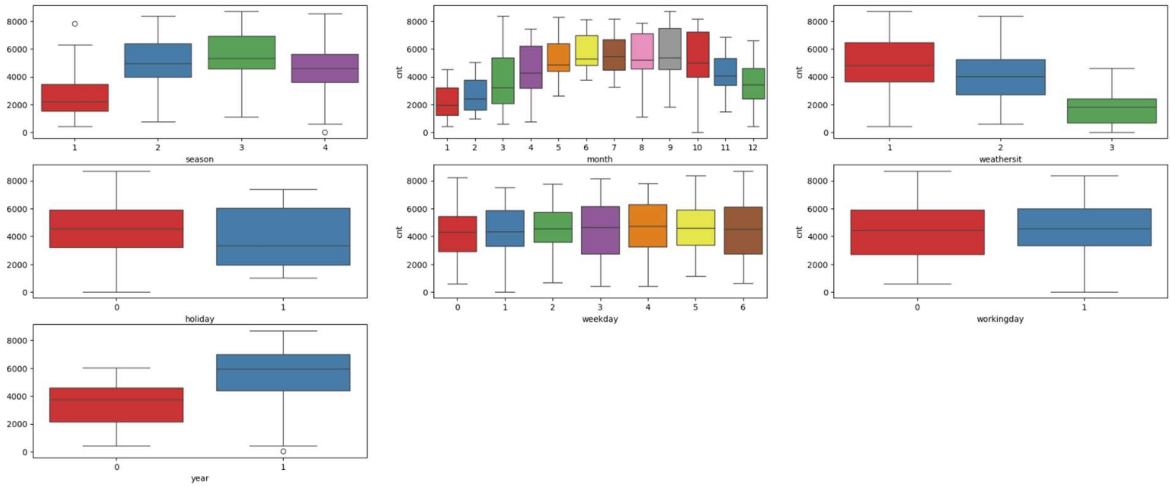
Assignment-based Subjective Questions

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)



Analysis of categorical variables (season, month, weathersit, holiday, weekday, workingday, year) using box plots revealed the following: (Fig. attached).

- **Season:** Fall exhibited the highest rental counts, followed by Summer and Winter, with Spring having the lowest.
- **Month:** September saw the highest rentals, while January had the lowest, potentially influenced by adverse winter weather conditions.
- **Weathersit:** Clear or partly cloudy conditions resulted in the highest rental counts, while heavy rain/snow significantly reduced usage.
- **Holiday:** Rentals typically decreased during holidays.
- **Weekday:** Rental counts were relatively consistent across weekdays.
- **Workingday:** Working days generally saw higher median rental counts.
- **Year:** 2019 experienced higher rental counts compared to 2018.

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)

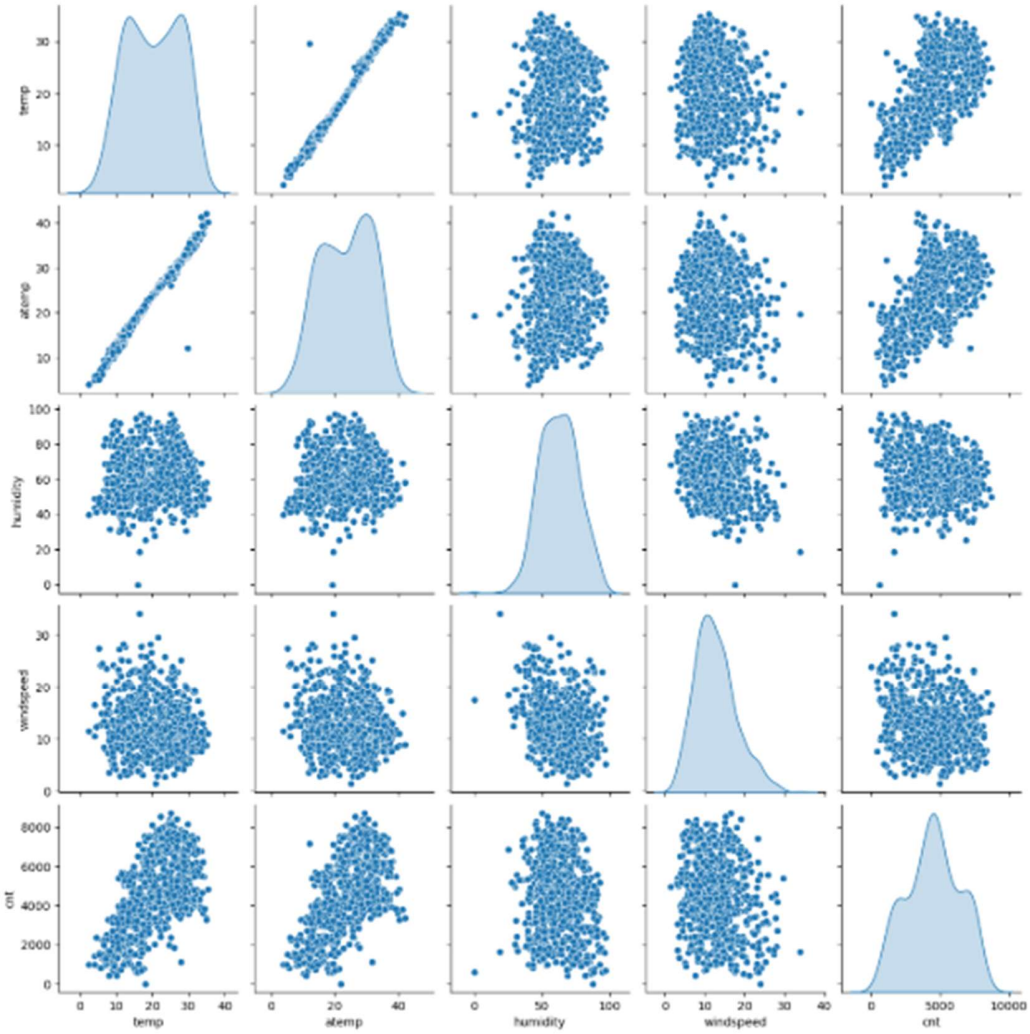
Answer: <Your answer for Question 2 goes below this line> (Do not edit)

When representing categorical variables in a model, it's crucial to avoid introducing redundant information that can lead to unstable and unreliable results. One-hot encoding, a common technique to convert categorical variables into numerical format, can create this redundancy if not handled carefully.

The **drop_first=True** parameter in one-hot encoding addresses this issue by removing one of the dummy variables created for each category. This effectively prevents multicollinearity, a situation where one variable can be perfectly predicted from others, making it difficult for the model to accurately estimate the impact of each category.

By dropping the first category, we simplify the model, improve its interpretability, and ensure that the model can accurately assess the relationship between each category and the outcome variable.

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)
Using the below pairplot it can be seen that , "temp" and "atemp" are the two numerical variables which are highly correlated with the 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)

- **Linearity**: We visually checked for linear relationships between the predictors and the outcome.

- **Normality of Errors**: We ensured the model's errors (the differences between actual and predicted values) were approximately normally distributed.
- **Multicollinearity**: We assessed for potential issues with predictor variables being highly correlated with each other by calculating the Variance Inflation Factor (VIF).

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)

The top 3 features are:

- Year: 0.2299
- Temp: 0.4680
- Weathersit_light_snow: -0.3396

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)

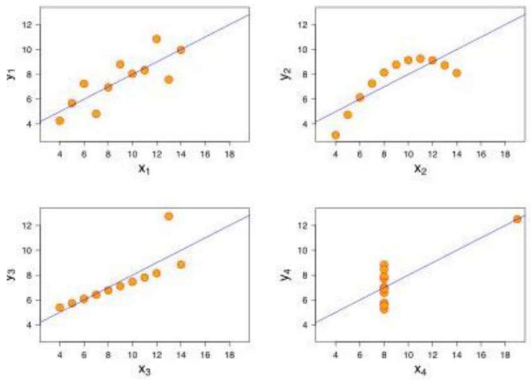
<Your answer for Question 6 goes here>

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables.

- Key Steps:
  - **Data Preparation**: Collect, clean, and prepare data, including handling missing values, scaling variables, and creating dummy variables for categorical features.
  - **Model Training**: Estimate the coefficients of the linear equation using the Ordinary Least Squares (OLS) method, which aims to minimize the difference between actual and predicted values.
  - **Prediction**: Use the fitted model to make predictions on new data.
  - **Evaluation**: Assess model performance using metrics like R-squared, Adjusted R-squared, Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE).
- Assumptions:
  - Linear regression relies on several key assumptions, including:
    - Linearity between predictors and the response.
    - Homoscedasticity (constant variance of errors).
    - Normally distributed errors.
    - Independence of errors.
    - No multicollinearity among predictors.
- Visualization:
  - Visual techniques, such as scatter plots, residual plots, and Q-Q plots, are crucial for understanding the model and validating its assumptions.
  - This summary provides a concise overview of the linear regression algorithm, its key steps, evaluation metrics, assumptions, and the importance of visualization.

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 was developed by statistician Francis Anscombe. It comprises four datasets with nearly identical summary statistics (mean, standard deviation, correlation, regression line). However, when visualized, they reveal vastly different underlying patterns. This highlights the crucial role of data visualization in understanding data beyond relying solely on numerical summaries.



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 correlation coefficient (often denoted by 'r') measures the strength and direction of the linear relationship between two variables.

- **Strength**:
  - r close to 1: Strong linear relationship
  - r close to 0: Weak or no linear relationship
- **Direction**:
  - r > 0: Positive correlation (as one variable increases, the other tends to increase)
  - r < 0: Negative correlation (as one variable increases, the other tends to decrease)
- **Key Points**:
  - **Range**: Pearson's r values range from -1 to 1.
  - **Linearity**: It specifically measures the strength of the linear relationship. Non-linear relationships may not be accurately reflected by Pearson's r.
- Assumptions:
  - The variables are continuous.
  - The relationship between the variables is linear.
  - The data is normally distributed.

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>

- Feature scaling is a crucial step in data preprocessing that aims to bring the features of a dataset to a common scale. This is essential because machine learning algorithms can be significantly influenced by the magnitude of feature values. Without scaling, algorithms may disproportionately weigh features with larger values, potentially leading to biased or inaccurate results.

- Normalization is a scaling technique that transforms features to a specific range, typically between 0 and 1. It is particularly useful when dealing with data that does not follow a Gaussian (normal) distribution, as it can improve the performance of algorithms that do not assume any specific data distribution, such as K-Nearest Neighbors and Neural Networks.

- Standardization is another scaling method that transforms features to have zero mean and unit variance. While often used with Gaussian distributed data, it can be applied to any dataset. Unlike normalization, standardization does not have a fixed upper or lower bound, making it less sensitive to outliers in the 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>
A Variance Inflation Factor (VIF) can become infinite when there is perfect multicollinearity among the predictor variables in a regression model. This occurs when one predictor variable can be perfectly explained by a linear combination of other predictors. Here's why this happens:

- **Reason for Infinite VIF**
  - When calculating VIF for a predictor variable X_i, the R^2 value represents how well X_i can be explained by other predictors. The formula for VIF is: VIF (X_i)=1/1−R_i^2
- **Perfect Multicollinearity**
  - Perfect Correlation: If X_i is perfectly correlated with other predictors, R_i^2 will be 1.
  - Division by Zero: Substituting R_i^2 = 1 into the formula results in division by zero: VIF (X_i) = 1 / 1−1 = 1 / 0 = ∞
  - This indicates an infinite VIF, meaning that the predictor variable X_i adds no unique information to the model because it is entirely predicted by other 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)

<Your answer for Question 11 goes here>

A Q-Q plot compares the quantiles of two data sets against each other, helping to compare their distribution shapes. It's a scatterplot where each point represents a pair of quantiles from the two data sets. If both sets are from the same distribution, the points will form a roughly straight line. The Q-Q plot helps answer:

- Do the two data sets originate from populations with a common distribution?
- Do the two data sets have similar location and scale?
- Do the two data sets share similar distributional shapes?
- Do the two data sets exhibit similar tail behavior?