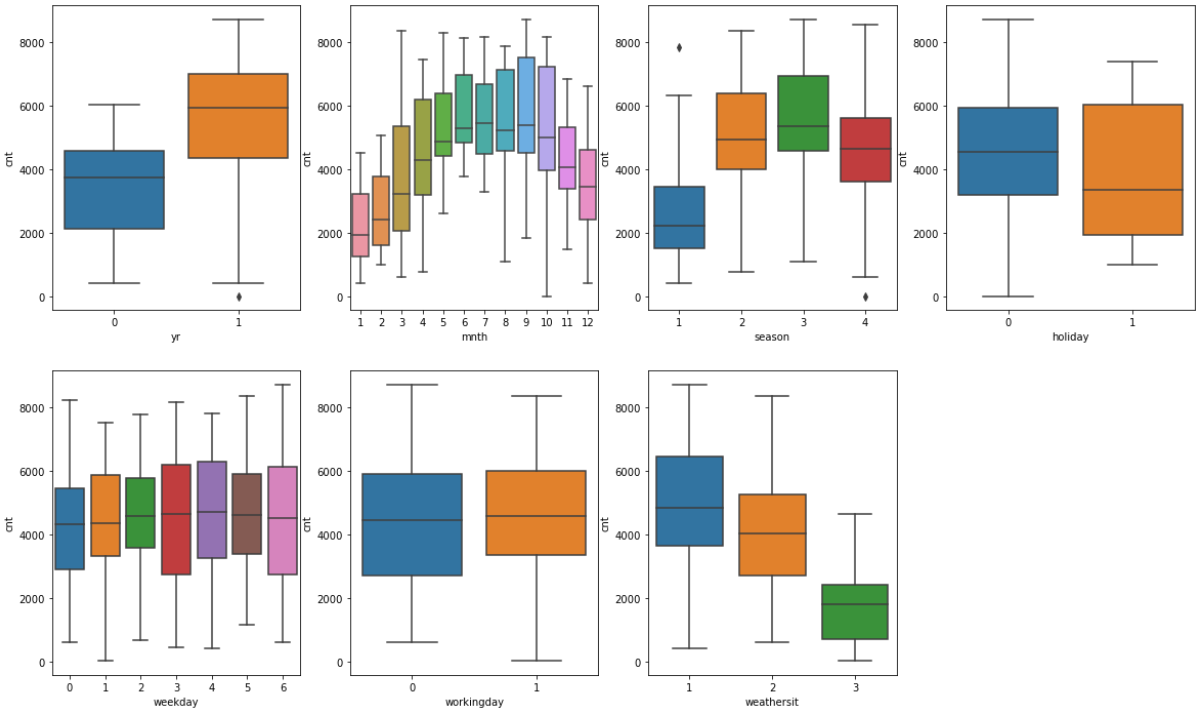
# Assignment-based Subjective

## Question 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: Looking at the Box plot of the categorical variable, below are the inferences




**Inferences**

Note: cnt is the dependent variable and is referred to as rentals in the below inferences

**Year**: Rentals increases as the year increases, I believe since the Bike sharing business was booming in 2018 2019, number of bikes rented in 2019 is surely higher than 2018

**Months**: Rentals was higher in May- September as compared to other months

**Seasons**: Rentals increased from Spring to Summer and went a little down in Winter

**Holiday**: Rentals reduced during Holidays

**Weekdays:** Rentals remained similar throughout the weekdays

**Working day:** Rentals increased on working day

**Weather situation:** Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog had 0 rentals stating this weather is extreme for bike, and highest rental was seen when the sky was Clear, Few clouds, Partly cloudy, Partly cloudy

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

If we decide not to exclude the initial column, the resulting dummy variables could end up being correlated or redundant. This might have adverse effects on certain models, particularly when dealing with smaller sets of unique values. For instance, iterative models could encounter challenges in achieving convergence, and lists detailing variable importance could become distorted.

So to say in more detail, below are the four factors

1) Multicollinearity
   When we encode a column we create different columns from one column, there do exists perfect collinearity among these columns and dropping one would break this perfect collinearity

2) Baseline Category
   One category of the variable is chosen as the baseline for reference category, the value of this baseline is implicitly captured
   For example if we split weekdays into dummies, we can drop first which is sumday and create 6 columns for each days, when its all zeros it would imply Sunday

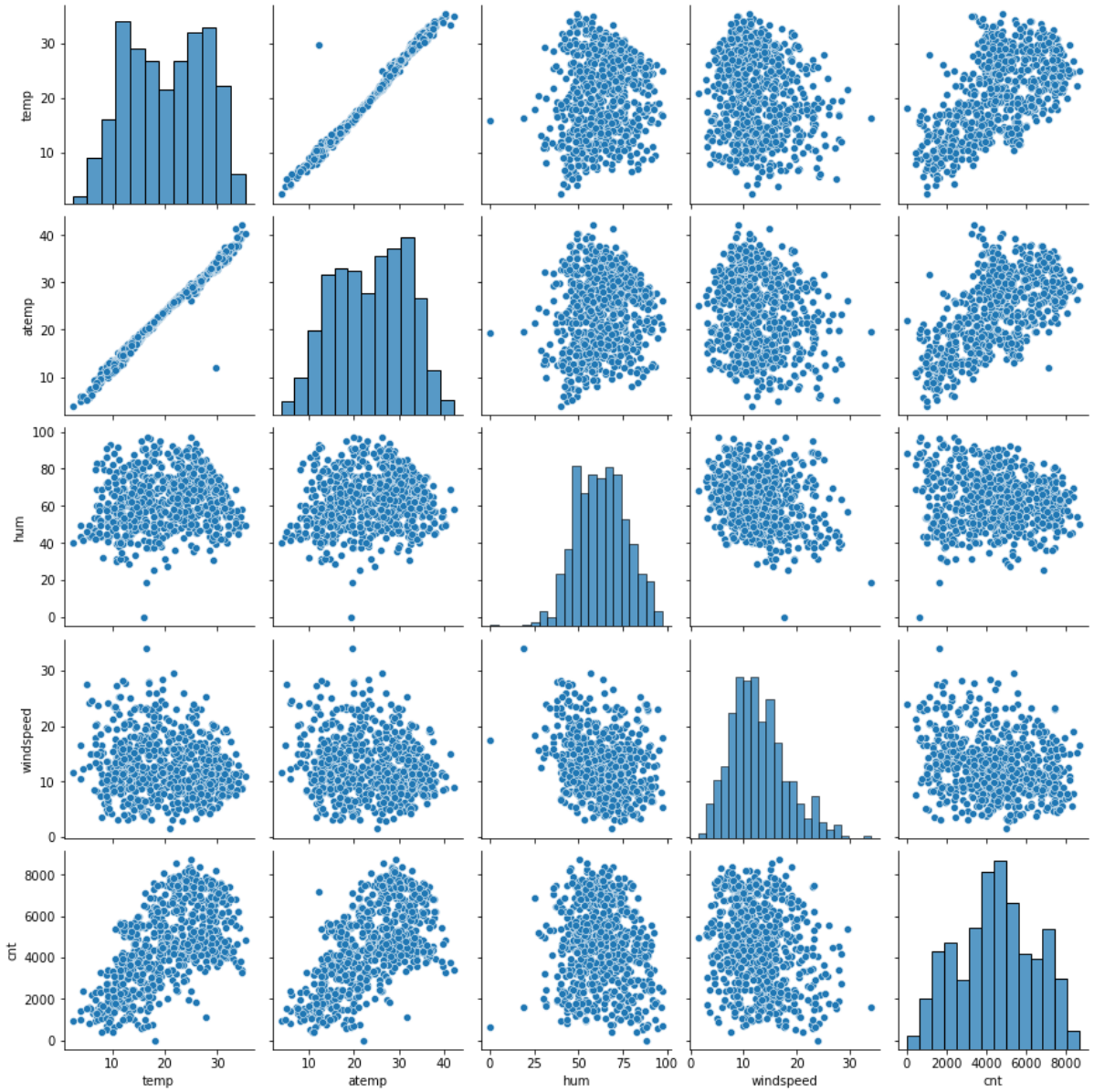
3) Interpretability

   It enhances interpretability of the model

4) Model efficiency
   Reducing columns means increased efficiency, less complex to train and evaluate

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

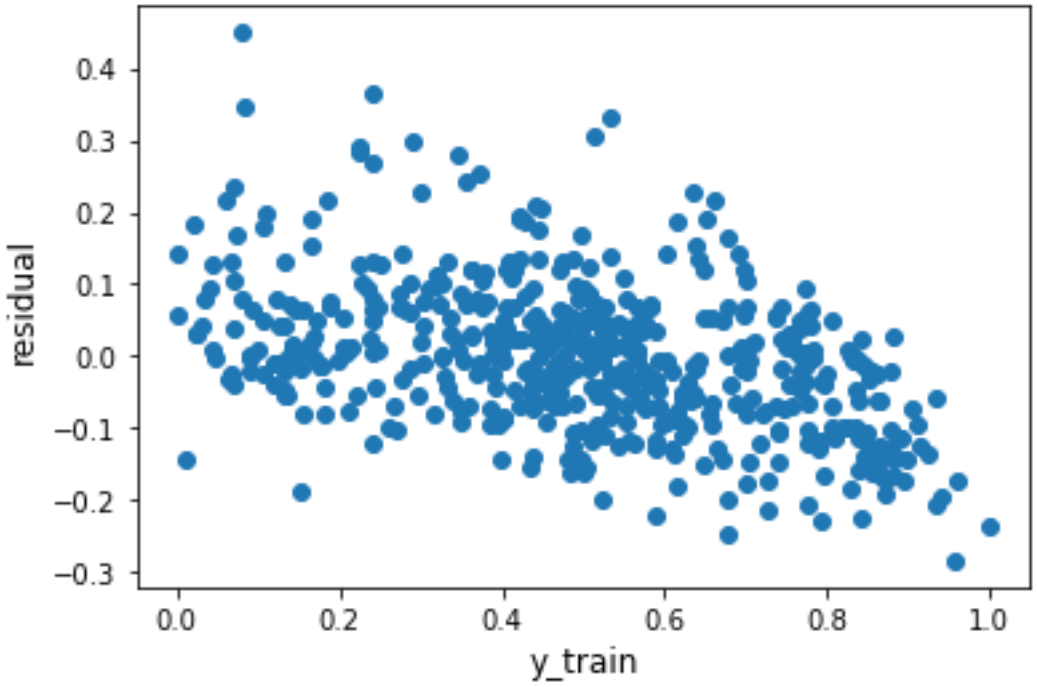
**Looking at the bellow pair plot graph numeric variables 'temp' and 'atemp' had highest corelation with target variable 'cnt'**



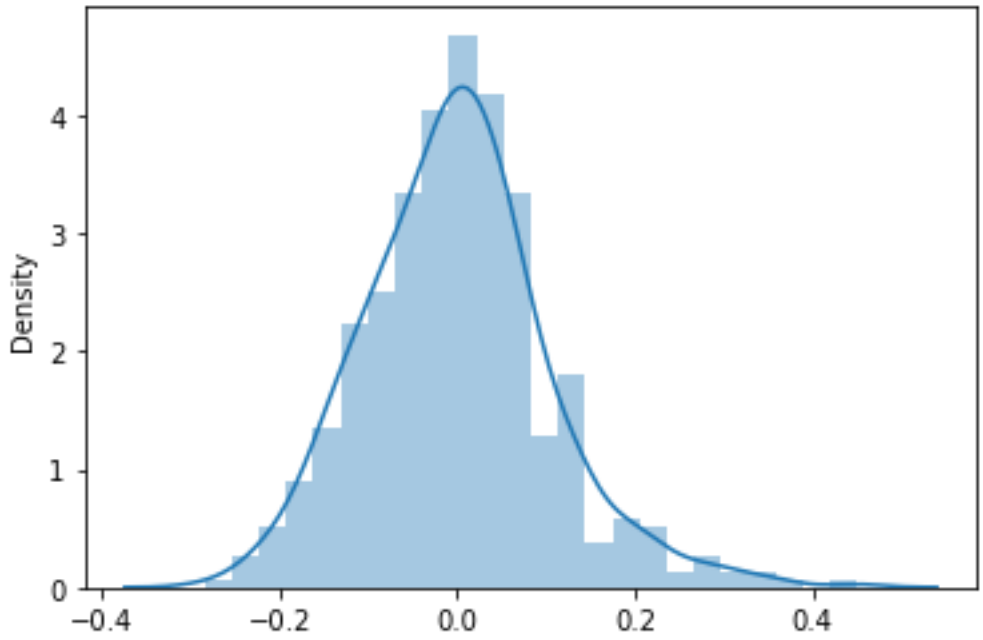
## Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Below tests were done to validate the assumptions of linear regression:

1. Linearity
   First, I did validate that there exists a linear relationship between one or more independent variable and dependent variable, we used pair-plot to visualize this (as explained in previous question)
2. Homoscedasticity
   Second I did validate that the variance of error terms are constant by plotting a residual plot




3. Thirdly, Residuals distribution should follow normal distribution and centred around 0 (mean = 0). I did validate this by plotting a distplot. The diagram below shows that the residuals are distributed about mean = 0.




4. Lastly, 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. I 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.

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

Top three features are

Temprature (0.657234)

weathersit : Light Snow, Light Rain + Mist & Cloudy (-0.257739)

year (0.238597)

## 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.

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

The equation for SLR will be: , where $b_o$ is the intercept, $b_1$ is coefficient or slope, x is the independent variable and y is the dependent variable.

$$y = b_o + b_1 x$$

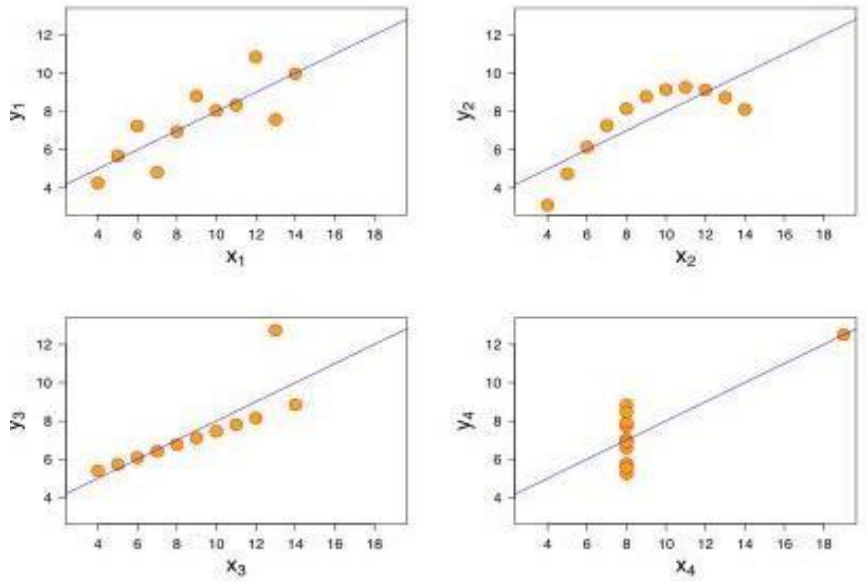
Multiple Linear Regression :MLR is used when the dependent variable is predicted using multiple independent variables. The equation for MLR will be:

where bo is the intercept, $b_1, b_2, b_3, b_4 \ldots, b_n$ are coefficients or slopes of the independent variables $x_1, x_2, x_3, x_4 \ldots, x_n$ and y is the dependent variable.

$$y = b_o + b_1 x_1 + b_2 x_2 + b_3 x_3 \ldots . + b_n x_n$$

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

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)

Pearson's r is a numerical summary of the strength of the linear association between the variables. It value 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

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

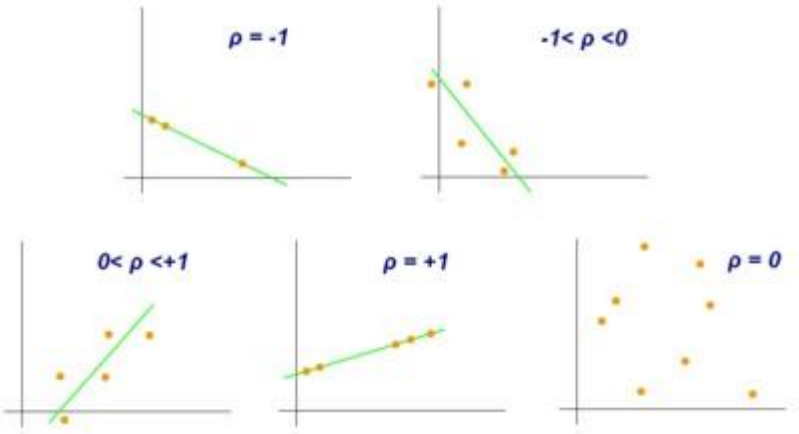
$r$ = correlation coefficient

$x_i$ = 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




As can be seen from the graph above

- 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)

**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 preprocessing 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)

VIF - Variance Inflation Factor

The VIF (Variance Inflation Factor) shows how much collinearity increases the uncertainty in our coefficient estimates. When variables are perfectly correlated, the VIF becomes infinite. It helps us understand how closely related our feature variables are. This is a key factor to consider when testing our linear model.

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

Where R-1 represents the R-square value of the specific independent variable under scrutiny, indicating how well this variable is influenced by other independent variables. If this variable is entirely explicable by other independent variables, it will exhibit perfect correlation and its R-squared value will equal 1. Consequently, VIF = 1/(1-1), yielding VIF = 1/0, resulting in a value of "infinity."

The numerical VIF value (in decimal form) indicates the degree by which the variance (i.e., the square of the standard error) is increased for each coefficient. For instance, a VIF of 1.9 signifies that the variance of a particular coefficient is 90% larger than expected when no multicollinearity — no correlation with other predictors — exists.

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)

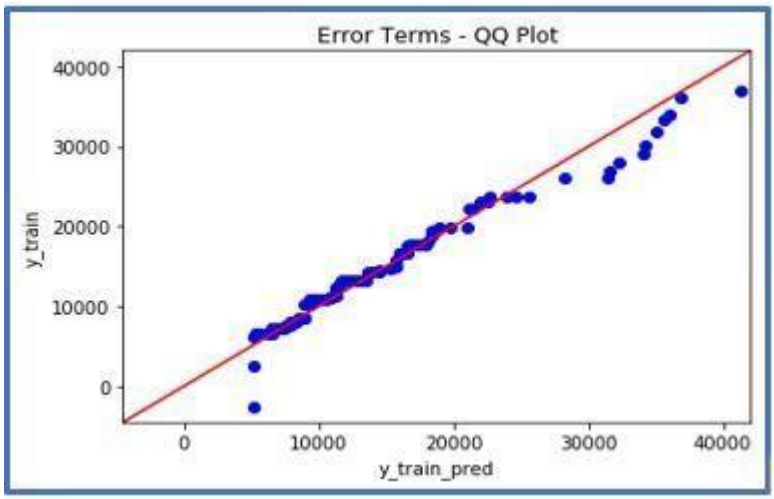
A quantile-quantile (q-q) plot presents the quantiles of one dataset against those of another dataset, facilitating a comparison of distribution shapes. This scatterplot pairs two sets of quantiles, aiming to reveal a linear alignment if both sets originate from the same distribution. The q-q plot serves to address the following inquiries:

- Do two datasets share a common underlying distribution?
- Do two datasets exhibit similarity in location and spread?
- Are the distribution shapes of two datasets comparable?

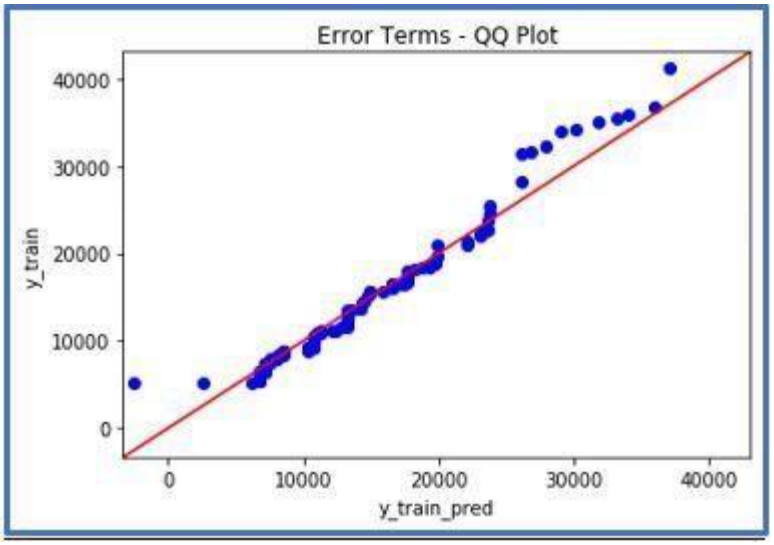
Do two datasets display akin behaviours in their tails?

The Q-Q plot allows for these interpretations of two datasets:

- Similar distribution: If quantile points align closely or directly along a 45-degree angle from the x-axis.
- Y-values < X-values: If the quantiles of the y-values are consistently lower than those of the x-values.




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


Error Terms - QQ Plot
40000
30000
20000
10000
0
y_train
0
10000
20000
30000
40000
y_train_pred