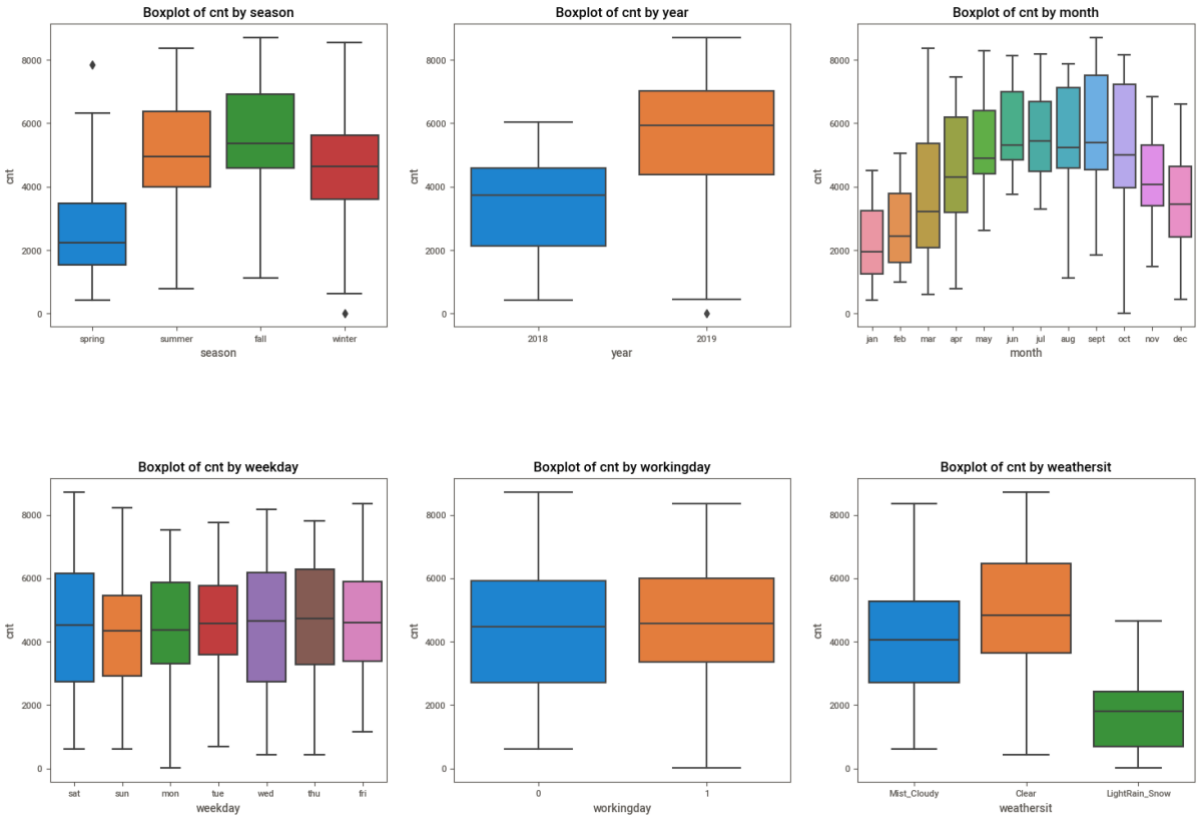
1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**



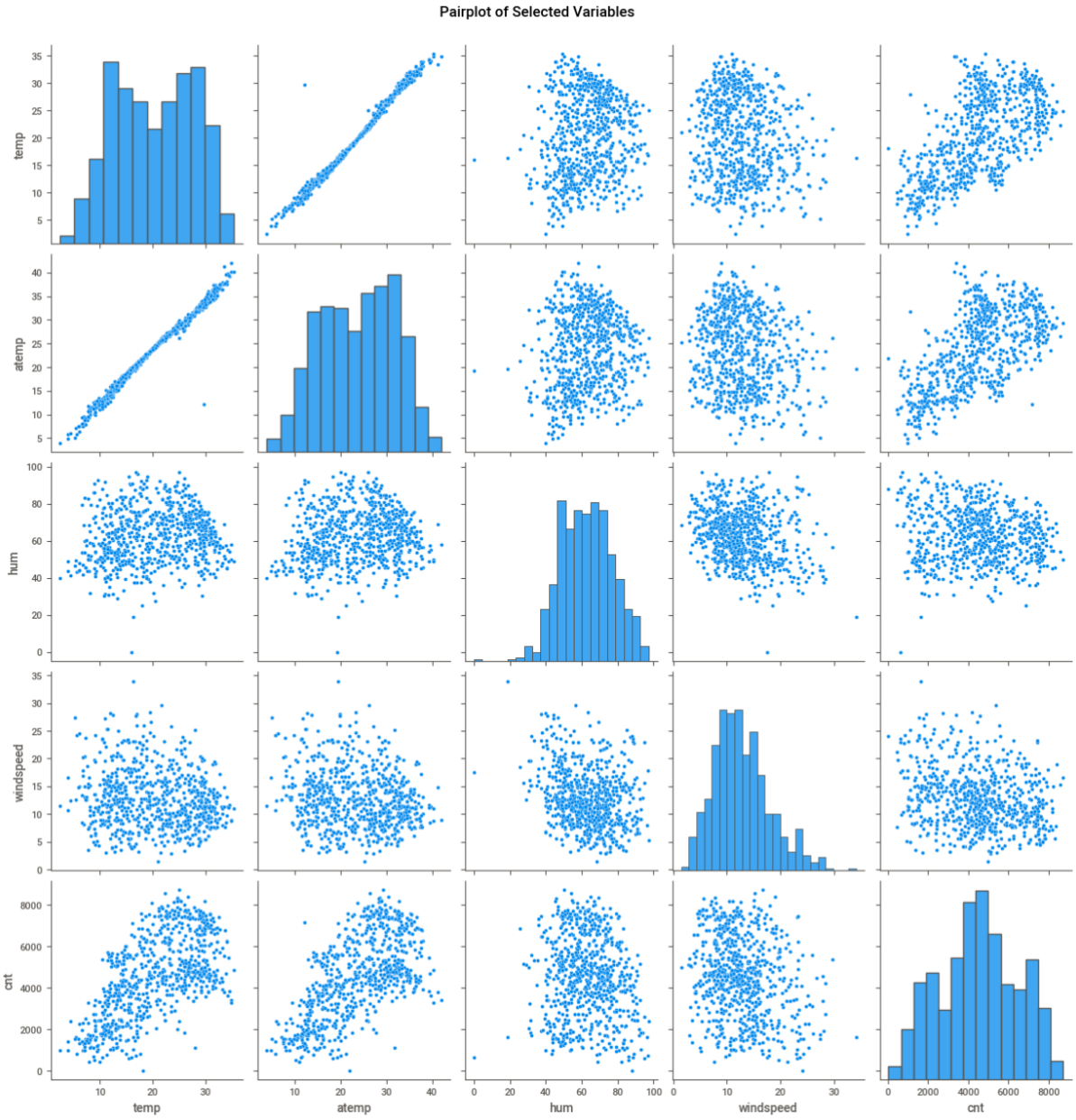
From the above boxplot we can observe following insights:

1. Summer and Fall are the most amount of counts that we can observe while winter and spring are relatively less.
2. Popularity has increased overtime compared to 2018 in 2019.
3. May, June and July are the most important months for the business.
4. Weather has huge influence to the business, during clear weather we could observe better count.

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

When creating dummy variables for a categorical variable with 'n' levels, using drop=True means that one of the levels will be omitted, resulting in 'n-1' dummy variables. By dropping one of the levels, we can prevent perfect multicollinearity because the information about that level is already captured by the remaining dummy variables. If all 'n' dummy variables were included, it could lead to a situation known as the "dummy variable trap," where the dummy variables are perfectly correlated, making it impossible for the model to distinguish the effect of each individual variable. In summary, using drop=True when creating dummy variables helps avoid multicollinearity and ensures that the model can effectively capture the information from the categorical variable without introducing redundancy or instability.

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



Among the numerical variables, we can clearly observe that atemp and temp are mostly correlated with the target variable. Also, we can observe linear relationship between each other.

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

To validate the assumptions of Linear Regression model, we check the correlation between the X variable which is also called multicollinearity using VIF<5 and P-value<0.05. We check for Homoscedasticity too.

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

The three most influential factors in elucidating the demand for shared bikes are the winter season (**season_winter**), temperature (**temp**), and the month of September (**month_sept**)

Subjective Questions:

1. Explain the linear regression algorithm in detail ?

Linear regression is a statistical method used for modeling the relationship between a dependent variable and one or more independent variables. The goal is to find a linear equation that best represents the data. In simpler terms, linear regression seeks to establish a straight line that predicts the values of the dependent variable based on the values of the independent variable(s). The process involves determining the slope and intercept of the line that minimizes the difference between the predicted and actual values. This is done through a technique called least squares, which aims to minimize the sum of squared differences between the observed and predicted values.

Linear regression assumes a linear relationship between the variables, meaning that a change in one variable is associated with a constant change in the other. It is widely used in various fields, including economics, finance, and science, to analyze and predict relationships between variables.

2. Explain the Anscombe's quartet in detail?

Anscombe's quartet consists of four datasets that have nearly identical statistical properties but differ significantly when graphically plotted. Created by statistician Francis Anscombe, the quartet highlights the importance of visualizing data. Each dataset includes 11 data points with two variables (X and Y). Despite having the same mean, variance, correlation, and regression line, the datasets reveal the limitations of relying solely on summary statistics. This underscores the need for graphical exploration and visualization to understand the underlying patterns and relationships within data.

3. What is Pearson's R?

r is calculated as the covariance of the variables divided by the product of their standard deviations. It is widely used to assess the strength and direction of the linear association between variables.

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

Scaling is the process of transforming variables to a standard range to ensure they contribute equally to a model. It is performed to avoid issues related to the different scales of variables in a dataset. Scaling helps models that rely on distance metrics or gradient-based optimization algorithms converge more quickly.

• Normalized Scaling: It scales variables to a range of 0 to 1.

• Standardized Scaling: It transforms variables to have a mean of 0 and a standard deviation of 1.

•The key difference is in the scale of the result.

Normalized scaling maintains the original distribution within the 0-1 range, while standardized scaling centers the distribution around 0 with a standard deviation of 1.

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

The Variance Inflation Factor (VIF) becomes infinite when perfect multicollinearity exists among predictor variables. Perfect multicollinearity occurs when one variable in a regression model can be exactly predicted by a linear combination of other variables. In such cases, the correlation matrix becomes singular, leading to an infinite VIF value for the affected variable. This hinders the stability of coefficient estimates and inflates their standard errors.

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

A Q-Q (Quantile-Quantile) plot is a graphical tool used to assess whether a dataset follows a theoretical distribution, typically the normal distribution. In a Q-Q plot, the quantiles of the observed data are compared to the quantiles of the expected theoretical distribution.

Use and Importance in Linear Regression:

- **Normality Check:** Q-Q plots are crucial for checking the normality assumption of residuals in linear regression. If residuals are normally distributed, the points in the Q-Q plot will approximately fall along a straight line.
- **Identifying Outliers:** Deviations from the straight line in a Q-Q plot can indicate outliers or non-normality in the data, which can impact the validity of linear regression results.
- **Model Assumption Validation:** Ensuring that residuals follow a normal distribution is essential for making valid statistical inferences based on linear regression results.

A well-behaved Q-Q plot supports the reliability of linear regression assumptions and enhances the model's interpretability.