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)

Analyzing the categorical variables reveals how they influence the dependent variable. For instance:
Weathersit: The dataset lacks records for extreme conditions like heavy rain or snow. While these events are rare, they could negatively impact the dependent variable when they occur.
Workingday: There is a notable increase in bike rentals on working days compared to non-working days.
Season: The data appears evenly distributed across different seasons.
Year and Month: The dataset is balanced across different years, ensuring no significant bias in temporal trends.

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 helps eliminate redundancy by creating k−1k-1k−1 dummy variables from a categorical variable with k levels. This prevents multicollinearity in the model, ensuring a more stable and interpretable regression output.

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)

Among numerical variables, temperature (temp or atemp) shows the strongest correlation with the target variable, indicating that bike rentals increase with temperature.

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 validate the assumptions of linear regression after training the model:

Residual Analysis: A histogram of residuals was plotted against fitted values, confirming an approximately normal distribution.
Multicollinearity Check: Variance Inflation Factor (VIF) was used to detect multicollinearity, ensuring most features had VIF < 5 and p-values < 0.05. Temperature had a slightly higher influence but remained within acceptable limits.

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)

The three most significant features influencing bike demand in the final model are:

Temperature (temp) – Higher temperatures correlate with increased rentals.
Year (yr_2019) – Rentals were higher in 2019 compared to 2018.
Weather Situation (weathersit_Light Snow/Rain) – Light rain or snow negatively impacts demand.

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)

< It is a supervised learning algorithm used for predicting continuous values based on input features. It establishes relationship between a dependent variable (target) and one or more independent variables (features) by fitting a straight line through the data points. It minimizes the sum of squared differences between actual and predicted values, assuming linearity and no multicollinearity among predictors
>

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 consists of four datasets that share identical statistical properties (such as mean, variance, and correlation) but have drastically different visual distributions. It highlights the importance of data visualization in statistical analysis.

Dataset 1: Displays a simple linear relationship with minor variance.
Dataset 2: Shows a non-linear pattern, implying that a linear regression model is not suitable.
Dataset 3: Appears linear but contains a single influential outlier.
Dataset 4: Data points are scattered randomly, yet the statistical measures resemble the other datasets.
This demonstrates that relying solely on summary statistics can be misleading without graphical analysis.

>

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 correlation coefficient (Pearson's R) measures the strength and direction of a linear relationship between two continuous variables. Its values range from -1 to 1, where:

+1 indicates a perfect positive correlation,
-1 indicates a perfect negative correlation,
0 signifies no linear relationship.
It is commonly used in various fields, such as economics, finance, and social sciences, to assess relationships 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)

< Scaling is the process of transforming numerical data to ensure that all features contribute equally to a model. It improves performance and numerical stability.

Reasons for Scaling:

Prevents larger values from dominating smaller ones.
Speeds up convergence in gradient-based algorithms.
Enhances the performance of distance-based models (e.g., KNN, SVM).
Improves interpretability by ensuring all variables have comparable ranges.
Difference Between Normalization and Standardization:

Normalization (Min-Max Scaling): Rescales values to a fixed range, typically [0,1] or [-1,1]. It is sensitive to outliers.
Standardization (Z-Score Scaling): Centers data around zero mean (0) and unit variance (1), making it less sensitive to outliers and preserving the shape of the original distribution.>

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 Variance Inflation Factor (VIF) value becomes infinite when perfect multicollinearity exists—meaning one predictor is an exact linear combination of others. This occurs when:

Two columns contain identical values.
A feature is completely predictable from other variables in the dataset.
To resolve this, redundant variables should be removed or transformed.>

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)

< A Q-Q (Quantile-Quantile) plot is a graphical method used to compare the distribution of a dataset against a theoretical distribution (often normal). It is particularly useful in:

Checking Normality: If residuals in a regression model follow a normal distribution.
Identifying Outliers: Extreme deviations from the reference line suggest outliers.
Detecting Skewness and Heavy Tails: If data points systematically deviate from the diagonal line.
In linear regression, normality of residuals is an important assumption. A Q-Q plot helps validate this assumption, ensuring model reliability.>