By Akshay Thakur
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 weathersit significantly affect the target variable cnt. For example, weathersit unit increase shows a decrease in bike hires by 0.3282 units. Other variables, such as season, also play an important role, affecting the model's R-squared values and explaining a portion of the variance in the data. These categorical variables help to capture how external factors like weather conditions and seasons influence bike-sharing demand, indicating that demand is significantly impacted by these factors.

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 prevent multicollinearity and redundant variables in the model. When creating dummy variables, each category is represented as a binary feature (0 or 1). If all categories are included, one category will be perfectly predicted by the others, leading to perfect multicollinearity. By dropping one reference category, we avoid this issue and ensure that the model's coefficients remain interpretable without redundancy.

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)

Initially, registered had the highest correlation with the target variable cnt at 0.95. However, after addressing multicollinearity, atemp (apparent temperature) emerged as the feature with the highest correlation (0.63) with cnt, indicating its stronger influence on the demand for shared bikes once multicollinearity was accounted for.

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)

After building the linear regression model, I validated its assumptions by:

Checking for Linearity: I ensured that the relationships between independent variables and the dependent variable were linear.

Normality of Errors: I used graphical methods such as Q-Q plots to check that the residuals followed a normal distribution.

Residual Analysis: I plotted residuals against predicted values to confirm homoscedasticity, ensuring that the residuals had constant variance.
Multicollinearity: I used Variance Inflation Factor (VIF) values and p-values to detect and address multicollinearity, ensuring that no predictors were highly correlated.

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)

Based on our final model, the top three predictor variables influencing bike bookings are as follows:

Temperature (temp): With a coefficient value of 0.4411, a unit increase in the temperature variable corresponds to an increase of 0.4411 units in bike hire numbers.

Weather Situation (weathersit) (Light Snow, Light Rain + Thunderstorm + Scattered Clouds, Light Rain + Scattered): A coefficient value of -0.3282 indicates that, relative to Weather Situation 3, a unit increase in this variable results in a decrease of 0.3282 units in bike hire numbers.

Year (yr): A coefficient value of 0.2310 signifies that a unit increase in the year variable leads to an increase of 0.2310 units in bike hire numbers.

It is recommended to prioritize these variables during planning to maximize bike bookings effectively

---

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 modeling technique used to predict the value of a dependent variable (target) based on the values of independent variables (predictors). The model assumes that there is a linear relationship between the independent and dependent variables, which can be represented by the equation:

$Y = \theta_0 + \theta_1X_1 + \theta_2X_2 + ... + \theta_nX_n$

Where:

Y is the dependent variable,
$X_1, X_2, ..., X_n$ are the independent variables, and
$\theta_0, \theta_1, ..., \theta_n$ are the coefficients to be estimated.
The model fits a line to the data by minimizing the sum of squared residuals (the difference

between the observed and predicted values). The coefficients are found using a method such as Ordinary Least Squares (OLS). The quality of the model is assessed using metrics like R-squared and p-values.

The algorithm assumes linearity, normality of residuals, independence of errors, and homoscedasticity (constant variance of 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 consists of four different datasets that have nearly identical descriptive statistics, such as mean, variance, and correlation, but differ significantly when plotted. The quartet was created to demonstrate the importance of visualization in data analysis. The four datasets include:

- A linear relationship between the variables.
- A quadratic curve, indicating a non-linear relationship.
- A linear relationship with one influential outlier that strongly affects the results.
- A dataset with no variation in the dependent variable, represented by a vertical line.

Anscombe's quartet teaches that relying on summary statistics alone can be misleading, and visualizing data is crucial to uncover 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, measures the strength and direction of the linear relationship between two variables. The coefficient ranges from -1 to +1:
A value of +1 indicates a perfect positive linear correlation.
A value of -1 indicates a perfect negative linear correlation.
A value of 0 indicates no linear correlation.
Pearson's R is commonly used in linear regression to assess the strength of the relationship between predictor variables and the target variable. It assumes that both variables are continuous and approximately normally distributed.

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 transforming features to ensure they are on a comparable scale, making them suitable for machine learning algorithms. It is particularly important when the features have different units or ranges, as algorithms like linear regression are sensitive to the magnitude of input variables.

- **Normalization:** This technique rescales features to a range between 0 and 1, by subtracting the minimum value and dividing by the range.
- **Standardization:** This method transforms the data to have a mean of 0 and a standard deviation of 1. It is useful when the data follows a Gaussian distribution.

The choice between normalization and standardization depends on the nature of the data and the algorithms used.

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>

An infinite VIF occurs when there is perfect multicollinearity between two or more independent variables, meaning that one predictor can be expressed as a linear combination of others. In such cases, the model cannot uniquely estimate the coefficients, leading to instability and infinite variance for the corresponding predictor. This usually indicates a redundancy among predictors, and addressing this issue may involve removing one of the highly correlated predictors.

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 whether the residuals of a model follow a specific distribution, typically the normal distribution. It compares the quantiles of the residuals with the quantiles of a normal distribution. If the residuals are normally distributed, the points in the Q-Q plot will align closely with the diagonal line.

In linear regression, a Q-Q plot is essential for validating the assumption of normality of residuals. If the residuals are not normally distributed, it can indicate model specification errors, the need for a transformation, or the presence of outliers that may affect the model's reliability.
