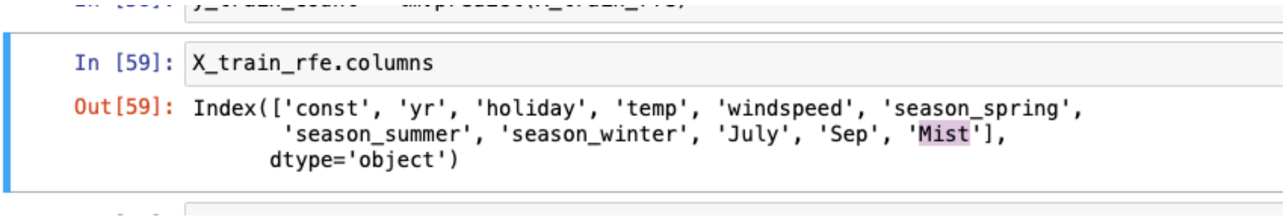
Assignment-based Subjective Questions

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

From the observations made in the assignment, the categorical variables have a less or no significance in the linear regression. As the final model built after analysis doesn't have the categorical variables.



```
In [59]: X_train_rfe.columns

Out[59]: Index(['const', 'yr', 'holiday', 'temp', 'windspeed', 'season_spring',
                'season_summer', 'season_winter', 'July', 'Sep', 'Mist'],
               dtype='object')
```

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

While doing the dummy variable creation, the number of columns that gets created is equal to the number of values present in the column. Like if we have seasons as Summer, winter and rainy there will be 3 columns that gets created.
But if we use the binary representation we can do it with just 2 columns instead of 3.
So to increase the performance of the model, we use this step.

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

Holiday has the highest correlation.

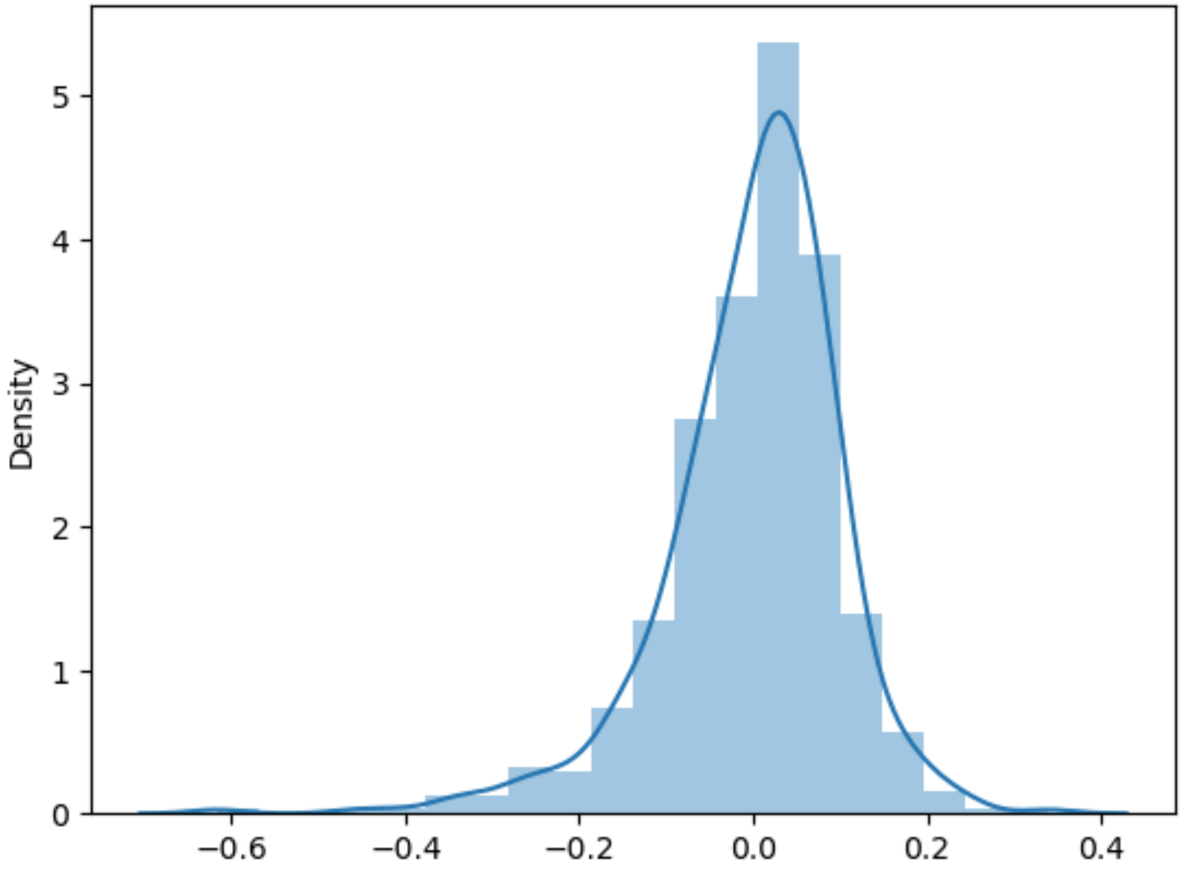
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Assumptions of linear regression were validated on the following criteria.

a. The p-value for all the variables is less than 0.05

b. VIF for all the variables is less than or equal to 5

c. The residual analysis shows the below graph with normal distribution.



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 to the model built are

1. Holiday
2. September month
3. Misty weather

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression algorithm is widely used in machine learning and statistics for modelling the relationship between a dependent variable and one or more independent variables.

The steps involved are

1. Problem formulation
2. Assumptions
3. Simple linear regression

*The equation for simple linear regression is Y = b0 + b1*X, where b0 is the y-intercept (constant term) and b1 is the slope (coefficient) of the line.*

4. Multiple linear regression

* *Multiple linear regression extends the concept of simple linear regression to multiple independent variables.*

* *The equation for multiple linear regression is Y = b0 + b1X1 + b2X2 + ... + bn*Xn, where b0, b1, b2, ..., bn are the coefficients for each independent variable.*

5. Training the model

Training the linear regression model involves finding the optimal values for the coefficients (b0, b1, ..., bn) that minimize the error between the predicted values and the actual values.

6. Evaluating the model

After training, the model's performance needs to be evaluated to assess its accuracy and generalization ability.

- *Common evaluation metrics for linear regression include the coefficient of determination (R-squared), mean squared error (MSE), root mean squared error (RMSE), and mean absolute error (MAE).*

7. Making Predictions

Once the model is trained and evaluated, it can be used to make predictions on new, unseen data.

- *Given the values of the independent variables, the model calculates the predicted value of the dependent variable using the learned coefficients.*

2. . Explain the Anscombe’s quartet in detail.

Here is a detailed description of each dataset in Anscombe's quartet:

1. Dataset I:

- x: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5
- y: 8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82, 5.68
- When graphed, this dataset exhibits a relatively linear relationship between x and y, with a slight positive slope.

2. Dataset II:
   - x: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5
   - y: 9.14, 8.14, 8.74, 8.77, 9.26, 8.10, 6.13, 3.10, 9.13, 7.26, 4.74
   - Similar to Dataset I, this dataset also shows a linear relationship between x and y, but with a different slope. However, there is also a distinct non-linear pattern in the data, particularly evident in the curved nature of the graph.
3. Dataset III:
   - x: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5
   - y: 7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42, 5.73
   - This dataset has a clear non-linear relationship, with an outlier that significantly affects the linear regression line.
4. Dataset IV:
   - x: 8, 8, 8, 8, 8, 8, 8, 19, 8, 8, 8
   - y: 6.58, 5.76, 7.71, 8.84, 8.47, 7.04, 5.25, 12.50, 5.56, 7.91, 6.89
   - Dataset IV demonstrates that even though all x-values are the same, the corresponding y-values exhibit a curved relationship, indicating that a simple linear model is not appropriate.

3. What is Pearson’s R?

Pearson's correlation coefficient, often denoted as "r" or "Pearson's r," is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It was developed by Karl Pearson in the late 19th century and is widely used in various fields, including statistics, social sciences, and data analysis.

Pearson's correlation coefficient takes values between -1 and +1, where:

- A value of +1 indicates a perfect positive linear relationship, meaning that as one variable increases, the other variable also increases proportionally.
- A value of -1 indicates a perfect negative linear relationship, meaning that as one variable increases, the other variable decreases proportionally.
- A value close to 0 indicates a weak or no linear relationship between the variables.

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

Scaling refers to the transformation of the data to a common scale. The different variables are represented using different scales, and it will give incorrect results if we use them as is.

So it is necessary to convert them to a common scale, in order to build a model using them

There are two common types of scaling: normalized scaling and standardized scaling.

1. Normalized Scaling (Min-Max Scaling):
    - Normalized scaling transforms the variable values to a predefined range, typically between 0 and 1.
    - It preserves the relative relationships and distribution of the data.
        - The formula for normalized scaling is:
        - X_scaled = (X - X_min) / (X_max - X_min)
        - Where X_scaled is the scaled value, X is the original value, X_min is the minimum value of the variable, and X_max is the maximum value of the variable.
    - Normalized scaling is useful when the distribution of the variable is known and bounded, and preserving the relationship between the values is important.
2. Standardized Scaling (Z-score Scaling):

- Standardized scaling transforms the variable values to have zero mean and unit variance.
- It centers the data around zero, with a standard deviation of 1.
    - The formula for standardized scaling is:
    - X_scaled = (X - X_mean) / X_std
    - Where X_scaled is the scaled value, X is the original value, X_mean is the mean of the variable, and X_std is the standard deviation of the variable.
- Standardized scaling is useful when the distribution of the variable is unknown or not necessarily bounded, and when comparing variables with different units or scales is important.

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

In some cases, the VIF can be computed as infinite, indicating a perfect multicollinearity issue. This happens when one or more independent variables in a regression model can be perfectly predicted by a linear combination of other independent variables. In other words, there is a complete linear dependency among the predictor variables.

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

A 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.

The Q-Q plot compares the quantiles of the observed data against the quantiles expected from the theoretical distribution. Here's how a Q-Q plot is constructed:

1. Sort the observed data in ascending order.
2. Compute the quantiles for both the observed data and the theoretical distribution. Typically, the quantiles are calculated using percentiles.
3. Plot the observed quantiles on the y-axis and the expected quantiles from the theoretical distribution on the x-axis.