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?

In the bike-sharing dataset, the categorical variable 'weathersit' was analyzed for its impact on the target variable 'cnt.' Through exploratory data analysis (EDA), it was observed that more bike rentals occurred during weather condition 1 (Clear, few clouds, partly cloudy). Similarly, the variables 'season' and 'yr' also influenced the target variable. Additionally, during model building, the inclusion of categorical variables such as 'yr' and 'seasons' resulted in a significant increase in the R-squared and adjusted R-squared values, indicating that these categorical variables helped explain more of the dataset's variance.

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

When creating dummy variables, it's important to set drop_first=True to prevent redundant features. Without this option, one dummy variable might act as a reference group, causing multicollinearity since the first column becomes a baseline for the others.

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

Among the numerical variables, 'temp' (temperature) has the highest correlation with the target variable 'count.' Since 'temp' and 'atemp' were highly correlated, 'atemp' was dropped to avoid multicollinearity.

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, several steps were taken to validate the model:

- The distribution of residuals was checked using a plot to ensure they followed a normal distribution.

- Independent variables were added or removed based on their VIF (Variance Inflation Factor) and p-values to manage multicollinearity.

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

The three most significant features that explain the demand for shared bikes are:

- **Temperature** : The perceived temperature in Celsius.
- **Year** : Differentiating between the years 2018 and 2019.
- **Holiday** : Whether the day is a holiday or not.

General Subjective Questions:

1. Explain the linear regression algorithm in detail.

Linear regression is a fundamental machine learning algorithm used to predict a continuous dependent variable based on one or more independent variables. The model fits a regression line, defined by the equation

$$Y = a + bX + e,$$

where:

- a is the intercept,
- b is the slope, and
- e is the error term.

The goal is to minimize the difference between the actual and predicted values.

2. Explain Anscombe's quartet in detail.

Anscombe’s quartet comprises four datasets with nearly identical statistical properties (mean, variance, correlation, etc.), but they appear very different when graphed. Created in 1973 by Francis Anscombe, the quartet highlights the importance of visualizing data before analysis and the influence of outliers.

3. What is Pearson's R?

- Pearson's R is a numerical measure indicating the strength of the linear relationship between two variables, ranging from -1 to +1.
- A positive correlation coefficient suggests that both variables tend to increase or decrease together.
- A negative correlation coefficient indicates that when one variable increases, the other decreases, and vice versa.
- An r value of 1 signifies a perfect positive linear relationship (both variables change in the same direction).
- An r value of -1 indicates a perfect negative linear relationship (the variables change in opposite directions).
- An r value of 0 means there is no linear relationship between the variables.
- An r value between 0 and 0.5 suggests a weak correlation.
- An r value between 0.5 and 0.8 indicates a moderate correlation.
- An r value above 0.8 represents a strong correlation.

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

- Scaling adjusts the range of independent variables.
- It is necessary to ensure variables are on comparable scales, especially for models sensitive to magnitudes.
- Normalization scales values between 0 and 1, whereas standardization adjusts values to have a mean of 0 and a standard deviation of 1.

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

An infinite VIF (Variance Inflation Factor) indicates severe collinearity between variables. This occurs when one predictor variable can be expressed as a linear combination of other predictor variables, leading to multicollinearity.

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 used to assess if a dataset comes from a theoretical distribution, such as a normal distribution. In linear regression, a Q-Q plot can be used to check whether the residuals follow a normal distribution, helping to validate the assumption of normality in the error terms.