Capstone Project-2

Seoul Bike Sharing Demand Prediction

ML SUPERVISED REGRESSION

VAIBHAV JAIN
NICKY GHAI

Contents

- Problem Statement
- Data Summary
- Data Analysis
- Analysis Details
- Feature Selection
- Data Preparation
- Implementing Various Regression Algorithms
- Challenges
- Conclusions

Problem Statements

- **Prediction of bike count required at each hour.**
- **Reduce waiting time of public.**

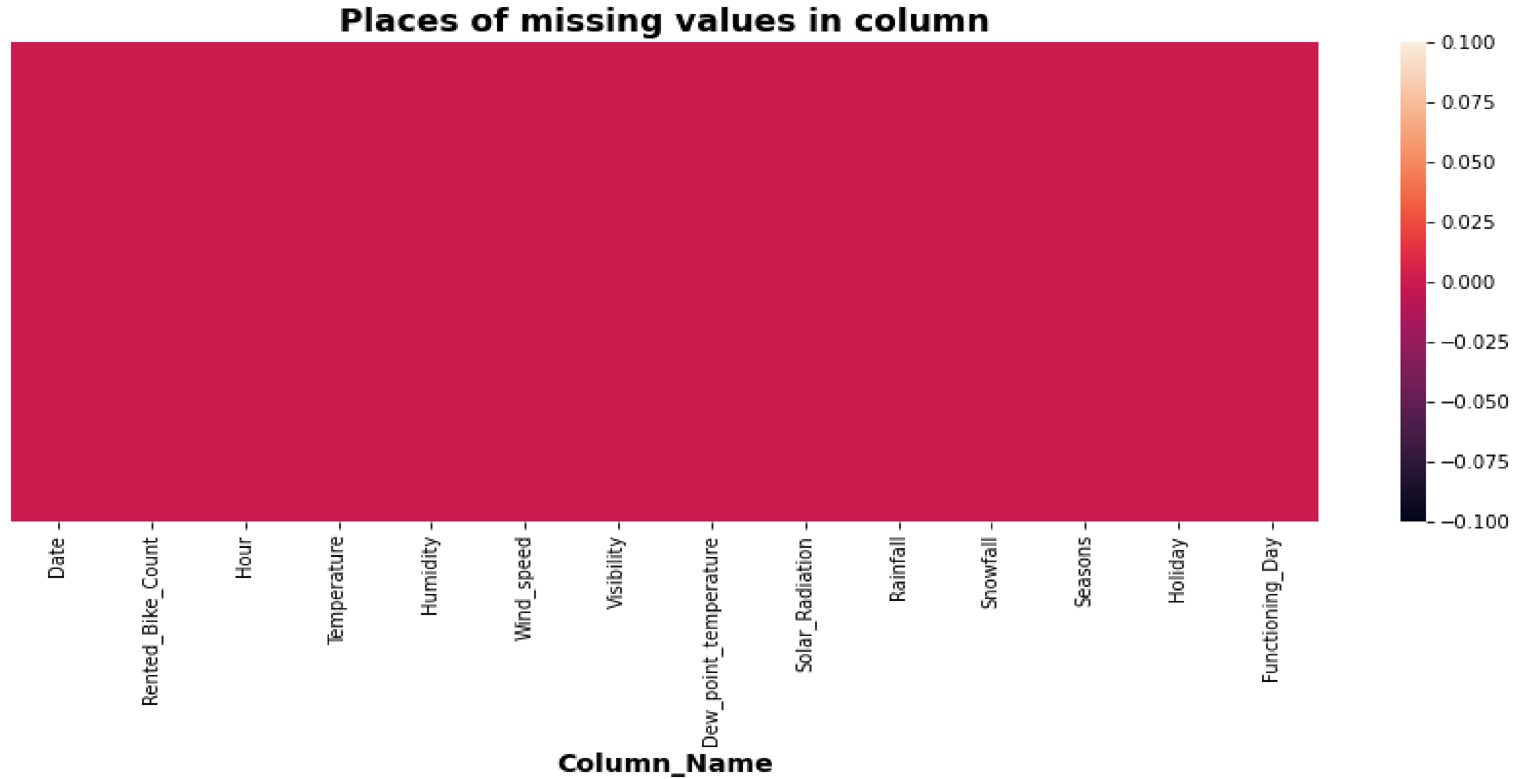
Data Summary

- Date : Year-Month-Day
- Rented Bike Count - Count of bikes rented at each hour
- Hour - Hour of the 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/No Holiday
- Functional Day - NoFunc(Non Functional Hrs),Fun(Functional Hrs)

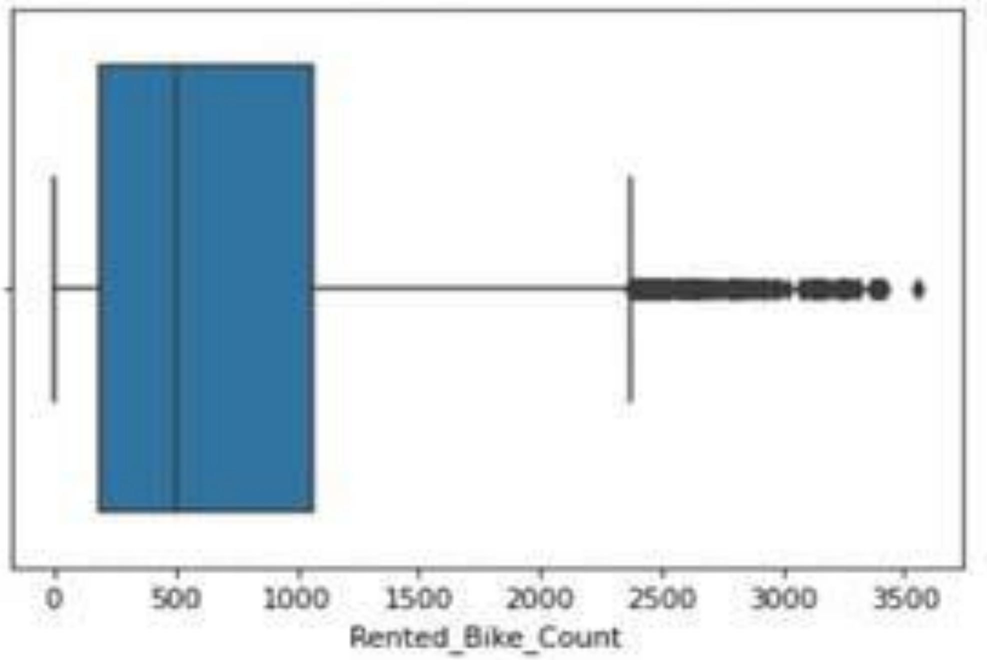
Basic Data Exploration

- The dataset has 8760 rows and 14 features(columns).
- Three categorical features ‘Seasons’, ‘Holiday’, & ‘Functioning Day’.
- One Datetime[ns] features ‘Date’.
- Outliers present only in dependent variable.
- No Missing Values.
- No Duplicated values.
- No null values.

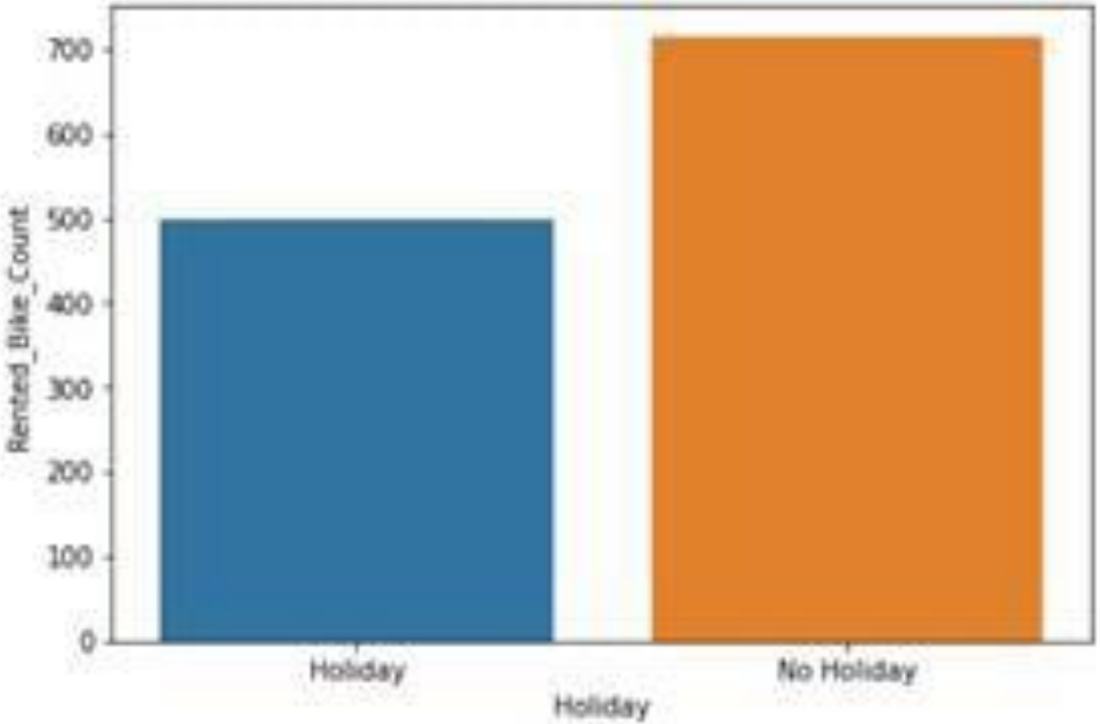
Missing Values



Checking for the outlier in our dependent variable



Division on rented bike on holiday and non holiday days



Distribution on rented bike according to different seasons

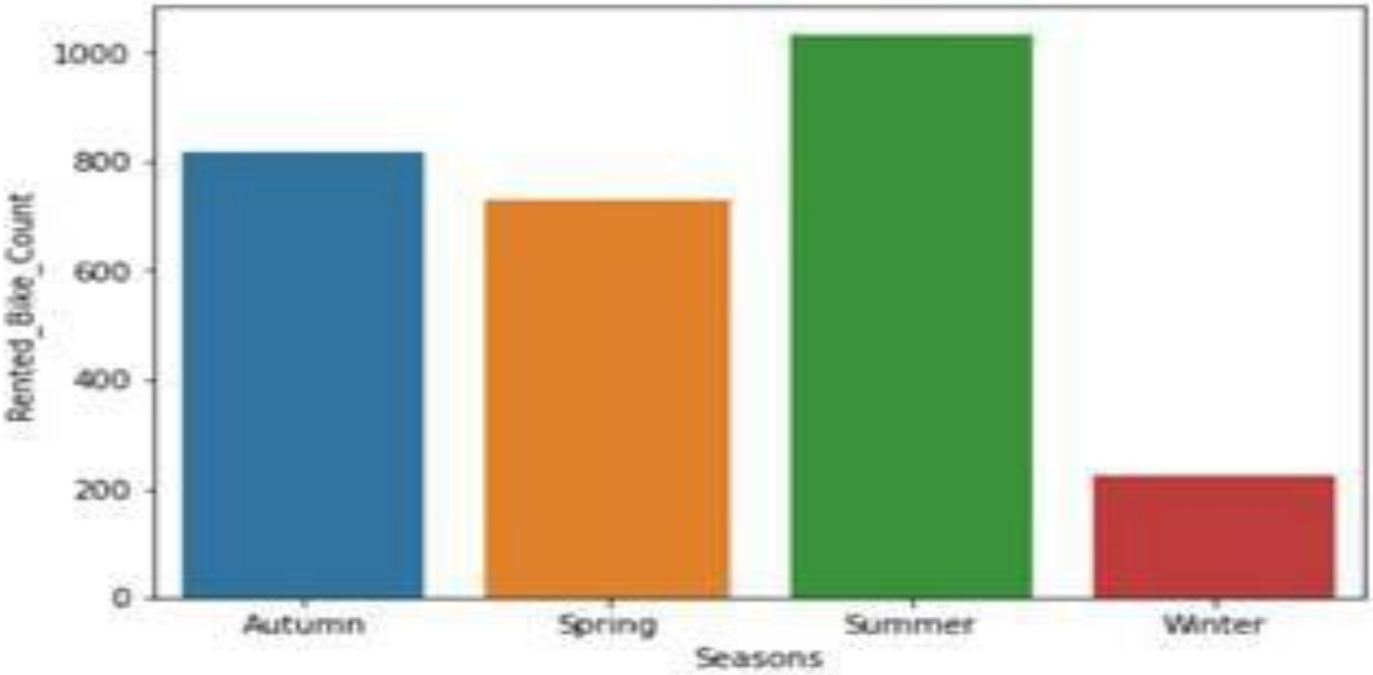
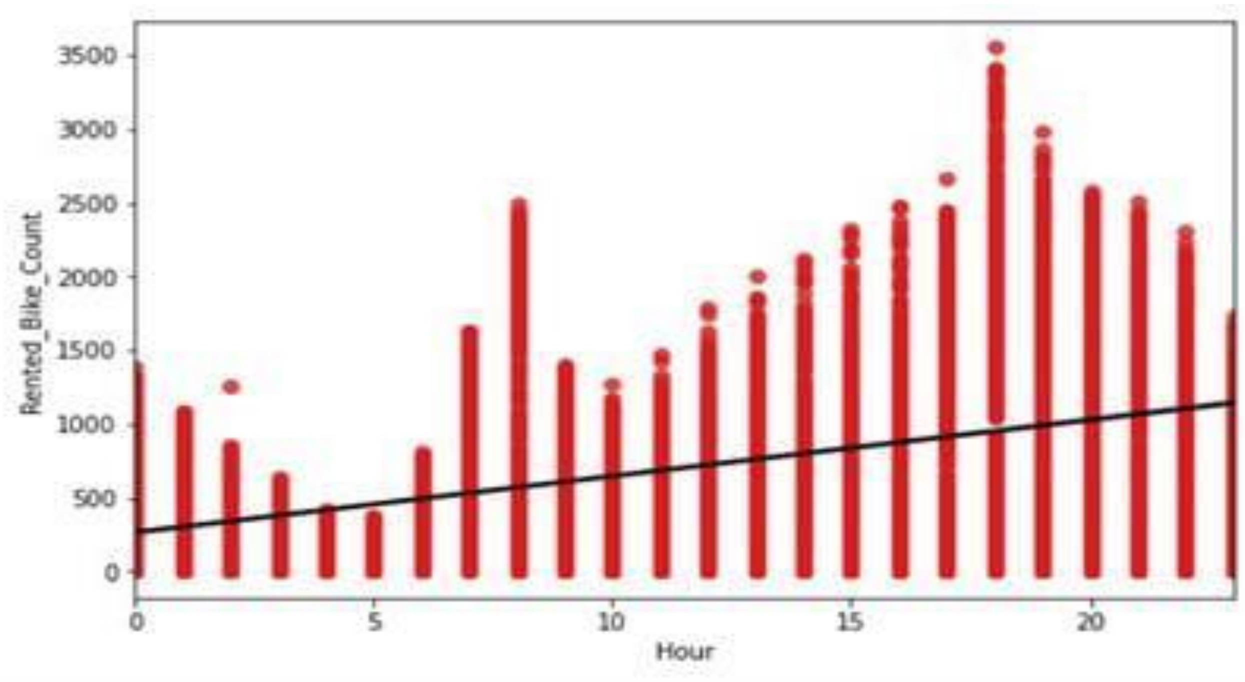
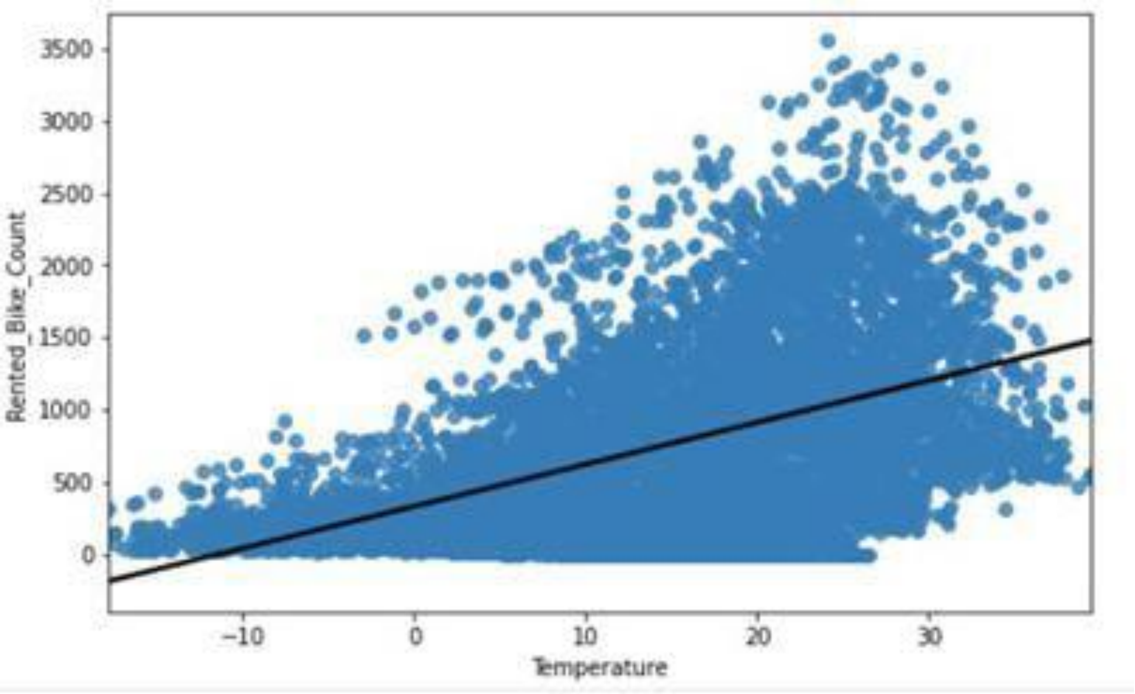


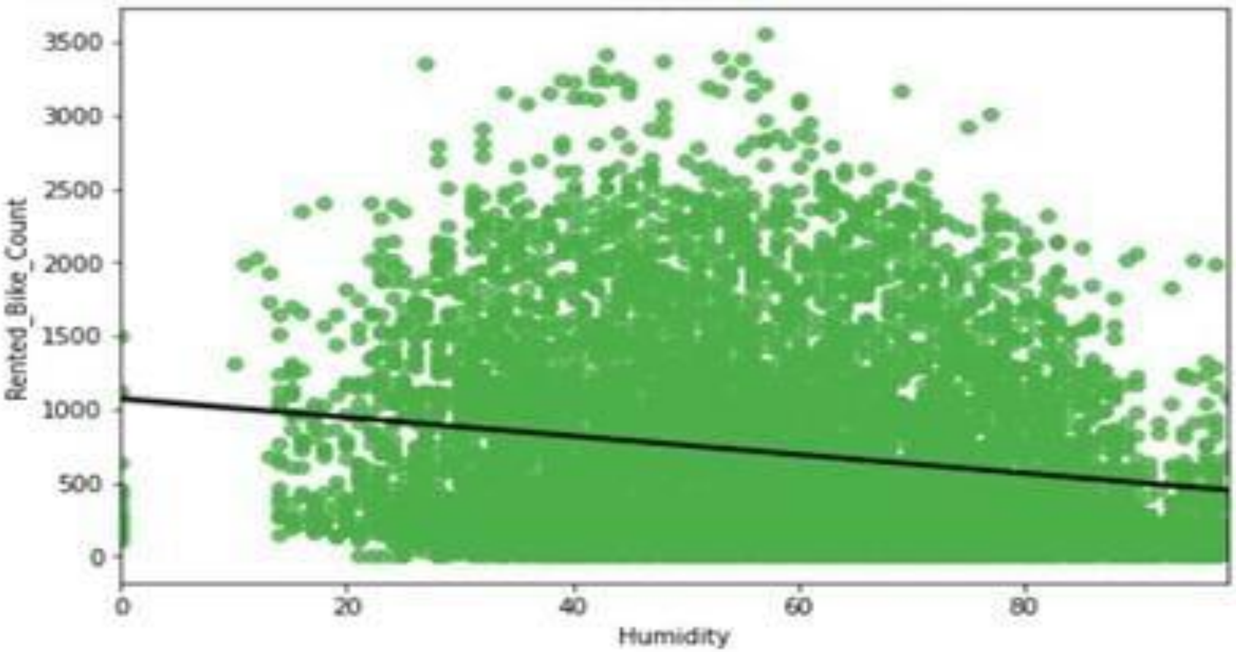
Chart showing distribution of Rented bike count per hour



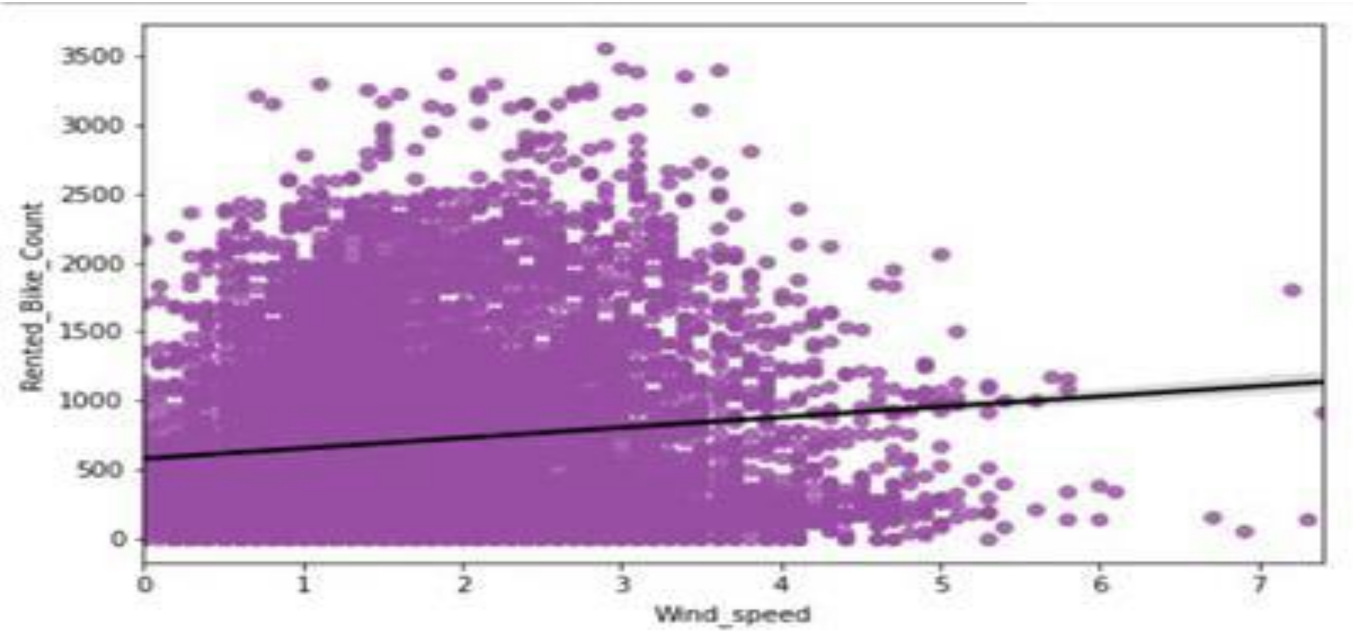
Relation of our dependent variable with Temperature



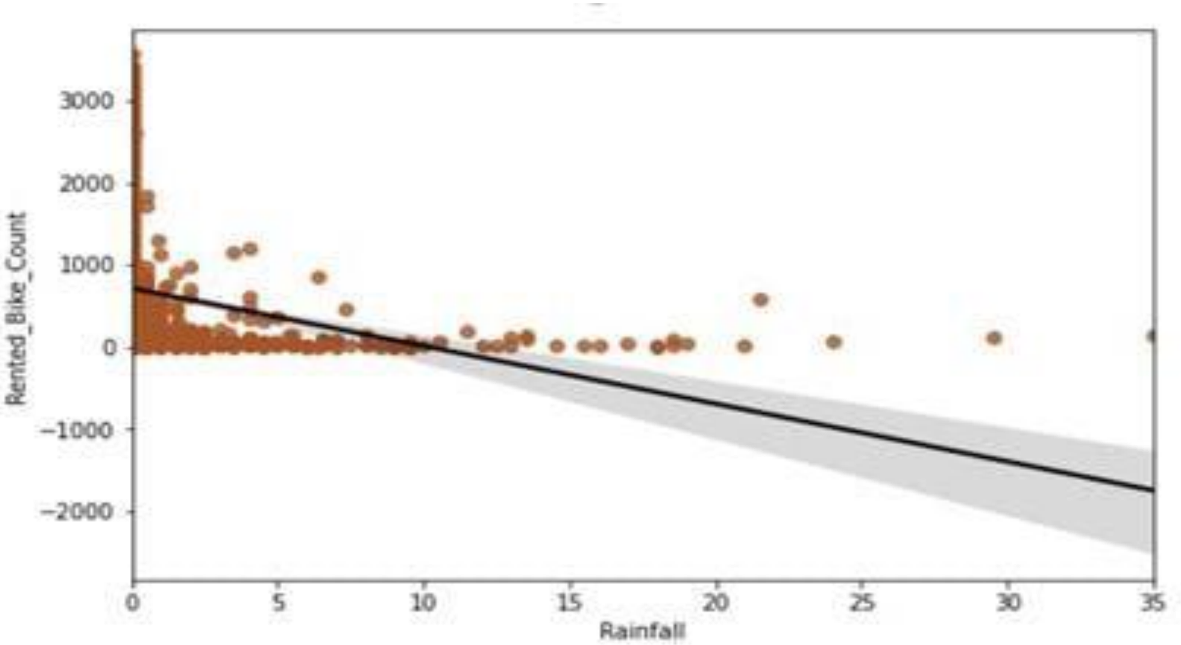
Relation of our dependent variable with Humidity



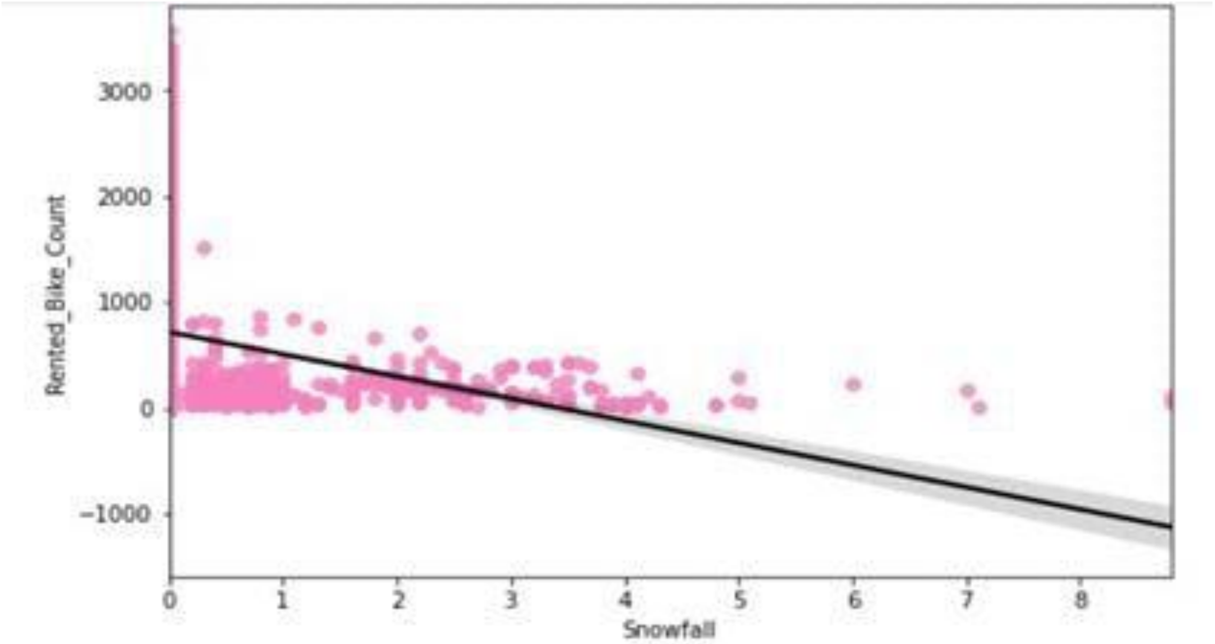
Relation of our dependent variable with wind speed



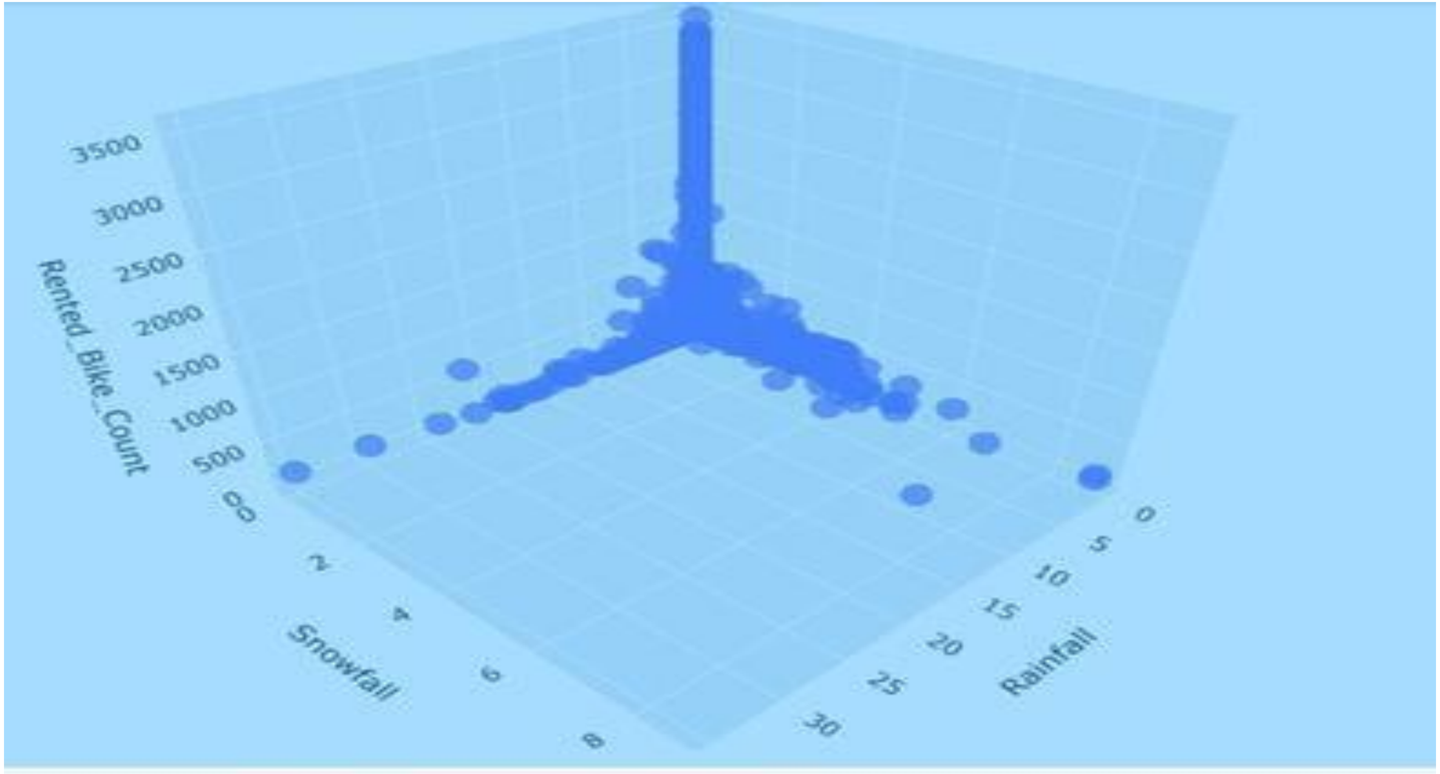
Relation of our dependent variable with Rainfall



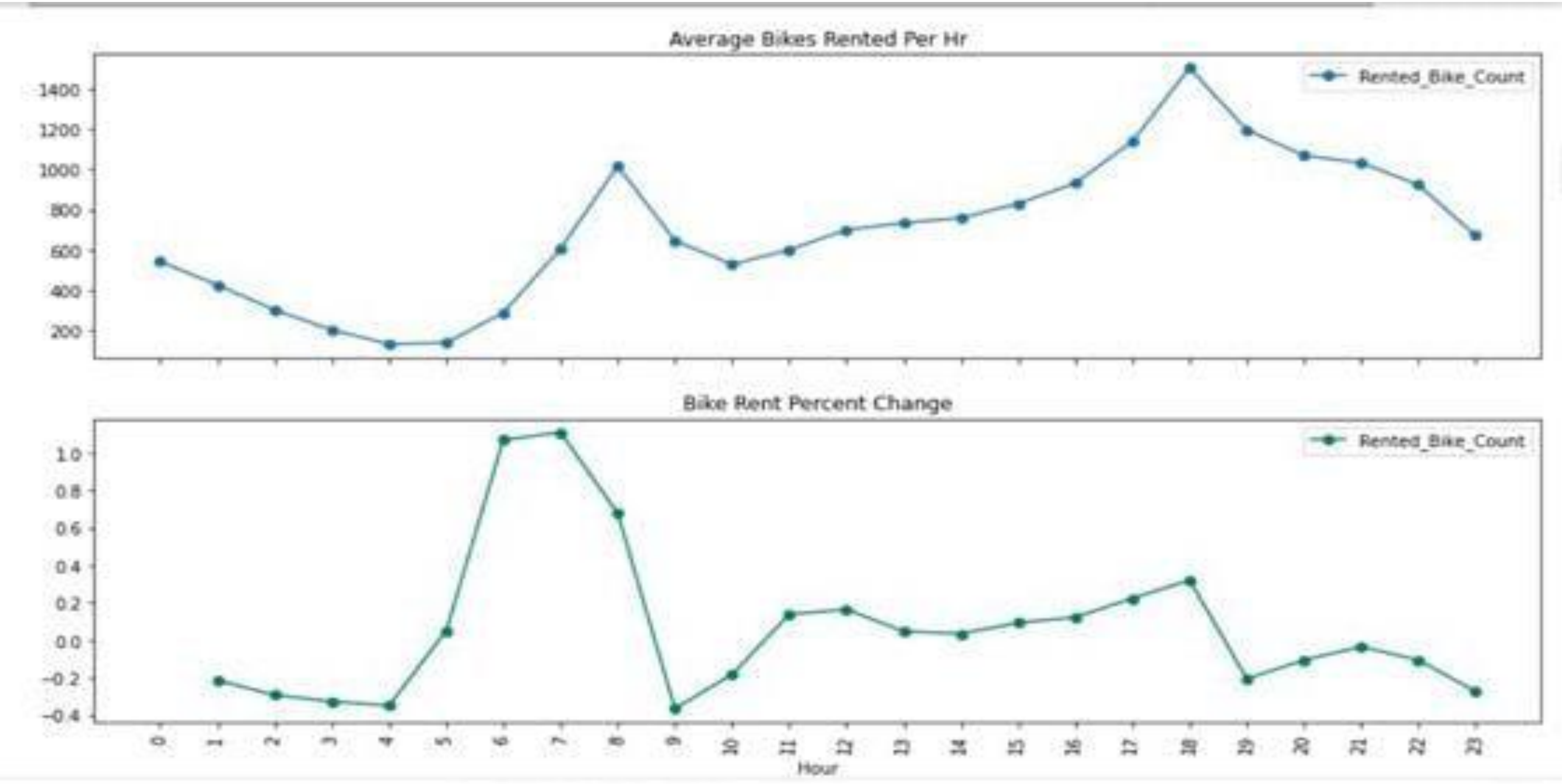
Relation of our dependent variable with Snowfall



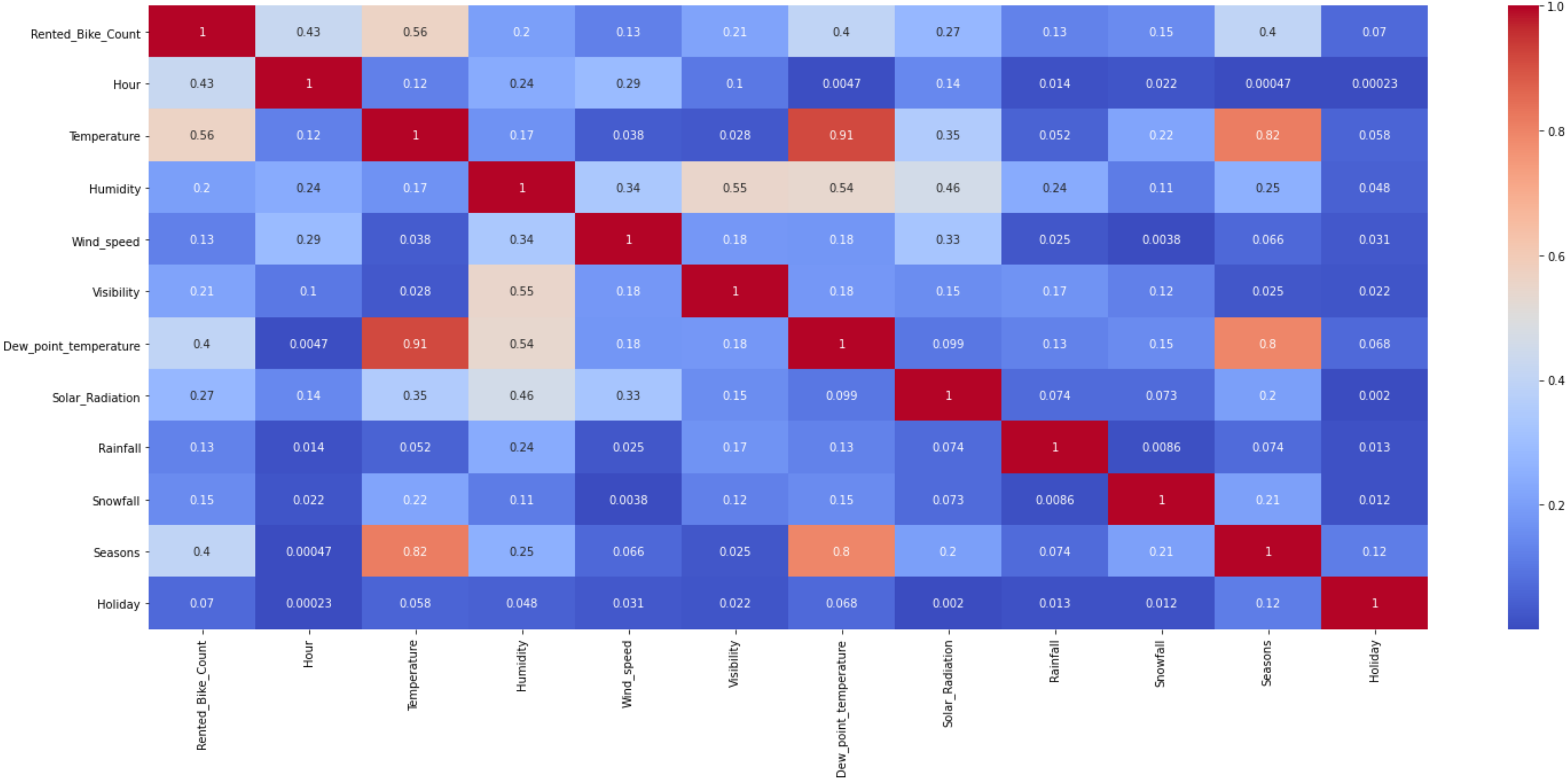
3-d plot showing relation between Snowfall , Rainfall and Rented bike count



Per hour distribution

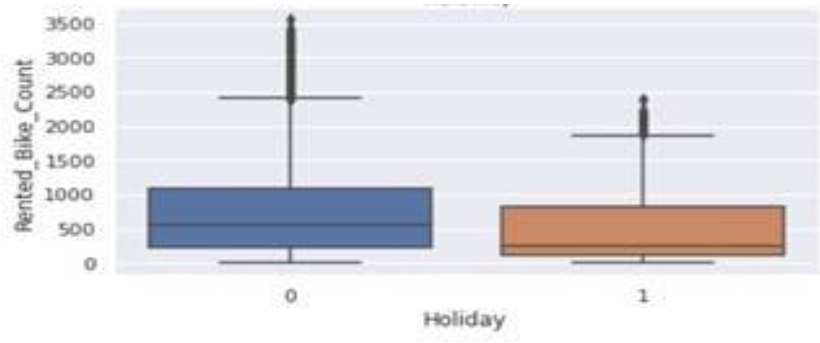


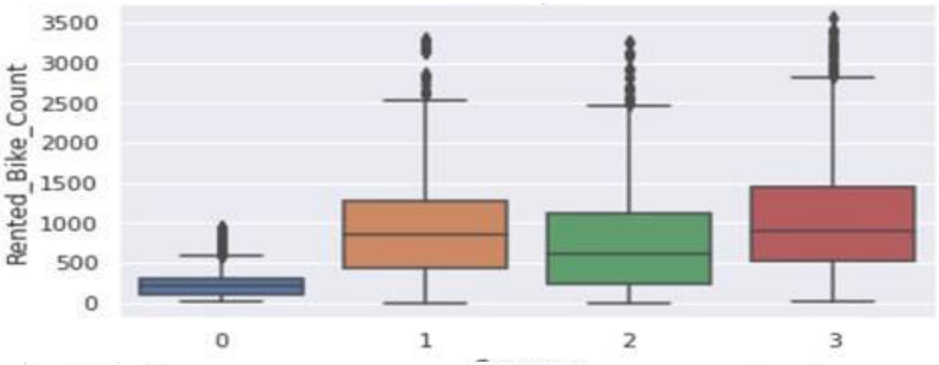
Correlation between different factors

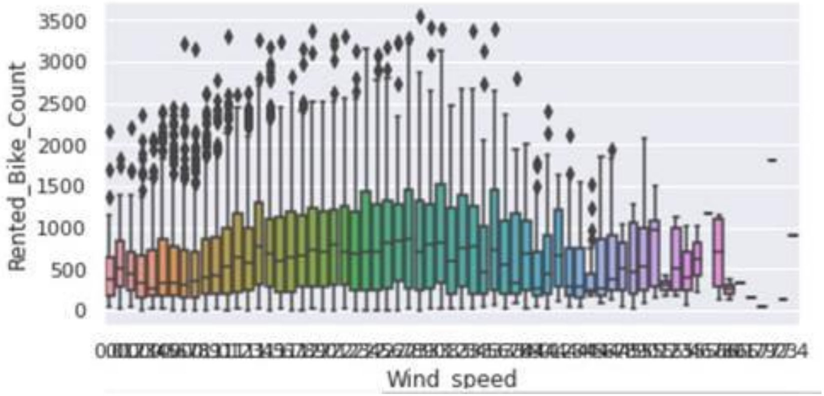


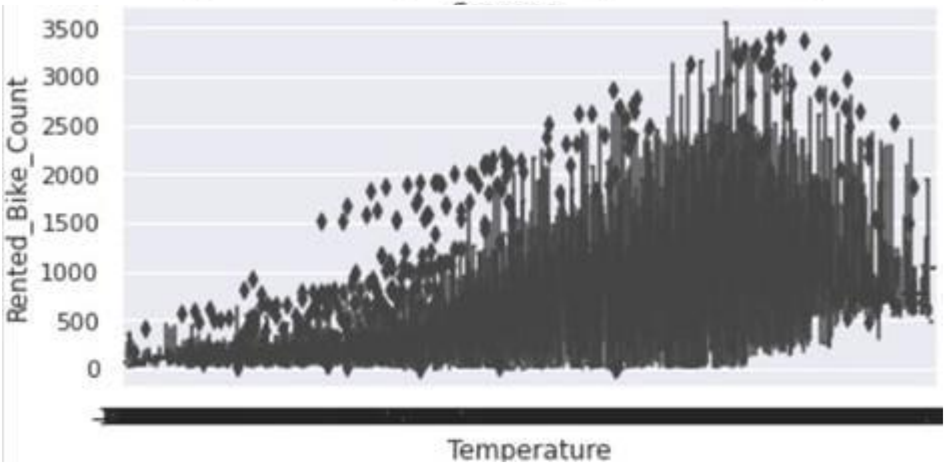
	Rented_Bike_Count	Hour	Temperature	Humidity	Wind_speed	Visibility	Dew_point_temperature	Solar_Radiation	Rainfall	Snowfall	Seasons	Holiday
Rented_Bike_Count	1	0.43	0.56	0.2	0.13	0.21	0.4	0.27	0.13	0.15	0.4	0.07
Hour	0.43	1	0.12	0.24	0.29	0.1	0.0047	0.14	0.014	0.022	0.00047	0.00023
Temperature	0.56	0.12	1	0.17	0.038	0.028	0.91	0.35	0.052	0.22	0.82	0.058
Humidity	0.2	0.24	0.17	1	0.34	0.55	0.54	0.46	0.24	0.11	0.25	0.048
Wind_speed	0.13	0.29	0.038	0.34	1	0.18	0.18	0.33	0.025	0.0038	0.066	0.031
Visibility	0.21	0.1	0.028	0.55	0.18	1	0.18	0.15	0.17	0.12	0.025	0.022
Dew_point_temperature	0.4	0.0047	0.91	0.54	0.18	0.18	1	0.099	0.13	0.15	0.8	0.068
Solar_Radiation	0.27	0.14	0.35	0.46	0.33	0.15	0.099	1	0.074	0.073	0.2	0.002
Rainfall	0.13	0.014	0.052	0.24	0.025	0.17	0.13	0.074	1	0.0086	0.074	0.013
Snowfall	0.15	0.022	0.22	0.11	0.0038	0.12	0.15	0.073	0.0086	1	0.21	0.012
Seasons	0.4	0.00047	0.82	0.25	0.066	0.025	0.8	0.2	0.074	0.21	1	0.12
Holiday	0.07	0.00023	0.058	0.048	0.031	0.022	0.068	0.002	0.013	0.012	0.12	1

Outliers present in our important independent features





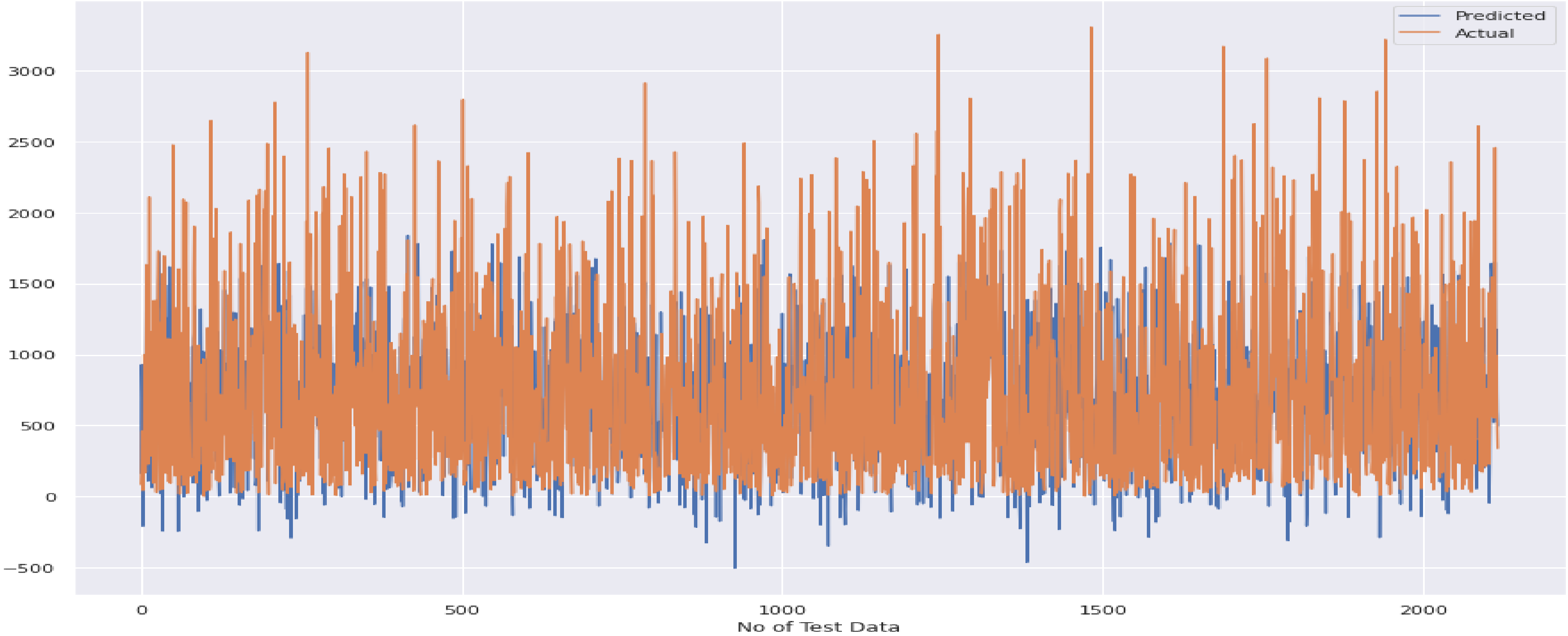




Linear Regression

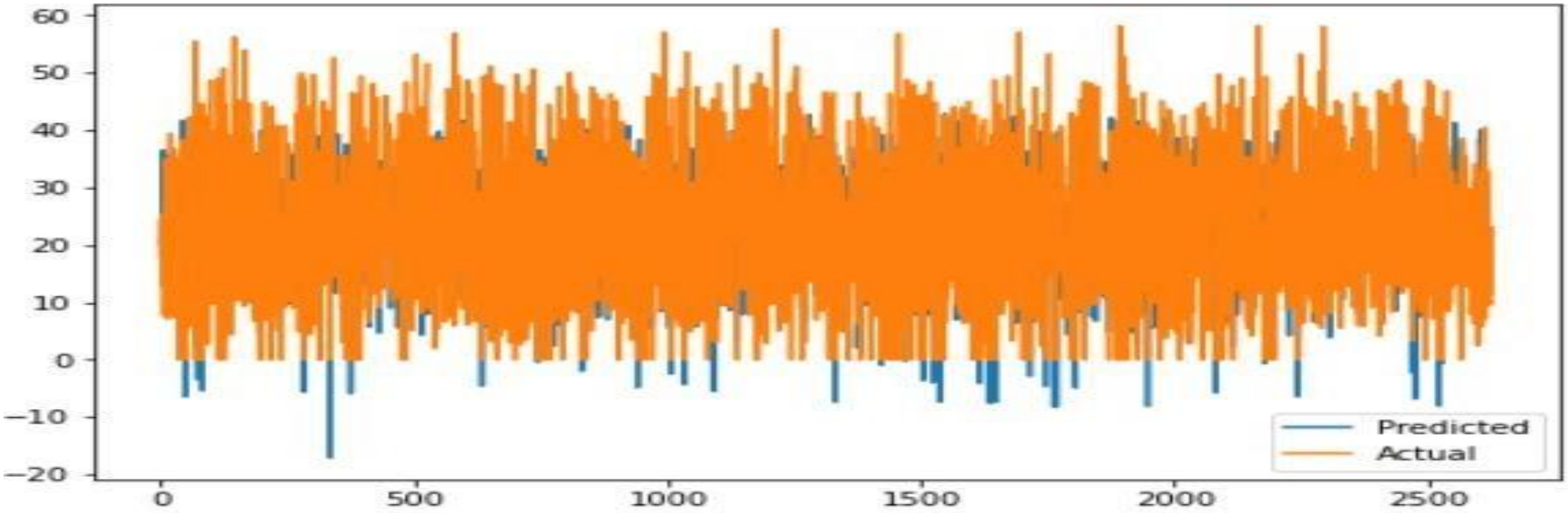
```
MSE : 198793.5341180045
RMSE : 445.8626852720515
Adjusted R2 : 0.5049660638596776
```

```
      MAE : 333.68919457334323
r2_score   0.5073055437091121
```



Lasso Regression

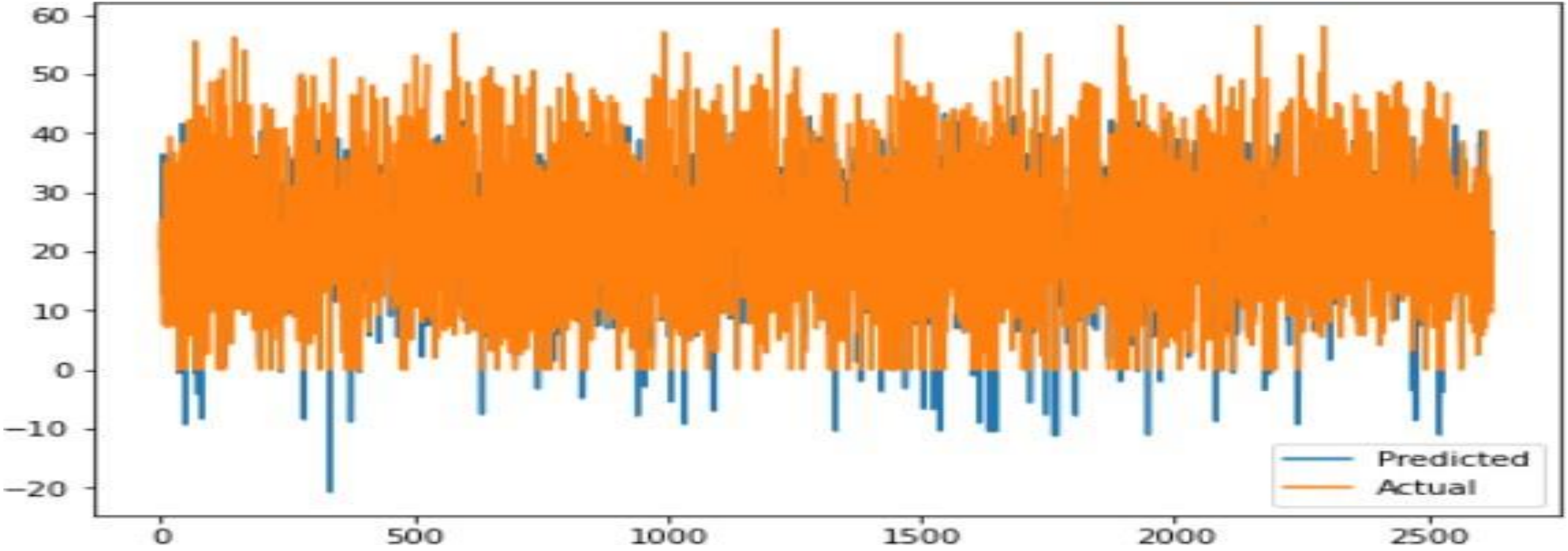
```
  MSE : 198793.663747306
  RMSE : 445.86283064111325
r2_score 0.5073052224328767
 MAE : 333.68926336070683
```



Ridge Regression

```
MSE : 198890.40226455292
RMSE : 445.97130206388044
MAE : 333.7678564764892
```

```
r2_score  0.5070654634720594
```



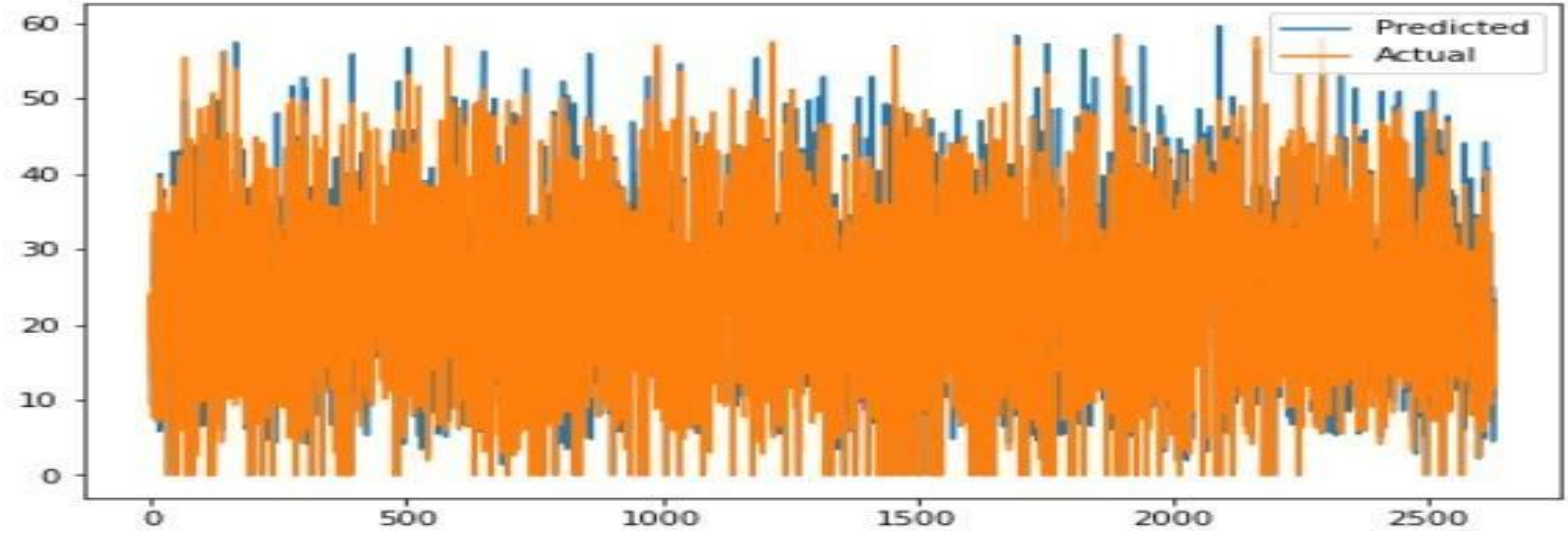
Decision Tree

```
MSE : 111943.4251299008
RMSE : 334.579475057722

MAE : 193.50543221539914
```

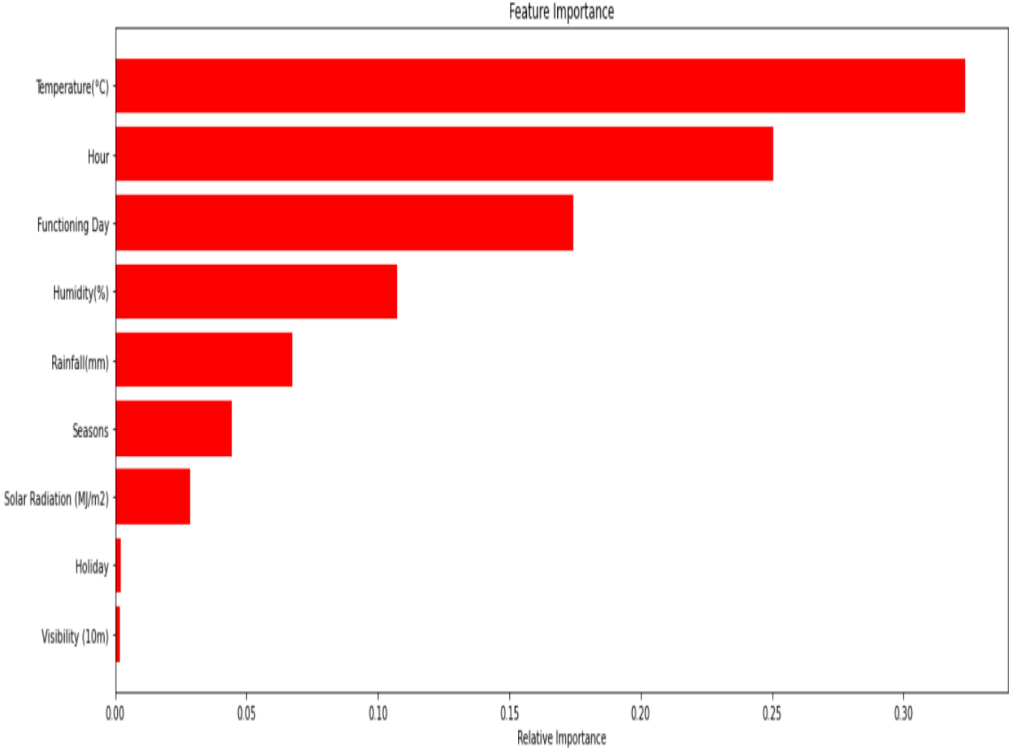
```
r2_score  0.7225568466076131

Adjusted R2 : 0.7212394527168611
```



Gradient Boosting Machine

```
MAE : 174.081134728031
MSE : 67935.3191486026
RMSE : 260.644046831310 9
Adjusted R2 : 0.830828056906927
r2_score      0.83162754624101 6
```



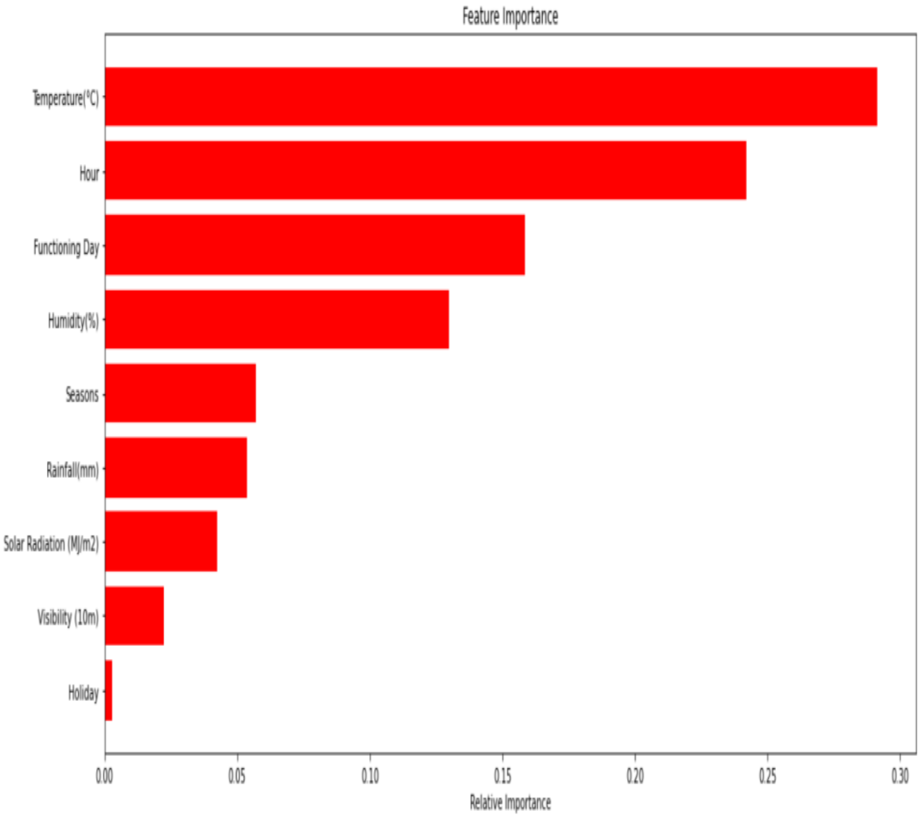
Random Forest

```
MSE : 60132.13303353803
RMSE : 245.2185413738896
```

MAE : 150.1287009919697

```
Adjusted R2 : 0.8502594833570604
    r2_score   0.8509671417532936
```

Feature Importance



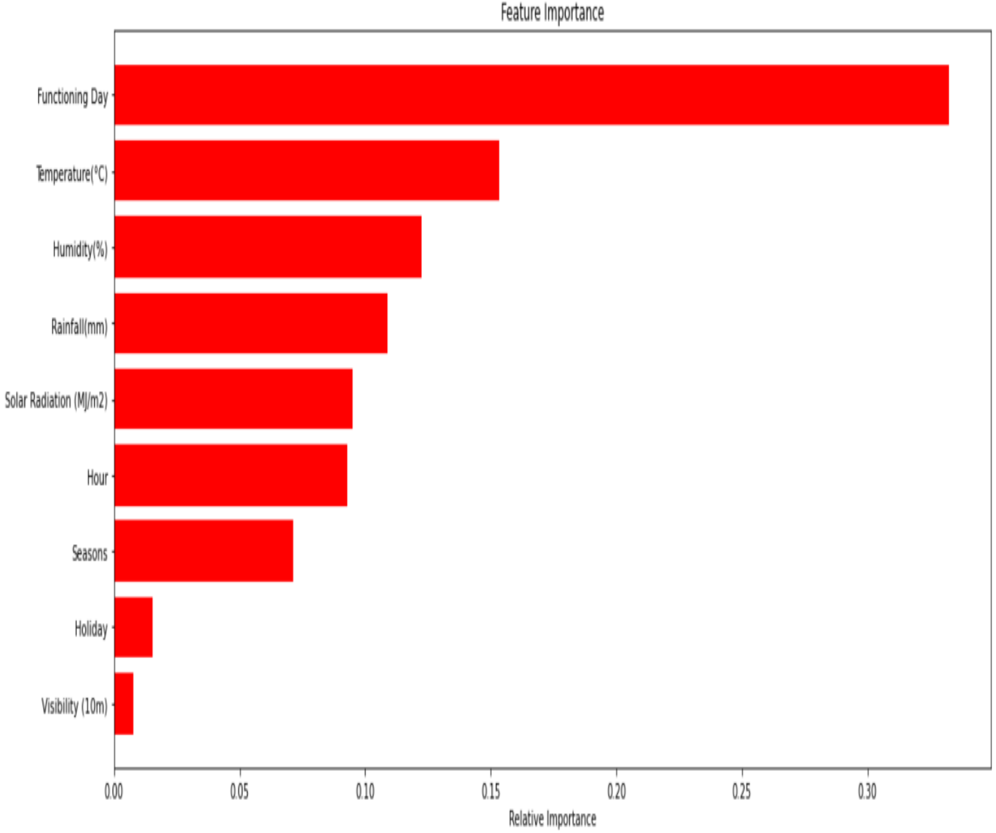
XGBoost

```
MSE : 54287.031544213925
RMSE : 232.9957758076612
 MAE : 143.48340080681663

adj_r2 0.8657453657658387

    r2    0.8662260483087465
```

Feature Importance



Challenges

- Large Dataset to handle.
- Needs to plot lot of Graphs to analyse.
- Carefully handled Feature selection part as it affects the R2 score.
- Carefully tuned Hyperparameters as it affects the R2 score.

Conclusion

- The Rented Bike Count has been increased from 2017 to 2018.

- No overfitting is seen.

- XGBoost Regressor gives the highest R2 score of 96.6% for Train Set and 89.4% for Test set.

- Feature Importance value for Random Forest, Gradient Boost, and XGBoost are different.

- We can deploy this model.

THANK YOU