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)

- Bike rentals tend to increase during the spring and summer months, but decrease during the fall and winter months.
- The demand for bike rentals was higher in 2019 compared to 2018.
- The months from June to September see the highest bike demand, while January experiences the lowest.
- Bike demand is lower on holidays compared to non-holiday periods.
- The demand for rental bikes remains relatively consistent throughout the weekdays.
- There is no significant difference in bike demand between working days and non-working days.
- The highest bike rental demand occurs during clear or partly cloudy weather, followed by misty cloudy weather, and the third-highest during light snow and light rain conditions.

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 multicollinearity in the model. When creating dummy variables from a categorical variable with n levels, we would end up with n columns, but one of those columns will be redundant. By dropping the first category (usually), we prevent the dummy variables from being perfectly correlated with each other. This reduces the risk of perfect multicollinearity, ensuring that the model does not face issues like inflated coefficients or biased estimates.

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 temp and atemp variables are strongly positively correlated, indicating that they provide similar information.
- The total_count, casual, and registered variables are also highly positively correlated with each other.

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: I used scatter plots to see if the relationship between the independent and dependent variables is linear.
- Independence: I checked if the residuals (errors) are independent.
- Homoscedasticity: I plotted the residuals against the predicted values to check if they have constant variance. If they are spread evenly, the assumption holds.
- Normality of residuals: I used a Q-Q plot to check if the residuals follow a normal distribution. A straight line in the Q-Q plot shows normally distributed errors.

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)

- Temperature
- Weather
- Year

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 utilized for modeling the relationship between a dependent variable and one or more independent variables. The objective is to identify the best-fitting line (or hyperplane in multiple regression) that minimizes the sum of the squared differences between the actual and predicted values. In simple linear regression, the model is represented by:

$$Y = \beta_o + \beta_1 X_1 + \epsilon$$

Where:

- Y represents the dependent variable (target).
- β_o denotes the intercept (the value of Y when X_1 is 0).
- β_1 signifies the coefficient for the independent variable X_1.
- ϵ symbolizes the error term (residual).

In multiple linear regression, the model extends to include more than one predictor variable, and the equation becomes:

$$Y = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta n Xn + \epsilon$$

The algorithm estimates the coefficients (β_o, β_1, …, βn) using methods such as Ordinary Least Squares (OLS) to minimize the residual sum of squares (RSS). Linear regression assumes that the relationship between the dependent and independent variables is linear, errors are normally distributed, and there is no multicollinearity among the independent variables.

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 consists of four datasets that have nearly identical simple descriptive statistics (mean, variance, correlation), yet the datasets appear vastly different when graphed. The datasets are designed to demonstrate the importance of visualizing data before analyzing it statistically. Here's a breakdown:

- Dataset 1: Shows a perfect linear relationship between the variables.
- Dataset 2: Shows a strong linear relationship but with one outlier.
- Dataset 3: Appears to have no relationship between the variables, despite having the same correlation.
- Dataset 4: Shows a curvilinear relationship between the variables.

Anscombe's quartet highlights that summary statistics alone cannot fully capture the underlying patterns in the data and emphasizes the importance of data visualization.

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 called the Pearson correlation coefficient) is a measure of the linear relationship between two variables. It ranges from -1 to +1:

- +1 indicates a perfect positive linear relationship (as one variable increases, the other also increases).
- -1 indicates a perfect negative linear relationship (as one variable increases, the other decreases).
- 0 indicates no linear relationship between the variables.

Pearson's R is computed as the covariance of the two variables divided by the product of their standard deviations:

$$r=\text{cov}(X,Y)/\sigma X\sigma Yr$$

It is widely used in statistics to assess the degree of linear association between two variables.

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 the process of transforming data to fit within a specific range or distribution, making the features comparable. It's important because many machine learning algorithms are sensitive to the scale of input features. Without scaling, features with larger numerical ranges can dominate the learning process.

- **Normalized Scaling (Min-Max scaling)**: This technique scales the data to a fixed range, usually [0, 1], using the formula:

$$Xnorm= (X−Xmin) / (Xmax−Xmin)$$

Normalization is useful when the data needs to be within a specific range.

- **Standardized Scaling (Z-score scaling)**: This technique scales the data to have a mean of 0 and a standard deviation of 1, using the formula:

$$X_{std} = (X - \mu)/\sigma$$

Standardization is useful when the data is normally distributed and we want to maintain the distribution's shape.

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 a regression coefficient is inflated due to collinearity with other independent variables. A VIF value of infinity typically happens when there is perfect multicollinearity, meaning one variable is a perfect linear function of another variable. This results in the matrix inversion process used to calculate the coefficients failing, leading to an infinite VIF. This issue can be detected using tools like correlation matrices or VIF itself and is addressed by removing or combining the highly collinear 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)

A Q-Q (Quantile-Quantile) plot is a graphical tool used to assess if a dataset follows a specific theoretical distribution, such as the normal distribution. It compares the quantiles of the sample data with the quantiles of a reference distribution (usually normal). In a Q-Q plot:

If the points lie along a straight line, the data follows the theoretical distribution.
If the points deviate from the straight line, it suggests that the data does not follow the expected distribution.

In the context of linear regression, a Q-Q plot is used to check the assumption that the residuals (errors) are normally distributed. This is important because many inference procedures in linear regression rely on this assumption. If the residuals are not normally distributed, it can affect the validity of confidence intervals and hypothesis tests.
