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 analysis revealed significant insights about categorical variables' impact on bike rental demand:

1. **Seasonal Variation**:
   - Different seasons show substantial variations in rental demand
   - Fall and Winter seasons demonstrated higher rental coefficients
   - Suggests seasonal preferences in bike usage

2. **Weather Situation**:
   - Clear weather (weathersit_1) positively influences rental demand
   - Misty or cloudy conditions (weathersit_2) slightly reduce rentals
   - This indicates that weather significantly affects bike rental decisions

3. **Working Day vs. Weekend**:
   - Slight variations in rental patterns between working days and weekends
   - Suggests different usage patterns based on day type
   - Potential impact of commuting and leisure activities on 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 ensures that the regression model remains free from multicollinearity, is easier to interpret, and is more computationally efficient by removing unnecessary dummy variables.

Avoiding Multicollinearity:
- Prevents the "dummy variable trap" by removing one reference category
- Eliminates perfect linear dependency between categorical variables

Model Interpretability:
- Reduces the number of features
- Allows easier interpretation of coefficients
- Prevents redundant information in the regression model

Computational Efficiency:
- Reduces computational complexity
- Helps in more stable model estimation

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)

Based on the pair-plot and correlation analysis, **temperature (temp)** has the highest correlation with the target variable (total bike rentals).

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)

By systematically evaluating these assumptions through visual diagnostics and statistical checks, I ensured that the linear regression model was appropriately specified and reliable for making inferences.

1. Linearity Check:
- 1.1.Residual plots examined for random scatter around zero
- 1.2.Plotted predicted vs. actual values to verify linear relationship

2. Homoscedasticity Verification:
- 2.1.Analysed residual plot for consistent variance
- 2.2.Checked for uniform spread of residuals

3. Normality of Residuals:
- 3.1.Examined distribution of residuals
- 3.2.Plotted histogram and Q-Q plot to assess normality
- 3.3.Verified residuals approximate normal distribution

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)

Top 3 Features Influencing Shared Bike Demand:

Temperature (temp): Highest positive coefficient
Feels-like Temperature (atemp): Second most important numeric feature
Humidity (hum): Significant negative correlation with 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)

Linear Regression is a supervised machine learning algorithm that models the relationship between dependent and independent variables through a linear equation.

Key Components:

1. **Mathematical Representation**:
   - $y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \varepsilon$
   - β_o: Intercept
   - β_n: Coefficients
   - x_n: Independent variables
   - ε: Error term

2. **Objective**:
   - Minimise the sum of squared differences between predicted and actual values
   - Uses Ordinary Least Squares (OLS) method

3. **Key Steps**:
   - Data collection
   - Feature selection
   - Model training
   - Coefficient estimation
   - Prediction

Estimation Process:

- Calculates best-fit line minimising prediction errors
- Finds coefficients that minimise residual sum of squares
- Provides interpretable relationships between variables

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 is a statistical phenomenon demonstrating the importance of visual data exploration.

Key Characteristics:

1. Four different datasets with identical:
   - Mean
   - Variance
   - Correlation
   - Linear regression line

2. **Visual Demonstration**:
   - Highlights limitations of relying solely on statistical summaries
   - Emphasises need for graphical data exploration

3. **Lesson**:
   - Always visualise data
   - Statistical summaries can be misleading
   - Graphical representation reveals hidden data characteristics

---

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 Correlation Coefficient (Pearson's R) measures linear relationship between two continuous variables.

Key Features:

1. **Range**: -1 to +1
   - +1: Perfect positive correlation
   - 0: No linear correlation
   - -1: Perfect negative correlation

2. **Calculation**:
   - Measures covariance between variables
   - Normalized by standard deviations
   - Indicates strength and direction of linear relationship

3. **Interpretation**:
   - Magnitude indicates correlation strength
   - Sign indicates relationship direction

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 is a data preprocessing technique used to adjust the range and distribution of feature values in a dataset. It transforms features to a common scale without distorting the differences in their ranges, ensuring that each feature contributes proportionately to the analysis.

By applying appropriate scaling techniques, models can achieve better performance, faster convergence, and more reliable predictions.

Purpose:

1. Normalize feature ranges
2. Prevent dominance of high-magnitude features
3. Improve model performance

Scaling Types:

1. **Normalized Scaling (Min-Max)**:
   - Scales features to [0, 1] range
   - Formula: (x - min) / (max - min)
   - Preserves zero values
   - Sensitive to outliers

2. **Standardized Scaling (Z-score)**:

- Centers data around mean (0)
- Scales to unit variance
- Formula: (x - mean) / standard deviation
- Handles outliers better
- Preferred for algorithms sensitive to scale

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)

Variance Inflation Factor (VIF) becomes infinite when:

1. **Perfect Multicollinearity**:
   - One feature is exactly linearly predictable from others
   - Indicates complete linear dependency

2. **Computational Reasons**:
   - Division by zero in VIF calculation
   - Occurs when feature's variance becomes zero

3. **Practical Implications**:
   - Signals redundant features
   - Requires feature selection or dimensionality reduction

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)

Q-Q (Quantile-Quantile) Plot compares sample data distribution to theoretical distribution.

Purpose:

1. Assess normality of residuals
2. Validate linear regression assumptions
3. Detect distribution deviations

Interpretation:

1. Straight line: Normal distribution
2. Curve deviations: Non-normal distribution
3. Helps identify model's statistical validity
