Assignment-based Subjective Questions

Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: <Your answer for Question 1 goes below this line> (Do not edit)

Categorical variables such as season, weathersit, holiday, and workingday significantly impact bike demand. For instance:
Season: Bike demand is highest during summer and fall, reflecting weather preferences for biking.
Weathersit: Clear weather conditions lead to increased demand, while adverse conditions (e.g., rain/snow) suppress it.
Holiday: Non-working days show higher casual usage compared to working days.
Workingday: Registered users dominate during working days, indicating a trend in daily commuting.

Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 2 goes below this line> (Do not edit)

Using drop_first=True ensures that one category of each categorical variable is used as the baseline (reference), avoiding the dummy variable trap (perfect multicollinearity). This helps the regression model interpret coefficients relative to the baseline category and prevents redundancy in the feature set.

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)
Total Marks: 1 mark (Do not edit)
Answer: <Your answer for Question 3 goes below this line> (Do not edit)

temp (normalized temperature) has the highest correlation with cnt (bike demand) at approximately 0.86, indicating that warmer temperatures strongly drive higher bike demand.

Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: <Your answer for Question 4 goes below this line> (Do not edit)

The assumptions of linear regression were validated as follows:
Linearity: Checked the Residuals vs. Fitted plot to ensure no systematic patterns.
Normality of Residuals: Verified using a Q-Q plot, which showed residuals following a straight line.
Homoscedasticity: Examined the spread of residuals in the Residuals vs. Fitted plot to confirm constant variance.
No Multicollinearity: Used the Variance Inflation Factor (VIF) to ensure all predictors had acceptable VIF values (below 10).

Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 5 goes below this line> (Do not edit)

The top three features contributing to bike demand are:
temp (temperature): Strongly correlated with higher demand.
yr (year): Indicates growth in demand over time.
season_Summer: Reflects high demand during summer compared to other seasons.

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 6 goes here>

Linear regression is a statistical method used to model the relationship between a dependent variable (target) and one or more independent variables (predictors). The algorithm works as follows:
Equation: Represents the target as a linear combination of predictors: , where is the intercept, are coefficients, and is the error term.
Objective: Minimize the sum of squared residuals (differences between observed and predicted values).
Solution: Uses methods like Ordinary Least Squares (OLS) to estimate coefficients .
Assumptions: Assumes linearity, independence, homoscedasticity, and normality of residuals for validity.

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 7 goes here>

Anscombe's quartet consists of four datasets with nearly identical summary statistics (mean, variance, correlation, and regression line), yet they have different distributions and patterns when visualized. This highlights the importance of:
Visualization: Demonstrates how relying solely on statistics can be misleading.
Interpretation: Emphasizes the need to visualize data to uncover patterns, outliers, or non-linear relationships.

Question 8. What is Pearson's R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 8 goes here>

Pearson's R, or the Pearson correlation coefficient, measures the linear relationship between two variables. It ranges from -1 to 1:
1: Perfect positive correlation.
0: No correlation.
-1: Perfect negative correlation.
Pearson's R is useful in understanding the strength and direction of relationships between variables.

Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 9 goes here>

Scaling: Adjusts the range of features to ensure uniformity.
Why: Prevents features with larger magnitudes from dominating models sensitive to scale (e.g., linear regression).
Normalized Scaling: Scales data to a fixed range, typically [0, 1].
Standardized Scaling: Centers data around mean (0) and standard deviation (1), maintaining variability.

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 10 goes here>

VIF becomes infinite when there is perfect multicollinearity, meaning one predictor is an exact linear combination of others. This can occur if:
Dummy variables are not created correctly.
Redundant features are included without elimination.

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 11 goes here>

A Q-Q plot (Quantile-Quantile plot) compares the distribution of residuals to a theoretical normal distribution. It is important because:
Validation: Helps verify the normality assumption of residuals.
Interpretation: Points lying along the diagonal line indicate normality, while deviations suggest skewness or kurtosis in the data.
