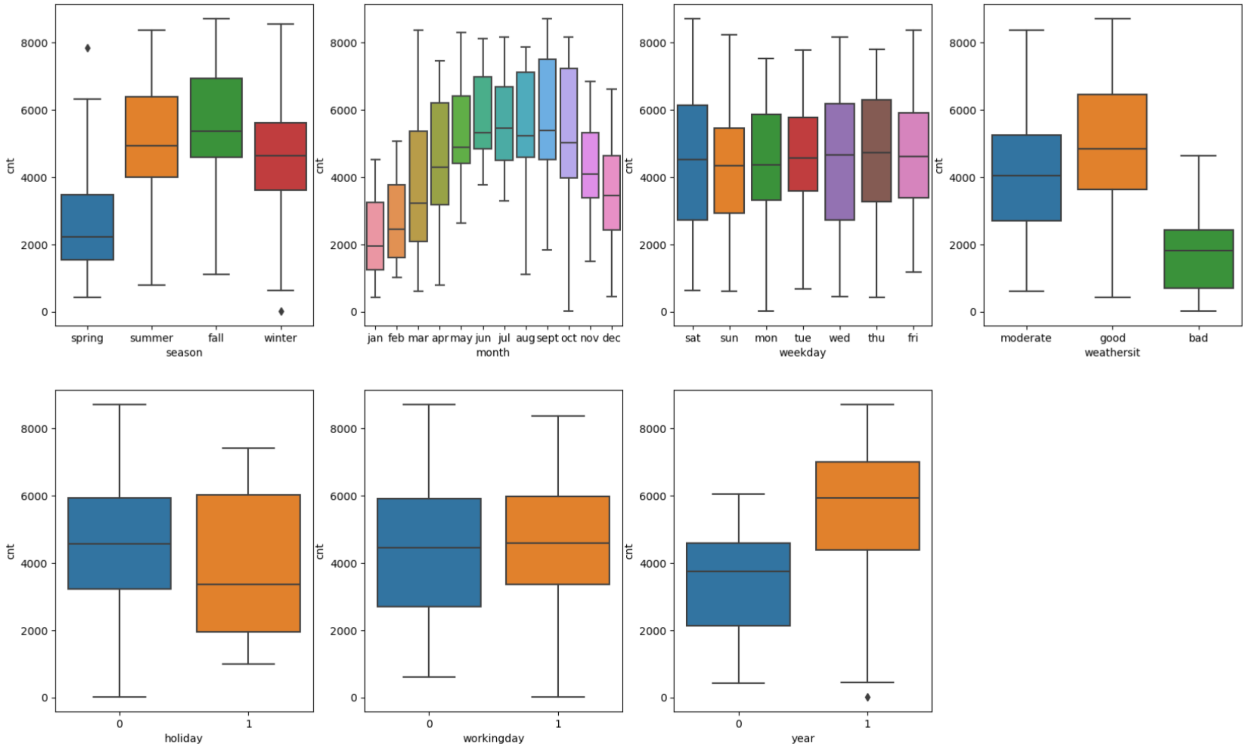
## Assignment-based Subjective Questions

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

Answer: From my analysis on categorical variables using box plot following are the few observations that we can infer.

- Season - Fall & Summer seasons attracted more bookings as people like to come out and enjoy the weather.
- Mnth- Most of the bookings happening from June to October months. Booking dipped in November, December, January and February months due to cold weather.
- Weekday - Saturday has a greater number of bookings when compared to other days in a week.
- Weathersit - Always a good weather attracts more bookings
- Holiday - Higher bookings are happening over the holidays.
- Yr - 50% percent of booking has increased in 2019 year when compared with 2018

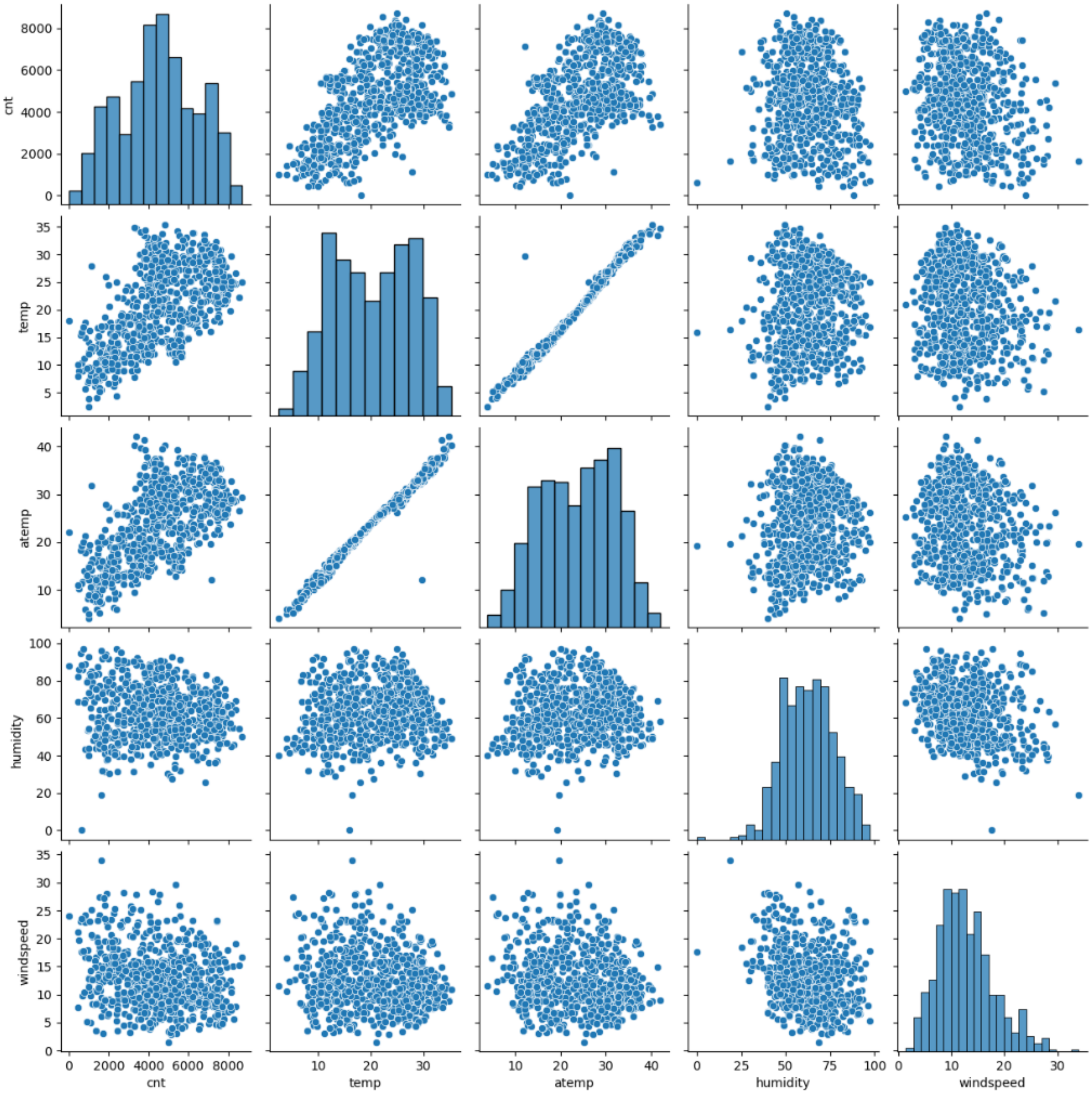


2. **Why is it important to use drop_first=True during dummy variable creation?**

Answer: While creating a dummy variable for any Categorical variable values, it’s always recommended to create n-1 variables (where n is the total number of distinct values of Categorical variable) to reduce the correlation between these dummy variables.

drop_first=True helps in dropping/reducing the extra column during the dummy value creation process.

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



4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**
   Answer: I have performed the linear regression validation based on the following assumptions.
   - Linear relation between the target and other variables
   - Multicollinearity: There is minimum or zero multicollinearity in the data. Multicollinearity occurs when there is a high correlation between the independent variables.

- Error terms: Error terms should be equally distributed. Residual distribution should have normal distribution

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**
   Answer:
   - Temp
   - Workingday
   - Windspeed

General Subjective Questions

1. Explain the linear regression algorithm in detail.
   Answer: Linear regression is a machine learning algorithm based on supervised learning. The process of predicting a continuous variable(y) based on one or many input variables(x) is a linear regression. It is mostly used for finding out the linear relationship between variables and forecasting.
   There are two types of Linear regressions:
   - **Simple Linear regression**: In simple linear regression there will be one dependent and one independent variable. The goal is to find out the linearity between these two variables.

     $Y= \beta0 + \beta1*X + \varepsilon$

     β0 is the constant and β1is the slope and X is the independent variable. whereas ε (epsilon) is the error term.

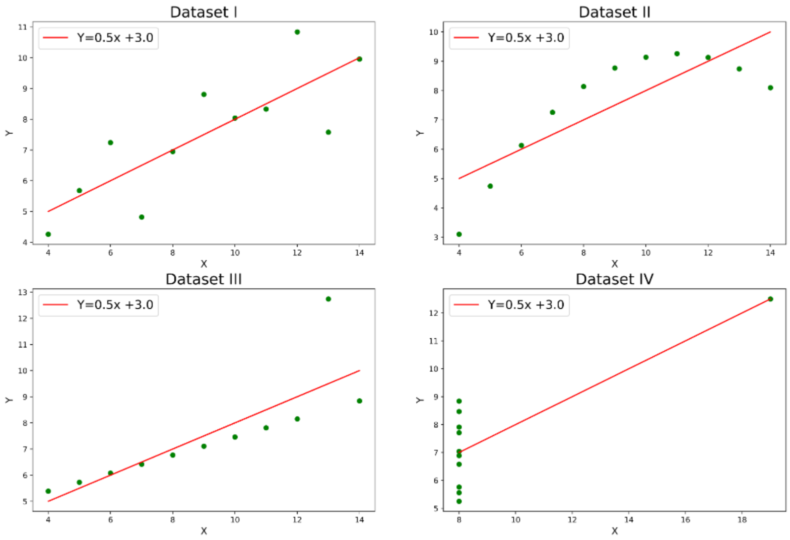
   - **Multi Linear regression**: In multiple linear regression there will be one dependent and multiple independent variables

     $Y= \beta0*X0 + \beta1X1 + \beta2X2+\ldots\ldots \beta nXn+ \varepsilon$

     As the number of independent variables increases, the constants also increase correspondingly.

2. **Explain the Anscombe's quartet in detail.**
   Answer: Anscombe's Quartet was developed by statistician Francis Anscombe. It contains four data sets that have almost identical statistical features, but they have a very different distribution and look totally different when plotted on a graph.
   It was developed to emphasize both the importance of graphing data before analyzing it and the effect of outliers and other influential observations on statistical properties.

- The first scatter plot (Dataset 1) appears to be a simple linear relationship.
- The second graph (Dataset 2) is not distributed normally; while there is a relation between them is not linear.
- In the third graph (Dataset 3), the distribution is linear, but should have a different regression line the calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient.
- The fourth graph (Dataset 4) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.



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

Answer: Pearson R is the numerical summary of the strength of strength of the linear associaΘon between the variables. It's value ranges between -1 to +1. It shows the linear relationship between two sets of data.

$$r = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i (x_i - \overline{x})^2} \sqrt{\sum_i (y_i - \overline{y})^2}}$$

If r=1 means a perfect positive relationship whereas r=-1 means a negative correlation.

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

scaling is a method used to standardize the range of independent variables. It is performed during the data pre-processing stage to deal with varying values in the dataset.

If scaling is not performed, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, irrespective of the units of the values.

**Normalized Scaling**:

- Minimum and maximum value of features are used for scaling
- It is used when features are of different scales.
- Range between -1 to 1
- Scaling is affected by outliers
- Scikit-Learn provides a transformer called MinMaxScaler for Normalization.

**Standardized Scaling:**

- Mean and standard deviation is used for scaling
- It is used when we want to ensure zero mean and unit standard deviation.
- This is not bounded to a certain range.
- It is much less affected by outliers
- Scikit-Learn provides a transformer called StandardScaler for standardization.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
   If there is perfect correlation, then VIF = infinity.
   VIF = $1/1-R^2$

   - VIF equal to 1 = variables are not correlated
   - VIF between 1 and 5 = variables are moderately correlated
   - VIF greater than 5 = variables are highly correlated