Assignment-based Subjective Questions

Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: <Your answer for Question 1 goes below this line> (Do not edit)

The final model predicts the Demand through this equation:

Final Model Equation = const * 0.1724 + yr * 0.2401 + holiday * -0.0985 + temp * 0.4623 + LightSnowRain * -0.2857 + MistCloudy * -0.0754 + spring * -0.1258 + winter * 0.0587

From the analysis of categorical variables, we can understand that:

1. **Year (yr):** This variable has a positive value (0.2401), meaning bike demand goes up over time. People may be using bikes more because they're aware of things like pollution or just find it convenient.

2. **Holiday:** The negative value (-0.0985) means bike demand is lower on holidays. This might be because fewer people are going to work or school, so they don't need bikes as much.

3. **Weather (LightSnowRain and MistCloudy):**
    a. **LightSnowRain** has a strong negative value (-0.2857), which shows bike demand drops when it's snowy or rainy. It makes sense because bad weather isn't good for biking.
    b. **MistCloudy** also has a negative value (-0.0754), though not as strong. So, bike demand slightly drops when it's misty or cloudy.

4. **Season (spring and winter):**
    a. **Spring** has a negative value (-0.1258), which means bike demand is lower in spring.
    b. **Winter** has a small positive value (0.0587), so bike demand goes up a bit in winter.

In short, bike demand is higher over the years, lower on holidays, drops in bad weather, and changes a little with seasons.

Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 2 goes below this line> (Do not edit)

Using *"drop_first = True"* when creating dummy variables is important to avoid something called "*the dummy variable trap*".

When we create dummy variables, each category of a categorical variable gets its own column with values of 0 or 1. If we don't drop the first column, all these columns together add up to the original variable, creating perfect multicollinearity (a situation where one variable can be perfectly predicted by others). This confuses the model and makes it hard to accurately estimate the effects of each variable.

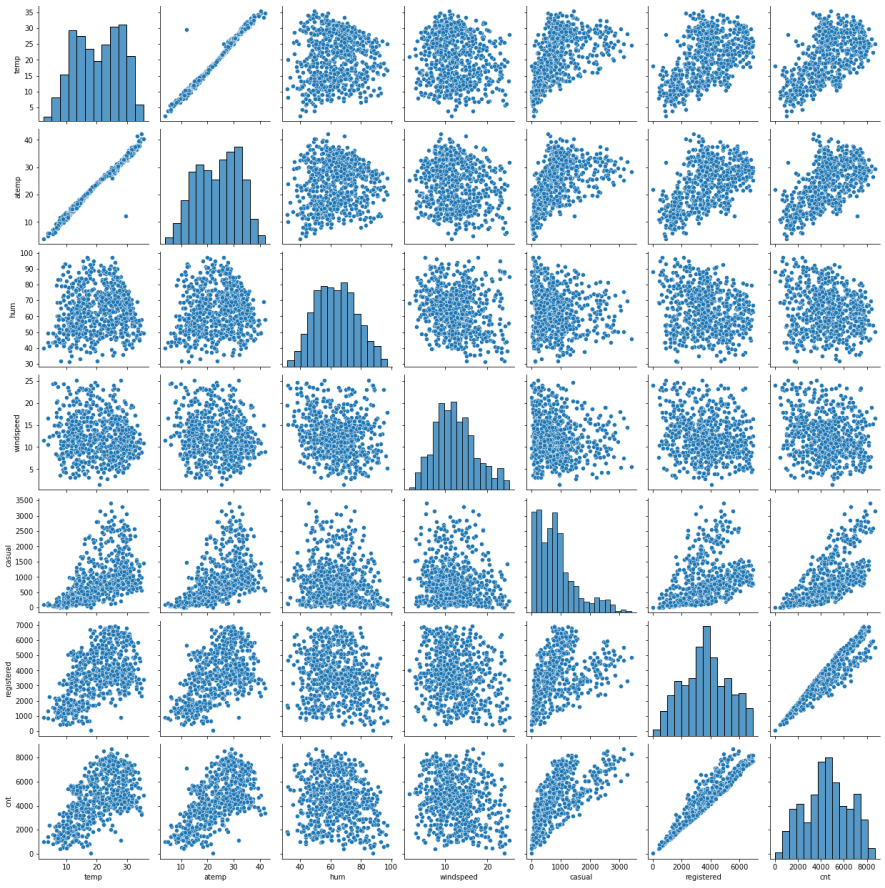
By setting *"drop_first = True"*, we drop one dummy column, which means we pick one category as a "reference" group. The model will compare all other categories to this reference group instead of

including all categories at once. This makes the model simpler and helps avoid errors due to multicollinearity.

So, using *"drop_first = True"* is a way to make sure our model doesn't have redundant data, which keeps the predictions more stable and reliable.

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)
Total Marks: 1 mark (Do not edit)
Answer: <Your answer for Question 3 goes below this line> (Do not edit)

Target Variable (cnt) has high correlation with Registered, Casual and Temp. The Highest correlation is with Registered.

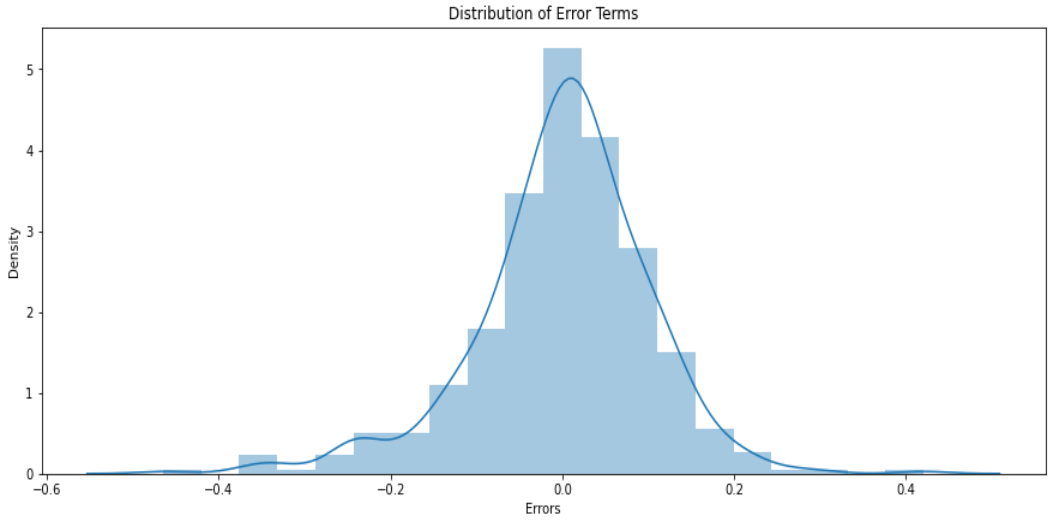


Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: <Your answer for Question 4 goes below this line> (Do not edit)

1. Multi-Collinearity:
   Verified and Validated that each predictor variable in the model doesn't have multi-collinearity. This was done using Variance Inflation Factor (VIF). After dropping all the high VIF values, the final values are as below:

	Features	VIF	P-Value
2	temp	2.24	0.000
0	yr	2.07	0.000
4	MistCloudy	1.53	0.000
6	winter	1.32	0.000
5	spring	1.25	0.000
3	LightSnowRain	1.06	0.000
1	holiday	1.04	0.001

2. Verified if the Error Terms are normally distributed, since this is one of the assumptions of Linear Regression. Observed that there is a linear relationship between predictor and the target variables and that there is a constant variance of residuals at all levels of x.



Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 5 goes below this line> (Do not edit)

The final model predicts the Demand through this equation:

Final Model Equation = const * 0.1724 + yr * 0.2401 + holiday * -0.0985 + temp * 0.4623 + LightSnowRain * -0.2857 + MistCloudy * -0.0754 + spring * -0.1258 + winter * 0.0587

Looking at the final model equation, the top 3 features with the highest absolute values of coefficients (positive or negative) are:

1. **Temperature (temp) with a coefficient of 0.4623**: This shows temperature has the biggest impact on bike demand. When it's warmer, demand goes up, which makes sense since people find it easier to bike in nice weather.

2. **LightSnowRain with a coefficient of -0.2857**: This has a strong negative impact, meaning that

when there's light snow or rain, bike demand drops a lot because of bad weather.

3. **Year (yr) with a coefficient of 0.2401**: This positive coefficient means that demand is rising over time, suggesting that bike-sharing has become more popular each year.

So, temperature, weather (light snow/rain), and year are the main factors affecting bike demand in this model.

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Linear Regression is a basic algorithm used to predict a continuous value based on one or more input features. Here's how it works:

1. Purpose of Linear Regression:
The goal of linear regression is to find the best-fitting line through the data points that predicts the dependent variable (like bike demand) based on the values of one or more independent variables (like temperature, year, etc.).

2. Line Equation:
In simple linear regression (one feature), the line equation is,

$$y = b_0 + b_1 \cdot x$$

where,
y is the predicted value.
x is the independent variable.
b_0 (intercept) and b_1 (slope) are parameters that the model learns.

In multiple linear regression (many features) it becomes,

$$y = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 + b_3 \cdot x_3 + \ldots\ldots\ldots\ldots\ldots + b_n \cdot x_n$$

where, each feature x_n has its own coefficient b_n

3. Finding the Best Line: The model tries to find values for b_0, b_1 and other coefficients that minimize the error between actual and predicted values. It does this by minimizing the sum of squared errors (SSE). This process is called **least squares**.

4. Gradient Descent: In some cases, especially with lots of data, the model uses a technique called gradient descent to adjust the coefficients step-by-step, moving towards the values that minimize the SSE.

5. Interpreting Coefficients: Each coefficient in the final equation shows how much yyy will change with a one-unit change in that feature, keeping all other features constant. A positive coefficient means the feature increases yyy, and a negative one means it decreases yyy.

6. Assumptions of Linear Regression:

a. Linearity: The relationship between features and the target is linear.
b. Independence: The residuals (errors) should be independent.
c. Homoscedasticity: Residuals have constant variance.
d. Normality: Residuals are normally distributed.

In summary, linear regression fits a line to predict values by finding the best coefficients for features, keeping assumptions in mind to ensure reliable predictions. It's simple but powerful for understanding relationships between variables.

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Anscombe's quartet is a group of four datasets created by statistician Francis Anscombe in 1973 to show the importance of data visualization. Each dataset has nearly identical statistical properties—like mean, variance, correlation, and regression line—yet they look very different when graphed.

1. The first dataset resembles a typical linear relationship.
2. The second dataset has a clear curve, not well-suited to a linear fit.
3. The third dataset has one outlier that heavily influences the regression line.
4. The fourth dataset has all points in a vertical line except one, showing minimal relationship.

Anscombe's quartet teaches that summary statistics alone can be misleading. Visualizing data helps spot patterns, outliers, and relationships that numbers might hide, making it essential for accurate data analysis.

Question 8. What is Pearson's R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Pearson's R, or Pearson correlation coefficient, measures the strength and direction of a linear relationship between two variables. It ranges from -1 to +1:

=> +1 indicates a perfect positive linear relationship (as one variable increases, the other does too).
=> -1 indicates a perfect negative linear relationship (as one variable increases, the other decreases).
=> 0 means no linear relationship exists between the variables.

Pearson's R is calculated by comparing how each variable deviates from its mean, standardized by their variances. It's widely used in statistics and data science to understand how two variables are related, but it only captures linear relationships.

Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Scaling is the process of adjusting the range of features in a dataset to a common scale, which is especially important for algorithms sensitive to the magnitude of features, like regression, k-nearest neighbors, and neural networks.

Why Scaling is Performed: When features have vastly different ranges, the algorithm may assign more importance to larger values, leading to biased results. Scaling helps ensure that each feature contributes equally to the model.

Normalized Scaling: This rescales features to a specific range, typically 0 to 1. It's useful when data needs to be bounded, like in image processing.

Standardized Scaling: This center features around the mean with unit variance, resulting in a mean of 0 and a standard deviation of 1. Standardization is ideal for algorithms that assume normal distributions, like PCA and linear regression.

Both techniques make models more efficient and accurate by equalizing feature importance.

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

A Variance Inflation Factor (VIF) becomes infinite when there is perfect multicollinearity among the independent variables in a regression model. This means that one or more of the independent variables can be perfectly predicted by other variables in the model.

In simpler terms, when there is exact linear dependence between variables, the model cannot distinguish between the effects of those variables. As a result, the variance of the estimated regression coefficients becomes extremely large, leading to an infinite VIF.

This happens because VIF is calculated as: $VIF = 1 / 1 - R^2$

where R^2 is the coefficient of determination when a feature is regressed against all other features. If R^2 is 1 (perfect correlation), the denominator becomes 0, causing the VIF to be infinite.

To fix this, we need to remove or combine the correlated variables.

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

A Q-Q plot (Quantile-Quantile plot) compares the distribution of a dataset to a theoretical distribution, like the normal distribution. If the points lie along a straight line, the data is approximately normally distributed.

Importance in Linear Regression:

1. Check Normality of Residuals: Linear regression assumes residuals are normally distributed. A Q-Q plot helps verify this.
2. Detect Outliers: Points away from the line indicate outliers.
3. Model Assumptions: Non-normal residuals can lead to unreliable regression results, so checking with a Q-Q plot is essential for accurate predictions.

In short, a Q-Q plot helps validate model assumptions and identify issues with the data.
