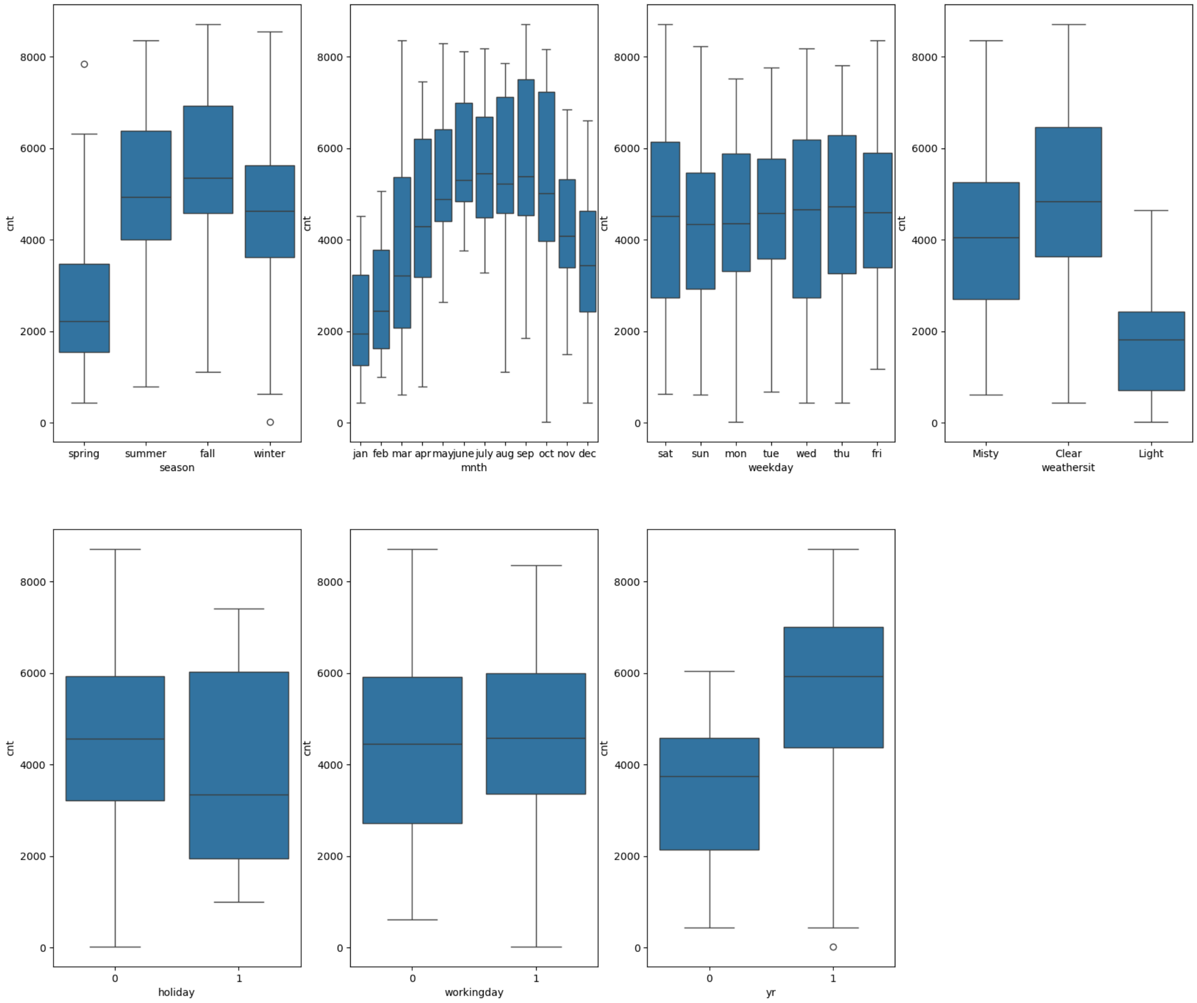
Assignment-based Subjective Questions

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)

Below is the visualization of categorical variables with target variables.



Season: Most people rent bikes in the fall season because the weather is mild, making it comfortable to ride.
Year: More people rented bikes in 2019 than in 2018. This is because a new company's growth usually increases in the second year.
Holiday: Bike rentals are higher on holidays as people spend time with family and friends.
Weekday: On the third day of the week, bike rentals are higher since it's midweek, and many people commute to work.
Working Day: More bikes are rented on non-working days compared to working days.
Weather: Bike rentals are higher when the weather is clear or partly cloudy.
Month: Most bike rentals happen in August and September compared to other months.

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)

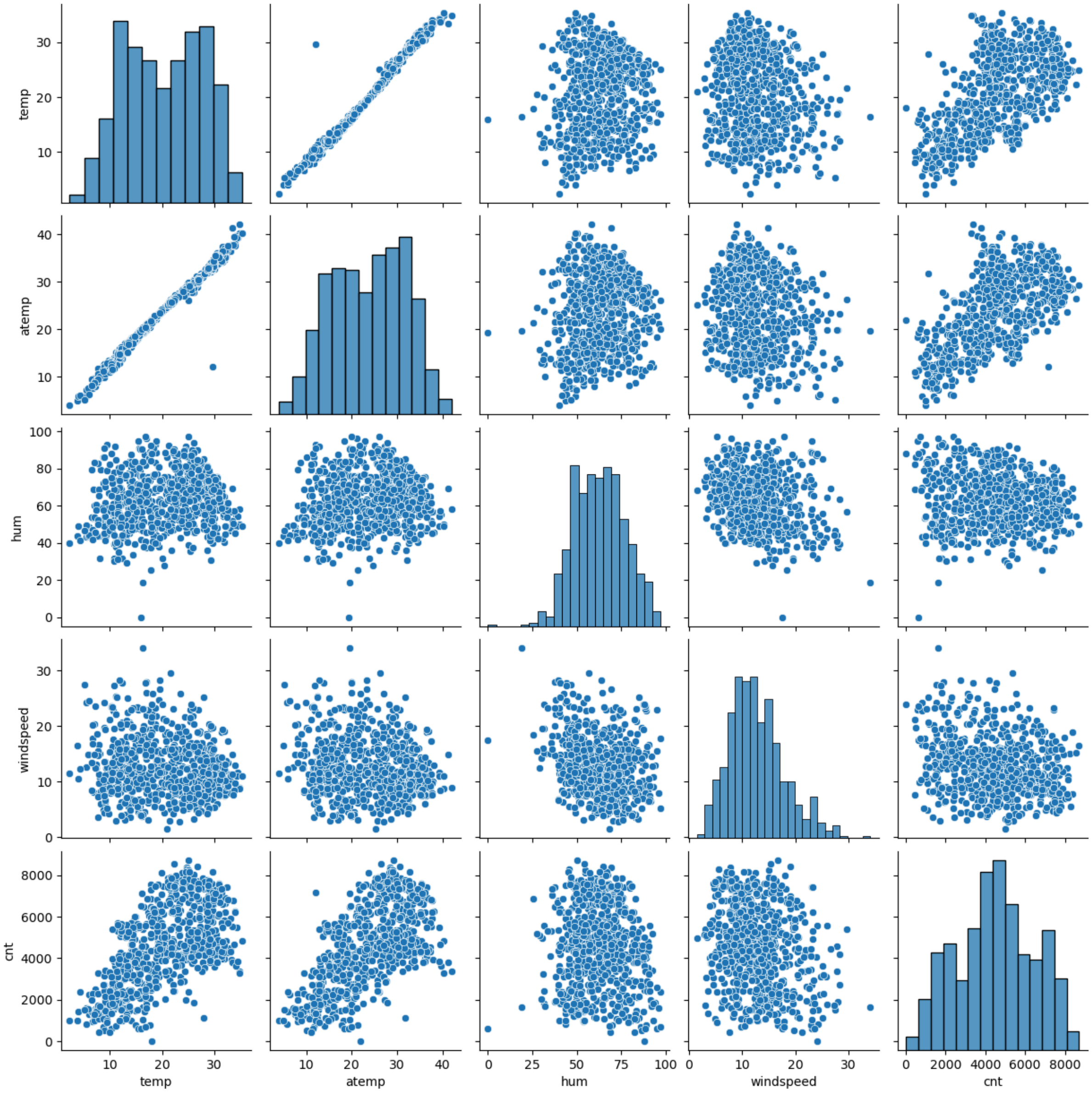
drop_first is the parameter used to create a dummy variable. When we create a dummy variable it generates N number of columns depending on categorical data. If we do drop_first=True it drops the first column which is easily determined using another dummy variable. It reduces multicollinearity and it won’t create unnecessary weight while building a 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)

From the pair plot of numerical variables, we observed that 'temp' and 'atemp' have the strongest correlation with the target variable 'cnt'. Both variables show a similar pattern and have a linear relationship with '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)
Linearity – Check scatter plot of predicted vs. actual values.
Independence – Use the Durbin-Watson test for autocorrelation.
Homoscedasticity – Plot residuals vs. predicted values.
Normality – Check residual distribution using a histogram or Q-Q plot.
No Multicollinearity – Use VIF; values should be **< 5**.

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)

1. Year
2. Holiday
3. Temp

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 fundamental statistical and machine learning algorithm used to establish a relationship between a dependent variable **Y** and one or more independent variables (**X**). It follows the mathematical equation:

$$y = c + mx1 + m2x2 + \ldots + mxxn + e$$

where **c** is the intercept, **m1..mn** are the coefficients representing the impact of each independent variable, and **e** is the error term. The objective of Linear Regression is to find the best-fit line that minimizes the difference between the actual and predicted values, achieved using the **stats** and **sklearn** library, which minimizes the **Mean Squared Error (MSE)**.

Assumptions of Linear Regression:

1. **Linearity** – The relationship between independent and dependent variables should be linear.
2. **Independence** – Observations should be independent of each other.
3. **Homoscedasticity** – The variance of residuals should remain constant across predictions.
4. **Normality of Residuals** – The residual errors should be normally distributed.
5. **No Multicollinearity** – Independent variables should not be highly correlated.

Linear Regression is widely used in fields like finance, economics, and research for making predictions and trend analysis.

<Your answer for Question 6 goes here>

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>

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 is also called Pearson correlation coefficient and it measures the strength and direction of the linear relationship between two continuous variables. Its value ranges from **-1 to 1**:

- **1:** Perfect positive linear relationship.
- **-1:** Perfect negative linear relationship.
- **0:** No linear relationship.

<Your answer for Question 8 goes here>

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 adjusts the range of data to make it consistent, which is important for algorithms sensitive to data magnitude.

Why Scaling is Performed

- It improves model performance.
- Ensures features are on a similar scale.
- Helps when features have different units.

Normalized Scaling vs. Standardized Scaling:

- **Normalized Scaling (Min-Max):**
  Scales data to a specific range (e.g., 0 to 1). Useful for algorithms like neural networks.
- **Standardized Scaling (Z-score):**

Scales data to have a mean of 0 and a standard deviation of 1. Useful when data follows a normal distribution.

Difference: Normalization adjusts data to a range, while standardization centers it around 0 with a standard deviation of 1.

<Your answer for Question 9 goes here>

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)
If R^2 is equal to 1 (perfect correlation), the denominator becomes zero, causing VIF to become infinite.
<Your answer for Question 10 goes here>

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 plot is used to check if residuals follow a normal distribution, which is important for valid regression results.
<Your answer for Question 11 goes here>
