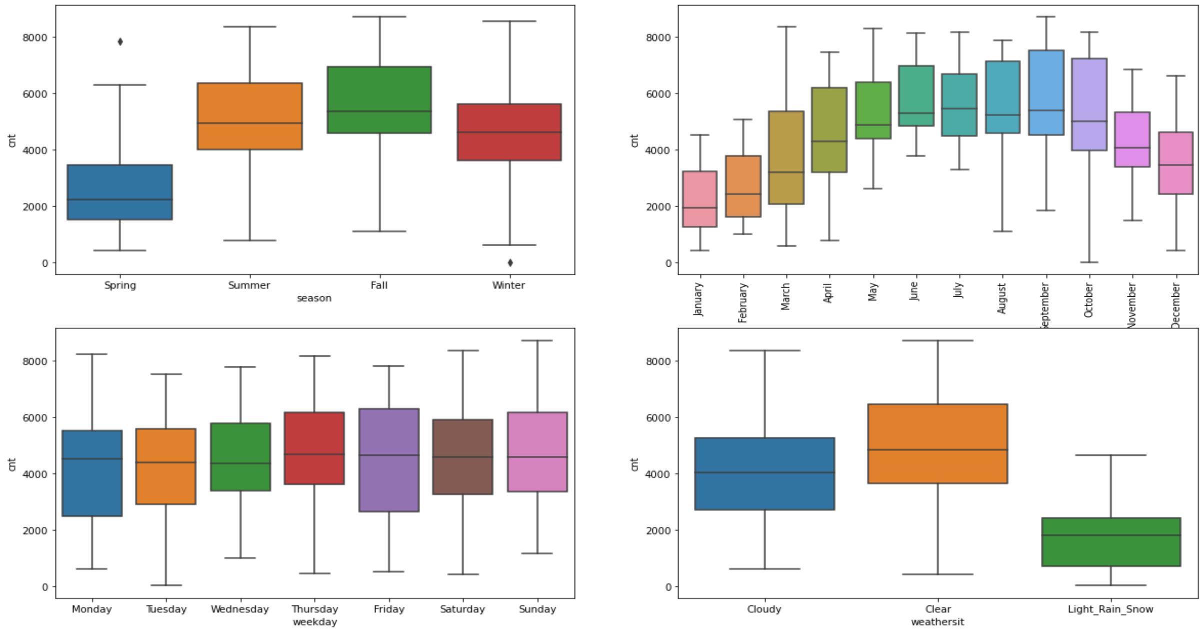
Assignment-based Subjective Questions

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

Ans. The Categorical Variables in data set are Season, Month, Weekday and Weathersit. I used, box plot for Univariate Analysis on the categorical variables. The analysis is as follows



Season – We can observe that, more number of bookings for bikes are during **Fall**, followed by **Summer, Winter** and **Spring.** There is a very huge drop in bookings in spring compared to remaining seasons.

Month – Here we observe that, bookings are increasing from the month of March till October and are again gradually decreasing.

Weekday – Here, we can observe that, bookings are mostly uniform as the median is somewhere around 5000 for all the days.

Weathersit – Here, we can observe that, booking are more on clear days and a bit reduced on cloudy days and very low on Rainy/ Snowy days which is understandable as there is no rain covers for bikes.

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

Ans. During creation of dummy variables, it is important to use drop_first = True, because we can get the information by using n-1 variables. For example, lets consider we have a category with 3 variables A, B, C. When we create dummy variables , the available combinations will be as follows

A	B	C
1	0	0
0	1	0
0	0	1

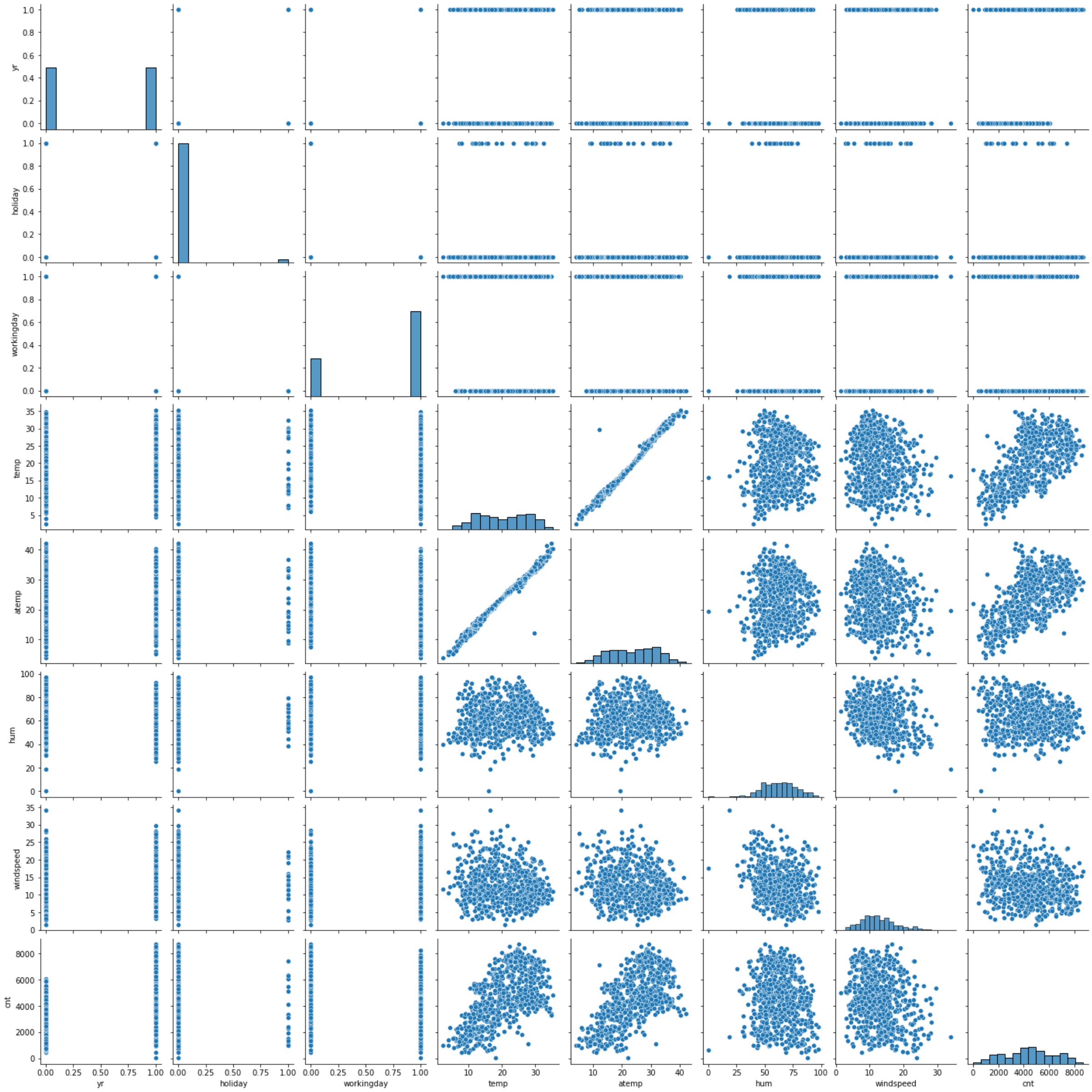
Same can be replicated by using 2 variables also

B	C
0	0
1	0
0	1

This helps us in reducing the number of variables used.

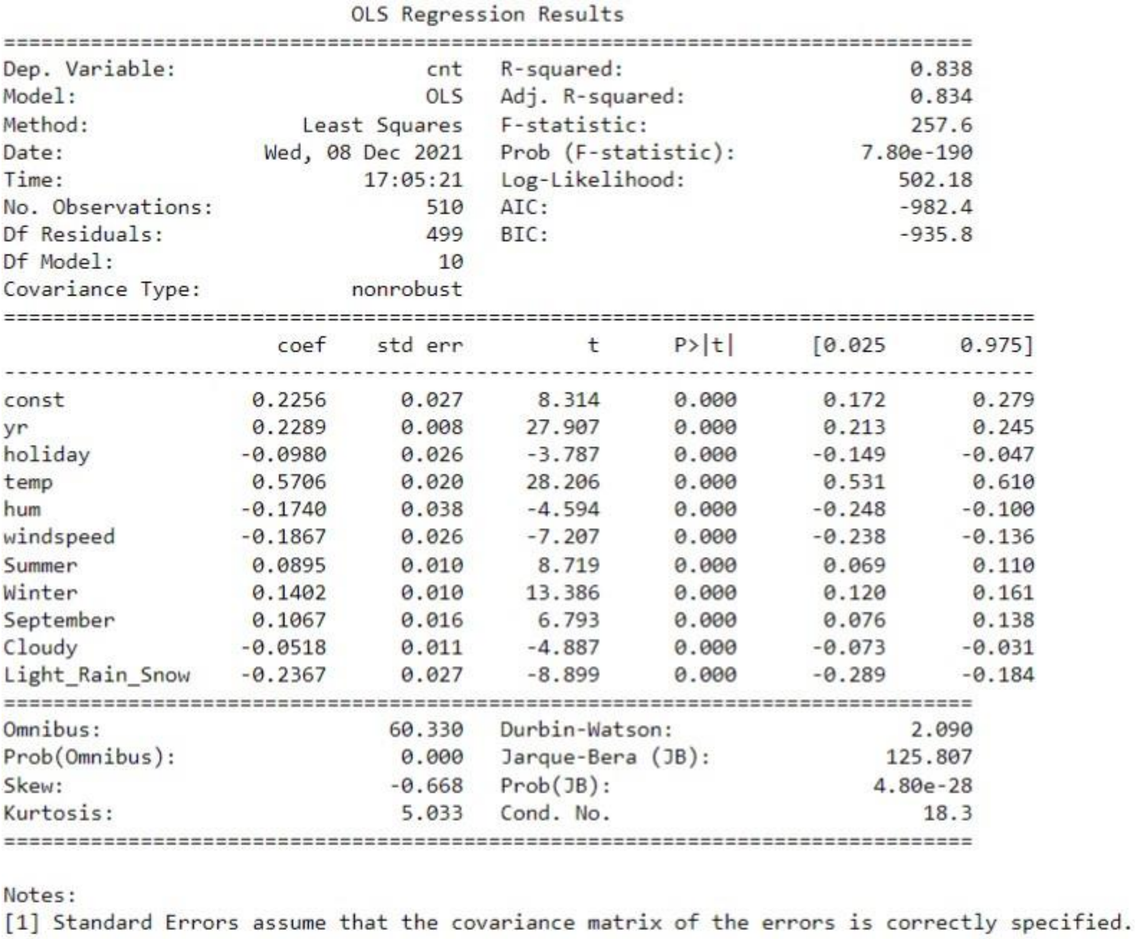
3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**

   Ans. Looking at the pair plot, we can observe the variable with highest correlation with target variable is temp and atemp. I have given both the variables because both of them looks highly correlated with each other,



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

   Ans. Here we can Validate the Linear Regression Model by saying the that, we got a Linear equation for the variables with the target variable and also, by looking at the coefficients, we can also say no two variables are highly correlated to each other.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    cnt   R-squared:                       0.838
Model:                            OLS   Adj. R-squared:                  0.834
Method:                 Least Squares   F-statistic:                     257.6
Date:                Wed, 08 Dec 2021   Prob (F-statistic):          7.80e-190
Time:                        17:05:21   Log-Likelihood:                 502.18
No. Observations:                 510   AIC:                            -982.4
Df Residuals:                     499   BIC:                            -935.8
Df Model:                          10
Covariance Type:            nonrobust
===================================================================================
                      coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------
const               0.2256      0.027      8.314      0.000       0.172       0.279
yr                  0.2289      0.008     27.907      0.000       0.213       0.245
holiday            -0.0980      0.026     -3.787      0.000      -0.149      -0.047
temp                0.5706      0.020     28.206      0.000       0.531       0.610
hum                -0.1740      0.038     -4.594      0.000      -0.248      -0.100
windspeed          -0.1867      0.026     -7.207      0.000      -0.238      -0.136
Summer              0.0895      0.010      8.719      0.000       0.069       0.110
Winter              0.1402      0.010     13.386      0.000       0.120       0.161
September           0.1067      0.016      6.793      0.000       0.076       0.138
Cloudy             -0.0518      0.011     -4.887      0.000      -0.073      -0.031
Light_Rain_Snow    -0.2367      0.027     -8.899      0.000      -0.289      -0.184
==============================================================================
Omnibus:                       60.330   Durbin-Watson:                   2.090
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              125.807
Skew:                          -0.668   Prob(JB):                     4.80e-28
Kurtosis:                       5.033   Cond. No.                         18.3
==============================================================================

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

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

Ans. The 3 features contributing significantly towards explaining the demand of shared bikes are

1. Temperature (temp) – Positively Correlated
2. Light_Rain_Snow (Weathersit) – Negatively Correlated
3. Year (yr) – Positively Correlated

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans. Linear Regression model is a Machine Learning used to predict the Value of a Target Variable based on the Dependant Variable by making a Linear Equation between them.

Basic Linear Equation is

Y (Target Variable) = Constant + (Coefficient of Dependant Variable) * X (Dependant Variable)

But if there are more than 1 dependant Variable, the equation shall be
y = constant + C1 * X1 + C2 * X2 + ………… + Cn * Xn

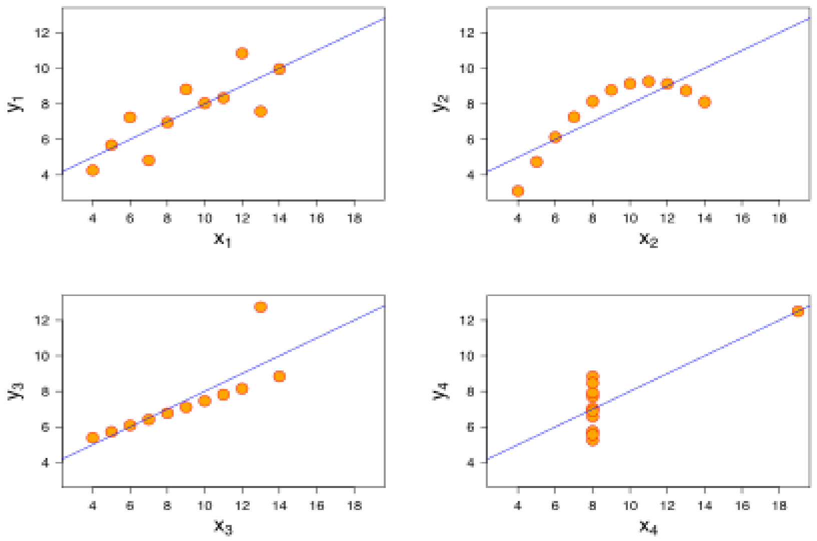
We can find a basic relation between a Target and Dependant Variable by using a Scatter Plot. The more linear the scattering of data, the more the correlation between both variables is higher.

For getting the Linear Regression, we take the data and perform univariate and bivariate analysis on the and the basic correlation between Variables, and then we perform Linear Regression to find the R- Squared Value. We also find out P Value and VIF Value. Based on the P Value and VIF we eliminate the Variables which doesn't give any proper results. We can eliminate variables with P Value more than 0.05 and VIF more than 5.

After completing analysis on the Training Data Set, we check the results on the test data set and get a final prediction.

2. Explain the Anscombe's quartet in detail.

Ans. Anscombe's quartet says that, we can get similar information from different data sets but we can find the difference when we plot the data and it can give a totally different perspective of data. An example for the same is shown below.



A more brief explanation can be statistically, the data for above 4 data sets can be similar, like mean, median etc but when we plot the data sets it may show a very different picture and show how different the data sets are.

This shows the importance of Visualization of data for getting accurate analysis.

3. What is Pearson's R?

Ans. Pearson's R is also called as Pearson's Correlation Coefficient. It is a measure of linear correlation between two data sets.

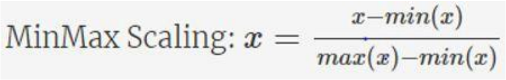
4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**

   Ans. Scaling is process to normalize the range of Independent Variables. It is generally done during data processing.

   Scaling is performed do that all kinds of Independent variables are brought into a certain range.

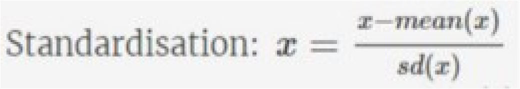
   There are two types of Scaling

   1. Min-Max Scaling (Normalized Scaling) – This kind of scaling brings all the values of a variable between 0 and 1.



$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

   2. Standardized Scaling – Replaces the values with their Z Scores.



$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

   In min max scaling, all values are between 0 and 1 which includes the outliers also in this range and they are also normalised and in standardised scaling, since the values are Z scores, the outliers will be visible.

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

   Ans. The formula for VIF is 1/(1-R^2). When the value of R^2 is 1, then the Value of VIF is infinity.

   This happens when there is a perfect correlation between two independent variables. We will have to drop a variable to avoid this.

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 quantile- quantile plot, which is a graphical plot between two probability distributions.

   They are required for getting a relation between two distributions.