Assignment-based Subjective Questions

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 : We have following categorical in our final model
1. Season : Winter help to increase the deamand while Spring impactively negatively
2. Month: In September demand increases while in Jan, July, Nov and Dec demand decreases
3. Year : Demand in 2018 is less than in 2019. Which is showing a increased demand. So we company should increase bikes stocks for next year so that customer demand of bikes can be fullfilled
4. Weathersit: Whether situation Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds **and** Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist decreasing my sales.

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

Answer : It helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

For Ex: We creating dummy variable for season column. Now as per data set we will have four types of dummy varriables(Columns in Data Set), 1. Spring, 2. Summer, 3. Winter and 4. Fall. Using these dummy variable we show season such as for Winter(0010) and Summer(0100). Each character denote columns. But we can show this status using three variables also such as

Spring 000
Summer 100
Winter 010
Fall 001

Thats why we do not required lets consider Summer column but can denote theri value.

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

Answer : After seeing the pair plot of all numeric variables(atemp, windspeed, hum, cnt), **atemp** have the highest correlation with **cnt(target variable).**

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

Answer : In last we can compare our final model with our initial pair plot and boxplot. We can see that **in pair plot atemp have highest correlation with target(cnt) variable**. When we see **boxplot season, yr, whethersit have a good correaltion with target(cnt) variable**.

Question 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 : Positively Significant: atemp(coffe: 0.4109) , yr_2019 (coffe: 0.2361), X Light Snow(coffe: -0.2881)

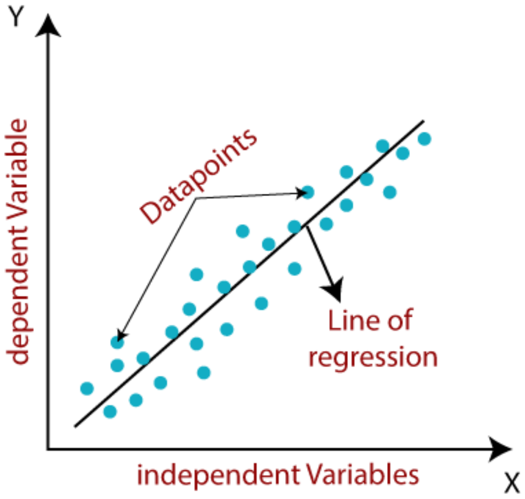
General Subjective Questions

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

Answer : Linear regression is one of the statistical method used for predictive analysis. Linear regression make predictions for categorical(continous/real) or numeric variables such **season, wethersituation, temprature, month etc.**

Linear regression algorithm shows a linear relationship between a dependent variable(y) and one or more dependent variables(X), hence called linear regression which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

We represent linerar regression using $y=B_0+B_1x+B_2x+.........+B_ix+ \varepsilon$, below is the referenced image to understand.



Here,

y= Dependent Variable (Target Variable)
X= Independent Variable (predictor Variable)
B_0= intercept of the line (Gives an additional degree of freedom)
B_i = Linear regression coefficient (scale factor to each input value).
ε = random error

The values for x and y variables are training datasets for Linear Regression model representation.

If there are only one indepent variable then its called **Simple Linear Regression(**$y=B_0+B_1x+ \varepsilon,$**)**, but independent variables are more than 1 then its called **Multiple Linear Regression(**$y=B_0+B_1x+B_2x+.........+B_ix+ \varepsilon,$**).**

When working with linear regression, our main goal is to find the best fit line that means errors between predicted and actual value should be minimized. Our best fit line should have least error. The best-fit line is obtained by minimising a quantity called **Residual Sum of Squares (RSS)**

RSS = $e1^2+e2^2+e3^2+e4^2+e5^2+..........++eN^2$

Where $eN=y_{actual}-Y_{predicted}$

$Y_{predicted}=B_0+B_1x$

$RSS=(y_1-B_0-B_1x)^2+(y_2-B_0-B_2x)^2+(y_3-B_0-B_3x)^2+..............+(y_N-B_0-B_Nx)^2$

Residuals: The distance between the actual value and predicted values is called residual. If the observed points are far from the regression line, then the residual will be high, and so cost function will high. If the scatter points are close to the regression line, then the residual will be small and hence the cost function.

Gradient Function: Gradient descent is used to minimize the MSE by calculating the gradient of the cost function. It is done by a random selection of values of coefficient and then iteratively update the values to reach the minimum cost function.

Model Performance: The process of finding the best model out of various models is called optimization. It can be achieved by below method:

1**. Prob(F-statistics):** It denotes the overall fitness of model. It should be less then 0.05(5 Significance level)

2. Adjusted R-Square: It denotes the goodness of fit of overall models. Its beasically calculated using R-Square using following formula 1/(1-R²) .
3. P-Value : P-Value of all independent variables should be less then 0.05. If its higher than drop that variable and do modeling again than check OLS summery again.
4. VIF(Variance Inflaction Factor): Check VIF value for all variables and for this assignment I take number which <=5 for good fit. VIF is basically show the multicollinearity between variables.

Assumptions of Linear Regression:

1. No multicollinearity between the features: Multicollinearity means high-correlation between the independent variables. Due to multicollinearity, it may difficult to find the true relationship between the predictors and target variables. It should not be >5 (or >10 in some business cases). If its greater than 5 for any variable then our model become overfit.

2. Homoscedasticity Assumption: Homoscedasticity is a situation when the error term is the same for all the values of independent variables. With homoscedasticity, there should be no clear pattern distribution of data in the scatter plot.

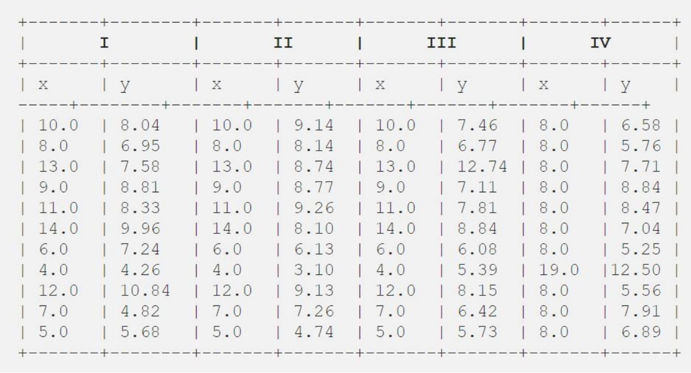
3. Normal distribution of error terms: If error terms are not normally distributed, then confidence intervals will become either too wide or too narrow, which may cause difficulties in finding coefficients.

4. No autocorrelations: The linear regression model assumes no autocorrelation in error terms. If there will be any correlation in the error term, then it will drastically reduce the accuracy of the model.

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

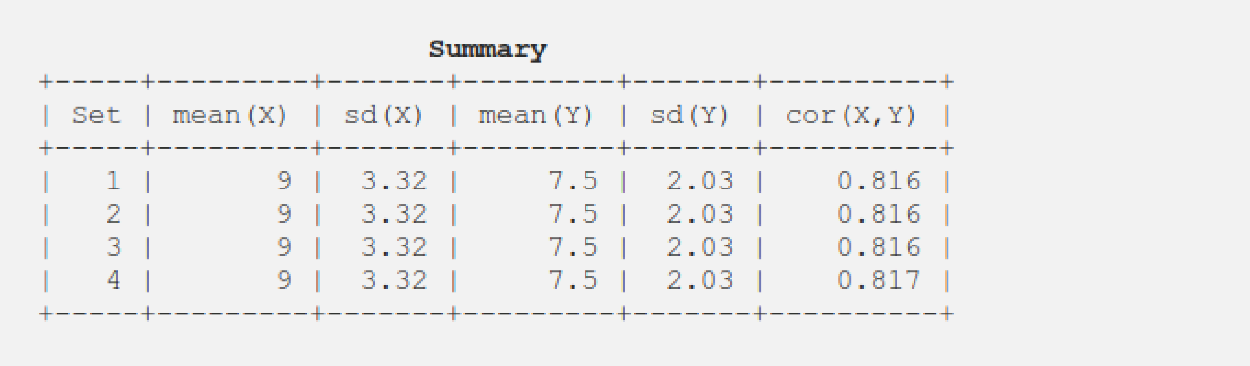
Answer: **Anscombe's quartet** a statistician of great repute always give a importance for visualizations with summary statistics. But People ignore that and show interest in only statistical information like mean, standar deviation etc.

So **Anscombe's quartet** comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.



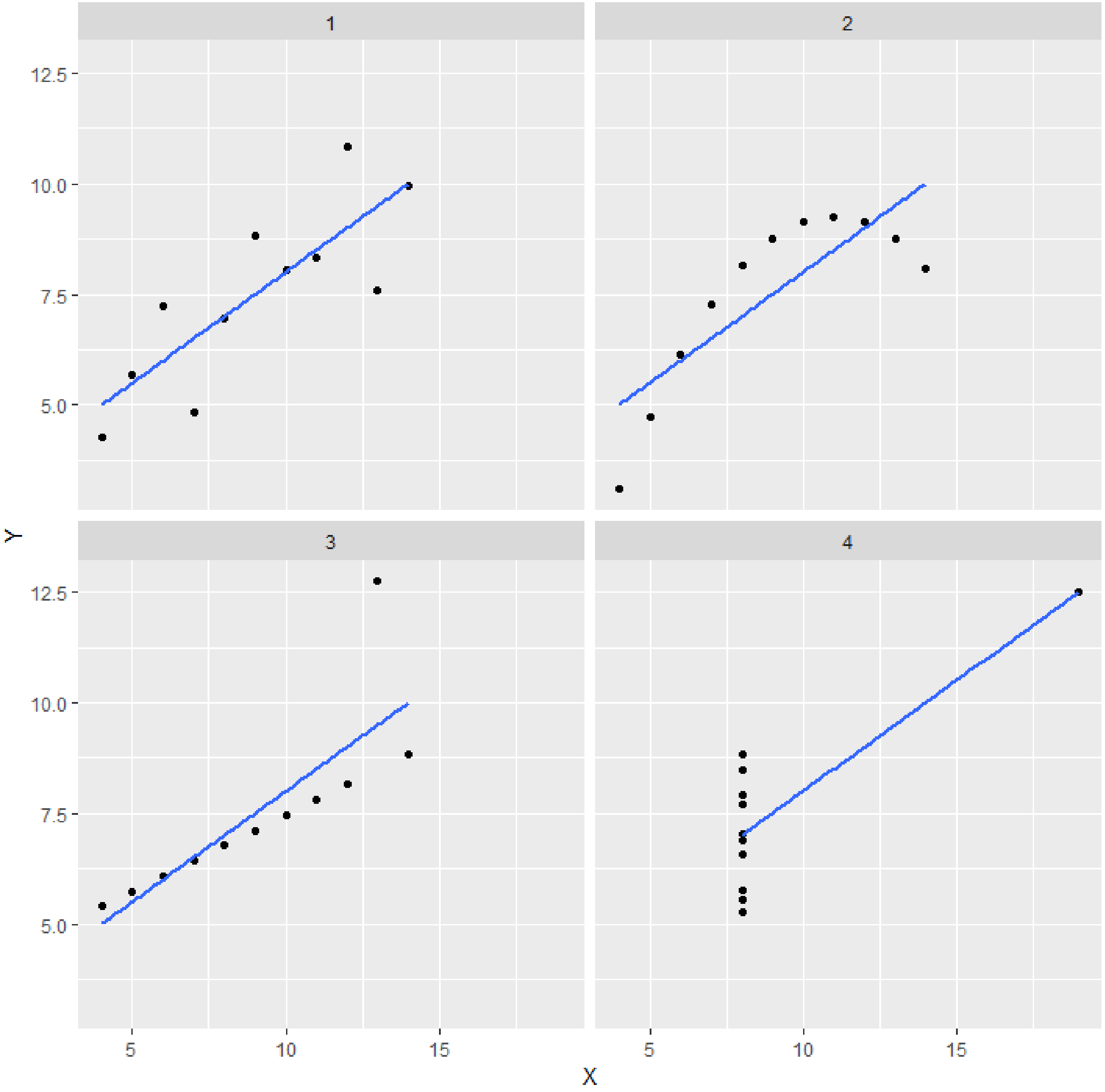
```
+-------+--------+-------+-------+-------+-------+-------+------+
|       I        |       II      |      III      |      IV      |
+-------+--------+-------+-------+-------+-------+-------+------+
| x     | y      | x     | y     | x     | y     | x     | y    |
-----+--------+-------+-------+-------+-------+-------+------+
| 10.0  | 8.04   | 10.0  | 9.14  | 10.0  | 7.46  | 8.0   | 6.58 |
| 8.0   | 6.95   | 8.0   | 8.14  | 8.0   | 6.77  | 8.0   | 5.76 |
| 13.0  | 7.58   | 13.0  | 8.74  | 13.0  | 12.74 | 8.0   | 7.71 |
| 9.0   | 8.81   | 9.0   | 8.77  | 9.0   | 7.11  | 8.0   | 8.84 |
| 11.0  | 8.33   | 11.0  | 9.26  | 11.0  | 7.81  | 8.0   | 8.47 |
| 14.0  | 9.96   | 14.0  | 8.10  | 14.0  | 8.84  | 8.0   | 7.04 |
| 6.0   | 7.24   | 6.0   | 6.13  | 6.0   | 6.08  | 8.0   | 5.25 |
| 4.0   | 4.26   | 4.0   | 3.10  | 4.0   | 5.39  | 19.0  |12.50 |
| 12.0  | 10.84  | 12.0  | 9.13  | 12.0  | 8.15  | 8.0   | 5.56 |
| 7.0   | 4.82   | 7.0   | 7.26  | 7.0   | 6.42  | 8.0   | 7.91 |
| 5.0   | 5.68   | 5.0   | 4.74  | 5.0   | 5.73  | 8.0   | 6.89 |
+-------+--------+-------+-------+-------+-------+-------+------+
```

Statistical Summery of above dataset



```
                         Summary
+-----+---------+--------+---------+-------+----------+
| Set | mean(X) | sd(X)  | mean(Y) | sd(Y) | cor(X,Y) |
+-----+---------+--------+---------+-------+----------+
|   1 |       9 |   3.32 |     7.5 |  2.03 |    0.816 |
|   2 |       9 |   3.32 |     7.5 |  2.03 |    0.816 |
|   3 |       9 |   3.32 |     7.5 |  2.03 |    0.816 |
|   4 |       9 |   3.32 |     7.5 |  2.03 |    0.817 |
+-----+---------+--------+---------+-------+----------+
```

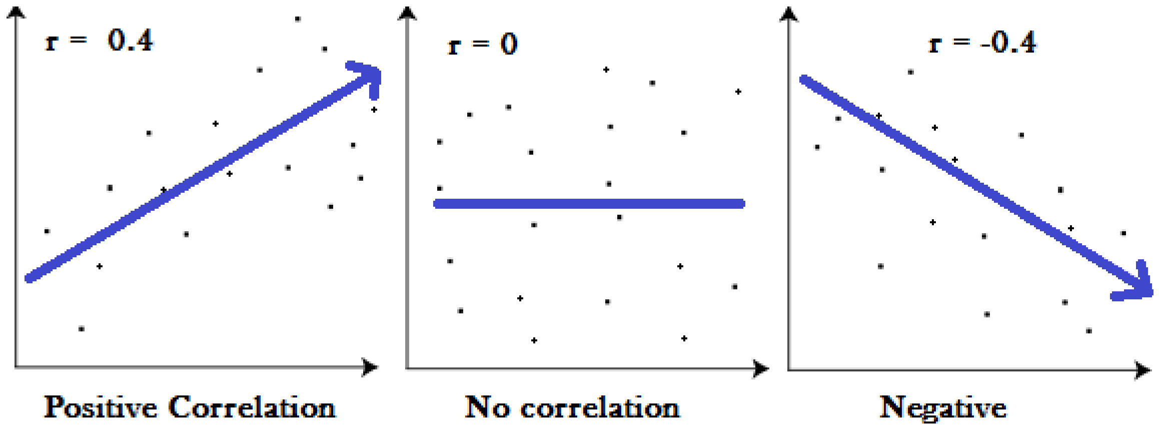
Grahpical Representation of data



Note: Graphs were completely different even though the summary was exactly similar

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

Answer : In statistics, the Pearson correlation coefficient also referred to as **Pearson's R**, or the **bivariate correlation**, is a statistic that measures linear correlation between two variables X and Y. It has a value between +1 and −1. A value of +1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation.



For a Population :

Pearson's correlation coefficient, when applied to a population, is commonly represented by the **Greek letter ρ (rho)** and may be referred to as the population correlation coefficient or the population Pearson correlation coefficient

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

where

cov is the covariance
sigma-x is the standard deviation of X
sigma-y is the standard deviation of Y

For a Sample:

Pearson's correlation coefficient, when applied to a sample, is commonly represented by $\mathbf{r_{xy}}$ and may be referred to as the **sample correlation coefficient**.

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

where

n is the sample size
x_i and y_i are the individual sample points indexed with i

$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the sample mean

Question 4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?(3 marks)

Answer : Scaling/DataScaling or 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. Sometimes, it also helps in speeding up the calculations in an algorithm.

Lets understand the problem with data and why we need Scaling.

We have a dataset in which my independent variables are Age and my Salary and one dependent variable Purchased.

Dataset = [Age:{30,27,29},Salary:{67000,80000,75000}].

This will cause some issues in our models since a lot of machine learning models such as k-means clustering and nearest neighbour classification are based on the Euclidean Distance.. If we have more than Crore range of data then my algorithm will also slow down.

If we scale these value between 0 to 1 or any other value depend on selected way. My data points are draw nearly and also memory consumption will also less.

We can do scalling using two algorithm Standarization and Max-Min

Standariztion: The result of standardization (or Z-score normalization) is that the features will be rescaled to ensure the mean and the standard deviation to be 0 and 1, respectively.

$$x_{\text{stand}} = \frac{x - \text{mean}(x)}{\text{standard deviation }(x)}$$

We use **from sklearn.preprocessing import StandardScaler** to do scaling

MinMax Normalization: This technique is to re-scales features with a distribution value between 0 and 1.. For every feature, the minimum value of that feature gets transformed into 0, and the maximum value gets transformed into 1.

$$x_{\mathrm{norm}} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

We use **from sklearn.preprocessing import MinMaxScaler** to do scaling

Question 5.. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
(3 marks)

Answer : In order to fit a multiple regression model, there are several assumptions that need to be made such as the model is linear, the residuals are normal, there is no multicollinearity between the independent variables, and the variance is constant.

Multicollinearity: Multicollinearity refers to the problem when the independent variables are collinear. Collinearity refers to a linear relationship between two explanatory variables. Two variables are perfectly collinear if there is an exact relationship between the two variables. If the independent variables are perfectly collinear, then our model becomes singular and it would not be possible to uniquely identify the model coefficients mathematically.

One way to address this issue is to check the correlation coefficient between the independent variables and if the correlation coefficient is high (either close to +1 or -1) then we conclude that the variables may be collinear, and we need to determine how best we can address this issue before we build the multiple regression model. If we have three inputs (X1, X2, and X3) then we would have to check the correlation coefficient between X1 and X2, between X1 and X3, and between X2 and X3. If any of these correlation coefficients are high, then we need to address this before building the multiple regression model.

However, correlation coefficient test works great if two variables are correlated but, in some cases, we can have a more complex situation where one independent variable may be related to more than two independent variables, so checking the variables two at a time does not catch the problem. For example, if we have X3 = 0.25X1 + 0.32X2. In this case, the matrix is still singular because of a relationship between X1, X2 and X3. To address these types of situations, we can use an index called VIF which will give an indication of multicollinearity.

What is VIF:
VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. In order to determine VIF, we fit a regression model between the independent variables.

$$y=B_0+B_1x+B_2x+.........+B_ix+ \varepsilon,$$

$$VIF=1/(1-R^2)$$

If all the independent variables are orthogonal to each other, then VIF = 1.0. If there is perfect correlation, then VIF = infinity.

A Genral rule of thumb we used that if VIF value is
<=5   Good
>5 and<10 Condierable on the basis of busniess decision
>10 Bad. Feature need to be dropped.

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

Answer : 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.

The advantages of the q-q plot are:

1. The sample sizes do not need to be equal.
2. 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:

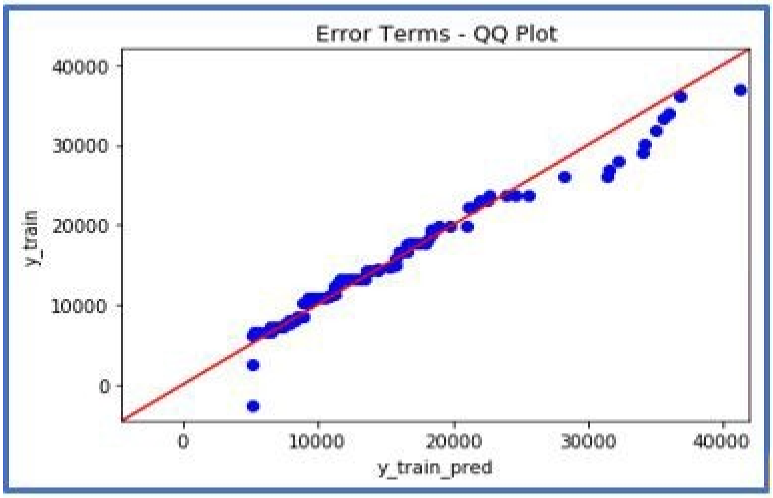
If two data sets —
i. come from populations with a common distribution
ii. have common location and scale
iii. have similar distributional shapes
iv. have similar tail behavior

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

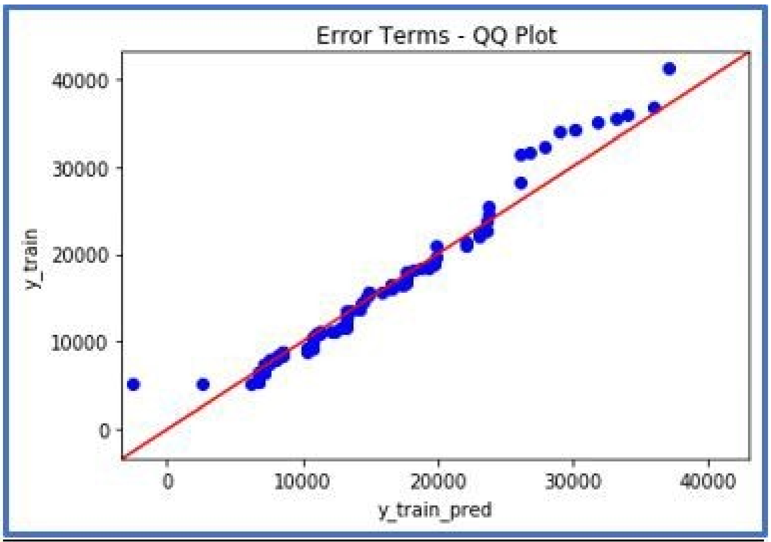
Below are the possible interpretations for two data sets.

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