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)

Based on our analysis, categorical variables showed significant effects on bike rentals:

1. Season: Highest demand in Summer and Fall, lowest in Winter
2. Year: 2019 showed ~23% higher demand than 2018, indicating growing adoption
3. Weather: Clear weather (category 1) showed 30% higher rentals compared to adverse weather conditions
4. Working Day: Working days showed more consistent rental patterns compared to holidays/weekends

These variables help explain the seasonal and temporal patterns in bike rental 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 is important for two reasons:

1. Avoids the dummy variable trap (perfect multicollinearity) by removing one category as the reference level
2. Prevents redundant information since n-1 dummy variables can fully represent n categories
3. Improves model stability and interpretation as coefficients represent the effect relative to the reference category

---

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)

Temperature ('temp') showed the highest correlation with the target variable 'cnt' with a correlation coefficient of 0.62. This indicates that temperature is the strongest numerical predictor of bike rental 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)

We validated the linear regression assumptions through:

1. Linearity: Residual plot showed no systematic patterns, confirming linear relationships
2. Normality: Q-Q plot demonstrated residuals following approximately normal distribution
3. Homoscedasticity: Plot of residuals vs predicted values showed relatively constant variance
4. Independence: Time-based plot of residuals showed no significant autocorrelation

5. Multicollinearity: VIF analysis on predictor variables ensured no severe multicollinearity

---

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 with highest absolute coefficients were:
1. Temperature (0.5 coefficient): Strongest positive impact on demand
2. Year_2019 (0.23 coefficient): Significant year-over-year growth
3. Clear Weather (0.19 coefficient): Strong positive impact during good weather conditions

These features showed the most substantial and statistically significant effects on bike rental demand.

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 models the relationship between a dependent variable (Y) and one or more independent variables (X) using the equation:

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \varepsilon$

Where:
- β_0 is the intercept
- β_i are coefficients representing the change in Y for one unit change in X_i
- ε is the error term

The algorithm:
1. Uses Ordinary Least Squares (OLS) to minimize the sum of squared residuals
2. Finds optimal coefficients that best fit the training data
3. Makes predictions using the linear combination of features and their coefficients

---

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 that have nearly identical statistical properties (mean, variance, correlation, linear regression line) but look very different when plotted. Key points:

1. Demonstrates importance of visualizing data before analysis
2. Shows limitations of summary statistics alone
3. Illustrates why checking model assumptions is crucial
4. Each dataset has:
- Same mean of x and y
- Same variance of x and y
- Same correlation coefficient
- Same linear regression line
But they represent very different relationships, including linear, nonlinear, and outlier cases.

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 (correlation coefficient) measures the strength and direction of linear relationship between two variables:

1. Ranges from -1 to +1
- +1 indicates perfect positive correlation
- -1 indicates perfect negative correlation
- 0 indicates no linear correlation

2. Properties:
- Scale-invariant
- Symmetric
- Only measures linear relationships

3. Formula: $r = cov(X,Y)/(\sigma x * \sigma y)$
Where cov is covariance and σ is standard deviation

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 features to a similar range:

1. Why Scaling is Important:
- Prevents features with larger ranges from dominating
- Improves convergence of gradient descent
- Makes features comparable

2. Normalized Scaling (Min-Max):
- **Scales features to [0,1] range**
- **Formula: X_norm = (X - X_min)/(X_max - X_min)**
- **Preserves zero values**
- **Better with non-normal distributions**

3. Standardized Scaling (Z-score):
- **Transforms to mean=0, std=1**
- **Formula: X_std = (X - μ)/σ**
- **Better for normal distributions**
- **Handles outliers better**

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's perfect multicollinearity:
1. Occurs when one feature is an exact linear combination of others
2. Common causes:
   - Including all dummy variables (dummy variable trap)
   - Duplicate features
   - Derived features that are linear combinations
3. Makes coefficient estimates unstable
4. Solution: Remove one of the perfectly correlated features

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>

Q-Q (Quantile-Quantile) plot is a visualization tool that:

1. Purpose:
- **Assesses if residuals follow normal distribution**
- **Compares empirical distribution to theoretical normal**

distribution

2. Interpretation:
- **Points following diagonal line suggest normality**
- **Deviations indicate non-normality**
- **S-shaped curve suggests skewness**
- **Tails deviating suggest kurtosis issues**

3. Importance in Linear Regression:
- **Validates normality assumption**
- **Helps identify potential outliers**
- **Guides potential data transformations**
