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)

The categorical variables such as 'season', 'year', 'month', 'holiday', 'weekday', 'workingday', and 'weathersit' significantly influence bike demand. From the model's coefficients:

- 'season_4' (Winter) has the highest positive effect on bike demand, increasing it by approximately 1855.35 units.
- 'yr_1' (Year 2019) increases demand by 1953.19 units, suggesting that bike usage increased in the later year.
- 'weathersit_3' (Bad Weather) reduces bike demand by -1874.02 units, indicating poor weather decreases bike rentals.

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' prevents the dummy variable trap, a scenario where multicollinearity arises due to the presence of highly correlated features. It ensures only (n-1) categories are included, allowing the model to interpret the reference category implicitly, leading to better model stability and 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)

From the output and VIF values, 'atemp' (feels-like temperature) has the highest correlation with the target variable (bike demand), as it also shows the highest VIF value (542.88), indicating a strong linear relationship.

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)

- Linearity: Checked by inspecting the relationship between independent variables and the target variable using pair-plots.
- Multicollinearity: Verified using the Variance Inflation Factor (VIF), where features like 'temp' and 'atemp' indicated high multicollinearity.

- Normality of Residuals: Assessed using a Q-Q plot to ensure residuals follow a normal distribution.

- Homoscedasticity: Evaluated by plotting residuals versus fitted values to check for constant variance.

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)

- 'yr_1' (Year 2019) – Increases bike demand by 1953.19 units.

- 'season_4' (Winter) – Increases bike demand by 1855.35 units.

- 'weathersit_3' (Bad Weather) – Decreases bike demand by -1874.02 units.

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 (features). The goal is to find the best-fit line represented by the equation:

$y = \beta0 + \beta1x1 + \beta2x2 + ... + \beta nxn + \varepsilon$

Where:

y is the target variable

β0 is the intercept

β1, β2,... are coefficients (slopes)

x1, x2,... are independent variables

ε is the error term

The model minimizes the sum of squared residuals (differences between actual and predicted values) using the Ordinary Least Squares (OLS) method. Assumptions include linearity, independence, homoscedasticity, and normality of residuals.

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 statistical properties (mean, variance, correlation) but visually different distributions. It highlights the importance of data visualization in addition to statistical summaries. The datasets show:

- A linear relationship
- A curve
- An outlier affecting regression
- A vertical outlier this emphasizes checking visual patterns to avoid misleading interpretations from statistics alone.

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 is the Pearson correlation coefficient, measuring the linear relationship between two continuous variables. It ranges from -1 to +1:

- +1 indicates a perfect positive correlation
- -1 indicates a perfect negative correlation
- 0 indicates no correlation

It is calculated as:

$R = \text{cov}(X,Y) / (\sigma X * \sigma Y)$
Where cov(X,Y) is the covariance and σ represents standard deviation. It assumes linearity, homoscedasticity, and normality.

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 transforms data to ensure all features contribute equally to the model. It helps algorithms converge faster and improves accuracy.

- Normalization (Min-Max Scaling): Scales values between [0,1] using:
  X_scaled = (X - min) / (max - min)

- Standardization (Z-score scaling): Centers data to mean 0 and standard deviation 1:
  X_scaled = (X - mean) / std

Normalization is suitable for bounded data; standardization is used when data follows a Gaussian distribution.

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>
Infinite VIF occurs due to perfect multicollinearity, where one feature is a linear combination of others. This happens if:

- Duplicate or highly correlated variables exist.

- Dummy variable trap (not using 'drop_first=True').

- Derived features (e.g., temperature and feels-like temperature) overlap.

Resolving it requires removing or combining correlated 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)

<Your answer for Question 11 goes here>
A Quantile-Quantile (Q-Q) plot compares the distribution of residuals to a normal distribution. In linear regression, it validates the assumption of normal residuals. If points lie along the 45-degree line, residuals are normally distributed. Deviations indicate skewness or heavy tails, impacting model accuracy. It's essential for ensuring unbiased coefficient estimates and reliable predictions.
