Assignment-based Subjective Questions

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

Categorical variables such as season, weather, and month have significant and varying impacts on the dependent variable. Seasonal patterns and weather conditions play a crucial role in influencing the count, with some seasons and poor weather conditions leading to decreases, while certain months like September and March contribute to increases. This understanding can be vital for making predictions or decisions based on these factors.

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

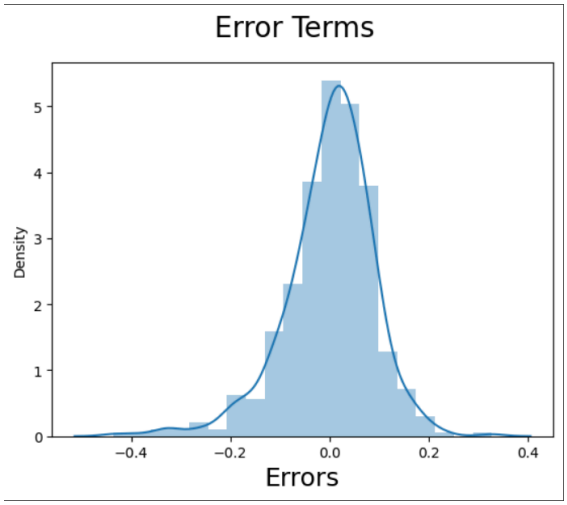
When we create dummy variables for a categorical variable with “n” categories, we end up with “n” dummy variables (one for each category). If we include all “n” dummy variables in a regression model along with the intercept (constant term), the sum of all the dummy variables will always equal 1. This introduces perfect multicollinearity because the sum of the dummy variables can perfectly predict the intercept (representing the baseline category).

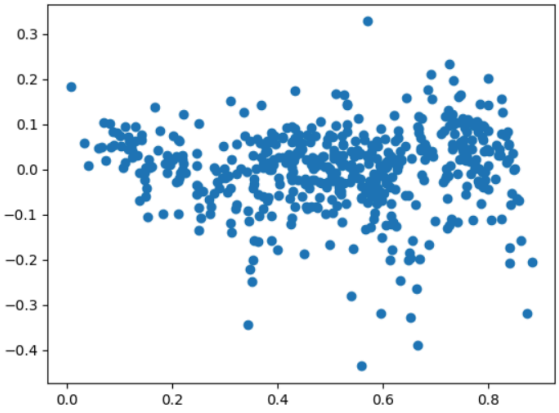
Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Temp has the highest correlation with the target variable among the numerical variables.

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

To validate the assumptions of the linear regression after building the model on the training set, the following were checked:

- 

- The residuals follow the normally distributed with a mean of 0.
- 

- plot of residuals versus predicted values should show a random scatter without any funneling or patterns.
- VIF values are less than 5 for all the variables it does not indicate any problematic multicollinearity.

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

Based on the final model's OLS regression results, the top three features contributing significantly towards explaining the demand for shared bikes are:

1. Year (yr): The coefficient is 0.2289, which indicates a strong positive influence on bike demand. The model suggests that bike demand increases significantly from one year to the next.
2. Temperature (temp): With a coefficient of 0.4563, temperature has a substantial positive impact on bike demand, meaning that as the temperature increases, the demand for shared bikes also increases.
3. Weather Situation 3 (weathersit_weathersit_3): The coefficient is -0.2645, indicating a strong negative impact on bike demand. This suggests that unfavorable weather conditions (such as heavy rain or snow) significantly decrease the demand for shared bikes.

These features have the largest absolute coefficients, showing their significant contributions to predicting bike demand.

General Subjective Questions

Explain the linear regression algorithm in detail.

Linear regression is a fundamental statistical and machine-learning technique used for modeling the relationship between a dependent variable and one or more independent variables. The main objective of linear regression is to predict the value of the dependent variable based on the values of the independent variables.

Types of Linear Regression

1. **Simple Linear Regression**: Involves one independent variable and one dependent variable.
2. **Multiple Linear Regression**: Involves multiple independent variables predicting a single dependent variable.

Assumptions of Linear Regression

1. **Linearity**: The relationship between the independent and dependent variables should be linear.
2. **Independence**: Observations should be independent of each other.

3. **Homoscedasticity**: The residuals (errors) should have constant variance at every level of X.
4. **Normality**: The residuals should be normally distributed.
5. **No Multicollinearity**: Independent variables should not be highly correlated with each other.

Steps in Linear Regression

1. Data Collection
2. Data Preprocessing
3. Model Building
4. Model Evaluation
5. Model Validation
6. Prediction

Explain the Anscombe's quartet in detail.

Anscombe's quartet is a collection of four datasets that are designed to illustrate the importance of graphing data before analyzing it and to demonstrate that summary statistics alone can be misleading. Each of the four datasets in Anscombe's quartet has nearly identical simple descriptive statistics, such as mean, variance, correlation, and linear regression lines, but they differ significantly in their graphical representation.
Anscombe's quartet underscores the need for comprehensive data analysis that goes beyond just calculating summary statistics. By graphing data and considering the broader context.

What is Pearson's R?

Pearson's R, also known as the Pearson correlation coefficient, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It is one of the most widely used methods for assessing correlations.

Example: Imagine you have two variables: hours studied and exam scores. If Pearson's R-value is 0.85, it indicates a strong positive linear relationship, meaning that as the number of hours studied increases, the exam scores tend to increase as well.

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

Scaling is a process in data preprocessing where numerical features are transformed to ensure that they operate on a similar scale. This is important because many machine learning algorithms are sensitive to the scale of the input data. Scaling can improve the performance of models and ensure that different features contribute equally to the analysis.

Scaling is performed for the following reasons:

- Algorithms like gradient descent (used in linear regression, logistic regression, neural networks, etc.) and distance-based algorithms (like k-nearest neighbors and k-means clustering) perform better when features are on a similar scale. Scaling ensures that each feature contributes proportionally to the calculation of distances or gradients.
- For algorithms that use iterative optimization methods, scaling can speed up convergence. Features with different scales can cause the algorithm to converge slowly or get stuck in suboptimal solutions.
- Scaling can make the data easier to interpret, especially when comparing feature importance or coefficients in models like linear regression.

The difference between normalization and standardization is that while normalization brings all the data points in a range between 0 and 1, standardization replaces the values with their Z scores.

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

Reasons for infinite VIF

- If one or more predictor variables are perfectly linearly dependent on other predictor variables, the matrix used to compute the VIF becomes singular (non-invertible). This causes the VIF to be infinite because the formula for VIF involves the inverse of this matrix.
- Even if the correlation between predictors is not perfect, very high correlations can lead to very large VIF values. This is because the presence of multicollinearity inflates the variance of the estimated coefficients.
- In cases of computational precision errors, the matrix inversion might result in extremely large VIF values. This can happen in some numerical calculations when dealing with very high or low magnitudes in the data.

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

A Q-Q plot (Quantile-Quantile plot) is a graphical tool used to assess if a dataset follows a specific theoretical distribution, often the normal distribution. It compares the quantiles of the dataset against the quantiles of the theoretical distribution.

Use and Importance in Linear Regression:

1. **Normality Check**: In linear regression, the residuals (errors) should ideally be normally distributed for the standard errors and confidence intervals to be valid. A Q-Q plot helps visually check if the residuals follow a normal distribution.

2. **Model Assumptions**: By examining the Q-Q plot of residuals, we can verify the assumption of normality, which is crucial for reliable hypothesis testing and the accuracy of confidence intervals for predictions.
3. **Diagnosing Problems**: If the points on the Q-Q plot deviate significantly from the reference line, it indicates that the residuals may not be normally distributed, suggesting potential problems in the model fit or the presence of outliers.

In summary, a Q-Q plot is a useful diagnostic tool in linear regression to assess whether the residuals are normally distributed, which is important for validating model assumptions and ensuring the reliability of statistical inferences.