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)

From the categorical variable year 1, i.e 2019 has very good rentals.
Season spring is negatively impacted the rentals.
In mid of the year months rentals are better than start and end of the year
Surprisingly holiday/working day are not so much impacted the rentals

---

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)

To avoid "dummy variable trap", which is resulting n columns for n categories
Including all categories leads to multicollinearity issues, which can inflate the variance of coefficient estimates and make them unstable

---

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)

Column "registered" is having the highest correlation with the target variable

---

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)
Plotting the predicted values against residuals should be linear
When plotted histogram error(y_pred-y_train), normal distribution, centered at zero.
Graph when plotted should not have any significant pattern.

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)

In my final model 'casual, 'year' and 'temp' contribute

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 by fitting a linear equation. Here depedent variables can be any number.

y=B0 + B1x1+ B2+x2… BnXn

Here idea is that algorithm minimizes the sum of the squared differences between observed and predicted values, known as the residuals, using the least squares method.

Mainly for linear Regression assumption must be satisfied.
1. Error must be normal distributed.
2. Values must be centered at zero
3. No significant pattern should be seen
4. Graph must be scattered around the fitted line.

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 datasets that have nearly identical simple descriptive statistics but very different distributions and relationships when graphed. Each dataset contains 11 pairs of x and y values. Despite having the same mean, variance, and correlation coefficient, visual inspection reveals that they exhibit distinct patterns:

The first dataset shows a linear relationship.
The second has a curved relationship.
The third contains an outlier that affects the linearity.
The fourth features a vertical line, indicating no relationship.

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, also known as the Pearson correlation coefficient, measures the linear correlation between two continuous variables. It ranges from -1 to +1:

+1 indicates a perfect positive linear correlation, meaning as one variable increases, the other also increases.

-1 signifies a perfect negative linear correlation, where one variable increases as the other decreases.

0 suggests no linear correlation.

Pearson’s R is calculated by dividing the covariance of the two variables by the product of their standard deviations. It's sensitive to outliers, so it's essential to visualize data before relying solely on this coefficient for correlation analysis.

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>

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>

Scaling is the process of transforming features to a similar range or distribution, which is essential in machine learning to improve model performance and convergence speed. It ensures that no single feature dominates others due to its scale

Normalized Scaling: Rescales the data to a fixed range, typically [0, 1]. This is done by subtracting the minimum value and dividing by the range (max - min). It’s useful when the distribution is not Gaussian and focuses on proportions.

Standardized Scaling: Centers the data around the mean with a unit standard deviation, resulting in a distribution with a mean of 0 and a standard deviation of 1.

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 (Quantile-Quantile) plot is a graphical tool used to assess if a dataset follows a specific theoretical distribution, typically the normal distribution. It compares the quantiles of the dataset against the quantiles of the reference distribution.

Use of Q-Q Plot:

Normality Check: In linear regression, one of the key assumptions is that the residuals (errors) are normally distributed. A Q-Q plot helps visually check this assumption.

Identifying Outliers: It can reveal outliers or deviations from normality that may influence model performance.