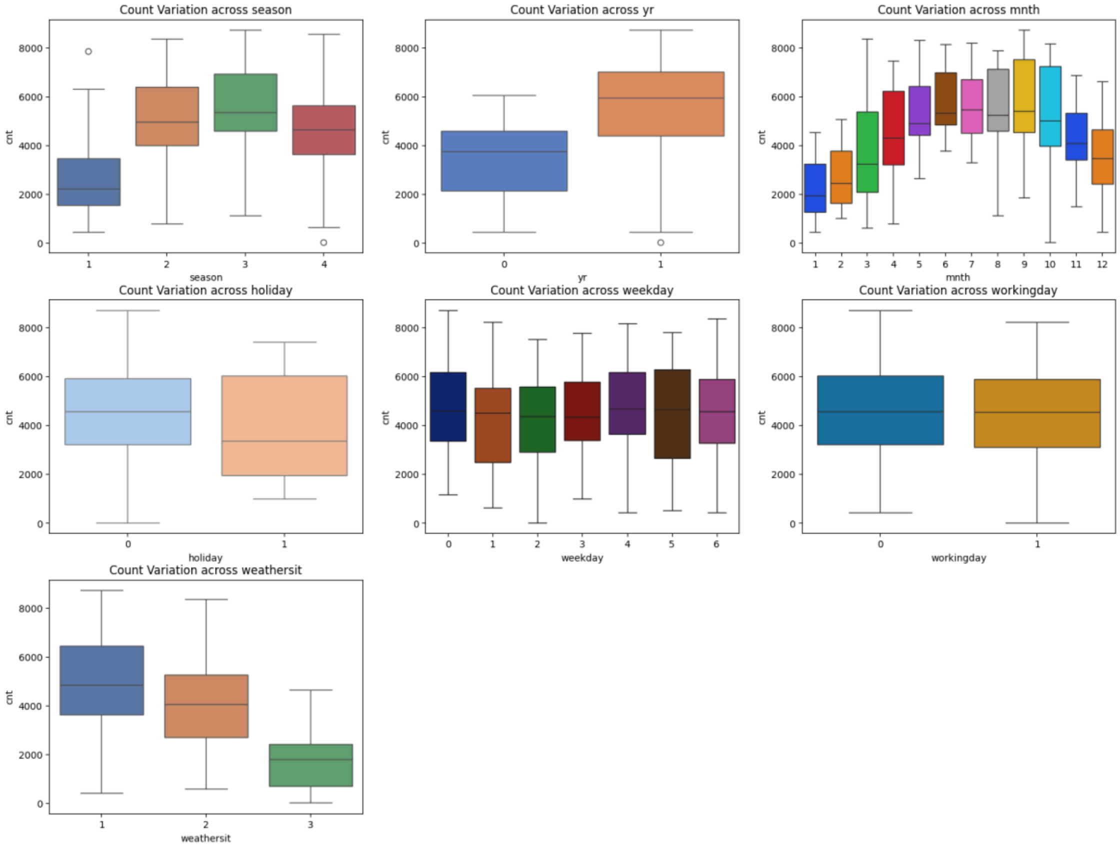
Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

   Ans:



- **Season:** Bike rentals peak in the fall and are lowest in spring.
- **Month:** Most rentals happen from May to October, with a rise in demand at the start of the year that declines toward the end.
- **Weekday:** Sunday sees the fewest rentals, while demand grows from Monday to Friday.
- **Weather:** Rentals are higher when the weather is clear.
- **Working Day:** Rentals are nearly the same whether it's a working day or not.
- **Holiday:** There are fewer rentals on holidays.
- **Year:** More rentals occurred in 2019 compared to 2018.

2. Why is it important to use **drop_first=True** during dummy variable creation?

   Ans:

   `drop_first=True` removes an extra column when creating dummy variables. This avoids a situation called the "dummy variable trap," which can cause multicollinearity, or perfect correlation between variables, making it difficult for models to give accurate predictions.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

   Ans:

   The variable **temperature** (temp) has the strongest correlation with the target variable, meaning it significantly affects bike rental demand.

4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans:

- **Linearity:** Ensure that the relationship between the dependent and independent variables is linear.
- **Normality of Errors:** Residuals (errors) should follow a normal distribution.
- **Multicollinearity:** Check that the independent variables are not too highly correlated using the Variance Inflation Factor (VIF).
- **Homoscedasticity:** Residuals should have consistent variance across all levels of the independent variables.
- **Independence of Residuals:** Residuals should be independent of each other (no autocorrelation).

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Ans:

- Temperature (temp) : 0. 403067
- Year (yr): 0. 232158
- Severe (weathersit): 0.266589

General Subjective Questions

1. Explain the linear regression algorithm in detail.

   Ans:

   - Linear regression is a predictive model that analyzes the relationship between a dependent variable and one or more independent variables.
   - It attempts to find a straight line that best fits the data points, minimizing the distance between the actual and predicted values.
   - The equation for simple linear regression is Y=Bo + B1X, where B1 is the slope and Bo is the intercept.
   - In multiple linear regression, this extends to include more independent variables. The model works best when the assumptions of linearity, normality of errors, and no multicollinearity are met.

2. Explain the Anscombe’s quartet in detail.

   Ans:

   - **Anscombe's Quartet** is a set of four datasets that were specifically designed by the statistician Francis Anscombe in 1973 to show the importance of data visualization in statistical analysis.
   - Each of the four datasets has nearly identical basic statistical properties (such as mean, variance, and correlation), yet they display drastically different patterns when graphed.
   - The quartet was created to demonstrate that relying only on summary statistics without visually inspecting the data can lead to misleading interpretations.

3. What is Pearson’s R?

Ans:

- Pearson’s R is a measure of the linear relationship between two continuous variables. It ranges from -1 (perfect negative correlation) to 1 (perfect positive correlation), with 0 meaning no linear correlation.
- The coefficient is calculated by dividing the covariance of the two variables by the product of their standard deviations.
- Pearson’s R is widely used in statistical analysis to understand how closely two variables are related.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans:

- Scaling is used in data preprocessing to standardize the range of independent variables, ensuring that all features contribute equally in the model.
- Without scaling, features with larger ranges can dominate the learning process, leading to biased or incorrect predictions. For instance, variables with larger values might overshadow those with smaller values, regardless of their importance. By scaling, we ensure that the model treats all features equally, leading to better model performance and faster convergence during training.
- There are two types of scaling technique:

  i **Normalization:** Scales values between 0 and 1, often used when the data doesn't follow a normal distribution.

  ii **Standardization:** Scales values to have a mean of 0 and a standard deviation of 1, generally used for data that follows a normal distribution.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans:

- The Variance Inflation Factor (VIF) measures how much the variance of a regression coefficient is inflated due to multicollinearity between independent variables. When independent variables are highly correlated, VIF values increase, indicating potential issues in the model.
- VIF becomes infinite when there is perfect multicollinearity, meaning one independent variable can be exactly predicted by a linear combination of other variables.
- An infinite VIF signals that the regression model has redundant information, which causes instability in the coefficient estimates.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Ans:

- A Q-Q (Quantile-Quantile) plot is a graphical tool used to assess whether the distribution of a dataset matches a theoretical distribution, typically a normal distribution. In linear regression, it helps to visually check if the residuals (the differences between predicted and actual values) are normally distributed.
- The Q-Q plot compares the quantiles of the residuals with the quantiles of a normal distribution. If the points in the plot lie approximately along a 45-degree reference

line, it suggests that the residuals follow a normal distribution. Deviations from this line indicate potential non-normality, which may affect the accuracy of the model's predictions and statistical tests.

- The Q-Q plot is crucial for diagnosing issues with the model's assumptions, such as non-normal residuals, which can impact hypothesis testing, confidence intervals, and predictions. It also helps detect outliers and other anomalies in the residuals.