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)

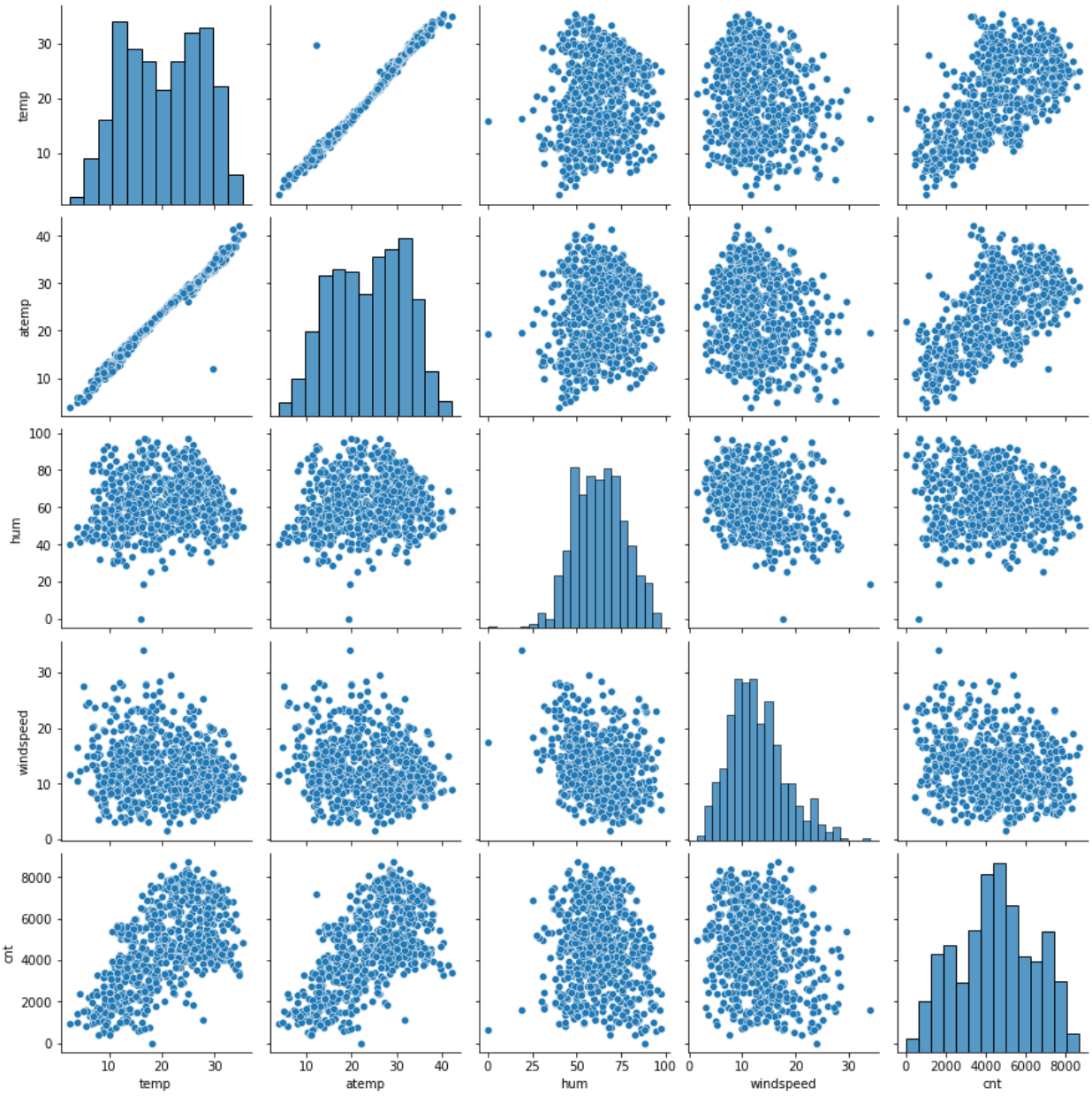
Some of the Categorical variables in datasets were : season, yr, mnth, holiday, weekday, workingday and weathersit. Following observations can be inferred with respect to dependent variable ,i.e., cnt from each categorical variables :

1. **season**: Demand of rental bikes in the season of spring were relatively less with compare to that in other seasons. Fall season saw the maximum demand of rental bikes, followed by summer and winter.
2. **yr**: In 2018, the mean number of rental bikes in a day was around 3800 whereas it was nearly 6000 in the year 2019. This represents an increase in business and also increase in customer size.
3. **mnth**: We can see that the peak in demand of rental bikes were in the middle months, i.e., from June till September. The demand keeps on steadily increasing from Jan till July and then till September it remains almost same and then towards the end of the year it decreases steadily.
4. **weathersit**: As expected, people do not use bikes in heavy rain/snow situation. They prefer to ride bikes mostly in clear and cloudy weather. Also, relatively less population uses rental bikes in light snow weather.
5. **holiday**: people uses rental bikes more on non-holidays than on holidays.

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

I think it is very important to use drop_first=True during dummy variable creating because of the following reasons:
a. Removes unnecessary extra columns which otherwise would occupy memory and make our model complex.
b. Removes collinearity between dummy variables, which is one of the most important things to be taken care while building a linear regression model.

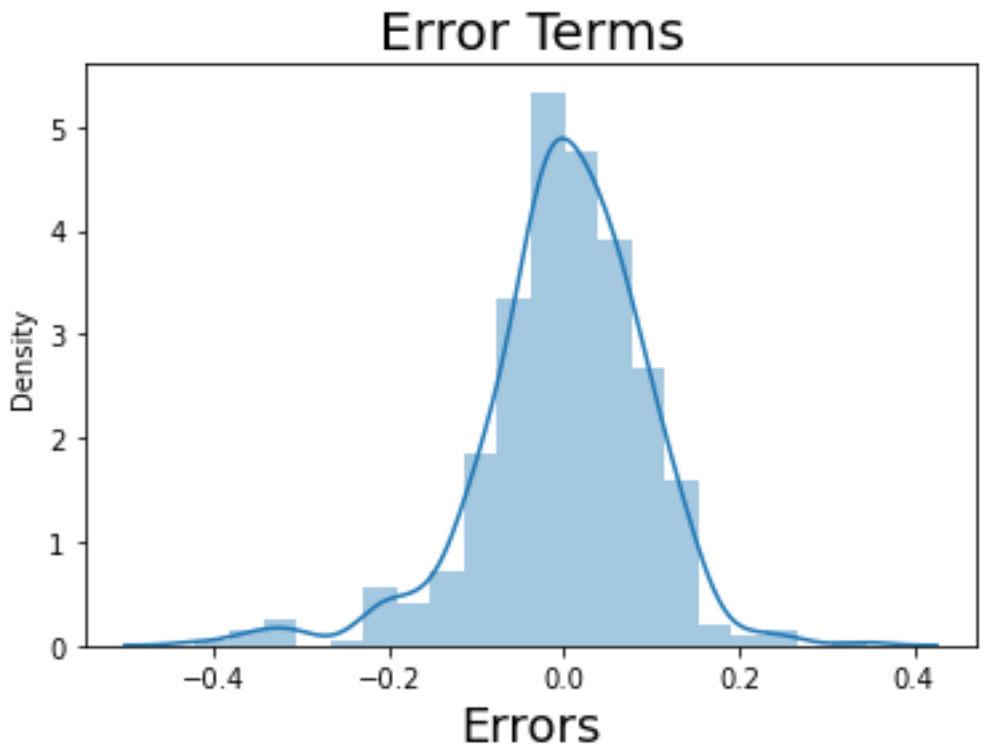
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)



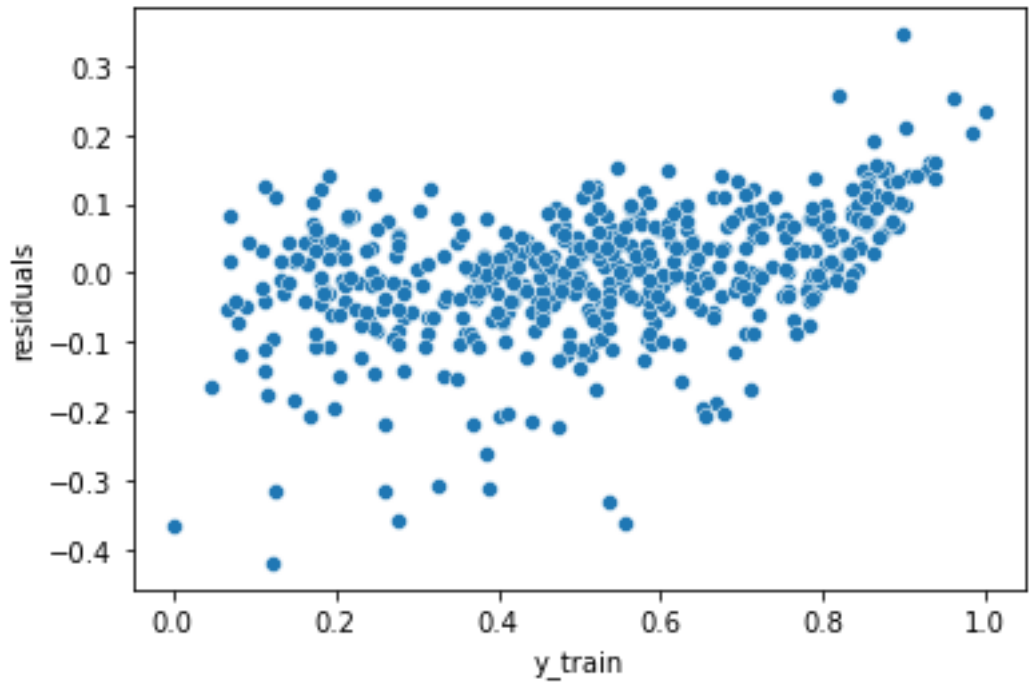
From the above pair plot we can observe that 'atemp' variable have the highest correlation with cnt (target variable).

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

To validate the assumptions of Linear Regression after the model was ready, I followed below approach:
1. To validate that residuals follow normal distribution with mean Zero, I plotted distribution plot of residuals.



2. To check if Observations are independent of each other and that the variance it constant across the residuals, I plotted the residuals/error terms with target variable, as shown below.



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

Following are the top 3 coefficients contributing significantly towards explaining the demand of shared bikes :

Features	Coefficients
atemp	0.4373
light snow	-0.2805
yr	0.2372

General Subjective Questions

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

- Linear regression is one of the very basic forms of machine learning where we train a model to predict a numeric target variable based on some predictor/independent variables.
- In linear regression as we can see the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.
- In regression, we calculate the best fit line which describes the relationship between the independent and dependent variable.
- Linear Regression is divided into simple linear regression and multiple linear regression.
  a) Simple Linear Regression: SLR is used when the dependent variable is predicted using only one independent variable.
  b) Multiple Linear Regression: MLR is used when the dependent variable is predicted using multiple independent variables.

Mathematically, we can write a linear regression equation as:

$$y = a + bx$$

Where a and b given by the formulas:

$$b\,(slope) = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

$$a\,(intercept) = \frac{n\sum y - b(\sum x)}{n}$$

Cost function that we try to minimize is:

$$minimize\frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2$$

Model Performance is calculated using :

$$R^2 = \frac{TSS - RSS}{TSS}$$

$$TSS = \textbf{Total Sum of Squares} = \sum_{i=1}^{N}(Y_i - \bar{Y})^2$$

$$ESS = \textbf{Explained Sum of Squares} = \sum_{i=1}^{N}(\hat{Y}_i - \bar{Y})^2$$

Equation of Multiple Linear Regression :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ ... + \beta_i X_i$$

Y : Dependent variable
β_0 : Intercept
β_i : Slope for X_i
X = Independent variable

2. Explain Anscombe's quartet in detail.

- Anscombe's quartet consist of 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 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.
- Those 4 sets of 11 data-points are given below:

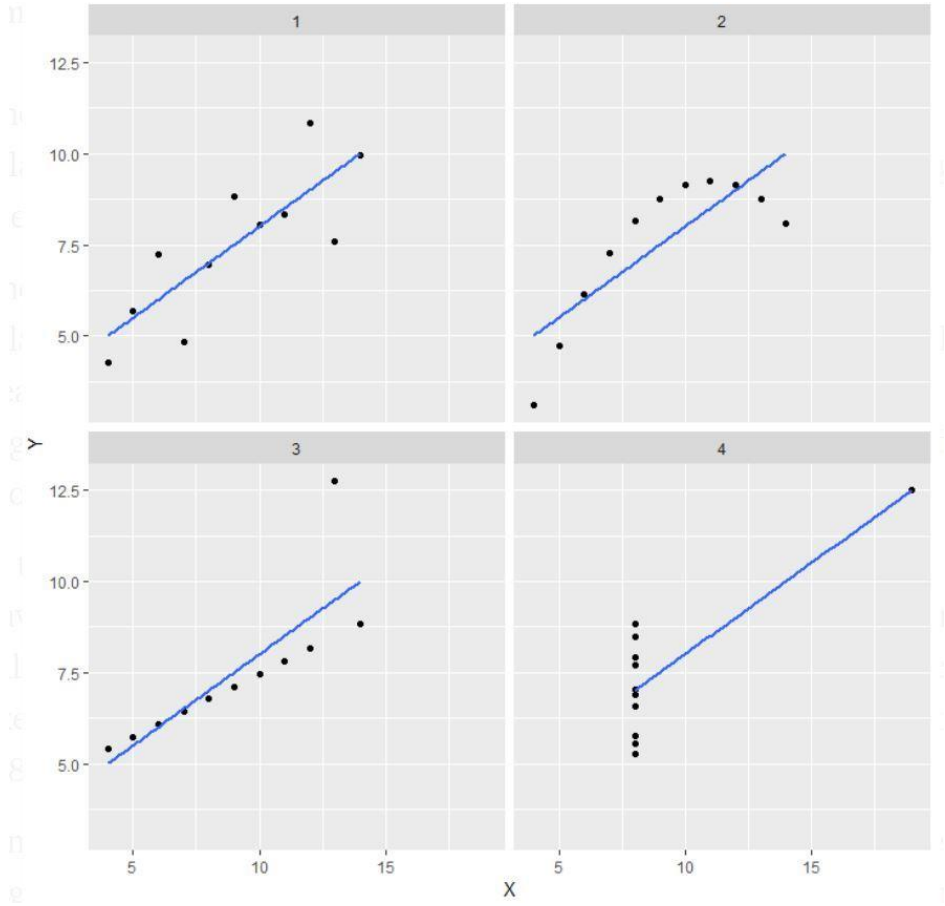
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	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 calculation on each data set :

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

Graphical representation of all 4 data sets



Explanation of this output:

- ✓ In the first one(top left) if you look at the scatter plot you will see that there seems to be a linear relationship between x and y.
- ✓ In the second one(top right) if you look at this figure you can conclude that there is a non-linear relationship between x and y.
- ✓ In the third one(bottom left) you can say when there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line.
- ✓ Finally, the fourth one(bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient.

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

- ✓ Pearson's r or Pearson's correlation coefficient is defined in statistics as the measurement of the strength of the relationship between two variables and their association with each other.
- ✓ In simple words, Pearson's correlation coefficient calculates the effect of change in one variable when the other variable changes.

- ✓ The Pearson coefficient correlation has a high statistical significance. It looks at the relationship between two variables. It seeks to draw a line through the data of two variables to show their relationship. This linear relationship can be positive or negative.
- ✓ The Pearson's R varies between -1 and +1 where:
    a) r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
    b) r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
    c) r = 0 means there is no linear association
- ✓ Pearson correlation coefficient formula:

$$r = \frac{N\Sigma xy-(\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2- (\Sigma x)^2][N\Sigma y^2 - (\Sigma y)^2]}}$$

Where:

N = the number of pairs of records/points
Σxy = the sum of the products of paired records/points
Σx = the sum of x records/points
Σy = the sum of y records/points
Σx2 = the sum of squared x records/points
Σy2 = the sum of squared y records/points

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

- ✓ 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 to handle highly varying values. 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, regardless of the unit of the values.

Why scaling is performed:

1. It helps in speeding up the calculations in an algorithm. It makes the model converge faster.
2. Dataset may contain features that are highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence it leads to incorrect modelling. So scaling is necessary.

3. It makes the model easier to interpret. If scaling in not done than there may be high difference in beta coefficients in our final model.

   ✓ It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

- Two most important techniques to perform Feature Scaling:

  ✓ **Min-Max/Normalization**:
  a) It brings all of the data in the range of 0 and 1. ***sklearn.preprocessing.MinMaxScaler*** helps to implement normalization in python.
  b) Normalization is good to use 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.

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

  ✓ **Standardization**:
  a) It is a very effective technique which brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).
  b) 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.

$$X' = \frac{X - \mu}{\sigma}$$

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

✓ VIF stands for Variance Inflation Factor

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

✓ In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. So, if there is a perfect correlation, then VIF = infinity. This shows a perfect correlation between one and other independent variables. 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

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

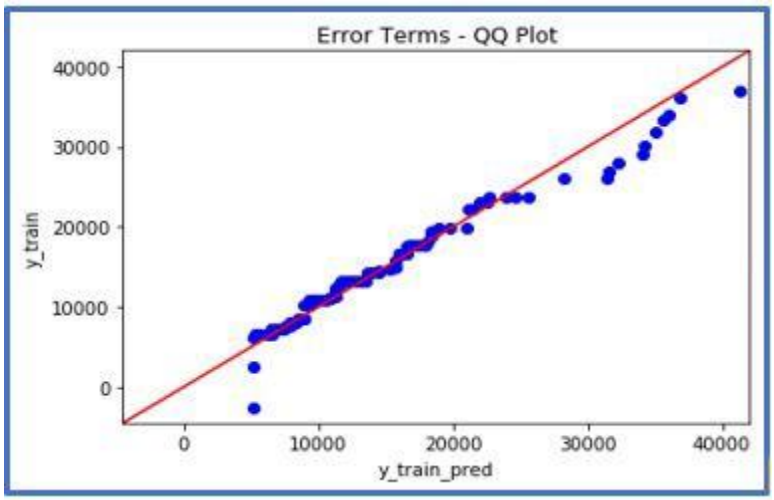
- ✓ A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.
- ✓ Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. 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.
- ✓ 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:

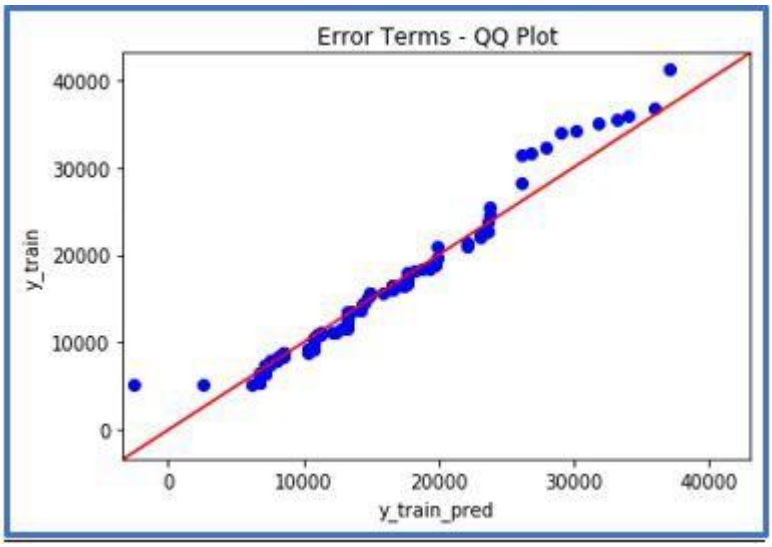
  If two data sets —

  a) come from populations with a common distribution
  b) have common location and scale
  c) have similar distributional shapes
  d) have similar tail behavior

- ✓ 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