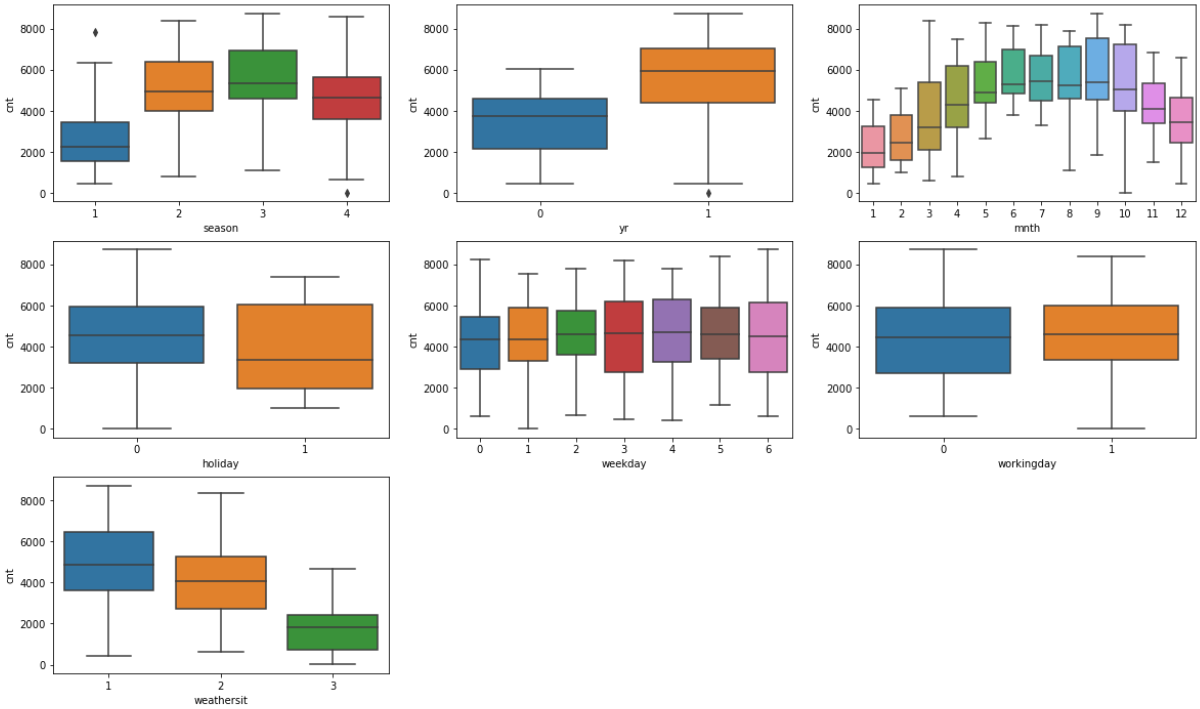
Assignment-based Subjective Questions

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

A. 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.

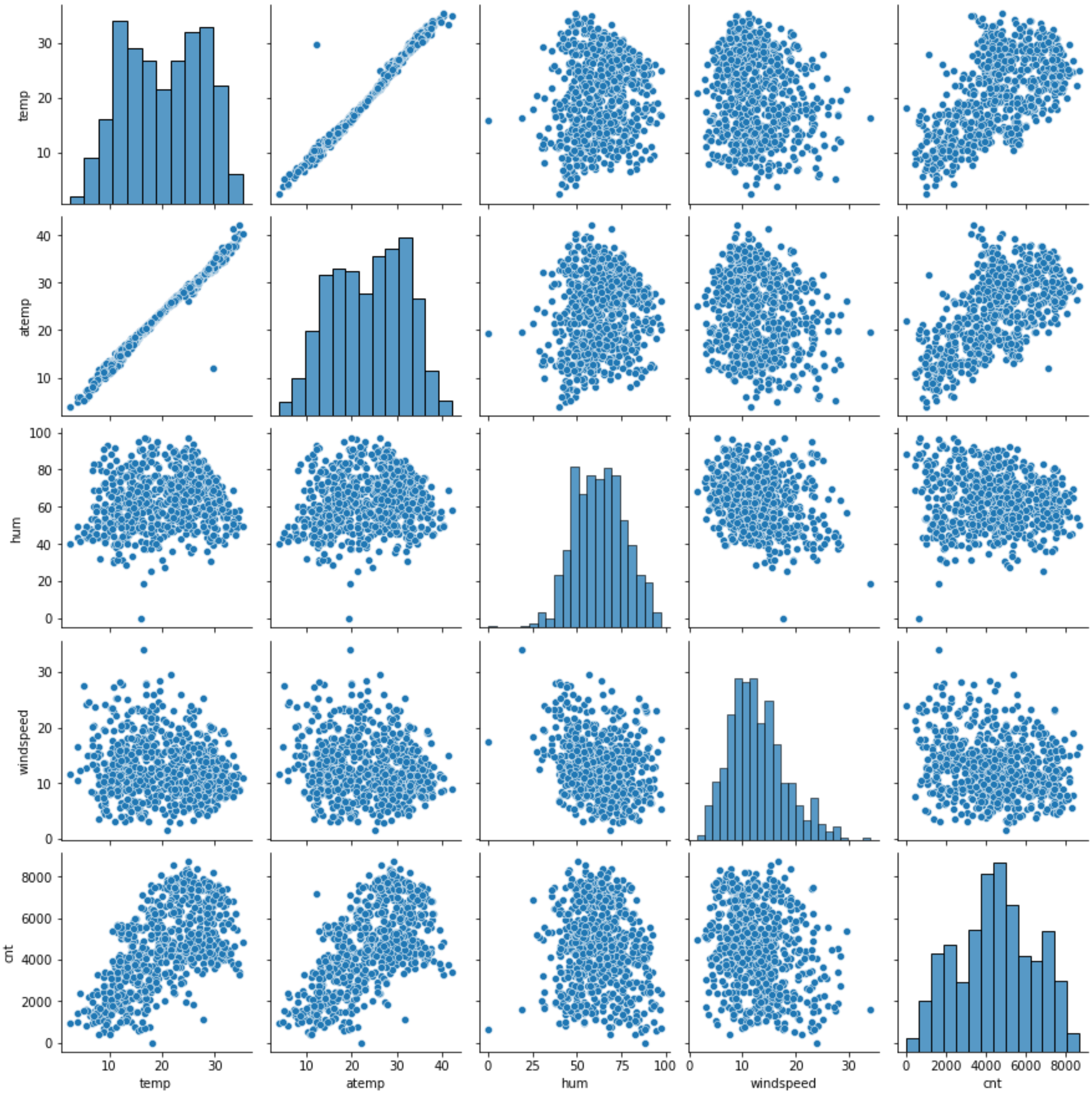
1) If we observe Season variable, the least number of counts produces in spring and it increases in summer and reaches maximum in fall and again decreases in winter it follows cycle.
2) If we observe yr variable the more number of count present in 2019 compared to 2018.which means the Customers are increased 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.

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

A. 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.

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

A. 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.

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

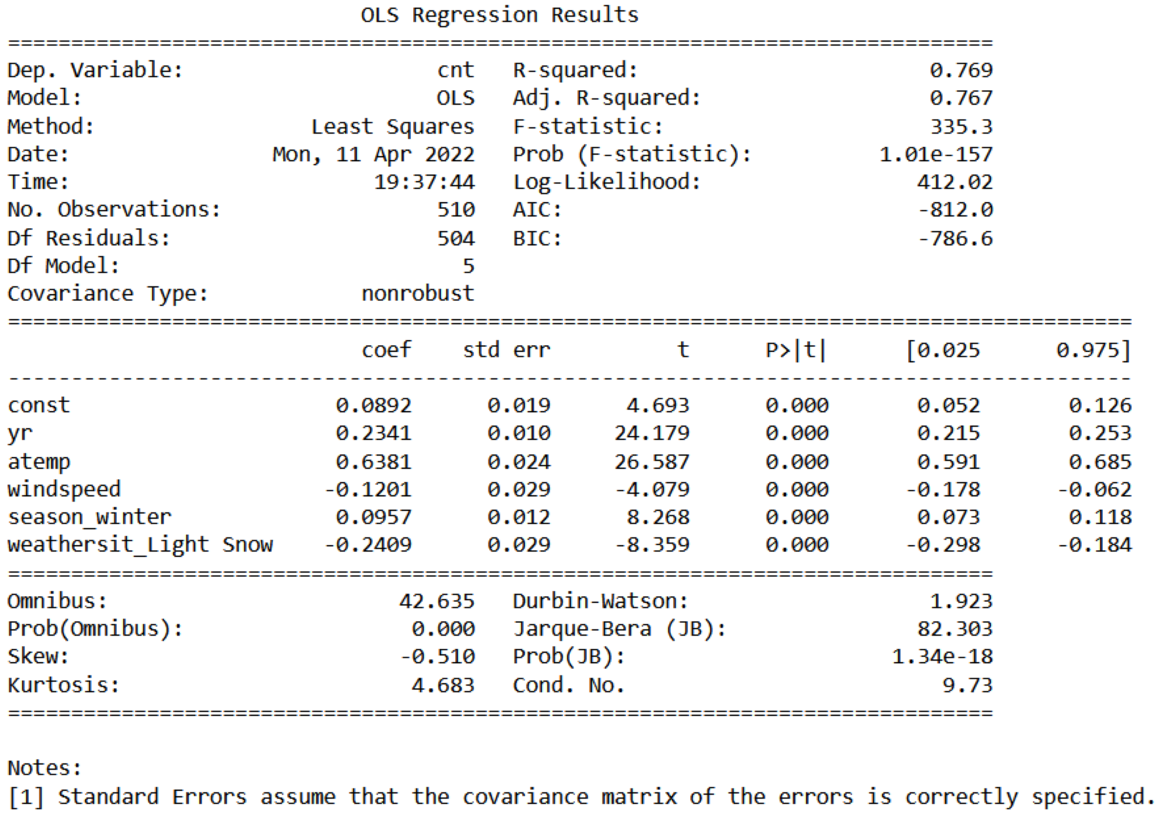
A. There are four Assumptions of Linear Regression:

   1. Linearity between dependent and Independent variables, This can be verified by using scatter plots between Independent and dependent variables to check for Linearity.

2. 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.
3. Little or no Multi collinearity between the features, This can be verified by using the Heat-plot of variables correlation matrix.
4. 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:



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    cnt   R-squared:                       0.769
Model:                            OLS   Adj. R-squared:                  0.767
Method:                 Least Squares   F-statistic:                     335.3
Date:                Mon, 11 Apr 2022   Prob (F-statistic):          1.01e-157
Time:                        19:37:44   Log-Likelihood:                 412.02
No. Observations:                 510   AIC:                            -812.0
Df Residuals:                     504   BIC:                            -786.6
Df Model:                           5
Covariance Type:            nonrobust
=========================================================================================
                            coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------
const                     0.0892      0.019      4.693      0.000       0.052       0.126
yr                        0.2341      0.010     24.179      0.000       0.215       0.253
atemp                     0.6381      0.024     26.587      0.000       0.591       0.685
windspeed                -0.1201      0.029     -4.079      0.000      -0.178      -0.062
season_winter             0.0957      0.012      8.268      0.000       0.073       0.118
weathersit_Light Snow    -0.2409      0.029     -8.359      0.000      -0.298      -0.184
==============================================================================
Omnibus:                       42.635   Durbin-Watson:                   1.923
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               82.303
Skew:                          -0.510   Prob(JB):                     1.34e-18
Kurtosis:                       4.683   Cond. No.                         9.73
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

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

a. atemp.
b. yr.
c. weathersit_Light Snow (negatively affecting).

General Subjective Questions

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

A. Linear Regression is an basic Supervised Machine Learning Model which is used to predict Continuous Variables. It is used to find out the Linear relationships present between Two Variables.

The basic Equation for Linear Regression is y = mX+c, where y = dependent variable, X is Independent variable, c is Constant.
We use Root Mean Square error as cost Function for Linear Regression. Our aim is to build the Line with minimising the Cost Function.

2. Explain the Anscombe's quartet in detail.

A. Anscombe's Quartet is a set o our data sets Which are identical in descriptive statistics(Mean, Variance, Standard Deviation), But when we plot them in Scatter plot they all are very different in their appearance Pattern. It can Simply fools the Regression model if built , So before Building the model it is necessary to graph the variables and see their Nature.

3. What is Pearson's R?

A. Pearson's R is a statistic that measures the Correlation between two variables . It has a Numerical value which ranges from -1 to +1.

Mathematically it is defined as covariance of two variables divide by product of their standard deviations.

If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

Pearson's R is failed to explain the slope of the correlation line.

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

A. Scaling is one of the data pre-processing step performed on variables to normalize the data within a particular range.

Mostly the data contains variables from different ranges which are highly variable, If scaling is not done Algorithm takes only magnitude in account not units which results in wrong modelling. to resolve this issue we need to bring all the variables to similar range.

And most of the algorithms like Gradient based Algorithms, Distance based and Tree Based Algorithms are sensitive to feature Scaling.

Normalization is a scaling technique in which values are rescaled ending up in the range between 1 and 0 , It is also known as Min-Max Scaling. The formulae or Normalization is as follows :

X' = X-Xmin/Xmax-Xmin.

Standardization is another scaling technique in which the values are centered around zero and with standard deviation of 1, so the range is -1 to 1.The formulae :

X' = X-mean/std.

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

A. When two variables are Perfectly correlated to each other the Their R2 =1 , And VIF Formulae is given by 1/(1-R2), so when R2=1 VIF becomes 1/0 which is Infinite.

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

A. A Q-Q plot or a quantile-quantile plot is an image method to determine whether two sets of data are from individuals with the same distribution. The Q-Q plot is a scatterplot formed by arranging two sets of quantiles against each other. If both sets of quantiles come from the same distribution, we should see points that form a nearly straight line.

The Q-Q plot is used to determine if points lie approximately in line. If they do not, then our residuals are not Normal and therefore, our errors are also not Normal.

Importance o Q-Q plot:
1. Sample sizes do not need to be equal.
2. Multiple distribution features can be tested at once. For example, shifts in space, scale changes, changes in symmetry, and the presence of outsiders.
3. The structure of q-q can provide more insight into the nature of the differences than analytical methods.