Linear Regression Assignment

Assignment-based Subjective Question -Answer

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

The barplot shows the relationship between **categorical variables** and a dependent variable as :

- Bike Rentals are more during the Fall season and then in summer
- Bike Rentals are more in the year 2019 compared to 2018
- Bike Rentals are more in clear weather
- Bike Rentals are more on Saturday,wednesday and Thursday
- Bike Rentals are more in June, August and September

The Scatterplot shows the relationship between **numerical variables** and dependent variable as

- Rentals are observed at higher temperatures
- Bike Rentals are observed at higher humidity

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

Each of the dummy variables has 'm' levels. So, to represent one categorical variable, we would require (m-1) levels. Hence, to represent 'n' categorical variables, you would need (m - 1) * n dummy variables. So to perform m-1 levels we need to drop the first column of the dummy variable.

For example,

	Clear	Light Snow	Mist+Cloudy
Clear	1	0	0
Light Snow	0	1	0
Mist+Cloudy	0	0	1

	Light Snow	Mist+Cloudy
Clear	0	0
Light Snow	1	0
Mist+Cloudy	0	1

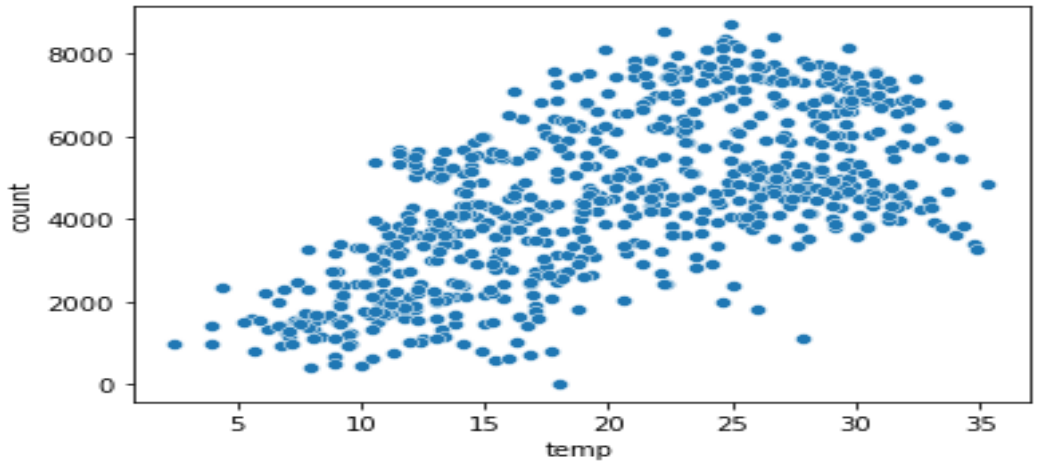
Clear predictor variable represents base value 00

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

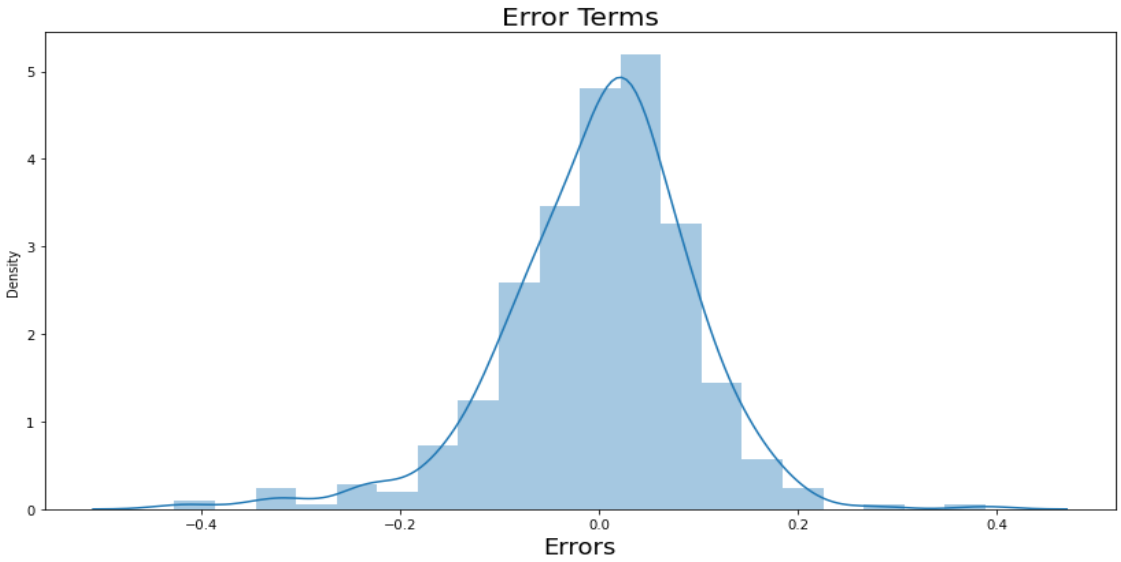
Registered variable has highest correlation with Count variable
Apart from registered and casual temperature variable is highest correlated with target variable

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

Linearity : By plotting Scatter plot to see whether the relationship between predictor variables (X) and target variable (Y) is linear



Normality : By plotting Histogram for error terms to check whether they are normally distributed



Homoscedascity : Error terms have *constant variance*

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

The top 3 features contributing significantly towards explaining the demand of the shared bikes are Temperature ,Year and Wheather