Linear_Regression_Subjective_Questions

Assignment-based Subjective Questions

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

Demand is high for next year 2019

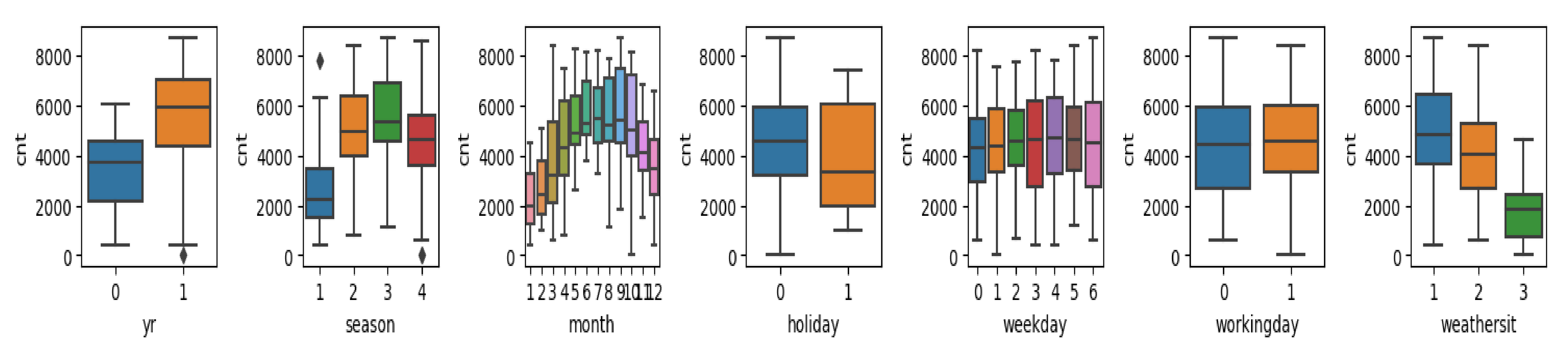
Season 3 has highest demand.

Month 5 to 10 has peak demand.

Year start and end has low demand and mid month has high demand

The clear weathershit highest booking

High booking on non holiday

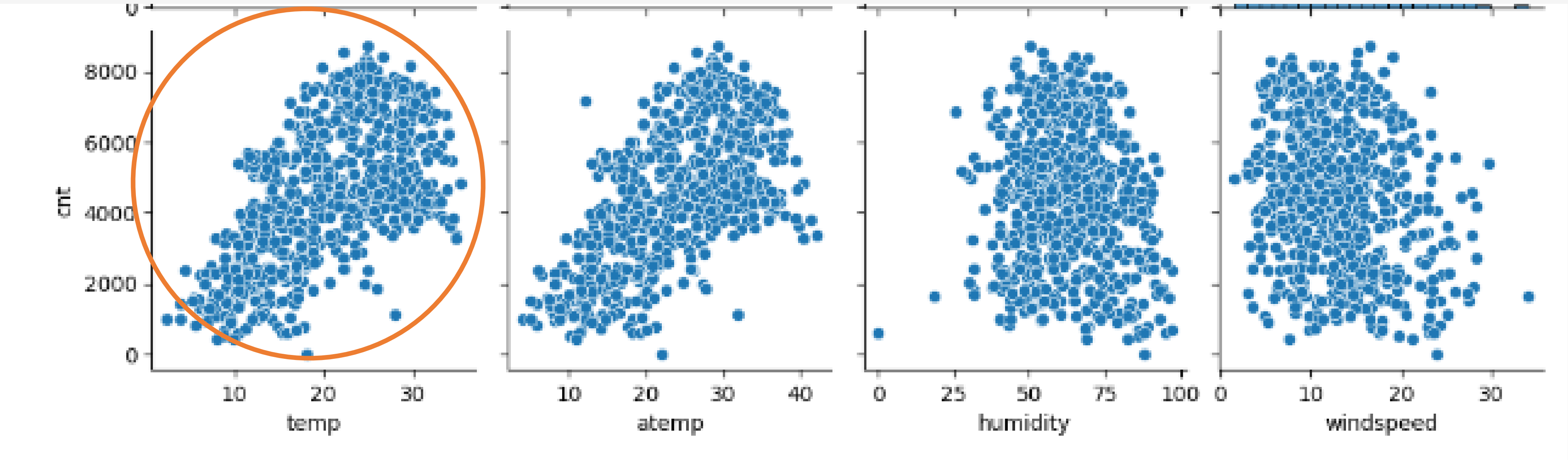


Why is it important to use drop_first=True during dummy variable creation?

Drop_first = True, this will drop the first dummy variable, thus it will give n-1 dummies out of n discrete categorical levels by removing the first level.
If we do not use **drop_first = True**, then n dummy variables will be created, and these predictors(n dummy variables) are themselves correlated which is known as multicollinearity and it, in turn, leads to Dummy Variable Trap

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

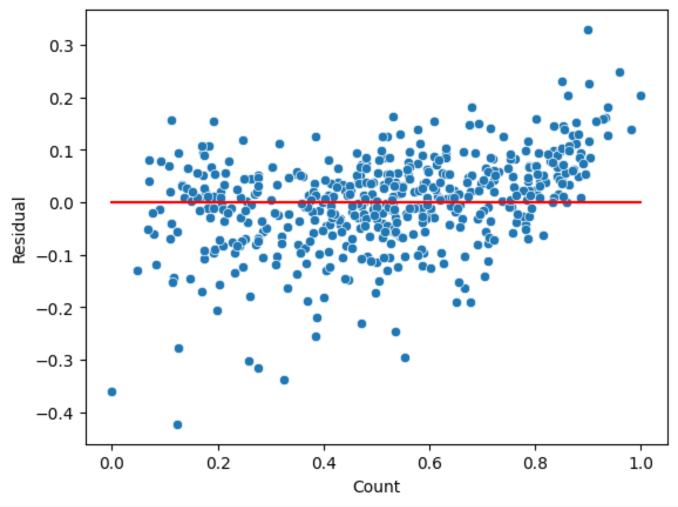
Cnt has strong correlation with **Temp**



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

Linearity: The relationship between X and Y is linear. Correlation matrix

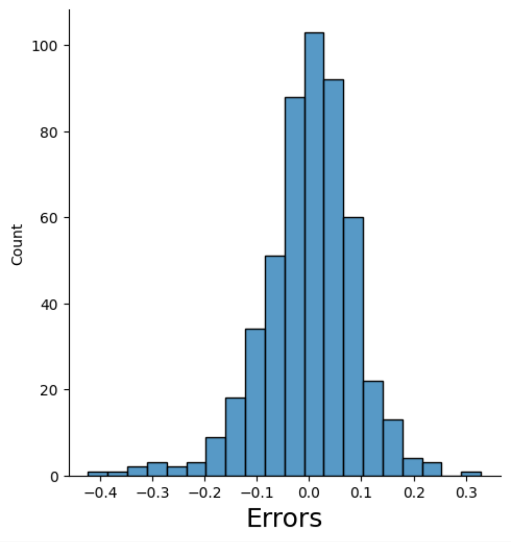
Homoscedasticity: The variance of residual is the same for any value of X.



Low Multi collinearity : Low multi co-relation check with VIF

Independence: Observations are independent of each other. Correlation matrix and VIF

Normality: For any fixed value of error distribution, Y_train - Y_train_predicted is normally distributed.



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

1. temp
2. Year
3. Light_snowrain

Explain the linear regression algorithm in detail.

Linear regression is a data analysis technique that predicts the value of unknown data by using another related and known data value. Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable

Types of Linear Regression

1. Simple Linear Regression:
   If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.

   - Y= β0*X + β1 + ε

   - β0 and β1 are two unknown constants representing the regression slope, whereas ε (epsilon) is the error term.

2. Multiple Linear regression:
   If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

Explain the linear regression algorithm in detail.

Multiple Linear regression:

If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

$$yi=\beta0+\beta1Xi1+\beta2Xi2+\cdots+\beta pXip+\varepsilon i,\quad i=1,\cdots,n,$$

As the number of predictor variables increases, the β constants also increase correspondingly.

Multiple linear regression models multiple variables and their impact on an outcome:

where

- n is the number of observations.
- yi is the ith response.
- βk is the kth coefficient, where β0 is the constant term in the model. Sometimes, design matrices might include information about the constant term. However, fitlm or stepwiselm by default includes a constant term in the model, so you must not enter a column of 1s into your design matrix X.
- Xij is the ith observation on the jth predictor variable, j = 1, ..., p.
- εi is the ith noise term, that is, random error.

If a model includes only one predictor variable (p = 1), then the model is called a simple linear regression model.
In general, a linear regression model can be a model of the form

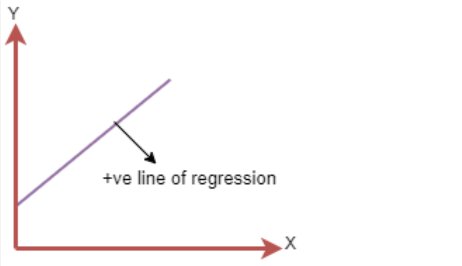
Explain the linear regression algorithm in detail.

Linear Regression Line

A linear line showing the relationship between the dependent and independent variables is called a regression line. A regression line can show two types of relationship:

Positive Linear Relationship:

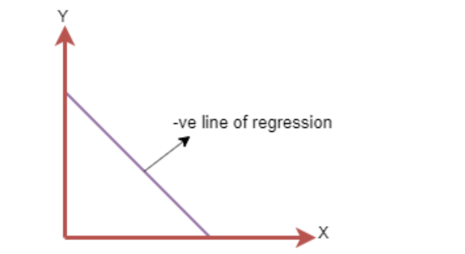
If the dependent variable increases on the Y-axis and independent variable increases on X-axis, then such a relationship is termed as a Positive linear relationship.



The line equation will be: $Y= a_0+a_1x$

Negative Linear Relationship:

If the dependent variable decreases on the Y-axis and independent variable increases on the X-axis, then such a relationship is called a negative linear relationship.

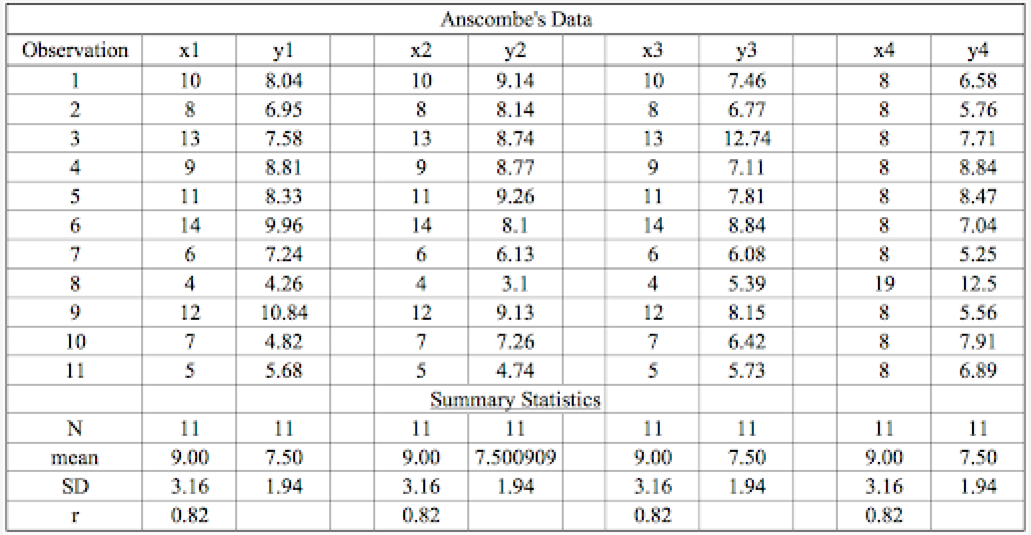


The line of equation will be: $Y= -a_0+a_1x$

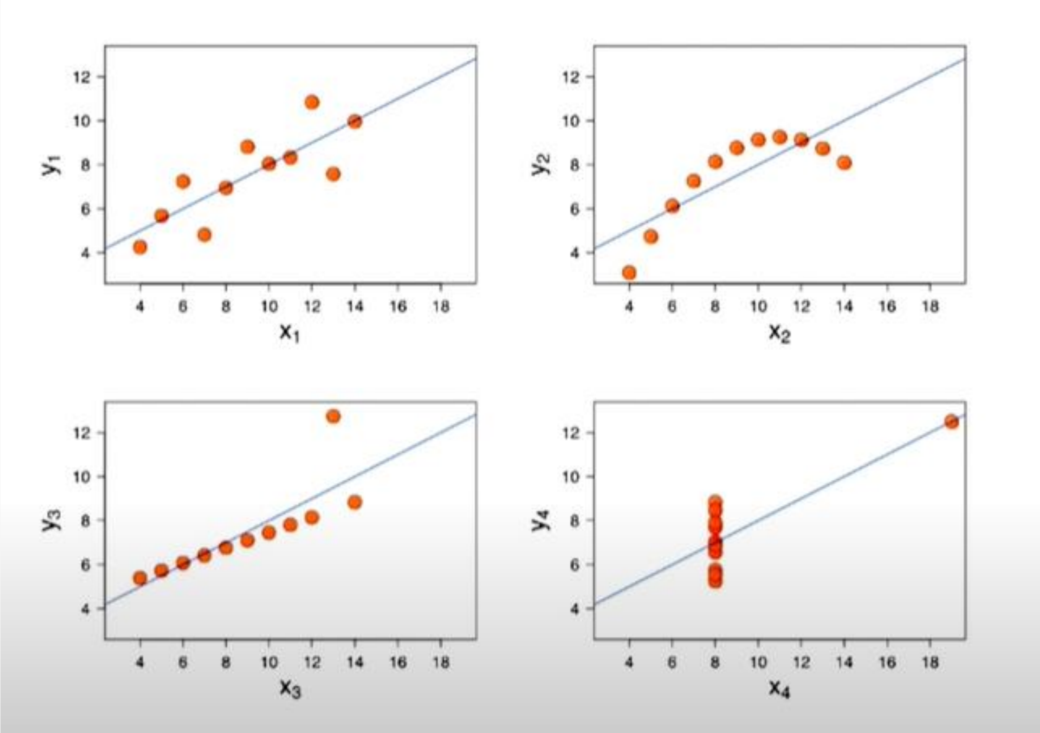
Explain the linear regression algorithm in detail.

Anscombe’s quartet was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting data before you analyze it and build your model. These four data sets have nearly the same statistical observations, which provide the same information (involving variance and mean) for each x and y point in all four data sets. However, when you plot these data sets, they look very different from one another.

Anscombe’s quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

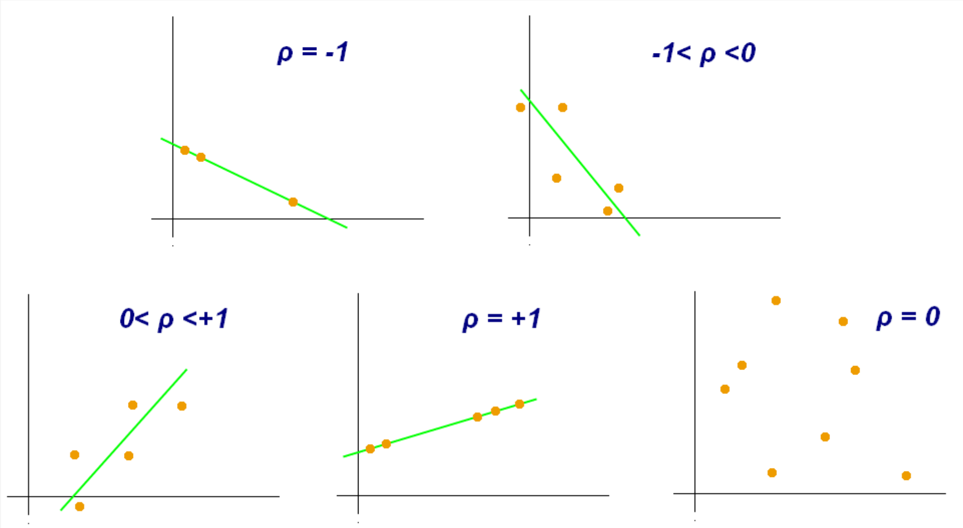


Anscombe's Data											
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		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	



What is Pearson’s R?

Pearson’s R is a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations; thus, it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1



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

What : 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.

Why? 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.

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).

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

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

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

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

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.

Interpretation:

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

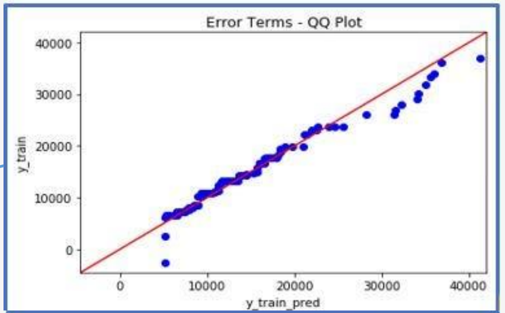
Below are the possible interpretations for two data sets.

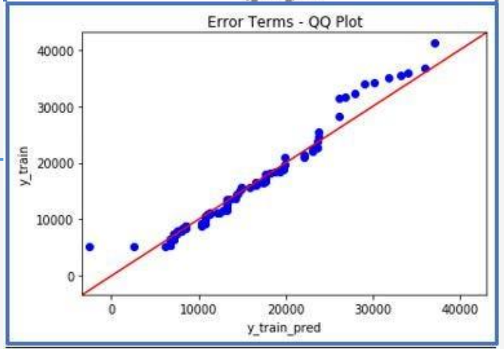
*a) **Similar distribution**: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis*

*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*


Error Terms - QQ Plot
y_train
y_train_pred


Error Terms - QQ Plot
y_train
y_train_pred