CAPSTONE PROJECT : 2

SEOUL BIKE SHARING DEMAND PREDICTION

BY

ABHAY DEEP SINGH

PROBLEM DESCRIPTION:

Currently Rental bikes are introduced in many urban cities for the enhancement of mobility comfort. It is important to make the rental bike available and accessible to the public at the right time as it lessens the waiting time. Eventually, providing the city with a stable supply of rental bikes becomes a major concern. The crucial part is the prediction of bike count required at each hour for the stable supply of rental bikes.

BUSINESS UNDERSTANDING

- Bike rentals have became a popular service in recent years and its seems pepole are using it more often. With relatively cheaper rates and ease of pickup and drop at own convenience is what making this business thrive.

- Therefore the business to strive and profit more it has to be always ready and supply no.of bikes at different location to fullfill the demand.

- Bicycle system provides user to rent a bike from one docking station, where user can ride and then return in another docking station.

- Our project goal is pre planned set of bike count values that can be a handly solution to meet all demands.

DATA SUMMARY

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
01/12/2017	254	0	-5.2	37	2.2	2000	-17.6	0.0	0.0	0.0	Winter	No Holiday	Yes
01/12/2017	204	1	-5.5	38	0.8	2000	-17.6	0.0	0.0	0.0	Winter	No Holiday	Yes
01/12/2017	173	2	-6.0	39	1.0	2000	-17.7	0.0	0.0	0.0	Winter	No Holiday	Yes
01/12/2017	107	3	-6.2	40	0.9	2000	-17.6	0.0	0.0	0.0	Winter	No Holiday	Yes
01/12/2017	78	4	-6.0	36	2.3	2000	-18.6	0.0	0.0	0.0	Winter	No Holiday	Yes

DATA SUMMARY

- This data set contains 8760 rows and 14 columns.
- There are 3 categorical feaure in this data set ‘Seasons’,’Holiday’,’Functioning day’.
- One datetime column ‘Date’.
- There are no missing value.
- There are no dupicate value.
- There are no null value.
- The dependent variable is ‘Rented Bike Count’ which we need to make prediction.
- The dataset shows hourly rental data for one year (1 dec 2017 to 31 nov 2018) .

FEATURE TYPES

- **NUMERIC FEATURE**

1. Hour
2. Temperature
3. Humidity
4. Wind
5. Dew point temperature
6. Solar radition
7. Rainfall
8. Snowfall

- **CATEGORICAL FEATURE**

1. Season
2. Holiday
3. Functioning day
4. Date time

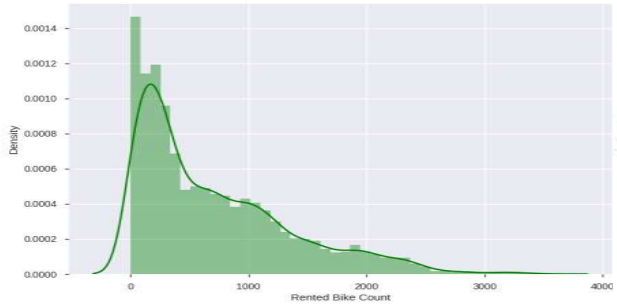
- **TARGET VARIABLE**

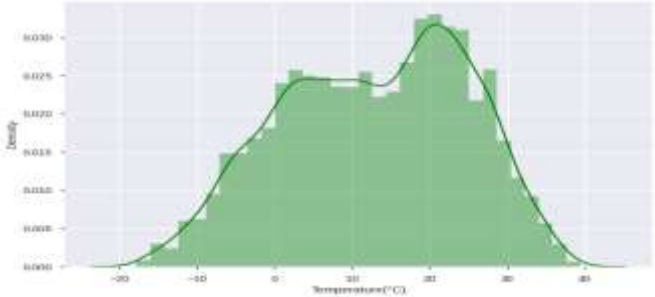
BIKE COUNT

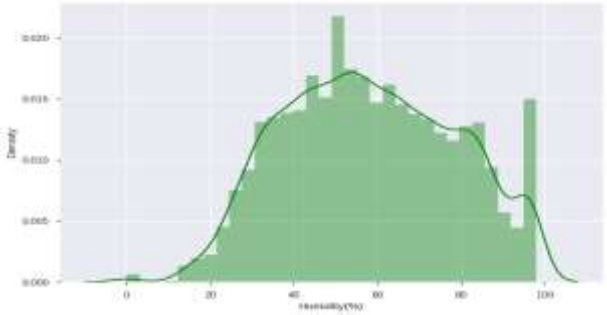
FEATURE SUMMARY

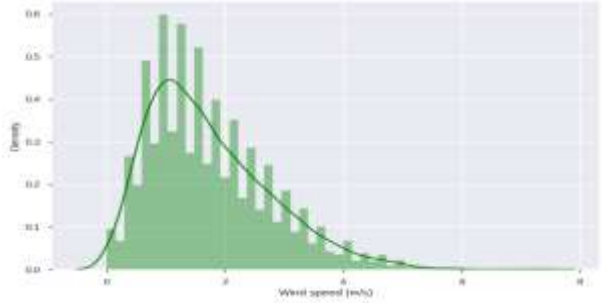
- **Date :** year-month-day
- **Hour -** Hour of he day
- **Temperature-** Temperature in Celsius
- **Humidity -** %
- **Windspeed -** m/s
- **Visibility -** 10m
- **Rented Bike count -** Count of bikes rented at each hour
- **Dew point temperature -** Celsius
- **Solar radiation -** MJ/m2
- **Rainfall -** mm
- **Snowfall -** cm
- **Seasons -** Winter, Spring, Summer, Autumn
- **Holiday -** Holiday/No holiday
- **Functional Day -** NoFunction(Non Functional Hours)

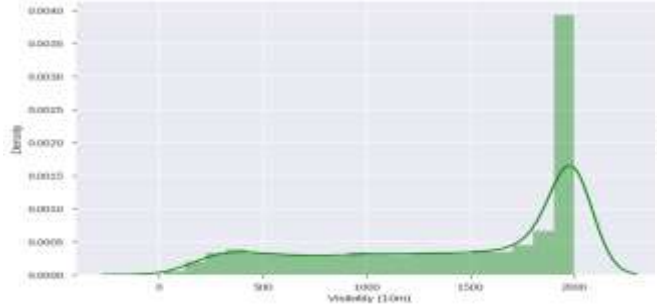
VISUALIZING DISTRIBUTIONS

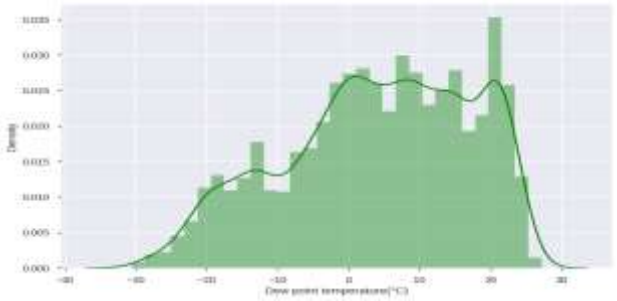


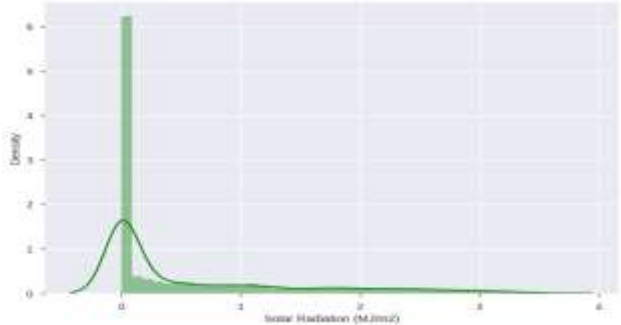


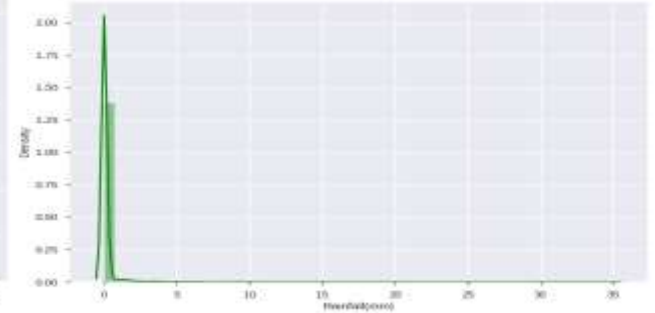


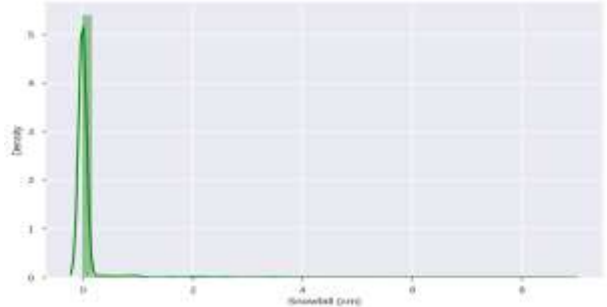




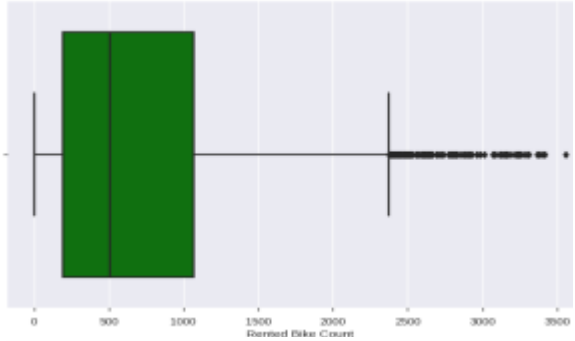


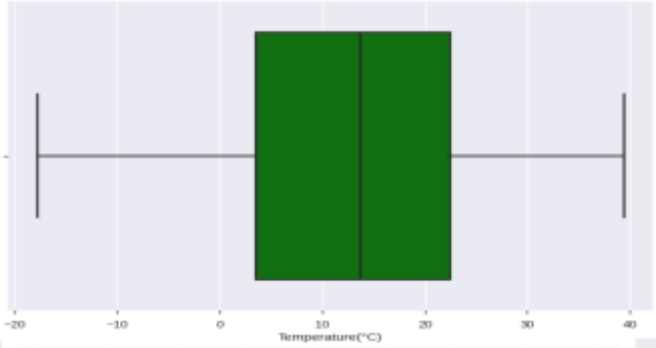


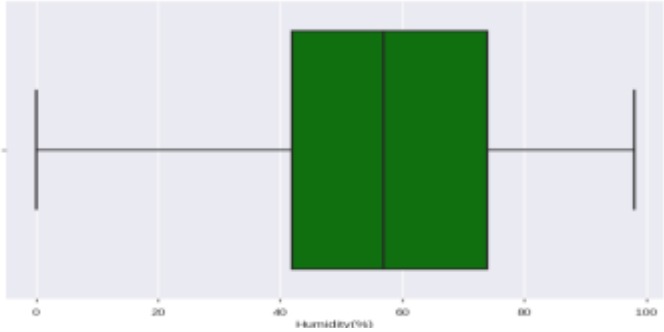


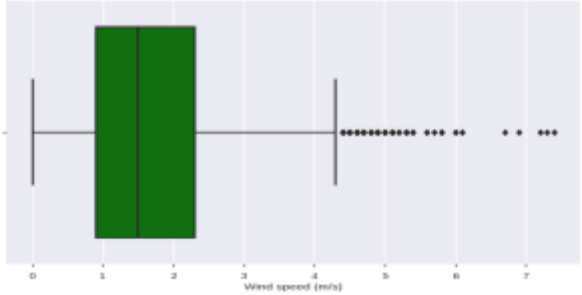


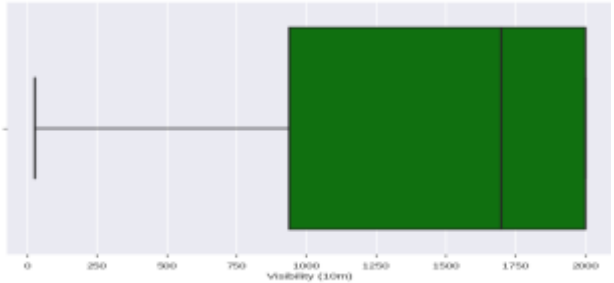
CHECKING OUTLIERS

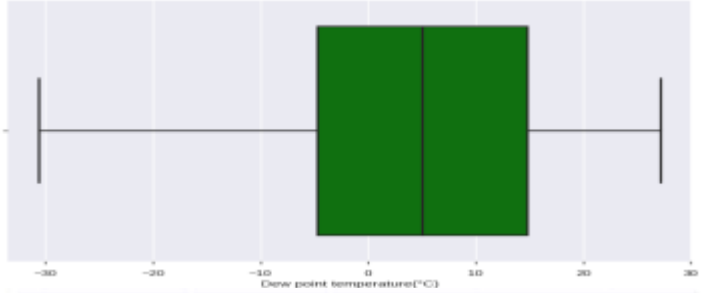


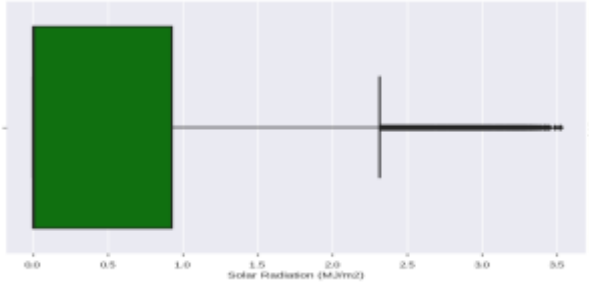


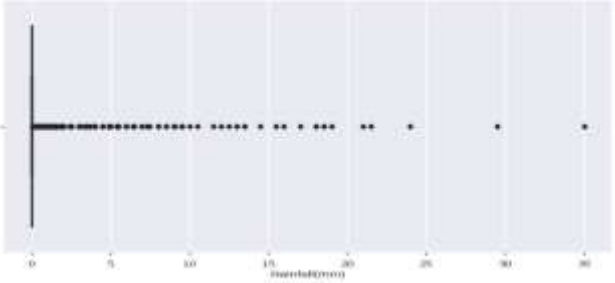


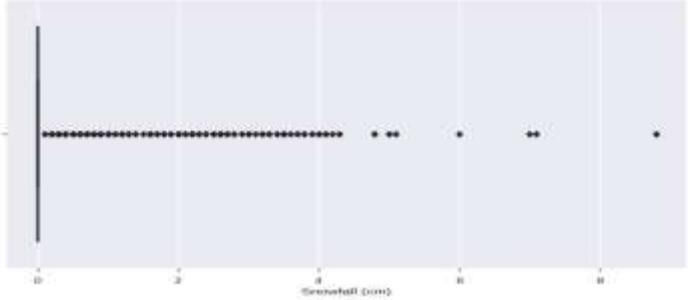










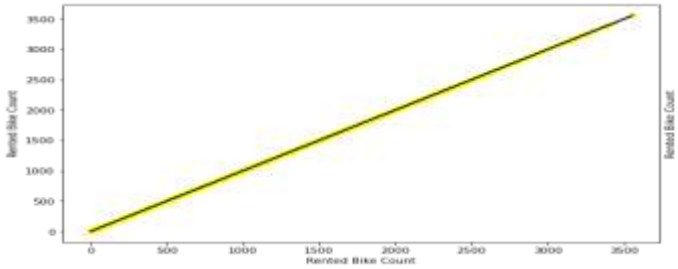


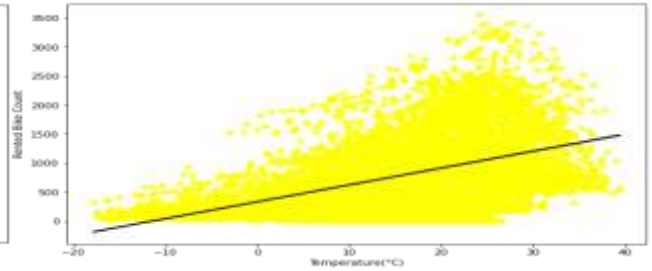
CHECKING OUTLIERS

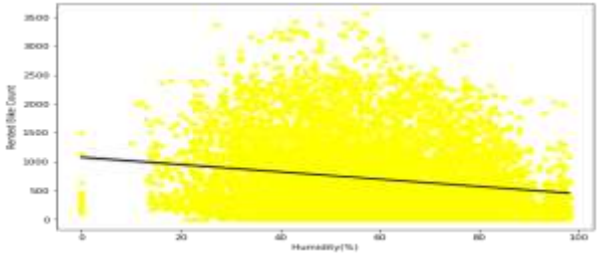
- We see outliers in some columns like Solar radition, Wind speed, Rainfall, and snowfall but lets not treat them because they may not be outliers as snowfall,rainfall.

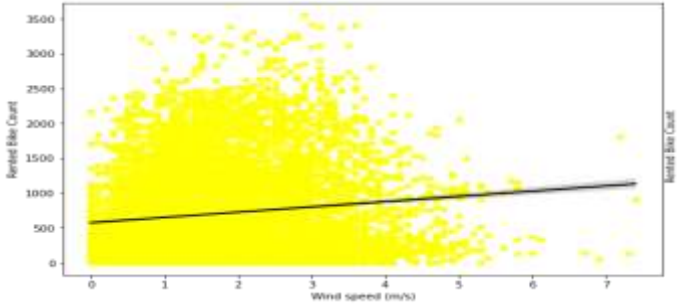
- We treated outliers in the target variable by capping with interquartile range limits.

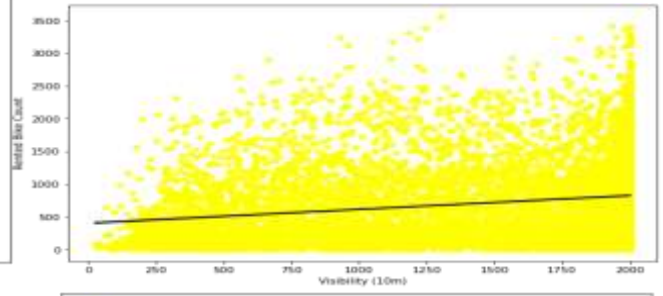
CHECKING LINEARITY IN DATA

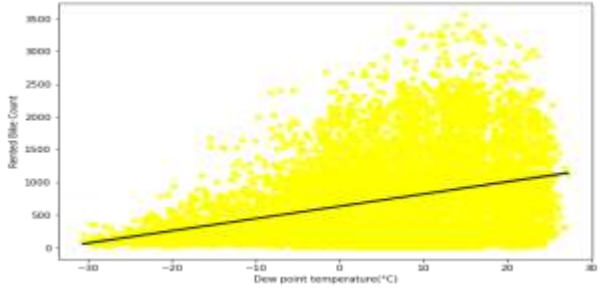


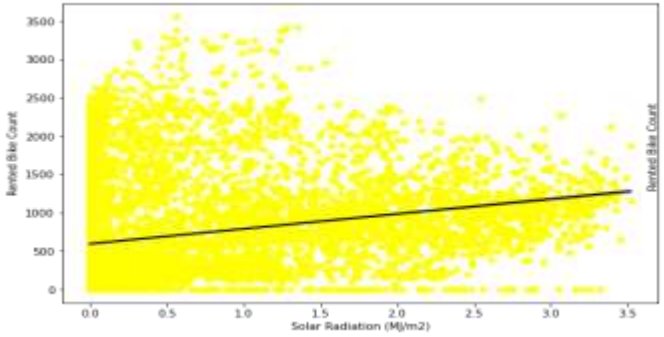


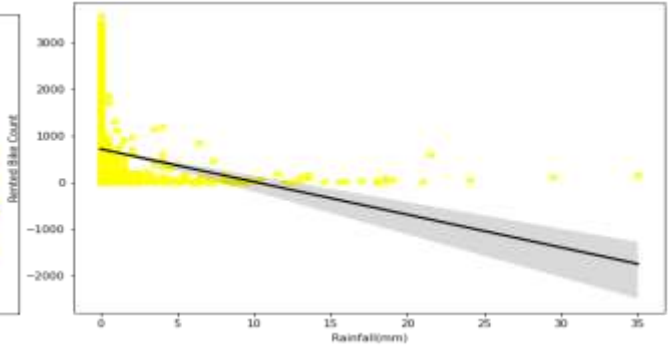


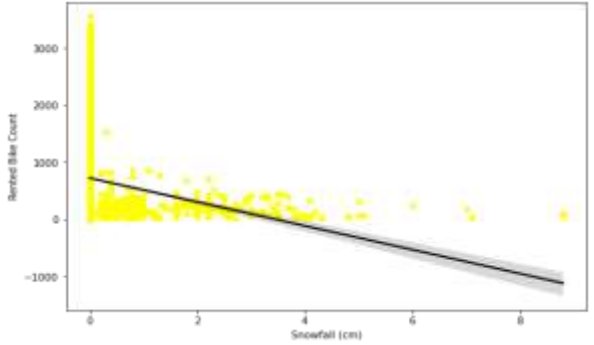












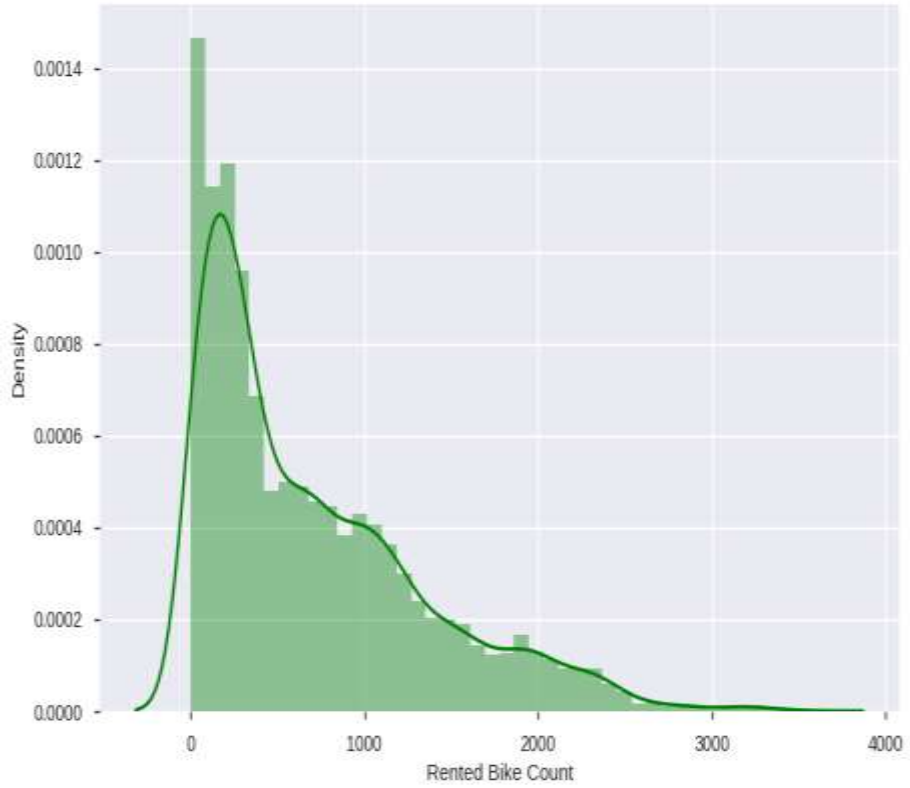
CHECKING LINEARITY IN DATA

- From regression plot we can see that features like temperature, wind speed, visibility, dew point temperature, solar radiation, are positively related to target variable.

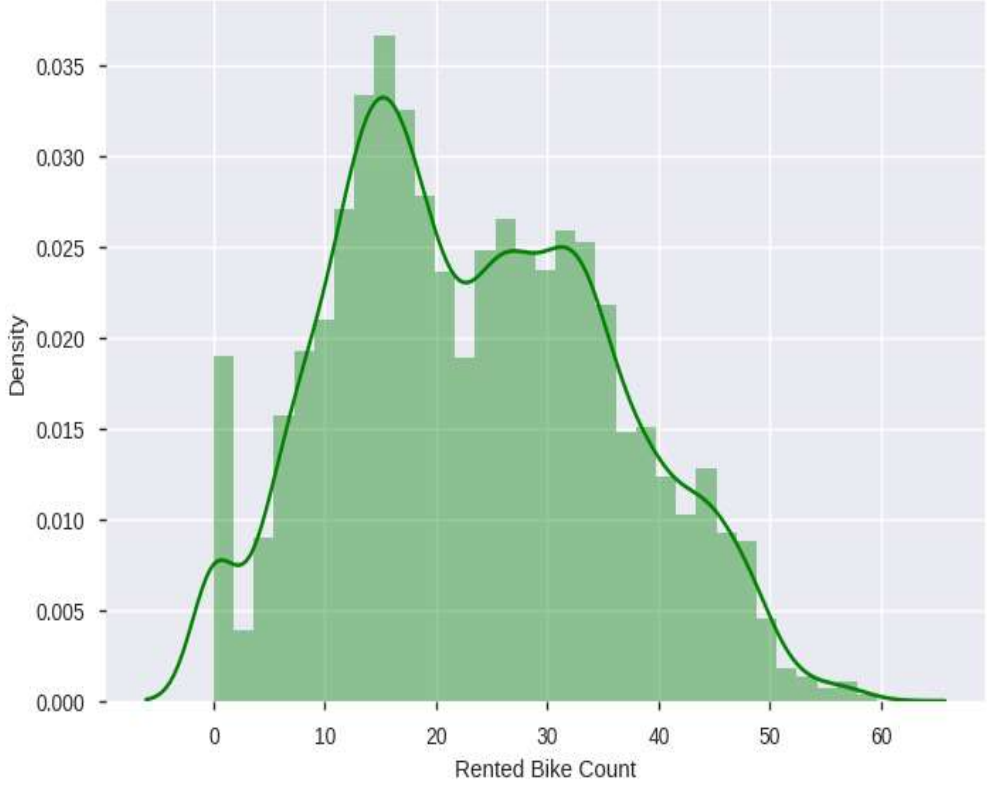
- Rainfall, snowfall and humidity are negatively related to target variable, rented bike count decreases with increase in values of these features.

Normalization of Target Variable

Bike count before transformation

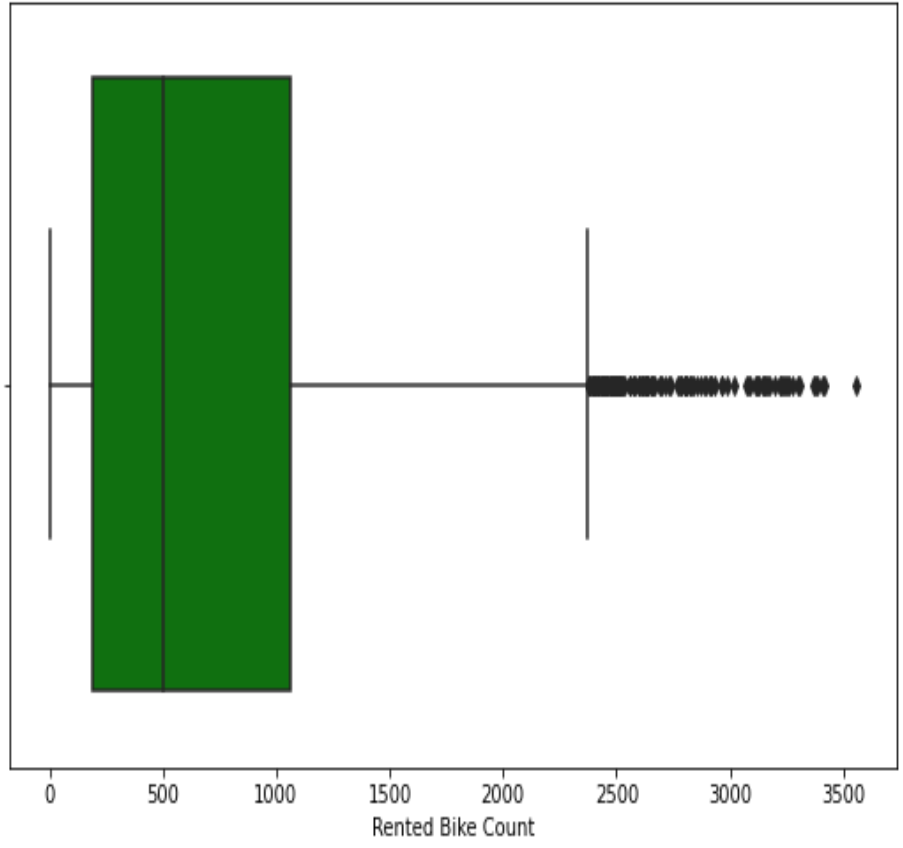


Bike count after transformation

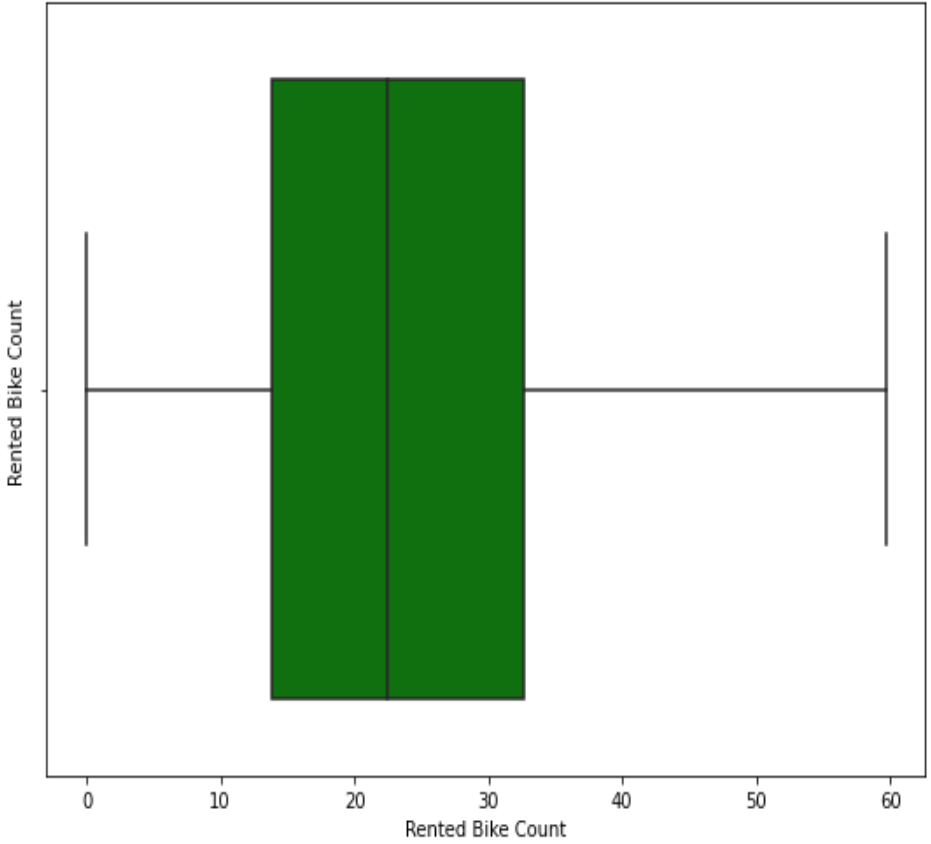


Normalization of Target Variable

Bike count before transformation



Bike count after transformation



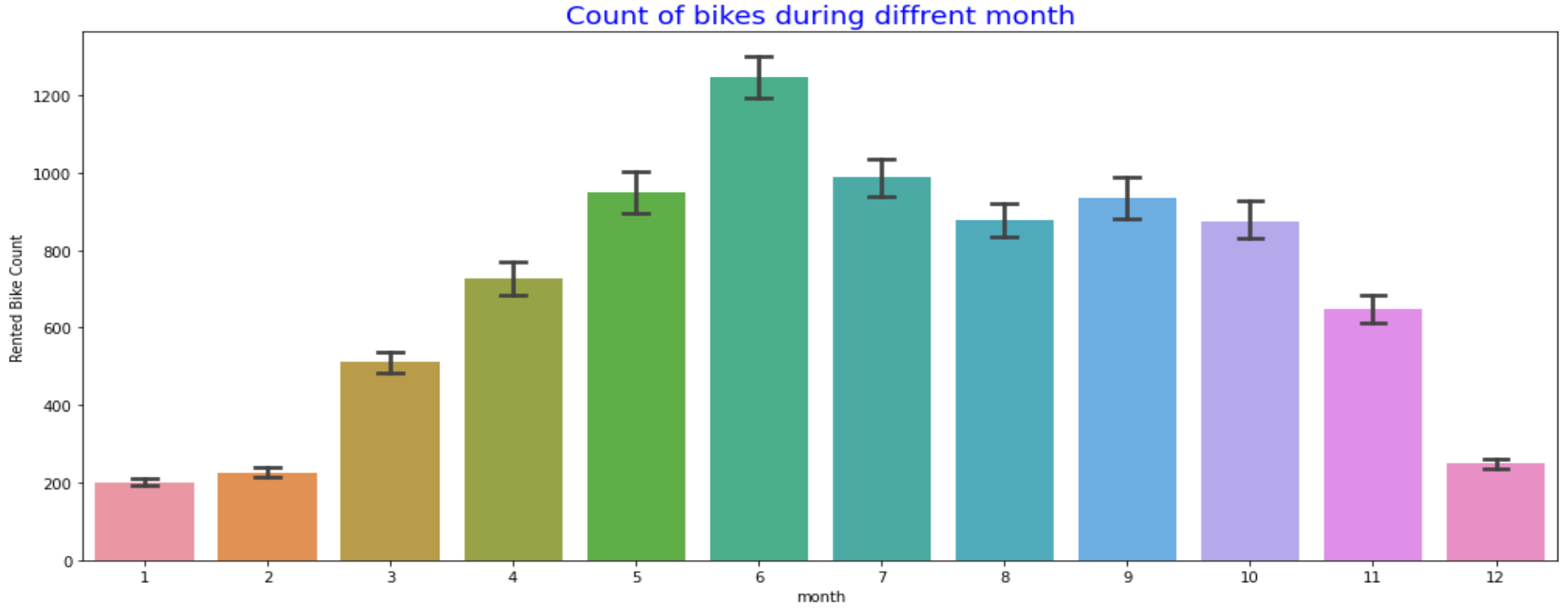
Normalization of Target Variable

- Earlier the distribution of the target variable was postively skwed with a skewness value 0.985. we tried to make this distribution some what close to normal distribution.

- We applied squre root transformation we got the best result, the skewness value was dropped to 0.153, which is comparatively closer to the normal distribution.

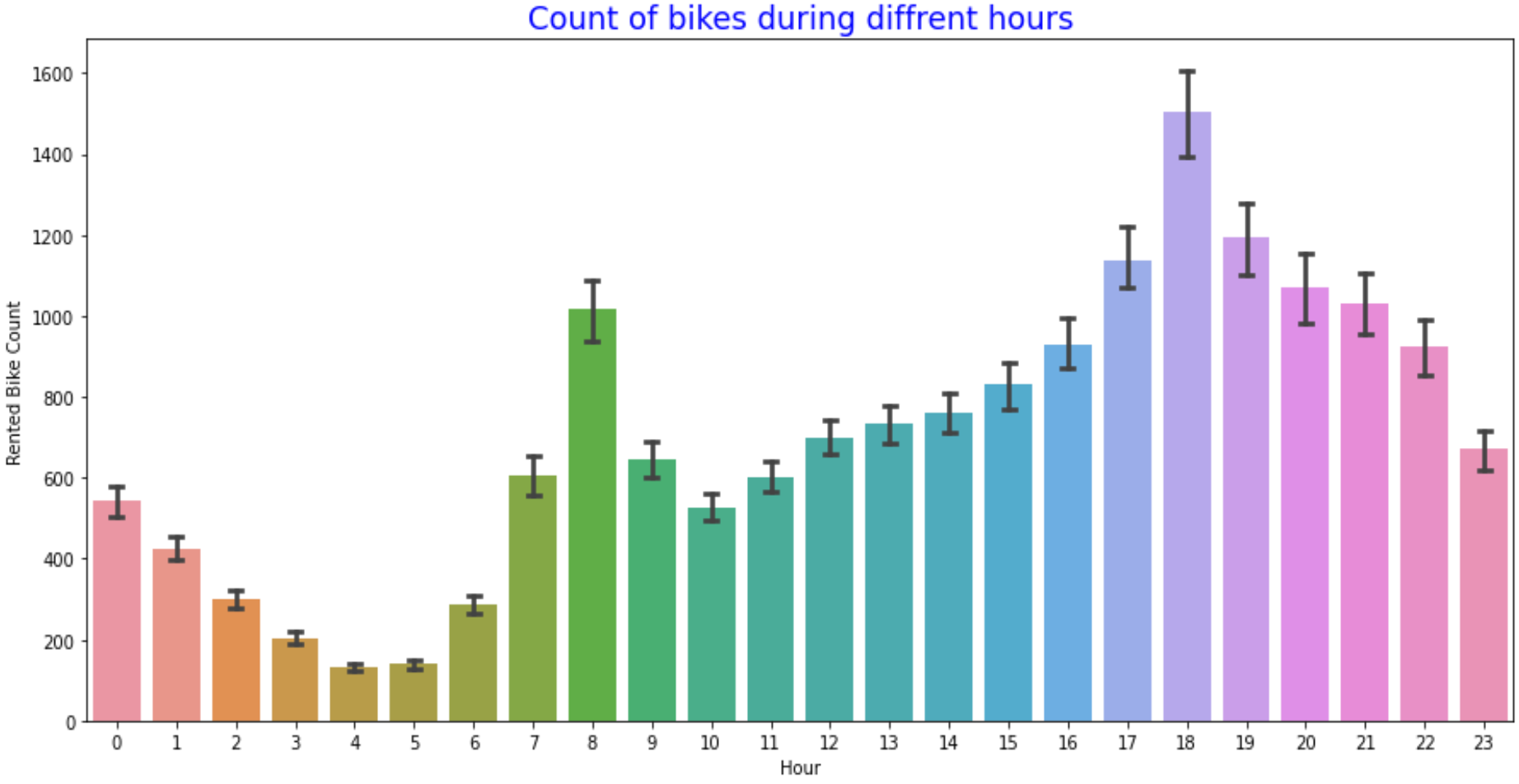
- The box plot shows the presence of outliers in target variables.

Count of Bikes during different of Month

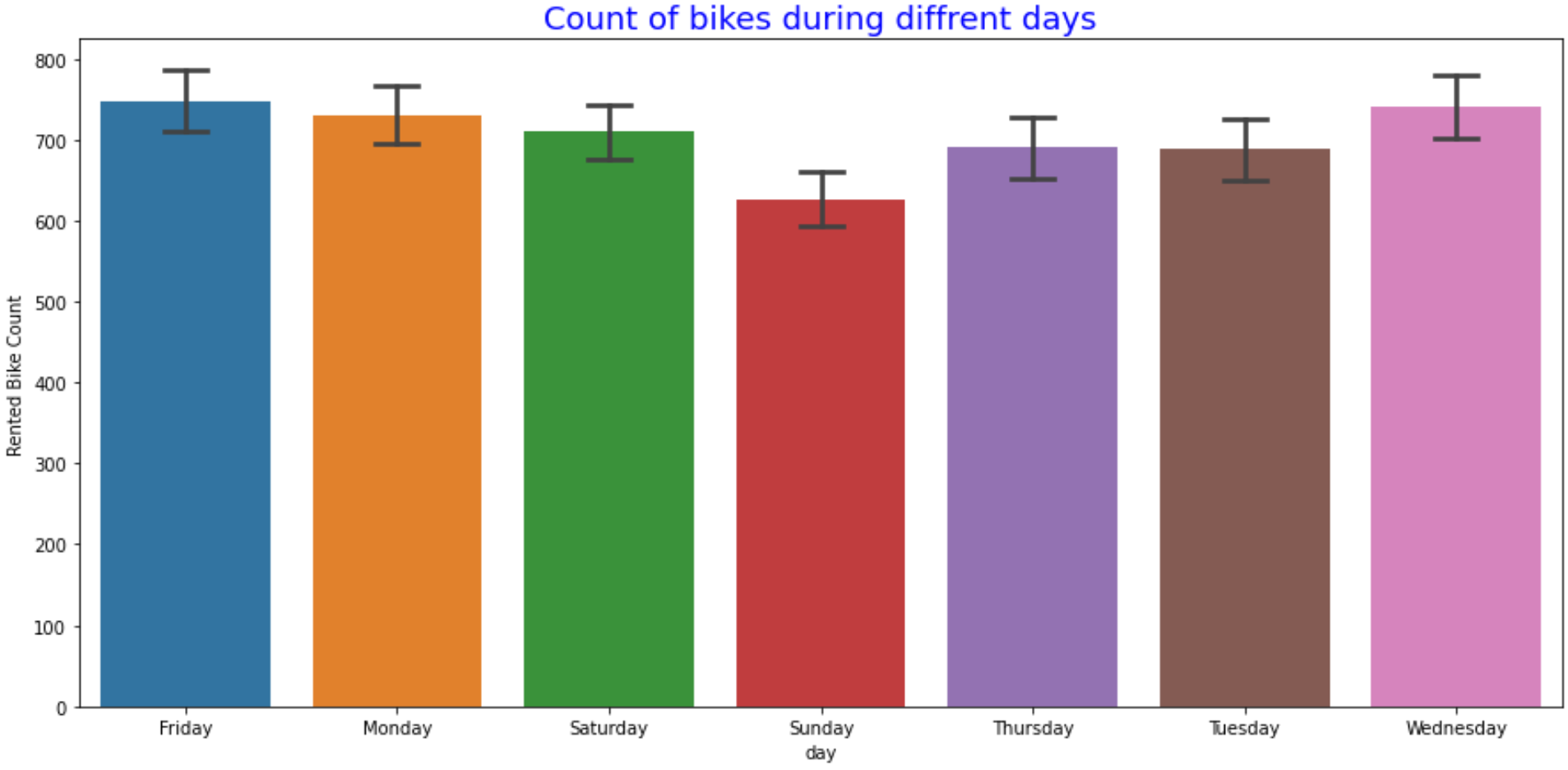


- Demand of rented bike is high in summer season which is from may to september.

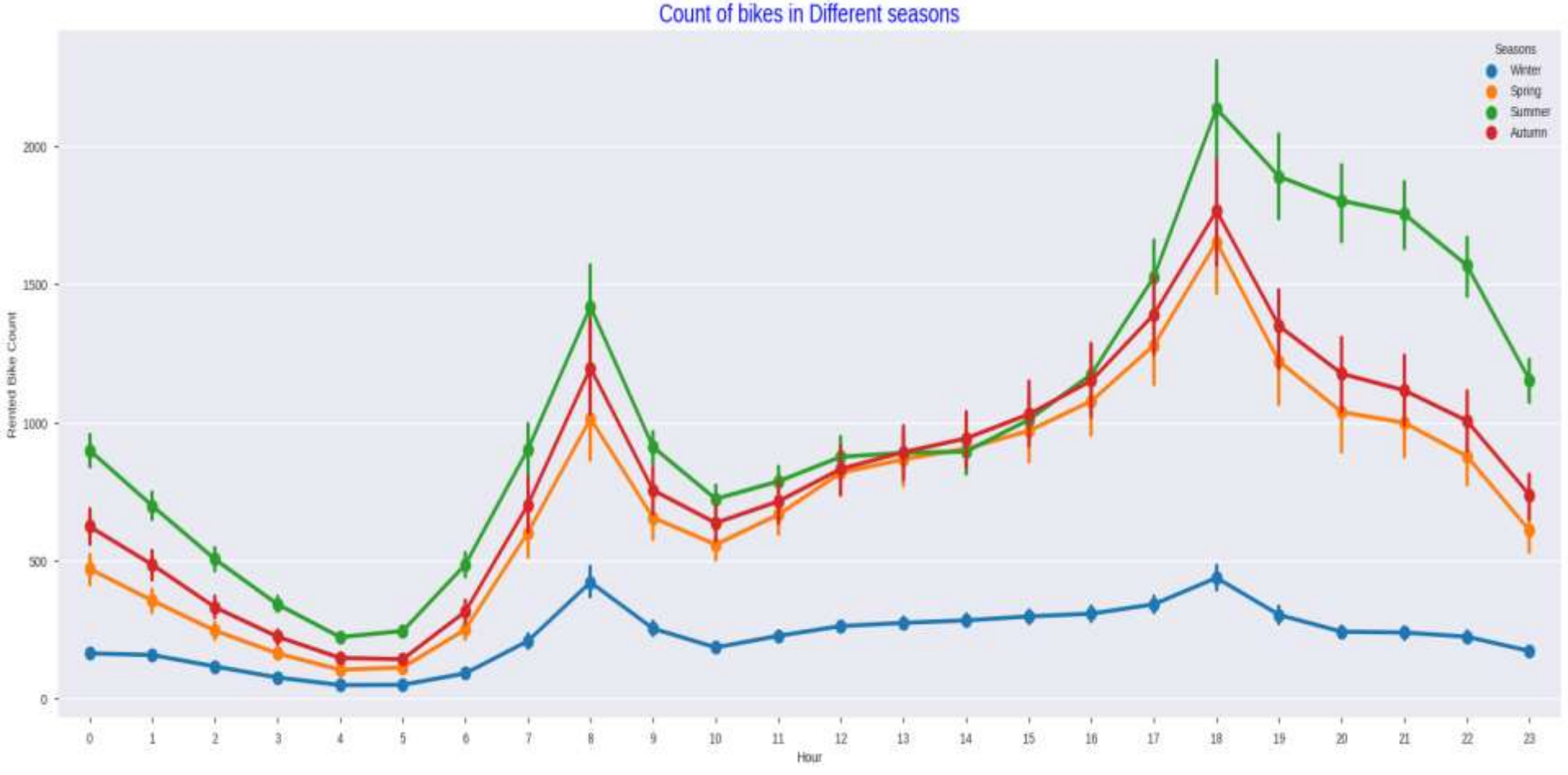
Count of Bikes for each Hour in a year



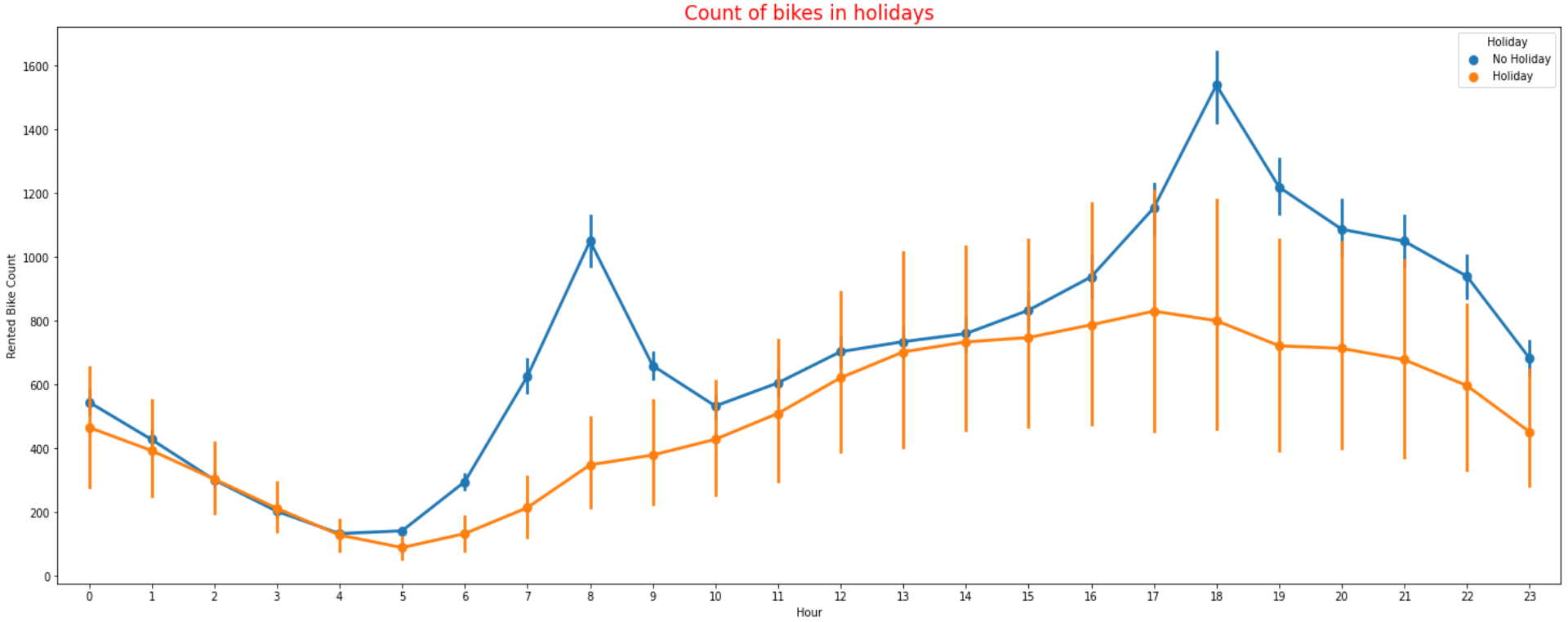
Count of bikes in different days



Count of bikes in Different seasons

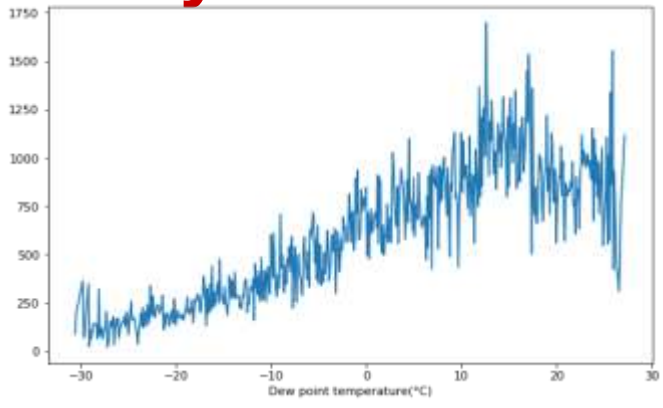


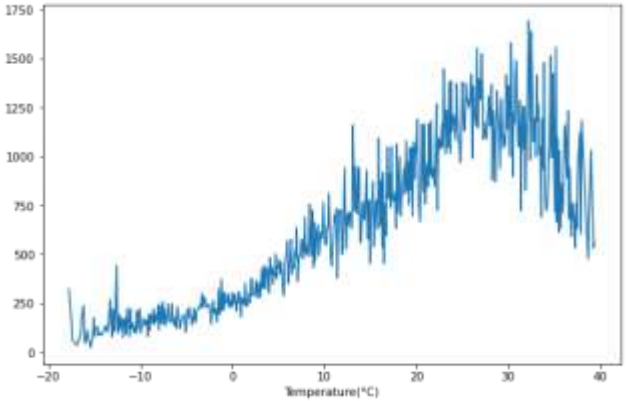
Count of bikes in holidays

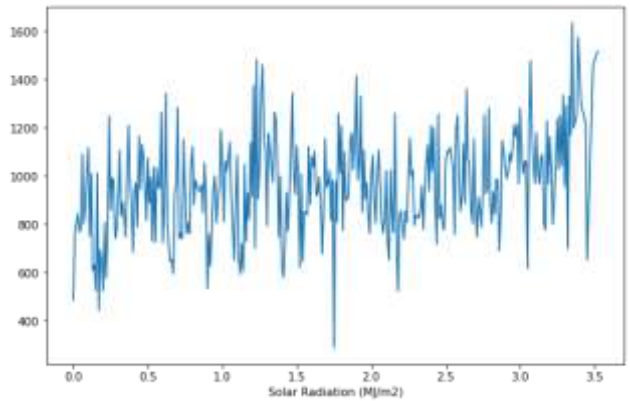


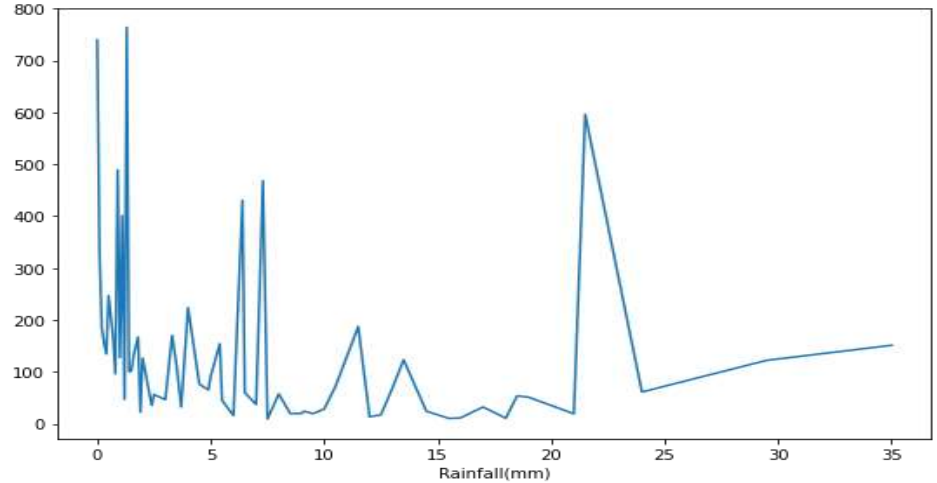
- On the no holiday demand of bike is high between 7 to 9 in morning & 5 to 7 in evening.

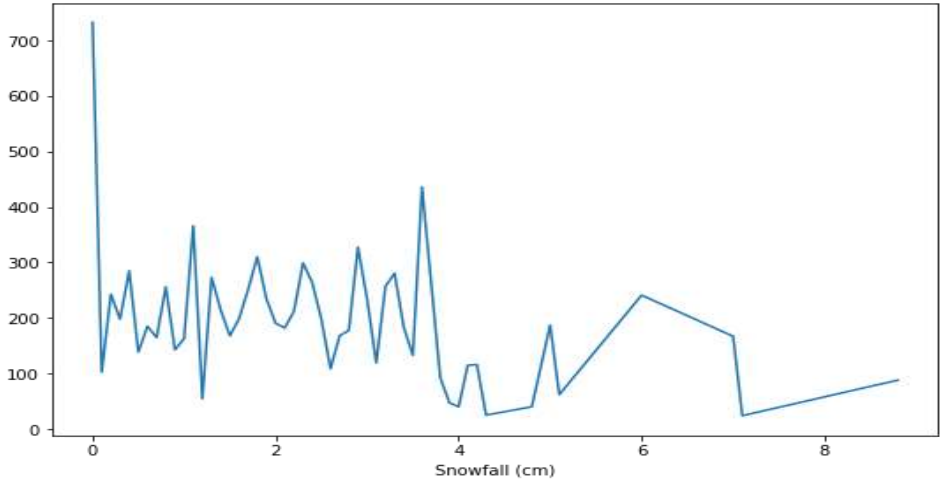
Analysis Numerical variable & Dependent Variable



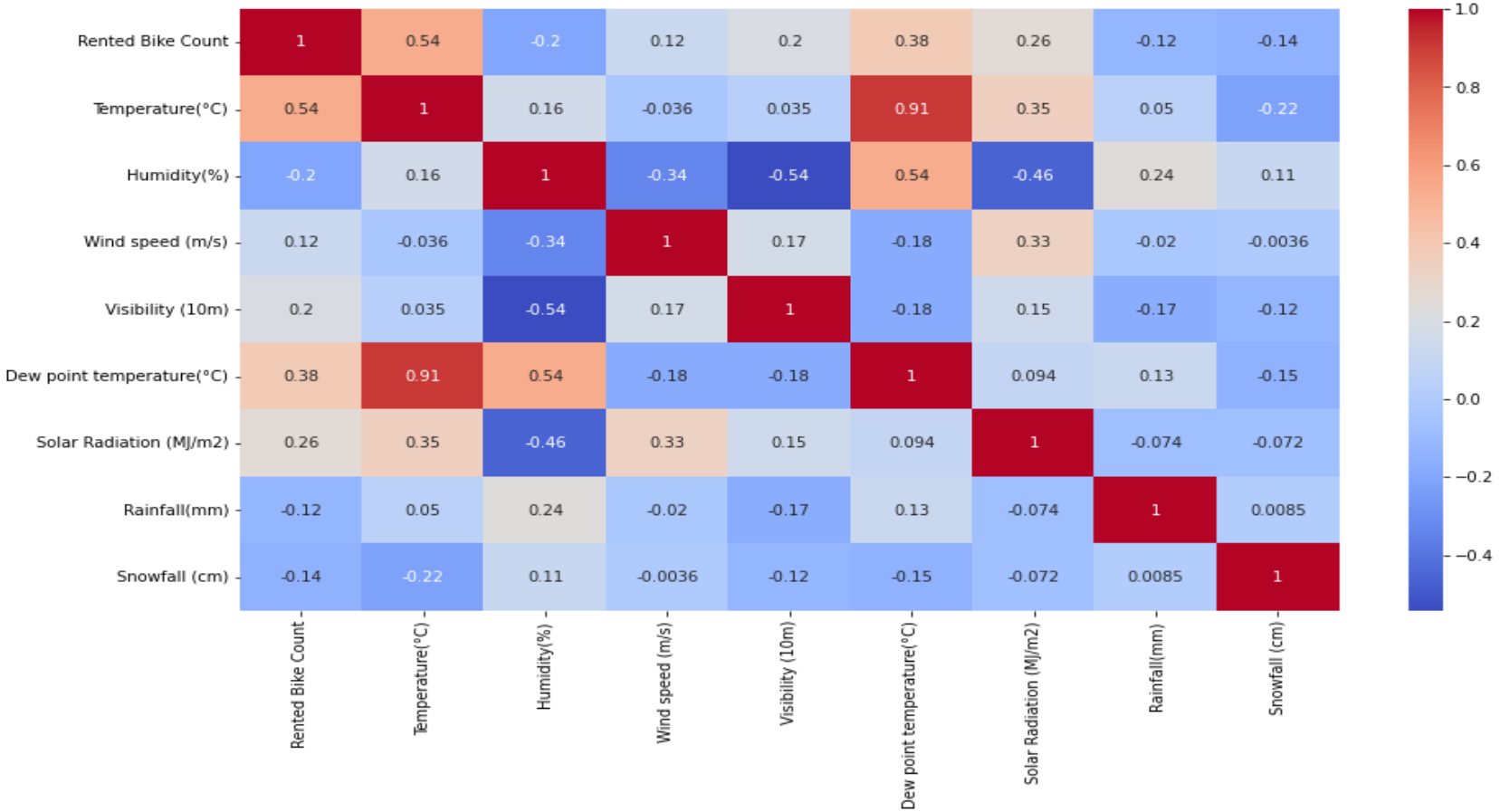








Correlation Matrix



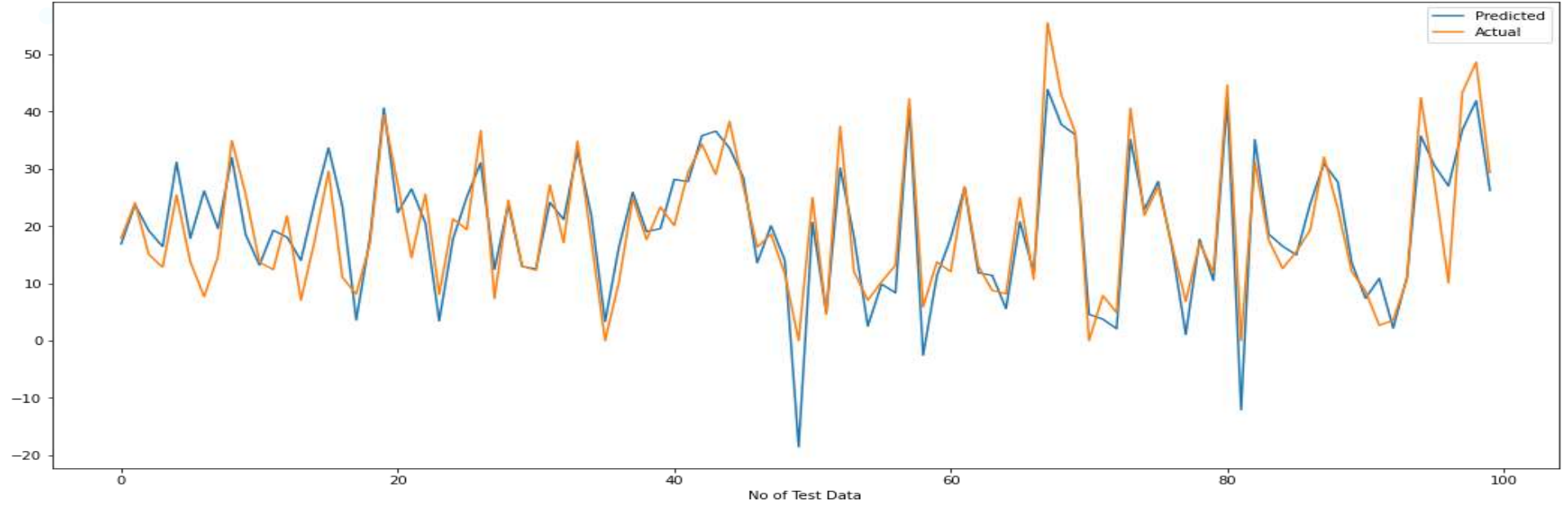
Linear Regression

Train result

MSE: 34.19388697470096
RMSE : 5.84755393089289
MAE: 4.398692851601256
R2 : 0.77844103464799
Adjusted R2 : 0.7713908377540576

Test result

MSE: 33.515173741831845
RMSE : 5.7892291146431445
MAE: 4.384428327762863
R2 : 0.787186676840818
Adjusted R2 : 0.7804147738056997



Lasso Regression

Train result

MSE : 91.68134019453143

RMSE : 9.575037346900086

MAE : 7.25236195572733

R2 : 0.4059516284119845

Adjusted R2 : 0.3870484981434207

Test result

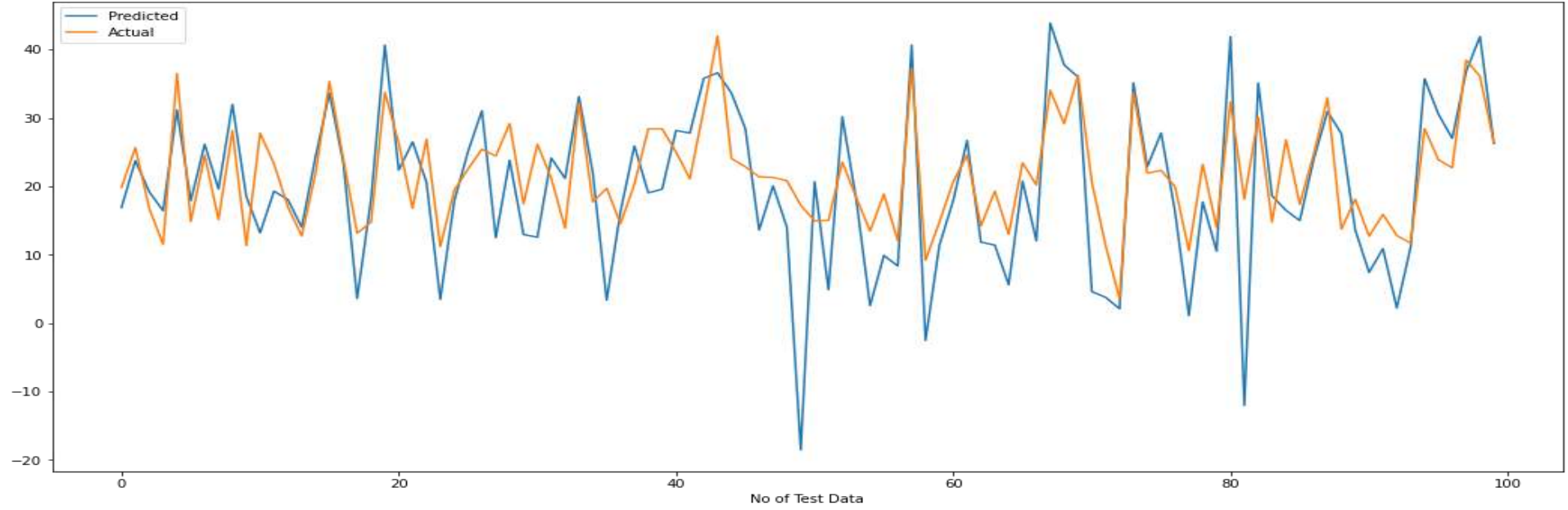
MSE : 97.08251039591772

RMSE : 9.853045742100141

MAE : 7.442888428031137

R2 : 0.38354931956674243

Adjusted R2 : 0.3639333285570807



Ridge Regression

Train result

MSE : 34.194914650252244

RMSE : 5.8476418024920305

MAE : 4.399430563227578

R2 : 0.7870901022812925

Adjusted R2 : 0.7713839670509107

Test result

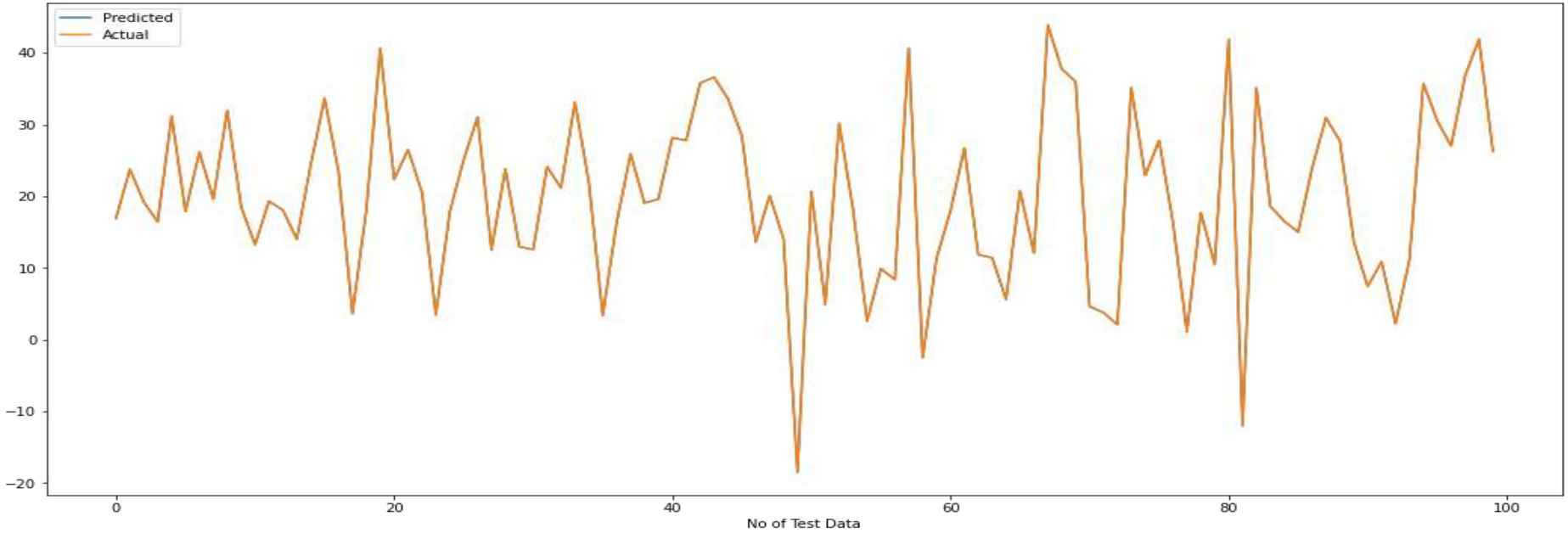
MSE : 33.53038290775007

RMSE : 5.790542540017305

MAE : 4.386514540627386

R2 : 0.7870901022812925

Adjusted R2 : 0.780315126160603



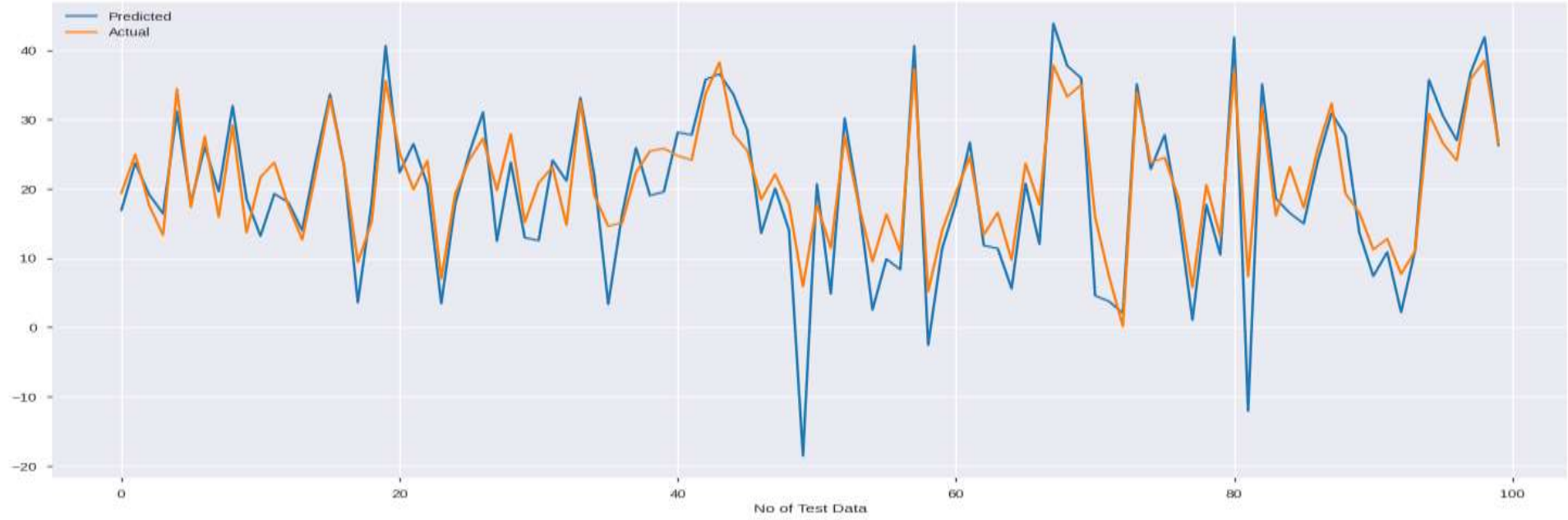
Elastic Regression

Train result

MSE : 57.079943909482175
RMSE : 7.5551269949274955
MAE : 5.7474859198465955
R2 : 0.6301510464635889
Adjusted R2 : 0.6183821345655534

Test result

MSE : 59.7362695667946
RMSE : 7.728924217948744
MAE : 5.840645682783431
R2 : 0.6206890008218877
Adjusted R2 : 0.6086189984909401



Decision Tree

AI

Train result

MSE : 49.62534842289796

RMSE : 7.044526131891198

MAE : 5.131970423194867

R2 : 0.678453027000266

Adjusted_R2 : 0.6682211256791195

Test result

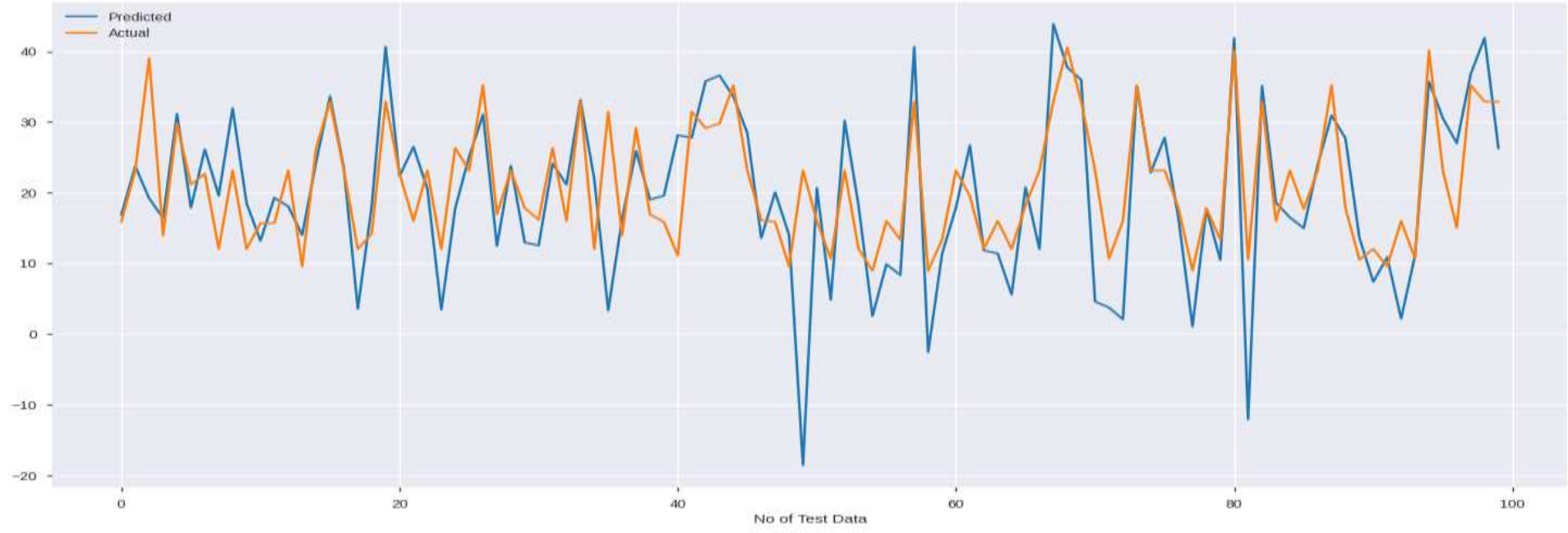
MSE : 57.21350239265634

RMSE : 7.563960760914637

MAE : 5.4177754415537285

R2 : 0.6367079679327508

Adjusted_R2 : 0.6682211256791195



Random Forest

Train result

MSE : 1.6900266899636223

RMSE : 1.3000102653300942

MAE : 0.8430042608885823

R2 : 0.9890494881403026

Adjusted_R2: 0.9887010334317441

Test result

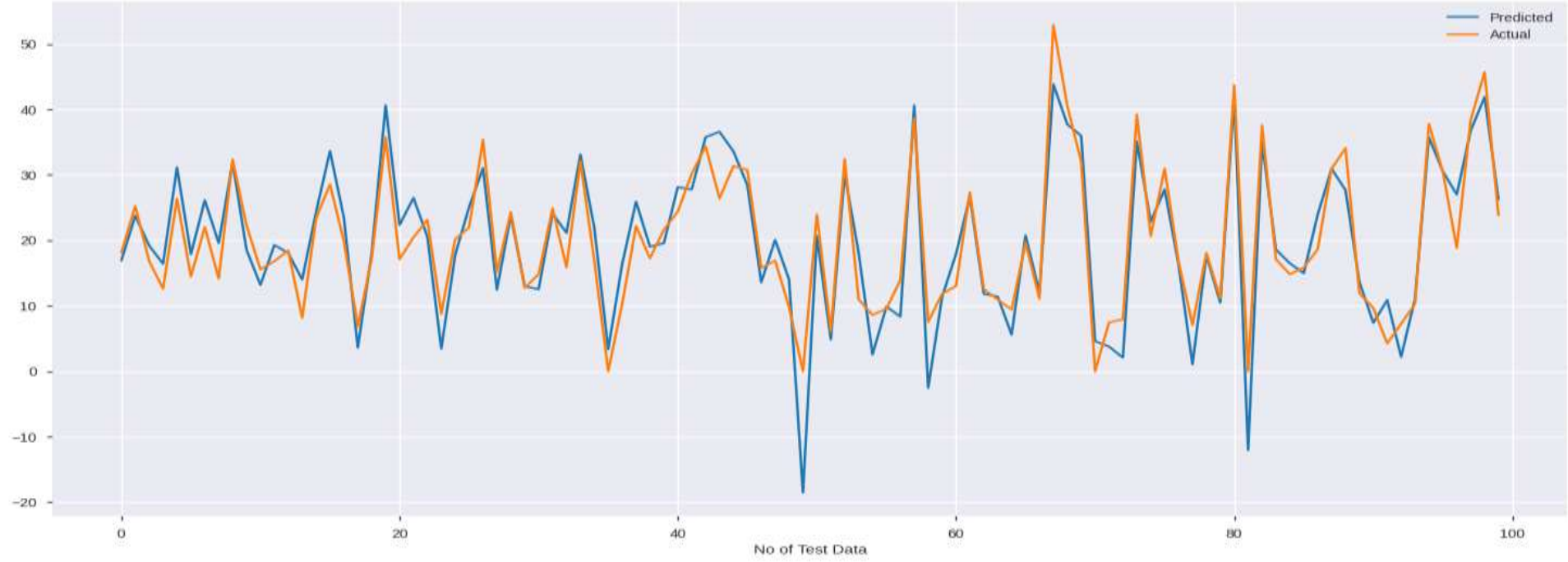
MSE : 13.675876221960827

RMSE : 3.698090888818287

MAE : 2.301005575576969

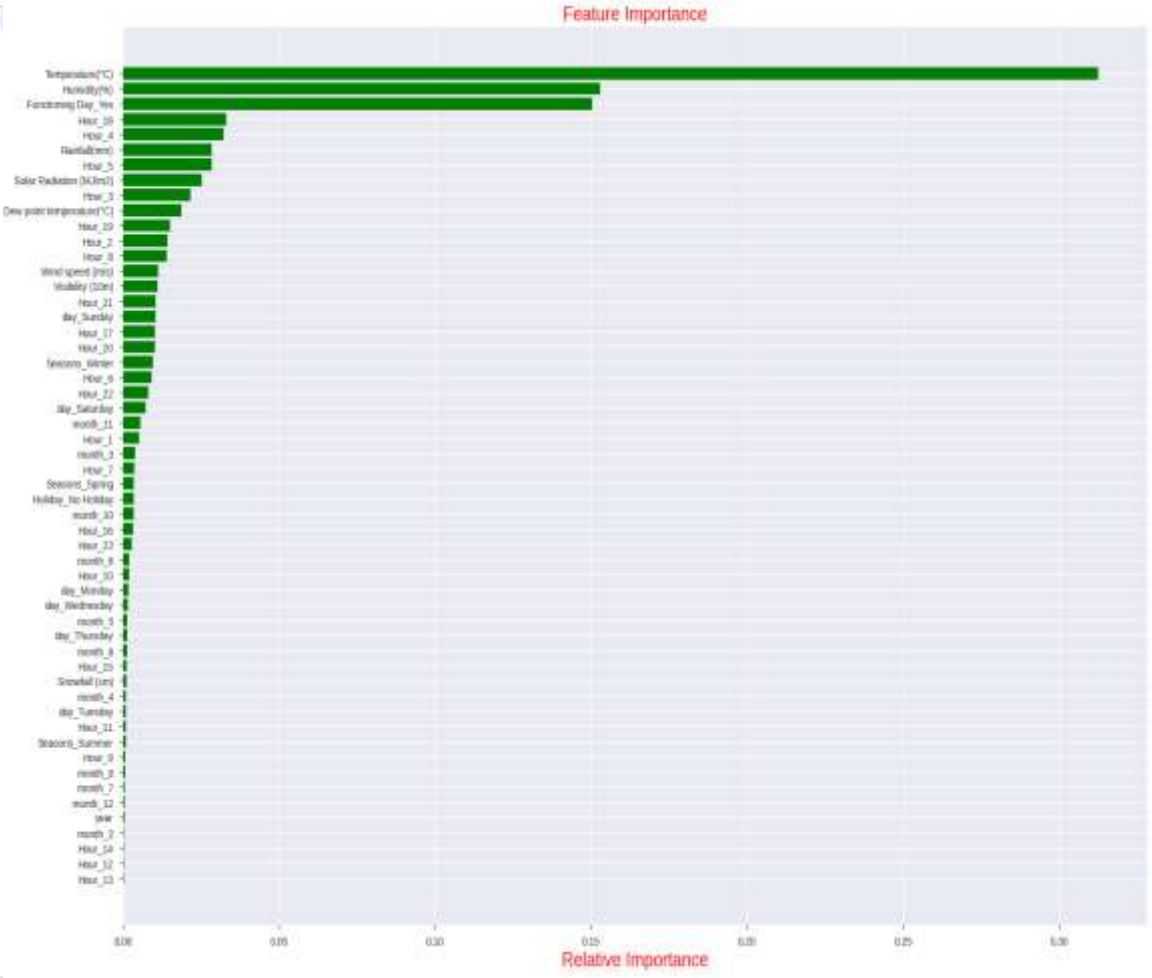
R2 : 0.91316146267574

Adjusted_R2: 0.9103981857072604



Random Forest

	Feature	Feature Importance
0	Temperature(°C)	0.31
13	Functioning Day_Yes	0.15
1	Humidity(%)	0.15
37	Hour_18	0.03
23	Hour_4	0.03
24	Hour_5	0.03
5	Solar Radiation (MJ/m2)	0.03
6	Rainfall(mm)	0.03
22	Hour_3	0.02
38	Hour_19	0.02
4	Dew point temperature(°C)	0.02
20	Hour_1	0.01
36	Hour_17	0.01
39	Hour_20	0.01
40	Hour_21	0.01
25	Hour_6	0.01
41	Hour_22	0.01
52	month_11	0.01
21	Hour_2	0.01
27	Hour_8	0.01



Overall conclusion

- After comparing the root mean squared error and mean absolute error of train and test result of all the applied model on the dataset, I found that Random forest gave the highest R2 score of 98% and 91% on train and test data respectively.

- Hence can be concluded that random forest is best model for predicting bike rental for each hour and lessens the waiting time and enhancing the. mobility comfort.