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)**

Categorical variables such as season, yr, mnth, holiday, weekday, workingday, and weathersit significantly affect bike demand. For instance, bike rentals are higher during certain seasons (season), and weekends (weekday) tend to see more rentals compared to weekdays. Additionally, weather conditions (weathersit) play a crucial role, with adverse weather leading to fewer rentals.

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

Using drop_first=True helps avoid multicollinearity by dropping one category from each categorical variable. This ensures that the dummy variables are independent and prevents the "dummy variable trap," where one category can be perfectly predicted from others.

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

The variable temp (temperature) has the highest correlation with the target variable cnt (total bike rentals), indicating that higher temperatures are associated with increased bike demand.

4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**

Linearity: Checked scatter plots between predicted values and residuals to ensure no patterns.

Normality: Used Q-Q plots to confirm that residuals follow a normal distribution.

Homoscedasticity: Ensured residuals have constant variance by inspecting residual plots.

Multicollinearity: Calculated VIF (Variance Inflation Factor) to ensure no multicollinearity among predictors.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)**

The top 3 features are temp (temperature), yr (year), and season_3 (summer season). These features have the highest coefficients, indicating their strong influence on bike demand.

General Subjective Questions

1. **Explain the linear regression algorithm in detail. (4 marks)**

Linear regression aims to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data. The equation is of the form y=β0+β1x1+β2x2+...+βnxn$y=\beta 0+\beta 1x1+\beta 2x2+...+\beta nxn$, where y$y$ is the dependent variable, xixi are the independent variables, and βiβi are the coefficients. The algorithm minimizes the sum of squared residuals (differences between observed and predicted values) to find the best-fitting line.

2. **Explain the Anscombe's quartet in detail. (3 marks)**

Anscombe's quartet consists of four datasets with nearly identical statistical properties (mean, variance, correlation) but different distributions and plots. It demonstrates the importance of visualizing data, as relying solely on summary statistics can be misleading.

3. **What is Pearson's R? (3 marks)**

Pearson's R, or Pearson correlation coefficient, measures the linear correlation between two 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)**

Scaling transforms data to a specific range or distribution. It is performed to ensure that all features contribute equally to the model. Normalized scaling rescales data to a range (e.g., 0 to 1), while standardized scaling transforms 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)**

An infinite VIF indicates perfect multicollinearity, meaning one predictor variable can be perfectly predicted by a linear combination of other predictors. This situation arises when there is redundant information in the predictors.

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

A Q-Q (quantile-quantile) plot compares the quantiles of the residuals to a theoretical normal distribution. It is used to check if the residuals are normally distributed, which is an assumption of linear regression. Deviations from the line in a Q-Q plot indicate non-normality in the residuals.