Capstone Project-II
Supervised ML
(Regression)
Bike Sharing Demand Prediction

- Team Members:
- Shubham Joshi
- Kanika Kakra
- Akshay Fasale
- Rishikesh Damale

Acknowledgement

We would express our gratitude towards the entire team of "*Almabetter*" for acknowledging us with such important domain and providing us an opportunity to work on real life problems through Capstone Project.

▪***Problem Statement***

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.

▪***Algorithm followed***

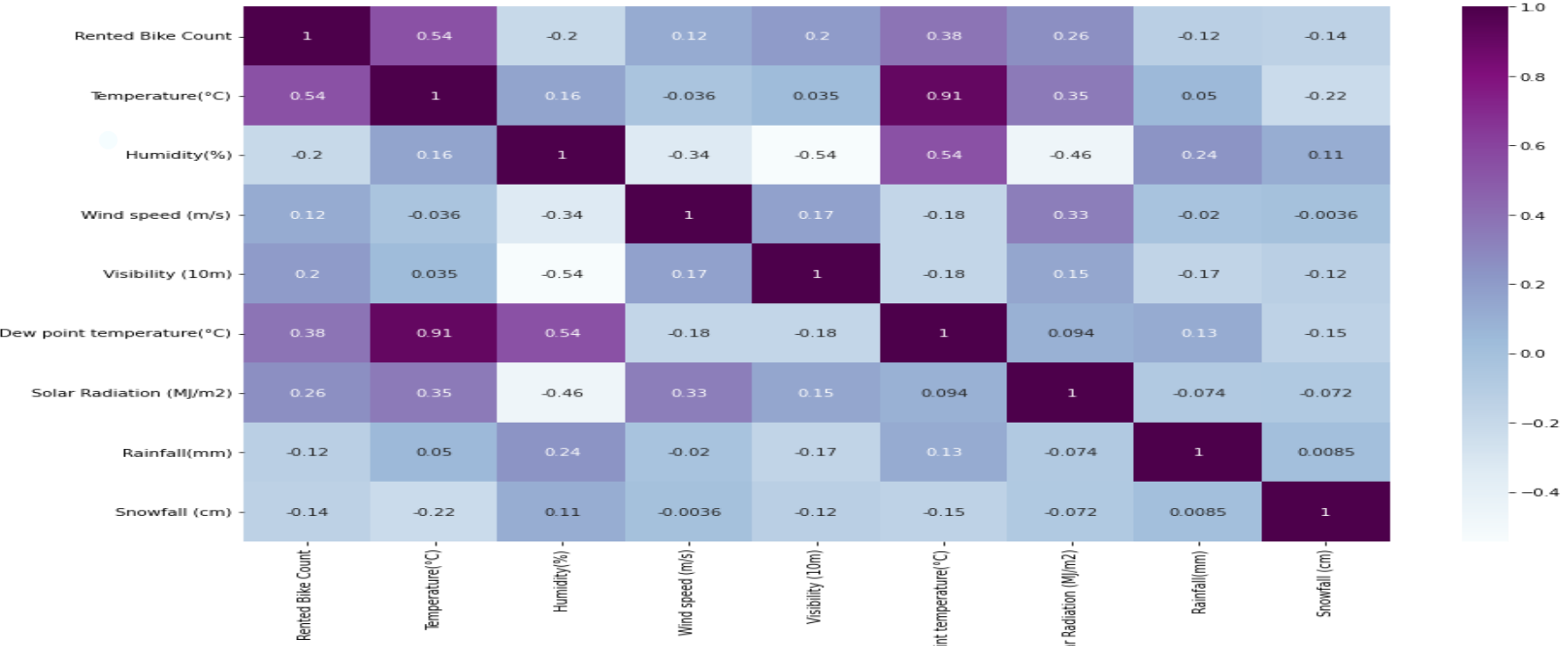
1.Importing the necessary packages and libraries.
2.Mounting the drive for importing the data.
3.Checking for missing,NaN values,Null values.
4.Observing the datatypes .
5.Observing the correlation among independent variables.
6. Exploring the data set.
7.Exploring the categorical values, numerical features from data set.
8.Exploring different target variable.
9.Splitting the data and training the data.
10.Observing the results.

▪Data Columns

Column Name	Description
Date	Year-month-day
Rented Bike count	Count of bikes rented at each hour
Holiday	Holiday/No holiday
Functional Day	NoFunc(Non Functional Hours), Fun(Functional hours

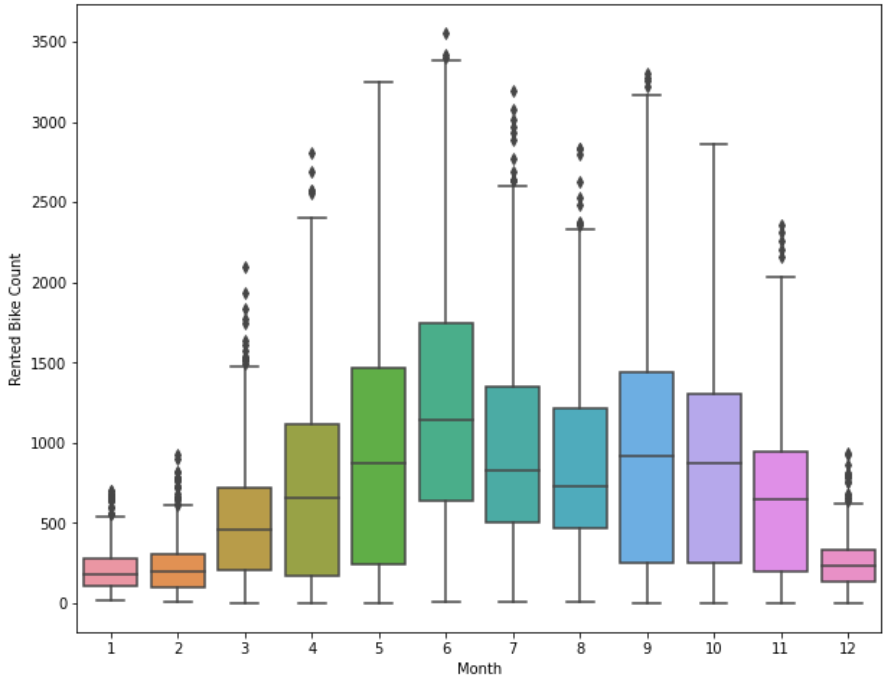
- ❖Hour - Hour of The day
- ❖Temperature-Temperature in Celsius
- ❖Humidity - %
- ❖Rainfall - mm
- ❖Snowfall - cm
- ❖Windspeed - m/s
- ❖Visibility - 10m
- ❖Dew point temperature - Celsius
- ❖Solar radiation - MJ/m2

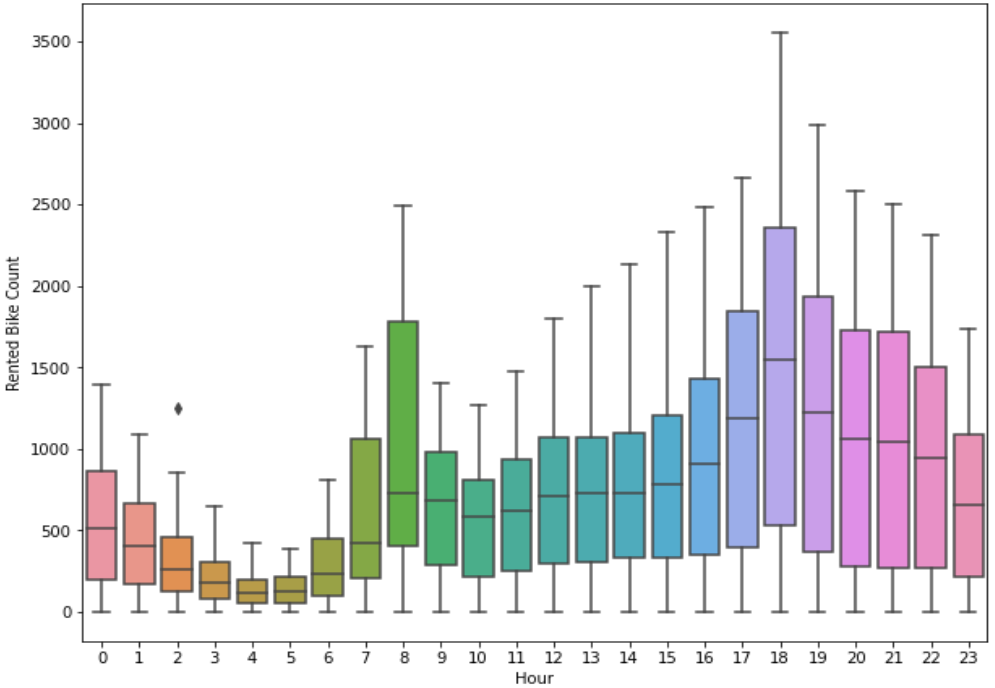
•Correlation

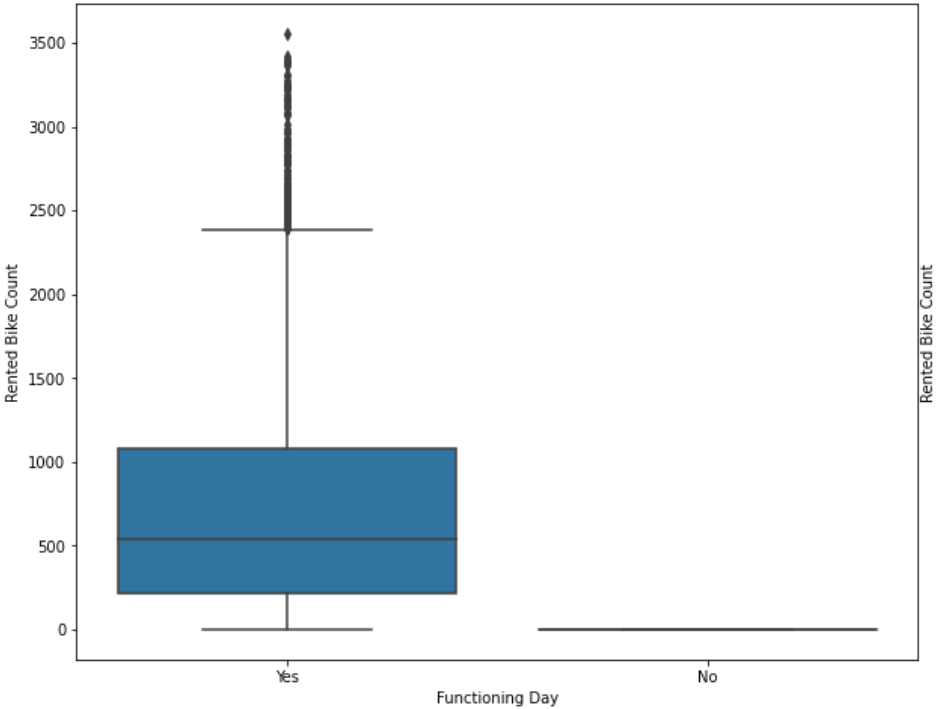


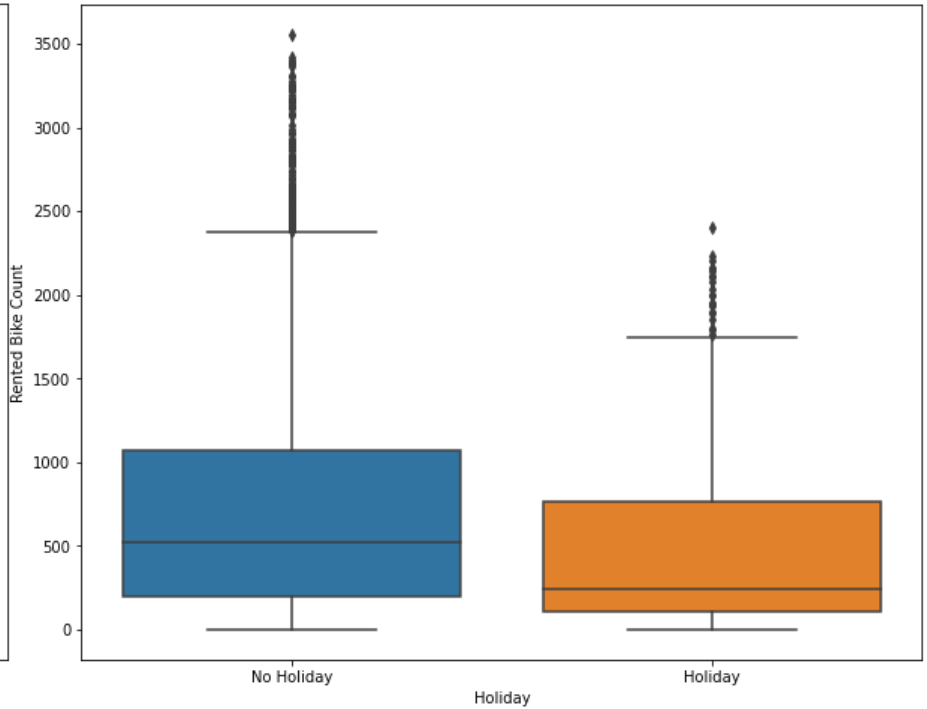
❖We observe that multicollinearity between Temperature and Dew point Temperature 0.91

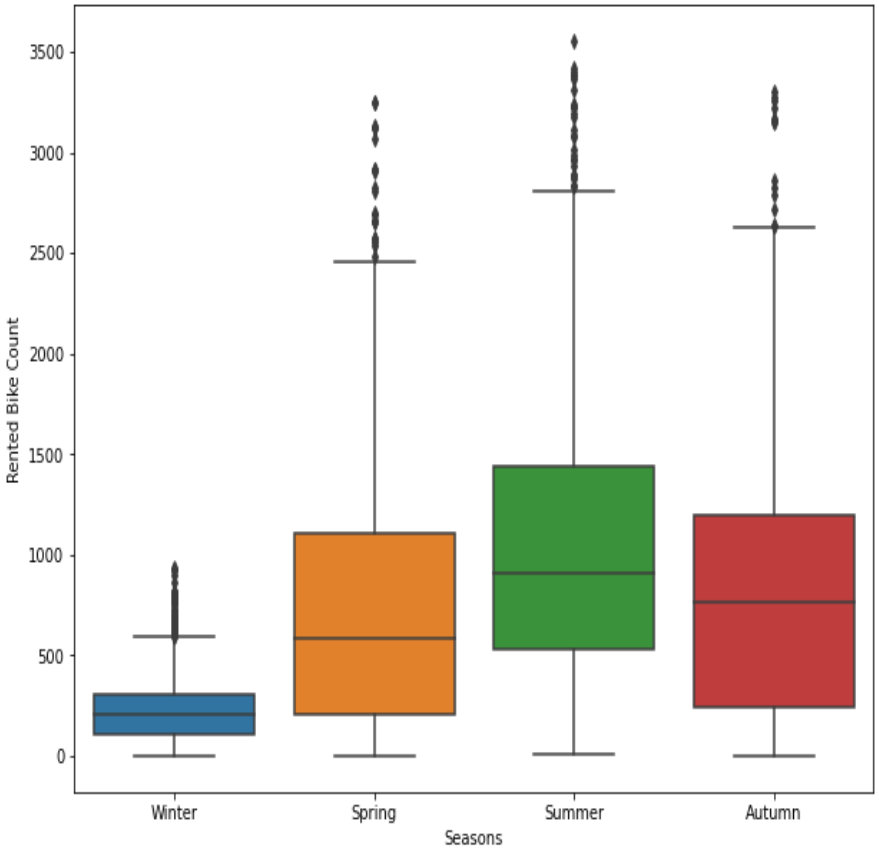
Exploring categorical features

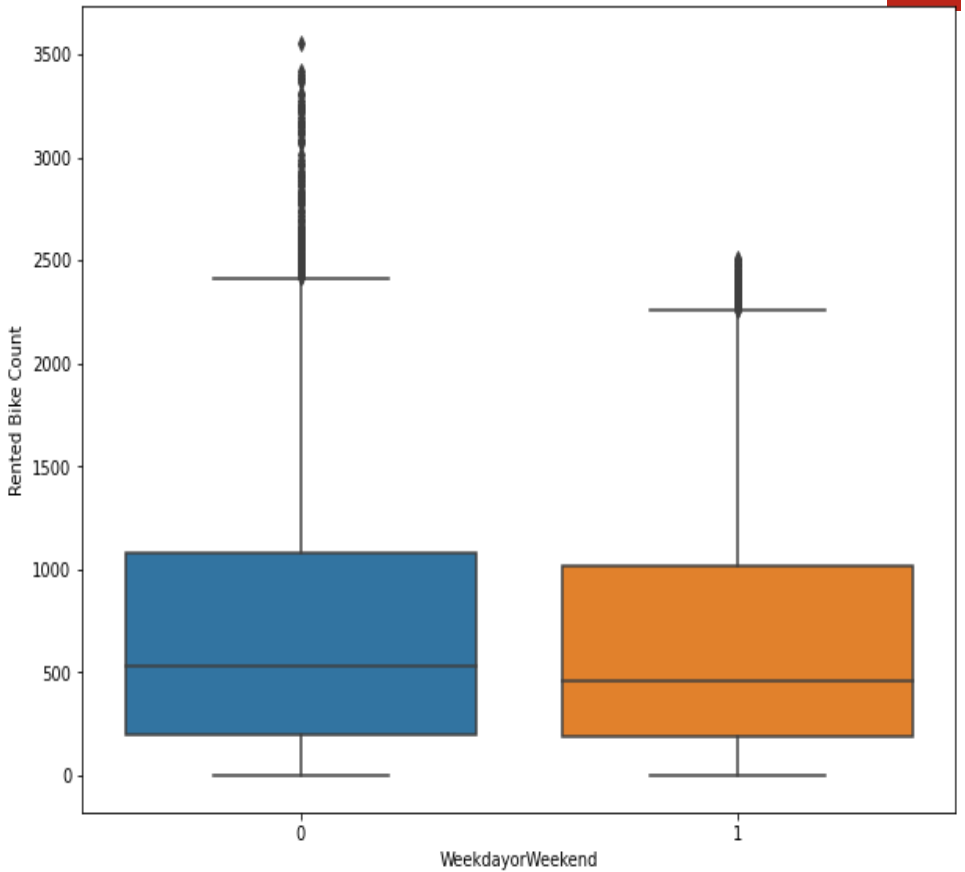





Rented Bike Count
3500
3000
2500
2000
1500
1000
500
0
Yes
No
Functioning Day


Rented Bike Count
3500
3000
2500
2000
1500
1000
500
0
No Holiday
Holiday
Holiday

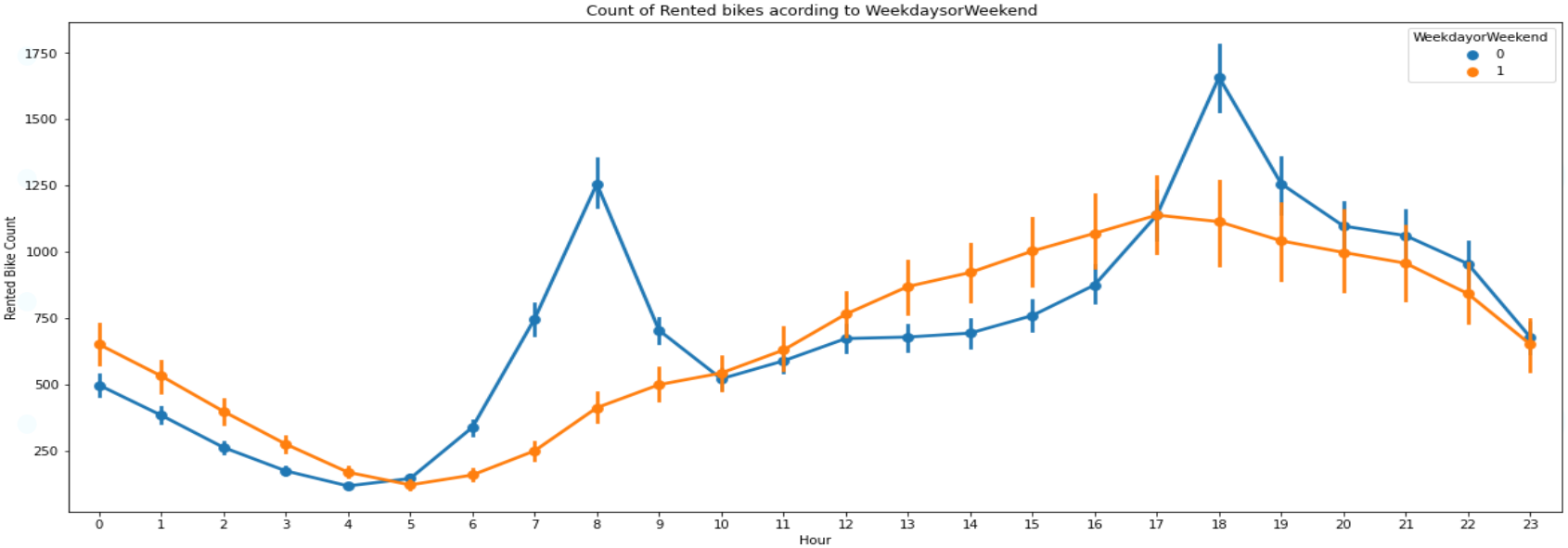

Rented Bike Count
3500
3000
2500
2000
1500
1000
500
0
Winter
Spring
Summer
Autumn
Seasons


Rented Bike Count
3500
3000
2500
2000
1500
1000
500
0
0
1
WeekdayorWeekend

Conclusions drawn

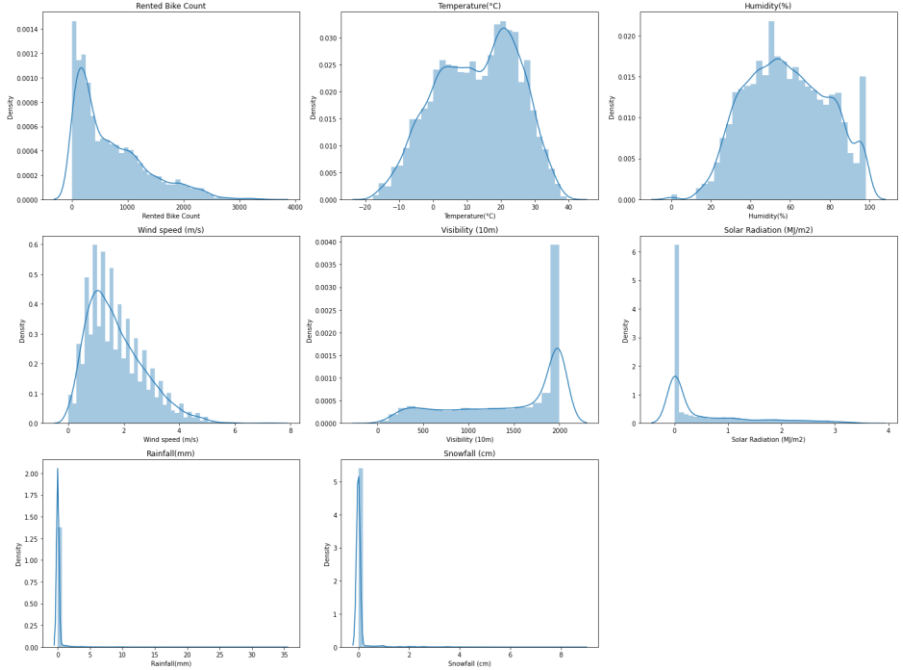
Less demand on winter season and more demand is in summer season.
Sligthly Higher demand during Non holidays.
Almost no demnad on non functioning day.
Most demand for bike is in between 7 to 9 AM and 5 to 8 PM .
More demand is in months May , June , July , August , september, November (Summer season) and less demand in december , January and February(Winter Season).
Weekday or weekend doesnt affect the rented bike count , we will try to see on the basis of hours how it affects.

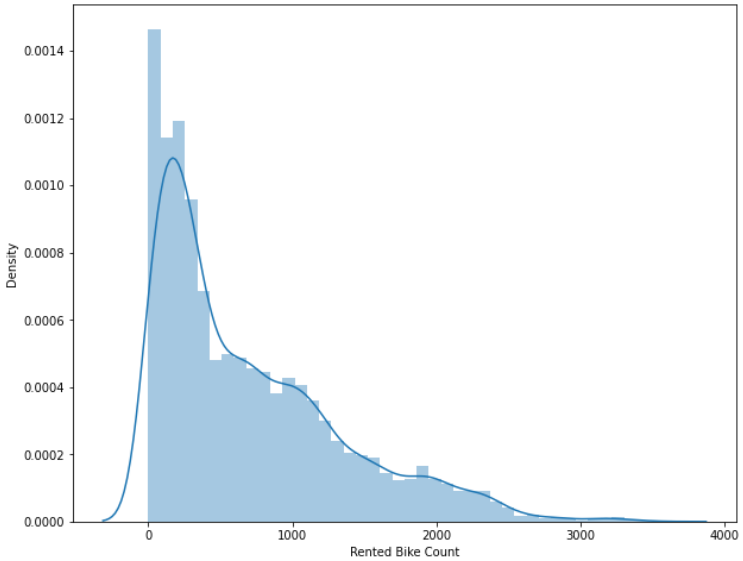
Hourly analysis for weekday or weekend



For Weekdays the count is higher in Morning 7 to 9 Am and then in the evening 5 to 8 PM

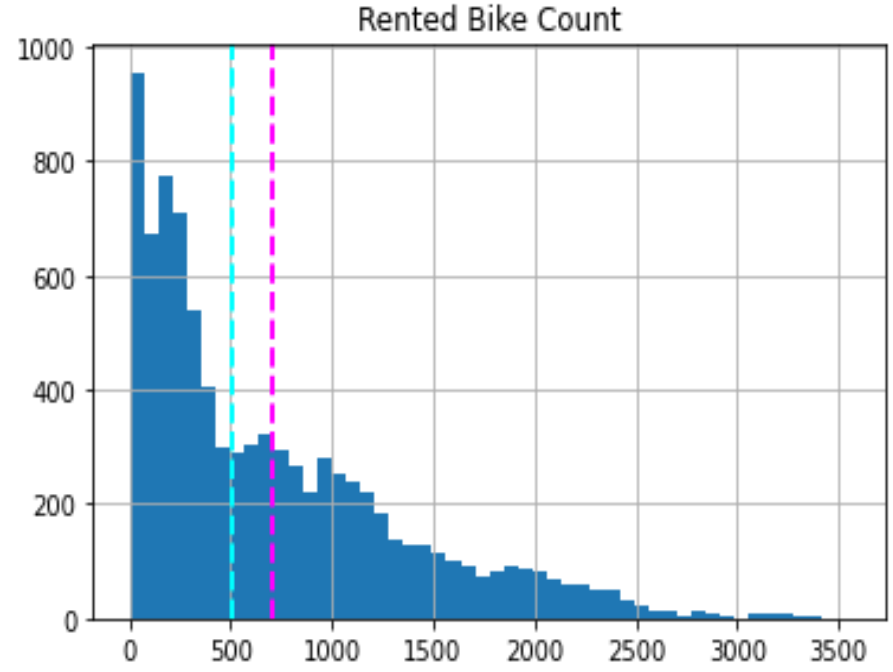
For weekends count is low in the morning but it gradually increases after 10 AM.

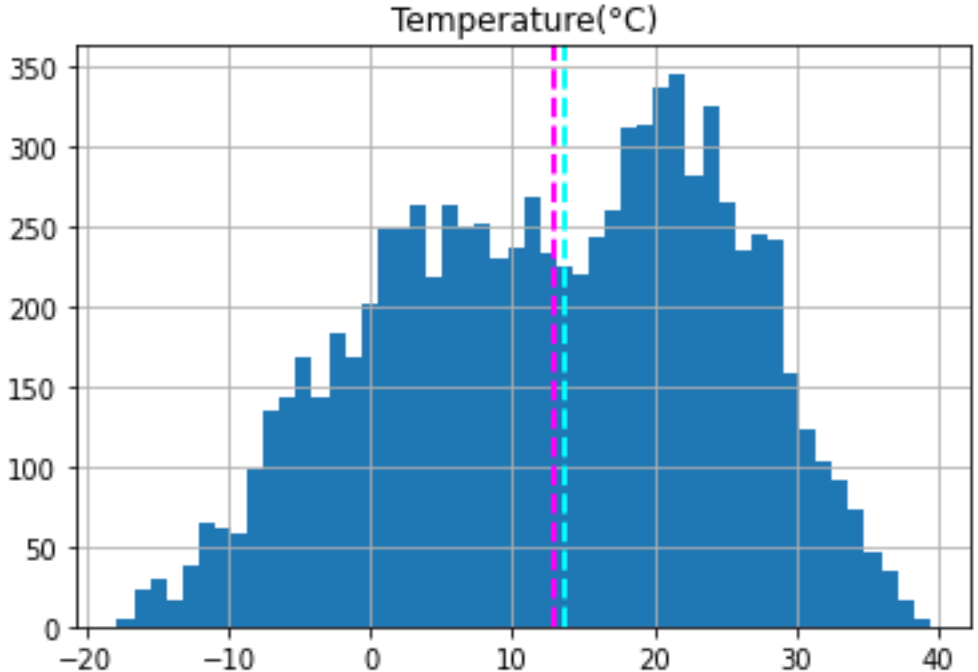


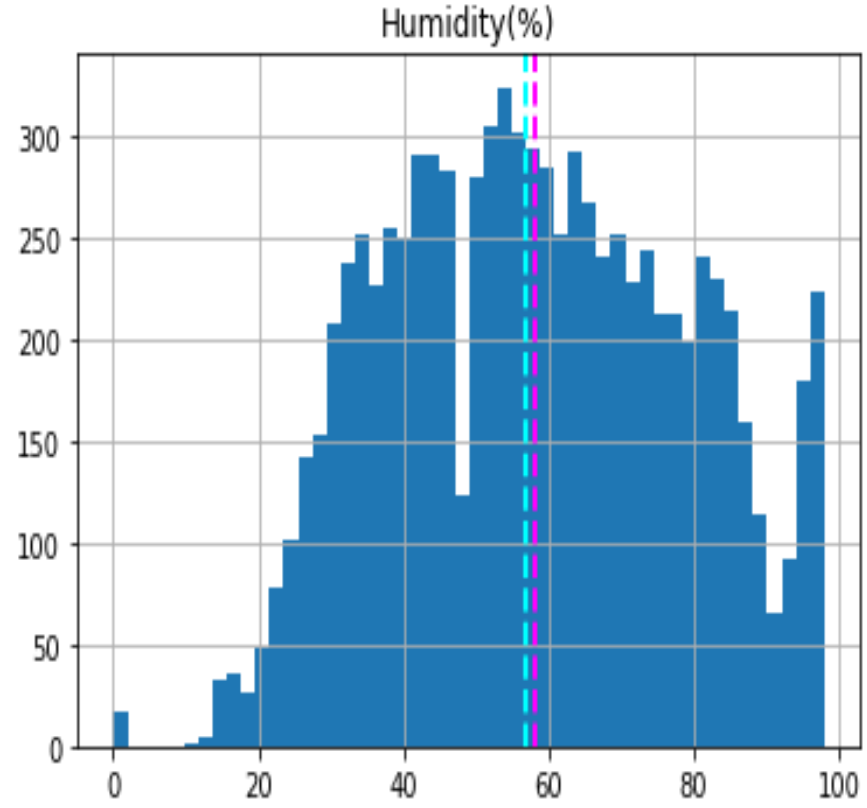


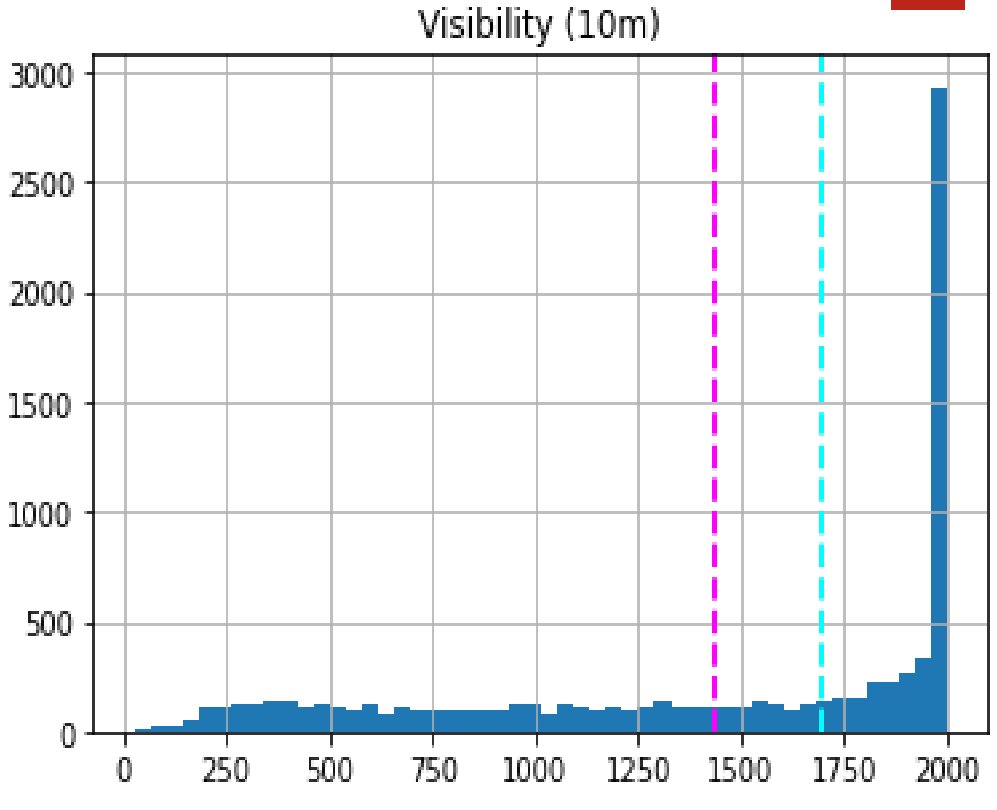
Skewness in numerical variables...

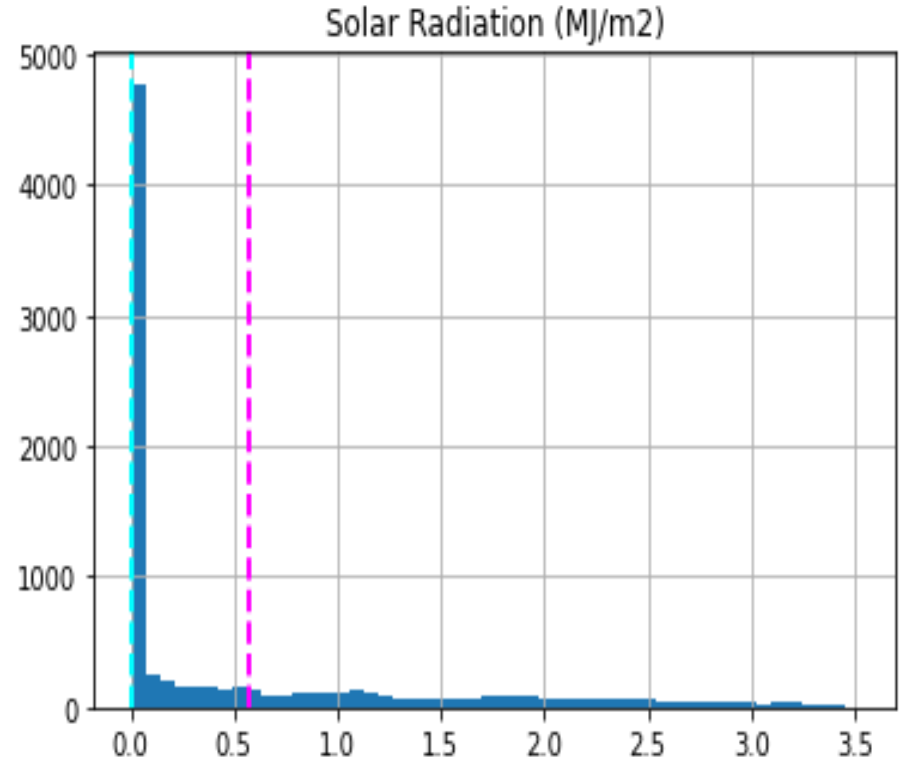
▪Plotting the histogram for numerical feature

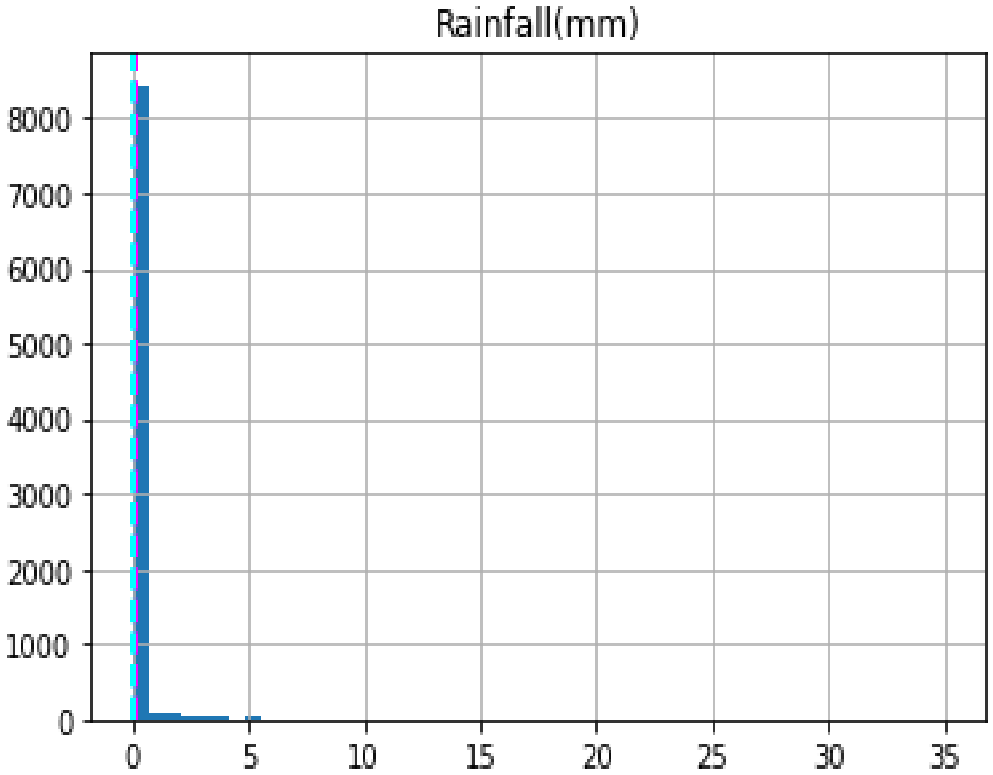


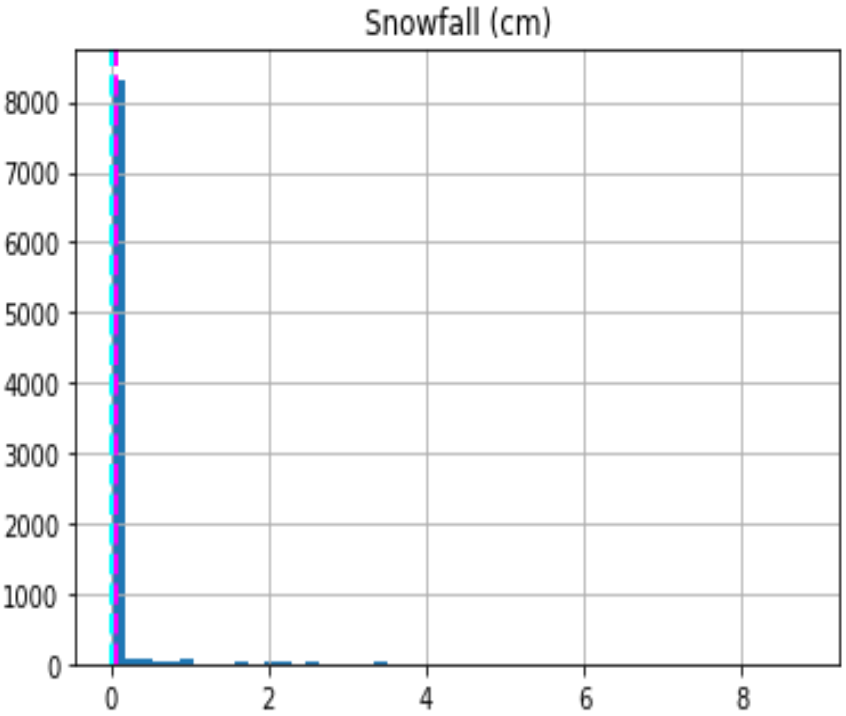


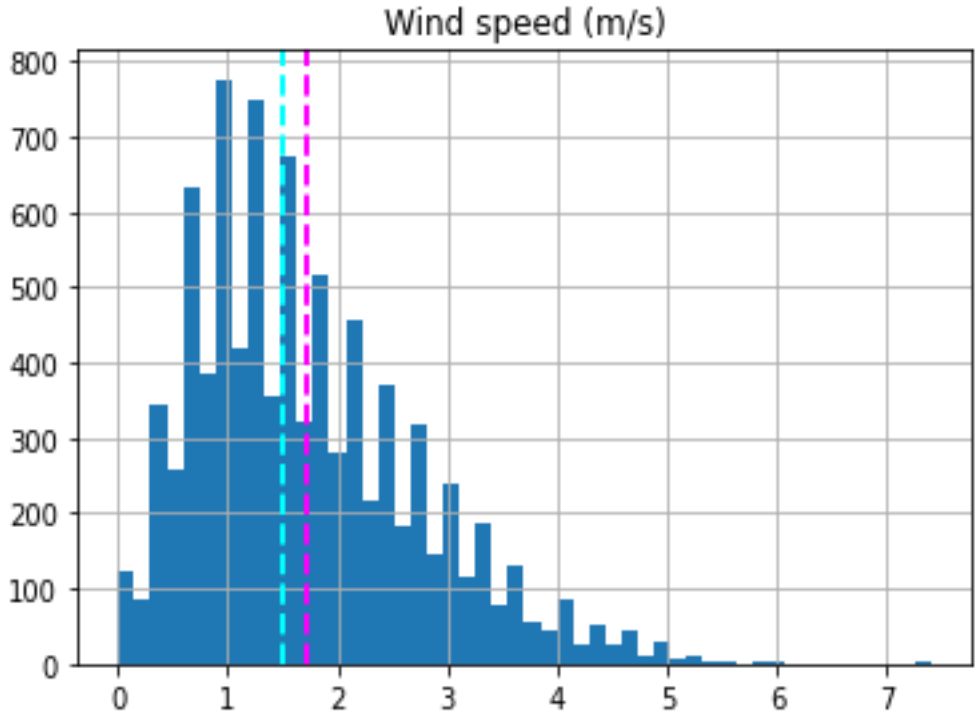

Humidity(%)
300
250
200
150
100
50
0
0
20
40
60
80
100


Visibility (10m)
3000
2500
2000
1500
1000
500
0
0
250
500
750
1000
1250
1500
1750
2000


Solar Radiation (MJ/m2)
5000
4000
3000
2000
1000
0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5


Rainfall(mm)
8000
7000
6000
5000
4000
3000
2000
1000
0
0
5
10
15
20
25
30
35





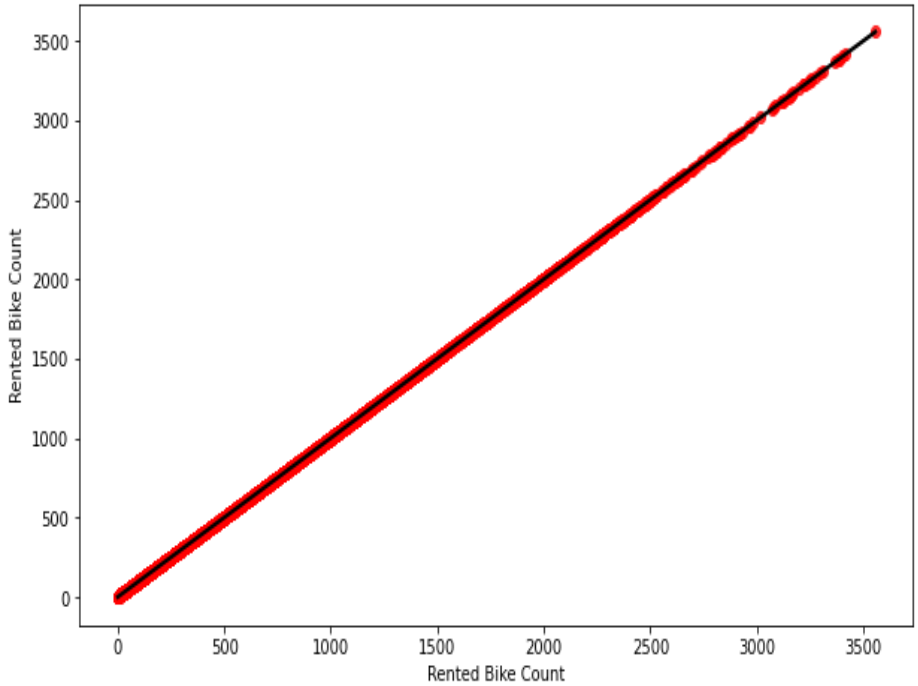
By observing 'Mean and Median' We can see the the Left or Right skewness in data.

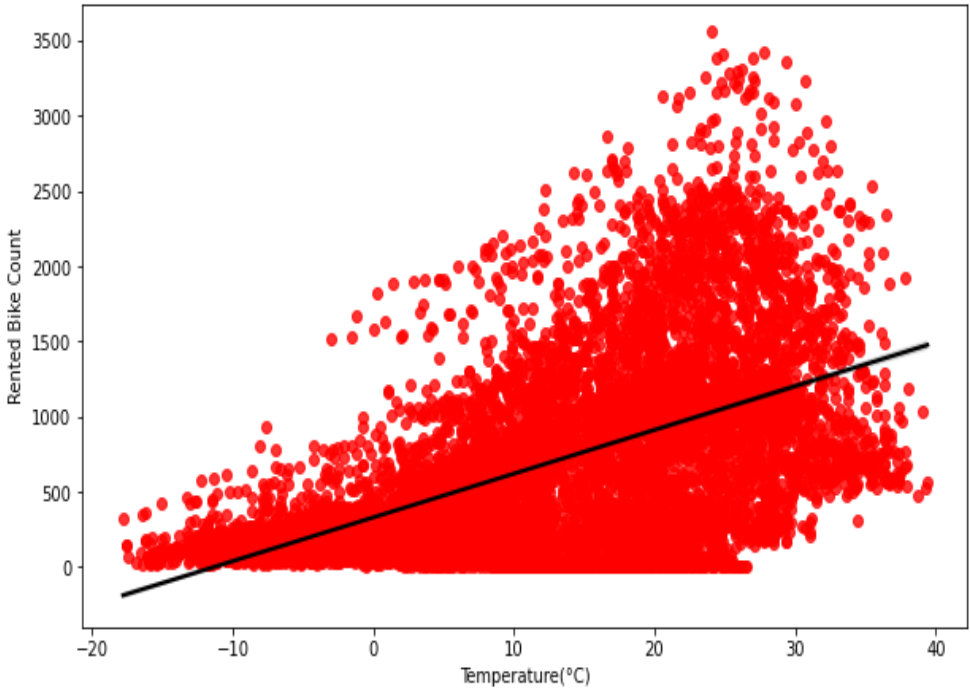
Right skewed columns are :

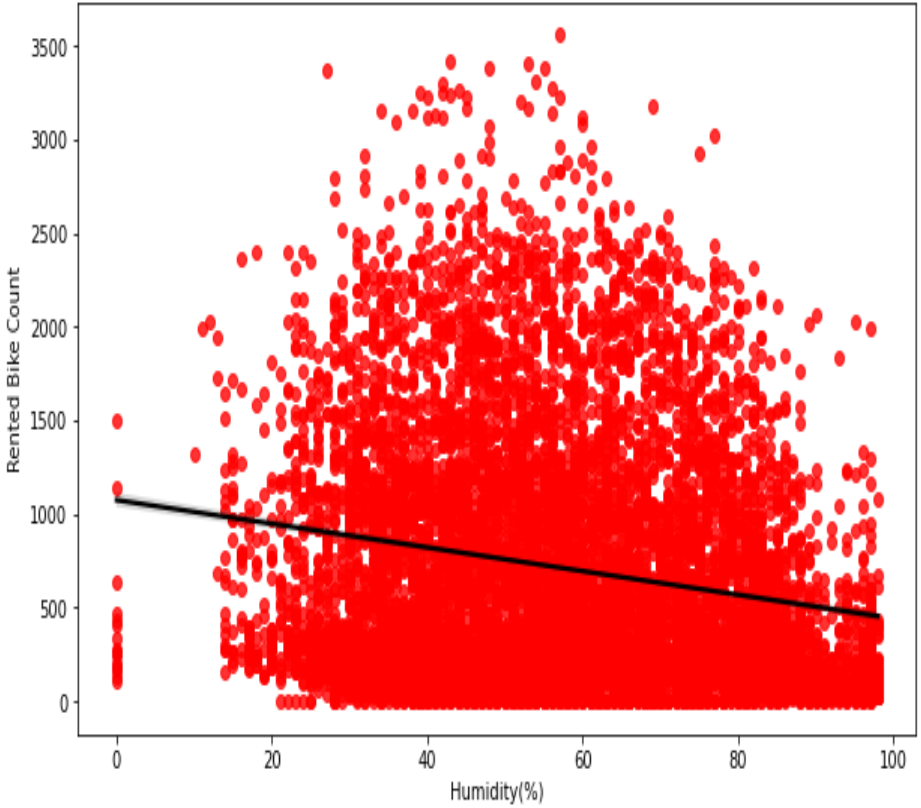
Rented Bike Count (Dependent variable),), Solar Radiation (MJ/m2),Snowfall (cm),

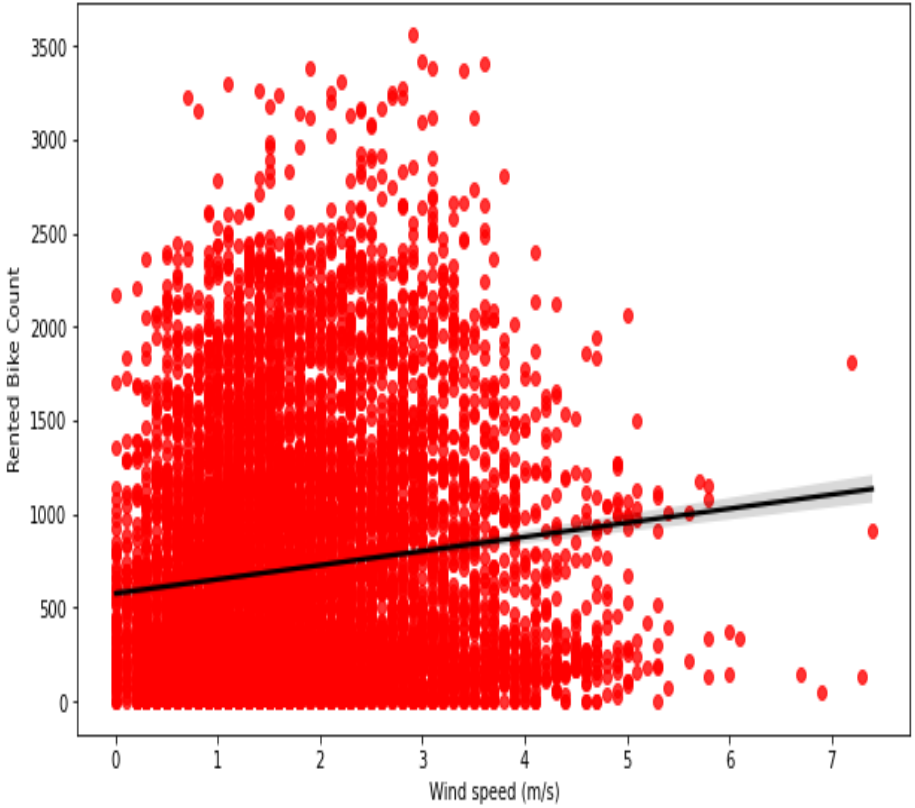
Left skewed columns are : Temprature(degree celcius),Visibility (10m), Rainfall(mm), Wind speed (m/s)

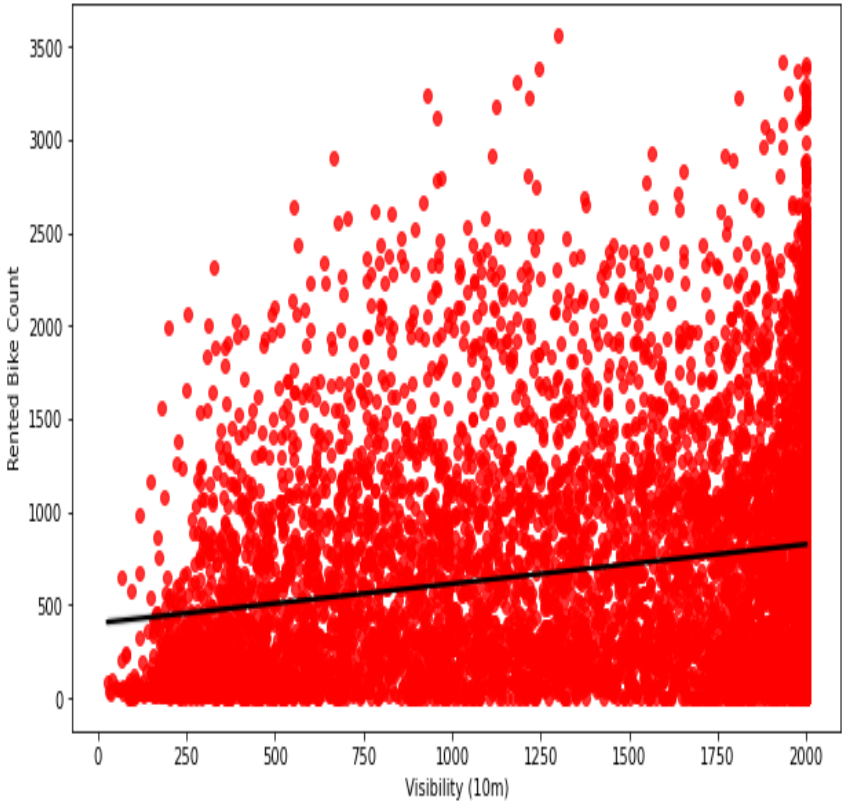
▪Regression plot for all numerical features

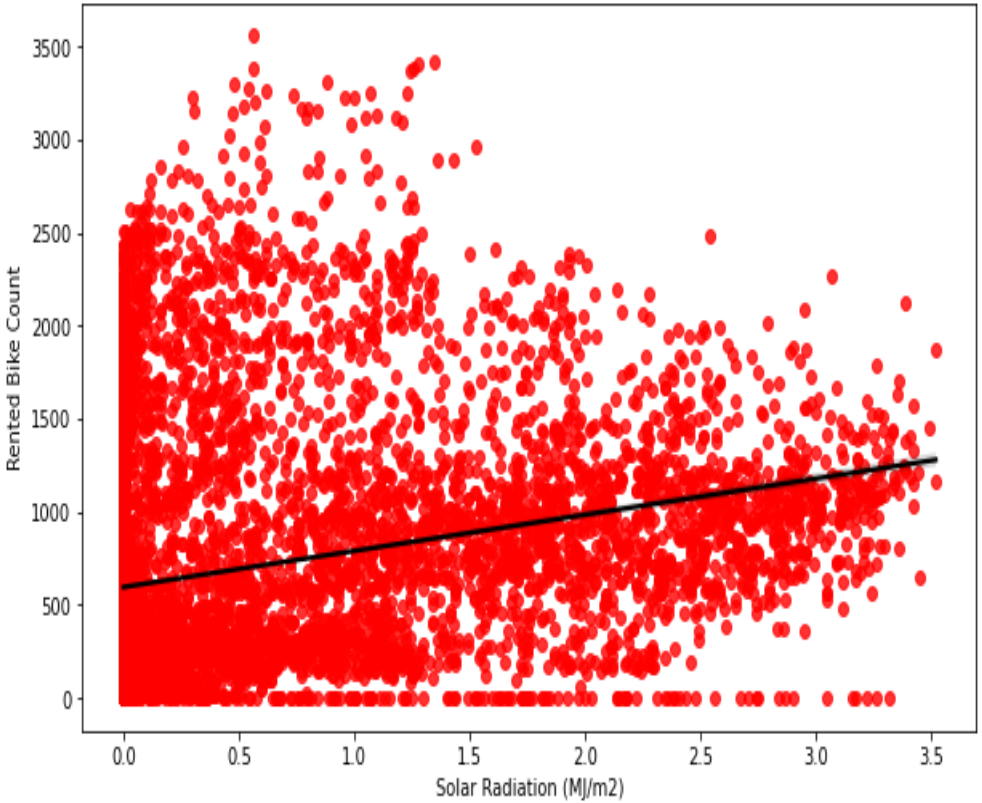


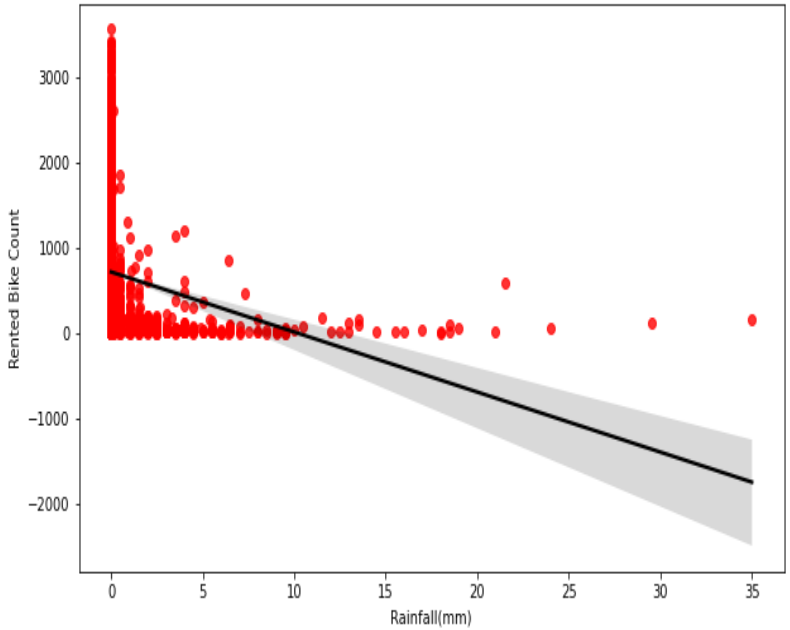


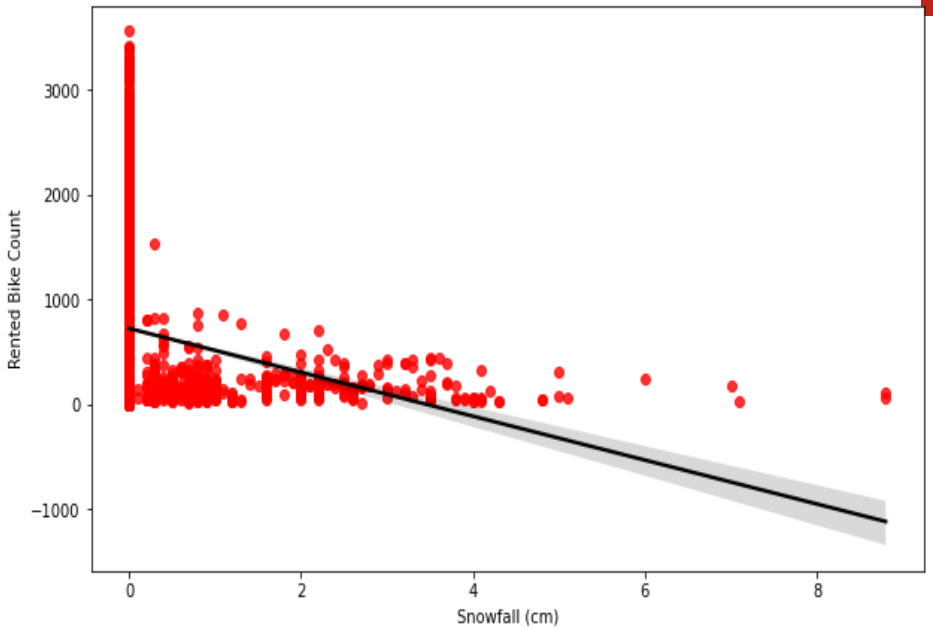

Rented Bike Count
3500
3000
2500
2000
1500
1000
500
0
0
20
40
60
80
100
Humidity(%)


Rented Bike Count
3500
3000
2500
2000
1500
1000
500
0
0
1
2
3
4
5
6
7
Wind speed (m/s)


Rented Bike Count
3500
3000
2500
2000
1500
1000
500
0
0
250
500
750
1000
1250
1500
1750
2000
Visibility (10m)


Rented Bike Count
3500
3000
2500
2000
1500
1000
500
0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Solar Radiation (MJ/m2)

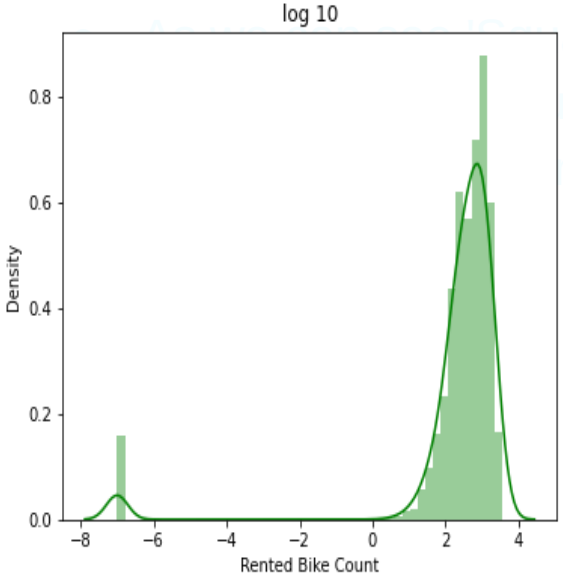


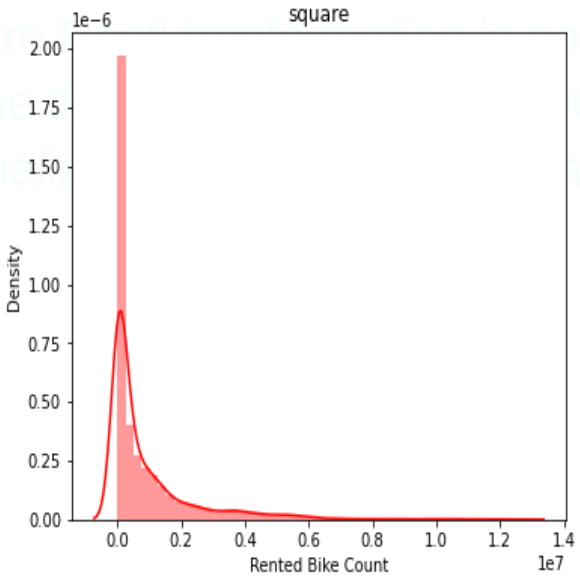


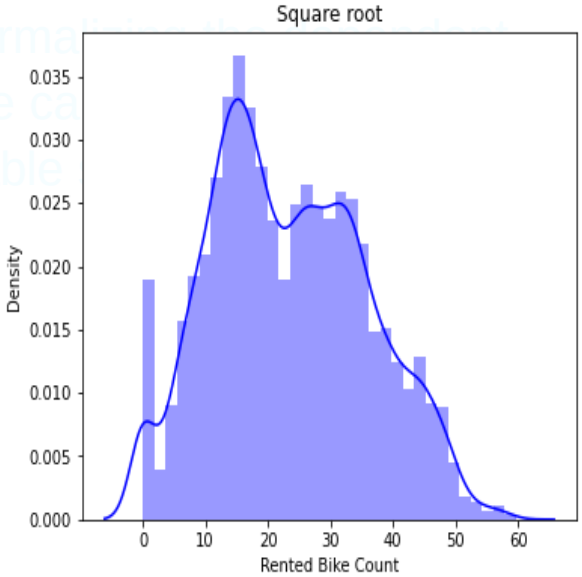
From above 'Regression Plots' we observe **'Temperature', 'Wind_speed' , 'Visibility' ,'Solar Radiation**' this features are positively related with our dependent variable.

And **'Rainfall' , 'Snowfall' , 'Humidity**' these features are negatively related with the dependent var iable or Target variable.

▪Feature Engineering

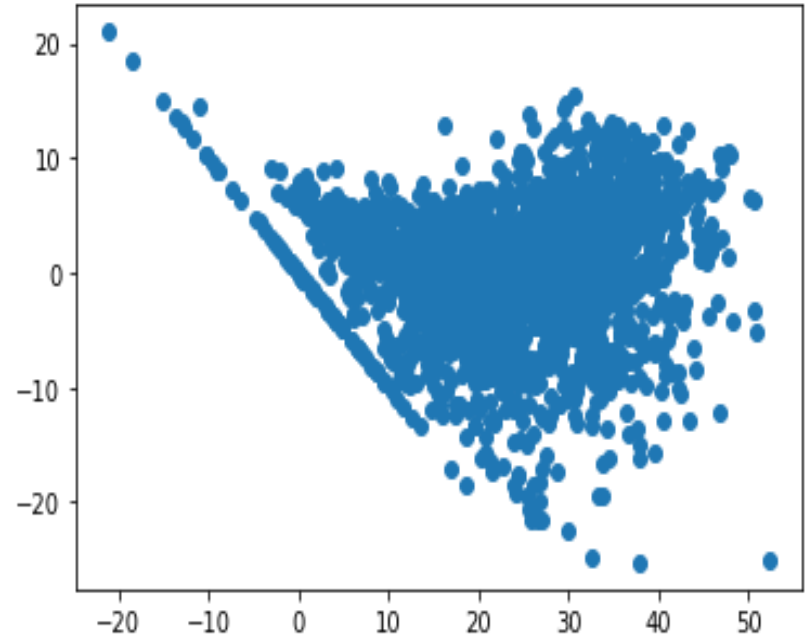


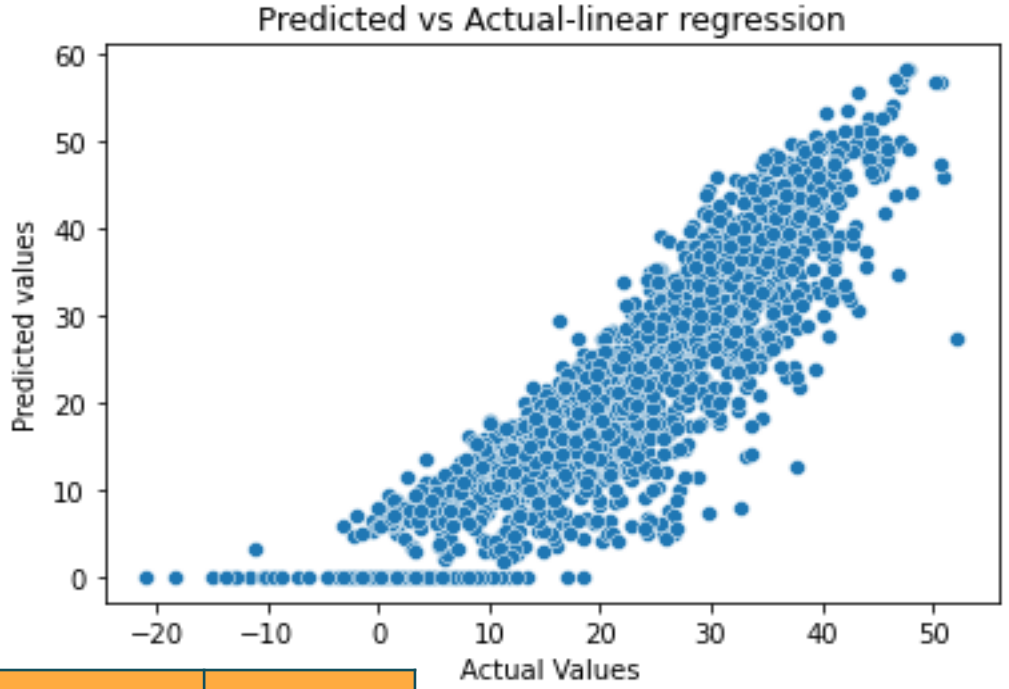




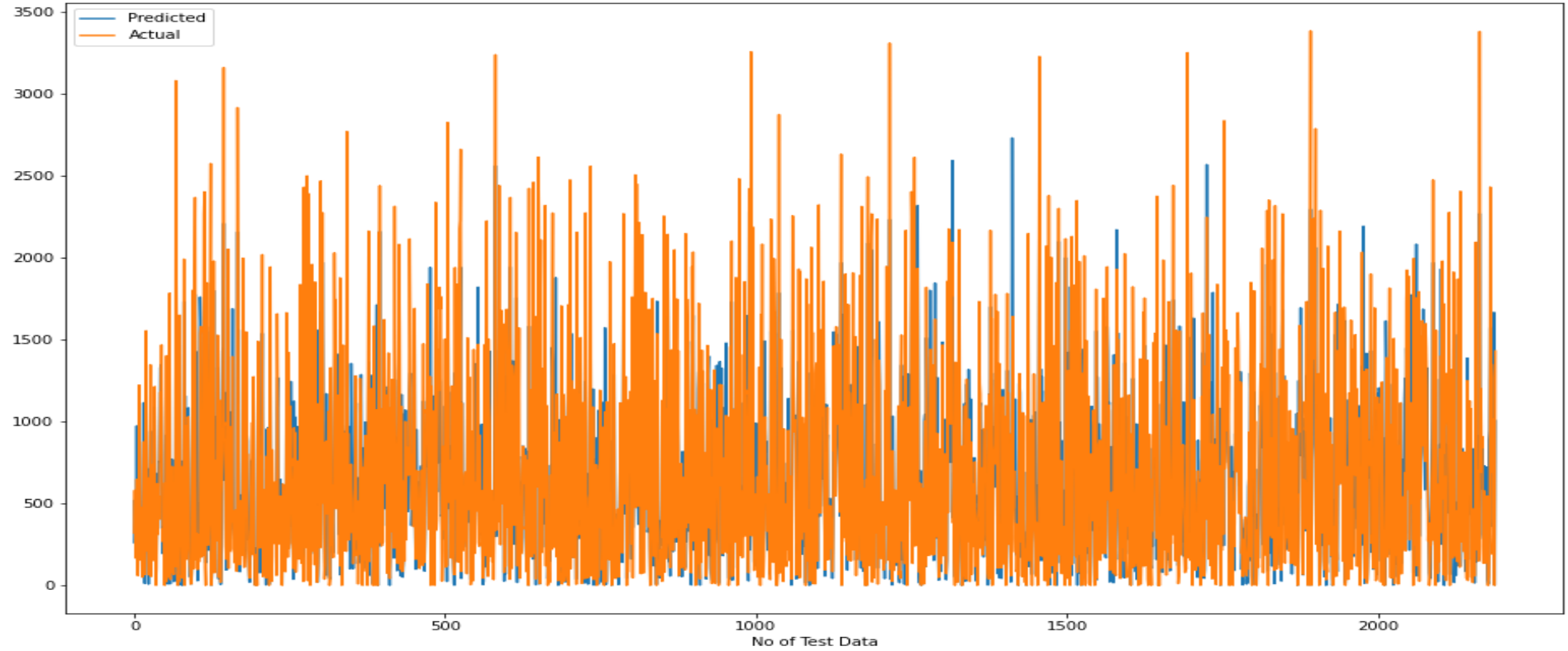
▪We observe **'Square Root'** tranformation is normalizing the dependent variable so we will use this transformation.

Linear Regression

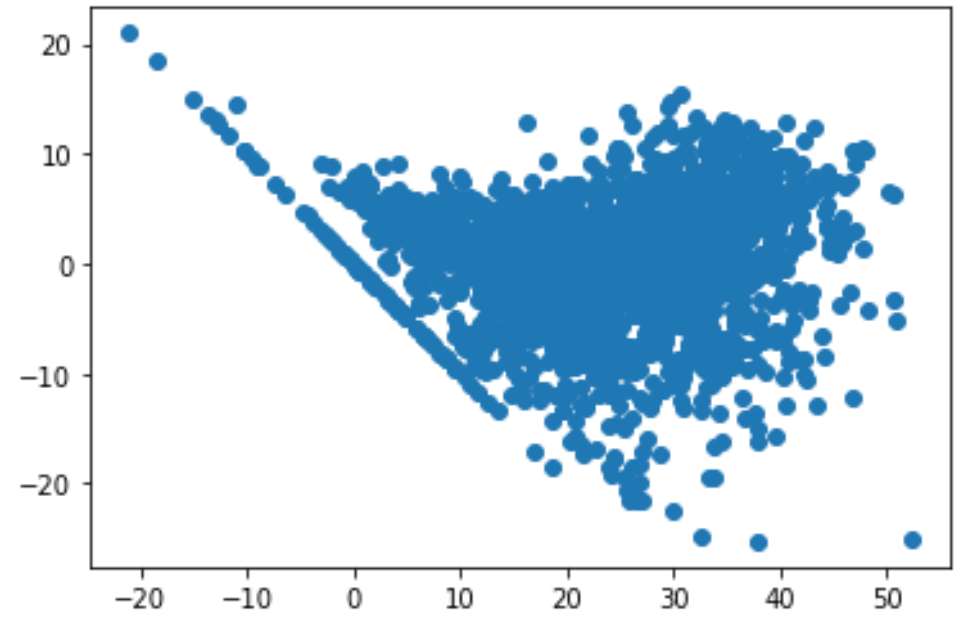


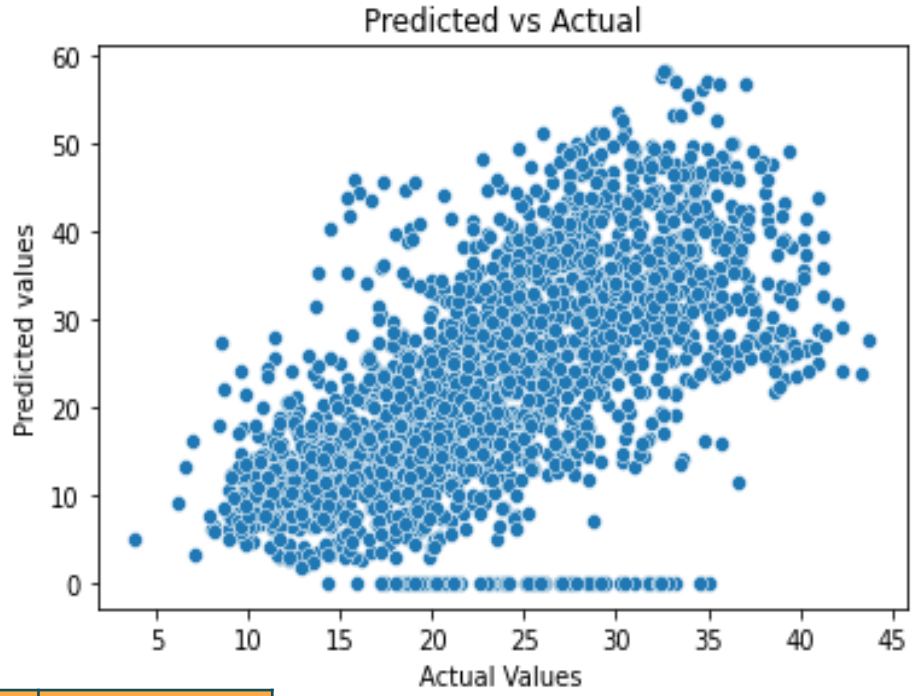


Model	MAE	MSE	RMSE	R2_score	Adjusted R2
Linear regression	4.474	33.275	5.768	0.772	0.77

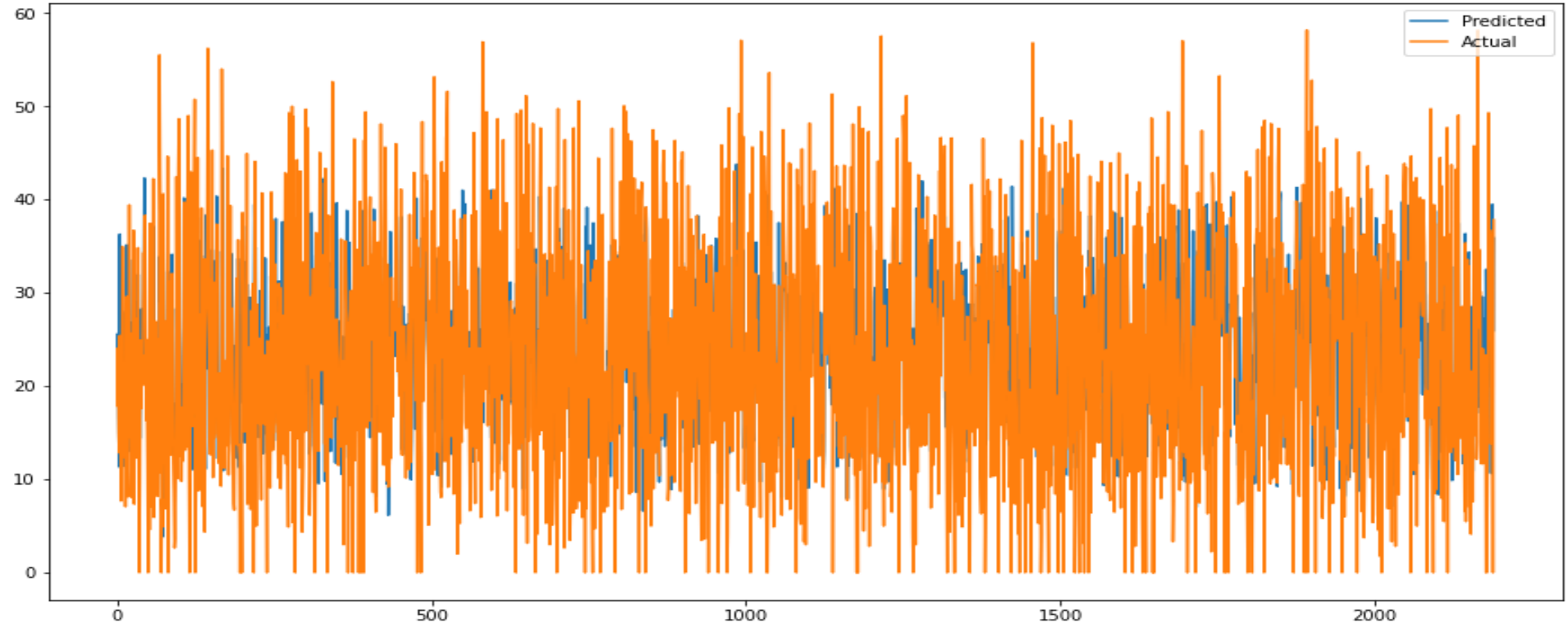

Predicted
Actual
3500
3000
2500
2000
1500
1000
500
0
0
500
1000
1500
2000
No of Test Data

▪Lasso regression

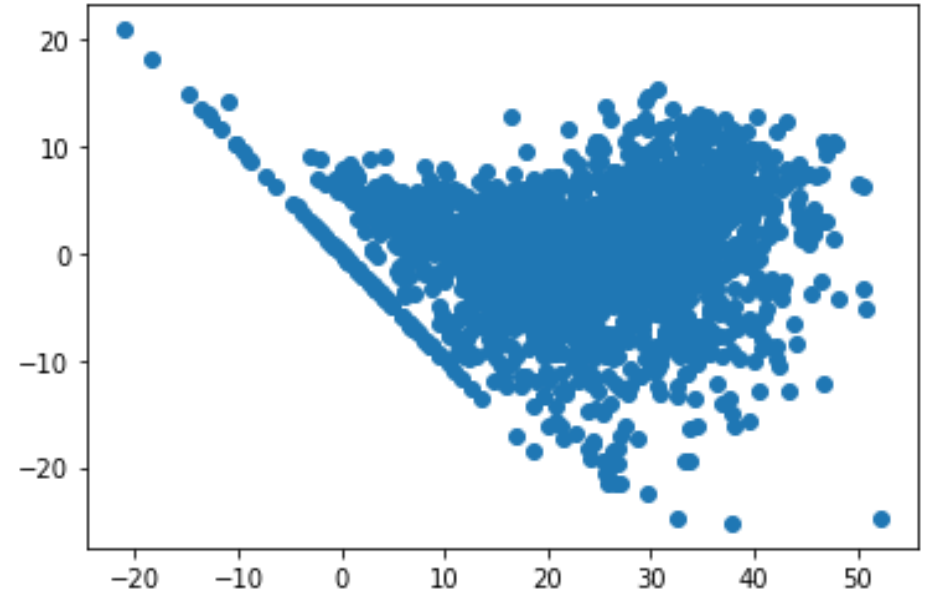


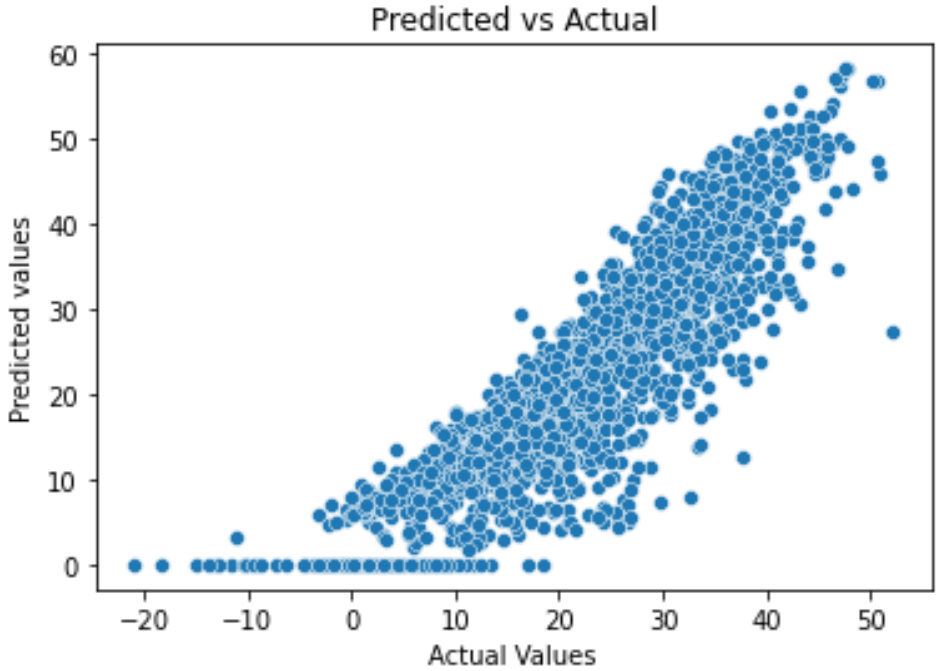


Model	MAE	MSE	RMSE	R2_score	Adjusted R2
Lasso regression	7.456	96.775	9.837	0.387	0.37

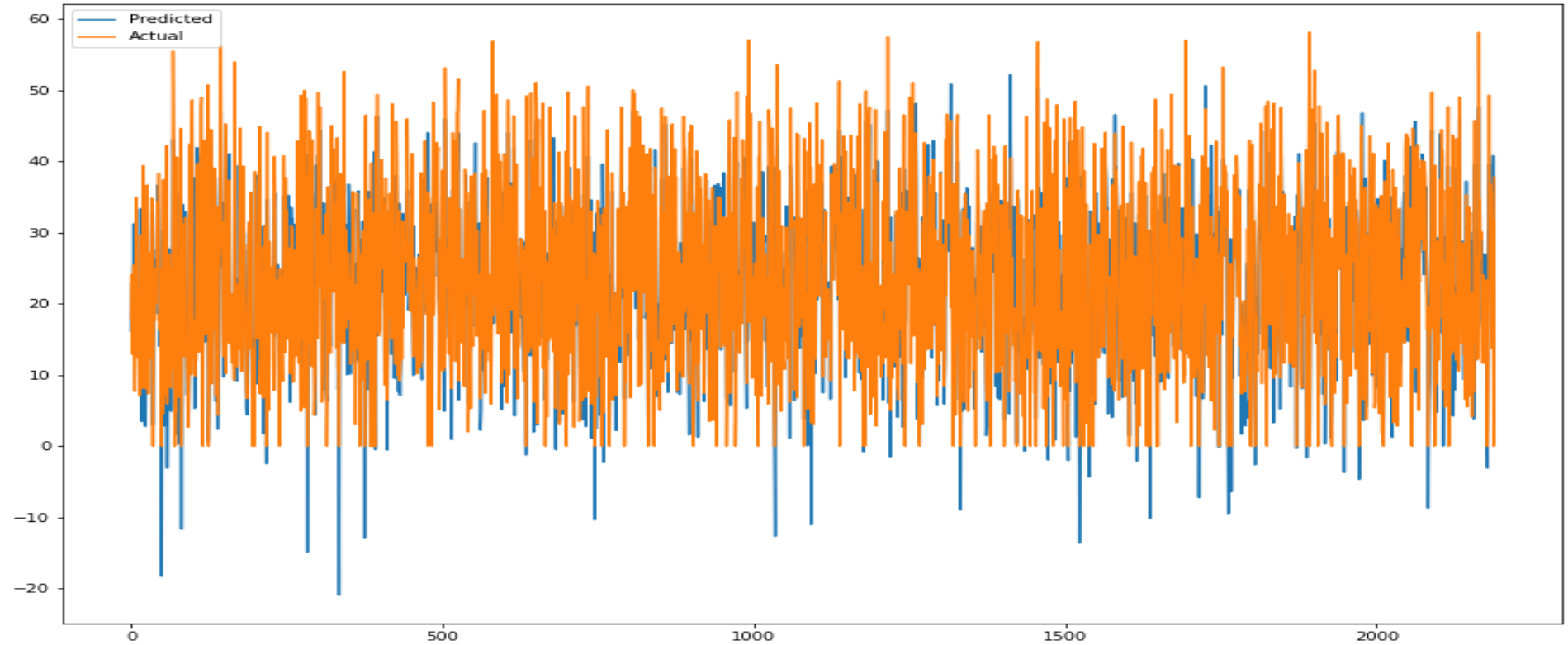

Predicted
Actual
60
50
40
30
20
10
0
0
500
1000
1500
2000

Cross Validation Lasso Regression

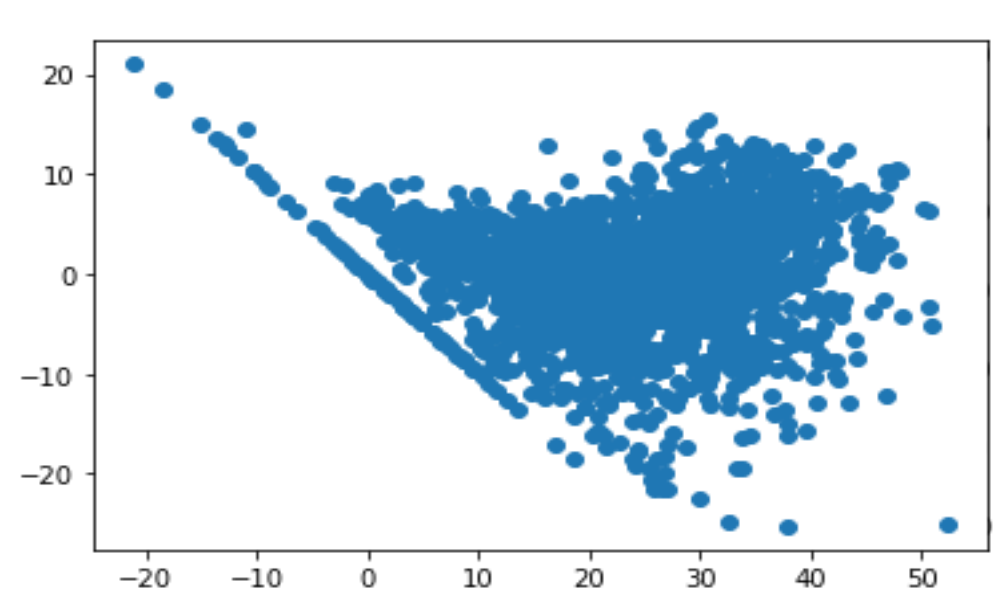


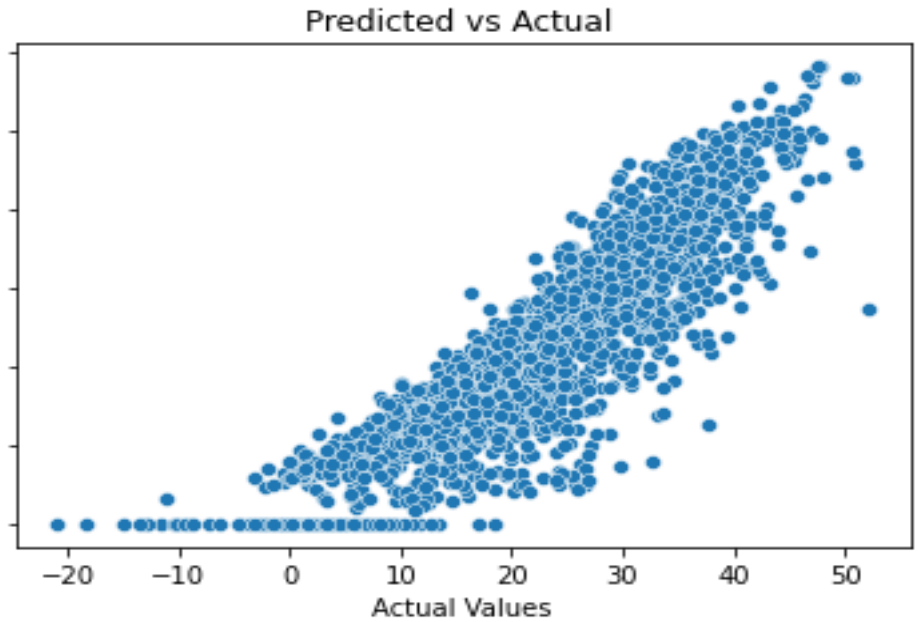


Model	MAE	MSE	RMSE	R2_score	Adjusted R2
Lasso regression Test with cross-validation	4.410	33.276	5.769	0.789	0.78

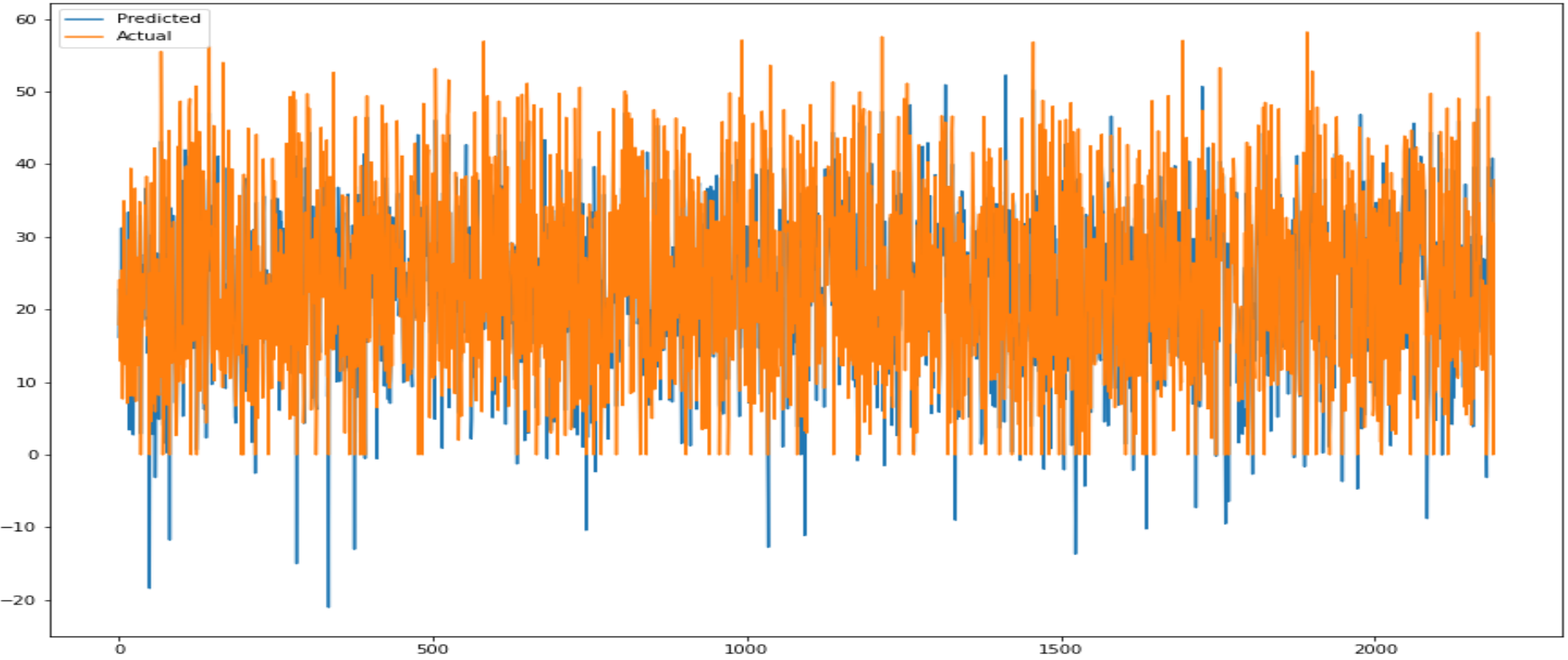

Predicted
Actual
60
50
40
30
20
10
0
−10
−20
0
500
1000
1500
2000

▪Ridge Regression

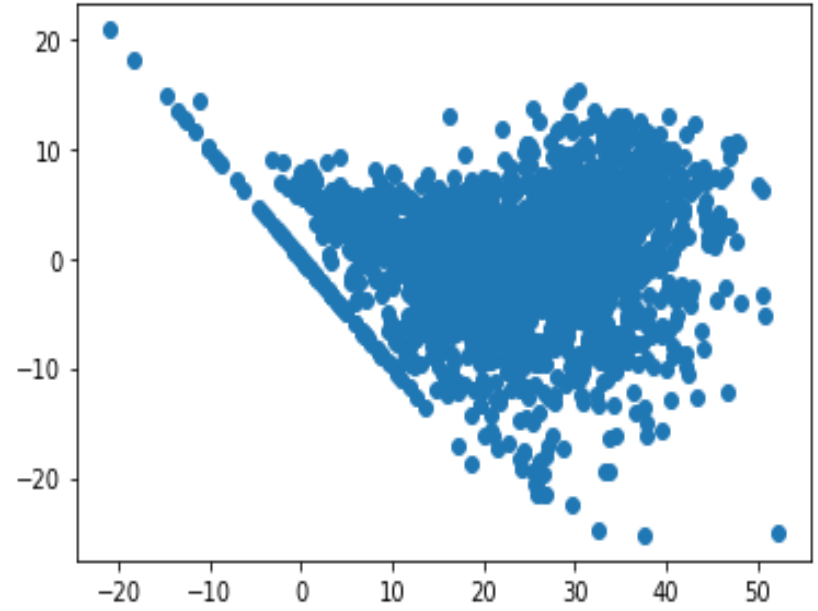


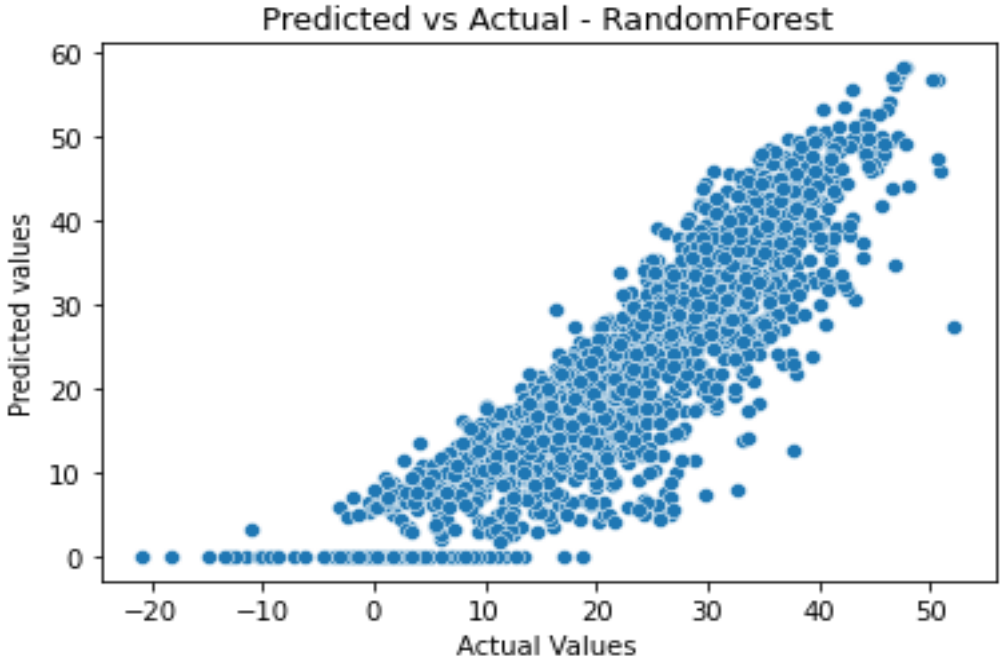


Model	MAE	MSE	RMSE	R2_score	Adjusted R2
Ridge regression	4.410	33.277	5.769	0.789	0.78

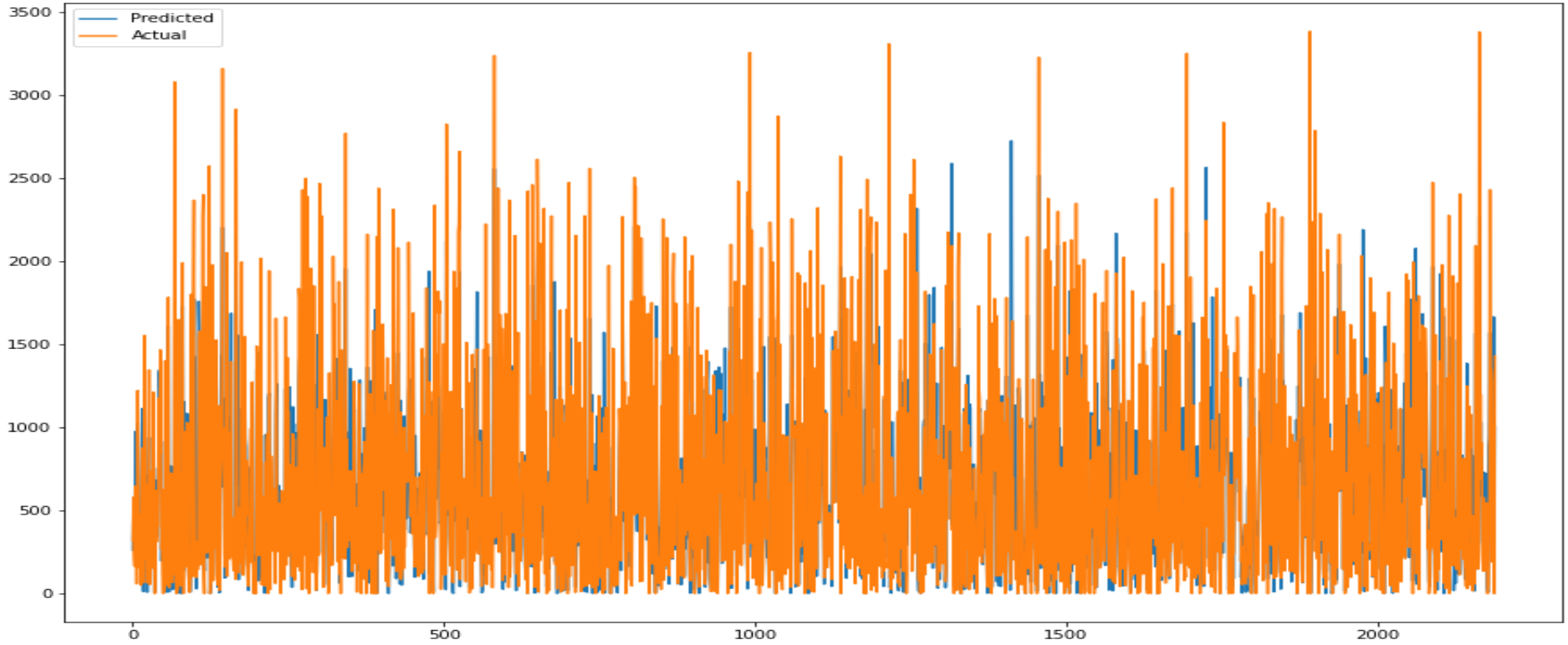

Predicted
Actual
60
50
40
30
20
10
0
−10
−20
0
500
1000
1500
2000

Cross Validation Ridge Regression

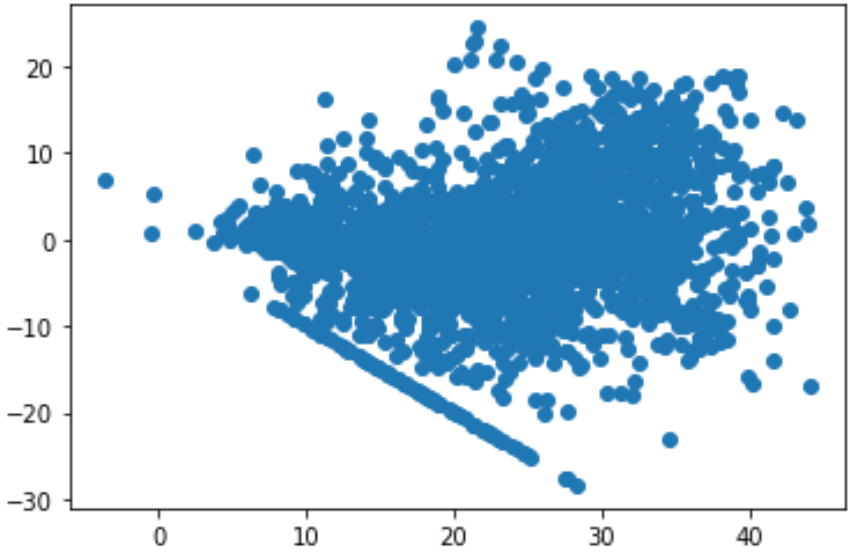


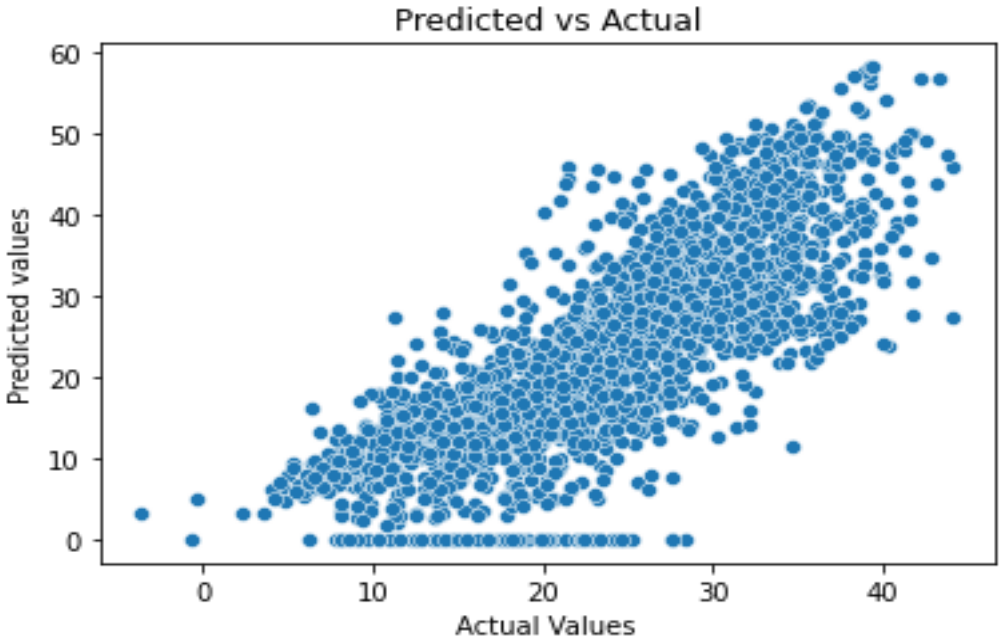


Model	MAE	MSE	RMSE	R2_score	Adjusted R2
Ridge regression Test with cross-validation	4.413	33.291	5.770	0.789	0.78

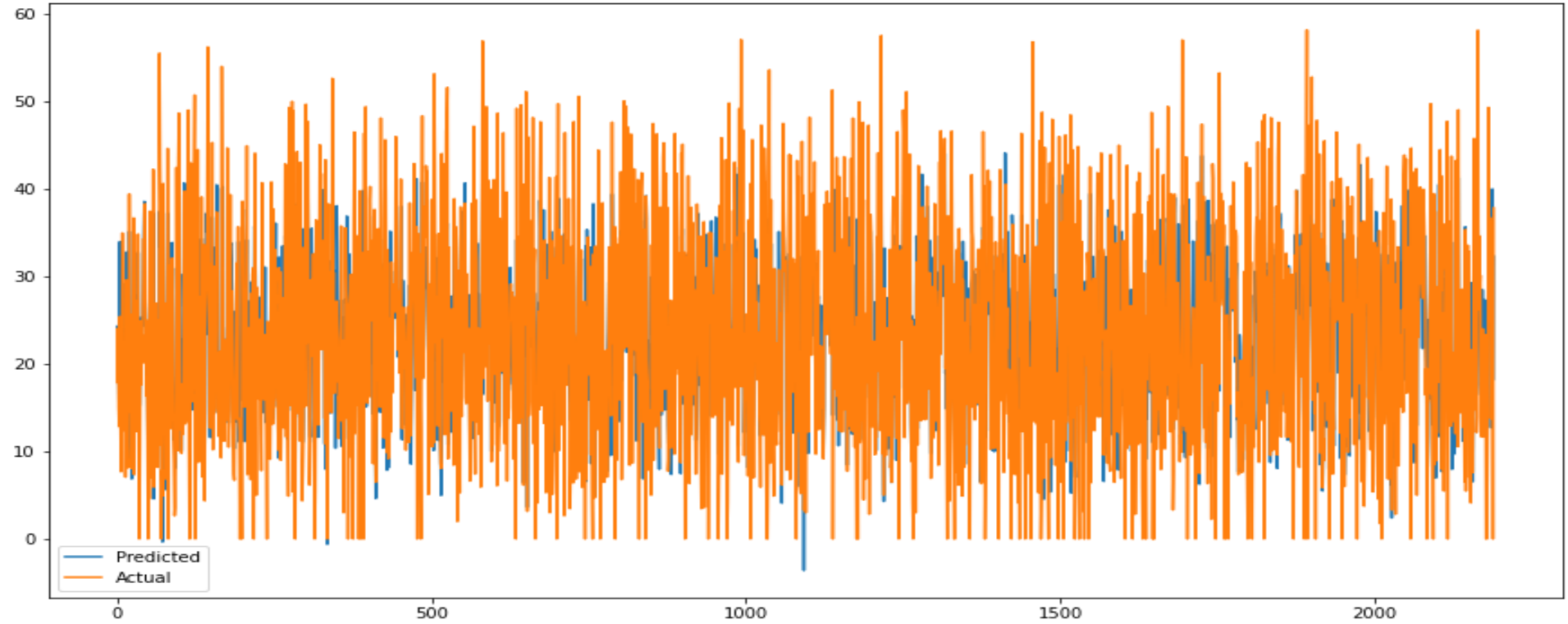

Predicted
Actual
3500
3000
2500
2000
1500
1000
500
0
0
500
1000
1500
2000

▪Elastic Net Regression

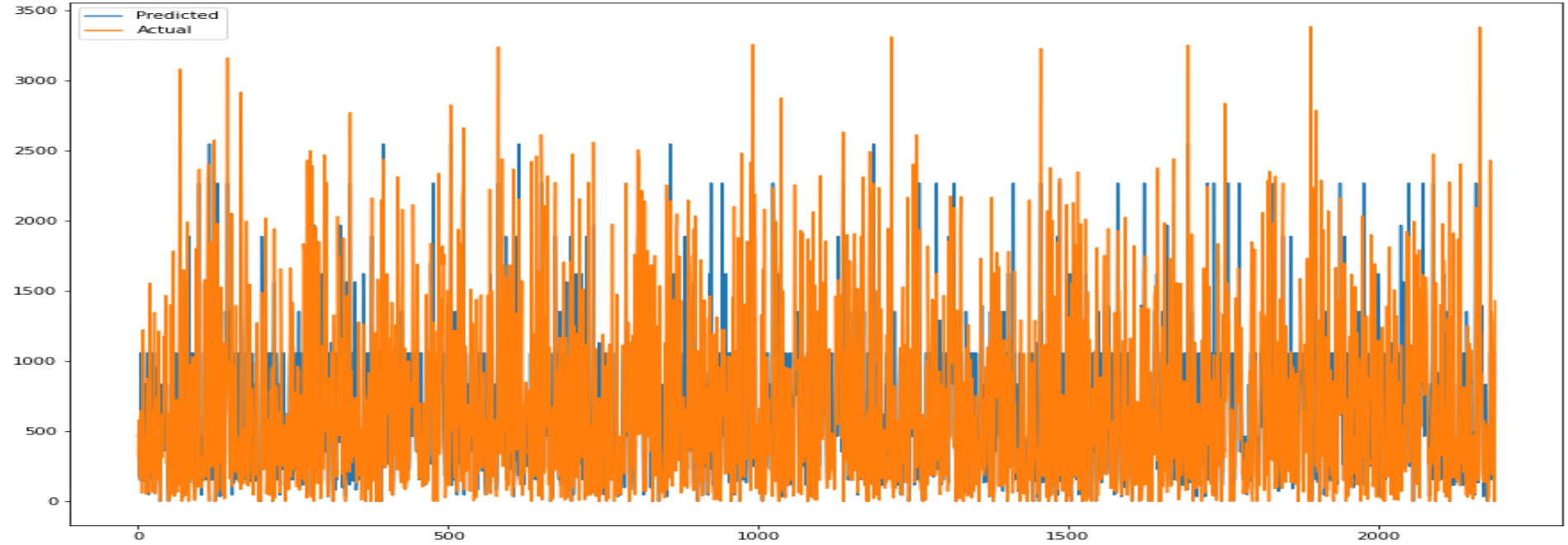




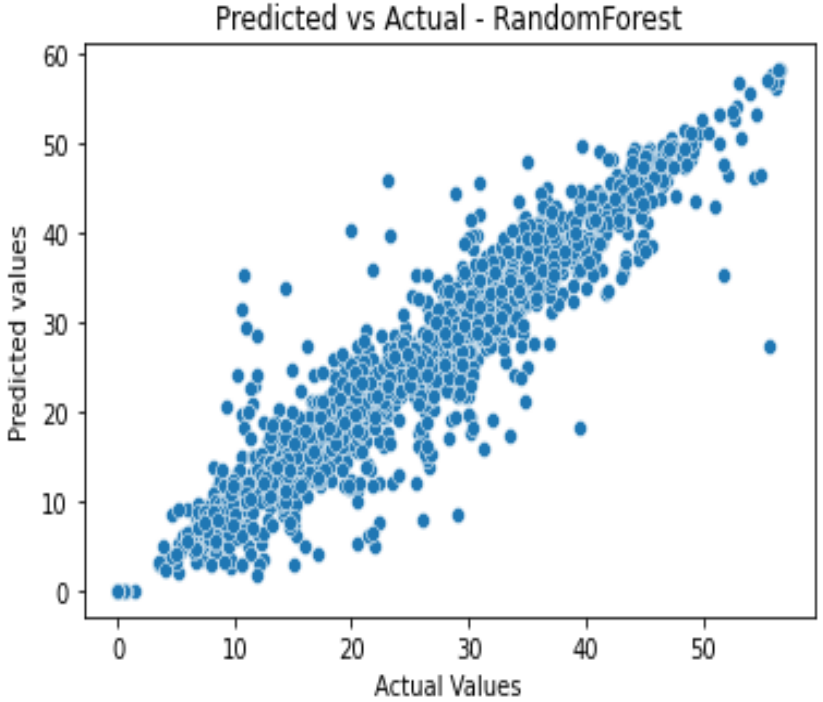
Model	MAE	MSE	RMSE	R2_score	Adjusted R2
Elastic net regression Test	5.874	59.451	7.710	0.624	0.62

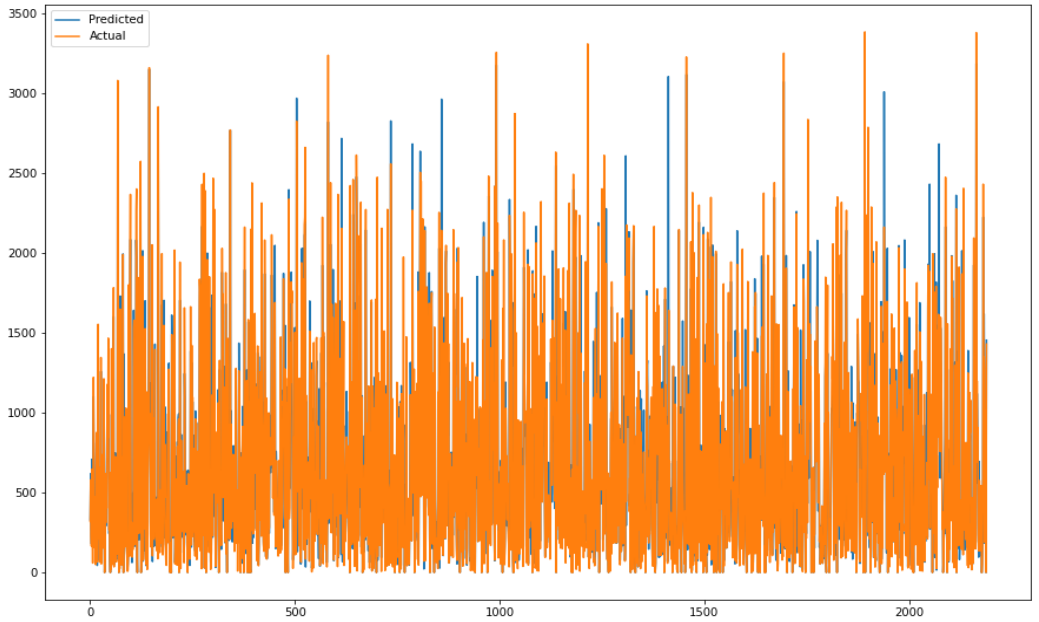

60
50
40
30
20
10
0
Predicted
Actual
0
500
1000
1500
2000

▪Decision Tree



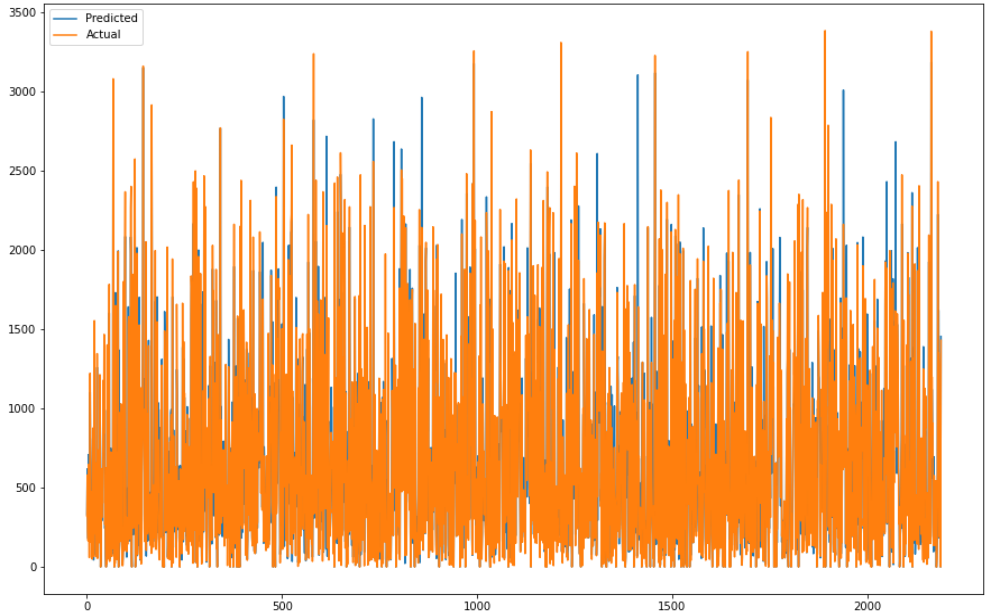
▪Random Forest

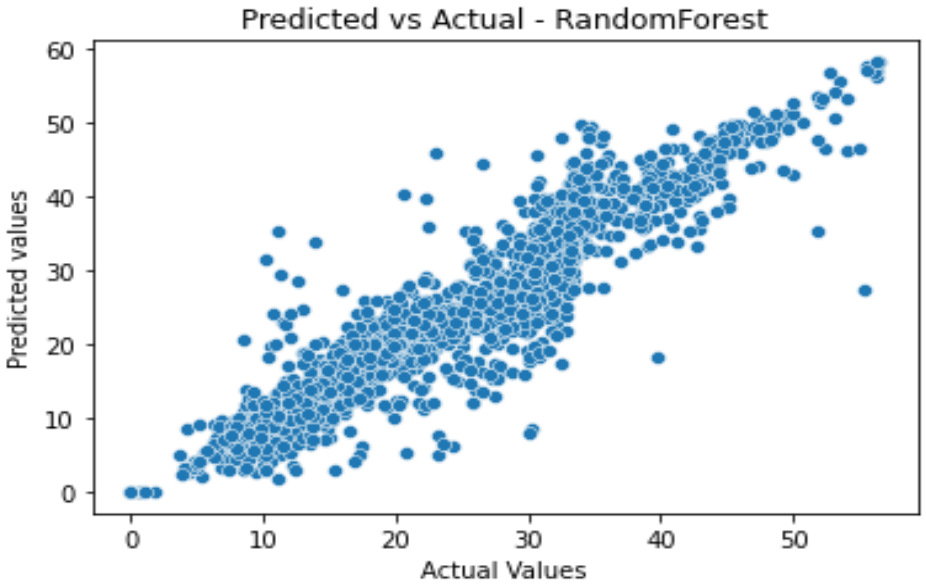




Model	MAE	MSE	RMSE	R2_score	Adjusted R2
Random forest regression	2.210	12.634	3.554	0.920	0.92

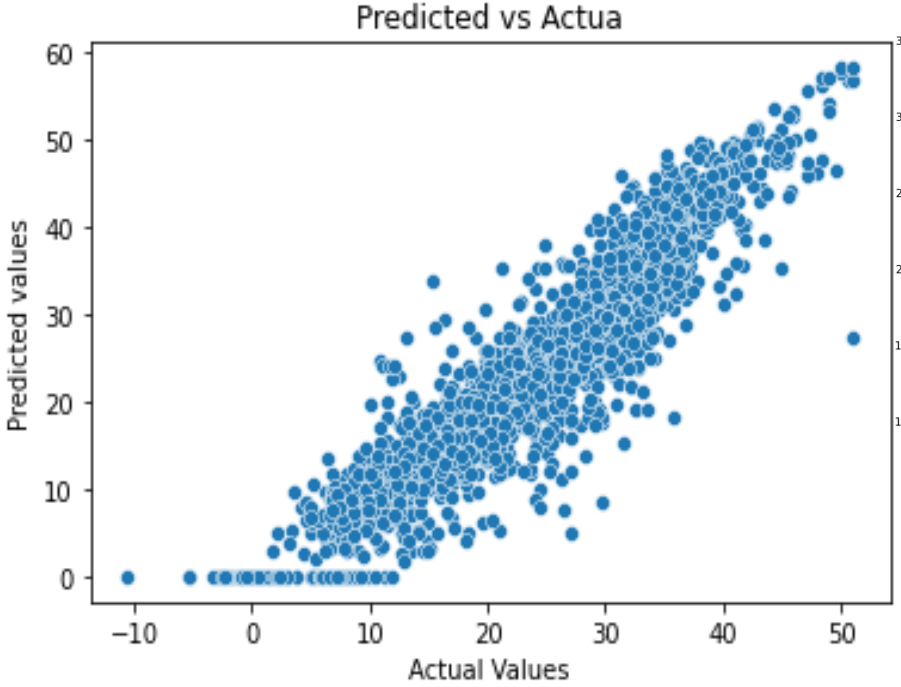
Cross Validation Random Forest

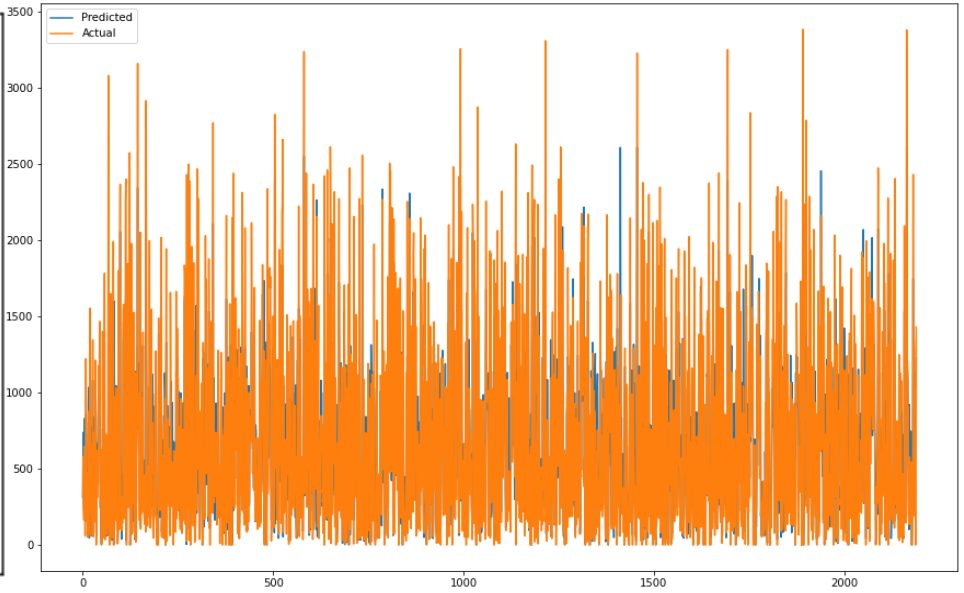




Model	MAE	MSE	RMSE	R2_score	Adjusted R2
Random forest regression with gridSearchCV	2.785	17.680	4.205	0.888	0.89

▪XGBoost





Model	MAE	MSE	RMSE	R2_score	Adjusted R2
XGBoost regression	3.508	21.640	4.652	0.863	0.86

Conclusion

- No overfitting is seen, as we can see the models are performing well with the test data with good results.
- • Linear Regression, Lasso , Ridge and Elastic net performed moderately and gave an R2 score of 77, 78, 78 and 62% respectively for test dataset.
- • Random forest Regressor, Random forest Regressor with gridsearchCV and XGB Regressor gives the highest R2 score of 92%, 91% and 86% respectively for test dataset.
- • Feature Importance value for Random Forest and Gradient Boost were different.
- Finally, we can say that Random Forest model performed best out of all the models.

References

- https://towardsdatascience.com/
- https://www.analyticsvidhya.com/
- https://www.geeksforgeeks.org/python-data-visualization-tutorial/

THANK YOU