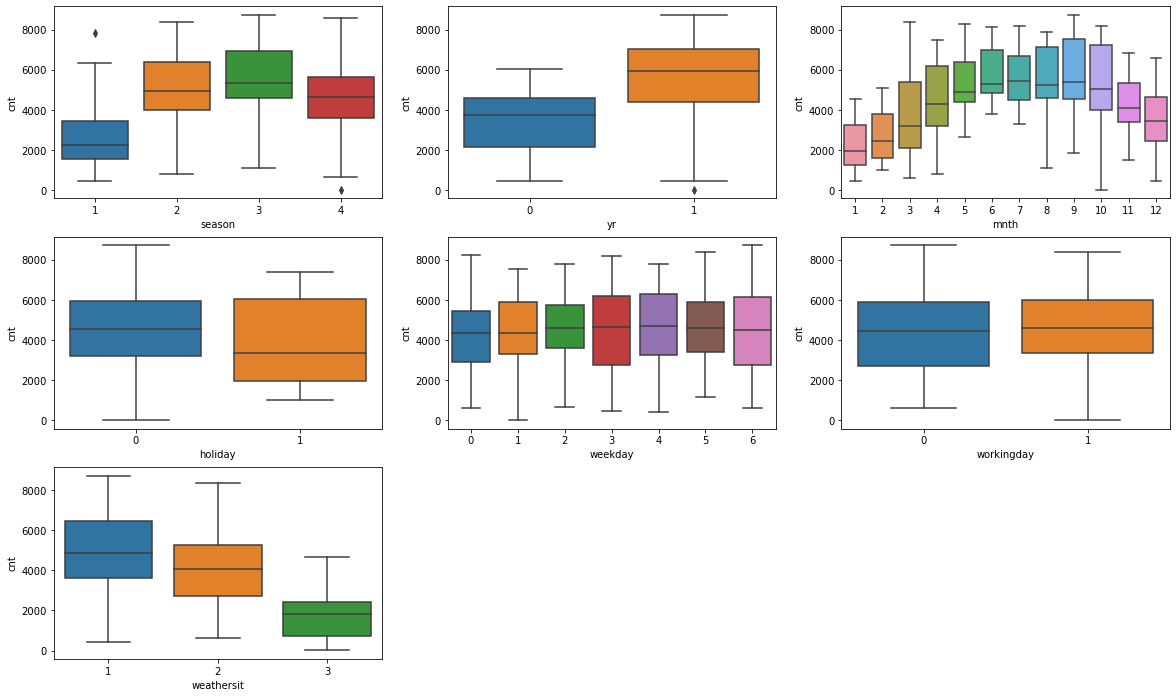
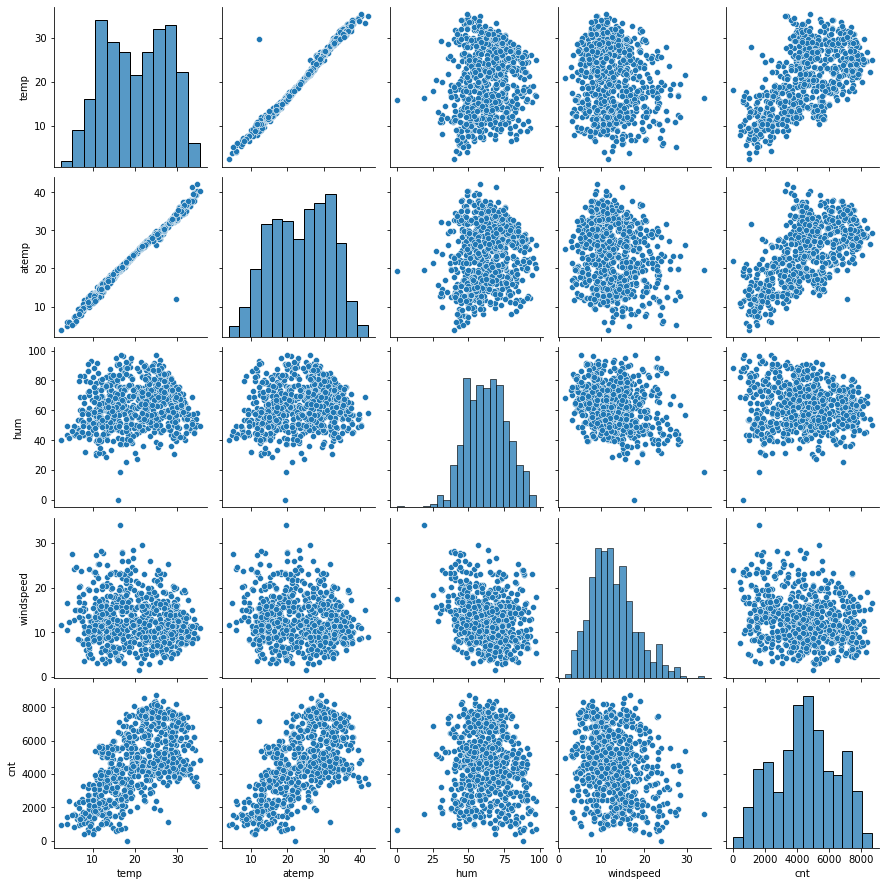
**Assignment-based Subjective Questions**

1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**
2. Lets see the Box Plot of all Categorical values present in the data.



From the Graph we can clearly say that Except Weekday and Workingday all other variables have effect on CNT variable.

1. If we observe Season variable, the least number of counts produces in spring and it increases in summer and reaches maximum in fall and again decreases in winter it follows cycle.
2. If we observe yr variable the more number of count present in 2019 compared to 2018.which means the Customers are increased in 2019 compared to 2018.
3. If we observe mnth variable, the least number of counts produces in January and it increases maximum in July and again decreases towards December, it follows cycle.
4. If we observe holiday variable, We observe that during holidays we have very less count compared to non holiday .
5. If we observe weathersit variable , more number of bikes will go for rent in when weather is clear , Few clouds, Partly cloudy, Partly cloudy and moderate during Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist and less during Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds.
6. **Why is it important to use drop\_first=True during dummy variable creation?**
7. By using drop\_first = True , it helps to reduce the extra column created during dummy variable creation , so that it helps to reduce the correlation among dummy variables , because the remaining variable can be calculated by remaining variables.
8. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**
9. If pair plot of all the Numerical variables in the data is as follows :



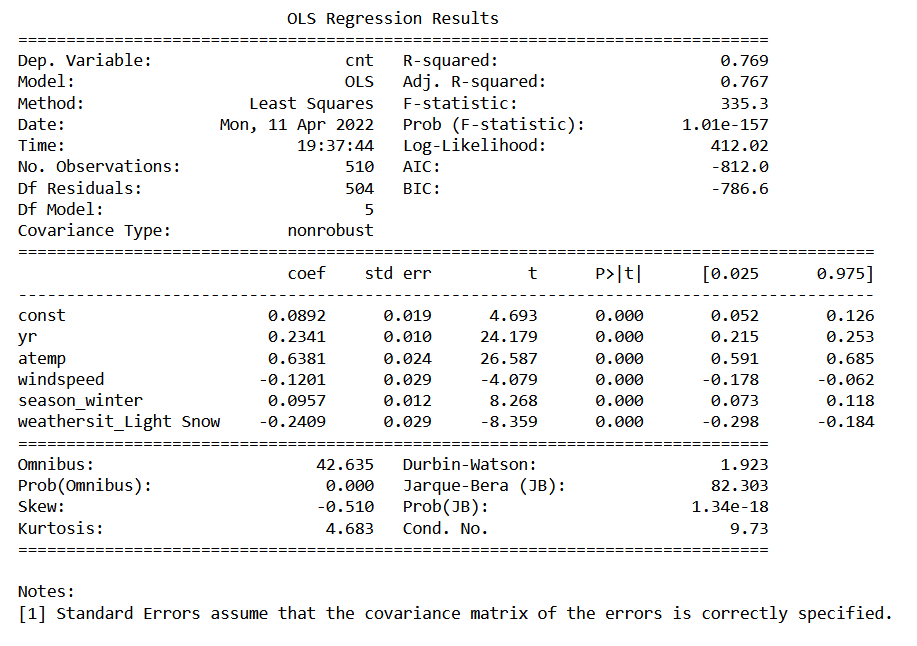
From the above Graph we can Say that temp and atemp are highly correlated to cnt variable.

In these two atemp Variable is highly correlated to cnt variable.

1. **How did you validate the assumptions of Linear Regression after building the model on the training set?**
2. There are four Assumptions of Linear Regression:
3. Linearity between dependent and Independent variables, This can be verified by using scatter plots between Independent and dependent variables to check for Linearity.
4. **Homoscedasticity , All values of the variable X(Independent variables) should have same variance, This can be verified by using Scatter plot of the Residuals of model, we should not see any pattern in the graph.**
5. Little or no Multi collinearity between the features, This can be verified by using the Heat-plot of variables correlation matrix.
6. Residuals should be Normally Distributed with mean 0 , This can be verified by using Q-Q plot or Histogram of Residuals.

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

A. Lets see the Model Summary:



From Above we can say that The Top 3 Features Contributing significantly towards explaining the demand of the shared bikes are :

1. atemp.
2. yr.
3. weathersit\_Light Snow (negatively affecting).

**General Subjective Questions**

1. **Explain the linear regression algorithm in detail.**

A.

**2. Explain the Anscombe’s quartet in detail.**

A.

**3.** **What is Pearson’s R?**

A.

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

A.

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

A.

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

A.