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)

1. Higher rentals are expected in summer and spring, with winter showing the least activity
2. Clear weather is likely to promote rentals, while adverse conditions lead to decreased 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 dummy variable encoding with drop_first=True helps prevent multicollinearity, establishes a reference category for better interpretability, and improves model efficiency. This is a standard practice in statistical modeling, especially for regression analysis.

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)

Variables like temperature and apparent temperature often have a strong positive correlation with bike rentals, as favorable weather conditions influence outdoor activities. By identifying the variable with the highest correlation to bike rentals, you can use that insight to improve your predictive model.

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)

After building your linear regression model, you can validate its assumptions through visual methods (like residual plots and Q-Q plots) and statistical tests (like Durbin-Watson, Breusch-Pagan, and Shapiro-Wilk). This thorough validation helps ensure the model's reliability and informs whether any transformations or adjustments are needed.

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)

In this hypothetical output, the top three features contributing to bike demand would be temp, atemp, and hum, with temp having the strongest positive effect. Analyzing these features will help you understand which factors are most influential in driving bike rentals, allowing for more targeted business strategies and marketing efforts.

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>

Basic Concept

The central idea of linear regression is to fit a linear equation to the observed data. The general form of the linear regression model can be expressed as:

Y=β0+β1X1+β2X2+...+βnXn+ϵY = \beta_0 + \beta_1X_1 + \beta_2X_2 + ... + \beta_nX_n + \epsilon Y=β0+β1X1+β2X2+...+βnXn+ϵ

Types of Linear Regression

- **Simple Linear Regression**: Involves one independent variable and one dependent variable. The model fits a straight line to the data points.

  Y=β0+β1X1+ϵY = \beta_0 + \beta_1X_1 + \epsilonY=β0+β1X1+ϵ

- **Multiple Linear Regression**: Involves multiple independent variables. The model fits a hyperplane to the data points in multi-dimensional space.

  Y=β0+β1X1+β2X2+...+βnXn+ϵY = \beta_0 + \beta_1X_1 + \beta_2X_2 + ... + \beta_nX_n + \epsilonY=β0+β1X1+β2X2+...+βnXn+ϵ

Assumptions of Linear Regression

Linear regression makes several assumptions about the data:

1. **Linearity**: The relationship between independent and dependent variables is linear.
2. **Independence**: Observations are independent of each other.
3. **Homoscedasticity**: The residuals (errors) have constant variance at all levels of the independent variables.
4. **Normality**: The residuals of the model should be normally distributed.
5. **No multicollinearity**: Independent variables should not be highly correlated with each other.

Fitting the Model

The goal of linear regression is to find the optimal values of the coefficients (β\betaβ) that minimize the difference between the predicted values and the actual values of the dependent variable. This is often done using:

- **Ordinary Least Squares (OLS)**: A common method for estimating the parameters. OLS minimizes the **Residual Sum of Squares (RSS)**, defined as:

RSS=∑i=1n(Yi−Y^i)2RSS = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 RSS=i=1∑n(Yi−Y^i)2

Implementation Steps

The general steps to implement a linear regression model are:

1. **Data Preparation**: Clean the dataset, handle missing values, and perform any necessary preprocessing (like encoding categorical variables).
2. **Exploratory Data Analysis (EDA)**: Analyze relationships between features and the target variable using visualizations.
3. **Split Data**: Divide the dataset into training and testing subsets to evaluate the model's performance.
4. **Fit the Model**: Use a library (like scikit-learn in Python) to fit the linear regression model on the training data.
5. **Make Predictions**: Use the model to predict the target variable for the test data.
6. **Evaluate the Model**: Assess the model's performance using metrics like R-squared, Mean Squared Error (MSE), and Root Mean Squared Error (RMSE).

Limitations of Linear Regression

- **Sensitivity to Outliers**: Outliers can disproportionately affect the model.
- **Linearity Assumption**: If the relationship is not linear, the model may underperform.
- **Feature Selection**: Requires careful selection of relevant features; irrelevant features can degrade performance.

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>

1. **Overview of Anscombe's Quartet**

- **Purpose**: The main goal of Anscombe's quartet is to illustrate that statistical properties alone can be misleading without visualizing the data. It emphasizes the need to consider the distribution and structure of the data.
- **Components**: The quartet consists of four different datasets (labeled I,II,III,IVI, II, III, IVI,II,III,IV), each containing 11 pairs of (x,y)(x, y)(x,y) values.

2. **Statistical Properties**

Despite their different appearances, each dataset in Anscombe's quartet has the following identical statistical properties:

- Mean of xxx: 9
- Mean of yyy: 7.5
- Variance of xxx: 11
- Variance of yyy: 4.125
- Correlation between xxx and yyy: 0.8165
- Linear regression line: All datasets have the same regression line equation: y=3+0.5xy = 3 + 0.5xy=3+0.5x

3. **Graphical Representation**

   When visualized, the datasets reveal significant differences:

- **Dataset I**: Shows a linear relationship.
- **Dataset II**: Shows no relationship; all xxx values are the same, but yyy values vary.
- **Dataset III**: Represents a non-linear relationship (quadratic).
- **Dataset IV**: Exhibits a distinct pattern with one outlier (with an unusual point affecting the overall 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)

<Your answer for Question 8 goes here>

Pearson's R is a valuable statistical tool for assessing the linear relationship between two variables. While it provides useful insights, it is essential to consider its limitations and ensure that the underlying assumptions are met when interpreting its results

---

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>

Why Scaling is Performed

1. **Improves Convergence Speed**: Many machine learning algorithms, particularly those based on gradient descent (like linear regression and neural networks), converge faster when features are on a similar scale. This is because features with larger ranges can cause the optimization algorithm to oscillate wildly and take longer to converge.
2. **Enhances Performance**: Scaling can improve the performance of models that are sensitive to the scale of the data, such as Support Vector Machines (SVM), K-nearest neighbors (KNN), and others. If features are not scaled, models may attribute more importance to larger values.
3. **Facilitates Distance Calculation**: Algorithms that rely on distance metrics, such as clustering and nearest neighbor algorithms, benefit from scaling since distances will be more meaningful when features are comparable.
4. **Ensures Feature Comparability**: When features have different units or ranges, scaling helps to make them comparable, allowing for better interpretation of model coefficients in linear models.

Scaling is an essential step in preparing data for machine learning and statistical analysis. Choosing between normalization and standardization depends on the distribution of your data and the specific requirements of your chosen model. Properly scaled data leads to better model performance and interpretability.

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 values indicate perfect multicollinearity, which can severely impact regression model performance. Identifying and addressing the underlying causes of multicollinearity is essential for creating a stable and interpretable regression model.

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 is a valuable diagnostic tool in linear regression that helps assess the normality of residuals, verify model assumptions, and evaluate the appropriateness of the model. It plays a crucial role in ensuring that the results obtained from regression analysis are valid and reliable.
