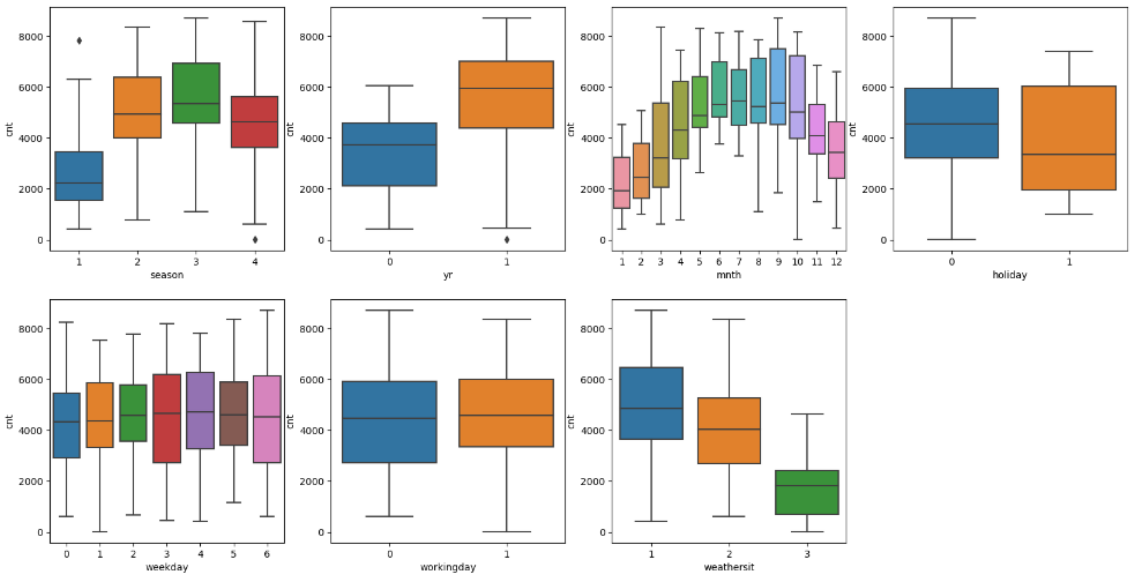
1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**

**Ans:** Refer to the figure:

- The target variable count has increased by a great margin for 2019.
- There is a sudden jump in the count during the summer and winter season. The count is relatively less during the spring season.
- There is a seasonal trend in the count w.r.t the month. The count is increasing and peaks in the period August-October.
- There is not significant trend for the weekday, the median is same in all the cases
- The median is same for both the conditions of the working day
- The count is relatively high during the weather sit 1 condition (Clear, Few clouds, Partly cloudy, Partly cloudy).




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

**Ans:** This is used to avoid predicting the value of the other variable. This also reduces the correlation between the dummy variables. It is preferred to have n-1 dummy variables. The dummy variables should explain all the aspects of the categorical variables. Also, the goal is to have minimum columns to work with. This increases the readability and agility of the workbook.

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

**Ans:** Temp and atemp are highly correlated. The correlation factor is of 0.99

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

**Ans:** We look for the following:

a. Error terms are normally distributed.
b. The predicted value have linear relationships (Significant R^2)
c. P-value for all the variables are less than 0.05
d. VIF is less than 5

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

**Ans:** The following could be the top features contributing significantly:

**a.** Year
**b.** Temperature
**c.** Light Rain

**General Subjective Questions**

1. Explain the linear regression algorithm in detail:
    a. Linear regression is a statistical method used for modeling the relationship between a dependent variable and one or more independent variables. The basic idea is to find a linear equation that best fits the data points and can be used to predict the value of the dependent variable for given values of the independent variables.
    b. Components of LR
        i. Dependent Variable(Y) : This is the target variable which we want to predict
        ii. Independent Variable(X) : These are used to predict the target variable
        iii. Equation : Y= Beta 0 + Beta1 *X + err (Beta 0 is the intercept, Beta 1 is the slope, err is the error)
        iv. Residuals : Difference b/w the actual and predicted values
        v. Hypothesis Testing : This is used to determine whether the Beta 1 is significant
    c. Steps for LR:
        i. Data Cleaning
        ii. Data Visualization
        iii. EDA
        iv. Train Test Split
        v. Scaling
        vi. RFE Model Creation
        vii. Model fitting using Stats Model
        viii. Model Evaluation
            1. P-value < 0.05
            2. Significant R^2
            3. VIF<5
        ix. Residual Analysis
            1. Error distribution
        x. Model Validation
            1. Test data v/s Pred Test data
            2. R^2

2. **Explain the Anscombe's quartet in detail**
    a. Anscombe's quartet consists of four datasets that, when scatter plotted on a graph, have different representations despite sharing the same descriptive statistical features (means, variance, R-squared, correlations, and linear regression lines). The

datasets were developed in 1973 by statistician Francis Anscombe to highlight the value of data visualization and highlight the limitations of summary statistics on their own.

b. Anscombe's quartet consists of four datasets, each containing eleven x-y pairs of data. Plotting each dataset appears to show a different relationship between x and y, with varied correlation strengths and variability patterns. The summary statistics for every dataset, including the x and y mean and variance, x and y correlation coefficient, and similar values, remain constant despite these differences.

**3. What is Pearson's R**

**Ans:** Pearson's R is a number between -1 and 1 that measures the strength and direction of the relationship between two variables.

a. **0 < R < 1:** R is positive and The one variable changes the other variable in a positive manner (i.e in the same direction)
b. **R = 0:** R is zero. There is no relationship
c. **-1 < R < 0:** R is negative. The one variable changes the other variable in a negative manner

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

**Ans:** Scaling is used to bring all the variables to the same scale. Variables on different scale might have very different coefficients making it difficult to interpret or converge.

Scaling is performed for the following reasons:

a. Ease of Interpretation
b. Faster convergence

**Standardized Scaling:** The variables are standardized in such a way that their mean is zero and S.D is 1

X = x-mean(x)/standard deviation (x)

**Normalized Scaling:** The variables are scaled in such all the values lies between 0 and 1 using max and min value of the data.

X = x-min(x)/(max(x)-min(x)

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

**Ans:** The VIF is defined as 1/(1-R^2). Here the value of VIF will tend to infinite when the denominator is 0. This will occur when the R value is 1 or -1. That is when two values are perfectly correlated to each other. This will mean that the model is overfitting. One issue of this could be, the EDA is not performed correctly or the constant is not added during the sm model

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

**Ans:** A Q-Q plot is a scatter plot used for comparing two probability distributions by plotting their quantities against each other.

The Q-Q Plot can be used for the following:

a. Behavior of two samples
b. Behavior of tails of two sample
c. Comparing of dataset to theoretical model
d. Shape of the distribution