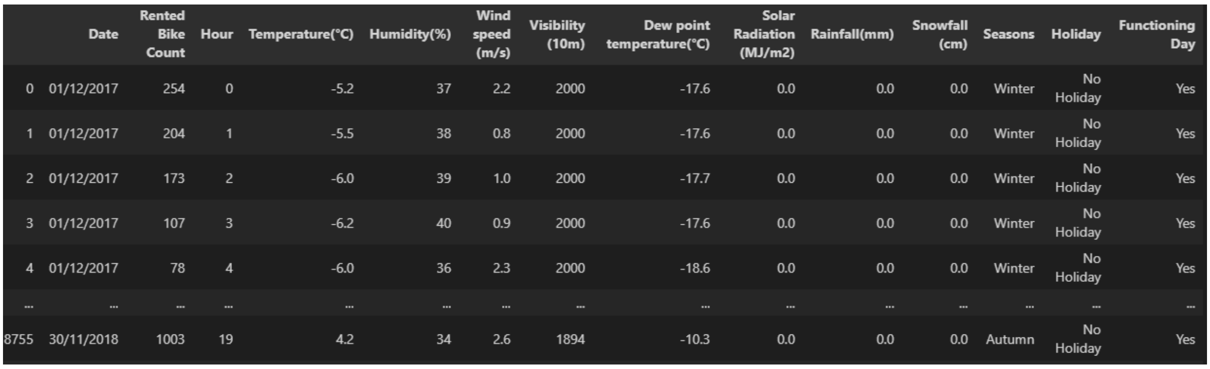
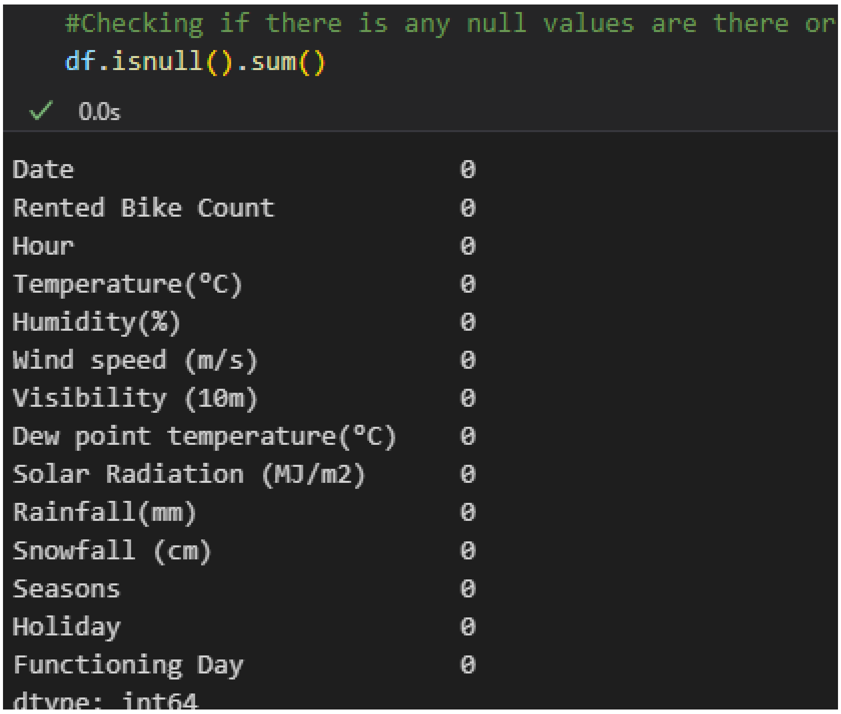
Final Lab Exam(23909-Sairaj)

- **The data is consists of the Seoul Bike sharing data.(Public vehicle in Seoul)**
- **The dataset looks like this:**



	Date	Rented Bike Count	Hour	Temperature(°C)	Humidity(%)	Wind speed (m/s)	Visibility (10m)	Dew point temperature(°C)	Solar Radiation (MJ/m2)	Rainfall(mm)	Snowfall (cm)	Seasons	Holiday	Functioning Day
0	01/12/2017	254	0	-5.2	37	2.2	2000	-17.6	0.0	0.0	0.0	Winter	No Holiday	Yes
1	01/12/2017	204	1	-5.5	38	0.8	2000	-17.6	0.0	0.0	0.0	Winter	No Holiday	Yes
2	01/12/2017	173	2	-6.0	39	1.0	2000	-17.7	0.0	0.0	0.0	Winter	No Holiday	Yes
3	01/12/2017	107	3	-6.2	40	0.9	2000	-17.6	0.0	0.0	0.0	Winter	No Holiday	Yes
4	01/12/2017	78	4	-6.0	36	2.3	2000	-18.6	0.0	0.0	0.0	Winter	No Holiday	Yes
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
8755	30/11/2018	1003	19	4.2	34	2.6	1894	-10.3	0.0	0.0	0.0	Autumn	No Holiday	Yes

- **The dataset contains 8760 rows and 14 columns.**
- **As from the below code we are seeing there is no null values are present in the data , so very less pre-processing is required to the data.**



```
#Checking if there is any null values are there or
df.isnull().sum()
```

✓ 0.0s

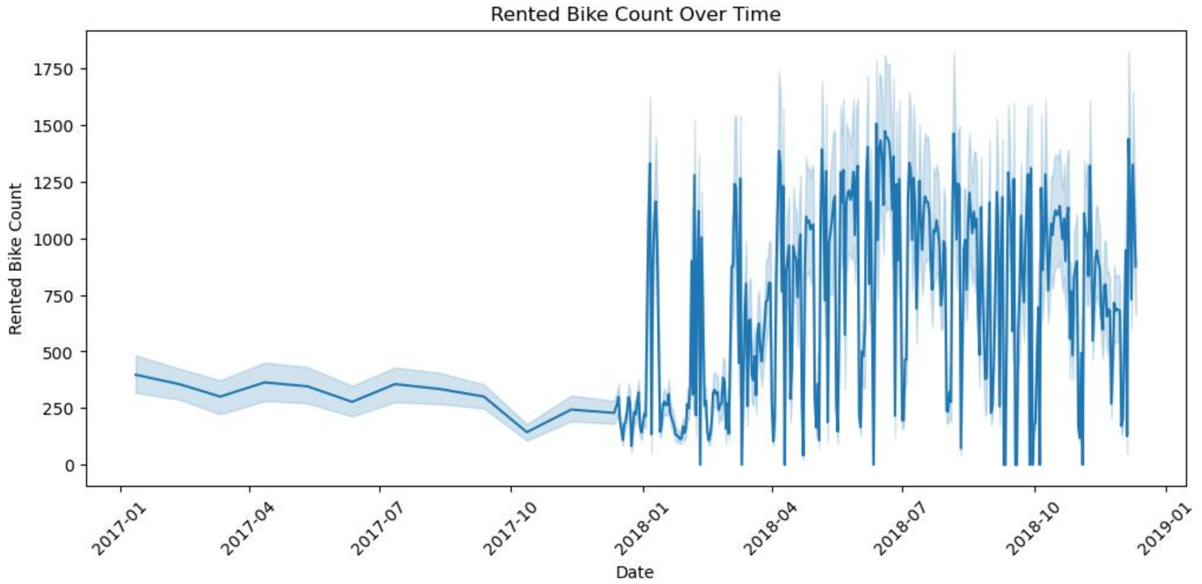
```
Date                         0
Rented Bike Count            0
Hour                         0
Temperature(°C)              0
Humidity(%)                  0
Wind speed (m/s)             0
Visibility (10m)             0
Dew point temperature(°C)    0
Solar Radiation (MJ/m2)      0
Rainfall(mm)                 0
Snowfall (cm)                0
Seasons                      0
Holiday                      0
Functioning Day              0
dtype: int64
```

-

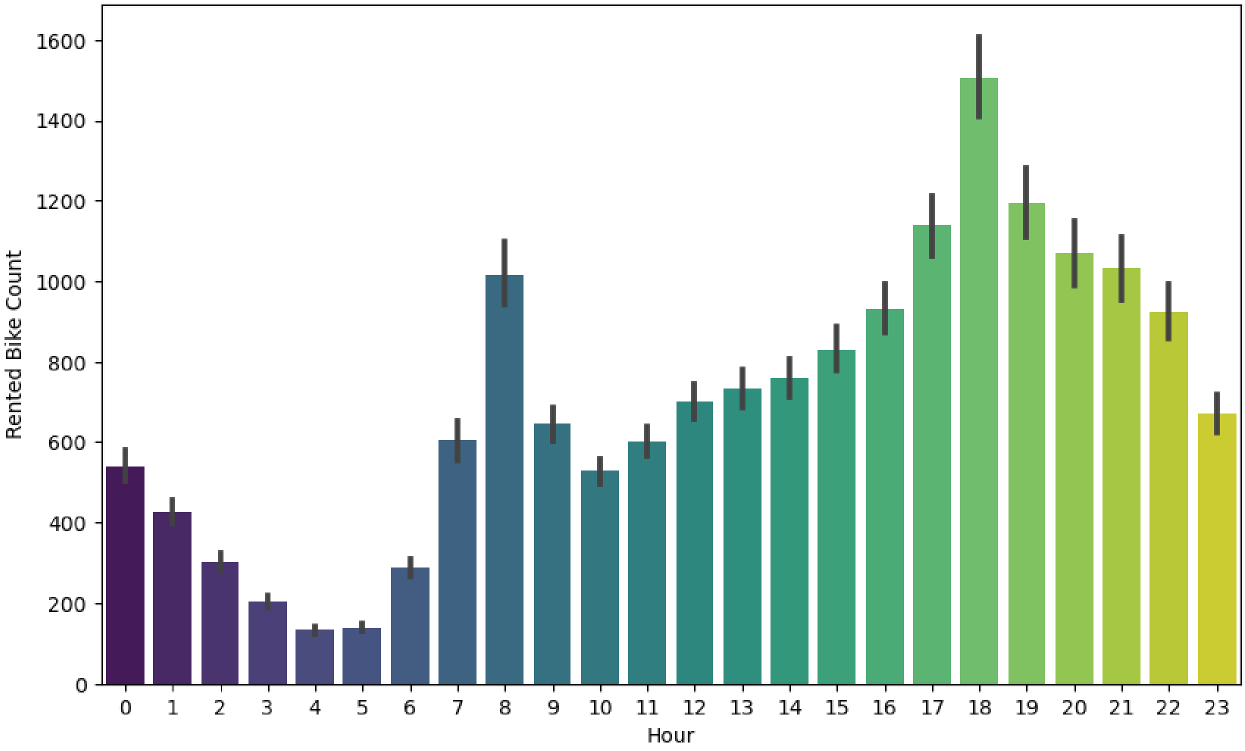
Inferences that can be made out of this data:

1. Line Plot:
    a. From this we can make out that in the year 2017 the number of the rented bikes decreased by the end of the year.

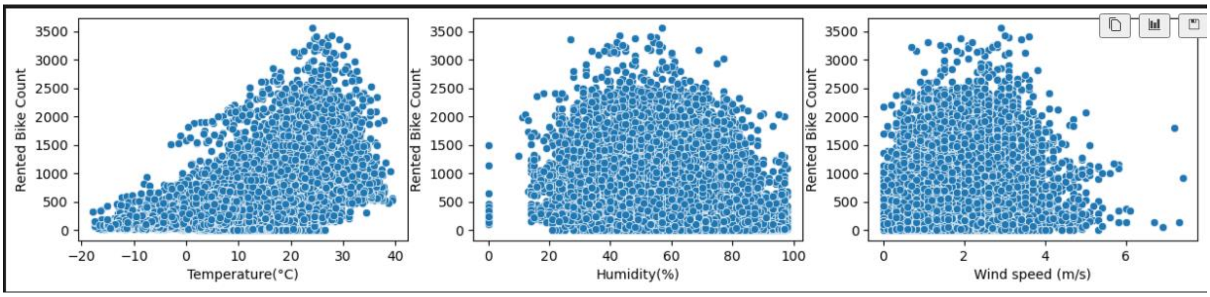
b. Again in the year 2018 the number of rented bikes significantly for some time and again there is a fluctuation the year ahead and it is difficult to make an inference the years ahead.



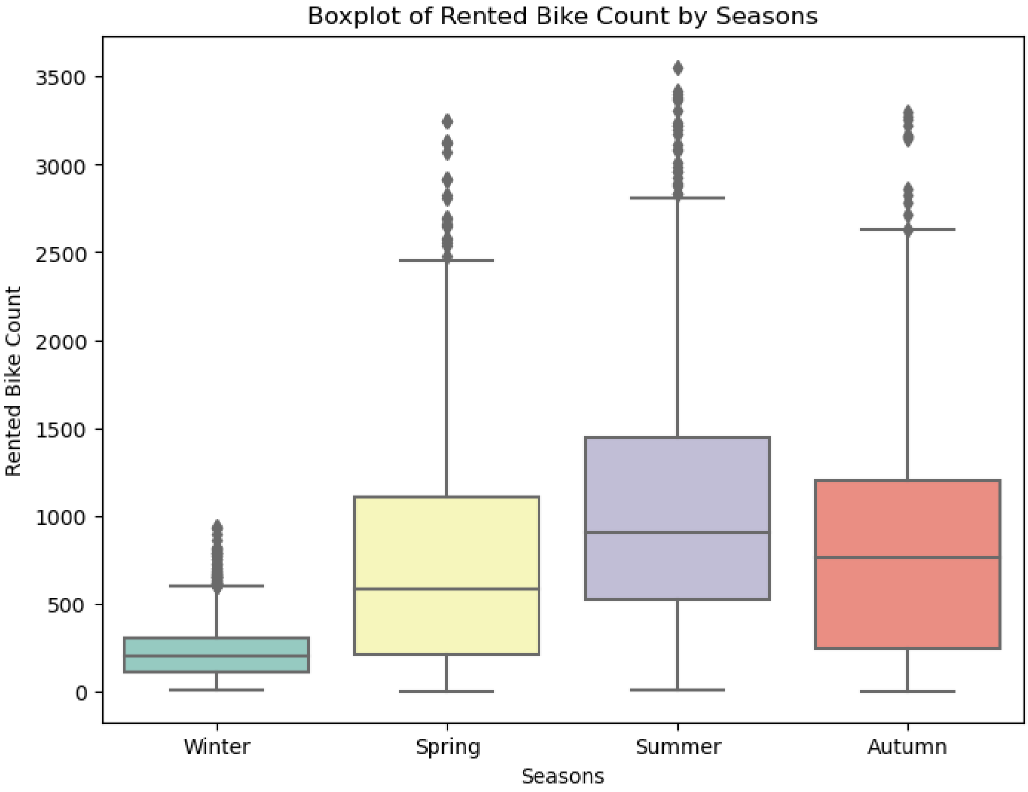
2. Bar Plot:
   a. From this we can clearly see that most of the bikes are rented for 18 hours.
   b. The least number of bikes are rented for only 4-5 hrs.
   c.



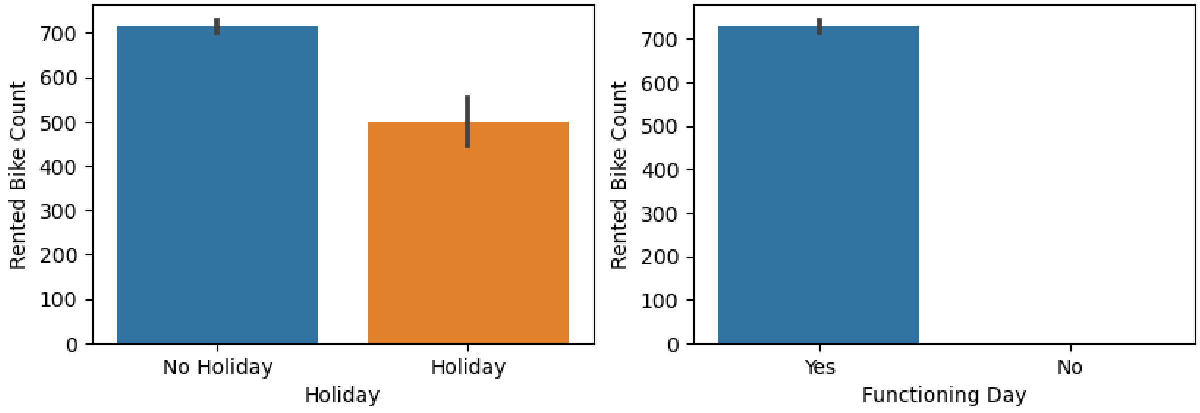
3. Scatter Plot :
   a. The plot between Temperature, Humidity and Wind Speed and Rented Bike Count, shows that:
      - Temperature: When temperature is between 10-30(degree cel.) most of the bikes are rented.
      - Humidity: when Humidity is between 40-80 less number of bikes are rented.
      - Wind Speed: More number of bikes are rented when the wind speed is less.



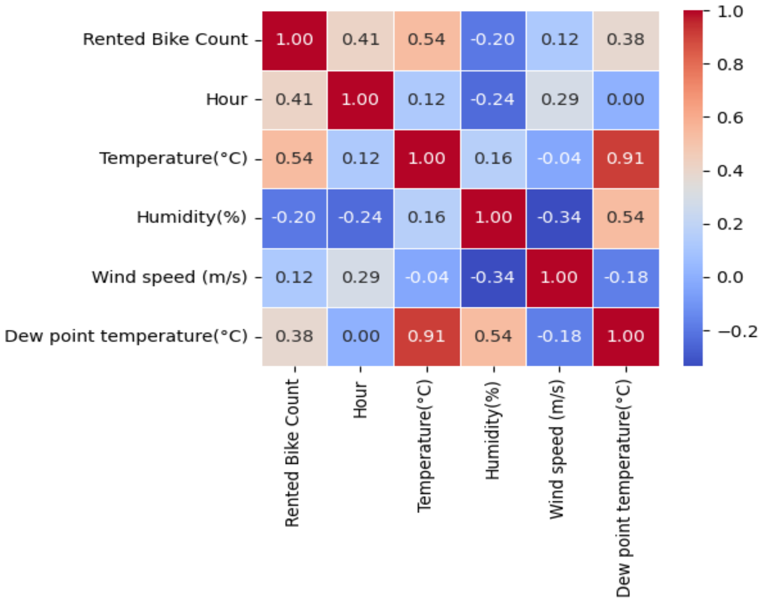
4. Box Plot :
   a. We can clearly see that, comparatively to all other seasons, In summer more number of bikes are rented.
   b. Very less number of Bikes are rented in winter.
   c. Here we can observe the outliers present in all the cases.



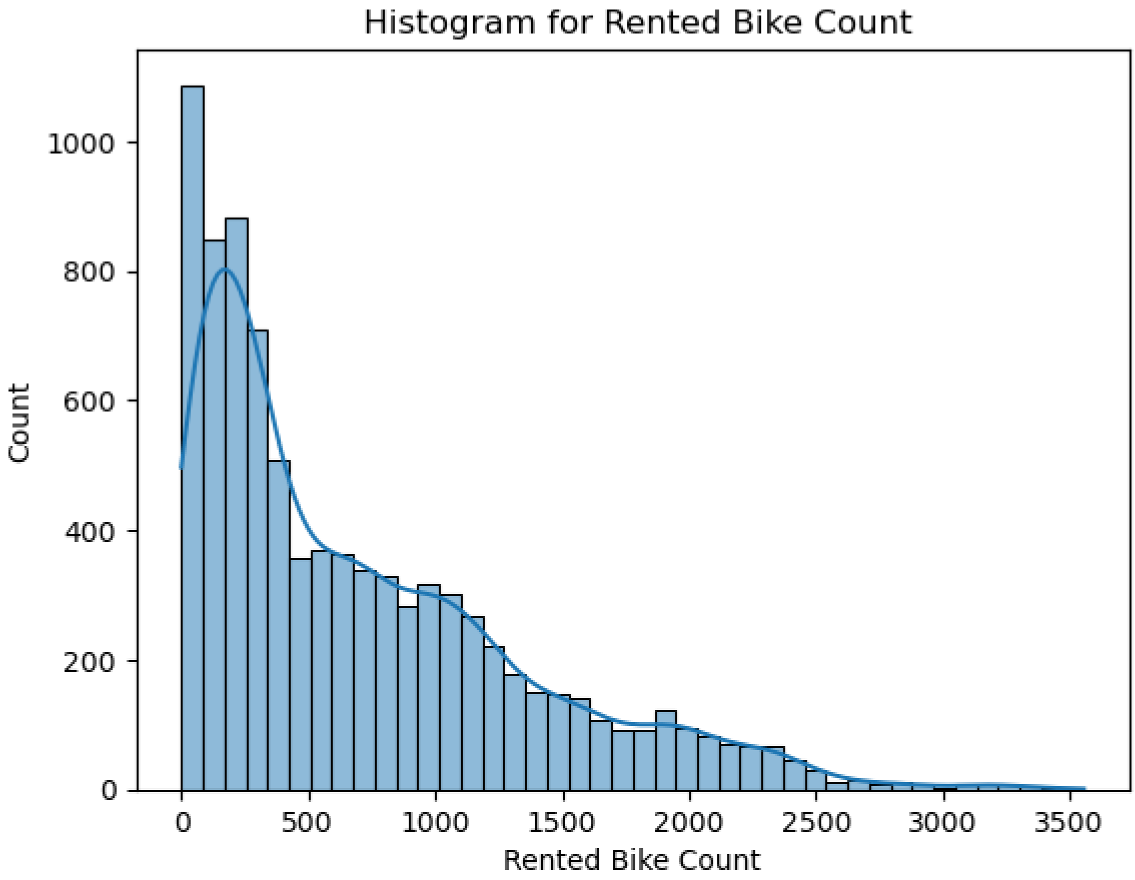
5. Bar plot:
    a. From this bar plot we can make out that on the Non-Holiday and Functioning day the Rented bike Count is more that the Holidays.



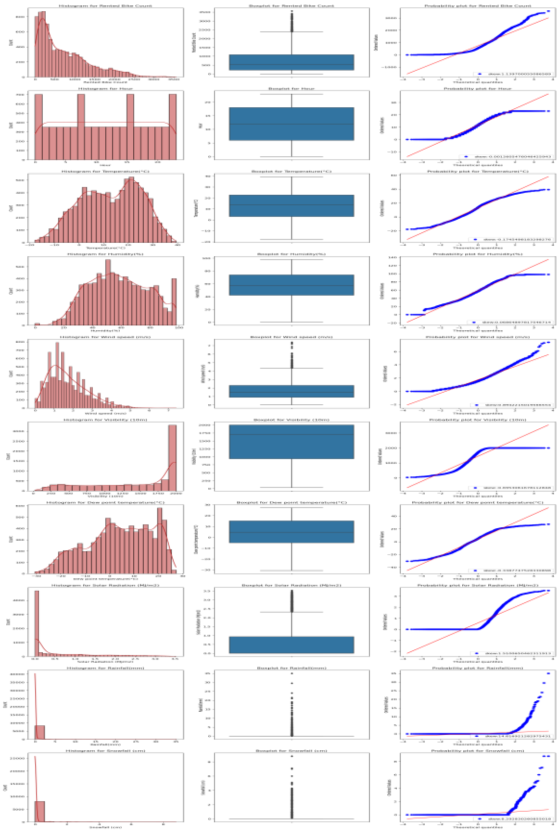
6. Heat Map:
    a. This heat map shows the following:
        - Whenever there is an increase in the hours the rented bike count also increases slightly.
        - Temperature also doesn't have a negative effect on the Bike count.
        - Whenever Humidity increases the Rented bike count is affected directly. The number of Rented bikes count decreases.
        - Wind speed and Due point temp. doesn't have much affect on the count, it is positive.



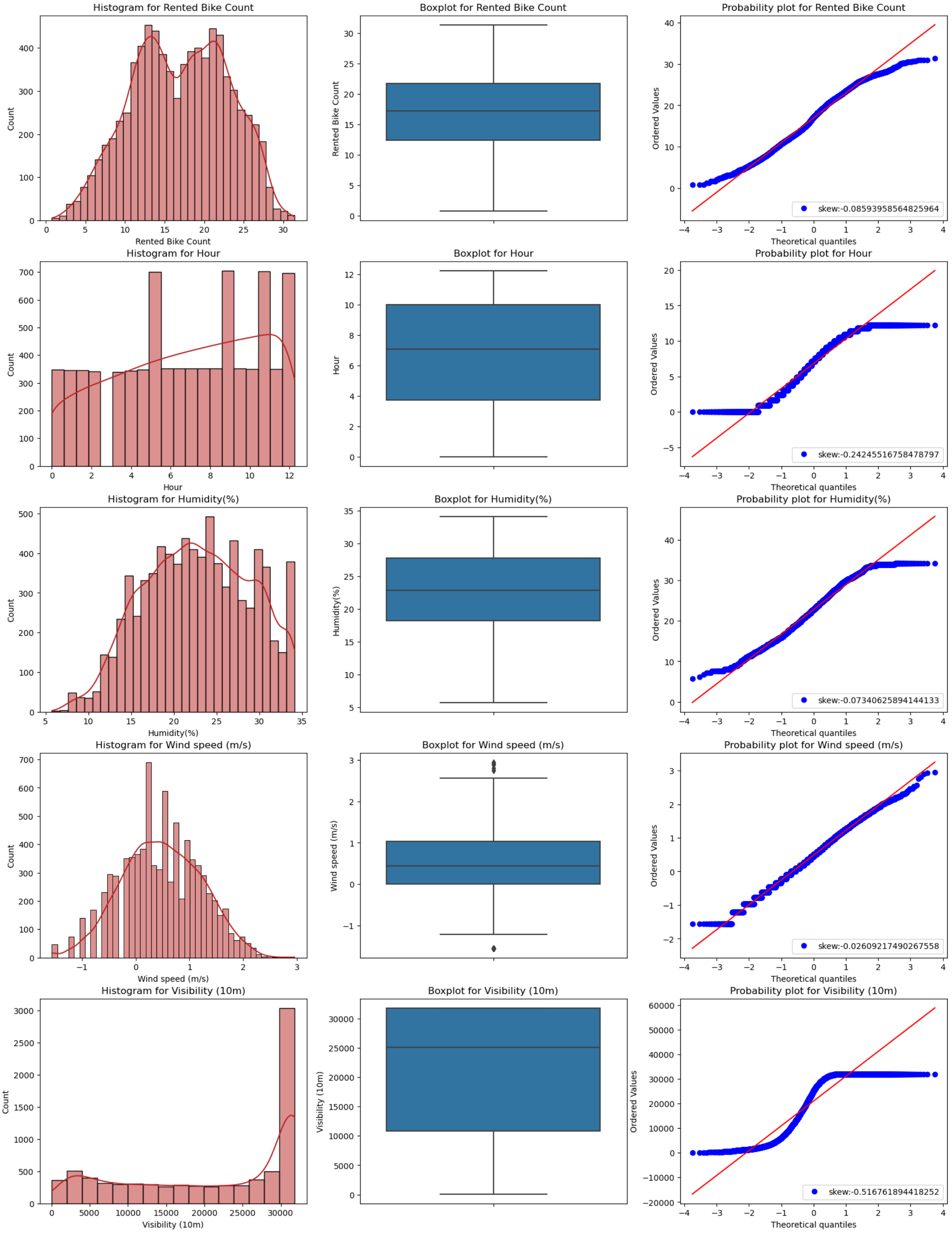
7. Histogram : (for Rented Bike Count)



Before Normalizing:

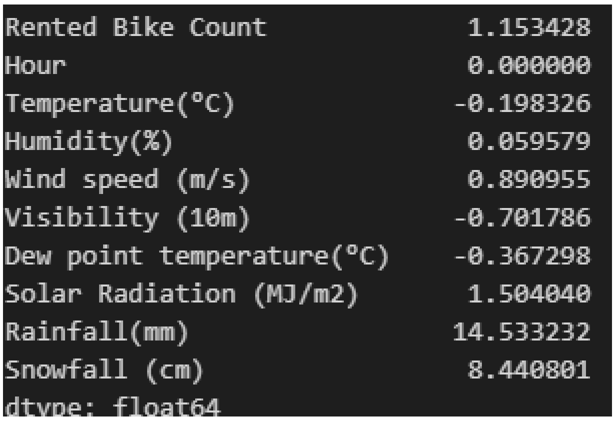


After Normalizing (skewness decreased):



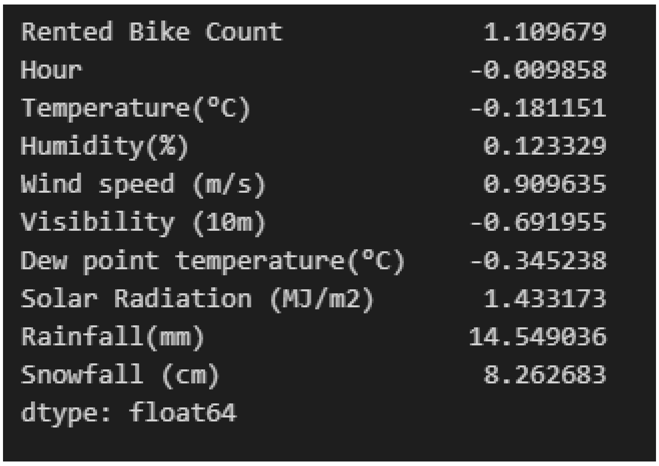
8. Checking the skewness of entire dataset:

   Before Normalising:



```
Rented Bike Count              1.153428
Hour                           0.000000
Temperature(°C)               -0.198326
Humidity(%)                    0.059579
Wind speed (m/s)               0.890955
Visibility (10m)              -0.701786
Dew point temperature(°C)     -0.367298
Solar Radiation (MJ/m2)        1.504040
Rainfall(mm)                  14.533232
Snowfall (cm)                  8.440801
dtype: float64
```

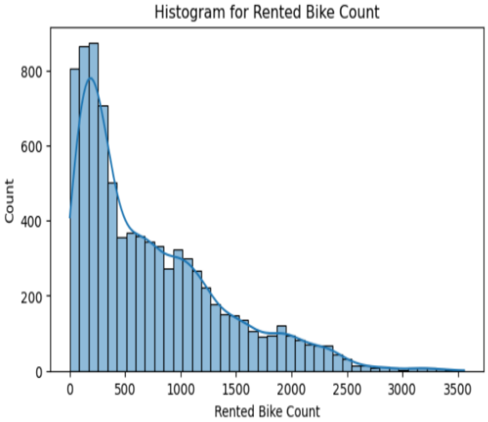
   After Normalising:

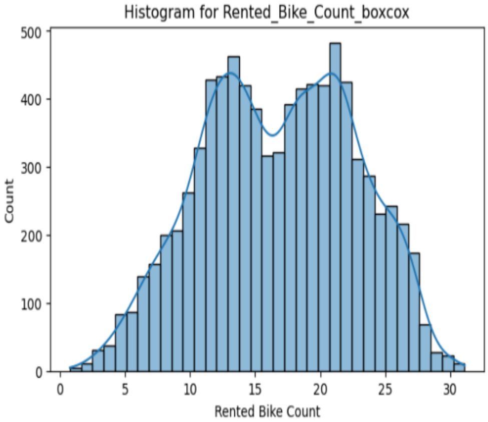


```
Rented Bike Count              1.109679
Hour                          -0.009858
Temperature(°C)               -0.181151
Humidity(%)                    0.123329
Wind speed (m/s)               0.909635
Visibility (10m)              -0.691955
Dew point temperature(°C)     -0.345238
Solar Radiation (MJ/m2)        1.433173
Rainfall(mm)                  14.549036
Snowfall (cm)                  8.262683
dtype: float64
```

9. From above as we can see that the skewness is more so, we'll normalise the dataset by using log, sqrt function.

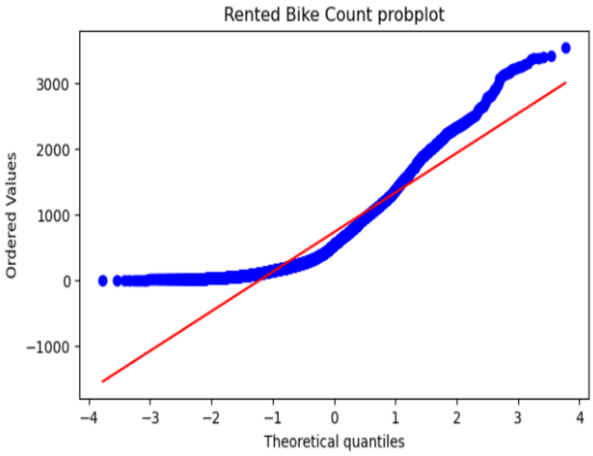
   After we normalise by using boxcox technique our target variable I e, Rented Bike Count :

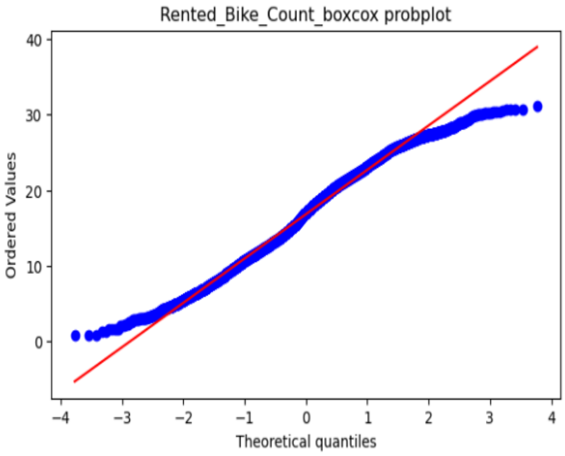




10. Probability plot:

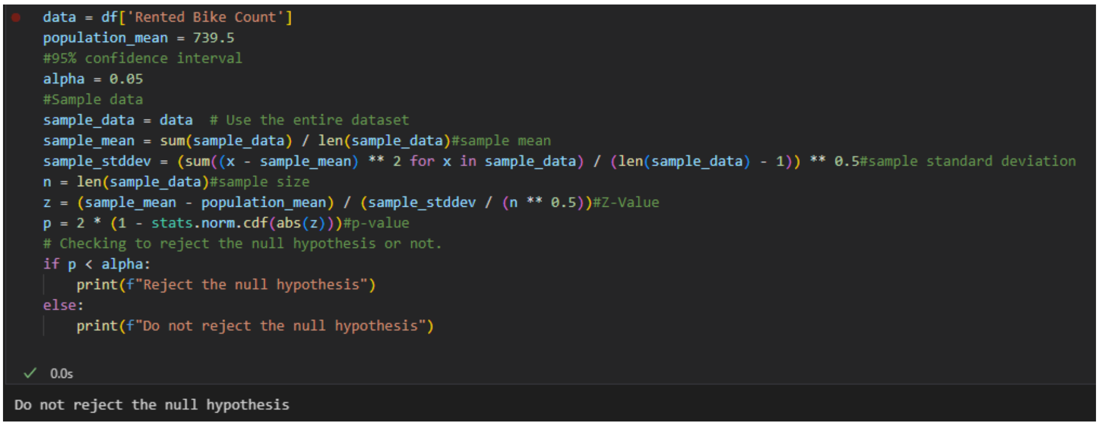
    a. We can clearly see that after using boxcox the datapoints came near to the mean.(The red line.)





11. Hypothesis Testing:

a. We are testing that if the mu is equal to 739.5 or not.
(Performing the testing on the Rented Bike Count column.)



```
data = df['Rented Bike Count']
population_mean = 739.5
#95% confidence interval
alpha = 0.05
#Sample data
sample_data = data  # Use the entire dataset
sample_mean = sum(sample_data) / len(sample_data)#sample mean
sample_stddev = (sum((x - sample_mean) ** 2 for x in sample_data) / (len(sample_data) - 1)) ** 0.5#sample standard deviation
n = len(sample_data)#sample size
z = (sample_mean - population_mean) / (sample_stddev / (n ** 0.5))#Z-Value
p = 2 * (1 - stats.norm.cdf(abs(z)))#p-value
# Checking to reject the null hypothesis or not.
if p < alpha:
    print(f"Reject the null hypothesis")
else:
    print(f"Do not reject the null hypothesis")
```

✓ 0.0s

```
Do not reject the null hypothesis
```

Our call is not to Reject the null Hypothesis.