AlmaBetter

Capstone Project 2

Seoul Bike Sharing Demand Prediction
ML SUPERVISED REGRESSION

CHETAN BHANGARE

NIKHIL BHANGARE

AlmaBetter

Contents

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

Problem Statement

- A bike-sharing system is a service in which bikes are made available for shared use to individuals on a short term basis on rent.
- A South korea city Seoul based rental bike provider wants to come up with a mindful business plan to be able to accelerate its revenue.
- The company wants to find out the key driving factors for the demand for shared bikes.
- These will help them to construct the business strategy to meet the demand levels and meet the customer’s expectations.
- This model will also help to reduce waiting time of public.
- Further , the model will be a good way for management to understaned the demand dynamics of a new market.

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

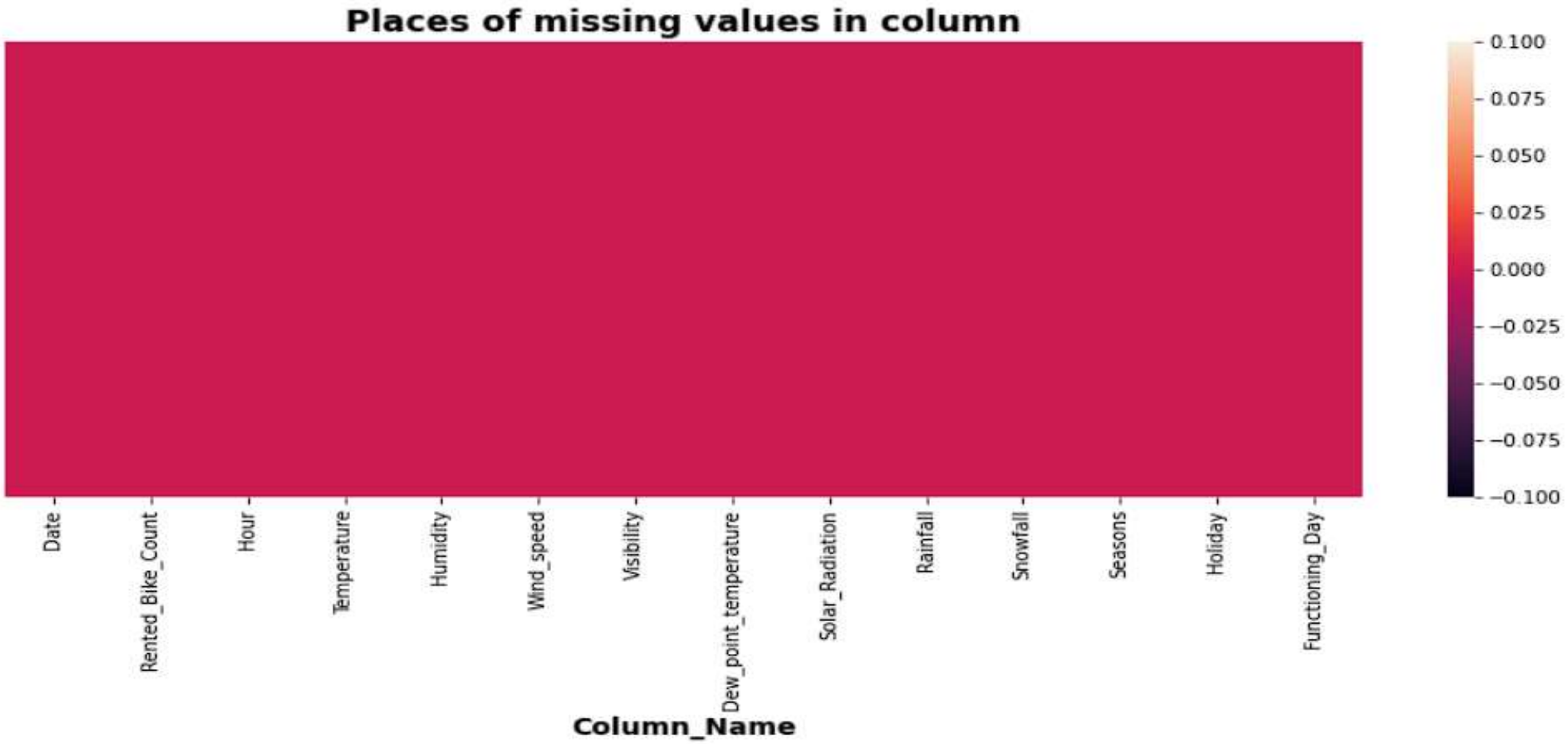
Data Summary

- 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)

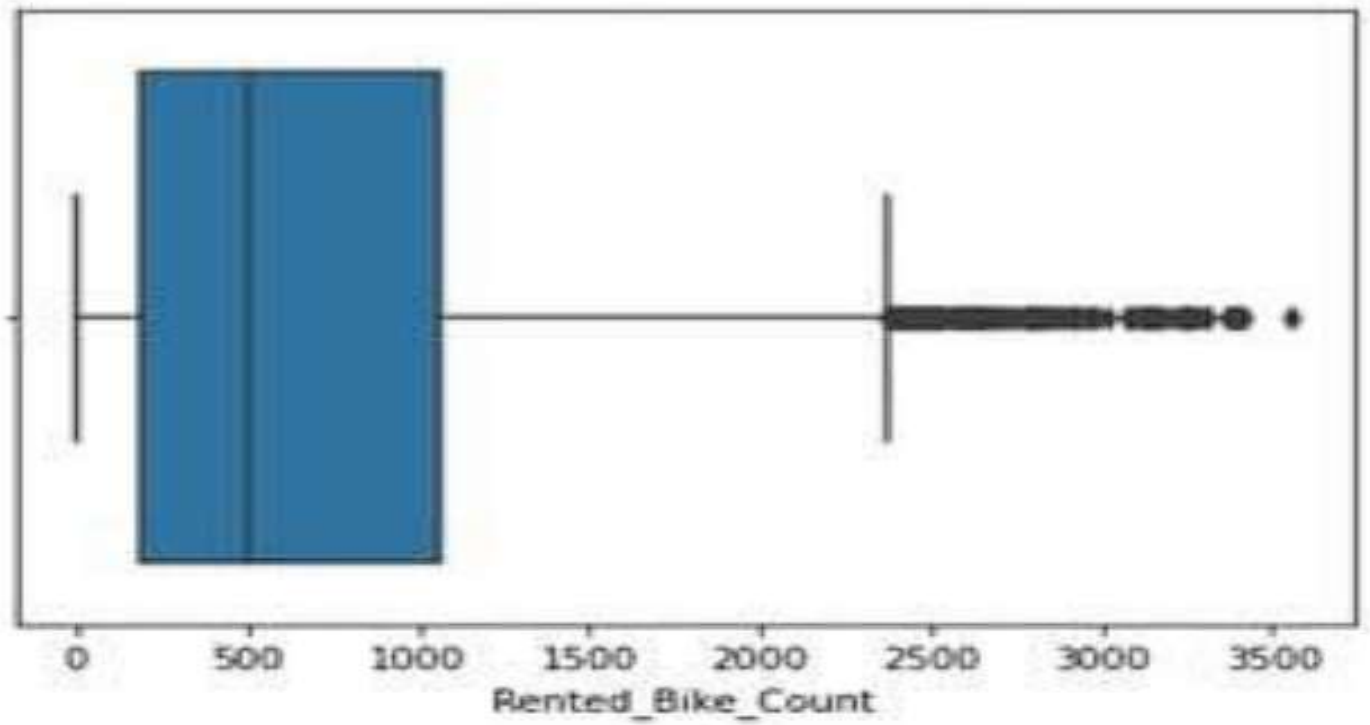
Basic Data Exploration

- 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.

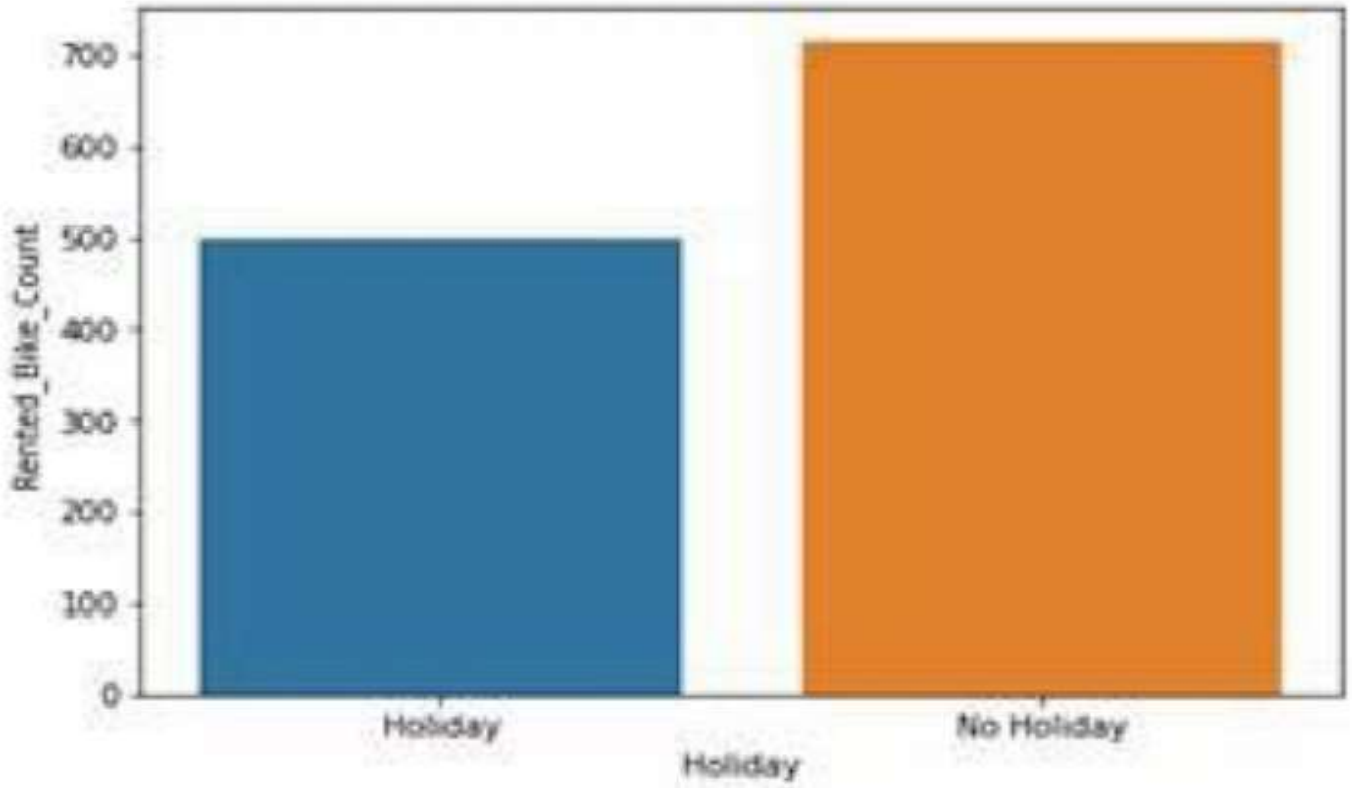
Missing Values



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

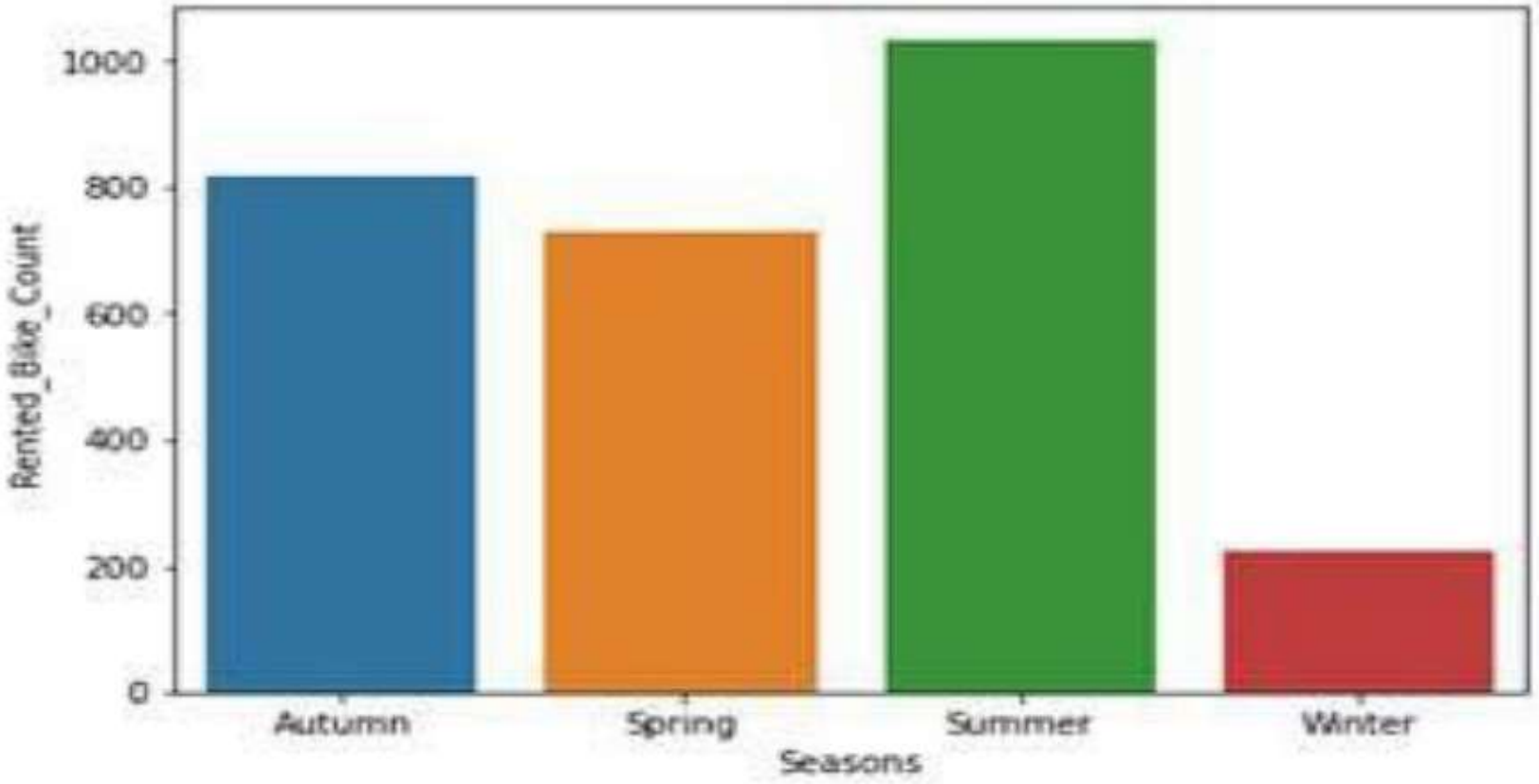
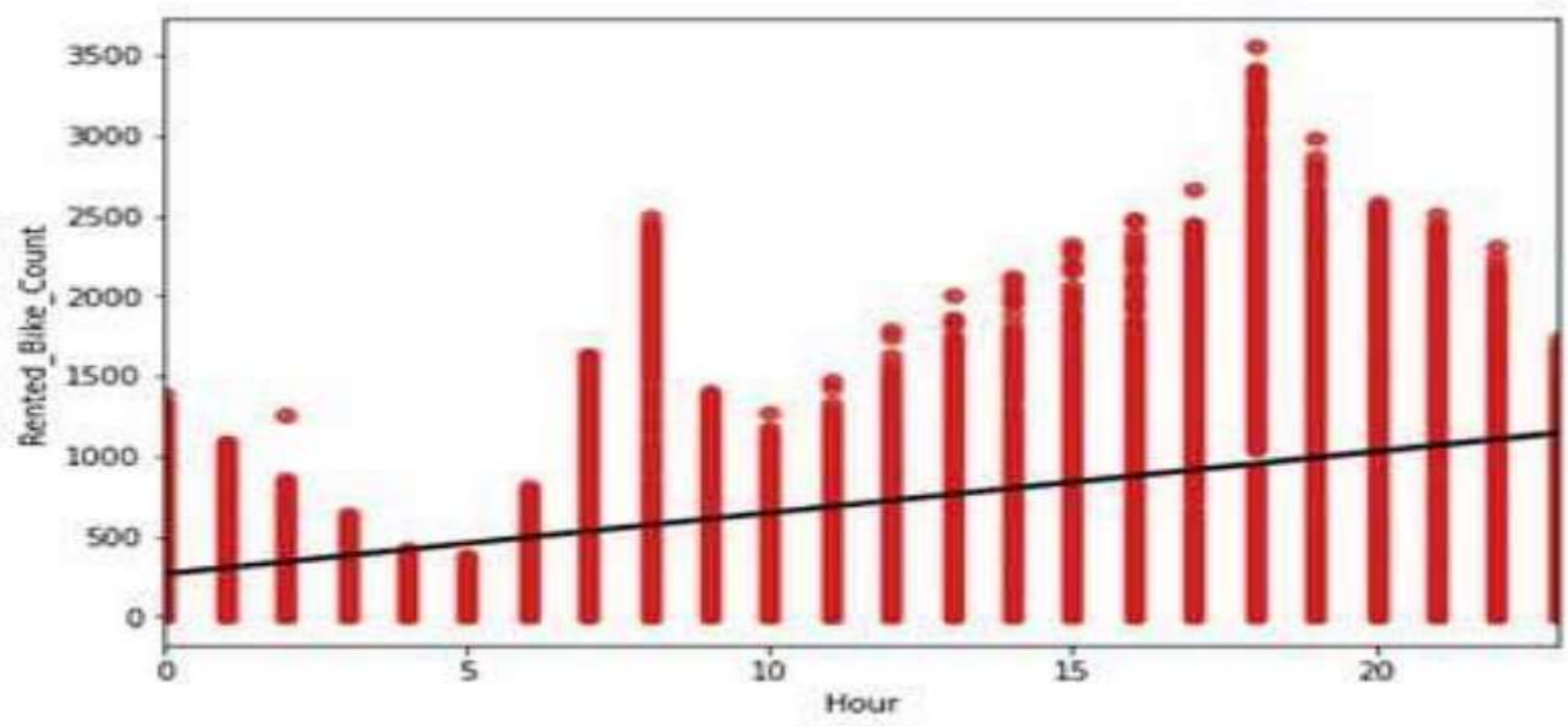
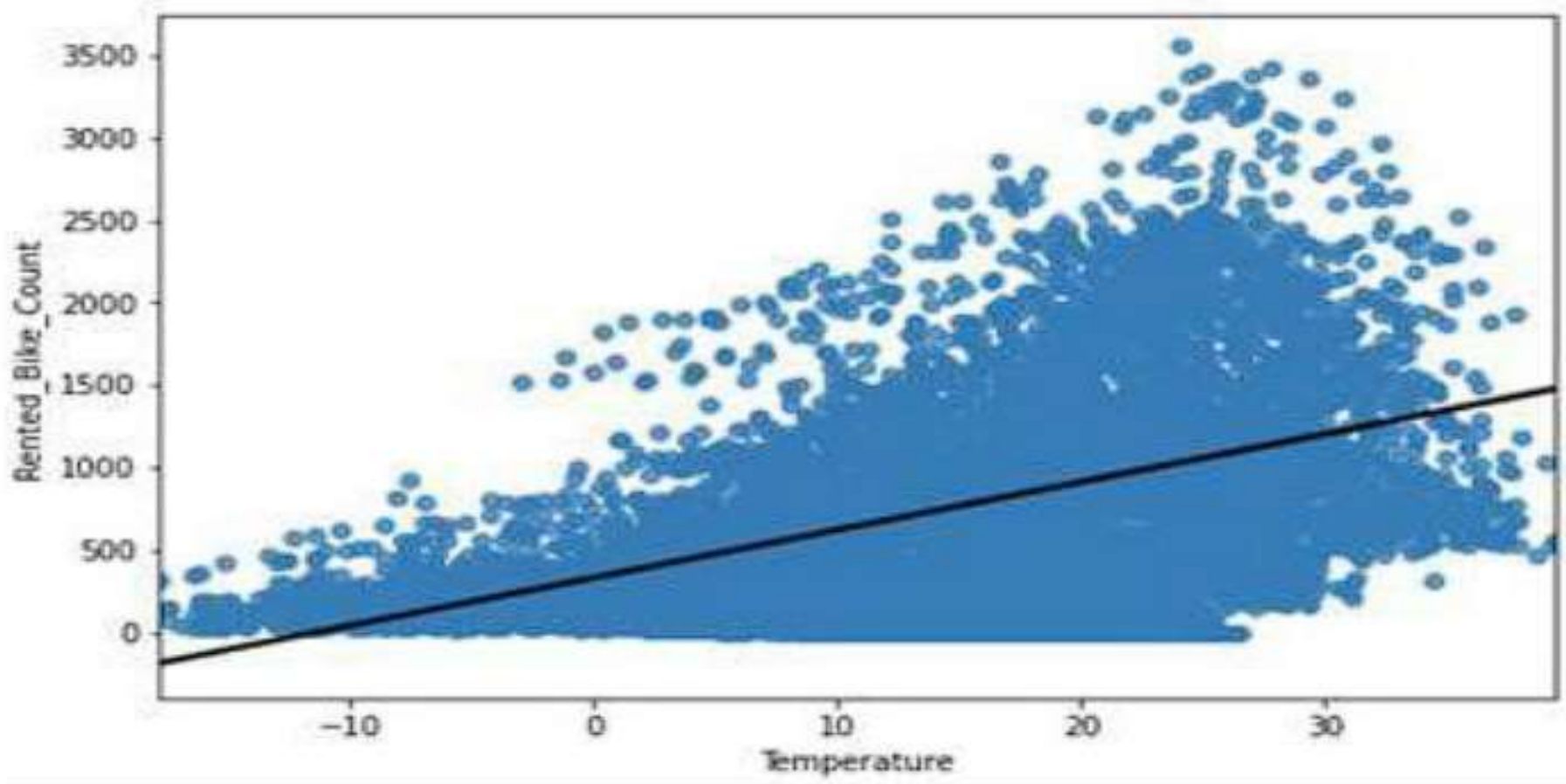


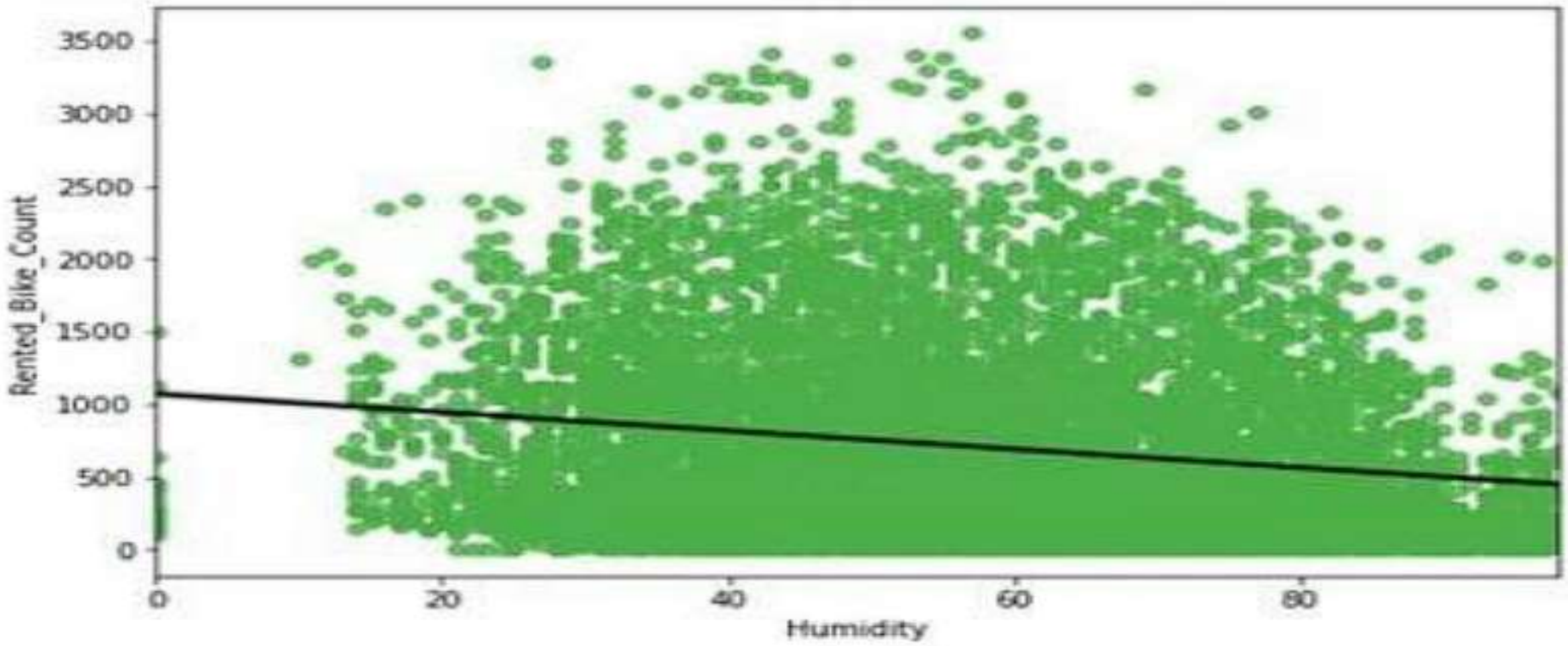
Chart showing distribution of Rented bike count per hour



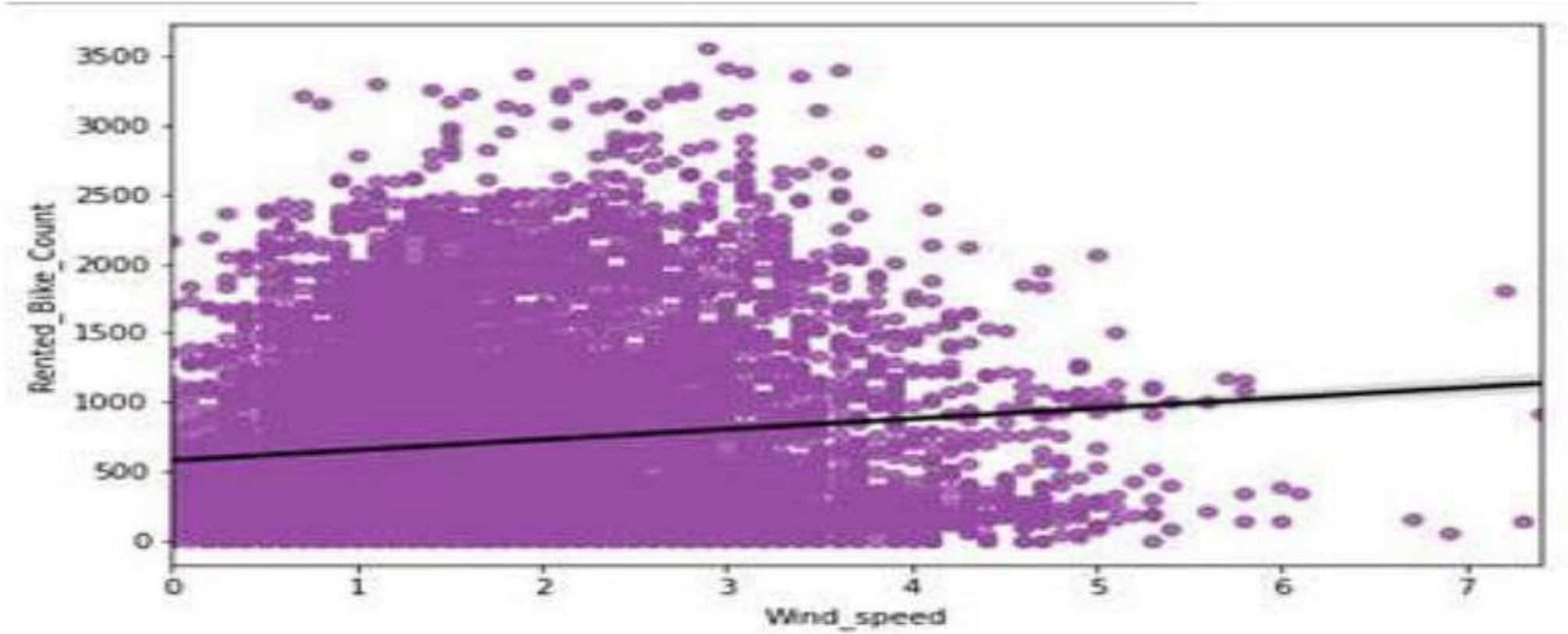
Relation of our dependent variable with Temperture



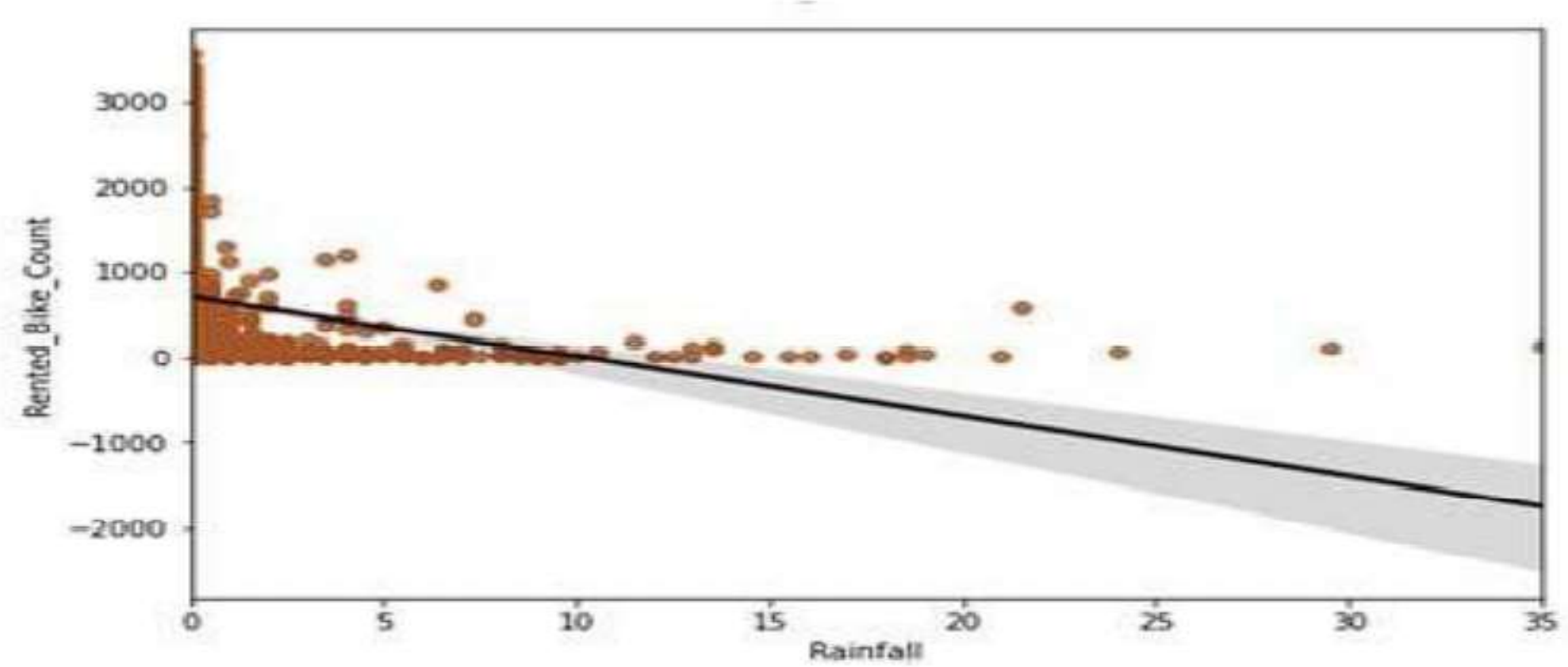
Relation of our dependent variable with Humidity



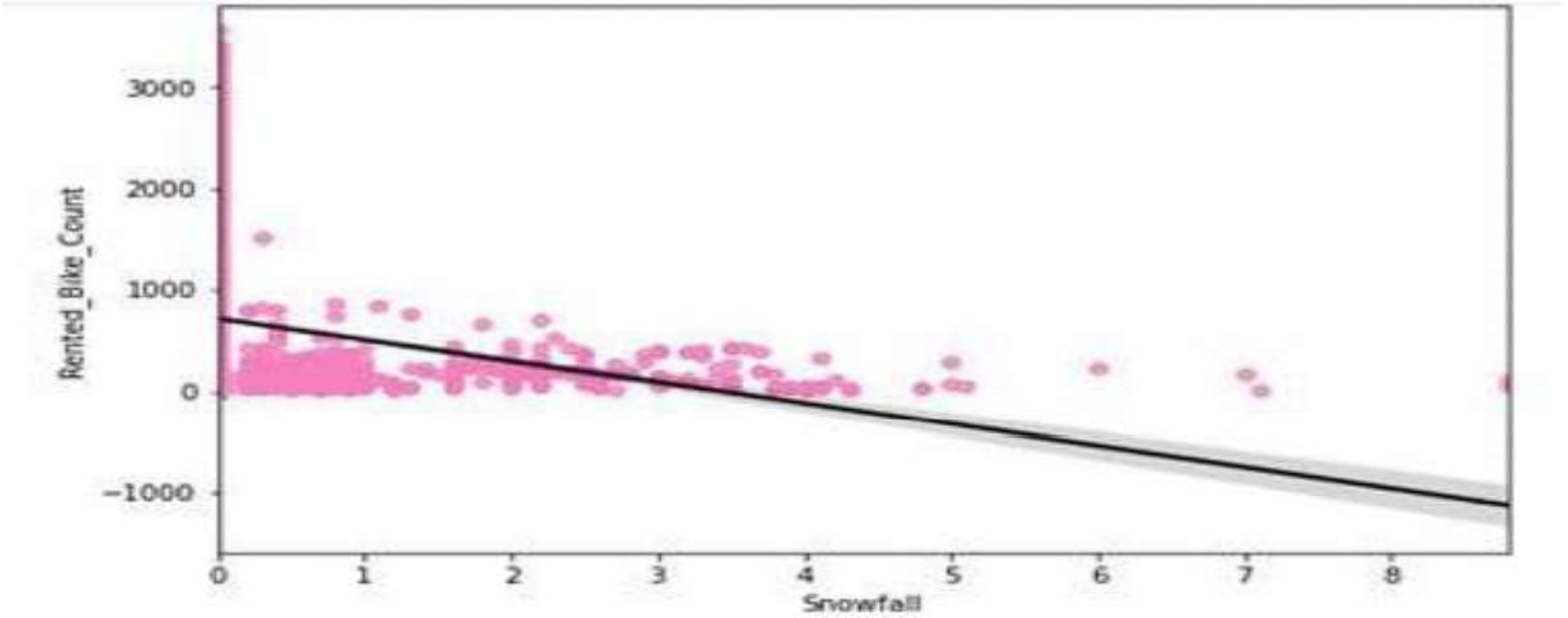
Relation of our dependent variable with wind speed



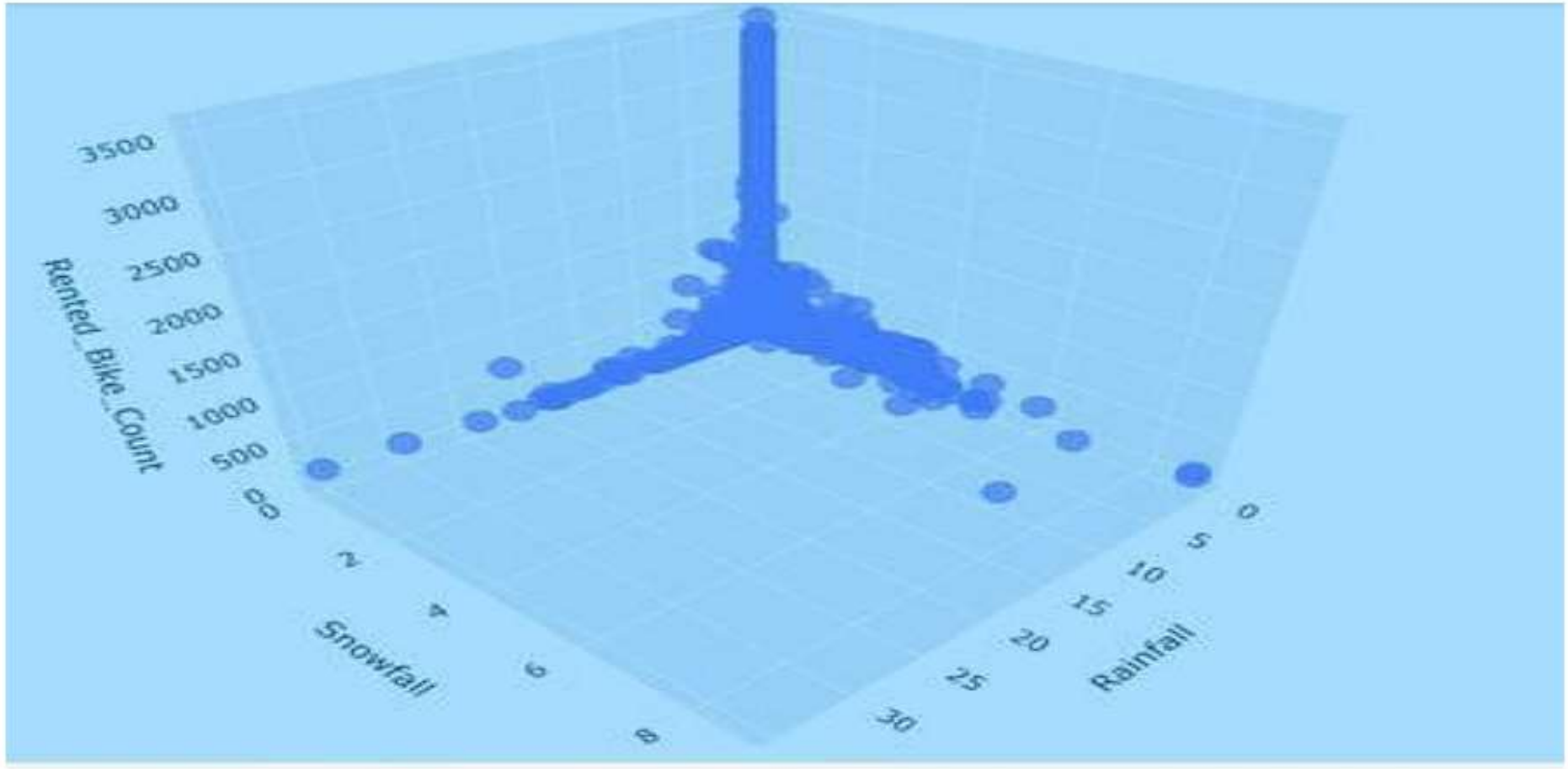
Relation of our dependent variable with Rainfall



Relation of our dependent variable with Snowfall

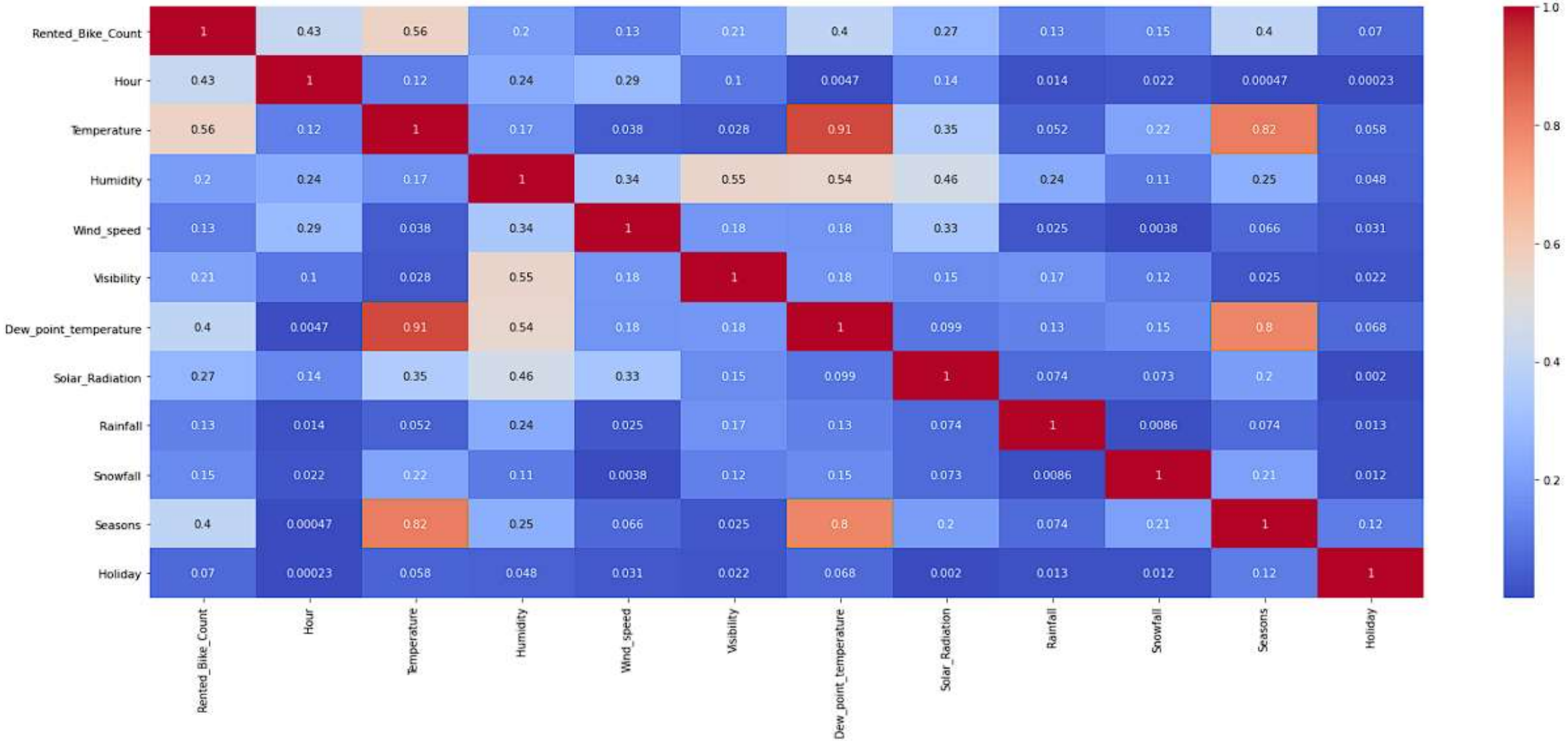


3-D plot showing relation between Snowfall, Rainfall and Rented bike count

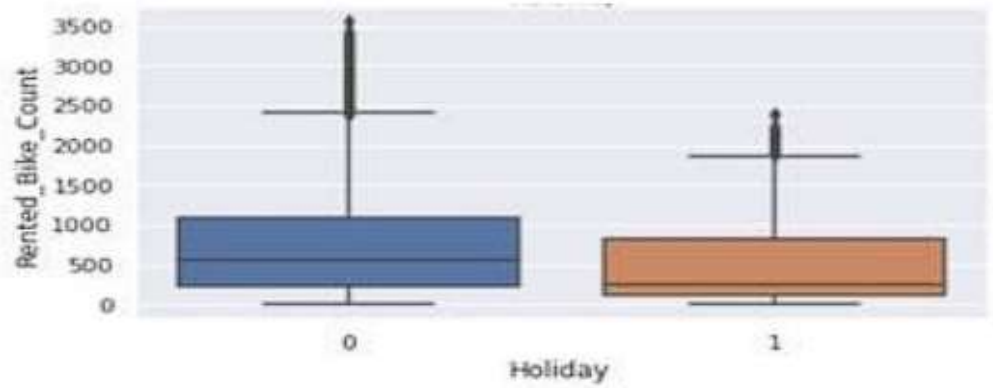


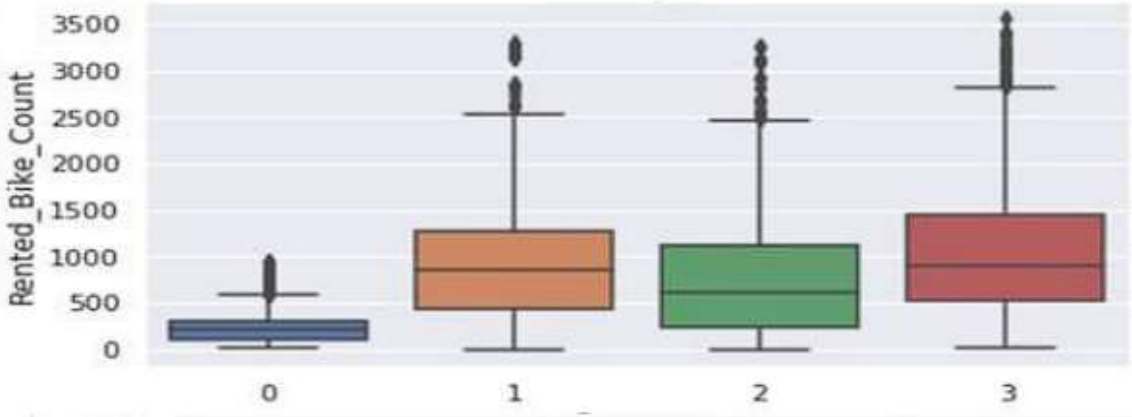
Per hour distribution

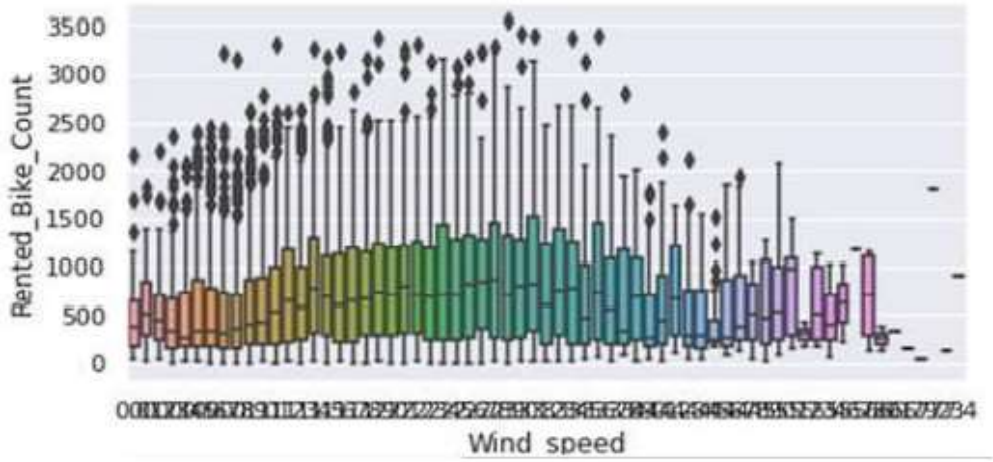
Correlation between different factors

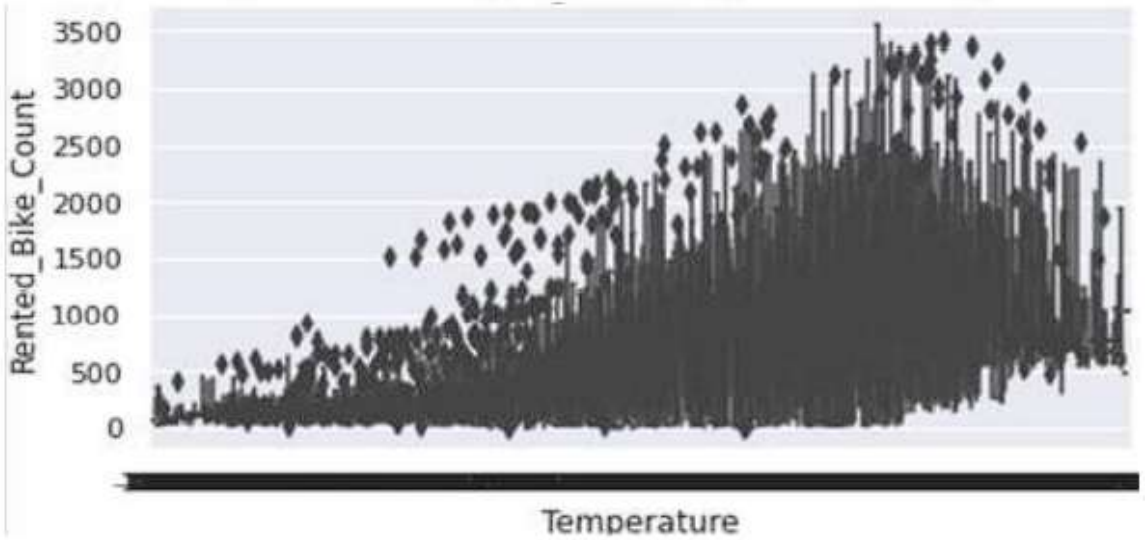


Outliers present in our important independent features





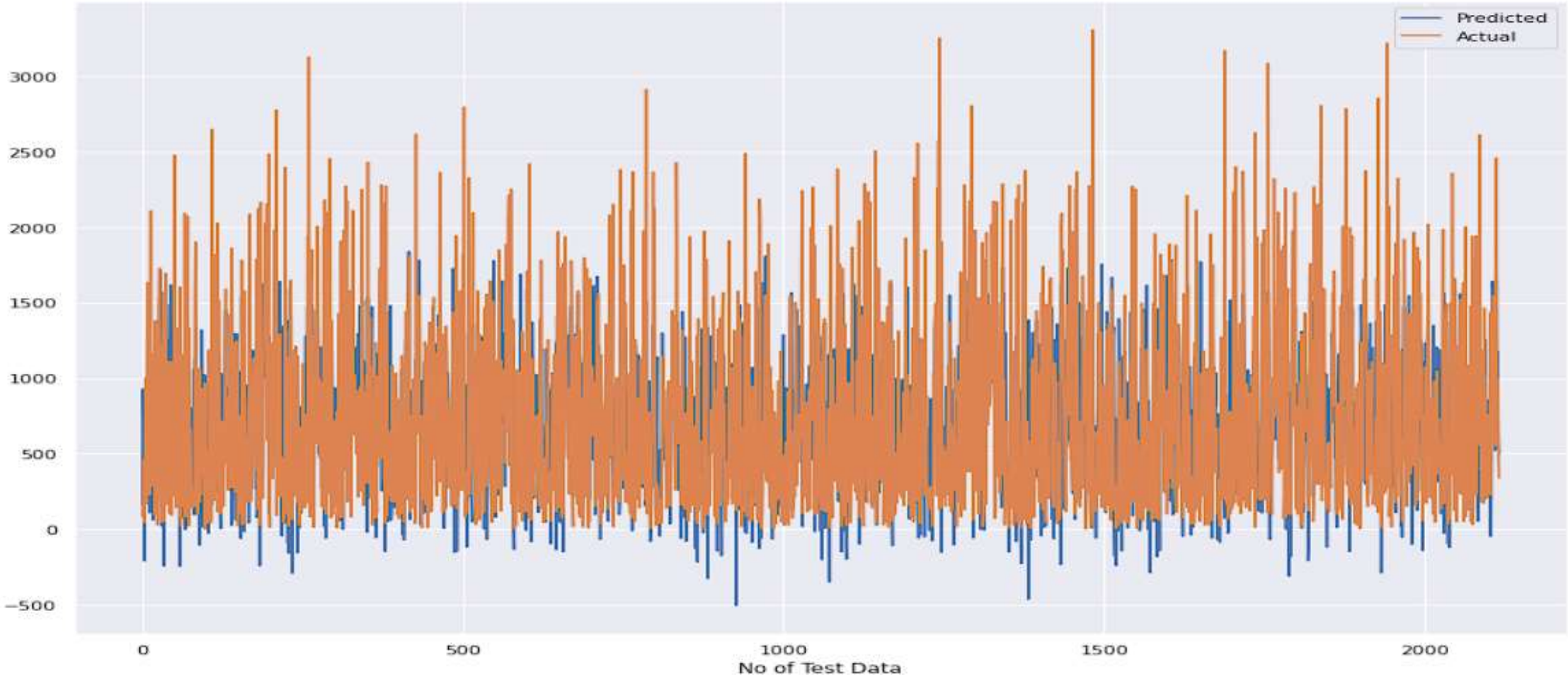




Linear Regression

MSE : 198793.5341180045
RMSE : 445.8626852720515
MAE : 333.68919457334323
Adjusted R2 : 0.5049660638596776

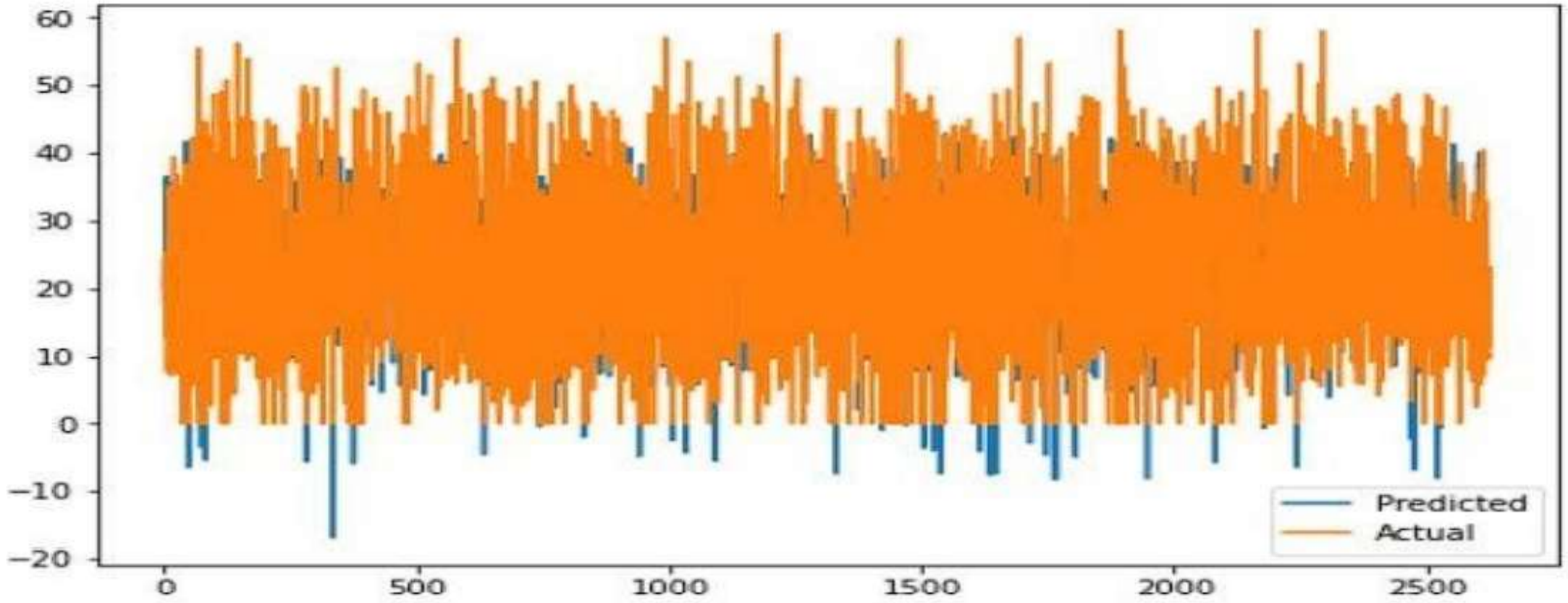
r2_score 0.5073055437091121



Lasso Regression

MSE : 198793.663747306
RMSE : 445.86283064111325
MAE : 333.68926336070683

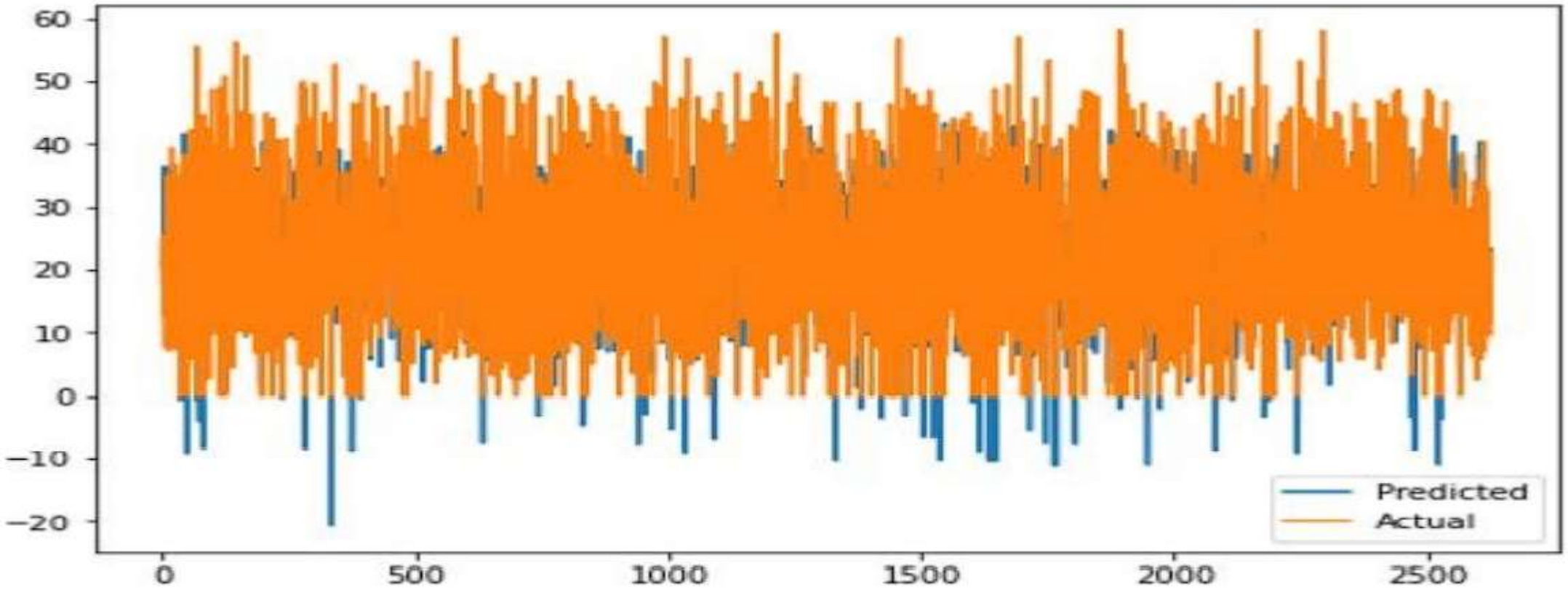
r2_score 0.5073052224328767



Ridge Regression

MSE : 198890.40226455292
RMSE : 445.97130206388044
MAE : 333.7678564764892

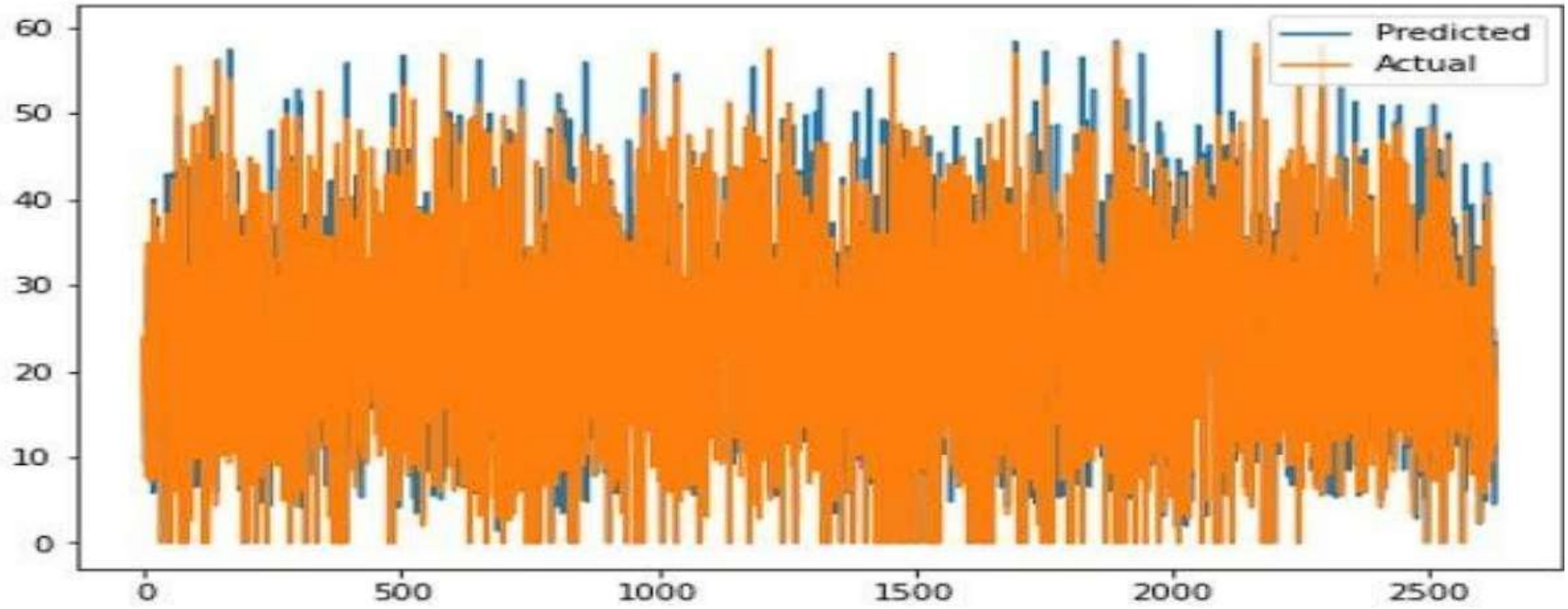
r2_score 0.5070654634720594



Decision Tree

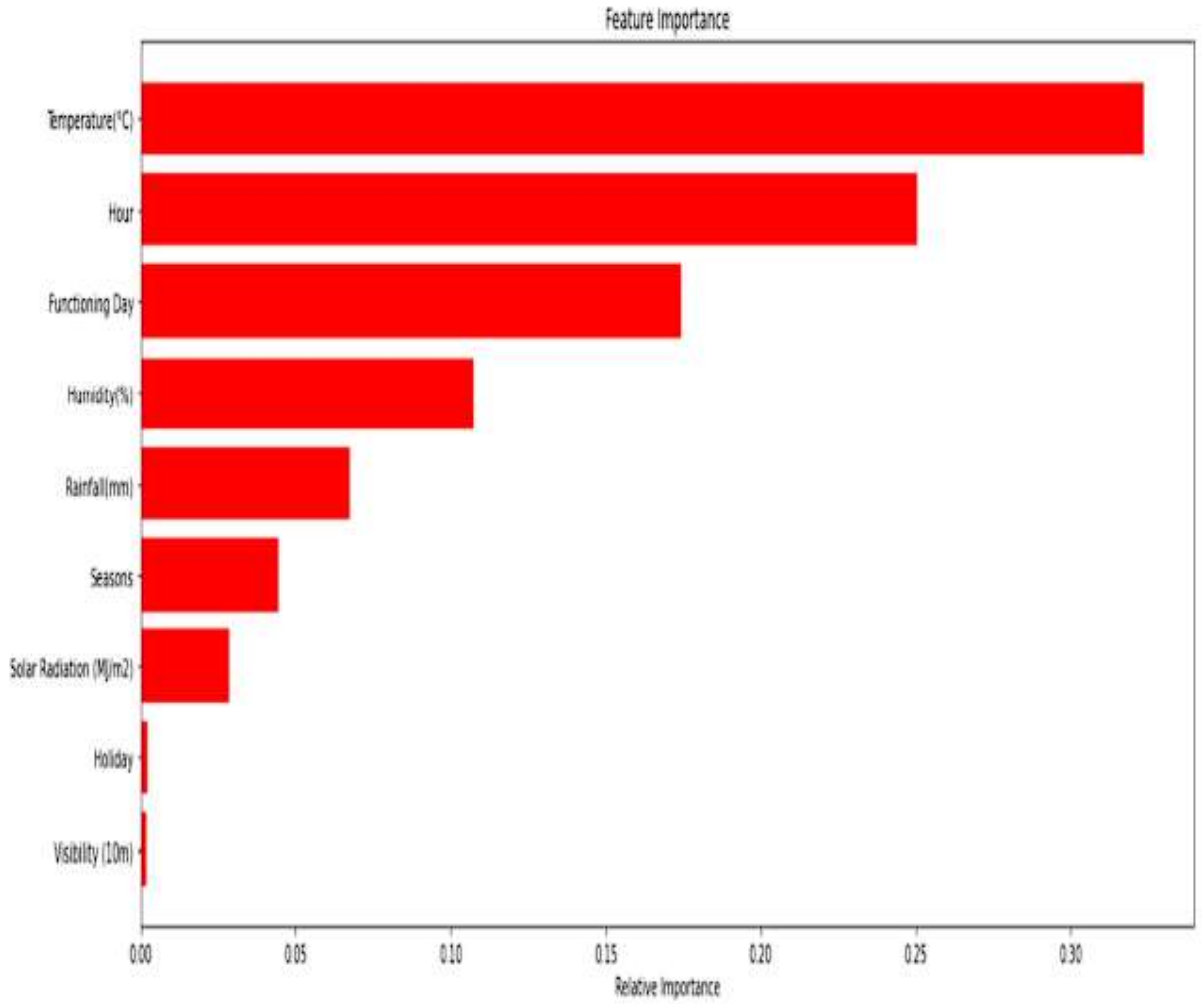
MSE : 111943.4251299008
RMSE : 334.579475057722
MAE : 193.50543221539914
Adjusted R2 : 0.7212394527168611

r2_score 0.7225568466076131



Gradient Boosting Machine

MAE : 174.081134728031
MSE : 67935.3191486026
RMSE : 260.6440468313109
Adjusted R2 : 0.830828056906927
r2_score 0.831627546241016

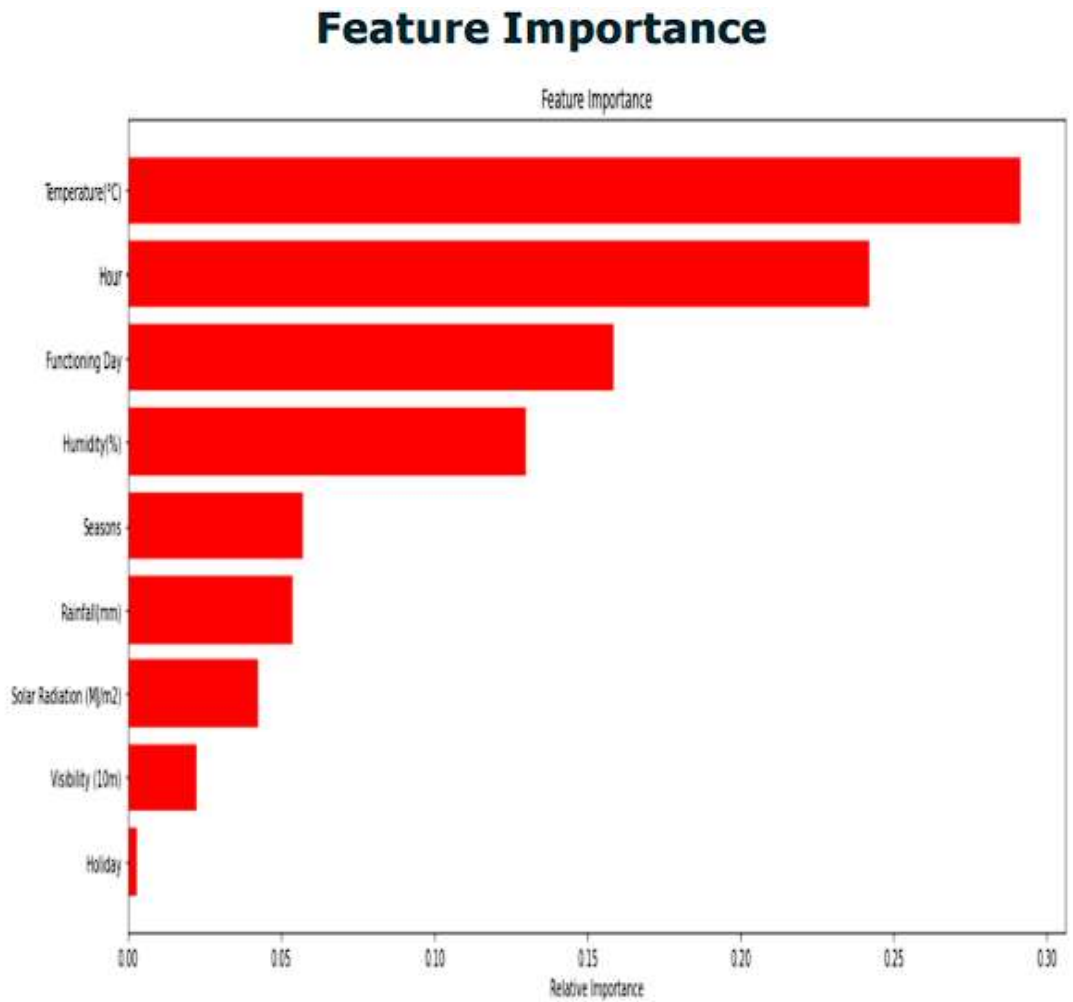


Random Forest

MSE : 60132.13303353803
RMSE : 245.21854137388965
MAE : 150.1287009919697

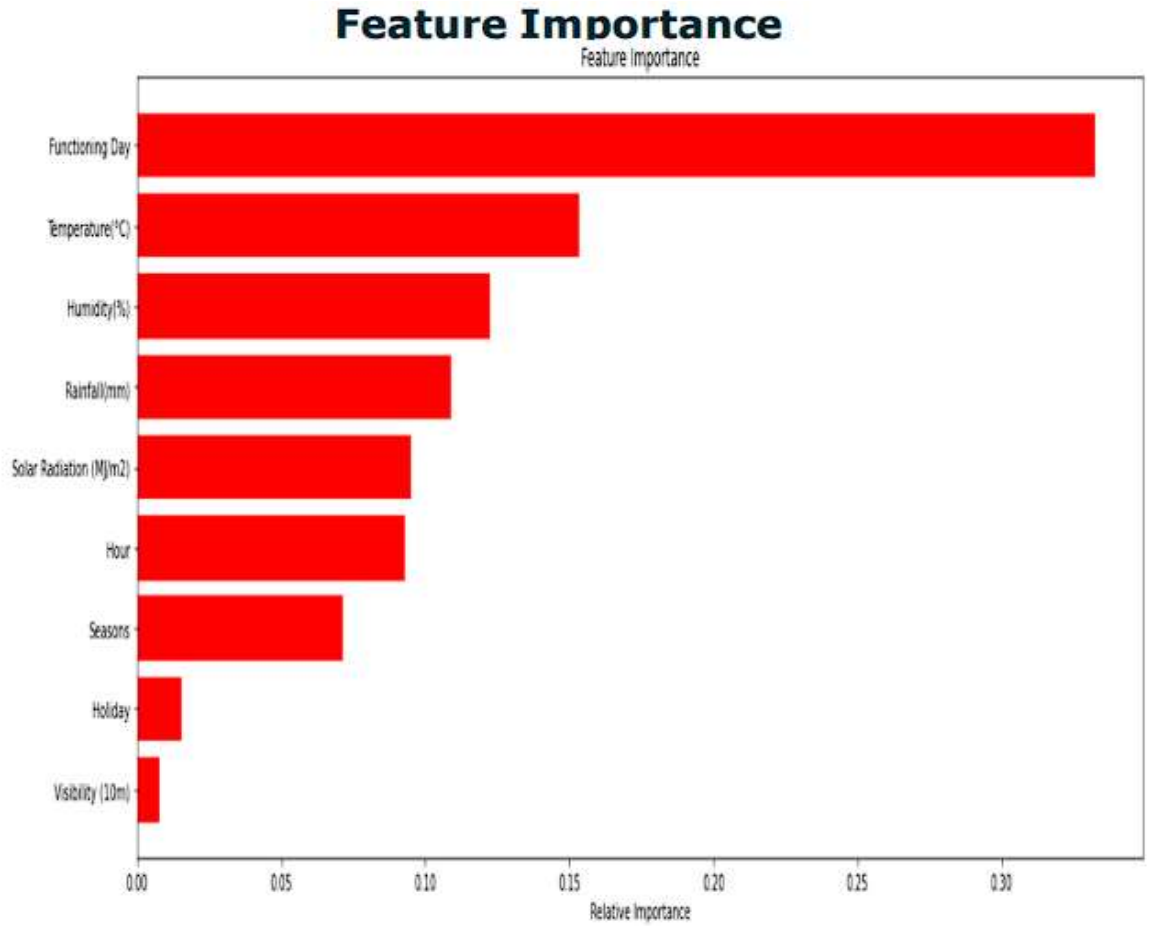
Adjusted R2 : 0.8502594833570604

r2_score 0.8509671417532936



XGBoost

MSE : 54287.031544213925
RMSE : 232.9957758076612
MAE : 143.48340080681663
adj_r2 0.8657453657658387
r2 0.8662260483087465



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.

AlmaBetter

THANK YOU