Capstone Project - 2

Seoul Bike Sharing Demand Prediction

Team Members

Ajinkya Jumde
Neha Gupta
Shahfaissal I Dharwad

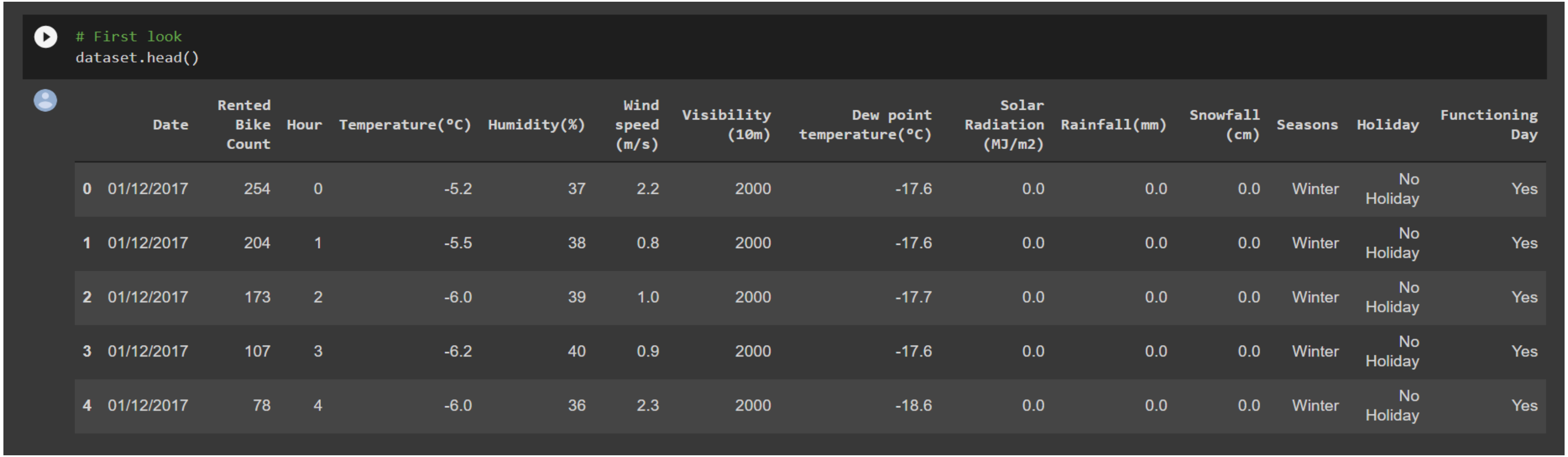
CONTENTS

- Business understanding
- Data summary
- Feature analysis
- Exploratory Data analysis
- Data pre-processing
- Implementing algorithms
- Conclusion

BUSINESS UNDERSTANDING

- Bike rentals have become a popular service in recent years and it seems people are using it more often. With relatively cheaper rates and ease of pick up and drop at own convince is what making the business thrive .
- Mostly used by people having no personal vehicles and also to avoid congested public transport which that’s why they prefer rental bikes
- Therefore the business to strive and profit more it has to be always ready and supply no of bikes at different locations, to fulfil the demand
- Our project goal is a pre planned set of bike count values that can be a handy solution to meet all demands.

DATA SUMMARY



```
# First look
dataset.head()
```

	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

- This dataset contains 8760 rows and 14 columns.
- Three categorical features ‘seasons’, ‘Holiday’ & functioning day.
- One datetime feature ‘Date’.
- We have some numerical type variables such as temperature, humidity,wind,visibility,dew point temperature, solar radiation , rainfall, snowfall which tells the environment conditions at that particular hour of the day.

ATTRIBUTE INFORMATION

- Date :year-month-day
- Rented Bike count - Count of bikes rented at each hour
- Hour - Hour of he day
- Temperature-Temperature in Celsius
- Humidity -%
- Windspeed -m/s
- Visibility -10m
- Dew point temperature - Celsius
- Solar radiation -MJ/m2
- Rainfall -mm
- Snowfall -cm
- Seasons - Winter, Spring, Summer, Autumn
- Holiday - Holiday/Noholiday
- Functional Day - NoFunc(Non Functional Hours), Fun(Functional hours)

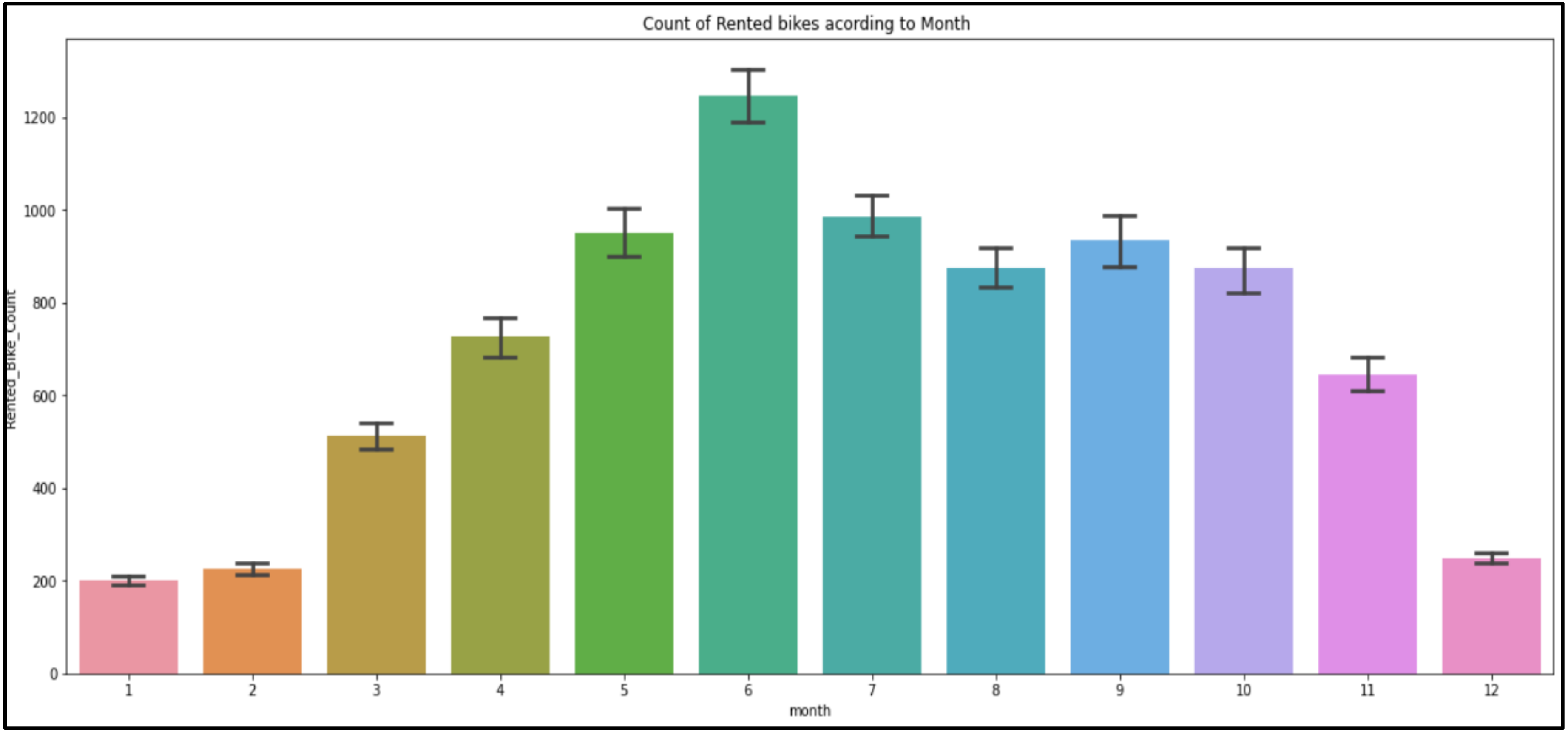
INSIGHTS FROM OUR DATASET

- There are no missing values present in the dataset
- There are no duplicate values present
- There are no null values
- And finally we have ‘Rented bike count’ variable which we need to predict for new observations.
- The dataset shows hourly rental data for one year (1st December 2017 to 31st November 2018(365 days).This is considered as a single year data.
- so we convert the “date” column into 3 different columns i.e “year”,”month”,”day”.
- For the convenience of use of dataset names where accordingly changed.

PROBLEM STATEMENT

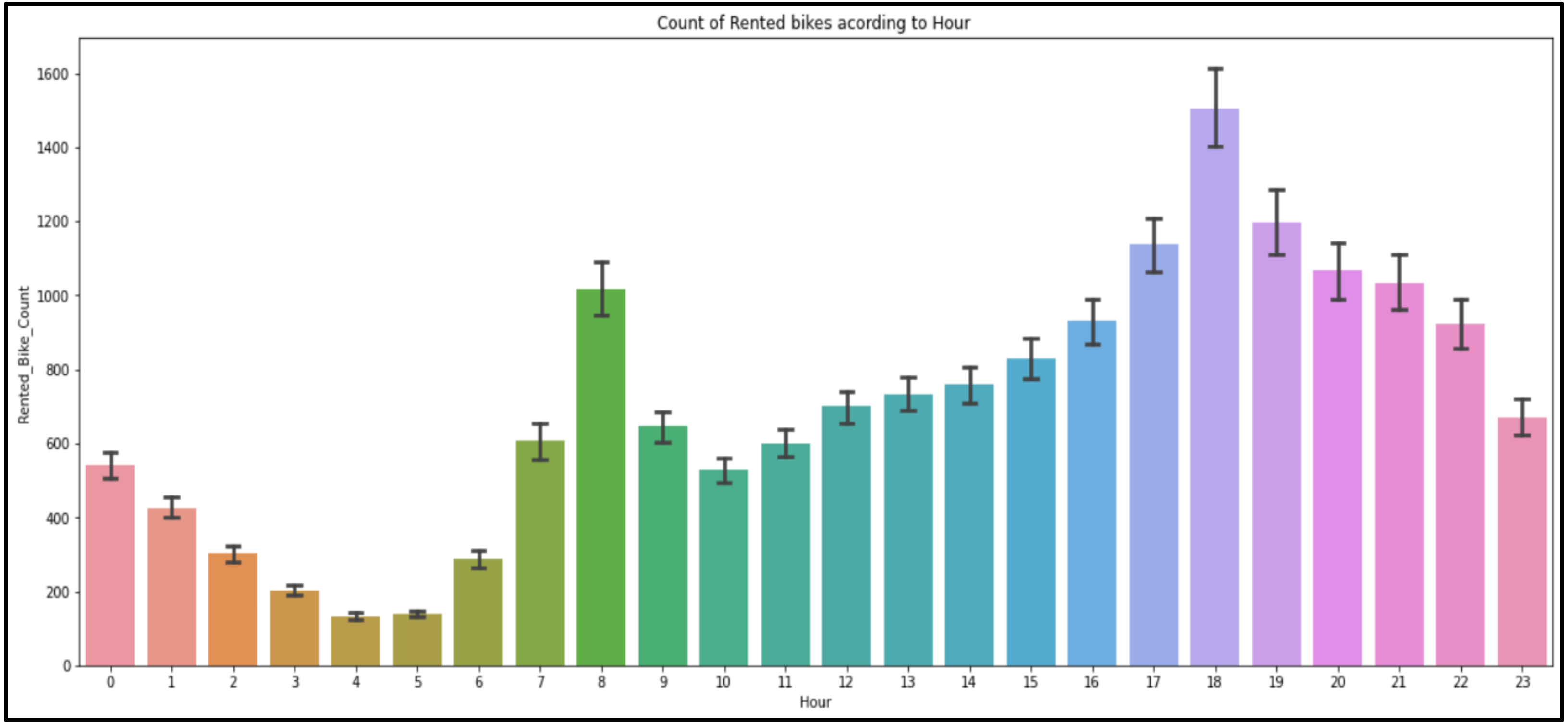
The project goal is to predict number of rental bikes required at a particular time of the day.

EXPLORATORY DATA ANALYSIS



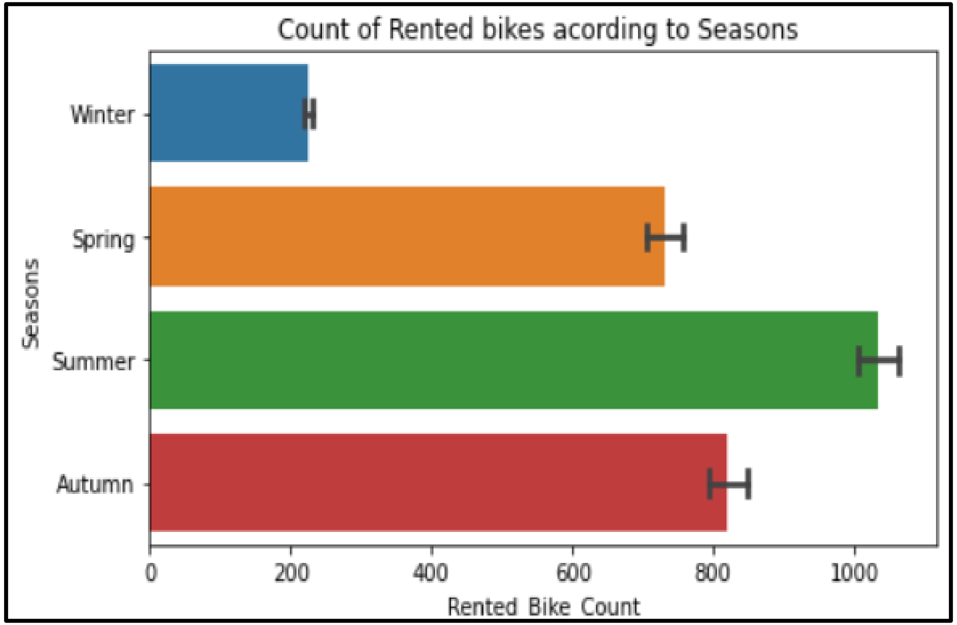
EXPLORATORY DATA ANALYSIS

Rented Bike counts across different hours in a day

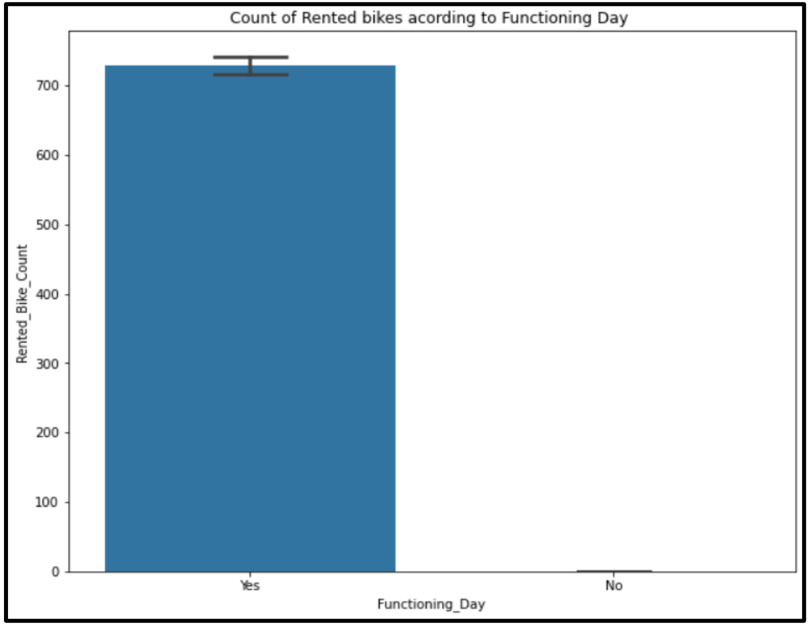


EXPLORATORY DATA ANALYSIS

Bike counts vs Seasons

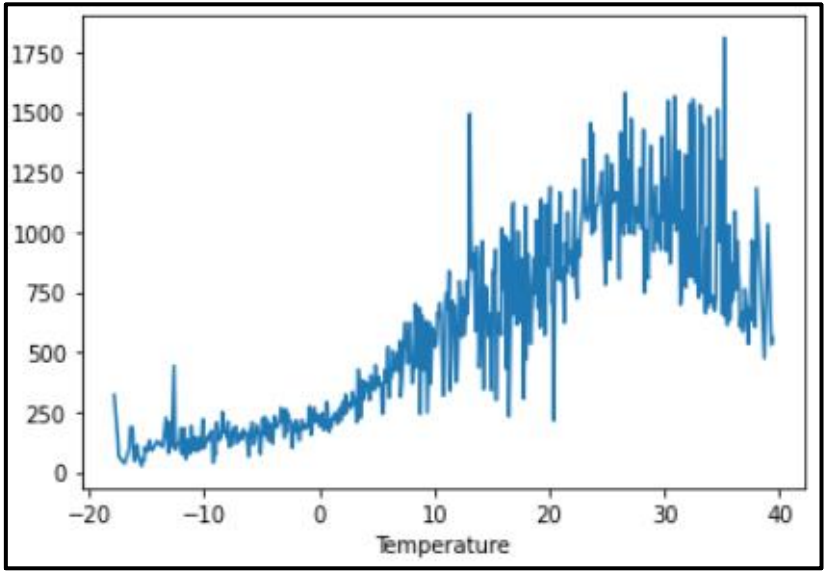


Bike counts vs Functionality day

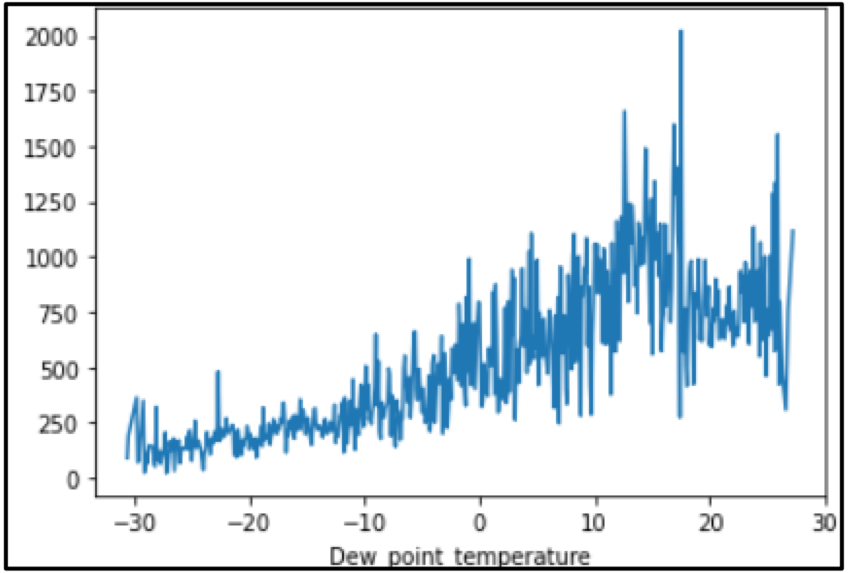


EXPLORATORY DATA ANALYSIS

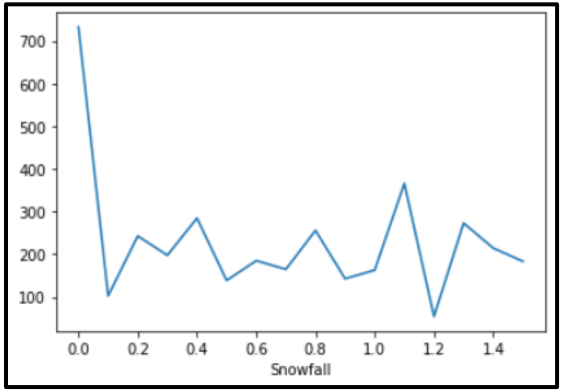
Bike counts vs Temperature

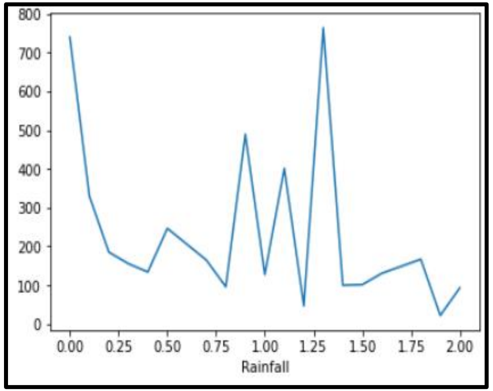


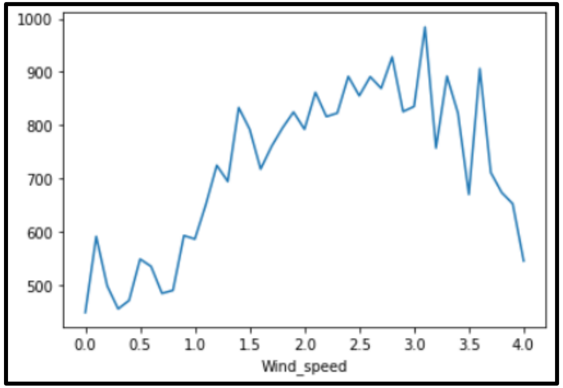
Bike counts vs Dew point temperature



EXPLORATORY DATA ANALYSIS

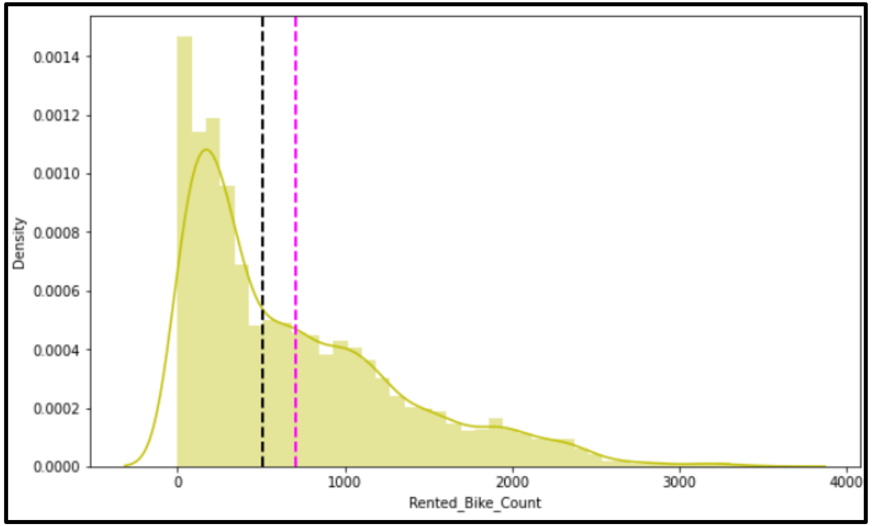


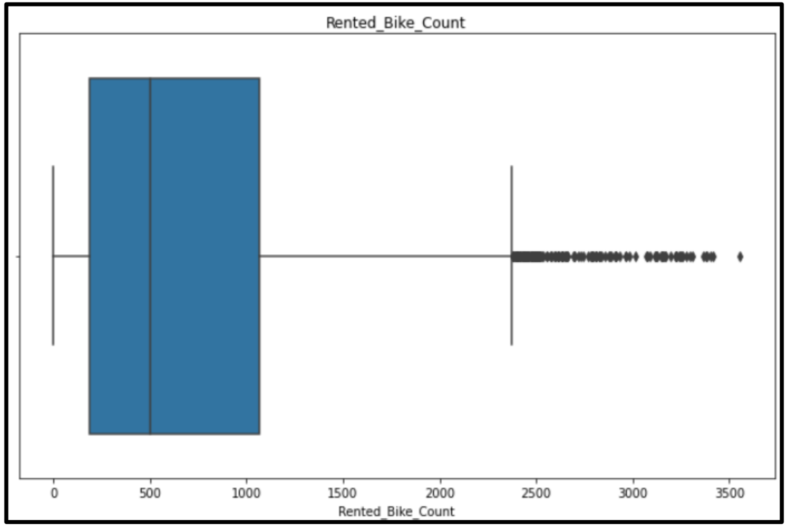




- In snowfall plot, on the y-axis, the amount of rented bike is very low when we have more than 4cm of snow, the bike rents is much lower.
- In rainfall plot, if it rains demand of rented bikes is not decreasing, here for example even if we have 20mm of rain there is a big peak of rented bikes.
- In wind speed plot that the demand of rented bike is uniformly distribute despite of wind speed but when the speed of wind was 7m/s then the demand f bike also increases which clearly tells that people love to ride bikes when its windy.

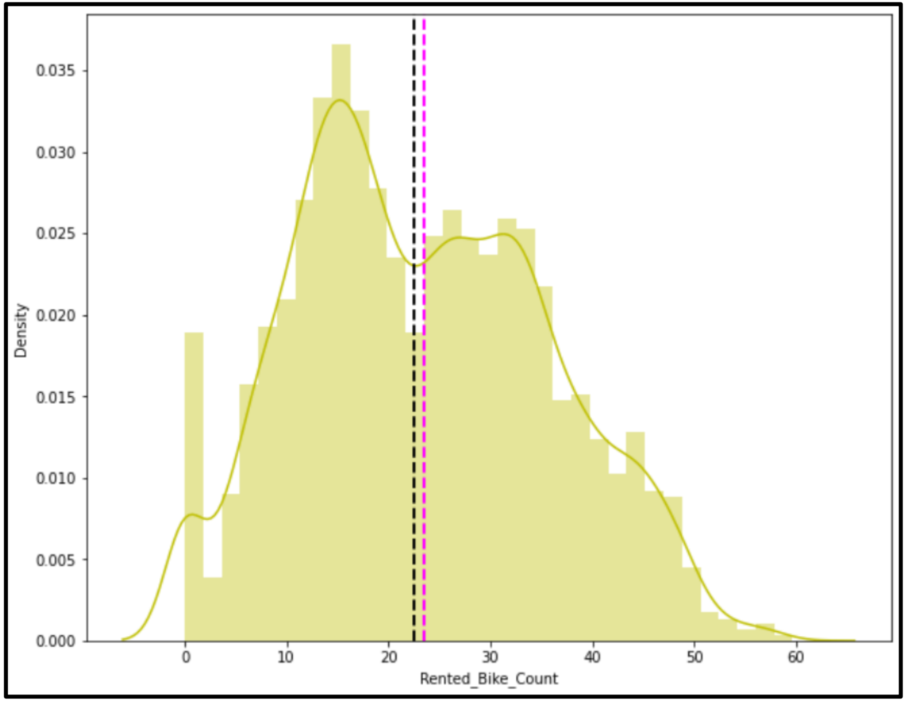
EXPLORATORY DATA ANALYSIS

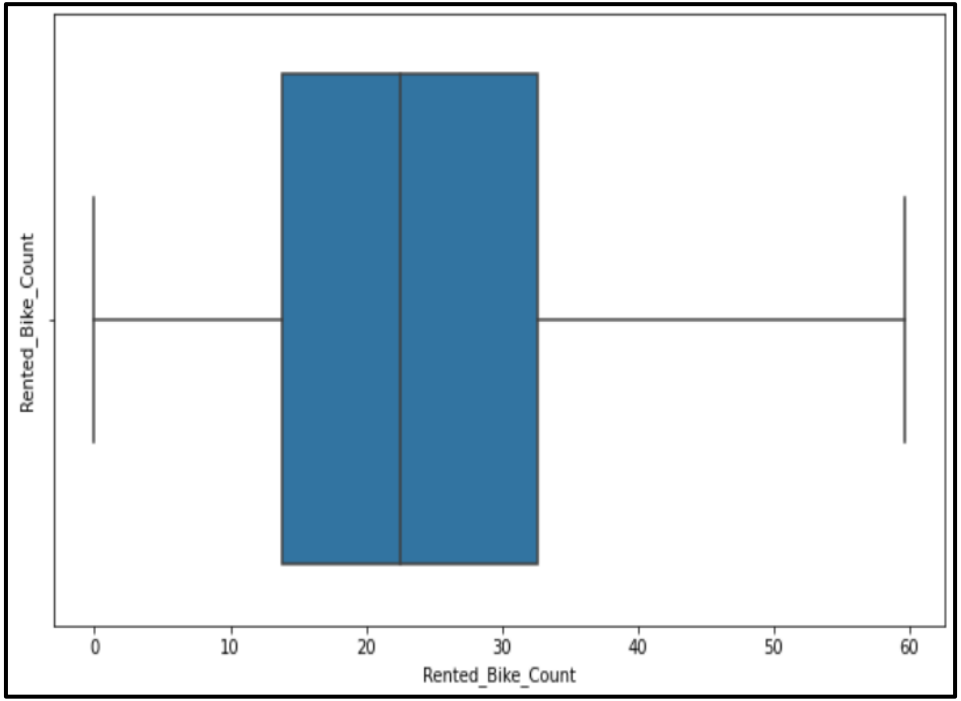




- The above graph shows that rented bike count has moderate right skewness.
- In the above boxplot outliers can be detected in rented bike count column.
- From the assumption of linear regression that the distribution of dependent variable has to be normal, so square root of the distribution has to be done to make it normal.

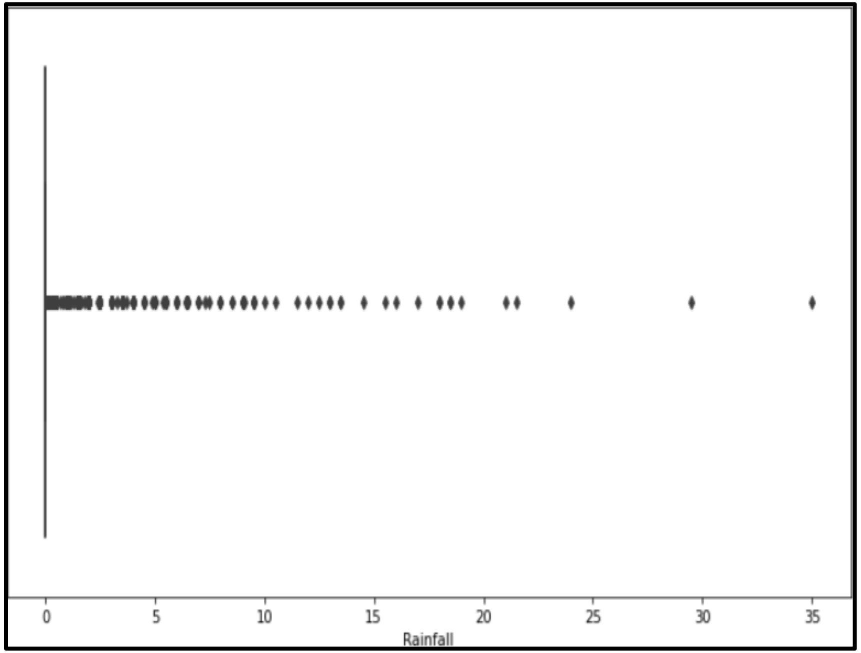
EXPLORATORY DATA ANALYSIS

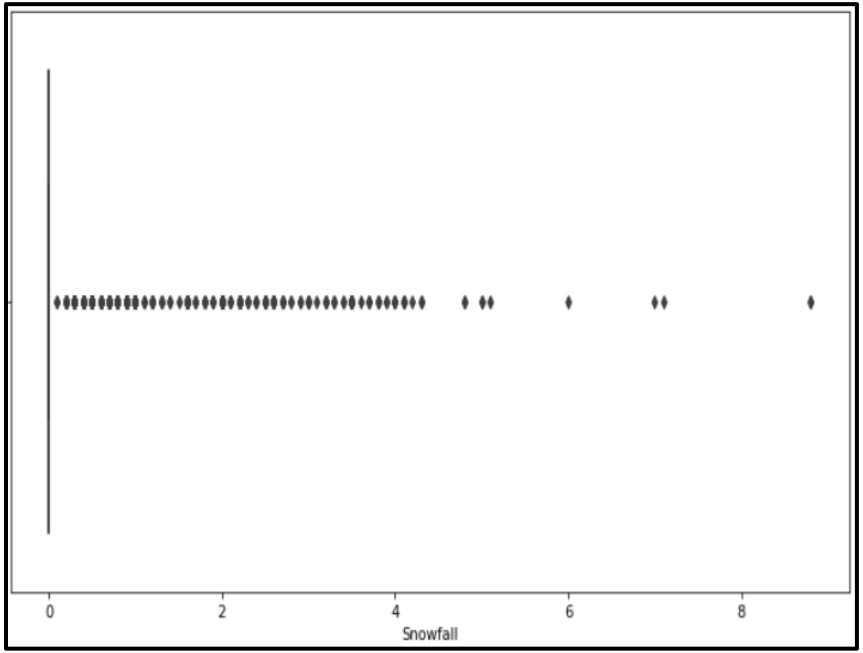


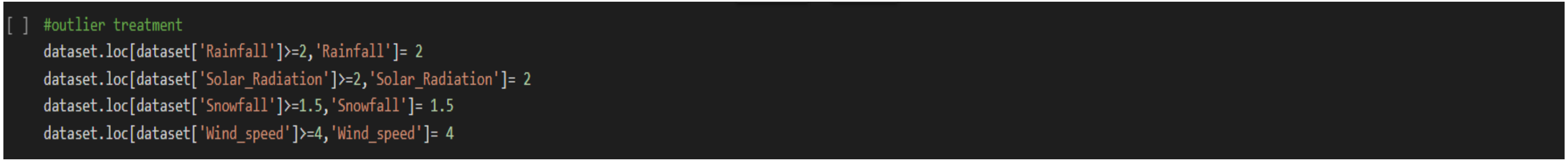


- The above graph shows that rented bike count after normalization
- The outliers have been treated and distribution is normal.

OUTLIERS DETECTION AND TREATMENT

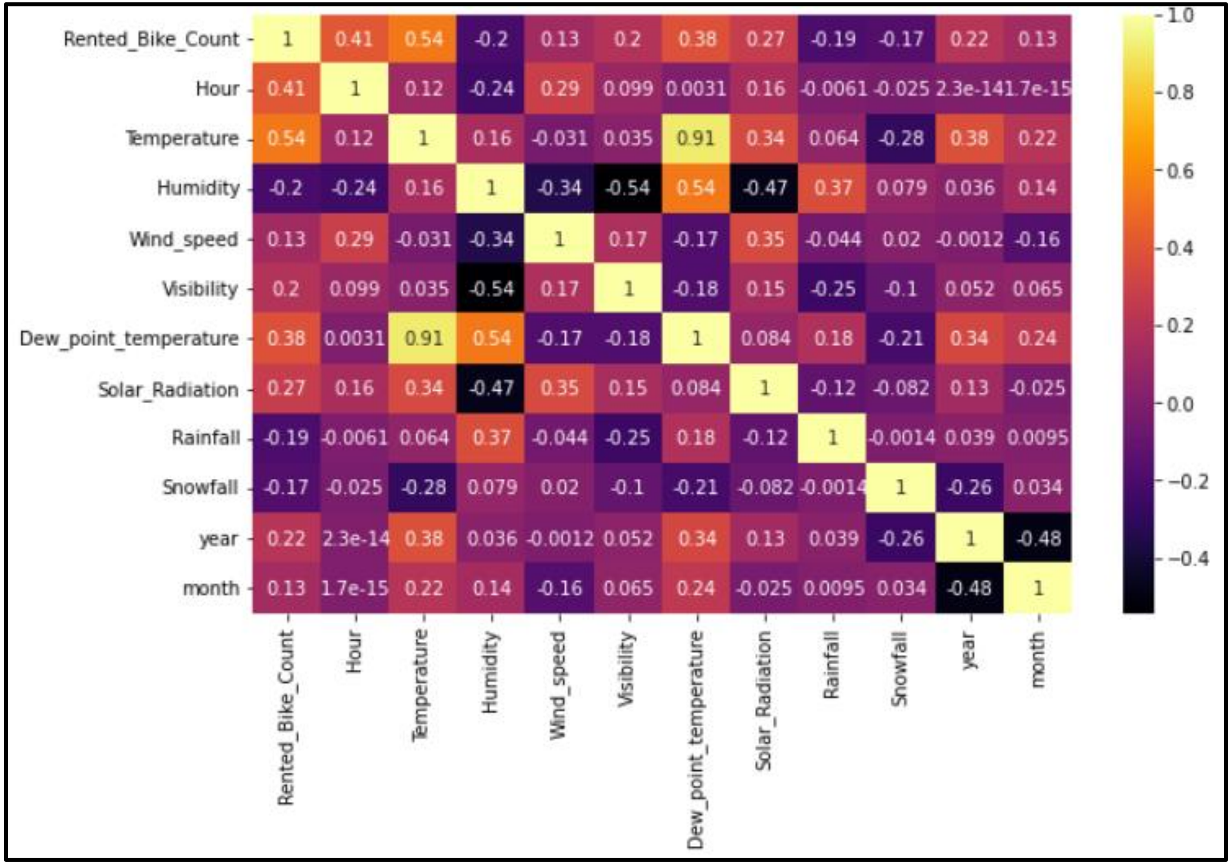






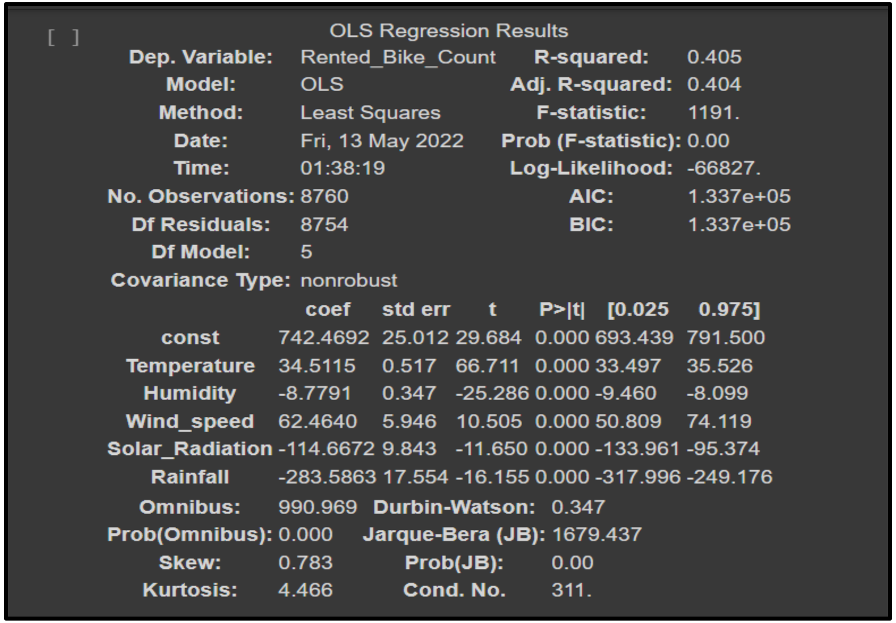
```
[ ]  #outlier treatment
     dataset.loc[dataset['Rainfall']>=2,'Rainfall']= 2
     dataset.loc[dataset['Solar_Radiation']>=2,'Solar_Radiation']= 2
     dataset.loc[dataset['Snowfall']>=1.5,'Snowfall']= 1.5
     dataset.loc[dataset['Wind_speed']>=4,'Wind_speed']= 4
```

CORRELATION HEATMAP

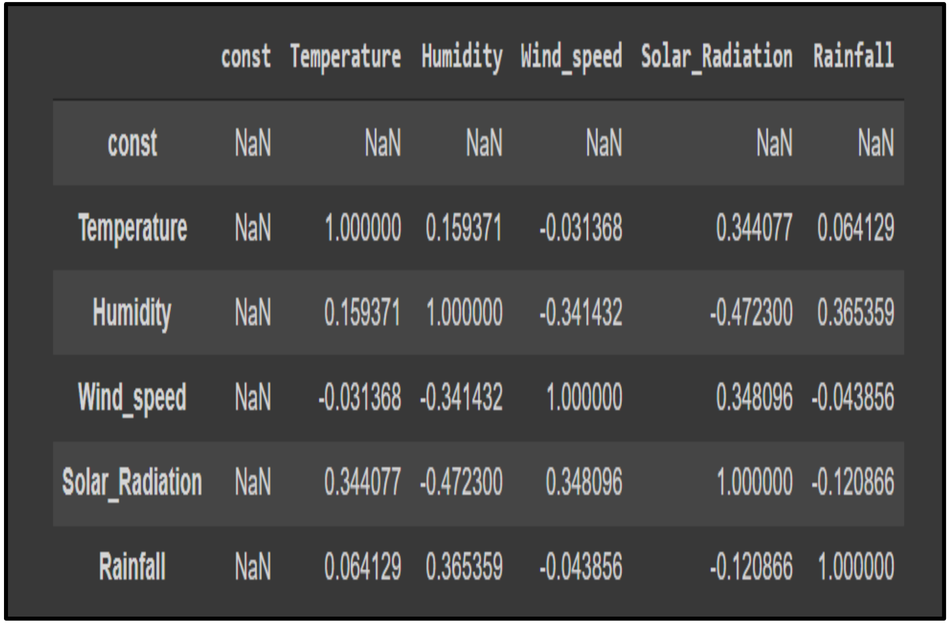


	Rented_Bike_Count	Hour	Temperature	Humidity	Wind_speed	Visibility	Dew_point_temperature	Solar_Radiation	Rainfall	Snowfall	year	month
Rented_Bike_Count	1	0.41	0.54	-0.2	0.13	0.2	0.38	0.27	-0.19	-0.17	0.22	0.13
Hour	0.41	1	0.12	-0.24	0.29	0.099	0.0031	0.16	-0.0061	-0.025	2.3e-14	1.7e-15
Temperature	0.54	0.12	1	0.16	-0.031	0.035	0.91	0.34	0.064	-0.28	0.38	0.22
Humidity	-0.2	-0.24	0.16	1	-0.34	-0.54	0.54	-0.47	0.37	0.079	0.036	0.14
Wind_speed	0.13	0.29	-0.031	-0.34	1	0.17	-0.17	0.35	-0.044	0.02	-0.0012	-0.16
Visibility	0.2	0.099	0.035	-0.54	0.17	1	-0.18	0.15	-0.25	-0.1	0.052	0.065
Dew_point_temperature	0.38	0.0031	0.91	0.54	-0.17	-0.18	1	0.084	0.18	-0.21	0.34	0.24
Solar_Radiation	0.27	0.16	0.34	-0.47	0.35	0.15	0.084	1	-0.12	-0.082	0.13	-0.025
Rainfall	-0.19	-0.0061	0.064	0.37	-0.044	-0.25	0.18	-0.12	1	-0.0014	0.039	0.0095
Snowfall	-0.17	-0.025	-0.28	0.079	0.02	-0.1	-0.21	-0.082	-0.0014	1	-0.26	0.034
year	0.22	2.3e-14	0.38	0.036	-0.0012	0.052	0.34	0.13	0.039	-0.26	1	-0.48
month	0.13	1.7e-15	0.22	0.14	-0.16	0.065	0.24	-0.025	0.0095	0.034	-0.48	1

OLS REGRESSION MODEL



```
[ ]                          OLS Regression Results
      Dep. Variable:   Rented_Bike_Count      R-squared:         0.405
          Model:       OLS                  Adj. R-squared:      0.404
         Method:       Least Squares           F-statistic:      1191.
          Date:        Fri, 13 May 2022    Prob (F-statistic):   0.00
          Time:        01:38:19              Log-Likelihood:     -66827.
    No. Observations:  8760                        AIC:          1.337e+05
      Df Residuals:    8754                        BIC:          1.337e+05
        Df Model:      5
    Covariance Type:   nonrobust
                       coef     std err   t      P>|t|  [0.025   0.975]
        const         742.4692  25.012  29.684   0.000  693.439  791.500
      Temperature     34.5115   0.517   66.711   0.000  33.497   35.526
        Humidity      -8.7791   0.347   -25.286  0.000  -9.460   -8.099
      Wind_speed      62.4640   5.946   10.505   0.000  50.809   74.119
    Solar_Radiation   -114.6672 9.843   -11.650  0.000  -133.961 -95.374
        Rainfall      -283.5863 17.554  -16.155  0.000  -317.996 -249.176
       Omnibus:       990.969   Durbin-Watson:    0.347
    Prob(Omnibus):    0.000    Jarque-Bera (JB):  1679.437
         Skew:        0.783        Prob(JB):      0.00
       Kurtosis:      4.466        Cond. No.      311.
```



	const	Temperature	Humidity	Wind_speed	Solar_Radiation	Rainfall
const	NaN	NaN	NaN	NaN	NaN	NaN
Temperature	NaN	1.000000	0.159371	-0.031368	0.344077	0.064129
Humidity	NaN	0.159371	1.000000	-0.341432	-0.472300	0.365359
Wind_speed	NaN	-0.031368	-0.341432	1.000000	0.348096	-0.043856
Solar_Radiation	NaN	0.344077	-0.472300	0.348096	1.000000	-0.120866
Rainfall	NaN	0.064129	0.365359	-0.043856	-0.120866	1.000000

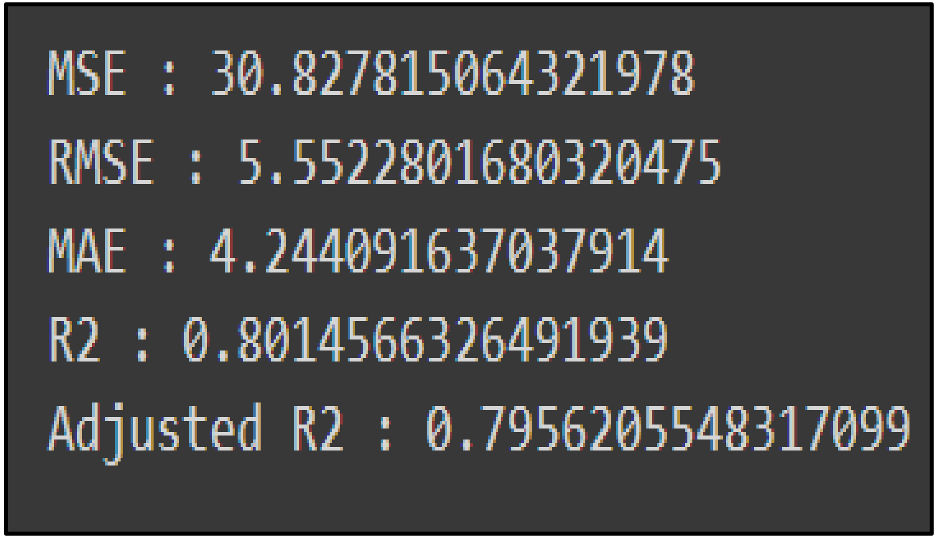
- R square and adjusted R-square are near to each other. 40% of variance in rented bike count is explained by the model

CONTINUED…..

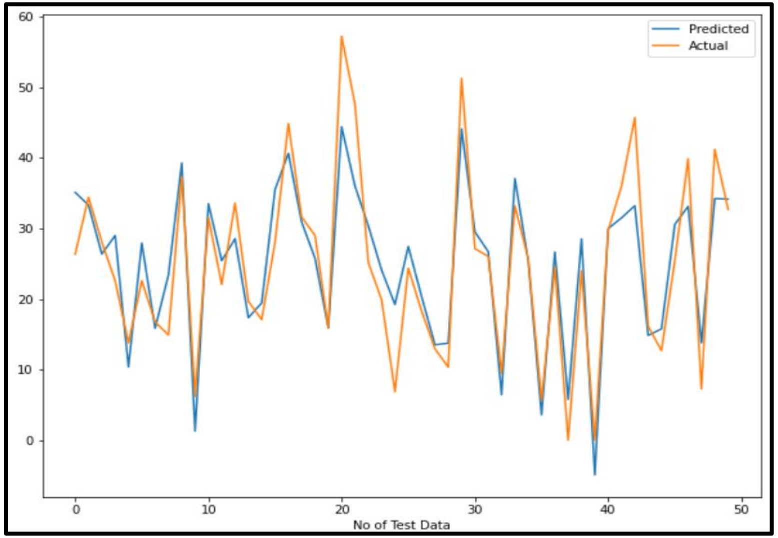
- According to the requirements of regression model many data features were dropped or modified.
- We assign the independent variables and dependent variable as a dataset to X and Y
- The model is fit for linear regression in terms of X_train and y_train
- The coefficients for each attribute of test and train dataset is computed.
- The regression model output is then verified in order to compute MSE, MAE, RMSE, R2 and Adjusted R2.
- R square and adjusted R-square are near to each other. 40% of variance in rented bike count is explained by the model

CONTINUED…..

- The plot between the predicted values and the actual values show that the model has predicted the rented bike count with high accuracy and with minimal residue.
- The values show that the model is able to capture 80% of the variance and so we might consider this model as an efficient model to compute the dependent variable.

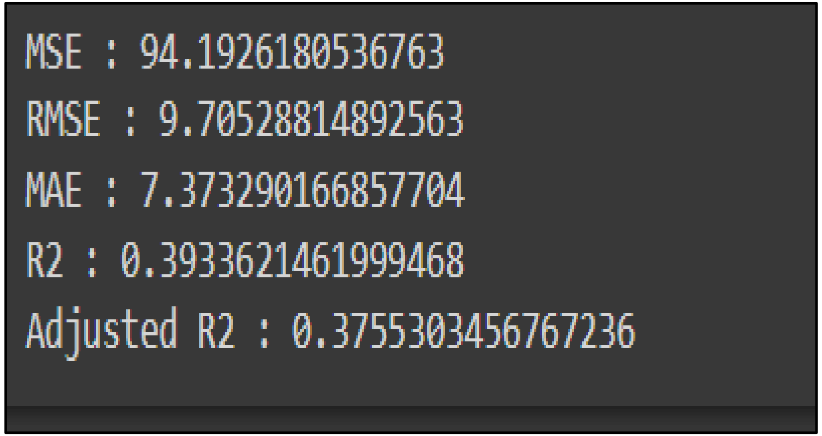


```
MSE : 30.827815064321978
RMSE : 5.5522801680320475
MAE : 4.244091637037914
R2 : 0.8014566326491939
Adjusted R2 : 0.7956205548317099
```

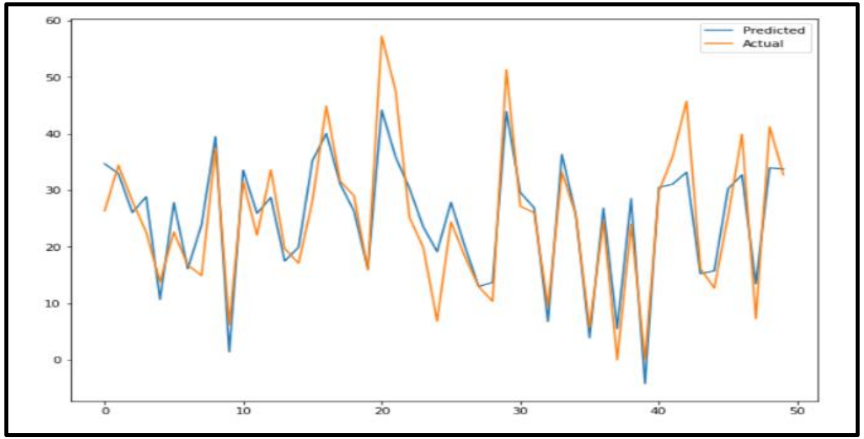


REGULARIZATION OF THE MODEL

- As the complexity of our model rises, variance becomes our primary concern.
- So regularization is done to shrink the size of coefficients by Ridge Regression and Lasso Regression.
- Here for our model lasso regression model to the same dataset and the values for the model were computed .
- The R2 value signifies that the Lasso model was also able to capture the variance of the data efficiently.

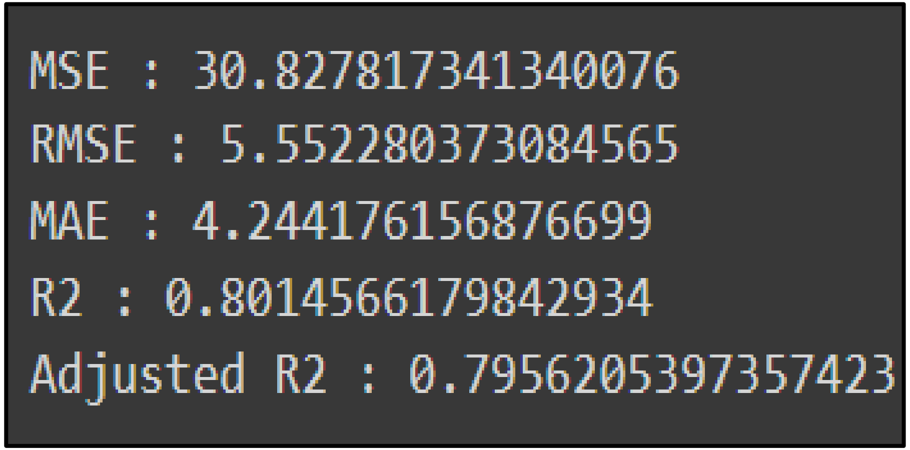


```
MSE : 94.1926180536763
RMSE : 9.70528814892563
MAE : 7.373290166857704
R2 : 0.3933621461999468
Adjusted R2 : 0.3755303456767236
```

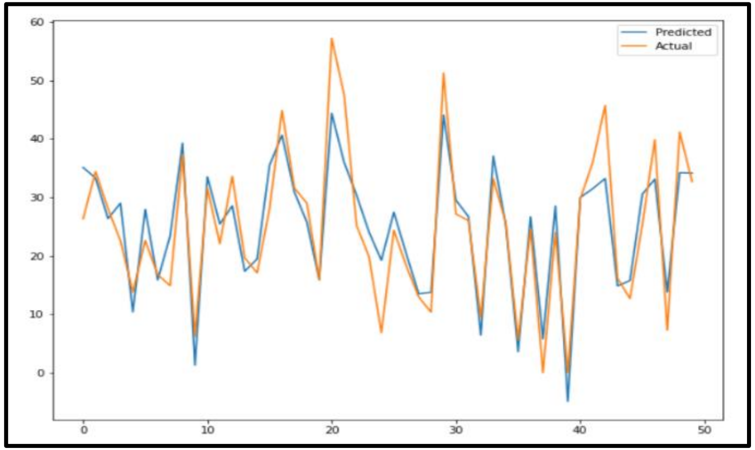


CONTINUED….

- In Ridge regression regularization was also applied for the dataset and values for the model were computed.
- The R2 value signifies that the Ridge model was also able to capture the variance of the data efficiently.
- So there is high accuracy and lower residue in the predicted data.



```
MSE : 30.827817341340076
RMSE : 5.552280373084565
MAE : 4.244176156876699
R2 : 0.8014566179842934
Adjusted R2 : 0.7956205397357423
```



CONCLUSION

- Hour of the day holds the most important feature.
- Bike rental count is mostly correlated with the time of the day as it is peak at 10 am morning and 8pm evening
- We observed that bike rental count is high during working days than non working day.
- We see that people generally prefer to rent bike at moderate to high temperatures, and when little windy.
- It is observed that highest number bike rental counts are in autumn and summer seasons and the lowest in winter season, we observed that the highest number of bike rentals on a clear day and the lowest on as snowy day or rainy day. We observed that it is increasing humidity, the number of bike rental counts decreases
- The predicted values were highly accurate and so the companies can rely on the model in order to predict the demand for the bikes on particular date.

Thank You