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)

Season, Year, and Weather Situation have the strongest effects on rentals.
Holidays negatively impact rentals, while weekends might increase them.
Warm months & clear weather significantly boost demand.
Multicollinearity (e.g., `season` & `mnth`) must be checked using **VIF** to avoid redundant predictors.

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` during **dummy variable creation** is important to avoid the problem of **multicollinearity** in your regression model.

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)

Temprature

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) There should be no strong correlation between the independent variables, as multicollinearity can inflate standard errors and make the model unstable.

2) The residuals should be approximately normally distributed for hypothesis testing (like t-tests for coefficients) to be valid.

3) The relationship between the independent variables (predictors) and the dependent variable (target) should be **linear**.

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)

temp (Temperature):

- **Coefficient**: 0.5038
- **t-value**: 14.860
- **p-value**: 0.000

- **Interpretation**: The temperature has a strong positive relationship with bike demand. As the temperature increases, the demand for bikes significantly rises.

Year:

- **Coefficient**: 0.2329
- **t-value**: 27.665
- **p-value**: 0.000
- **Interpretation**: The year feature has a significant positive influence on the demand, indicating that the demand has increased over time (possibly due to factors like increasing popularity of bike-sharing or infrastructure development).

Light Snow:

- **Coefficient**: -0.3032
- **t-value**: -12.045
- **p-value**: 0.000
- **Interpretation**: Light snow has a strong negative effect on bike demand. As light snow increases, the demand for bikes significantly decreases, which is expected as adverse weather reduces outdoor activity.

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 model used to establish a relationship between a dependent variable (target) and one or more independent variables (predictors). The goal is to fit a linear equation that minimizes the difference between observed and predicted values.

Steps in Linear Regression:

2. **Fitting the Model**: The model coefficients (\(\beta_0, \beta_1, \dots, \beta_n\)) are estimated by minimizing the **sum of squared errors** (difference between actual and predicted values).

3. **Prediction**: Once the model is trained, it can predict new values of \(Y \) based on the learned coefficients and input features.

Key Assumptions:
- Linearity: The relationship between the variables is linear.
- Independence: Observations are independent.
- Homoscedasticity: Constant variance of errors.
- Normality of errors for hypothesis testing.

Linear regression is a simple yet powerful tool for predictive analysis, assuming the relationships between variables are linear.

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 identical summary statistics but very different underlying distributions and relationships. The key takeaway is the importance of visualizing data rather than relying solely on summary statistics, as they can be misleading.

- **Dataset I** shows a classic linear relationship with no outliers.
- **Dataset II** has a non-linear relationship, with an outlier affecting the regression line.
- **Dataset III** has nearly constant X values with one outlier, making the regression line less meaningful.
- **Dataset IV** shows a linear relationship with one extreme outlier, distorting the regression line.

The quartet emphasizes that statistical measures like correlation and regression coefficients alone may not provide a true representation of data, and visual inspection is crucial for understanding underlying patterns.

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 (or Pearson correlation coefficient) is a measure that quantifies the strength and direction of a linear relationship between two continuous variables. It ranges from -1 to 1:

- **1** indicates a perfect positive linear relationship,
- **-1** indicates a perfect negative linear relationship,
- **0** means no linear relationship.

The closer the value is to 1 or -1, the stronger the 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)

<Your answer for Question 9 goes here>
Scaling is the process of adjusting the range or distribution of features in a dataset, ensuring that they are on a similar scale. It is crucial for machine learning models, as many algorithms are sensitive to the magnitude of features.

- **Normalized Scaling** (Min-Max Scaling) rescales data to a fixed range, typically [0, 1], by subtracting the minimum and dividing by the range.

- **Standardized Scaling** (Z-score Scaling) transforms data to have a mean of 0 and a standard deviation of 1 by subtracting the mean and dividing by the standard deviation.

The main difference is that normalization focuses on a specific range, while standardization centers and scales the data based on its 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>

The **Variance Inflation Factor (VIF)** becomes **infinite** when there is perfect multicollinearity between two or more predictor variables in a dataset. This occurs when one predictor variable is a perfect linear combination of others, meaning that the predictor variables are highly correlated (correlation coefficient of 1 or -1).

In such cases, the regression model cannot distinguish between the predictors, leading to an inflated variance estimate for the coefficients, which results in an infinite VIF. Essentially, the model struggles to separate the individual contributions of the 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 **Q-Q (Quantile-Quantile) plot** is a graphical tool used to compare the distribution of a dataset to a theoretical distribution (commonly the normal distribution). It plots the quantiles of the observed data against the quantiles of the theoretical distribution. If the points fall roughly along a straight line, it indicates that the data follows the theoretical distribution.

Use and Importance in Linear Regression:
- **Normality of Residuals**: In linear regression, one of the key assumptions is that the residuals (errors) are normally distributed. A Q-Q plot helps visualize this assumption by comparing the distribution of residuals to a normal distribution.
- **Identifying Deviations**: Significant deviations from the straight line suggest that the residuals are not normally distributed, indicating a potential issue with the model (e.g., non-linearity, outliers, or heteroscedasticity).

In summary, a Q-Q plot is essential for diagnosing the normality assumption in linear regression, which affects the validity of statistical tests and confidence intervals.
