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' and 'weathersit' have significant effects on bike demand. For example:

1) 'season' : Higher demand is observed in warmer seasons('season_3' and 'season_4')

2) 'weathersit' : favorable weather conditions('weathersit_1') lead to hihger bike rentals, while poor weather conditions('weather_3') drastically reduce demand.

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` avoids the dummy variable trap, where multicollinearity arises due to redundant dummy variables. By dropping one category, the model can uniquely identify each group while maintaining interpretability.

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` (temperature) shows the highest positive correlation with the target variable `cnt`, indicating that bike demand increases with warmer temperatures.

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)

1. Linearity: Checked residual plots to ensure residuals were randomly distributed.
2. Homoscedasticity: Verified constant variance of residuals across fitted values.
3. Normality: Used Q-Q plots to ensure residuals were approximately normally distributed.
4. Multicollinearity: Calculated VIF to confirm the absence of highly correlated predictors.

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 3 features are:

1. `temp` - Indicates bike demand increases with temperature.
2. `yr` - Suggests increasing demand over years.

---

3. `season_4` - Reflects the high demand in the fall season.

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)

Linear regression predicts a dependent variable (`y`) using one or more independent variables (`x`). It assumes a linear relationship of the form:
$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \varepsilon$
Where:
- β_0 is the intercept.
- β_1, β_2, ... are coefficients.
- ε is the error term.
The algorithm minimizes the sum of squared residuals to find the best-fit line.

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)

Anscombe's quartet consists of four datasets with identical statistical properties (mean, variance, correlation, regression line), but with drastically different distributions and patterns when plotted. It highlights the importance of visualizing data rather than relying solely on summary statistics.

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)

Pearson's R is a measure of the linear correlation between two variables, ranging from -1 to +1. A value of +1 indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and 0 indicates no correlation.

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)

Scaling adjusts feature magnitudes to a similar range for improved model performance.
- **Normalization** rescales data to [0, 1].

- **Standardization** transforms data to have zero mean and unit variance. Scaling is essential for models sensitive to feature magnitude, like linear regression.

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)

Infinite VIF occurs when there is perfect multicollinearity, i.e., one predictor is a perfect linear combination of others. This can be addressed by removing redundant variables.

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)

A Q-Q plot compares the quantiles of residuals to a normal distribution. It helps assess whether residuals follow a normal distribution, a key assumption in linear regression.
