Bike Sharing Assignment

Ritesh Sanghavi

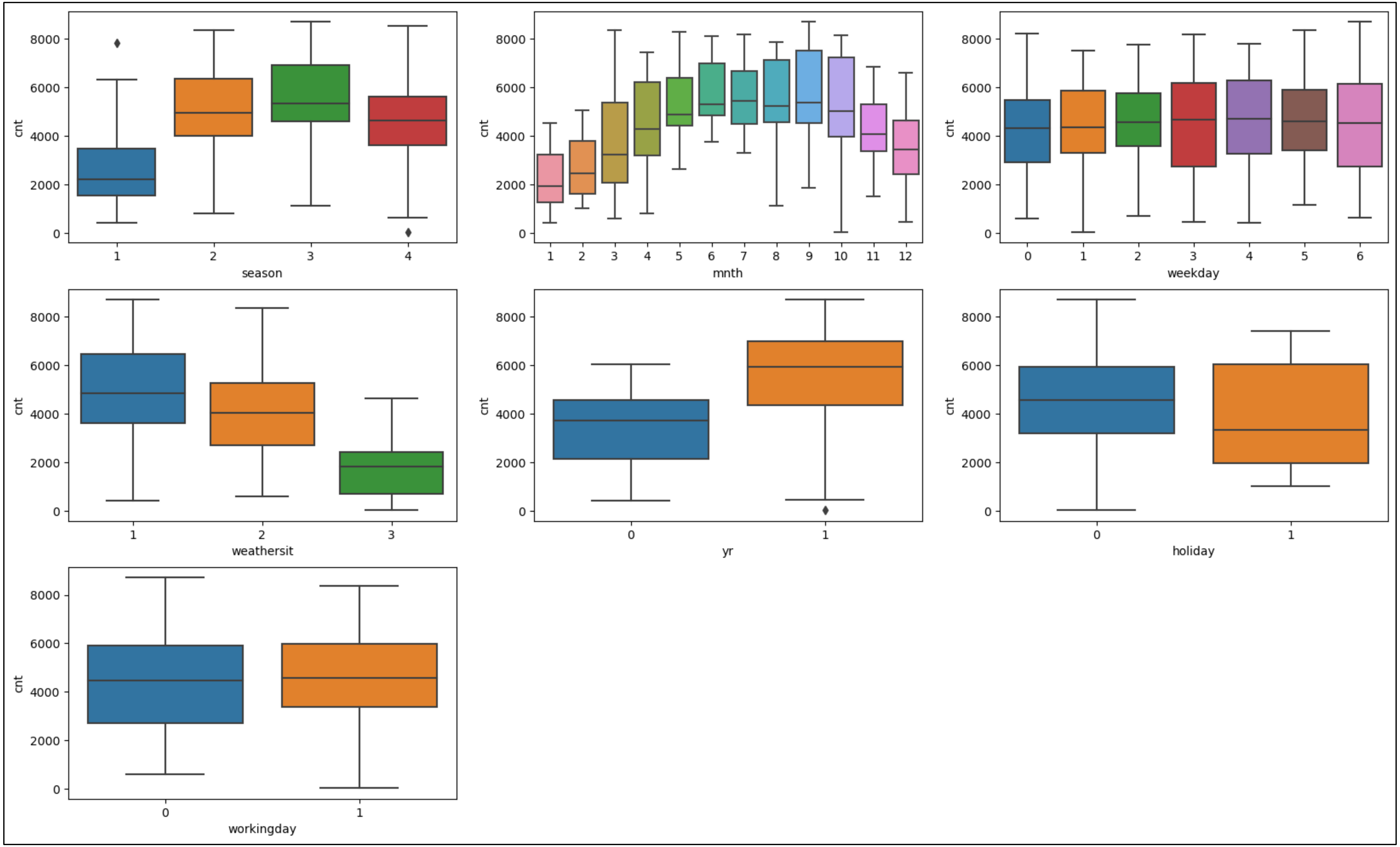
September 2023

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?

Ans: Season, Month, Weekday, Weathersit, Year, Holiday and Workingday are categorical variables in the data set. Categorical variables are displayed with box plots(*image in following slide*).

- Season 2, 3 and 4 shows good booking with mean value greater then 4000 compared to 1
- Month 8, 9,1 0 show good booking with mean value greater then 4000.
- Weekday - no significant difference
- WeatherSit 1 is significantly better then others having mean value greater then 4000 booking
- yr 1 (2019) is better year of booking bike compared to 0 (2018)
- Holiday shows no significant change. Data is consistent.
- Working day shows no significant observation. Data is consistent.


cnt
0
2000
4000
6000
8000
1
2
3
4
season
mnth
5
6
7
8
9
10
11
12
weekday
0
weathersit
yr
holiday
workingday

Assignment-based subjective questions

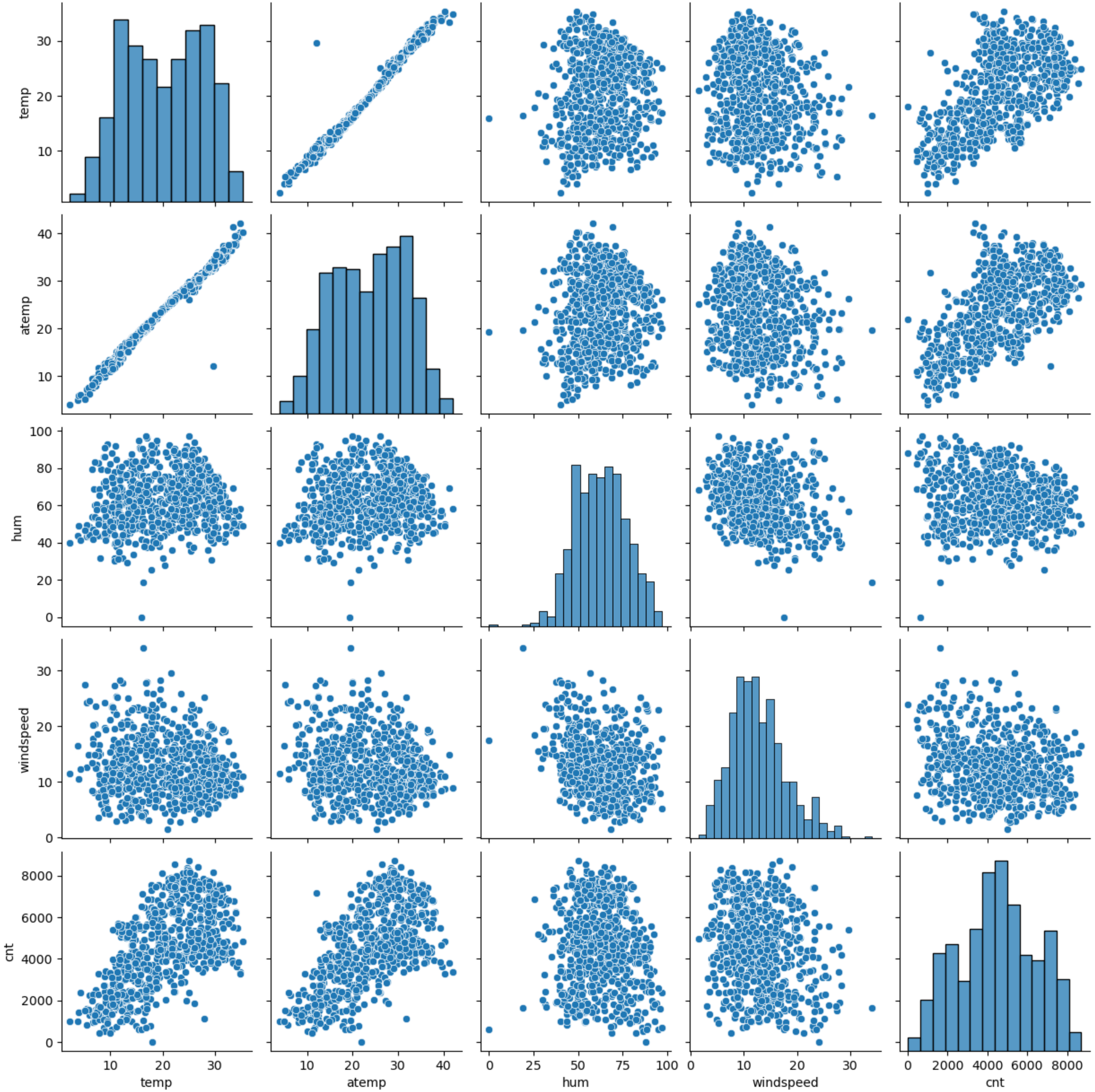
2. Why is it important to use **drop_first=True** during dummy variable creation?

Ans: Drop first is used to reduce the multicollinearity in Multiple Linear regression. Since k columns for k levels of a categorical variable is a good idea, there is a redundancy of one level, which is a separate column. Since one of the combinations will uniquely represent the redundant column. So it's good to drop one of the columns and keep k-1 columns to present k levels.

Assignment-based subjective questions

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

Ans: Based on the pair plot, temp and atemp variable has highest correlation with cnt target variable(image in following slide)

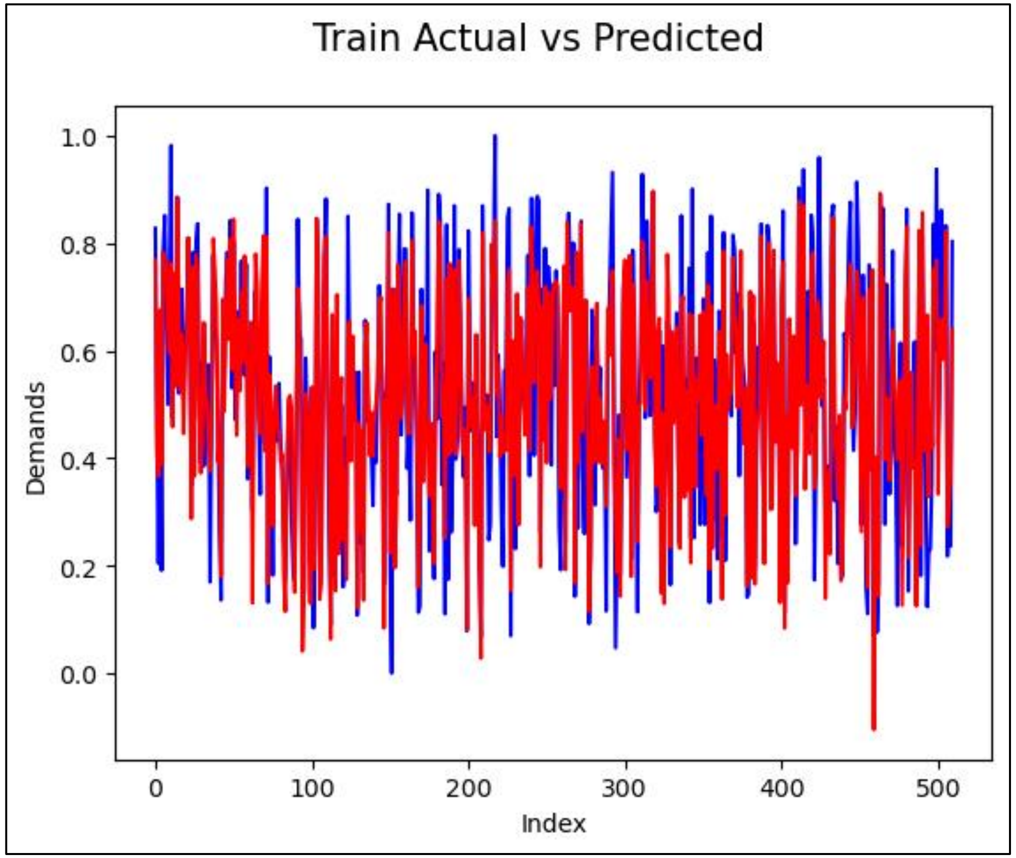

temp
30
20
10
atemp
40
30
20
10
hum
100
80
60
40
20
0
windspeed
30
20
10
0
cnt
8000
6000
4000
2000
0
10
20
30
temp
10
20
30
40
atemp
0
25
50
75
100
hum
0
10
20
30
windspeed
0
2000
4000
6000
8000
cnt

Assignment-based subjective questions

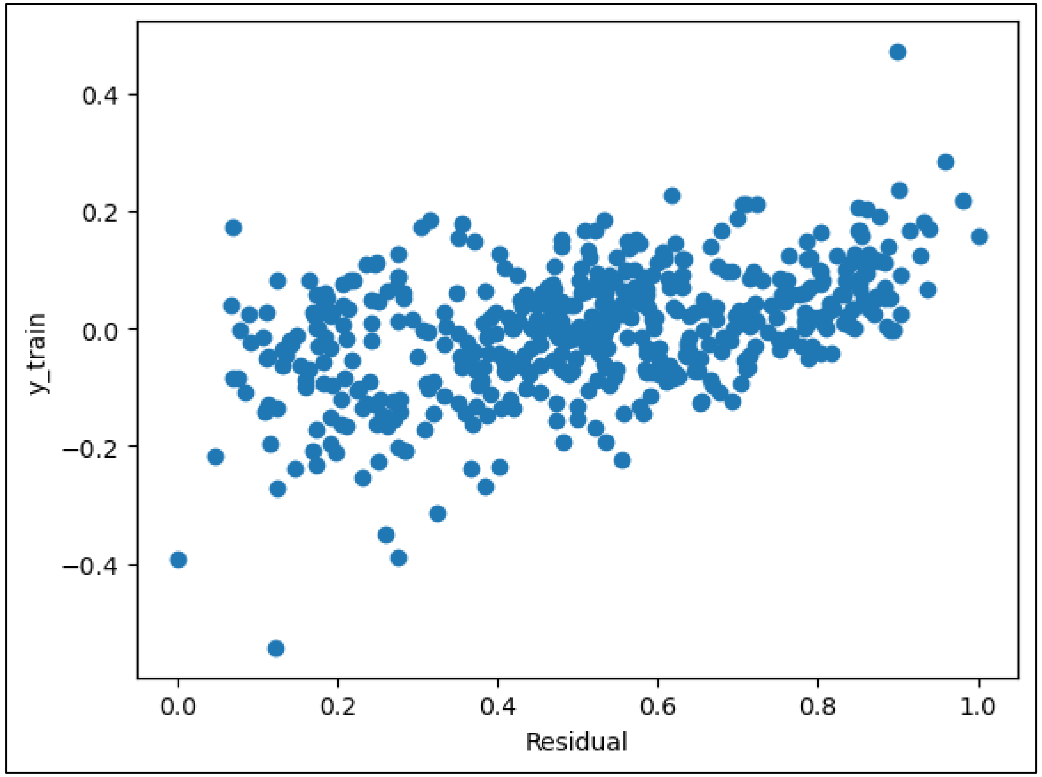
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans: The first step in the process is to check the R-squared, which in our case is 0.7790717032929404. This is considered relatively high for a model and is well within acceptable range. Second, we need to check the predicted vs actual graph. Both the lines are significantly overlapped, thus able to explain the demand very well.

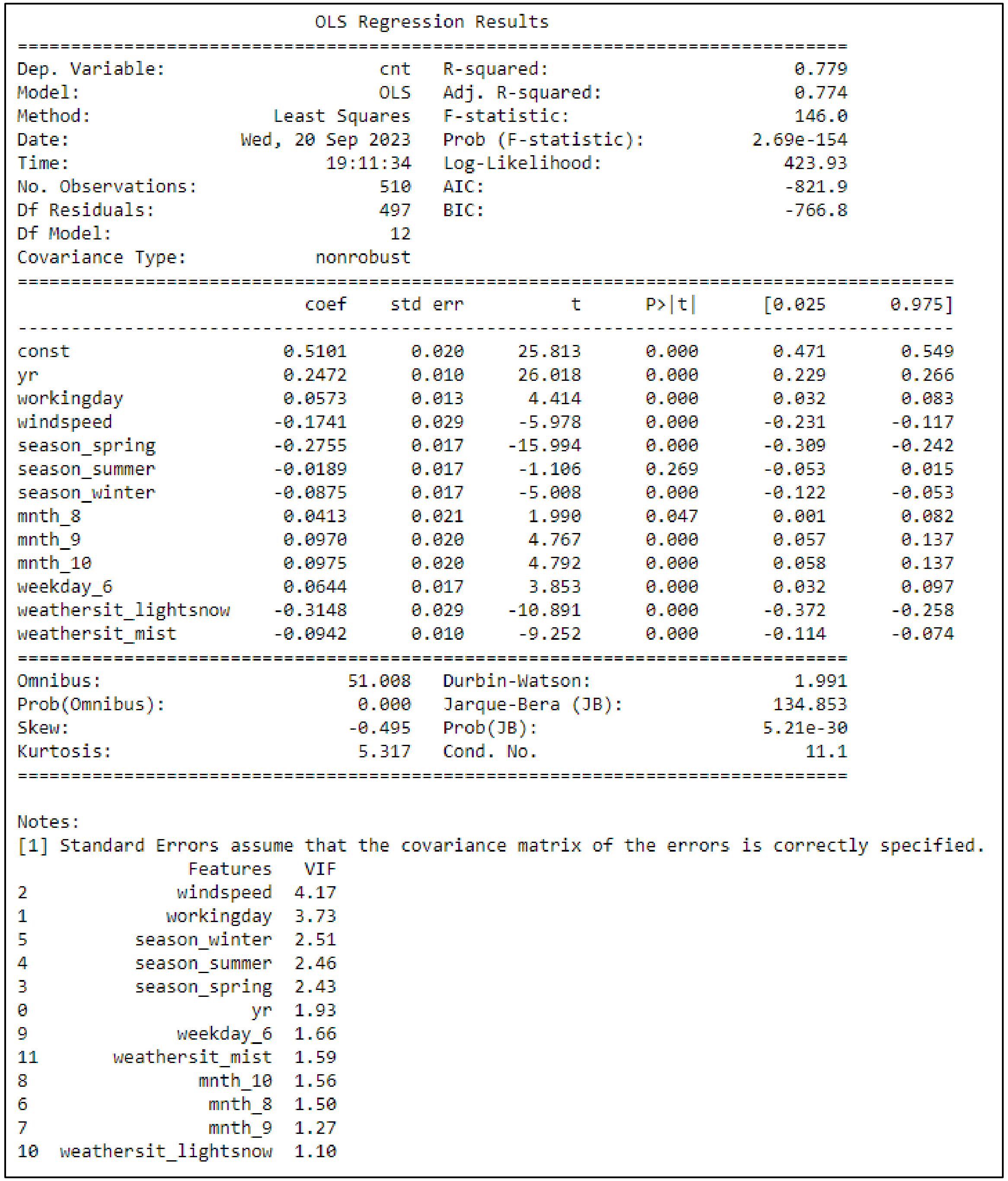
Contd…



Third, the error term representing the difference between predicted and actual values. As we notice, they are normally distributed again meeting criteria with mean = 0.



Other parameters are F-statistic and Prob(F-statistic). Higher F-statistic suggests that the regression model as a whole is statistically significant in predicting the dependent variable. When Prob (F-statistic) < 0.05, we can conclude that there is a statistically significant relationship between the independent variables as a group and the dependent variable. In our case respective values are 146 and 2.69e-154.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    cnt   R-squared:                       0.779
Model:                            OLS   Adj. R-squared:                  0.774
Method:                 Least Squares   F-statistic:                     146.0
Date:                Wed, 20 Sep 2023   Prob (F-statistic):          2.69e-154
Time:                        19:11:34   Log-Likelihood:                 423.93
No. Observations:                 510   AIC:                            -821.9
Df Residuals:                     497   BIC:                            -766.8
Df Model:                          12
Covariance Type:            nonrobust
========================================================================================
                           coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------
const                    0.5101      0.020     25.813      0.000       0.471       0.549
yr                       0.2472      0.010     26.018      0.000       0.229       0.266
workingday               0.0573      0.013      4.414      0.000       0.032       0.083
windspeed               -0.1741      0.029     -5.978      0.000      -0.231      -0.117
season_spring           -0.2755      0.017    -15.994      0.000      -0.309      -0.242
season_summer           -0.0189      0.017     -1.106      0.269      -0.053       0.015
season_winter           -0.0875      0.017     -5.008      0.000      -0.122      -0.053
mnth_8                   0.0413      0.021      1.990      0.047       0.001       0.082
mnth_9                   0.0970      0.020      4.767      0.000       0.057       0.137
mnth_10                  0.0975      0.020      4.792      0.000       0.058       0.137
weekday_6                0.0644      0.017      3.853      0.000       0.032       0.097
weathersit_lightsnow    -0.3148      0.029    -10.891      0.000      -0.372      -0.258
weathersit_mist         -0.0942      0.010     -9.252      0.000      -0.114      -0.074
==============================================================================
Omnibus:                       51.008   Durbin-Watson:                   1.991
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              134.853
Skew:                          -0.495   Prob(JB):                     5.21e-30
Kurtosis:                       5.317   Cond. No.                         11.1
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
                Features   VIF
2              windspeed  4.17
1             workingday  3.73
5          season_winter  2.51
4          season_summer  2.46
3          season_spring  2.43
0                     yr  1.93
9              weekday_6  1.66
11       weathersit_mist  1.59
8                mnth_10  1.56
6                 mnth_8  1.50
7                 mnth_9  1.27
10  weathersit_lightsnow  1.10
```

Assignment-based subjective questions

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

Ans: weathersit_lightsnow (-0.314818), season_spring (-0.275540) and yr (0.247188)

General subjective questions

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

Linear regression is a statistical technique used to model the relationship between a dependent variable (Y) and one or more independent variables (X) by fitting a linear equation. It aims to find the coefficients (β) that minimize the sum of squared errors (SSE) between predicted (Ŷ) and actual (Y) values. Once trained, the model can make predictions, and its performance is evaluated using metrics like Mean Squared Error (MSE). Linear regression makes assumptions like linearity, independence, homoscedasticity, normality, and no multicollinearity. It's widely used in fields like economics, finance, and machine learning for prediction and inference.

Cont…

Linear regression encompasses simple and multiple linear regression techniques.

Simple Linear Regression models the relationship between one independent variable (X) and a dependent variable (Y) using the equation $Y = \beta 0 + \beta 1 * X + \varepsilon$. The coefficients $\beta 0$ and $\beta 1$ are estimated using OLS:

$$\beta 1 = \frac{\Sigma(Xi - \bar{X})(Yi - \bar{Y})}{\Sigma(Xi - \bar{X})^2}$$
$$\beta 0 = \bar{Y} - \beta 1 * \bar{X}$$

Multiple Linear Regression extends this to multiple independent variables (X1, X2, ..., Xn) with the equation $Y = \beta 0 + \beta 1 * X1 + \beta 2 * X2 + ... + \beta n * Xn + \varepsilon$. The coefficients are determined similarly, and it allows modeling complex relationships.

Assumptions:

Linear regression makes several assumptions, including:

- Linearity: The relationship between independent and dependent variables is linear.
- Independence: Errors (ε) are independent of each other.
- Homoscedasticity: The variance of errors is constant across all levels of independent variables.
- Normality: The errors follow a normal distribution.
- No multicollinearity: Independent variables are not highly correlated with each other.

General subjective questions

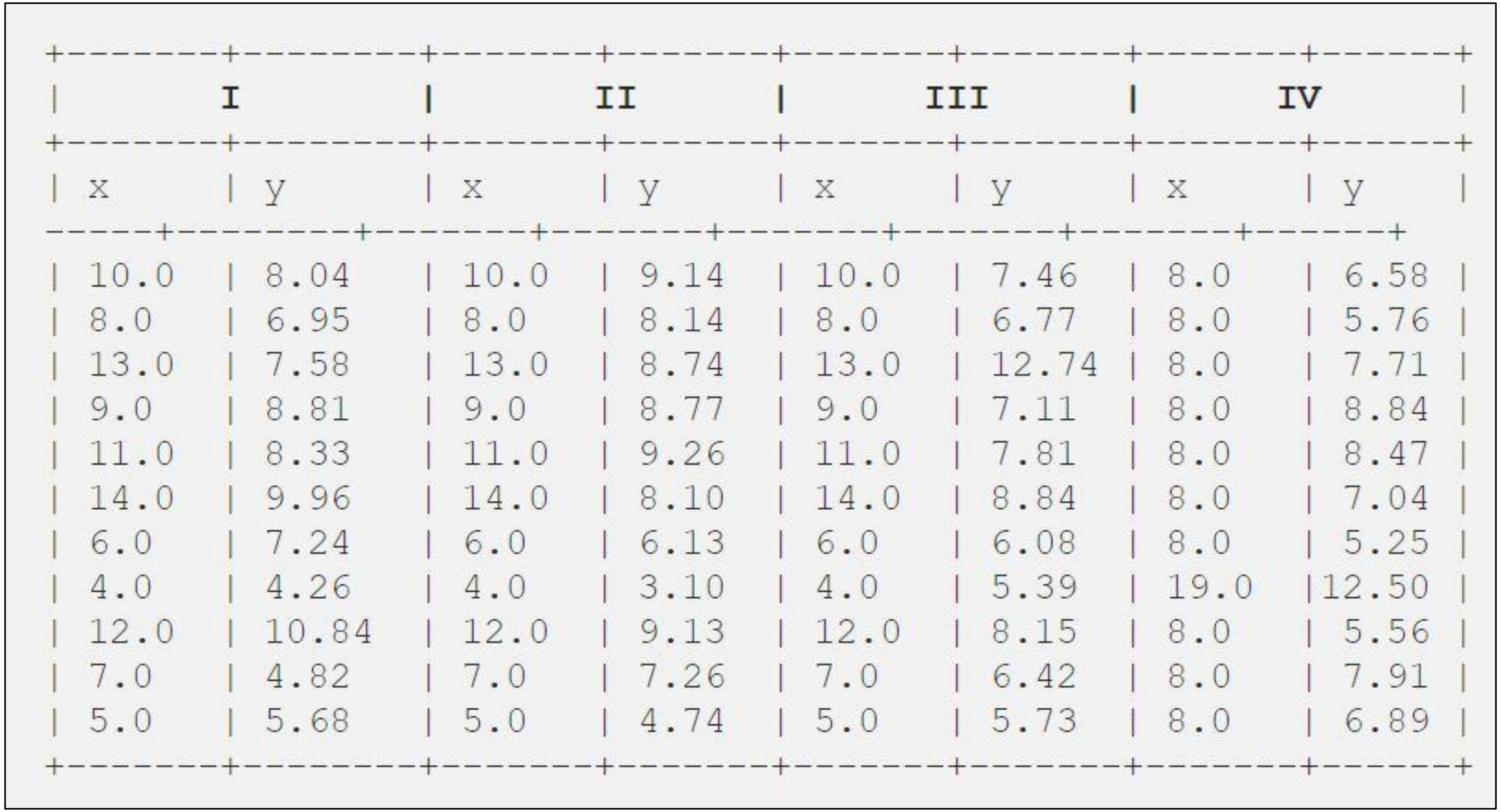
2. Explain the Anscombe's quartet in detail. (3 marks)

Anscombe's quartet comprises a set of four dataset, having identical descriptive statistical properties in terms of means, variance, R-Squared, correlations, and linear regression lines but having different representations when we scatter plot on graph. The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.

The four datasets that make up Anscombe's quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.

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.

The four datasets of Anscombe’s quartet:

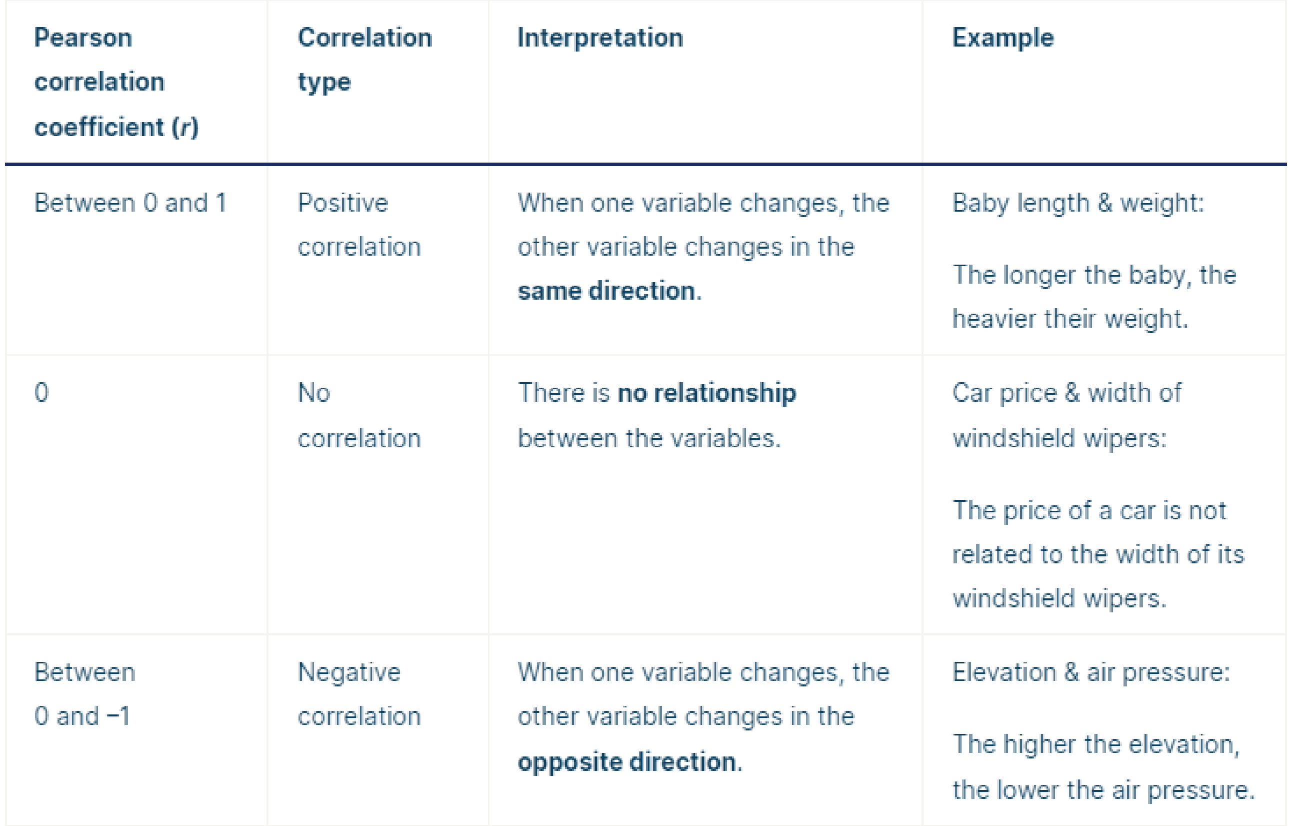


I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0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

General subjective questions

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

The **Pearson correlation coefficient (*r*)** is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.



Pearson correlation coefficient (*r*)	**Correlation type**	**Interpretation**	**Example**
Between 0 and 1	Positive correlation	When one variable changes, the other variable changes in the **same direction**.	Baby length & weight: The longer the baby, the heavier their weight.
0	No correlation	There is **no relationship** between the variables.	Car price & width of windshield wipers: The price of a car is not related to the width of its windshield wipers.
Between 0 and –1	Negative correlation	When one variable changes, the other variable changes in the **opposite direction**.	Elevation & air pressure: The higher the elevation, the lower the air pressure.

General subjective questions

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

When we have a lot of independent variables in a model, a lot of them might be on very different scales which will lead to a model with very weird coefficients that might be difficult to interpret. So we need to scale features because of two reasons: 1. Ease of interpretation 2. Faster convergence for gradient descent methods You can scale the features using two very popular method: 1. Standardizing: The variables are scaled in such a way that their mean is zero and standard deviation is one. 2. MinMax Scaling: The variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data. It is important to note that scaling just affects the coe

General subjective questions

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

VIF = infinity when there is perfect correlation between two independent variables. Since R2 =1, which leads 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).

General subjective questions

6. 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.

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.

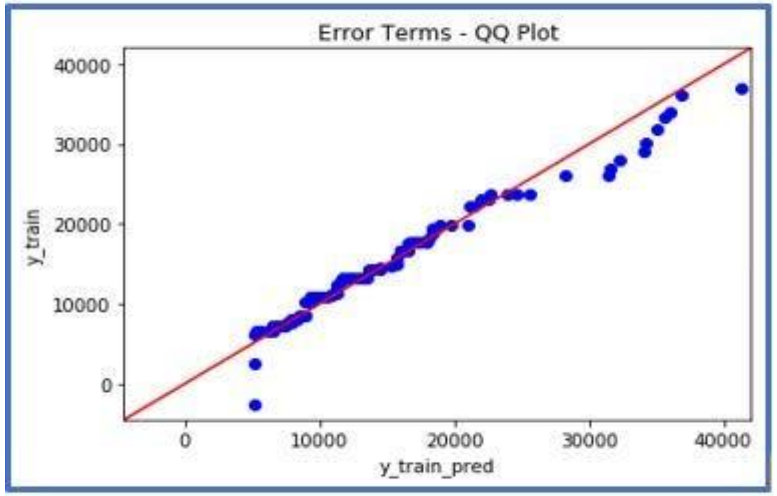
It is used to check following scenarios:

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

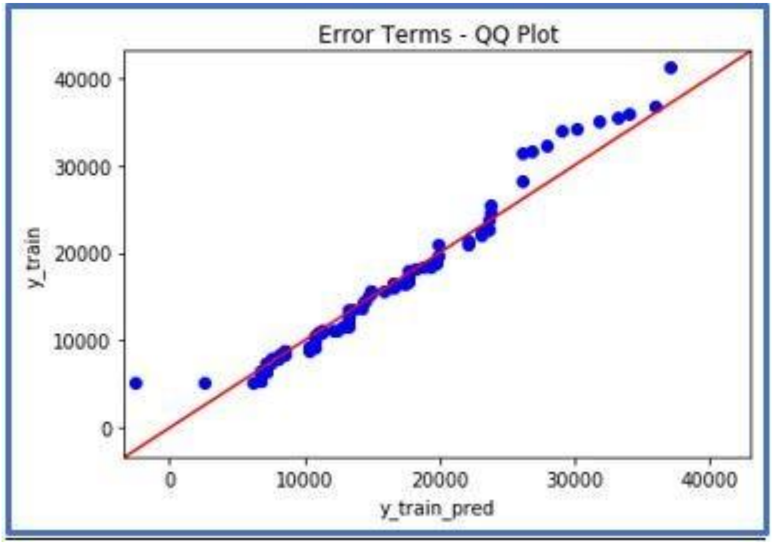
Interpretation:
A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. Below are the possible interpretations for two data sets.

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*