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 analysis of categorical variables, the following insights are derived:

- **Season:** Bike rentals are notably higher in summer and fall compared to spring and winter. This indicates that warmer and milder weather conditions drive higher bike rental demand. The higher mean counts in summer and fall suggest these seasons are more favorable for biking.
- **Month:** August and June exhibit the highest bike rental rates, while January shows the lowest. This suggests that bike rentals peak during warmer months and drop during colder periods, reflecting seasonal variations in bike usage.
- **Weekday:** Rentals are highest on Saturdays and lower on Mondays. This pattern suggests that weekends, particularly Saturdays, are associated with increased recreational activities, leading to higher bike rentals.
- **Weather Situation:** Clear weather significantly increases bike rentals, while conditions such as light snow and misty/cloudy weather decrease demand. This highlights the substantial impact of weather on bike rental preferences, with clear conditions encouraging more bike usage.
- **Year:** There is a noticeable increase in bike rentals from 2018 to 2019, indicating a growing trend in bike-sharing usage over time.
- **Working Day:** Bike rentals are slightly higher on working days compared to non-working days, though the difference is not very pronounced. This could be due to the mix of commuting and recreational use.
- **Holiday:** Bike rental rates are higher on non-holidays compared to holidays. This suggests that people may use bike-sharing services more frequently during regular days as opposed to holidays, possibly due to different activity patterns.

2. **Why is it important to use drop_first=True during dummy variable creation?**
   Using drop_first=True during dummy variable creation is crucial to prevent the issue of multicollinearity in regression models. Multicollinearity occurs when independent variables are highly correlated, leading to redundancy in the model. By dropping the first category, you create a reference category, avoiding perfect multicollinearity and ensuring that the dummy variables accurately represent the categorical variable without introducing redundancy.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**
   In the pair-plot analysis, the numerical variable with the highest correlation with the target variable is Temperature (temp). This indicates that temperature has the strongest relationship with bike rental demand among the numerical variables.

4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**
   To validate the assumptions of Linear Regression, the following steps were taken:

- **Linearity:** Checked scatter plots of residuals versus predicted values to ensure a linear relationship between predictors and the target variable. A linear pattern suggests that the model is appropriate for the data.
- **Normality of Residuals:** Evaluated the distribution of residuals using a Q-Q plot and a histogram. The residuals should be approximately normally distributed, which can be assessed visually through these plots.
- **Homoscedasticity:** Analyzed residuals versus fitted values to ensure that the variance of residuals is constant across all levels of the predictor variables. A consistent spread of residuals indicates homoscedasticity.
- **Independence:** Assessed the residuals for autocorrelation using statistical tests such as the Durbin-Watson test. Residuals should be independent of each other, with no discernible patterns or correlations.
- **Multicollinearity:** Examined variance inflation factors (VIFs) for each predictor to ensure that predictors are not highly correlated with each other, which could inflate the variance of coefficient estimates.

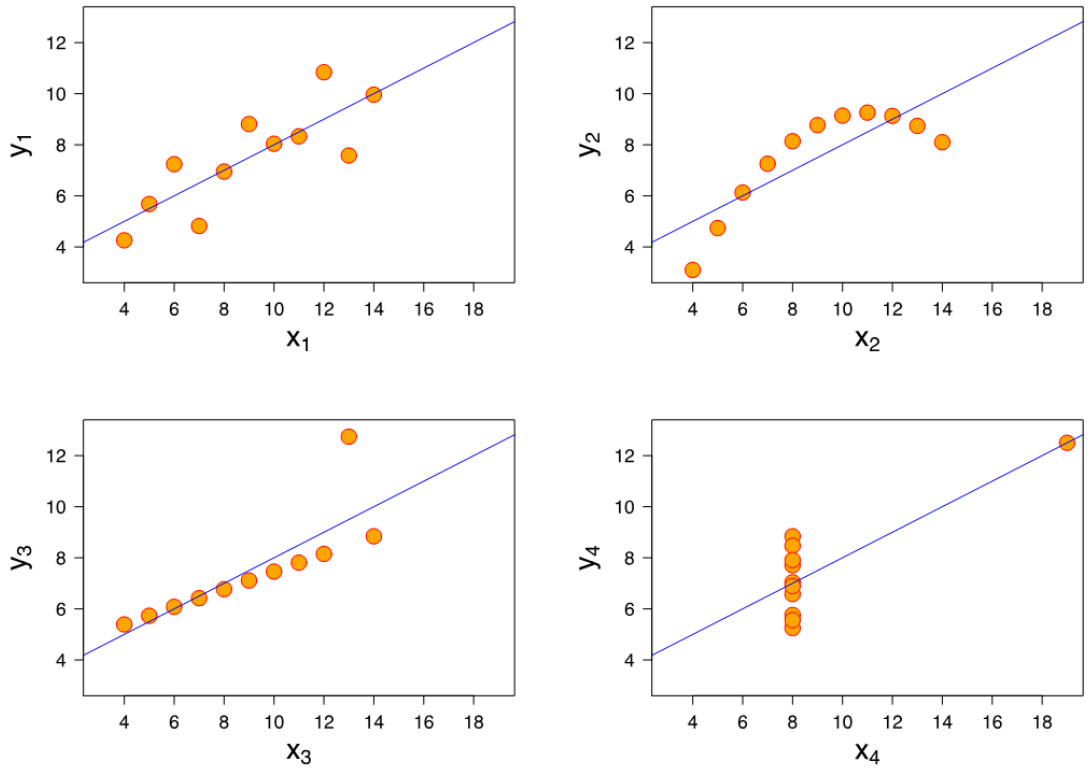
5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand for shared bikes?**
   Based on the regression equation, the top 3 features contributing significantly to bike demand are:

- **Temperature (temp):** With a coefficient of 0.4777, temperature has the highest positive impact on bike demand. Each degree increase in temperature results in a notable increase in bike rentals.
- **Year (yr):** With a coefficient of 0.2341, this feature reflects a positive trend in bike demand over time. Each additional year increases bike demand, indicating growing popularity.
- **Light Snow (Light Snow):** With a coefficient of -0.2850, light snow has the most significant negative impact on bike demand among the weather-related variables. Snowy conditions lead to a substantial decrease in bike rentals.

  These features are significant predictors of bike demand, with temperature and year contributing positively, while light snow decreases demand.

1. **Explain the linear regression algorithm in detail. (4 marks)**
   Answer: Linear regression is a technique used to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to the observed data.
   - **Objective:** The main goal is to understand and predict the dependent variable (Y) based on the independent variables (X).
   - **Simple Linear Regression:** This involves one independent variable. The model is represented as:
     - **Equation:** $Y = \beta0 + \beta1X + \varepsilon$
       - Y is the dependent variable.
       - β0 is the intercept (the value of Y when X is zero).
       - β1 is the slope (how Y changes with X).
       - X is the independent variable.
       - ε is the error term (difference between observed and predicted Y).
   - **Multiple Linear Regression:** This extends the model to multiple independent variables. The equation is:
     - **Equation:** $Y = \beta0 + \beta1X1 + \beta2X2 + ... + \betanXn + \varepsilon$
       - X1, X2, ..., Xn are independent variables.
       - β1, β2, ..., βn are their corresponding coefficients.
   - **Fitting the Model:** The model is fit to data using Ordinary Least Squares (OLS). OLS finds the coefficients that minimize the sum of squared differences between the observed and predicted values:
     - **Objective:** Minimize the sum of $(Y_i - \hat{Y}_i)^2$, where $\hat{Y}_i$ is the predicted value for the i-th observation.
   - **Assumptions:**
     - **Linearity:** The relationship between Y and X is linear.
     - **Independence:** Observations are independent.
     - **Homoscedasticity:** Constant variance of residuals across all levels of X.
     - **Normality:** Residuals are normally distributed.
   - **Evaluation Metrics:**
     - **R-squared:** Represents the proportion of variance in Y explained by X. Values range from 0 to 1; higher values indicate a better fit.
     - **Mean Squared Error (MSE):** Measures the average squared difference between observed and predicted values. Lower MSE indicates a better model fit.

2. **Explain Anscombe's quartet in detail. (3 marks)**
   Answer: Anscombe's Quartet is a collection of four datasets that have identical statistical properties but different distributions when graphed.
   - **Purpose:** To show that summary statistics alone can be misleading and that visualizing data is crucial for understanding its distribution.
   - **Datasets:** Each dataset in the quartet has nearly the same means, variances, and Pearson correlation coefficient:



-
    - **Dataset 1:** Exhibits a clear linear relationship.
    - **Dataset 2:** Shows a non-linear pattern.
    - **Dataset 3:** Contains a significant outlier that skews the regression line.
    - **Dataset 4:** Shows a vertical line with one outlier affecting the regression analysis.
- **Importance:** Highlights the necessity of graphical analysis to detect underlying patterns and anomalies that summary statistics alone might not reveal.

3. **What is Pearson's R? (3 marks)**
   Answer: Pearson's R, or Pearson correlation coefficient, measures the strength and direction of the linear relationship between two continuous variables.
    - **Definition:** Quantifies how closely the data points fit a linear trend. It is calculated as:
        - **Formula:** r = cov(X, Y) / (σ_X * σ_Y)
            - cov(X, Y) is the covariance between X and Y.
            - σ_X and σ_Y are the standard deviations of X and Y.
    - **Range:**
        - **1:** Perfect positive linear relationship.
        - **-1:** Perfect negative linear relationship.
        - **0:** No linear relationship.
    - **Interpretation:** Helps to understand the degree of linear association between variables, providing insight into their relationship.

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)**
   Answer: Scaling adjusts the range and distribution of feature variables to ensure they contribute equally to the analysis, particularly in machine learning algorithms.
   - **Purpose of Scaling:**
     - **Equal Contribution:** Prevents features with different scales from dominating the analysis.
     - **Algorithm Efficiency:** Can enhance performance and speed up convergence in algorithms using optimization techniques.
   - **Types of Scaling:**
     - **Normalization (Min-Max Scaling):** Adjusts features to a fixed range, usually [0, 1], using the formula:
       - **Formula:** X_norm = (X - min(X)) / (max(X) - min(X))
     - **Standardization (Z-score Normalization):** Centers data around the mean with unit variance using:
       - **Formula:** X_std = (X - μ) / σ
         - μ is the mean.
         - σ is the standard deviation.
   - **Difference:** Normalization scales features to a specific range, useful for algorithms sensitive to the scale of input features. Standardization transforms features to have zero mean and unit variance, which is helpful for algorithms assuming normally distributed data.

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)**
   Answer: An infinite Variance Inflation Factor (VIF) indicates perfect multicollinearity among predictor variables.
   - **Definition of VIF:** Measures how much the variance of a regression coefficient increases due to collinearity with other predictors.
     - **Formula:** VIF_i = 1 / (1 - R_i^2)
       - R_i^2 is the R-squared value from regressing the i-th predictor on other predictors.
   - **Cause of Infinite VIF:** Occurs when R_i^2 equals 1, meaning the predictor is perfectly linearly dependent on other predictors. This causes the regression matrix to be singular or nearly singular.
   - **Impact:** Results in unreliable or non-existent coefficient estimates due to perfect multicollinearity.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)**
   Answer: A Quantile-Quantile (Q-Q) plot is used to assess if a dataset follows a specified theoretical distribution, such as the normal distribution.
   - **Construction:** Compares the quantiles of the data against the quantiles of a theoretical distribution. If the data follows the theoretical distribution, the points will lie on a straight line.
   - **Use in Linear Regression:** Helps check if the residuals (errors) are normally distributed, which is a key assumption of linear regression.
   - **Importance:** Ensures model validity and helps detect issues with the normality of residuals. Deviations from the line may indicate violations of the normality assumption, leading to unreliable statistical inferences.
   -