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)

- Year (yr): A positive coefficient for yr indicates that demand in 2019 is higher than in 2018, aligning with a trend of increasing popularity for bike-sharing.
- Holiday (holiday): A negative coefficient for holiday suggests lower bike demand on holidays, possibly because fewer people commute or use bikes on non-working days.
- Season (season): Positive coefficients for summer, fall, and winter (relative to spring, the baseline) imply higher demand in these seasons. This suggests that demand increases in warmer or more active seasons.
- Weather (weathersit): Negative coefficients for conditions like mist and light snow indicate reduced demand under these weather conditions, likely due to safety concerns or reduced comfort.

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 avoids the dummy variable trap, which occurs when all categories are included, leading to multicollinearity. By dropping the first category, we ensure that the remaining dummies represent relative comparisons, making the model more stable and interpretable.

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)

The highest correlation is of 0.63 between 'cnt' and 'temp'

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**:** Verified by plotting predicted vs. actual values to check if the relationship between predictors and the target variable, cnt, is approximately linear. Normality of Residuals**:** Checked to ensure residuals are normally distributed.

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, yr and light snow. Here all p-values are 0.00, indicating that all features are statistically significant and the higher the absolute value of a coefficient, the greater its impact on

---

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 statistical method used for modeling the relationship between a dependent variable and one or more independent variables. The core principle of linear regression is to find the best-fitting line (or hyperplane in multiple dimensions) that minimizes the differences between the observed values and the values predicted by the model. This is achieved by minimizing the sum of squared errors, known as the least squares method. The linear regression equation is typically expressed as (Y = beta_0 + beta_1X_1 + beta_2X_2 + ... + beta_nX_n + epsilon). The algorithm involves several steps: initializing the coefficients, predicting the output based on the input features, calculating the error, updating the coefficients to minimize this error, and repeating the process until convergence is achieved. Linear regression assumes a linear relationship between the variables and is sensitive to outliers, which can significantly affect the model's performance.
<Your answer for Question 6 goes here>

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 set of four datasets that were created by statistician Francis Anscombe in 1973 to illustrate the importance of graphing data before analyzing it and to demonstrate how summary statistics can be misleading. Each dataset contains 11 pairs of \((x, y) \) values and has nearly identical statistical properties, including the same mean, variance, and correlation coefficient, yet the datasets exhibit very different distributions and relationships when graphed. The first dataset shows a strong positive linear relationship, resembling a straight line. The second dataset presents a nonlinear relationship that curves upwards. The third dataset has a clear linear relationship but includes an outlier that distorts the linearity. The fourth dataset displays a vertical line of points with a single outlier that suggests no linear relationship at all. Despite their similar statistical summaries, the visualizations reveal distinct patterns, emphasizing that reliance solely on statistical metrics can lead to erroneous interpretations. Anscombe's quartet serves as a valuable educational tool in statistics, highlighting the necessity of visual data exploration to uncover underlying trends and anomalies.
<Your answer for Question 7 goes here>

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 R, also known as the Pearson correlation coefficient, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It ranges from -1 to 1, where a value of 1 indicates a perfect positive linear relationship, meaning that as one variable increases, the other also increases proportionally. Conversely, a value of -1 indicates a perfect negative linear relationship, where an increase in one variable corresponds to a proportional decrease in the other. A value of 0 indicates no linear correlation between the variables. The formula for calculating Pearson's R is (R = {Cov(X, Y)}/{sigma_X * sigma_Y}). Pearson's R assumes a linear relationship, normal distribution of the variables, and measurement on an interval or ratio scale. It is sensitive to outliers, which can significantly influence the correlation coefficient. Pearson's R is widely used across various fields, such as psychology, economics, and biology, to assess relationships between variables and guide decision-making processes.

<Your answer for Question 8 goes here>

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 preprocessing technique used in data analysis and machine learning to adjust the range and distribution of feature values, ensuring that algorithms relying on distance metrics do not give undue weight to features with different scales. This adjustment is crucial for improved convergence, enhanced performance, and better interpretability of the data. There are two main types of scaling: normalized scaling (or min-max scaling), which rescales feature values to a fixed range, typically [0, 1], using the formula (X' = {X - X_{min}}/{X_{max} - X_{min}}); and standardized scaling (or z-score normalization), which rescales feature values to have a mean of 0 and a standard deviation of 1, calculated with (X' = {X - mu}/{sigma}). Normalization is particularly useful when data needs to be bounded within a specific range, while standardization is beneficial for data that follows a Gaussian distribution or has different units.

<Your answer for Question 9 goes here>

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)

The Variance Inflation Factor (VIF) measures how much the variance of an estimated regression coefficient increases when your predictors are correlated. A VIF value is calculated for each predictor in a regression model, and a value of 1 indicates no correlation between the predictor

and the other variables. However, when the VIF is infinite, it typically occurs due to perfect multicollinearity among the predictors.

Perfect multicollinearity happens when one predictor variable is a perfect linear combination of one or more other predictor variables. In such cases, the matrix used to calculate the coefficients in the regression becomes singular, meaning it cannot be inverted. This leads to an undefined VIF value, which is represented as infinite. Consequently, infinite VIF indicates that the predictor does not provide any unique information to the model, and addressing this issue may involve removing or combining correlated predictors to improve the regression model's performance and interpretability.

<Your answer for Question 10 goes here>

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)

A Q-Q plot, or Quantile-Quantile plot, is a graphical tool used to compare the distribution of a dataset to a theoretical distribution, such as the normal distribution, by plotting the quantiles of the dataset against the quantiles of the theoretical distribution. In linear regression, a Q-Q plot is primarily used to assess the normality of the residuals, which are the differences between observed and predicted values. By visualizing the residuals in a Q-Q plot, analysts can determine if they follow a normal distribution, which is a key assumption of linear regression. If the data points deviate significantly from the straight line, it indicates non-normality, suggesting potential issues such as outliers or violations of the regression assumptions. Ensuring that the residuals are normally distributed is crucial, as non-normality can lead to biased estimates of regression coefficients, affect hypothesis tests, and result in inaccurate confidence intervals. Thus, the Q-Q plot serves as an important diagnostic tool in linear regression, helping to validate the robustness and reliability of the model's results.

<Your answer for Question 11 goes here>
