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? (3 marks)

   Ans - Categorical variables impact on the dependent variable is inferred from regression coefficients, positive coefficients indicate increased dependent variable values with category presence, while negative coefficients imply the opposite, with significance gauged by low p-values.

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

   ANS- Using **drop_first=True** during dummy variable creation mitigates multicollinearity by eliminating redundant information from the dropped category, ensuring stable parameter estimates in regression analysis.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

   ANS - holiday

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

   training set? (3 marks)

   Ans - Assumptions of Linear Regression are validated by examining residuals for normality, linearity, and homoscedasticity using residual plots, along with checking for multicollinearity through Variance Inflation Factor (VIF) analysis.

5. Based on the final model, which are the top 3 features contributing significantly towards

explaining the demand of the shared bikes?

Ans - holiday , instant , atemp

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans - Linear regression algorithm models the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data. It estimates the coefficients of the linear equation using the method of least squares.

2. Explain the Anscombe's quartet in detail.

   Ans - Anscombe's quartet consists of four datasets with identical statistical properties, yet they exhibit very different patterns when plotted. This highlights the importance of visualizing data and not relying solely on summary statistics.

3. What is Pearson’s R?

Ans - 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.

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

   and standardized scaling? (3 marks)

Ans - Scaling is the process of transforming data to a common scale. It's performed to ensure that features with different units and scales contribute equally to the analysis. Normalized scaling scales the data to a range of 0 to 1, while standardized scaling scales 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? (3 marks)

Ans - VIF becomes infinite when there is perfect multicollinearity between predictor variables, meaning one predictor variable can be perfectly predicted by another. This makes it impossible to estimate the coefficient of the affected predictor variable.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Ans - A Q-Q plot is a graphical tool to assess if a set of data follows a specific distribution, such as the normal distribution. In linear regression, Q-Q plots are used to check if the residuals of the model are normally distributed, which is an assumption of linear regression. It helps validate the model's accuracy and reliability.