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)

Season: Higher activity in spring and summer, with winter often showing lower demand or higher default rates.

Weather Situation: Poor weather reduces demand or could indirectly correlate with higher defaults due to economic strain.

Month: December may see higher defaults due to holiday spending; early months might show better financial planning.

Weekday: Weekdays generally have higher, more consistent activity, while weekends may show reduced engagement**.**

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)

It avoids multicollinearity, which occurs when one dummy variable can be perfectly predicted by the others. By dropping the first category, we ensure that the dummy variables are independent of each other, preventing 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)

➔ atemp

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)

I reviewed the **Variance Inflation Factor (VIF)** for each predictor VIF value less than 5 indicate its good . High VIF values (typically above 5 or 10) indicate multicollinearity, which may need to be addressed by removing or combining variables.

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)

1) Year
2) Season -Spring
3) Windspeed

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 data analysis technique that predicts the value of unknown data by using another related and known data value.
→ A linear regression model describes the relationship between a dependent variable, y, and one or more independent variables, X. The dependent variable is also called the response variable. Independent variables are also called explanatory or predictor variables.
→ step 1) Generate or load data.
Step 2) Split the data into training and testing set
Step 3) Create the linear regression model
Step 4) Train the model on the training data
Step 5) Make Predictions using the test data
Step 6) Visualize the data and the regression line
Step 7) Evaluate the model

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 comprises a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph.

Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

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)

The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

It's measure of linear association between the two variables: r^2 is the proportion of the total variance (s^2) of Y that can be explained by the linear regression of Y on x. $1-r^2$ is the proportion that is not explained by the regression.

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 the range and distribution of data values to a standard scale. It's an essential part of linear regression optimization because it can affect the speed, accuracy, and stability of the optimization algorithm.
- ➔ 1) Improve the model performance , 2) Enhances convergence speed & 3) Facilitates Distance calculation

➔ Difference Between Normalized Scaling and Standardized Scaling

1. **Normalized Scaling (Min-Max Scaling)**:
    - Transforms features to a fixed range, typically [0, 1].
    - Useful when you need a bounded range, especially for neural networks.
2. **Standardized Scaling (Z-score Scaling)**:
    - Centers the data around the mean with a standard deviation of 1.
    - Useful for datasets with normally distributed features.

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)
A VIF value becomes infinite when there is perfect multicollinearity among predictor variables, meaning one variable is a perfect linear combination of others. This makes it impossible to separate their effects, leading to inflated standard errors. Removing one of the collinear variables usually resolves this issue.

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)

Q–Q plots are commonly used to compare a data set to a theoretical model. This can provide an assessment of goodness of fit that is graphical, rather than reducing to a numerical summary statistic. Q–Q plots are also used to compare two theoretical distributions to each other.

It’s important in linear regression because it helps determine if data sets come from populations with the same distribution: