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)

Categorical variables like season, weather conditions, weekday, and working days significantly affect the bike demand. For instance, bike rentals tend to be higher on weekends (e.g., Saturdays) compared to weekdays. Weather conditions also show a notable influence, with adverse weather (e.g., light snow/rain) resulting in fewer rentals. Seasonality is another important factor, where demand varies between seasons like winter and spring. The year variable suggests an increase in rentals over time, reflecting growth in the service's usage.

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 there is perfect multicollinearity among the dummy variables. This means that one variable can be perfectly predicted from the others, leading to redundant information. By dropping the first category, we remove one dummy variable from the model, ensuring that the remaining variables provide independent information, preventing multicollinearity.

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 variable with the highest correlation with the target variable (bike demand) is temperature. As temperature increases, bike demand tends to rise, indicating that favorable weather conditions drive more 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)

The assumptions of linear regression were validated by:
Residual Analysis: A residuals vs. predicted values plot was used to check for homoscedasticity (constant variance of residuals). No pattern in the residuals suggested homoscedasticity.
Normality of Residuals: A Q-Q plot and a histogram of residuals were used to verify that the residuals followed a normal distribution.
Independence of Errors: A lag plot of residuals was used to check for autocorrelation, ensuring that the residuals are independent of each other.

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 top 3 features contributing significantly to bike demand are:

Temperature: A higher temperature correlates strongly with higher bike demand.
Year: The increase in rentals over the years reflects the growing popularity of the service.
Humidity: Higher humidity has a negative impact on demand, reducing the number of rentals.

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 method used to model the relationship between a dependent variable and one or more independent variables. The goal is to fit a line (in the case of one independent variable) or a hyperplane (for multiple independent variables) that minimizes the sum of squared residuals (differences between the actual and predicted values).

The equation for linear regression is:
$y=\beta0+\beta1x1+\beta2x2+...+\beta nxny$

Where:

- y is the dependent variable,
- x1,x2,.. xn are independent variables,
- β0 is the intercept, and

- β1,β2,...,βnare the coefficients representing the weight of each independent variable.

The algorithm uses the Ordinary Least Squares (OLS) method to estimate the coefficients, minimizing the sum of squared residuals. Linear regression assumes a linear relationship, independence of residuals, no multicollinearity, and normally distributed residuals.

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 that have nearly identical simple descriptive statistics (mean, variance, correlation, and linear regression results) but appear very different when graphed. Each dataset in the quartet reveals different aspects of the relationship between the variables, highlighting the importance of visualizing data before analyzing it.

The quartet demonstrates that relying solely on summary statistics like correlation or R-squared can be misleading, and it's essential to graph data to uncover patterns, outliers, or non-linear relationships.

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 (Pearson's correlation coefficient) is a measure of the strength and direction of the 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, and
- **0** indicates no linear relationship.

Pearson's R is calculated by dividing the covariance of the two variables by the product of their standard deviations. It is widely used to understand correlations in datasets.

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 of features in a dataset to ensure that they are on a similar scale. It is important because many machine learning algorithms (like linear regression, SVM, etc.) perform better when input features are of similar magnitudes.

- **Normalization (Min-Max Scaling):** Rescales the data to a range of 0 to 1 using the formula: $X_{scaled} = X - X_{min} X_{max} - X_{min}$
- **Standardization:** Transforms the data so that it has a mean of 0 and a standard deviation of 1: $X_{standardized} = (X - \mu) / \sigma$

Normalization is preferred when the features are not normally distributed, whereas standardization is used when data follows 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, meaning one variable is a perfect linear combination of the others. In this case, the regression algorithm cannot isolate the effects of each predictor, leading to instability in the model. To resolve this, one of the perfectly collinear variables should be removed from the 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** (Quantile-Quantile plot) is a graphical tool used to compare the distribution of a dataset to a theoretical distribution (usually normal). It plots the quantiles of the residuals against the quantiles of a normal distribution.

In linear regression, the Q-Q plot is important for verifying the **normality assumption** of the residuals. If the residuals are normally distributed, the points in the Q-Q plot will align along the 45-degree reference line. Deviations from this line suggest departures from normality, which could affect the reliability of the regression model.
