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 target variable is highly dependent on categorical variables like yr, Weathersit, Season, and Weekdays. For example:

1. Weather and seasonality (month) strongly influence bike rentals, with favorable weather and warmer seasons driving higher usage.
2. Rentals have increased over time, possibly reflecting trends or changes in external factors like infrastructure or population behavior.
3. Weekday variations are minimal, indicating stable demand throughout the week, possibly for regular commuting or daily errands.

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)

By setting drop_first=True, one category is dropped (the reference category), leaving k−1 dummy variables. This eliminates multicollinearity while retaining the necessary information. The coefficients of the remaining dummy variables are interpreted 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)

After processing the data, we can see that the target variable (cnt) has the highest correlation of 0.63 with 'temp' and 'atemp' variables.

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: A scatter plot of residuals vs. predicted values was examined to ensure no clear patterns, confirming a linear relationship between predictors and the target variable.

Multicollinearity: I calculated Variance Inflation Factor (VIF) values to ensure no multicollinearity among the predictors (VIF < 5).

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 highest contributors for the demand of bike rentals are:
Temperature
Year
Season

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 supervised learning algorithm that is used to analyze the relationship between a dependent variable and one or more independent variables by fitting a straight line. The line is represented as $y = \beta 0 + \beta 1x1 + \ldots + \beta nxn$, where β values are coefficients. It minimizes the sum of squared differences (errors) between predicted and actual values (least squares method). The model assumes linearity, independence, homoscedasticity, and normality of residuals for accurate predictions.

Linear Regression is divided into basically two types:

Simple Linear Regression: It is used to analyze the relationship between one independent variable and one dependent variable using a straight line.

Multiple Linear Regression: It extends simple linear regression and is used to analyze the relationship between multiple independent variables and one dependent variable.

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 is a set of four datasets which have nearly identical statistical properties, such as mean, variance, correlation, and regression line, but differ significantly in their graphical representation. It demonstrates the importance of visualizing data rather than relying completely on summary statistics. The datasets highlight how outliers, non-linear relationships, or other patterns can be hidden when analyzing data numerically. This quartet emphasizes that statistical measures like mean, variance, and correlation might not fully describe a dataset, making visual inspection crucial for understanding and interpreting data effectively.

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 the Pearson correlation coefficient, is a statistical measure that quantifies the linear relationship between two continuous variables.
It ranges from − 1 −1 to 1 1, where:
1: Perfect positive linear correlation (as one variable increases, the other increases proportionally).
0: No linear correlation.
−1: Perfect negative linear correlation (as one variable increases, the other decreases proportionally).

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 a data preprocessing technique used to adjust the range or distribution of numerical features so that they fit within a specific scale (e.g., 0–1 or a standard normal distribution). This ensures that all features contribute equally to the model and prevents features with larger magnitudes from dominating the learning process.

Scaling is performed for the following reasons:
It Improves Model Performance: Many machine learning algorithms (e.g., gradient descent, SVM, k-NN) are sensitive to feature magnitudes. Scaling improves convergence speed and accuracy.
Equal Feature Contribution: Ensures features are treated equally by models, especially those based on distance (e.g., k-means, PCA).
Prevents Numerical Instability: Large feature values can cause instability in calculations.

Normalized scaling rescales data to a specific range, usually [0, 1], making it suitable for comparisons based on magnitude. It is calculated by subtracting the minimum value and dividing by the range of the data. Standardized scaling, on the other hand, centers data around the mean and scales it to have unit variance, making it ideal for data with different distributions or outliers. It is calculated by subtracting the mean and dividing by the standard deviation. Normalization is useful when comparing feature magnitudes, while standardization is used for algorithms that assume a normal 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 value of Variance Inflation Factor (VIF) becomes infinite when there is perfect multicollinearity among the predictor variables in a regression model. This occurs when one predictor variable is a perfect linear function of another, meaning that one variable can be exactly predicted using the others. In this case, the matrix used to calculate VIF (the covariance matrix) becomes singular or non-invertible, leading to an infinite value for VIF.

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 (Quantile-Quantile plot) is a graphical tool used to assess if a dataset follows a particular theoretical distribution, such as the normal distribution. It compares the quantiles of the observed data with the quantiles of a specified distribution. If the data follows the theoretical distribution, the points on the Q-Q plot will lie roughly along a straight line.

Use in Linear Regression: In linear regression, one of the key assumptions is that the residuals (the differences between the observed and predicted values) are normally distributed. A Q-Q plot helps to visually assess whether the residuals meet this assumption.
