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 dependent variable (Count) is related and depends on several other factors termed as the Independent Variables. From the model built, it can be seen that few variables like Temperature, Year, Season etc has positive coefficients which in turn reflects that for a unit increase in those variables, the dependent variable (Count) will also increase by the corresponding factor, i.e. the dependent variable is directly proportional to the dependent variables like Temperature, Year, Season.
On the other hand, independent variables like Windspeed, holiday, Winter season (Dec, Jan, Feb), Snow etc have negative coefficients and this it implies that the dependent variable 'Count' is inversely correlated with these variables.

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

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

Let's say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not furnished and semi furnished, then It is obvious unfurnished. So, we do not need 3rd variable to identify the unfurnished.

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

The 'Registered' variable has the highest corelation (corr. = 0.95) with the target variable ('Count'). The scatter plot reveals that the shape is almost close to a straight line thus giving us the insight that Count of the people using the bike rental is completely related to the number of people registered to the service.

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

The assumptions are validated in the following steps:

- Before building the model, the Target Variable is plotted against other dependent variables to check if the relation is close to linear and any linear regression model can be fitted to the dataset. The assumption that the relation of the dependent and independent variable should be somewhat linear for a linear regression model to be used.
- After building the model and predicting the Target Variable using the model, the error values are calculated and the same is plotted. The error values are seen to be normally distributed with the mean = 0. This helps us ensure the assumption that the error or residual terms should be normally distributed.
- The Predicted values of the Target Variable from the test dataset is plotted against the actual values of the Target variable from the Test Data to notice any irregularity in the variance of the datapoints. The scatter plot shows that the variance is constant and thus ensures homoscedasticity which is again one of the assumptions of Linear regression.
- Before building the model, we see the relationships between various variables in the dataset to ensure and eliminate any multicollinearity among the variables. The same is ensured by finding out the VIF of the all the dependent variables used in the final model and the VIF value is checked. As per industry standard, anything less than 5 as VIF value is acceptable and ensures that there is no significant multicollinearity among the variables that may affect the model.

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

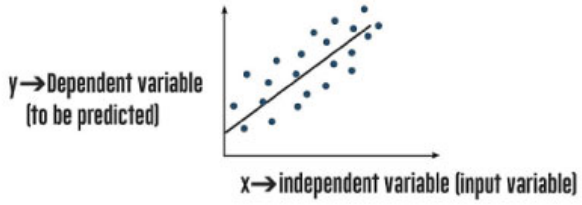
The model developed for the problem very efficiently takes care of the dependent variables responsible for the increase or decrease in demand of the bikes rented (target variable = ‘Count’). Based on the inference of the fitted model, the below 3 variables explains the demand of the shared bikes:

- Season: The company should focus on expanding its business in the Summer and the Fall season.
- Weather: The users prefer to rent a bike when the weather is pleasant i.e. either clear or cloudy and is comparatively less windy
- Temp: The users prefer to ride or rent a bike in a moderate temperature and not on the freezing days.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Majority of the machine learning algorithms fall under the supervised learning category. It is the process where an algorithm is used to predict a result based on the previously entered values and the results generated from them. Suppose we have an input variable 'x' and an output variable 'y' where y is a function of x (y=f{x}). Supervised learning reads the value of entered variable 'x' and the resulting variable 'y' so that it can use those results to later predict a highly accurate output data of 'y' from the entered value of 'x'. A regression problem is when the resulting variable contains a real or a continuous value. It tries to draw the line of best fit from the data gathered from a number of points.



What is Linear Regression?

Let's say we have a dataset which contains information about the relationship between 'number of hours studied' and 'marks obtained'. A number of students have been observed and their hours of study along with their grades are recorded. This will be our training data. Our goal is to design a model that can predict the marks if number of hours studied is provided. Using the training data, a regression line is obtained which will give minimum error. This linear equation is then used to apply for a new data. That is, if we give the number of hours studied by a student as an input, our model should be able to predict their mark with minimum error.

Hypothesis of Linear Regression

The linear regression model can be represented by the following equation:

$$Y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + ... + \theta_n x_n$$

where,

Y is the predicted value

θ_0 is the bias term.

$\theta_1,\ldots,\theta n$ are the model parameters

$x_1, x_2,\ldots,xn$ are the feature values.

The above hypothesis can also be represented by

$$Y = \theta^T x$$

Where, θ is the model's parameter vector including the bias term θ_0; x is the feature vector with $x_0 = 1$

Y (pred) = $b_0 + b_1 * x$

The values b_0 and b_1 must be chosen so that the error is minimum. If sum of squared error is taken as a metric to evaluate the model, then the goal is to obtain a line that best reduces the error.

$$\text{Error} = \sum_{i=1}^{n} (actual\_output - predicted\_output) ** 2$$

If we don't square the error, then the positive and negative points will cancel each other out.

For a model with one predictor,

$$b_0 = \bar{y} - b_1 \bar{x}$$

$$b_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Exploring 'b_1'

If $b_1 > 0$, then x (predictor) and y(target) have a positive relationship. That is an increase in x will increase y.

If $b_1 < 0$, then x (predictor) and y(target) have a negative relationship. That is an increase in x will decrease y.

Exploring 'b_0'

If the model does not include x=0, then the prediction will become meaningless with only b_0. For example, we have a dataset that relates height(x) and weight(y). Taking x=0 (that is height as 0), will make the equation have only b_0 value which is completely meaningless as in real-time height and weight can never be zero. This resulted due to considering the model values beyond its scope.

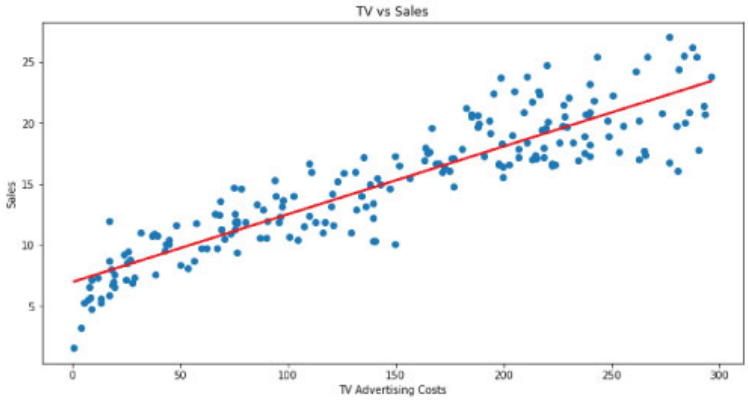
If the model includes value 0, then 'b_0' will be the average of all predicted values when x=0. But, setting zero for all the predictor variables is often impossible.

The value of b_0 guarantees that the residual will have mean zero. If there is no 'b_0' term, then the regression will be forced to pass over the origin. Both the regression coefficient and prediction will be biased.

How does Linear Regression work?

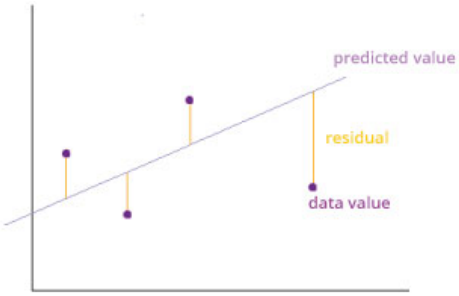
Let's look at a scenario where linear regression might be useful: losing weight. Let us consider that there's a connection between how many calories you take in and how much you weigh; regression analysis can help you understand that connection. Regression analysis will provide you with a relation which can be visualized into a graph in order to make predictions about your data. For example, if you've been putting on weight over the last few years, it can predict how much you'll weigh in the next ten years if you continue to consume the same amount of calories and burn them at the same rate.

The goal of regression analysis is to create a trend line based on the data you have gathered. This then allows you to determine whether other factors apart from the amount of calories consumed affect your weight, such as the number of hours you sleep, work pressure, level of stress, type of exercises you do etc. Before taking into account, we need to look at these factors and attributes and determine whether there is a correlation between them. Linear Regression can then be used to draw a trend line which can then be used to confirm or deny the relationship between attributes. If the test is done over a long time duration, extensive data can be collected and the result can be evaluated more accurately. By the end of this article we will build a model which looks like the below picture i.e, determine a line which best fits the data.



How do we determine the best fit line?

The best fit line is considered to be the line for which the error between the predicted values and the observed values is minimum. It is also called the ***regression line*** and the errors are also known as ***residuals***. The figure shown below shows the residuals. It can be visualized by the vertical lines from the observed data value to the regression line.



When to use Linear Regression?

Linear Regression's power lies in its simplicity, which means that it can be used to solve problems across various fields. At first, the data collected from the observations need to be collected and plotted along a line. If the difference between the predicted value and the result is almost the same, we can use linear regression for the problem.

Assumptions in linear regression

If you are planning to use linear regression for your problem then there are some assumptions you need to consider:

- The relation between the dependent and independent variables should be almost linear.

- The data is homoscedastic, meaning the variance between the results should not be too much.
- The results obtained from an observation should not be influenced by the results obtained from the previous observation.
- The residuals should be normally distributed. This assumption means that the probability density function of the residual values is normally distributed at each independent value.

You can determine whether your data meets these conditions by plotting it and then doing a bit of digging into its structure.

Few properties of Regression Line

Here are a few features a regression line has:

- Regression passes through the mean of independent variable (x) as well as mean of the dependent variable (y).
- Regression line minimizes the sum of "Square of Residuals". That's why the method of Linear Regression is known as "Ordinary Least Square (OLS)". We will discuss more in detail about Ordinary Least Square later on.
- B_1 explains the change in Y with a change in x by one unit. In other words, if we increase the value of 'x' it will result in a change in value of Y.

2. Explain the Anscombe's quartet in detail.

Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x,y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.

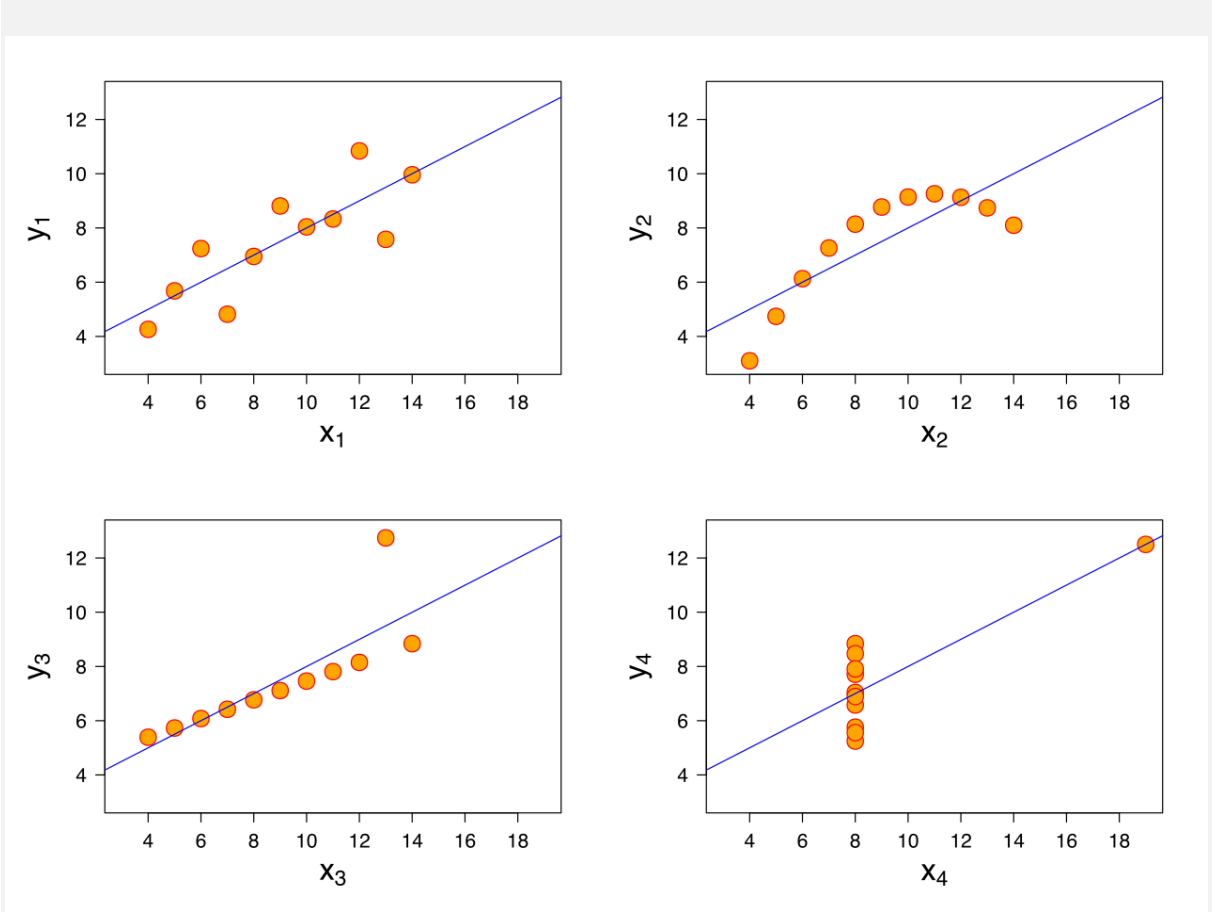
	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	8	5,56
	7	4,82	7	7,26	7	6,42	8	7,91
	5	5,68	5	4,74	5	5,73	8	6,89
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

Quartet's Summary Stats

The summary statistics show that the means and the variances were identical for x and y across the groups :

- Mean of x is 9 and mean of y is 7.50 for each dataset.
- Similarly, the variance of x is 11 and variance of y is 4.13 for each dataset
- The correlation coefficient (how strong a relationship is between two variables) between x and y is 0.816 for each dataset

When we plot these four datasets on an x/y coordinate plane, we can observe that they show the same regression lines as well but each dataset is telling a different story :



- Dataset I appears to have clean and well-fitting linear models.
- Dataset II is not distributed normally.
- In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.
- Dataset IV shows that one outlier is enough to produce a high correlation coefficient.

This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.

3. What is Pearson's R?

Pearson's product-moment correlation coefficient

This was introduced by Karl Pearson (1867- 1936). Pearson's correlation coefficient between two variables is defined as the covariance of the two variables divided by the product of their standard deviations

Karl Pearson's Coefficient of Correlation

Pearson's 'r' is the most common correlation coefficient.

Karl Pearson's Coefficient of Correlation denoted by- 'r' The coefficient of correlation 'r' measure the degree of linear relationship between two variables say x & y

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

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

Procedure for computing the correlation coefficient

- Calculate the sum of the two series 'Σx' & 'Σy'.
- Square each deviation of 'x' & 'y' then obtain the sum of the squared deviation i.e. ∑x 2 & ∑y2.
- Multiply each deviation under x with each deviation under y & obtain the product of 'xy'. Then obtain the sum of the product of x, y i.e. ∑xy.
- Substitute the value in the formula.

For Example: -

Concentration (X)	Absorbance (Y)	XY	X^2	Y^2
0	0	0.00	0.00	0.00
1	0.0778	0.0778	1	0.0060528
2	0.1543	0.3086	4	0.0232808
3	0.2286	0.6858	9	0.052258
4	0.3045	1.218	16	0.092720
5	0.3756	1.878	25	0.141075
ΣX= 15	ΣY= 1.1405	ΣXY= 4.1682	ΣX^2 =55	ΣY^2 =0.31538

Ans = 0.9995

Real Life Example

Pearson correlation is used in thousands of real life situations. For example, scientists in China wanted to know if there was a relationship between how weedy rice populations are different genetically. The goal was to find out the evolutionary potential of the rice. Pearson's correlation between the two groups was analysed. It showed a positive Pearson Product Moment correlation of between 0.783 and 0.895 for weedy rice populations. This figure is quite high, which suggested a fairly strong relationship.

Interpretation of Correlation Coefficient (r)

- The value of correlation coefficient 'r' ranges from -1 to +1.
- If r = +1, then the correlation between the two variables is said to be perfect and positive.
- If r = -1, then the correlation between the two variables is said to be perfect and negative.
- If r = 0, then there exists no correlation between the variables. 10

Limits of Correlation coefficient

- The correlation coefficient lies between -1 & +1 symbolically ($-1 \leq r \geq 1$).
- The correlation coefficient is independent of the change of origin & scale.

Value of r	Correlation
1.00	Perfect or ideal
0.90 -- 0.99	Excellent
0.80 -- 0.89	Very high correlation
0.60 -- 0.79	High correlation
0.40 -- 0.59	Medium correlation
0.20 -- 0.39	Low correlation
0.00 -- 0.19	Negligible correlation
0.00	No correlation

Merits

- Degree of correlation: Karl Pearson's methods gives us exact measure of degree of correlation between two variables.
- Direction of correlation: It provides the information whether the correlation is positive or negative.

Demerits

- Tedious calculations: Calculations of coefficient of correlation by this method is long, tedious and time consuming.
- Quantitative measurements: Pearson's correlation can be used only for those attributes which have quantitative measurements.

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

Scaling is the procedure of measuring and assigning the objects to the numbers according to the specified rules. In other words, the process of locating the measured objects on the continuum, a continuous sequence of numbers to which the objects are assigned is called as scaling.

Scale is important simply because the magnitude of the problems faced in areas such as poverty reduction, the environment, gender issues and healthcare require solutions at scale. By their nature they are often cross-border or not focused solely on one location.

Building programmes that are designed for scalability accordingly enhances the potential for impact. In areas such as the environment the nature of the problems we face injects an urgency that necessitates both scale and accelerated roll-out.

Creating common understanding about problems creates both broader awareness of the need to develop scalable solutions and engagement. Mindsets will need to be changed to develop strategies where people automatically think big. Building at scale also addresses corporations' need to ensure their investment is strategic and long-term and thus meets business goals.

Difference between Normalisation and Standardisation

S.NO.	NORMALISATION	STANDARDISATION
1	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4	It is really affected by outliers.	It is much less affected by outliers.
5	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.
6	This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.
7	It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
8	It is a often called as Scaling Normalization	It is a often called as Z-Score Normalization.

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

Variance Inflation Factors (VIFs) provide a one-number summary description of collinearity for each model term. Given an experiment with multiple factors, the variance inflation factor associated with the ith factor reflects the increase in the variance of the estimated coefficient for that factor compared to if the factors were orthogonal, and is defined as $VIFi$ = 1/1-R2i where R2i is the coefficient of determination of a regression model where the ith factor is treated as a response variable in the model with all of the other factors. $VIFi$ can range from one to infinity. Values equal to one imply orthogonality, while values greater than one indicate a degree of collinearity between the ith factor and one or more other factors. The square root of the VIF indicates how much larger the standard error is (and therefore, how much larger the confidence intervals will be), compared to a factor that is uncorrelated with the other factors. As a rule of thumb, values greater than 5 suggest that collinearity may be unduly influencing coefficient estimates. A variance inflation factor is calculated for each factor in the experiment. A shortcoming of relying solely on as a measure of merit is that it does not provide detailed correlation information between the the factor and other specific factors. For this, we must turn to the correlation coefficient matrix.

Keeping this in view, why is VIF infinite?

If there is perfect correlation, then **VIF** = **infinity**. A large value of **VIF** indicates that there is a correlation between the variables. If the **VIF** is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.

What does infinite VIF mean?

The user has to select the variables to be included by ticking off the corresponding check boxes. 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).

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

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.

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

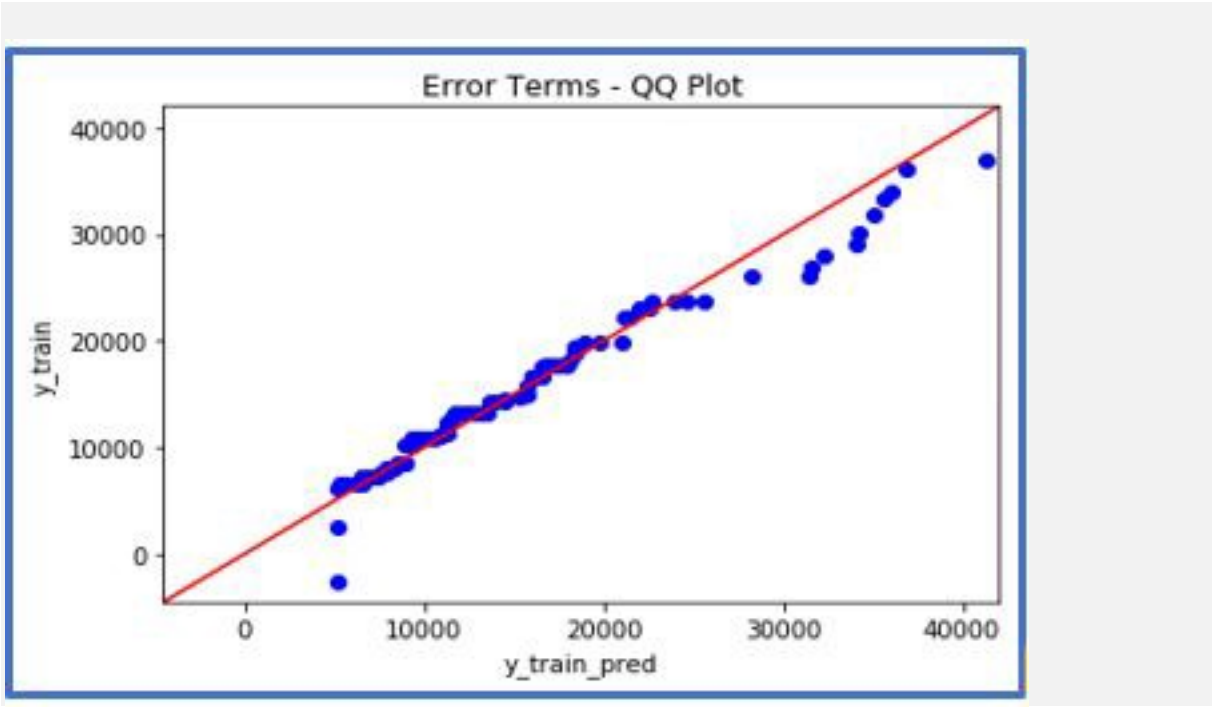
Interpretation:

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

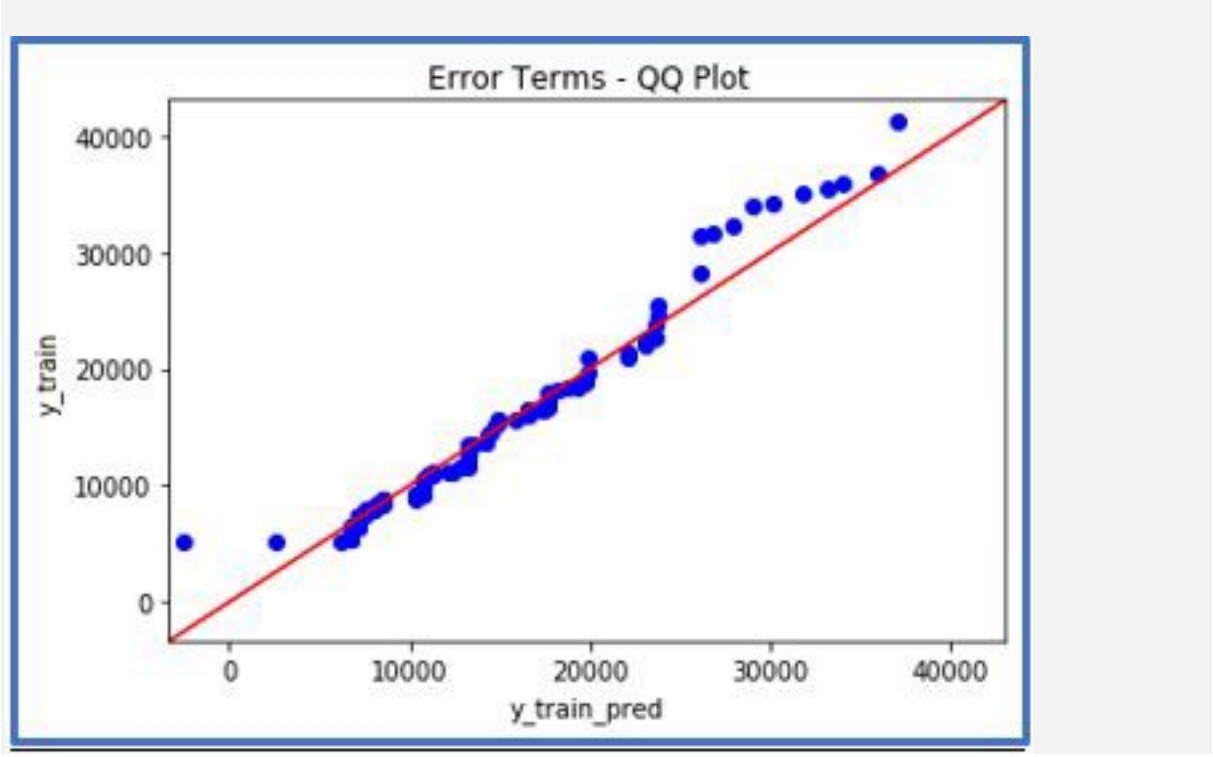
Below are the possible interpretations for two data sets.

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