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?**
   Analyzing categorical variables can reveal their impact on the dependent variable. For example, if a dataset contains a categorical variable like "Day of the Week" in a sales prediction model, you might find that sales are significantly higher on weekends than on weekdays. This suggests that the "Day of the Week" variable influences the sales (dependent variable) differently across its categories.

2. **Why is it important to use drop_first=True during dummy variable creation?**
   Using drop_first=True during dummy variable creation avoids multicollinearity, which occurs when one of the dummy variables can be predicted from the others. For instance, if you have a categorical variable "Color" with categories "Red," "Blue," and "Green," creating dummy variables without dropping one would result in redundancy. If you drop the "Red" category, you can infer its presence when both "Blue" and "Green" are 0, thus avoiding multicollinearity.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**
   Suppose you have a dataset where you're trying to predict house prices (target variable). After plotting pair-plots between numerical variables like "Square Footage," "Number of Rooms," and "Age of the House," you notice that "Square Footage" has the steepest linear relationship with house prices. This would indicate that "Square Footage" has the highest correlation with the target variable.

4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**
   To validate linear regression assumptions, you might:
   - Check for **linearity** by plotting residuals against predicted values to ensure there is no pattern.
   - Check for **homoscedasticity** by ensuring that the variance of residuals is constant across all levels of the independent variables.
   - Check for **independence** by examining the Durbin-Watson statistic for autocorrelation.
   - Check for **normality** of residuals using a Q-Q plot, where residuals should follow a straight line if they are normally distributed.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**
   If the final model includes features like "Temperature," "Hour of the Day," and "Day of the Week," and the coefficients indicate that "Temperature" has the largest positive effect, followed by "Hour of the Day" and "Day of the Week," these would be the top 3 features influencing bike demand. For example, higher temperatures and certain hours of the day (e.g., morning rush hour) might lead to higher bike demand.

General Subjective Questions:

1. **Explain the linear regression algorithm in detail.**
   Linear regression predicts the value of a dependent variable (Y) based on one or more independent variables (X). The algorithm fits a line (in simple linear regression) or a hyperplane (in multiple linear regression) to the data by minimizing the sum of squared differences between observed and predicted values (residuals). For instance, predicting house prices based on square footage, the linear equation might look like Price = 50,000 + 150*Square_Footage.

2. **Explain the Anscombe's quartet in detail.**
   Anscombe's quartet comprises four datasets that have nearly identical simple descriptive statistics (mean, variance, correlation, etc.), yet when plotted, they reveal very different relationships. This illustrates the importance of graphical analysis in statistics. For example, one of the datasets might show a linear relationship, another a perfect quadratic curve, and a third could have an outlier that distorts the statistical summary.

3. **What is Pearson's R?**
   Pearson's R measures the strength and direction of the linear relationship between two variables. For example, if you are analyzing the relationship between study hours and exam scores, and you calculate Pearson's R as 0.85, it indicates a strong positive correlation—meaning more study hours generally lead to higher exam scores.

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**
   Scaling adjusts data to a specific range or distribution. For instance, in a dataset with features like "Height" (in cm) and "Weight" (in kg), scaling ensures that one feature doesn't dominate because of its units. **Normalized scaling** transforms the data to fit within a range (e.g., 0 to 1), whereas **standardized scaling** adjusts data so that it has a mean of 0 and a standard deviation of 1. Normalization might be useful in image processing, where pixel values are scaled between 0 and 1, while standardization is common in algorithms like SVM.

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen?**
   An infinite VIF occurs when there is perfect multicollinearity, meaning one predictor variable is an exact linear combination of others. For example, if you include both "Monthly Income" and "Annual Income" as independent variables, the VIF for either will be infinite because they are perfectly correlated (Annual Income is 12 times Monthly Income).

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.**
   A Q-Q plot compares the quantiles of a dataset with the quantiles of a theoretical distribution, usually the normal distribution. It helps assess whether the residuals of a linear regression model are normally distributed—a key assumption. If the points in a Q-Q plot lie on a straight line, it suggests the residuals are normally distributed. For example, in a well-fitted linear regression model, the residuals' Q-Q plot should form a diagonal line, indicating that the residuals conform to normality.