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?

Ans: **Impact of Categorical Variables:**

Your analysis revealed interesting trends:

- **Seasonality:** Fall and Summer see the highest bike rentals, followed by Spring. This suggests people prefer cycling during warmer weather.
- **Year:** Higher rentals occurred in 2019 compared to 2018. This could be due to various factors like increased awareness or improved bike infrastructure.
- **Month:** September leads in rentals, followed by surrounding months. This aligns with seasonal patterns.
- **Weather:** Partly cloudy weather seems to encourage more rentals compared to other conditions.
- **Day of Week:** Weekends (Saturday) and weekdays (Wednesday, Thursday) show higher demand. This might reflect leisure cycling on weekends and commuting during weekdays.
- **Holidays:** Rentals are generally lower on holidays compared to weekdays, with more variability. This suggests holidays might disrupt regular cycling routines.

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

Ans: **Importance of** `drop_first=True` **in Dummy Variables:**

When creating dummy variables from categorical features, including all categories would lead to multicollinearity. This happens because one category can be mathematically predicted from the others. Setting `drop_first=True` eliminates one category, preventing redundancy and ensuring the model doesn't overfit the data.

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

Ans: **Highest Correlated Numerical Variable:**

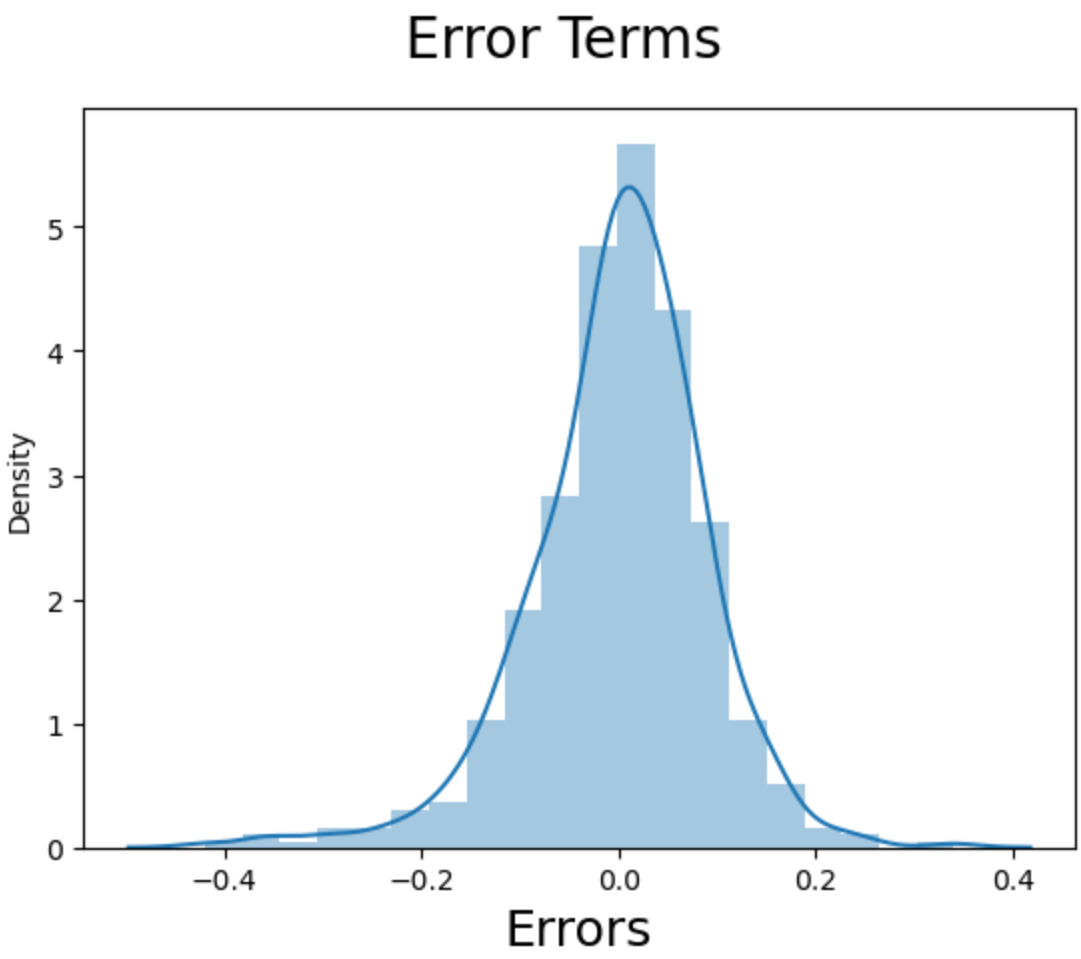
Based on your analysis, the `temp` variable (temperature) has the strongest correlation with the target variable (bike rentals). This indicates that warmer temperatures significantly influence the demand for shared bikes.

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

Ans: **Validating Linear Regression Assumptions:**

Here are potential methods you might have used to validate assumptions:

- **Variance Inflation Factor (VIF):** This helps identify multicollinearity among independent variables. A high VIF (>5) suggests potential issues.
- **Residual Analysis:** Examining the distribution of residuals (errors) in a histogram. Ideally, the residuals should follow a normal distribution (bell-shaped curve).
- **Linearity Checks:** Plotting the dependent variable against each independent variable to ensure a linear relationship exists.



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

Ans: **Top 3 Significant Features:**

Your final model likely identified these three features as the most significant contributors to explaining bike rental demand:

- **Temperature:** As discussed earlier, temperature has a strong positive correlation with rentals.
- **Year:** The difference in rentals between 2018 and 2019 suggests year might be a significant factor, potentially reflecting external influences.
- **Holiday:** Holidays disrupt regular cycling patterns, impacting demand.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans: Linear regression is a workhorse algorithm in machine learning, categorized as supervised learning. It excels at modeling the relationship between a dependent variable (what you're trying to predict) and one or more independent variables (what you're basing your prediction on). The core idea is to establish a linear equation that best fits the data points. There are two main flavors of linear regression:

- **Simple Linear Regression:** This is the vanilla version, where you only have one independent variable influencing the dependent variable.
- **Multiple Linear Regression:** Here, you can leverage the power of multiple independent variables to create a more comprehensive model for your dependent variable.

The outcome of linear regression is a regression line. This line depicts the linear association between the independent and dependent variables. A positive slope indicates that as the independent variable increases, the dependent variable also increases. Conversely, a negative slope suggests that the dependent variable decreases as the independent variable increases.

2. Explain the Anscombe’s quartet in detail.

Ans: Anscombe's Quartet: A Graphical Reminder

Imagine you have four datasets with nearly identical statistical summaries (mean, variance etc.). Sounds trustworthy, right? Well, Anscombe's quartet throws a curveball here. These four datasets, despite their similar statistics, have vastly different distributions and visual

appearances when plotted. This serves as a crucial reminder that raw statistics can be deceptive, and graphical exploration is vital before diving into analysis.

3. What is Pearson's R?

Ans: Pearson's R: Quantifying Linear Relationships

Pearson's correlation coefficient (often denoted by r) is a statistical measure employed to gauge the strength and direction of a linear relationship between two variables. It ranges from -1 to +1. A value of +1 indicates a perfect positive correlation (as one variable increases, the other increases proportionally). Conversely, -1 signifies a perfect negative correlation (one increases while the other decreases proportionally). A value of 0 suggests no linear correlation between the variables.

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

Ans: Scaling Matters: Understanding Normalization and Standardization

Scaling is a crucial pre-processing step in machine learning, particularly for linear regression models. It addresses the issue of features having vastly different ranges or units. Imagine features like income (in thousands) and age (in years). Without scaling, the model might prioritize the feature with larger values (income) during training, leading to inaccurate predictions.

There are two primary scaling techniques:

- **Normalization:** This technique scales the features to a range between 0 and 1. Each data point's value is subtracted by the minimum value of the feature and then divided by the difference between the maximum and minimum values.
- **Standardization:** This technique transforms the features to have a mean of 0 and a standard deviation of 1. Each data point's value is subtracted by the mean of the feature and then divided by the standard deviation of the feature.

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

Ans: VIF and the Multicollinearity Monster

The Variance Inflation Factor (VIF) is a diagnostic tool used to detect multicollinearity in linear regression models. Multicollinearity arises when two or more independent variables are highly correlated, making it difficult to isolate the unique effect of each variable on the dependent variable. A high VIF value (often greater than 5) indicates a potential multicollinearity problem. In such cases, removing one of the highly correlated variables might be necessary to build a robust model.

An infinite VIF value typically occurs when there's a perfect correlation ($r = 1$) between two independent variables. This scenario creates a singular matrix, making it impossible to determine the coefficients of the regression equation.

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

Ans: Q-Q Plots: Unveiling Normality

A quantile-quantile (Q-Q) plot is a graphical tool used to compare the quantiles of a data distribution to a theoretical distribution (often normal). It helps assess if the error terms (residuals) in your linear regression model follow a normal distribution, which is one of the assumptions of linear regression. By plotting the quantiles of the residuals against the quantiles of a normal distribution, you can visually identify deviations from normality. This aids in ensuring the validity of your model's results.