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

Based on the provided analysis of the categorical variables in the dataset, here are some potential inferences about their effects on the dependent variable (presumably the count of bike rentals):

1. **Year (yr):** The coefficient for the year variable suggests that there is a significant positive effect on bike rentals over time. This indicates that, on average, the number of bike rentals increases from one year to the next.

2. **Season (season_spring, season_winter)**: The negative coefficient for spring suggests that, on average, there are fewer bike rentals during spring compared to the reference season (possibly fall). Conversely, the positive coefficient for winter implies that there are more bike rentals during winter compared to the reference season.
3. **Month (mnth_jul, mnth_sept)**: The negative coefficient for July suggests that, on average, there are fewer bike rentals in July compared to the reference month (possibly January). However, the positive coefficient for September implies that there are more bike rentals in September compared to the reference month.
4. **Day of the Week (weekday_sun)**: The negative coefficient for Sunday suggests that, on average, there are fewer bike rentals on Sundays compared to the reference day (possibly Monday).
5. **Weather Situation (weathersit_bad, weathersit_moderate)**: The negative coefficients for bad weather and moderate weather indicate that, on average, there are fewer bike rentals during these weather conditions compared to the reference weather condition (possibly clear weather).

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

It is important to use **drop_first=True** during dummy variable creation to avoid multicollinearity issues in the regression model. By dropping the first category, we ensure linear independence among the dummy variables, which helps prevent multicollinearity, making the interpretation of the coefficients more straightforward and avoiding redundancy in the model. This approach also prevents the dummy variable trap, where the presence of redundant variables can lead to inaccurate and unstable coefficient estimates.

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

To determine which numerical variable has the highest correlation with the target variable, you would typically look at the correlation coefficients between each numerical variable and the target variable (often referred to as the Pearson correlation coefficient). The numerical variable with the highest absolute correlation coefficient value is the one with the highest correlation with the target variable. You

can identify this by plotting a correlation matrix or examining the correlation coefficients directly.

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

After building the linear regression model on the training set, you would typically validate the assumptions of linear regression through various diagnostic tests:
a. **Linearity**: You can assess linearity by plotting the observed values against the predicted values. A scatter plot should show a random distribution around a diagonal line.
b. **Homoscedasticity**: You can check for homoscedasticity by plotting the residuals against the predicted values. The plot should not exhibit any discernible pattern, indicating constant variance across all levels of the predictor variables.
c. **Normality of Residuals**: You can examine the distribution of residuals using a histogram or a Q-Q plot. The residuals should follow an approximately normal distribution.
d. **Independence of Residuals**: You can assess the independence of residuals by examining autocorrelation plots or Durbin-Watson statistics. Residuals should not show any patterns or trends.

5) 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, the top 3 features contributing significantly towards explaining the demand of the shared bikes are typically identified by examining the coefficients associated with each feature in the regression output. Features with larger absolute coefficient values indicate stronger contributions to the model. These features represent the variables that have the most substantial impact on the predicted demand for shared bikes.

1) Explain the linear regression algorithm in detail.

Linear Regression Algorithm: Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data. Here's a detailed explanation:

- **Assumption of Linearity**: Linear regression assumes that there is a linear relationship between the independent variables (predictors) and the dependent variable (response).
- **Model Representation**: In simple linear regression, the relationship between the dependent variable Y and one independent variable X is represented by the equation: $Y = \beta0 + \beta1*X + \varepsilon$, where $\beta0$ is the intercept, $\beta1$ is the coefficient for the independent variable, X is the independent variable, and ε is the error term.

- **Model Fitting**: The goal of linear regression is to find the best-fitting line (or hyperplane in multiple dimensions) that minimizes the sum of squared differences between the observed and predicted values. This is typically achieved using the method of least squares.
- **Estimation of Coefficients**: The coefficients (β0, β1, etc.) are estimated using mathematical techniques such as ordinary least squares (OLS), which minimize the sum of the squared residuals.
- **Model Evaluation**: Linear regression models are evaluated based on metrics such as R-squared, adjusted R-squared, mean squared error (MSE), and others to assess the goodness-of-fit and predictive accuracy.
- **Assumptions**: Linear regression assumes that the residuals (errors) are normally distributed, have constant variance (homoscedasticity), are independent of each other, and have a mean of zero

2) Explain the Anscombe's quartet in detail.

Anscombe's Quartet is a set of four datasets that have nearly identical statistical properties but different graphical representations. The datasets were created by Francis Anscombe to illustrate the importance of visualizing data before drawing conclusions. Despite having similar summary statistics (e.g., mean, variance, correlation), the datasets differ significantly when plotted, showcasing the limitations of relying solely on summary statistics and the importance of data visualization in understanding relationships and patterns.

3) What is Pearson's R?

Pearson's correlation coefficient (often denoted as r) measures the linear relationship between two continuous variables. It ranges from -1 to 1, where:

- r = 1 indicates a perfect positive linear relationship.
- r = -1 indicates a perfect negative linear relationship.
- r = 0 indicates no linear relationship. Pearson's R is calculated by dividing the covariance of the two variables by the product of their standard deviations. It is widely used to assess the strength and direction of the linear association between variables.

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

Scaling is the process of transforming data to a standard scale to ensure that different features are comparable and have a similar influence on the analysis. Scaling is performed to:

- **Normalize the Range**: Scaling ensures that all features have values within a similar range, preventing features with large magnitudes from dominating the model.
- **Improve Convergence**: Scaling can help improve the convergence rate of iterative optimization algorithms, such as gradient descent, by making the optimization landscape more uniform.

- **Enhance Model Performance**: Some machine learning algorithms, such as k-nearest neighbors (KNN) and support vector machines (SVM), are sensitive to the scale of features. Scaling helps these algorithms perform better.
- **Types of Scaling**:
    - **Normalized Scaling**: Normalization scales each feature to a range between 0 and 1. It is calculated as (x - min(x)) / (max(x) - min(x)).
    - **Standardized Scaling**: Standardization scales each feature to have a mean of 0 and a standard deviation of 1. It is calculated as (x - mean(x)) / std(x).

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

The Variance Inflation Factor (VIF) measures the degree of multicollinearity in a regression model. A VIF of infinity indicates perfect multicollinearity, where one or more independent variables can be perfectly predicted from the others. This often occurs when there is redundant information in the dataset, such as duplicate variables or linearly dependent predictors.

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

A Q-Q (Quantile-Quantile) plot is a graphical tool used to assess whether a set of data follows a specified distribution, such as the normal distribution. It compares the quantiles of the observed data against the quantiles of a theoretical distribution. The Q-Q plot is important in linear regression for:

- **Checking Normality Assumption**: By plotting the observed quantiles against the expected quantiles of a normal distribution, we can visually assess whether the residuals (errors) of the regression model are normally distributed.
- **Identifying Outliers**: Outliers in the data can be identified as points that deviate significantly from the diagonal line in the Q-Q plot, indicating potential violations of the assumption of normality.
- **Assessing Model Fit**: A well-fitting linear regression model should have residuals that follow a normal distribution, as indicated by a Q-Q plot where the points closely follow the diagonal line.