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 weather variable, if weather situation is light_snow and season is spring there is decrease in bike hiring.

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)

Dummy encoding creates binary columns for each category of a categorical variable.
By default, one dummy column is created for each category (which can cause redundancy).
drop_first=True removes one dummy column, preventing the dummy variable trap.

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)

temp and atemp has high correlation with cnt 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)

Plotted the heat map for correlation.
VIF for collinearity.
Plotting histogram on error term, normally distributed with mean zero.
Scatter plot between y-actual and y-predicted 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)

Temperature (temp)
weathersit_Light Snow
Year (yr) are the 3 features contributing significantly.

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 supervised learning algorithm used for predicting continuous values based on independent variables. It establishes a linear relationship between input (X) and output (Y).

Formula for Simple Linear Regression (one independent variable):

Y=mX+c

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 (mean, variance, correlation, regression line) but drastically different distributions when plotted graphically. It was created by Francis Anscombe in 1973 to emphasize the importance of visualizing data rather than relying only on summary statistics.

🔷 **It shows that statistics alone can be misleading. Even if datasets have the same mean, variance, correlation, and regression line, their actual distributions can be completely different.**

🔷 **It highlights the need for data visualization (scatter plots, box plots, histograms, etc.) before making conclusions.**

🔷 **It demonstrates the importance of checking for outliers, trends, and patterns rather than blindly trusting summary statistics.**

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 **correlation coefficient (r)**, also known as **Pearson's R**, is a **statistical measure** that quantifies the **strength and direction** of a **linear relationship** between two continuous variables. It is one of the most commonly used correlation measures.

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 the process of transforming numerical features in a dataset so that they fall within a specific range. It is essential in machine learning because many algorithms perform better when features are on similar scales.
Improves Model Performance → Many machine learning algorithms (e.g., gradient descent, KNN, SVM) perform better and faster when numerical values are on a similar scale.
Prevents Features from Dominating → Features with large values can dominate smaller ones, affecting model accuracy.
Improves Convergence → Optimization algorithms like gradient descent converge faster when data is scaled.
Better Distance-Based Calculations → Algorithms like K-Means clustering, KNN, and PCA rely on distance metrics, which are sensitive to different scales.
Avoids Numerical Instability → Some models (e.g., Linear Regression) may produce incorrect coefficients if features have very different magnitudes.

Normalization (Min-Max Scaling) is useful when you want data between 0 and 1 and when the range is known.
Standardization (Z-score Scaling) is better when dealing with outliers or unknown ranges.
Scaling is necessary for distance-based models and gradient-based optimization.

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>

Variance Inflation Factor (VIF) measures how much the variance of a regression coefficient is inflated due to multicollinearity (high correlation between independent variables).

A VIF value is considered infinite (∞) when one variable is a perfect linear combination of other variables in the dataset. This means that the variable can be exactly predicted from other independent variables, making the denominator of the VIF formula zero, leading to an undefined (infinite) value.

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 **compare the distribution of a dataset to a theoretical distribution** (usually a normal distribution). It helps assess **whether a variable follows a particular distribution** by plotting **quantiles** of the sample data against **quantiles** of the reference distribution.

Linear regression assumes that residuals (errors) are normally distributed. A Q-Q plot helps validate this assumption:

✓ If residuals follow a normal distribution → Model assumptions hold → Reliable predictions.
❌ If residuals deviate from normality → Model may be misspecified, and predictions may be unreliable.