Subjective question:

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?
2. From the categorical variables,
3. 2019 has the more bike usage compared to 2018.
4. LowRainandSnow and MistandCloudy the lower bike due to climate scenerios
5. Spring and Winter has the hard the climate so lower bike usage
6. July has high temperature and Sep, Mar of spring or snow seasons
7. Why is it important to use drop\_first=True during dummy variable creation?
8. drop\_first = True, then it will drop the first category.
9. So if you have K categories, it will only produce K – 1 dummy variables.
10. The first category is easily able to identified using the other category data and it act as redundant.

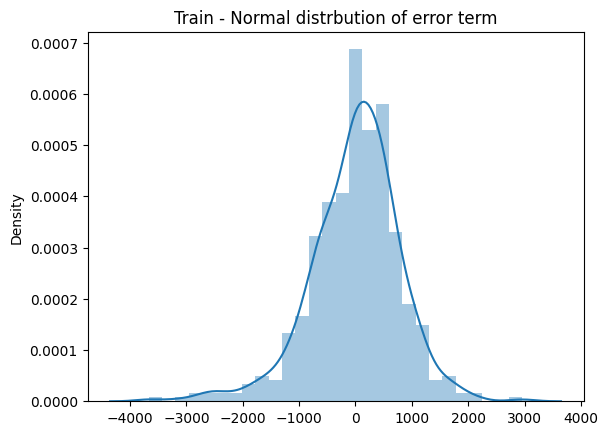
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

A) from the pair plot the ‘ atemp‘ has the highest correlation with the target variable ‘cnt’

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

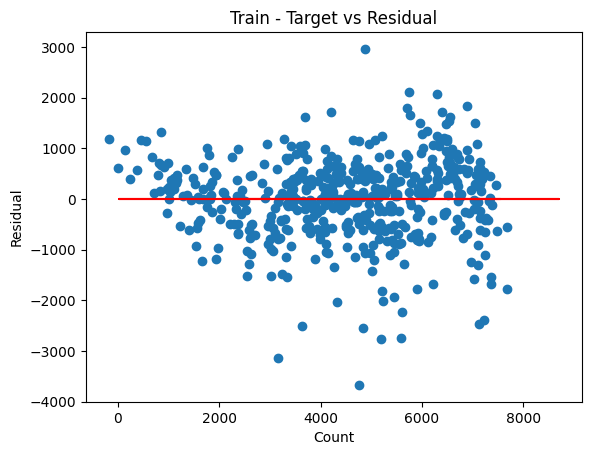
A) Mulitcolinearity between variables.

B) Mean of disrtibuted is 0 and the distribution of errors is normal and plot the histogram of the errors



C) The variance of errors is constant across all levels of the independent variables, plot the residuals versus the predicted values of y is show below

D)Error terms are not be dependent on one another. No auto correlation was calculated by Durbin-Watson\_statistic: 2.002 and in range between 1.50 to 2.50



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

A) Based on the coefficent and below three feature analysis.

B) The first three features are

C) temp

D) 2019

E) Sep

F) Feature: 2019,Score: 2033.64173

G) Feature: holiday, Score: -819.74379

H) Feature: LowRainandSnow, Score: -2510.57190

I) Feature: windspeed, Score: -1254.66710

J) Feature: spring, Score: -1010.25520

K) Feature: Jul, Score: -606.66553

L) Feature: Sep, Score: 520.28944

M) Feature: temp, Score: 3937.52655

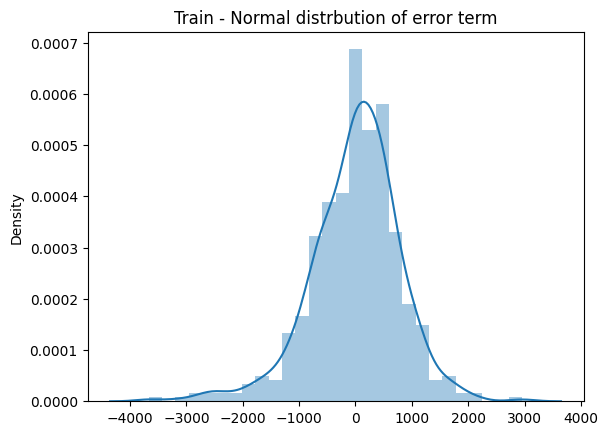
N) Feature: Mar, Score: 319.21269

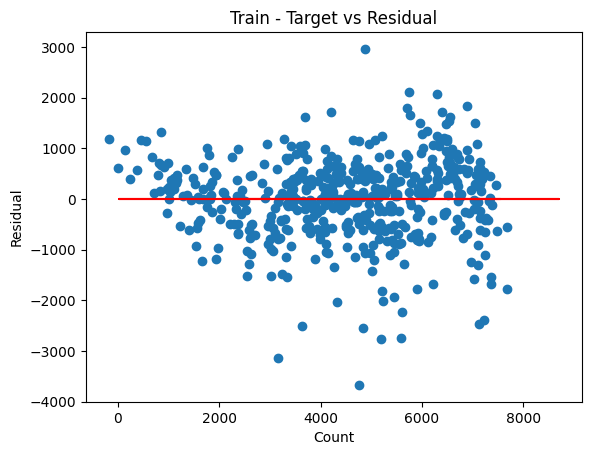
O) Feature: MistandCloudy, Score: -703.98768

P) Feature: winter, Score: 440.59056

GENERAL SUBJECTIVE QUESTIONS:

1. Explain the linear regression algorithm in detail?
2. Linear Regression is supervised classifier.
3. It used to predict or forecast the output for the continuous variables.
4. It predict the value using the historical data of the independent variables.
5. The Linear Regression coefficient of independent variables and try to fit the best fit lines using the least square method for set of paired data
6. Assumption of linear regression:
7. Mulitcolinearity between variables.
8. Mean of disrtibuted is 0 and the distribution of errors is normal and plot the histogram of the errors.
9. The variance of errors is constant across all levels of the independent variables, plot the residuals versus the predicted values of y is show below. Check for homoscedasticity
10. Error terms are not be dependent on one another. No auto correlation was calculated in range between 1.50 to 2.5





1. Explain the Anscombe’s quartet in detail?
2. Anscombe’s quartet consist of the four dataset which has the similar statical properties while plotting it shows different in the visualization.
3. The below dataset shows similar mean and standard deviations, but while plot it show different
4. What is Pearson’s R?
5. Pearson’s R is the coefficent that used to measure the strength between different variables and their relationship +1 – Perfect positive relationship between two variables.
6. One increase and other also will increase -1 – Perfect negative relationship between two variables.
7. One increase and other will decrease. 0 – no relationship between varibales. Its range from the +1 to -1.

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

A) The scaling is used to normalized the variables into different scale into similar scale. so that computation and memory for algorithm can be efficiently optimized.

Normalization:

1. The scale value between range of 0 to 1
2. Sensitive to outliers
3. Shape of the original distribution is not changed
4. It may not preserve the relationships between the data points
5. Equation: (x – min)/(max – min)

Standardization:

1. It centres data around the mean and scales to a standard deviation of 1
2. Less sensitive to outliers
3. Changes the shape of the original distribution
4. it will preserve the relationships between the data points
5. Equation: (x – mean)/standard deviationsss