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)

- Year 2019 has a higher count of bike Rentals compared to 2018
- Bike Rentals are higher when the weather situation is Clear. The Rentals drop as it becomes Cloudy or Light Rain/Snow
- Bike Rentals peak during the fall season and are at lowest during the spring

---

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)

By setting **drop_first=True**, you remove one of the dummy variables, usually the first or reference category. This effectively reduces the number of predictors and ensures that the remaining variables represent the difference relative to the reference category. As a result, the model is more stable, interpretable, and free from multicollinearity.

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 field has the highest correlation with target variable **cnt**

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)

Linear relationship scatterplot between predictors and target

Error terms are normally distributed: Residual Analysis to find if Errors Terms have a normal distribution

Error terms have constant variance: Scatter Plot for spread of residuals over index to ensure the variance is constant and centered towards zero.

Error terms are independent of each other: Scatter Plot to understand the spread of residuals across fitted values. The spread is random and does not have any patterns.

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)

Temp has a positive coefficient of 0.4667

Yr (Year) has a positive coefficient of positive 0.234

Weathersit_3 LightRainSnow has a negative coefficient of 0.279

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 statistical method used to model the relationship between a dependent variable (target) and one or more independent variables (predictors). It assumes a linear relationship between the variables, expressed as a straight line (or hyperplane in higher dimensions). The algorithm estimates the coefficients (weights) of the model by minimizing the sum of squared residuals (differences between predicted and actual values). This is typically done using methods like Ordinary Least Squares (OLS). Once the model is trained, it can predict the target variable for new input data by applying the learned coefficients.>

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 is a set of four datasets that have identical simple descriptive statistics (mean, variance, correlation, etc.) but vastly different distributions and relationships between variables. Despite having the same correlation and regression line, the datasets display different patterns, highlighting that summary statistics alone can be misleading. Each dataset consists of two variables, and when plotted, they reveal distinct patterns like linearity, curvilinearity, or outliers. Anscombe's quartet emphasizes the need for both statistical analysis and data visualization.>

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 **r** (correlation coefficient) is a measure of the linear relationship between two continuous variables. It ranges from **-1** to **+1**, where **+1** indicates a perfect positive linear correlation, **-1** indicates a perfect negative linear correlation, and **0** indicates no linear correlation. The closer the value is to **+1** or **-1**, the stronger the linear relationship. Pearson's r assumes that the relationship is linear and that both variables are normally distributed. It is widely used in statistics to assess the strength and direction of a linear relationship.>

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 features to ensure they are on a similar scale, which helps improve the performance of machine learning algorithms. **Normalization** (Min-Max scaling) resizes data to a specific range, typically [0, 1], while **Standardization** (Z-score normalization) centers data around 0 and scales it to have a standard deviation of 1. Normalization is used when a fixed range is needed, and standardization is used when the model assumes a normal distribution of data. Both techniques help improve convergence speed and model accuracy.>

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)** can be infinite when there is perfect multicollinearity between predictor variables, meaning one predictor is a perfect linear function of another. This occurs when a variable can be exactly predicted by other variables in the model, causing the model to have redundant information. In such cases, the matrix used in the regression becomes singular, and the inverse cannot be computed, leading to an infinite VIF. This suggests that the model is not stable, and the predictor variables are highly correlated, making it difficult to estimate unique effects. To resolve this, you might need to remove or combine collinear variables>

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 tool used to compare the distribution of a dataset to a theoretical distribution, often a normal distribution. It plots the quantiles of the data against the quantiles of the theoretical distribution. In linear regression, a Q-Q plot is important for assessing the **normality of residuals**—a key assumption for valid inference. If the points in the plot fall along a straight line, it indicates that the residuals are normally distributed. Deviations from the line suggest non-normality, which may indicate model issues or the need for data transformation >
