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)

In this analysis, the following categorical variables are considered:

1. **Season**: Spring, Summer, Fall, and Winter
2. **Weather Situation (Weathersit)**: Clear, Mist, Light Snow, and Heavy Rain

The target variable, **cnt**, represents the total count of rental bikes (including both casual and registered users). The effects of these categorical variables on the target variable, **cnt**, are summarized as follows:

Season (Categorical Variable):

- The highest average **cnt** is observed in the **Fall** season, followed by **Summer** and **Winter**.
- The lowest average **cnt** occurs in **Spring**.

Weathersit (Categorical Variable):

- The highest average **cnt** is observed during **Clear** weather, followed by **Mist**.
- The lowest average **cnt** is recorded during **Light Snow**.
- No rentals (i.e., **cnt** = 0) are observed during **Heavy Rain**

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)

drop_first=True drops one of the dummy variables, leaving n−1 variables.

- This ensures that the dummy variables are **independent** of each other.
- It facilitates proper model interpretation, as **coefficients** of dummy variables are relative to a reference category
- It enhances **efficiency** by reducing the number of variables included in the model

---

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)

Highest correlation among the numerical variables with Target variable **cnt** is
atemp – 0.65

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)

- Use scatterplots for each independent variable against the dependent variable to check for linearity.
- Validate The relationship between the independent variables and the dependent variable is linear
- Evaluate R-Square and adjusted R-square.
- Analyze the p-values of coefficients to ensure they are statistically significant.
- Calculate the Variance Inflation Factor (VIF) for each independent variable. A VIF value > 5 (or 10) indicates multicollinearity

---

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)

Top 3 factors are
1. Yr
Coefficient: 0.2350 (positive effect) P-value: <0.001 (highly significant) VIF: 2.07 (low multicollinearity)
2. temp
Coefficient: 0.4673 (positive effect) P-value: <0.001 (highly significant) VIF: 5.26 (moderate multicollinearity)
3. Light Snow (Weather Situation)
Coefficient: -0.2838 (negative effect) P-value: <0.001 (highly significant) VIF: 1.08 (no multicollinearity)

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 to model the relationship between one or more independent variables (X) and a dependent variable (Y) by fitting a linear equation to the data. The goal is to predict the value of Y based on the values of X.

The equation of a simple linear regression model (with one independent variable) is:

$$Y=\beta 0+\beta 1X+Error$$

- Y : Dependent variable (target).
- X: Independent variable (feature).
- $\beta 0$: Intercept (value of Y when X=0).
- $\beta 1$: Slope (rate of change in Y per unit change in X).
- Error : captures the variation not explained by the model

For multiple linear regression (with multiple independent variables), the equation becomes:

- $Y=\beta 0+\beta 1X1+\beta 2X2+\cdots+\beta nXn$ + Error

Steps in the Linear Regression Algorithm
1. Define the Problem
2. Fit the Model (Training the Algorithm)
3. Model Evaluation
4. Prediction

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 highlights the critical role of visualization in data analysis. Visualizing data ensures that relationships, patterns, and anomalies are properly understood, leading to more accurate and informed decision-making.
Anscombe's Quartet is a group of four datasets. These datasets are widely used to demonstrate the importance of data visualization.

1. Dataset 1: The data points are scattered around a straight line.
2. Dataset 2: The data points form a perfect quadratic curve.
3. Dataset 3: Most of the data points are clustered, but there is a single outlier.
   This outlier heavily influences the regression line and correlation coefficient, making them misleading.
4. Dataset 4: All data points are aligned vertically, except for one outlier

---

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, is a statistical measure used to quantify the strength and direction of the linear relationship between two continuous variables. It has ranges r from -1 to 1.
The value of r indicates both the strength and direction of the 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)

<Your answer for Question 9 goes here>

Scaling is the process of transforming the values of numerical features in a dataset so that they fall within a specific range or follow a standard distribution. This ensures that all features contribute equally to the analysis, regardless of their original units or scales.

Why Scaling performs
- Scaling ensures all features are on a comparable scale, so that they will make Equal Contribution to the Model
- Preventing Bias

Normalized Scaling - Normalization rescales the data to a specific range, often [0, 1].
Standardized Scaling - Standardization transforms data to have a mean of 0 and a standard deviation of 1

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>

The value of Variance Inflation Factor (VIF) becomes infinite when there is perfect multicollinearity between one independent variable and the other independent variables in a dataset

VIF(Xi) = 1/(1-R-square)
when R-square = 1

The coefficient of determination (R-squared) obtained by regressing (Xi) on all the other independent variables.
when there is perfect linear relationship between (Xi) so the R-square = 1 hence VIF(Xi) become infinity

---

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 Plot is a graphical representation used to compare the distribution of a dataset to a theoretical distribution . It helps assess whether the data follows the expected distribution by plotting the data distribution against the theoretical distribution.

This is used to
Assessing Normality of Residuals
Detecting Skewness
Identifying Outliers