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)

1. **Season:**
   - **season_spring:** Coefficient is -0.1517, p < 0.001. This means that, compared to the reference season, rentals are significantly lower in spring.
   - **season_winter:** Coefficient is 0.0530, p < 0.001. This means that, compared to the reference season, rentals are significantly higher in winter.
   - **Inference:** There are clear seasonal effects on bike rentals. Spring sees lower rentals, and winter sees higher rentals compared to the reference season.

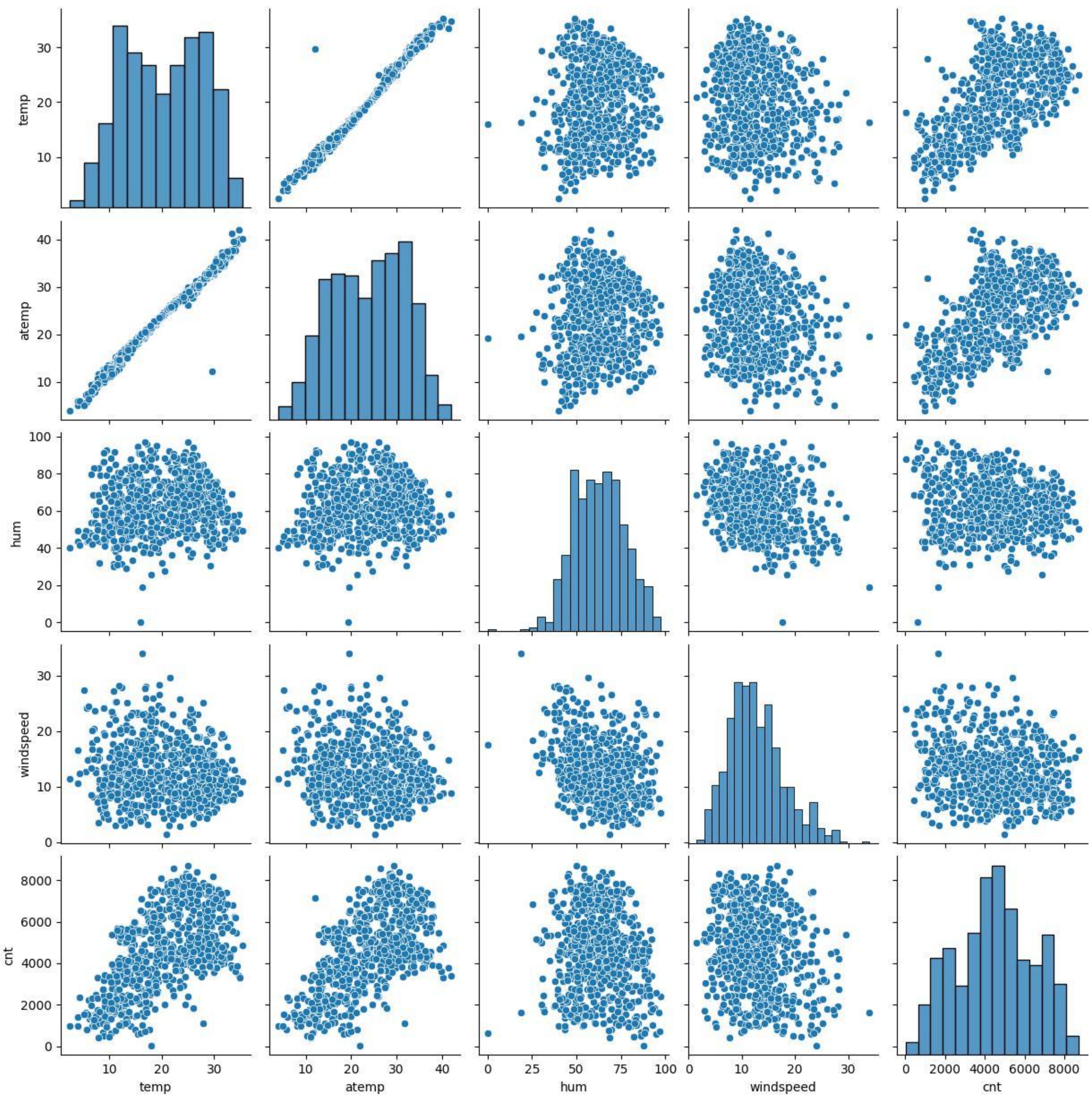
2. **Holiday:**
   - **holiday:** Coefficient is -0.0849, p = 0.002. This means that rentals are significantly lower on holidays compared to non-holidays.
   - **Inference:** Holidays have a negative impact on bike rentals, possibly due to changes in commuting patterns or people engaging in other activities.

3. **Weather Situation (weathersit):**
   - **weathersit_Light Snow/Rain:** Coefficient is -0.2652, p < 0.001. This means that rentals are significantly lower during light snow/rain compared to clear weather.
   - **weathersit_Mist + Cloudy:** Coefficient is -0.0768, p < 0.001. This means that rentals are significantly lower during mist/cloudy weather compared to clear weather.
   - **Inference:** Weather conditions have a substantial impact on bike rentals. Adverse weather conditions (light snow/rain, mist/cloudy) significantly reduce rentals compared to clear weather.

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)

The **drop_first=True** parameter in dummy variable creation (often used with functions like pd.get_dummies() in pandas) is crucial for avoiding a statistical problem called multicollinearity. This ensures the validity of regression analysis when working with categorical variables.

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)



- **Temperature (temp) is a Key Factor**: Both **temp** and **atemp** show a strong positive correlation with **cnt**, suggesting that temperature is a significant factor in bike rental demand.
- **Humidity and Windspeed have Weaker Influence**: **hum** and **windspeed** show weaker correlations with **cnt**. While high humidity and high wind speeds might slightly reduce rentals, their impact is less pronounced than 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)

After building a linear regression model on the training set, validating its assumptions is crucial for ensuring the model's reliability and generalizability. Here's how you can validate the key assumptions of linear regression:

1. **Linearity:** Plotted the residuals (the difference between predicted and actual values) against the predicted values or each independent variable. A random scatter of residuals suggests linearity.
2. **Independence of Errors (No Autocorrelation):** Checked for autocorrelation in the residuals. A value close to 2 indicates no significant autocorrelation. Values significantly below 2 suggest positive autocorrelation, and values significantly above 2 suggest negative autocorrelation.
3. **Homoscedasticity (Constant Variance of Errors):** Plotted the residuals against the predicted values. A random scatter of points with no clear pattern suggests homoscedasticity. A funnel-shaped pattern or other systematic changes in variance indicate heteroscedasticity.
4. **Normality of Errors:** Plot a histogram of the residuals. It should resemble a bell-shaped curve.
5. **No or Little Multicollinearity:** Validated Multicollinearity using VIF and Correlation Matrix.
    a. **Correlation Matrix:** Calculate the correlation matrix of the independent variables. High correlation coefficients (close to +1 or -1) indicate multicollinearity.
    b. **Variance Inflation Factor (VIF):** Calculate the VIF for each independent variable. VIF values greater than 5 or 10 suggest significant 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)

Based on the final model, below are top 3 features predicting the demand of shared bike.

1. **temp (Temperature):** The coefficient is 0.4300 and highly significant ($p < 0.001$). This confirms the strong positive relationship: higher temperatures lead to more rentals.
2. **yr (Year):** The coefficient is 0.2420 and highly significant ($p < 0.001$). This confirms the strong positive trend in bike rentals from 2018 to 2019.
3. **weathersit_Light Snow/Rain and weathersit_Mist + Cloudy**: These weather situation variables have negative and highly significant coefficients. This means that these weather conditions are associated with fewer rentals compared to clear weather.

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 fundamental and widely used statistical technique that models the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data.

The goal of linear regression is to find the "best-fitting" linear relationship between the variables. This relationship is expressed as an equation that can be used to predict the value of the dependent variable based on the values of the independent variables.

Types of Linear Regression:

1. **Simple Linear Regression**:
    - Involves only one independent variable.
    - The equation takes the form: $y = \beta_0 + \beta_1 x + \varepsilon$
        - y: Dependent variable
        - x: Independent variable
        - β_0: Y-intercept (the value of y when x is 0)
        - β_1: Slope (the change in y for a one-unit change in x)
        - ε: Error term
2. **Multiple Linear Regression**:
    - Involves two or more independent variables.
    - The equation takes the form: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + \varepsilon$
        - y: Dependent variable
        - $x_1, x_2, ..., x_p$: Independent variables
        - $\beta_0, \beta_1, \beta_2, ..., \beta_p$: Coefficients
        - ε: Error term

Key Aspects of the Algorithm:

1. **Finding the Best-Fitting Line:**
    - The "best-fitting" line is determined by minimizing the difference between the observed values of the dependent variable and the values predicted by the linear equation.
    - The most common method for this is the "least squares" method.
2. **Least Squares Method:**
    - This method minimizes the sum of the squared residuals.
    - A residual is the difference between the actual value of y and the predicted value of y.
    - By minimizing the squared residuals, the algorithm finds the coefficients (β_0, β_1, etc.) that result in the line that best fits the data.
3. **Assumptions of Linear Regression:**
    - Linearity: The relationship between the independent and dependent variables is linear.
    - Independence: The residuals are independent of each other.
    - Homoscedasticity: The residuals have constant variance.
    - Normality: The residuals are normally distributed.
4. **Evaluating Model Performance:**
    - R-squared: Measures the proportion of the variance in the dependent variable that is explained by the independent variables.

- Adjusted R-squared: Similar to R-squared, but adjusts for the number of independent variables in the model.
- P-values: Used to assess the statistical significance of the coefficients.
- Residual analysis: Used to check the assumptions of linear regression.

5. **Implementation:**
    - Linear regression can be implemented using various programming languages and statistical software, including:
        - Python (scikit-learn, statsmodels)
        - R
        - SPSS
        - Excel

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 group of four datasets that have nearly identical simple descriptive statistics, such as mean, variance, correlation, and linear regression line. However, when you visualize these datasets, they reveal significantly different patterns. This quartet was created by statistician Francis Anscombe in 1973 to demonstrate the importance of graphing data before performing statistical analysis.

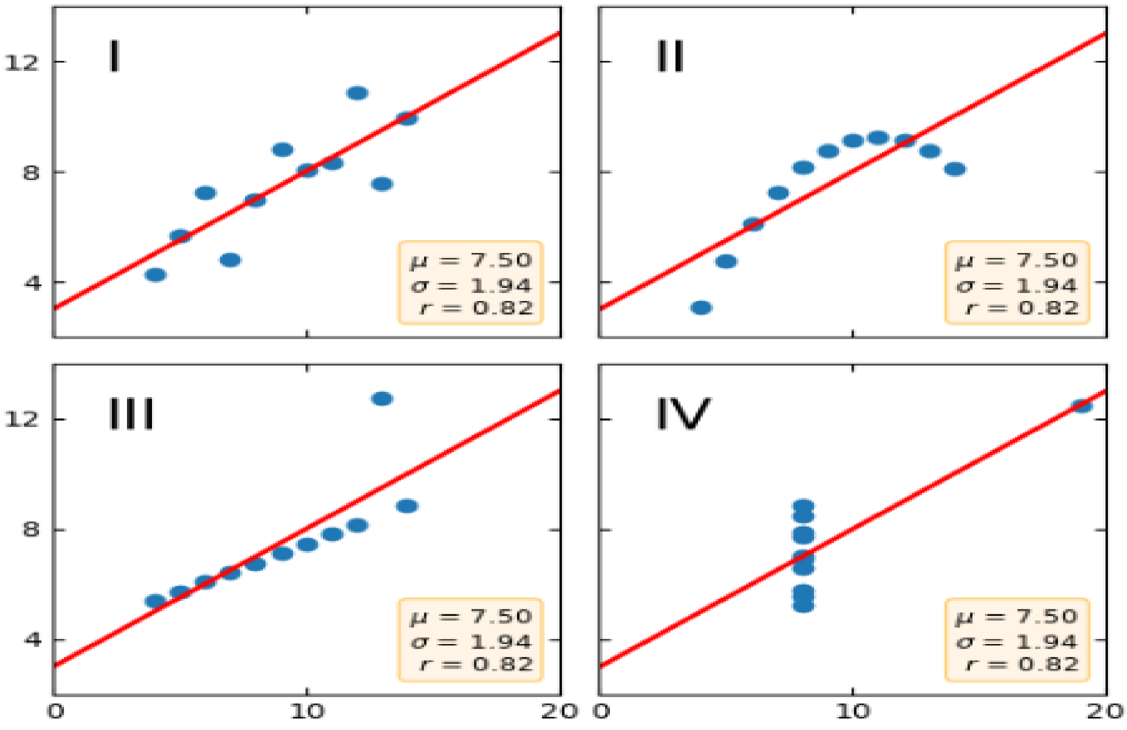
Despite these identical statistics, the four datasets display very different behaviours when plotted.

Dataset 1: A Simple Linear Relationship
Characteristics:
The points in this dataset follow a nearly perfect linear relationship.
The data fits well with the linear regression line, and there are no significant outliers.

Dataset 2: Non-linear Relationship
Characteristics:
This dataset features a non-linear (curved) relationship between x and y.
Although the summary statistics are identical to Dataset 1, the actual relationship is quadratic rather than linear.

Dataset 3: Linear Relationship with an Outlier
Characteristics:
In this dataset, most of the data points lie along a linear trend, but one point is a significant outlier.
The presence of the outlier greatly affects the linear regression line, even though the summary statistics are the same as those in Dataset 1.

Dataset 4: No Relationship with an Outlier
Characteristics:
This dataset appears random, with no clear relationship between x and y.
There is one outlier that significantly influences the correlation and regression line, but without it, there would be no discernible pattern in the data.



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, more formally known as the Pearson correlation coefficient, is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. Here's a breakdown of its key aspects:

What it Measures:

1. **Linear Relationship:**
   - Pearson's R specifically assesses how well a straight line can describe the relationship between two variables. It's designed to detect linear associations.
2. **Strength:**
   - The absolute value of Pearson's R indicates the strength of the relationship. A value closer to 1 or -1 signifies a strong relationship, while a value closer to 0 indicates a weak or no linear relationship.
3. **Direction:**
   - The sign of Pearson's R indicates the direction of the relationship:
     - A positive value (+1) means a positive linear relationship (as one variable increases, the other tends to increase).
     - A negative value (-1) means a negative linear relationship (as one variable increases, the other tends to decrease).

Key Characteristics:

- **Range:**
    - Pearson's R ranges from -1 to +1, inclusive.
- **Interpretation:**
    - +1: Perfect positive linear correlation.
    - -1: Perfect negative linear correlation.
    - 0: No linear correlation.
- **Assumptions:**
    - Both variables should be continuous.
    - The relationship between the variables should be linear.
    - The variables should be approximately normally distributed.
    - There should be no significant outliers.

When to Use It:

- Pearson's R is appropriate when you want to measure the linear association between two quantitative variables.
- It's commonly used in various fields, including psychology, sociology, economics, and biology.

Important Considerations:

- **Correlation vs. Causation:**
    - It's crucial to remember that correlation does not imply causation. A strong Pearson's R value only indicates a linear association, not that one variable causes the other.
- **Non-linear Relationships:**
    - Pearson's R is not suitable for detecting non-linear relationships. If the relationship between the variables is curved, Pearson's R may be close to 0, even if there's a strong association.
- **Outliers:**
    - Outliers can significantly affect Pearson's R. It's important to examine scatter plots and identify potential outliers.

In summary, Pearson's R is a valuable tool for quantifying linear relationships between variables, but it's essential to use it appropriately and interpret the results cautiously.

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>

In machine learning, "**scaling**" refers to the process of transforming numerical features to a similar scale. This is a crucial preprocessing step that can significantly impact the performance of many algorithms.

Essentially, **scaling** adjusts the range of independent variables (features) so that they all contribute equally to the analysis.

Why is Scaling Performed?

- **Algorithm Sensitivity:**
  - Many machine learning algorithms, especially those based on distance calculations (like k-nearest neighbors) or gradient descent (like neural networks), are sensitive to the scale of the input features.
  - Without scaling, features with larger ranges can dominate the calculations, leading to suboptimal model performance.
- **Faster Convergence:**
  - Gradient descent-based algorithms converge much faster when features are on a similar scale. This can significantly reduce training time.
- **Improved Accuracy:**
  - Scaling can improve the accuracy of models by ensuring that all features contribute equally to the learning process.
- **Preventing Domination:**
  - It prevents features with higher magnitudes from dominating the other features.

Difference between Normalized Scaling and Standardized Scaling:-

Feature	**Normalized Scaling (Min-Max Scaling)**	**Standardized Scaling (Z-score Normalization)**
Range	0 to 1	-∞ to ∞
Mean	Preserved	0
Standard Deviation	Not preserved	1
Outlier Sensitivity	Sensitive	Less sensitive
Use Cases	When data range is important (e.g., image pixels)	When data distribution is skewed or outliers are present

In summary:

- Normalization scales data to a specific range, while standardization scales data to have a mean of 0 and a standard deviation of 1.
- The choice between the two depends on the specific characteristics of your data and the requirements of your machine learning algorithm.
- It is very important to remember that when you scale training data, that you must use the same scaler to scale any test or production data.

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) encounters infinite values, due to a specific condition within the data i.e. **perfect multicollinearity.**

Why Infinite VIF Occurs:

1. **Perfect Correlation ($R^2 = 1$):**
   - An infinite VIF arises when the predictor variable in question can be perfectly predicted from the other predictor variables.
   - In the VIF formula, if $R^2 = 1$, then $(1 - R^2) = 0$, and $1 / 0$ is undefined (or infinite).
2. **Dummy Variable Trap (Most Common Cause):**
   - The most frequent cause of infinite VIF is the "dummy variable trap." This occurs when you include all categories of a categorical variable as dummy variables in your regression without dropping one.
   - For example, if you have a categorical variable "color" with categories "red," "green," and "blue," and you create dummy variables for all three, they will be perfectly multicollinear. If "red" = 0 and "green" = 0, then "blue" must be 1.
   - When this happens, one of the dummy variables can be perfectly predicted from the others, leading to an infinite VIF.
3. **Exact Linear Combinations:**
   - Infinite VIF can also occur when one predictor variable is an exact linear combination of other predictor variables.
   - For example, if you have variables "x1," "x2," and "x3," and "x3" is always equal to "2 * x1 + 3 * x2," then "x3" will have an infinite VIF.

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**, or quantile-quantile plot, is a graphical tool used to assess whether a dataset follows a particular theoretical distribution, such as a normal distribution. In the context of linear regression, it's primarily used to check the normality assumption of the residuals.

Use and Importance in Linear Regression:

1. **Checking Normality of Residuals:**
   - One of the key assumptions of linear regression is that the residuals (the differences between the observed and predicted values) are normally distributed.
   - A Q-Q plot of the residuals helps visually assess whether this assumption is met.
   - If the residuals are normally distributed, the points on the Q-Q plot will closely follow a straight diagonal line.
   - Deviations from the line indicate departures from normality.

2. **Identifying Departures from Normality:**
   - Q-Q plots can help identify specific types of departures from normality:
   - Skewness: A curved pattern in the Q-Q plot indicates skewness in the residuals.
   - Heavy Tails: Points at the ends of the Q-Q plot that deviate from the line indicate heavy tails (more extreme values than expected).

- **Light Tails:** Points at the ends of the Q-Q plot those are closer to the center than the line indicates light tails (fewer extreme values than expected).

3. **Assessing Model Validity:**
   - Violations of the normality assumption can affect the validity of statistical inferences made from the linear regression model.
   - If the residuals are not normally distributed, the p-values and confidence intervals associated with the coefficients may be unreliable.

4. **Guiding Model Refinement:**
   - If the Q-Q plot reveals significant departures from normality, it may be necessary to transform the dependent variable or include additional predictor variables in the model.

How to Interpret a Q-Q Plot:

1. **Straight Line:**
   - If the points closely follow a straight diagonal line, the residuals are approximately normally distributed.
2. **Deviations from the Line:**
   - Curved patterns or points that deviate significantly from the line indicate departures from normality.