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. From the analysis of categorical variables, we can infer their effect on the dependent variable by examining their coefficients in the regression model. Categorical variables with statistically significant coefficients indicate that they have a significant effect on the dependent variable. Additionally, comparing the magnitude and direction of coefficients helps understand the strength and direction of the relationship between each categorical variable and the dependent variable.

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

Ans. It is important to use *`drop_first=True`* during dummy variable creation to avoid multicollinearity issues. When creating dummy variables, dropping the first category helps prevent perfect multicollinearity, where one dummy variable can be perfectly predicted from the others, improving the interpretability of the model coefficients.

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

Ans. Looking at the pair-plot among the numerical variables, the one with the highest correlation with the target variable is identified by observing the scatter plots between each numerical variable and the target variable. The numerical variable with the scatter plot exhibiting the strongest linear relationship with the target variable has the highest correlation.

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

Ans. After building the linear regression model on the training set, the assumptions of linear regression are validated through various diagnostic tests, including checking for linearity, normality of residuals, homoscedasticity, and independence of residuals. Techniques such as residual analysis, Q-Q plots, scatter plots of residuals against predicted values, and statistical tests like Jarque-Bera test can be employed to validate these assumptions.

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

Ans. Based on the final model, the top 3 features contributing significantly towards explaining the demand for shared bikes are identified by examining the coefficients associated with each feature in the regression model. Features with larger absolute coefficients indicate a stronger impact on the dependent variable, hence contributing significantly to explaining the demand for shared bikes.

General Subjective Questions:

1. Explain the linear regression algorithm in detail.
Ans. Linear regression algorithm is a statistical method used for modeling the relationship between one or more independent variables and a dependent variable by fitting a linear equation to observed data. It aims to find the best-fitting line that minimizes the sum of squared differences between the observed and predicted values. The algorithm assumes a linear relationship between the independent and dependent variables, and it estimates the coefficients of the linear equation using the least squares method.

2. Explain the Anscombe's quartet in detail.
Ans. Anscombe's quartet comprises four datasets that have nearly identical simple descriptive statistics but differ greatly when graphed. It highlights the importance of visualizing data and the limitations of relying solely on summary statistics. Despite having similar means, variances, correlations, and linear regression lines, the datasets have different distributions and relationships between variables when plotted.

3. What is Pearson's R?
Ans. Pearson's correlation coefficient (Pearson's R) measures the linear relationship between two continuous variables. It ranges from -1 to 1, where 1 indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and 0 indicates no linear relationship. Pearson's R is sensitive to outliers and assumes a linear relationship between variables.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
Ans. Scaling is the process of transforming data to a standard scale to ensure that all features contribute equally to the analysis. It is performed to prevent features with larger scales from dominating the modeling process. Normalized scaling rescales the data to a range of [0, 1], while standardized scaling transforms the data 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?
Ans. The value of VIF (Variance Inflation Factor) becomes infinite when perfect multicollinearity exists between predictor variables. Perfect multicollinearity occurs when one predictor variable is a perfect linear function of other predictor variables, making it impossible to estimate unique coefficients for each predictor variable. VIF measures the extent of multicollinearity in regression analysis, and an infinite VIF indicates that one or more predictor variables can be perfectly predicted from other variables.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
Ans. A Q-Q plot (Quantile-Quantile plot) is a graphical tool used to assess whether a dataset follows a specific probability distribution, such as the normal distribution. It compares the quantiles of the observed data against the quantiles of a theoretical distribution. In linear regression, Q-Q plots help validate the assumption of normality of residuals by visually inspecting whether the residuals follow a straight line pattern, indicating normality. Departures from linearity suggest deviations from normality.