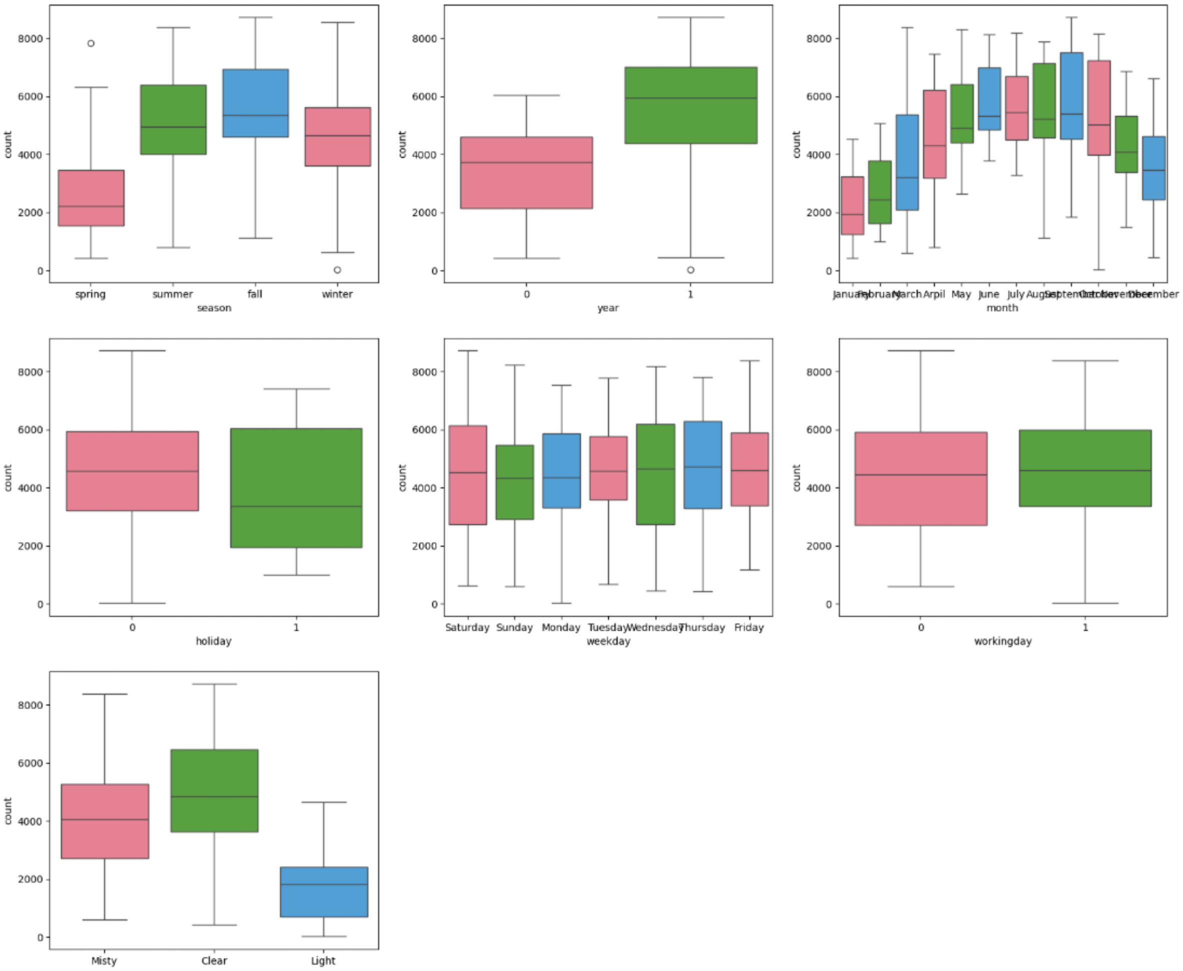
Assignment-based Subjective Questions

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)



- **Season:** Spring has the lowest median count, while Summer and Fall have the highest. Winter is in between. This suggests a strong seasonal influence on the count.
- **Year:** The '1' year (likely the later year) shows a higher median count than the '0' year, indicating an overall increase in count over time.
- **Month:** Counts generally rise from January to June, peak in the summer months (June-August), and then decline towards December. This reinforces the seasonal trend.
- **Holiday:** Holidays (1) have higher median counts than non-holidays (0), suggesting that holidays are associated with increased counts.
- **Weekday:** Weekends (Saturday and Sunday) have noticeably higher median counts than weekdays, indicating a weekly cyclical pattern.
- **Working day:** Non-working days (0) have higher median counts than working days (1), which aligns with the weekday analysis (weekends being non-working days).
- **Weather:** Clear weather has the highest median count, followed by Light (likely light rain/snow), and then Misty, which has the lowest median count. This shows a strong influence of weather on the count.

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)

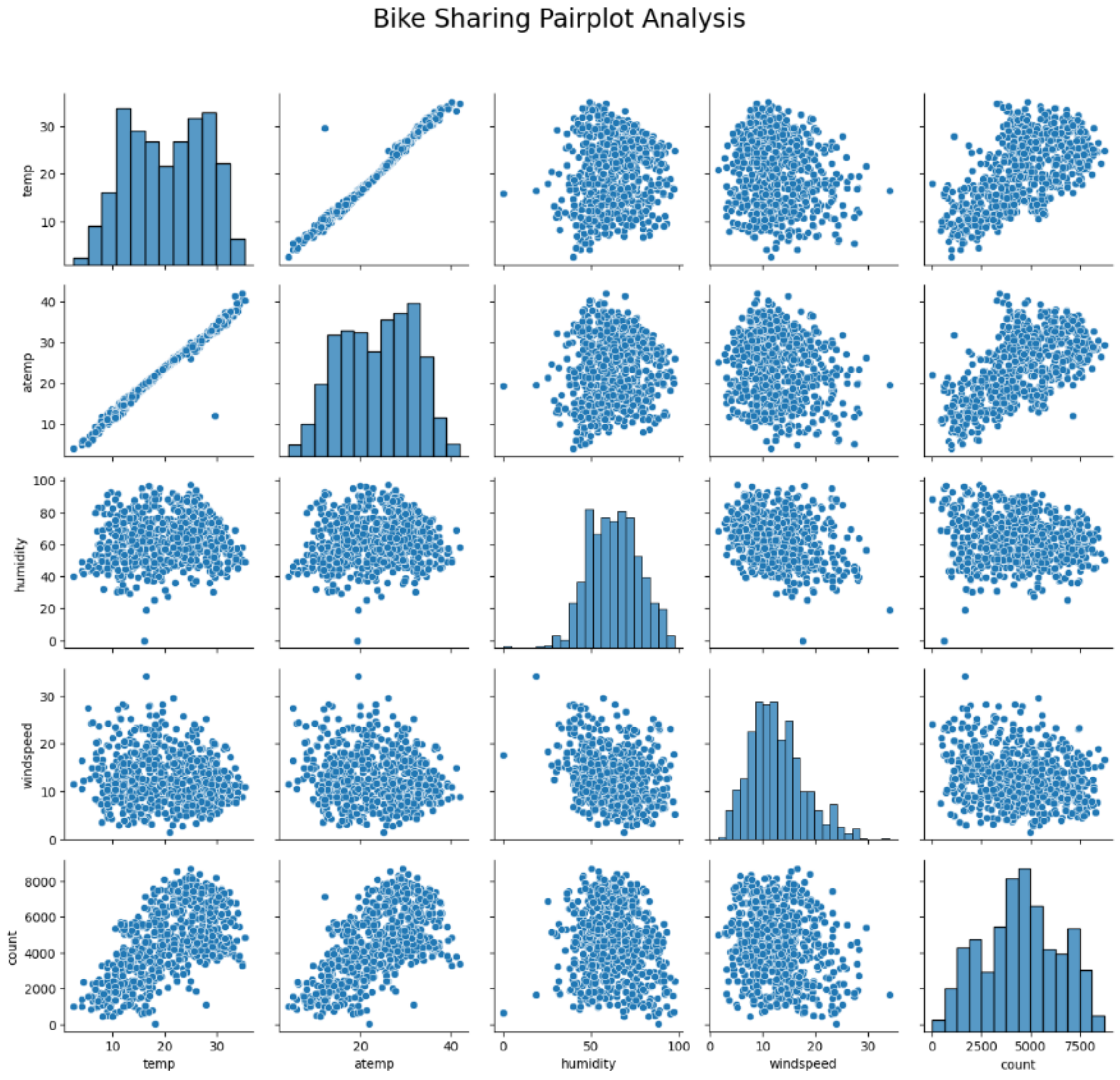
Using "***drop_first=True***" in dummy variable creation helps to avoid multicollinearity in regression models.

When creating dummy it helps to avoid extra, unnecessary information that can confuse the model. It removes one category from the dummy variables because the model can figure it out from the others. For example, if we have a "Color" column with Red, Blue, and Green, we only need two dummy columns (like "Blue" and "Green"), because if both are 0, we know it's Red. This makes the model simpler and prevents problems with too much overlapping information.

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, we can see that atemp has the highest correlation with the target variable count. Temp being the close second.

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 : Checking scatter plots of predictions vs. actual values.
No Multicollinearity : Using VIF (Variance Inflation Factor) to ensure independent variables aren't too correlated.
Homoscedasticity : Plotting residuals to check for constant variance.
Normality of Residuals : Using a histogram or Q-Q plot to see if errors follow a normal distribution.
Independence of Errors : Checking the Durbin-Watson statistic to detect autocorrelation.

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 features contributing significantly towards demand of shared bikes

- Year
- holiday
- temperature

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 statistical method used to predict a continuous outcome (Y) based on one or more independent variables (X). It finds the best-fitting straight line that minimizes errors between actual and predicted values.
Key Components:
Equation: Y=β0+β1X1+...+βnXn+ϵ
Learning Process:
Uses Ordinary Least Squares (OLS) to minimize the sum of squared errors.
Assumptions:

Linearity, no multicollinearity, homoscedasticity, normality of residuals, and independence of errors.

Evaluation Metrics:

R^2, Adjusted R^2, Mean Squared Error (MSE), and significance of 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)

<Your answer for Question 7 goes here>

Anscombe's Quartet consists of four datasets with nearly identical statistics (mean, variance, correlation) but look very different when graphed. It proves that relying only on numbers can be misleading, and visualizing data is crucial.

- One looks linear
- One is curved
- One has an outlier
- One is driven by a single point

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 (Correlation Coefficient) measures the strength and direction of the relationship between two variables. It ranges from -1 to 1:

- +1 → Perfect positive correlation (both increase together)
- 0 → No correlation (no relationship)
- -1 → Perfect negative correlation (one increases, the other decreases)

It helps understand how two variables are related but doesn't imply causation!

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>

What is Scaling?

Scaling resizes numerical data to a similar range, improving model performance.

Why is Scaling Performed?
To prevent large values from dominating and improve model accuracy.
Normalization vs. Standardization

- Normalization: Rescales data between 0 and 1.
- Standardization: Centers data with mean = 0 and std = 1.

Key Difference: Normalization is for fixed ranges, standardization for normal distribution.

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>

A VIF value becomes infinite when there is perfect multicollinearity between two or more independent variables. This means that one variable is an exact linear function of another, making it impossible to estimate the variance of the coefficients properly. In such cases, the VIF calculation can't be performed accurately, leading to an infinite value.
To resolve this issue, you should remove or combine highly correlated features in your dataset.

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 (Quantile-Quantile plot) is a graphical tool used to compare the distribution of your data to a theoretical distribution, typically the normal distribution.
Use and Importance in Linear Regression:
Check Normality of Residuals: One key assumption of linear regression is that the residuals (errors) are normally distributed. A Q-Q plot helps check this by comparing the residuals to a normal distribution.
Interpreting the Plot: If the points on the Q-Q plot lie along a straight line, the residuals are normally distributed. Deviations from this line indicate departures from normality, such as skewness or kurtosis.
Why It Matters:
Ensuring that residuals are normally distributed helps validate the linear regression model, making predictions more reliable.
It helps identify potential issues with the model, such as non-linearity, heteroscedasticity, or outliers.
By using a Q-Q plot, you can diagnose and address these issues, ensuring a more robust and accurate linear regression model.