Assignment based subjective questions :

Question 1: From your analysis of the categorical variables from the dataset, what could you infer about their effect on independent variable?

Below are the categorical variables which has the effect on independent variable in decreasing order as below.

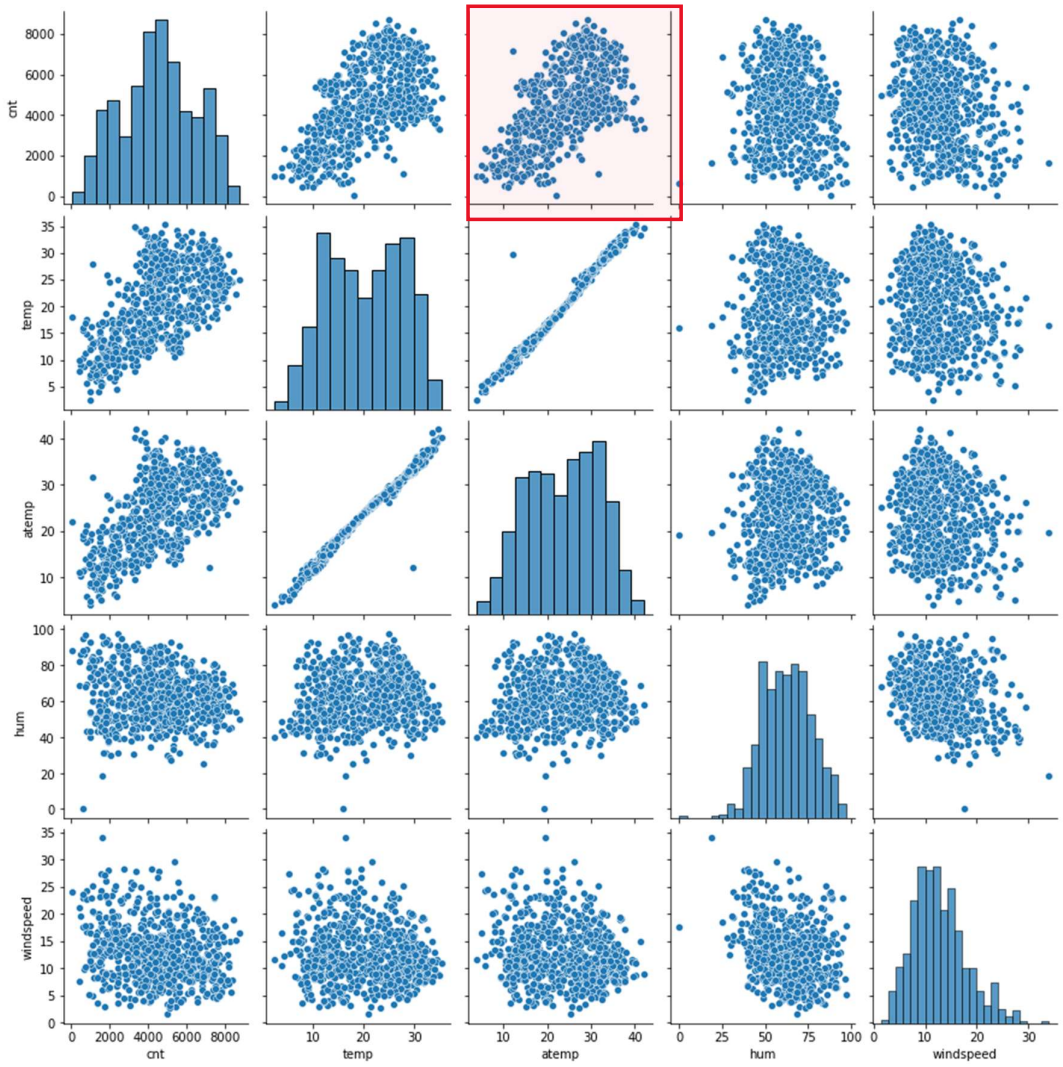
Variable	Coefficient of the variable	Inference
Year_2019	0.2342	Year to Year there is an increase of 23% in the business
Season – SPRING	-0.1238	During spring there is 12% reduction from the demand.
Season – Winter	0.0476	During winter there is an increase of 4% count.
Workingday	0.0255	During working day there is an increase of 2.55% demand.
Weathersit – LIGHT_RAIN	-0.2888	During light rain conditions, there is a significant decrease in the demand around 29%
Weathersit - MIST	-0.0746	During mist, there is a decrease in demand of around 7.5%

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

The category with n variable can be created/identified with n-1 bits. So it is not necessary to include one more bit for the categorical variable. If we have dummy for all the categorical variables, then that will lead to multicollinearity. So it is necessary to drop one of the categories which is done using drop_first=True.

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

The atemp has highest correlation with the target variable count. The next variable with highest correlation is temp.



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

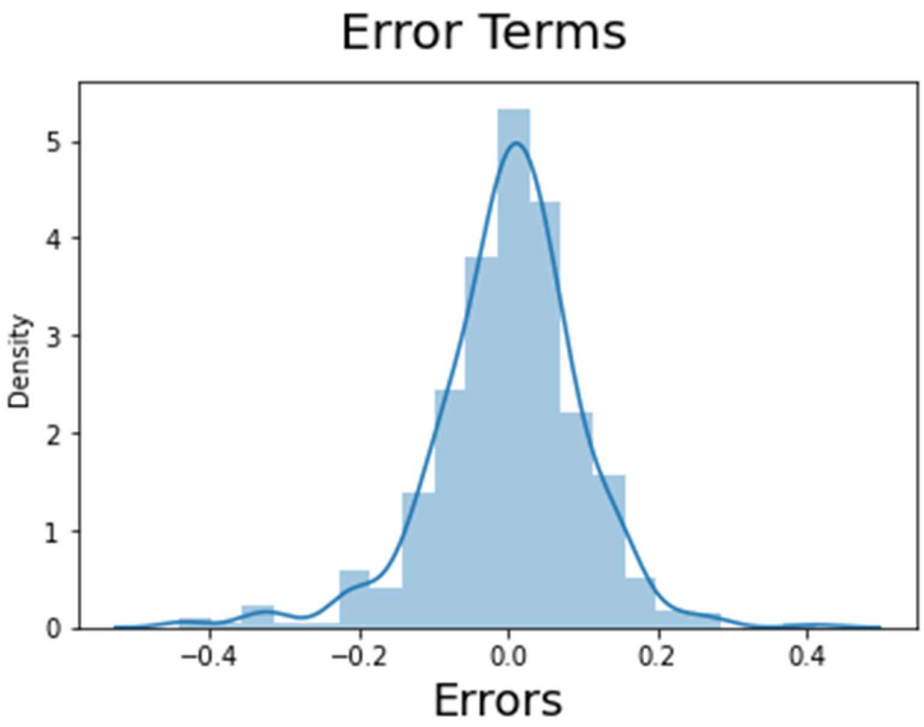
Below are the assumptions of the linear regression model

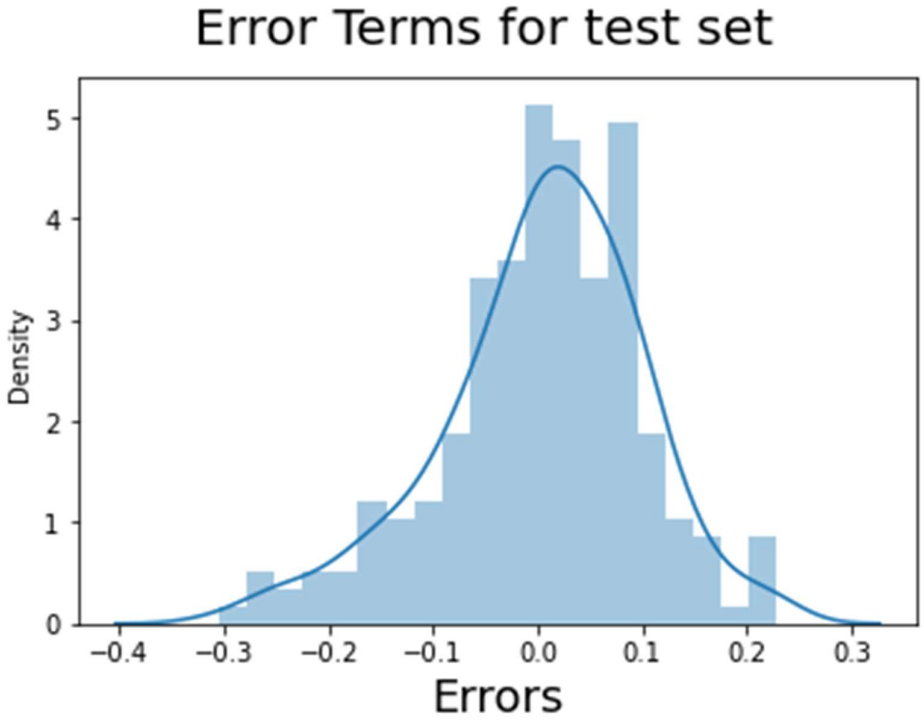
1.Error terms should be normally distributed.

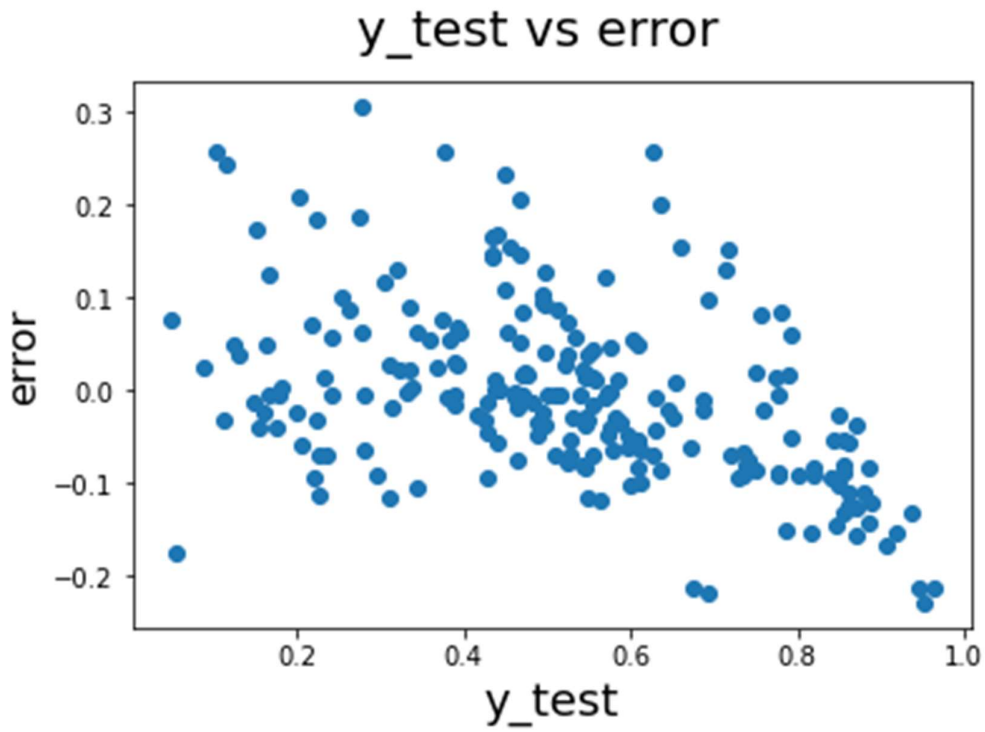
2. Error terms should be having mean zero and independent of each other.No visible pattern in the residuals.

3. Error terms have constant variance.

Attaching the screenshot of the above conditions.


Error Terms
Density
0
1
2
3
4
5
−0.4
−0.2
0.0
0.2
0.4
Errors


Error Terms for test set
Density
0
1
2
3
4
5
−0.4
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
Errors



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

- atemp - Based on the feeling of temperature, the demand raises upto 46%

- YEAR_2019 - Year on Year the business improvement is around 23%

- WINTER – During winter there is a 4% increase in the demand.

These three factors contribute to around 73% percent of demand and thus help in predicting the demand.

General subjective questions.

1. Explain the linear regression algorithm in detail
    - Do the data understanding, data quality and data visualization steps.
    - Algorithm for multiple linear regression.
        - Check for multicollinearity between different features.
    - For categorical variables, handle the columns and get dummies.
    - Get the model features using Recursive feature engineering.
    - Do the model assessment on the RFE supported features and select a model with best performance based on VIF and p value.
    - Confirm the assumptions on the linear regression on residuals.
    - Get the value of R2 score and mean square error score to confirm the optimized model.
2. Explain the Anscombe's quartet in detail.

    Anscombe's quartet consists of four datasets, which has the same descriptive statistics, but looks different when graphed. It is used to demonstrate the importance of visualization when analyzing a dataset. It also portrays the effect of outliers on statistical properties. This is mainly to counter the impression that numerical calculations are exact but graphs are rough.

3. What is Pearson's R?

    Correlation coefficient is used to measure the relationship strength between two variables. The value range is between -1 and 1.

    Say if we have two variables x and y.

    - If x increases, then y also increase then it is said to be positive correlation. The value will be in the range of 0> value >1.
    - If x increases and y remains constant and vice versa, then it is no correlation and the correlation coefficient would be 0.
    - If x increases and y decreases, then it is negative correlation. The value will be in the range of -1< value <0.
    - The strength of the relationship is measured by the absolute value of the correlation coefficient.

    The pearson's R can be calculated as shown in the formula below.

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

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

   Scaling : Scaling is the technique used to bring the variables to the closer range of one another.

   Why? : It is needed for the proper interpretation of coefficients. Without scaling the variables with high value range might have less coefficient value and variables with low value range would be having high coefficient value even if the impact by high range variable is more. This might lead to misinterpretation of coefficient by the business.

   Normalized scaling vs Standardized scaling.: Normalized scaling which is also called min-max scaling which shifts the value in the range of 0 to 1.

   Standardized scaling is the technique of shifting values with mean as 0 and standard deviation as 1.

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

   VIF with infinity means it has perfect multicollinearity. That means that the feature can be perfectly expressed with the help of other features, hence not required. Ie., Square of R is 1. To mitigate the multicollinearity situation, one of the features which contributes to the issue should be dropped.

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

   Q-Q plot is a probability plot where we compare two probability distributions. It done by plotting quantiles of both against each other. It is used to check whether the observations come from the same distribution.
   This can be used to check the normality of residuals in the linear regression.