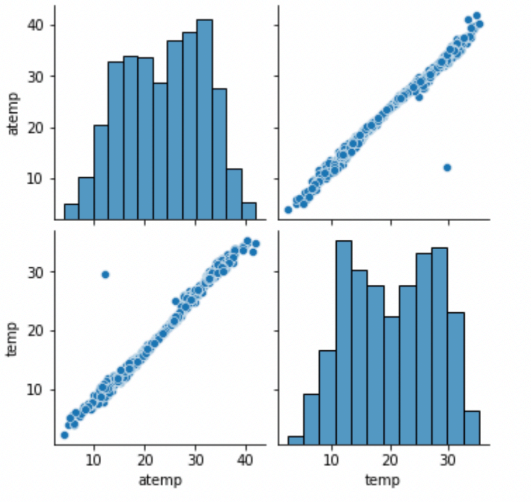
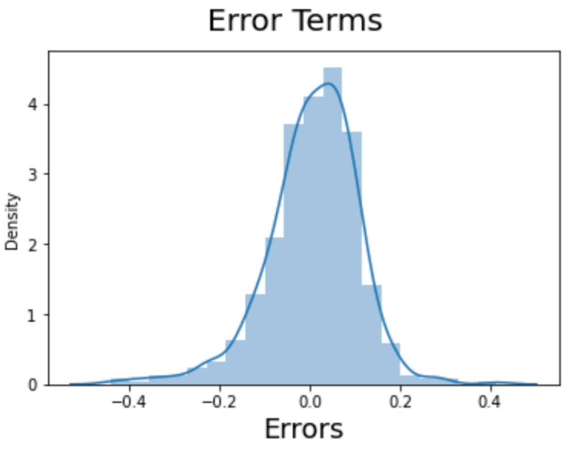
Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?
   Ans: ***Season****: The boxplot showed that Summer and Fall season has more value of cnt.*
   Month*: 5th to 10th month seem to have more customers.*
   Holidays*: Holidays seems to have lesser rentals than the other days.*
   WeatherSituation*: Clear and partly cloud weather situarion attracts more customers that mist, snow and heavy rain.*
   Year*: The number of rentals of year 2019 is more than 2018*

2. Why is it important to use **drop_first=True** during dummy variable creation?
   Ans: *During the dummy variable creation, it leads to Multicollinearity between the dummy variables. We drop one column using drop_first to keep it under control and avoid multicollinearity.*

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
   Ans: *Column **temp** and **atemp** are the two numerical variables which are highly correlated with **cnt***





4. How did you validate the assumptions of Linear Regression after building the model on the training set?
   Ans: *Using Residual distribution with the predicted Y we could see there is a **normal distribution entered around 0.***

5. Based on the final model, which are the top 3 features contributing significantl towards

explaining the demand of the shared bikes?

Ans: *Below 3 variables have more significance with higher coefficient:*

yr	*0.2420*
atemp	*0.7576*
weathersit_LightSnow	*-0.2728*

General Subjective Questions

1. Explain the linear regression algorithm in detail.
   Ans: *Linear regression is a supervised learning algorithm used to predict the numeric value of continuous and constant slope. We use this methodology to predict values within a continuous range. It does not deal with classification.*
   Linear regression is based on the concept of finding the best fit line which is
   Y = c + mX (where c is the intercept and m is the slope/coefficient)
   There are two types of linear regression:
   ***Simple linear regression** is used when the dependent variable is predicted using only one independent variable.*
   ***Multiple Linear Regression** is used when the dependent variable is predicted using multiple independent variables.*

2. Explain the Anscombe's quartet in detail.
   Ans: ***Anscombe's** Quartet is a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.*

3. What is Pearson's R?
   Ans: *It is defined as the strength of the linear association between the variables.*
   Its value ranges between -1 to +1. It shows the linear relationship between two sets of data. With this we can identify if we can draw a line graph to represent the data or not
   r = 1 means the data is perfectly linear with a positive slope
   r = -1 means the data is perfectly linear with a negative slope r = 0 means there is no linear association

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
   Ans: *Scaling is used to normalize the set of independent variables of the*

data within a certain range in order to maintain closer coefficients. Without this machine learning algorithm can weigh greater values, higher and consider smaller values as the lower values, irrespective of the units of the values. Different type of scaling used in ML
-Absolute Maximum Scaling.
-Min-Max scaling (Used in our algorithm)
-Normalisation
-Standardisation
-Robust scaling

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
   Ans: *The VIF can happen to be infinite(inf) which happens because of any variable being perfectly correlated.*
   Where (VIF) =1/(1-R^2) and R is supposed to be 1 in case of a perfect correlation
   Hence VIF = 1/0 = infinity

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

Ans: *Q-Q plot is also know as Quantile-Quantile plot which plots the quantiles of a sample distribution against quantiles of a theoretical distribution. Which helps us understand the probability distribution like normal, uniform, exponential.*