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)

We see a month-specific seasonal trend on bike sharing demand. Demand is highest from May to October, driven by warmer, favorable weather, while January and February see lower usage due to colder conditions. Also, the analysis shows the growing popularity in demand of sharing bikes in 2019 compared to 2018. Like the months, we can see favorable weather conditions (clear or partly cloudy days) also boost demand. While the median user count is same across all the 7 days of the week, we see a higher distribution of demand on the 3rd and 6th days of the week, which are Friday and Monday as per data.

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** when creating dummy variables helps prevent including too much information in the model. Dropping the first category gives the model just enough information, making it simpler and easier to understand.

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)

Registered user count has the highest correlation followed by casual user count. Since these 2 variables make up the target variable, it's obvious to see this correlation. Other than these, we can see a positive correlation 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)

To check if the Linear Regression assumptions were met, residuals (differences between actual and predicted values) were calculated. First, we checked if the residuals followed a normal distribution, showing that errors were spread evenly. Then, we plotted residuals against predicted values to make sure they were randomly scattered, confirming linearity and constant variance. These checks helped ensure our model was reliable.

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 with absolute coefficient 3895; Weather Situation: 3395 and Month: 2354

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 an algorithm used to predict a continuous target based on one or more input variables by finding the best-fitting line. The algorithm aims to find values for the intercept and slope (coefficients) that minimize the difference between predicted and actual values. This difference is minimized using the Ordinary Least Squares (OLS) method, which calculates the line that has the smallest possible sum of squared errors. Linear Regression makes some assumptions, such as a linear relationship between input variables and the target, constant error variance, and normally distributed errors. Once the model is built, each variable's coefficient shows how much the target changes with a one-unit increase in that variable. The model's performance is then evaluated using metrics like R-squared or Mean Squared Error (MSE) to check prediction accuracy.

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 statistical properties (like mean, variance, and correlation) but look very different when graphed. Each dataset tells a unique story: one is linear, another is curved, one has an outlier, and the last has no relationship at all. This shows that summary statistics alone can be misleading, as they don't capture the true nature of the data. The quartet emphasizes the importance of visualizing data to understand its structure and relationships fully.

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 a linear relationship between two variables. It ranges from -1 (perfect negative correlation) to +1 (perfect positive correlation), with 0 meaning no linear relationship. A positive value means that as one variable increases, the other does too, while a negative value means one variable decreases as the other increases. Pearson's R only captures linear relationships, so it won't detect non-linear patterns between 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)

<Your answer for Question 9 goes here>

Scaling is a way to adjust numbers in a dataset so they fall within a similar range, often between 0 and 1, or around an average of 0. This helps models perform better, as large differences in numbers can cause some variables to have too much influence. Scaling also makes calculations faster, especially for models that measure distance between data points. Normalization (or Min-Max Scaling) squeezes values into a specific range, usually 0 to 1. Standardization centers values around 0 and adjusts them to have a similar spread, which is helpful if data resembles a bell curve (normal distribution) Formula: Min-Max Scaling = X – Xmin / Xmax – Xmin | Standardized scaling = X – Xmean / Xstd

---

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>

A VIF of infinity happens when one independent variable is perfectly correlated with one or more other variables, a situation known as perfect multicollinearity. This causes the VIF calculation to divide by zero, resulting in an infinite value. Common causes include duplicate variables or including all categories of a categorical variable. To fix this, remove insignificant variables or drop one category from categorical variables to reduce multicollinearity.

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 is a graph that compares the distribution of data to a theoretical distribution, usually the normal distribution. In linear regression, it helps check if the residuals (errors) follow a normal distribution, which is an important assumption for accurate predictions. If the points in the Q-Q plot form a straight line, it means the residuals are likely normal; if they deviate, the residuals may not be normal. This helps identify if adjustments are needed to make the model more reliable.
