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)

Based on box plot, season and weather had impact on rental, clear and fall season was the most favored one.

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)

This is because we want to define a base state, which allows us to save one extra column representing that state. When drop_first=True is used, the base state can be inferred when the rest of the columns are 0. This reduces the number of parameters in the model, making it simpler and improving the accuracy of metrics like adjusted R-squared by not adding unnecessary predictors.

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)

Temperature and Perceived Temperature have the highest and similar collinearity with the rental.

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 model on the training set, I validated the assumptions of Linear Regression as follows:

1. Normality of Errors: I checked the normality of residuals (the difference between actual and predicted values) to ensure they are approximately normally distributed, a key assumption for linear regression. This was done using a histogram or Q-Q plot of the residuals.

2. Patterns in Residuals: I created a scatter plot of residuals vs. predicted values to check for any patterns. Ideally, residuals should be randomly scattered around zero with no discernible pattern, indicating that the model correctly captures the relationship in the data.

3. Homoscedasticity: I looked for constant variance in the residuals across all levels of predicted values (homoscedasticity). A consistent spread in the residual scatter plot confirms this assumption, whereas a funnel shape might indicate issues with the model.

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, year and humidity

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)

The linear regression algorithm assumes that there is one target variable (y) and one or more predictor variables (x). If there is one target variable and one predictor variable, you can find a line that minimizes the residuals (differences between actual values and predicted values). In this case, the model aims to find the line with the smallest possible sum of squared residuals, which is also called the "line of best fit."

If there are two predictor variables, then the model finds a plane in three-dimensional space where all the values of x1 and x2 (predictor variables) best predict the value of y (target variable). This concept extends to cases with multiple predictor variables, where the model finds a hyperplane in an n-dimensional space that best fits the data points.

In general, the model can be represented as: y = b0 (intercept) + b1 * x1 + b2 * x2 + ... + bn * xn, where each "b" (beta) represents a coefficient for a predictor variable, showing its contribution to the prediction of y. The algorithm finds the optimal values for these betas (coefficients) by minimizing the sum of squared residuals (errors), which gives the best-fit equation for predicting y from the predictors. This method of finding the best-fit line, plane, or hyperplane is called Ordinary Least Squares (OLS).

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 illustrates that four datasets can have the same mean, standard deviation, and correlation, but vastly different distributions, which impacts the understanding of linearity.

1. First Set: This is a classic example of a dataset that fits a linear relationship well. The data points align closely along a straight line, making linear regression appropriate.

2. Second Set: This dataset has a curved relationship. While a linear regression line can be drawn, it does not capture the pattern accurately, as the data is not truly linear.

3. Third Set: Although there is a linear trend, a single outlier has a strong influence on the

regression line, skewing the summary statistics. This demonstrates how outliers can impact the model's accuracy.

4. Fourth Set: Although the correlation and linear regression line appear significant in summary statistics, the x-values lack spread and are almost constant. This makes the data unsuitable for reliable linear modeling since it lacks variability.

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 coefficient R is a value ranging from -1 to 1 that measures how strongly two variables are correlated and how changes in one affect the other. If R is positive, it means that as one variable increases, the other does too. If R is negative, when one variable increases, the other decreases. The closer R is to 1 or -1, the stronger the relationship, meaning one variable's change is more likely to impact the other due to their correlation.

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 means adjusting values to a different scale, either up or down, while keeping the original information intact. For example, 1200 grams can be scaled down to 1.2 kg to make the number easier to work with but still hold the same meaning. In MLR models, we scale variables to similar levels to better understand their impact. When predictors are on a similar scale, it's easier to see how each one influences the target variable, allowing us to compare them and assess which predictors have stronger effects.

For instance, the area's impact on house rent might appear different if it's measured in square meters versus square kilometers—it might look far more sensitive in square kilometers. By scaling, we bring all predictors onto the same scale, which removes this confusion and helps in making fair comparisons.

Difference between Normalized Scaling and Standardized Scaling:

- Normalization: This scales data to a fixed range, often 0 to 1, keeping each value in proportion to the minimum and maximum values. It's useful when you want all values to fall within a specific range.
- Standardization: This scales data based on the mean and standard deviation, giving it a mean of 0 and a standard deviation of 1. This method is useful when you need a consistent spread, especially in models sensitive to the distribution of values, like regression.

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)

This means that there is a perfect collinearity with one or more variables. This will occur when one variable is a exact linear combination of other 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)

In linear regression, we assume that residuals are normally distributed. A Q-Q plot helps us see if this is true by showing how closely the residuals follow a normal distribution. It generates a diagonal that represent a perfectly normal (or some other) distribution points and then plots our data (in this case errors) to help us see how well they are distributed. This is a good way to make sure our MLR assumption for normality is correct.
