# Capstone Project

## Bike Sharing Demand Prediction

Nakul Pradeep

# PROBLEM STATEMENT

- 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.
- Explore the Seoul Bike demand dataset to deduce how the bike demand varies with weather conditions, date and time.
- We have to predict the number  of bikes required at each hour for the stable supply of rental bikes.

# Summary

- Bike rental system is a transportation system in cities in which bikes are available to individuals for a short period of time.

- It is popular in metro cities due to environmental, economical reasons and the presence of separate bike lanes helps in reducing the traffic.

# 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 - %
- Wind speed - 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 - Non Functional Days and Functional Days

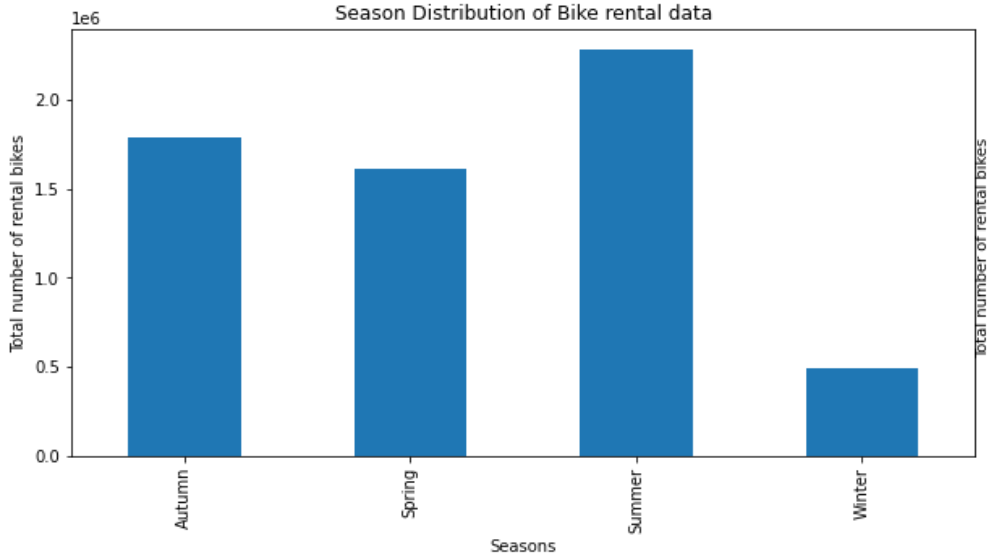
# Steps of our Data Analysis

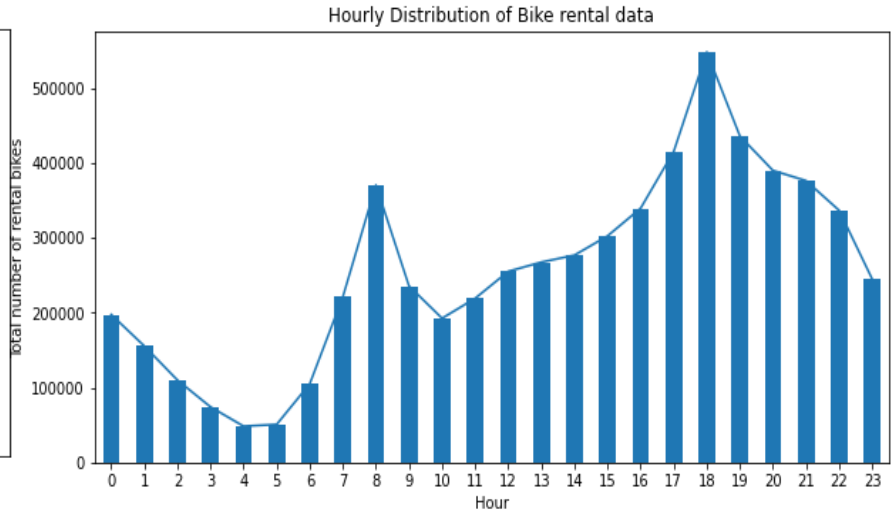
1. Problem Description and Data Description
2. Import the Libraries and Dataset
3. Outliers /Null value Treatment
4. Exploratory Data Analysis
5. Splitting the dataset into training and test sets.
6. Dataset transformation before applying to models which are parametric (i.e. models having prerequisites)
7. Models and Model Selection.
8. Evaluation of All the models and Comparison
9. Feature Importance of optimal model
10. Conclusion

# EXPLORATORY DATA ANALYSIS

- 8760 rows and 14 columns
- No null values or duplicates
- Date, Hour , Seasons, Functioning Day and Holiday are the categorical
- Maximum number of Bikes were rented on the Date: 13th June 2018 and the total number of bikes rented on that day is 36149.
- There where 12 days where there was zero rented bikes and on data analysis we found that all these where non functioning days.
- Day, month , year and weekday extracted from date column and made new columns out of it
- Bike rental demand is less on holidays. This indicates that people prefer to use these bikes as mode of transportation to work.

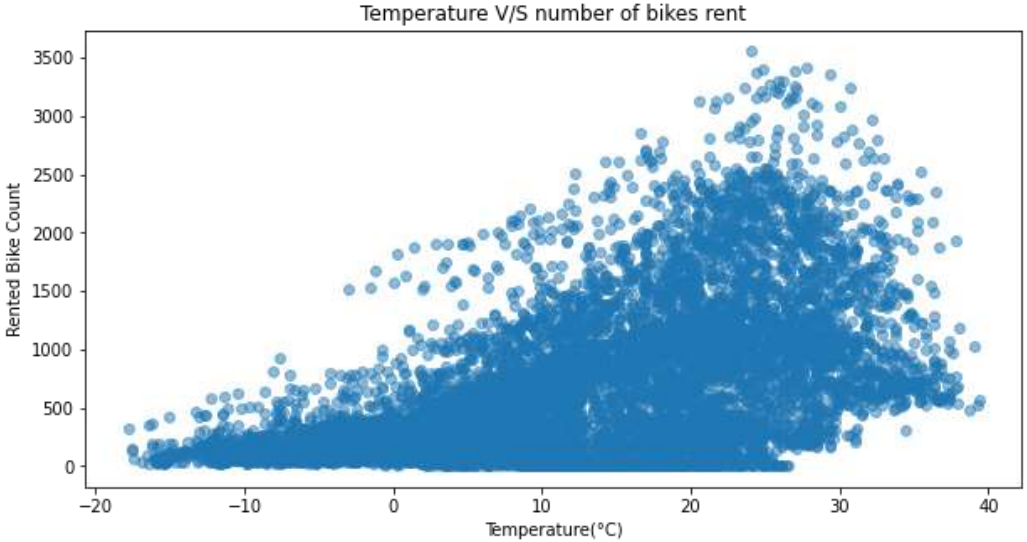
# Hourly and Seasonal Distribution of Total Bike Rental

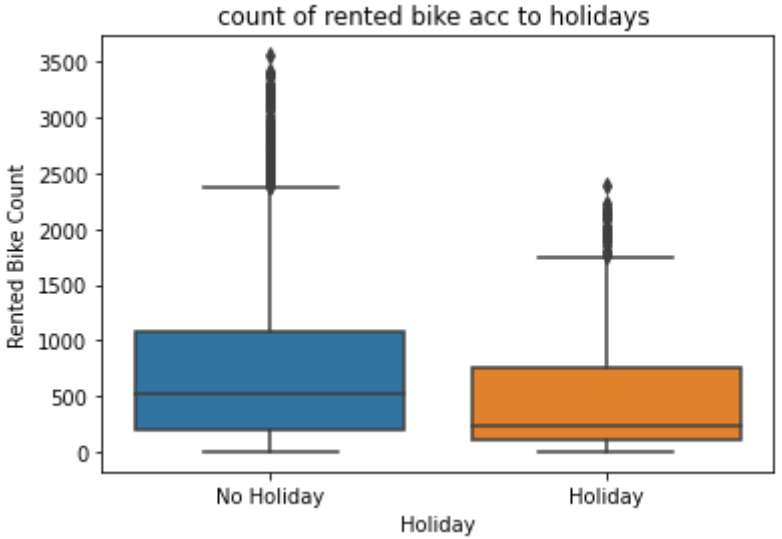



- Bike demand is maximum between 17th hour to 19th hour ie between 5pm to 7pm.
- People use rental bikes most during Summer and least during Winter

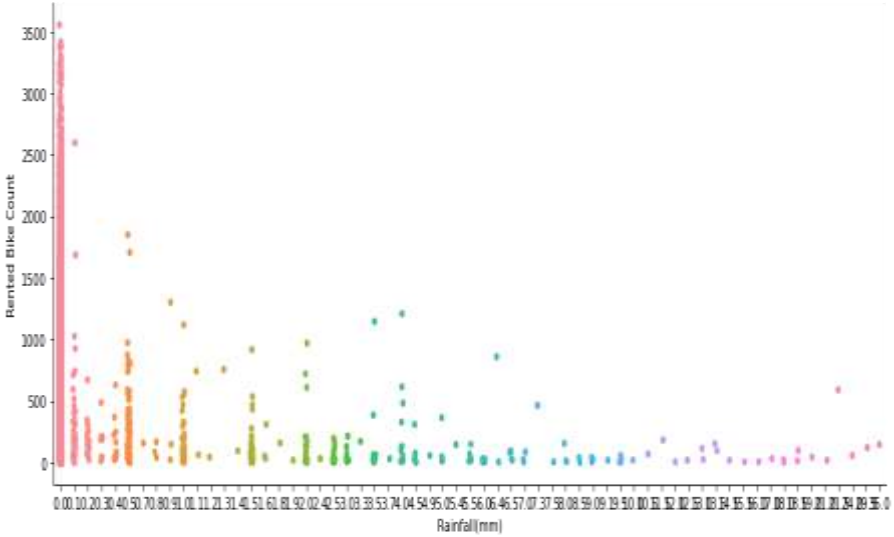
# Temperature ,Holiday relation with Rental Bike

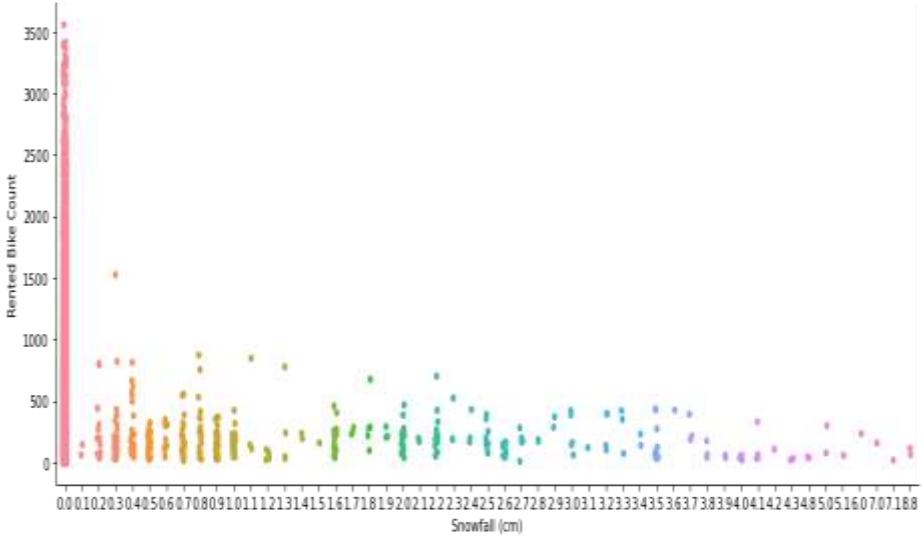



- From the scatterplot it is clear that when temperature is high then rental bike count is high. Also we found out that summer is favourable for high bike rentals and temperature is high during summer
- It is clear from the boxplot that people use rental bikes lesser during holidays.

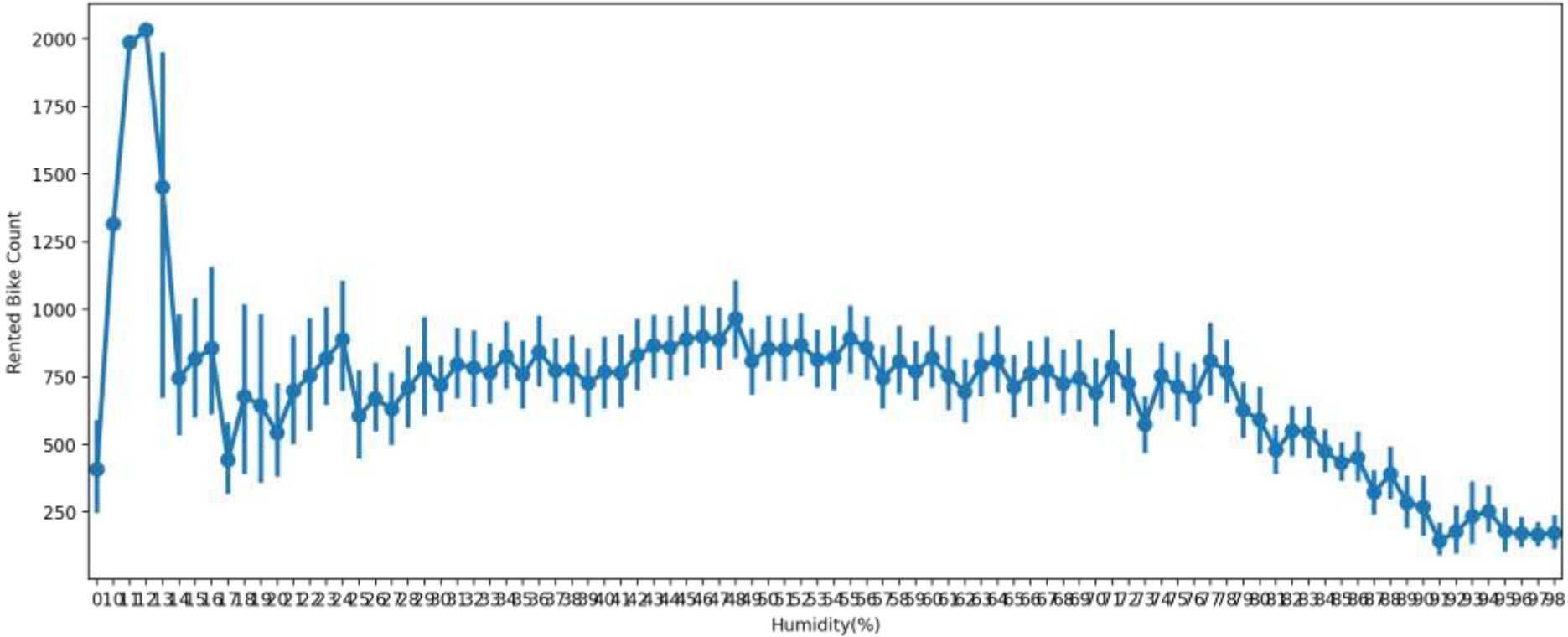
# Rainfall and Snowfall relation with Rental Bike Count



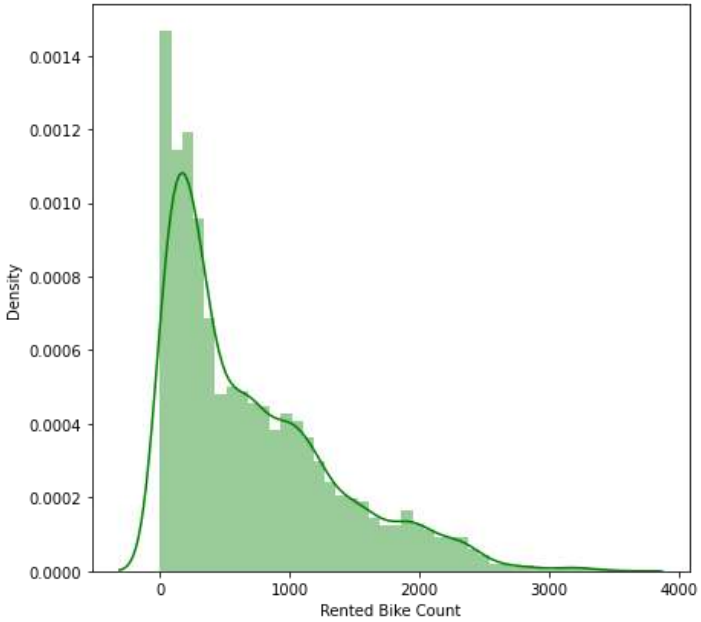

It is clear from the above categorical plot that maximum bike rentals takes place when there is zero or minimum rainfall or snowfall

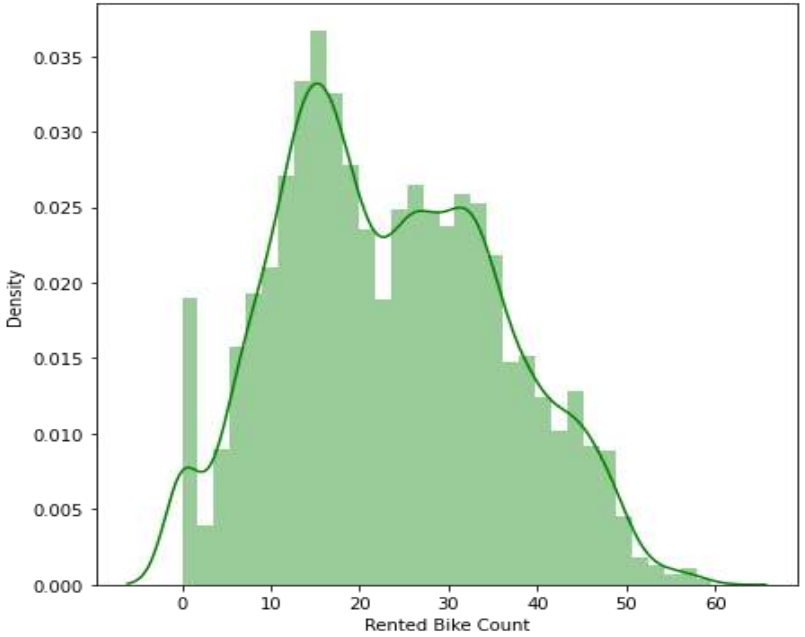
# Humidity relation with Rental Bike Count




From the point plot it is seen that as humidity increases the bike rental reduces

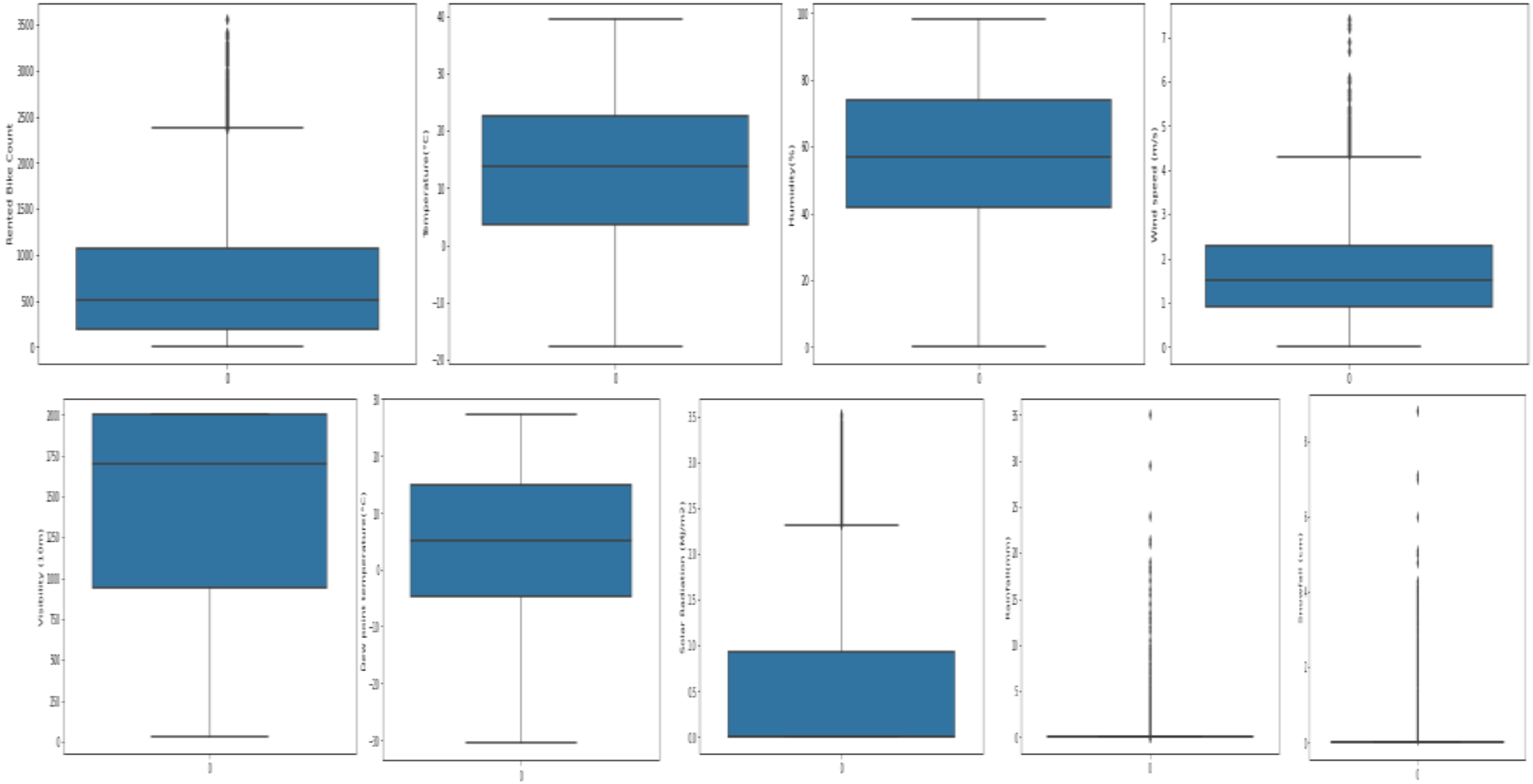
# Distribution Plot of Dependent Variable



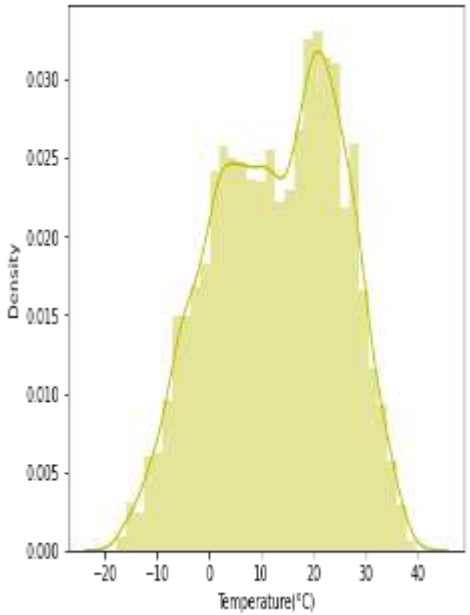



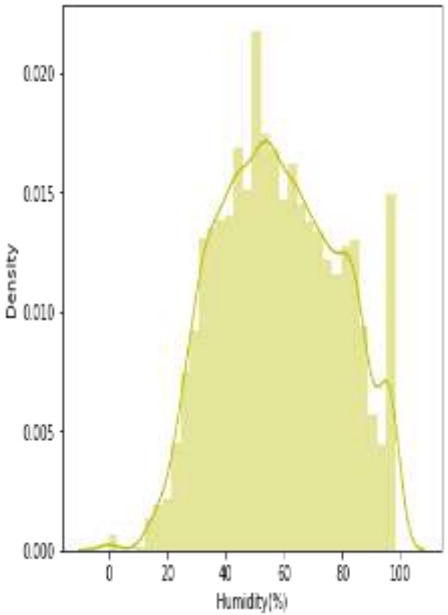

Skewness of the dependent variable is reduced by the square root transformation better than log transformation .

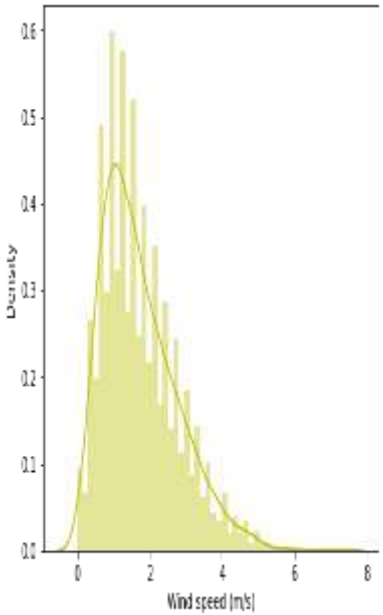
# Outlier Checking

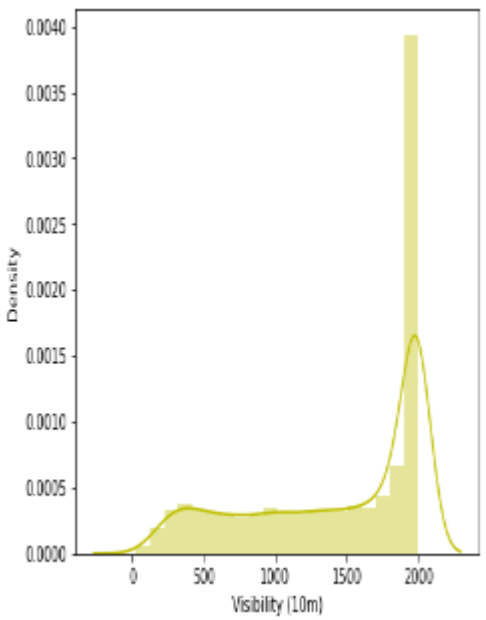


# Distribution of independent Variables

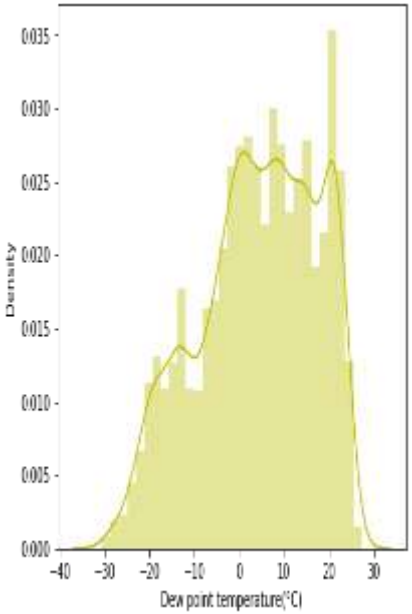


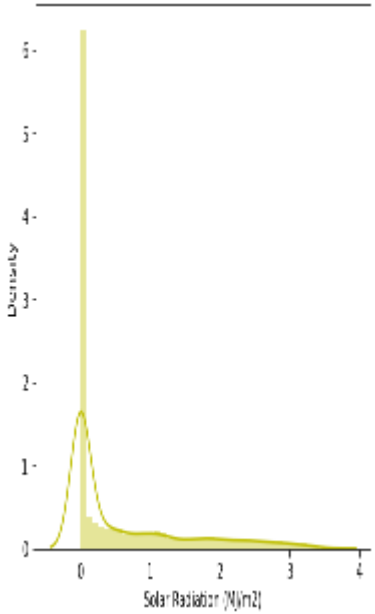



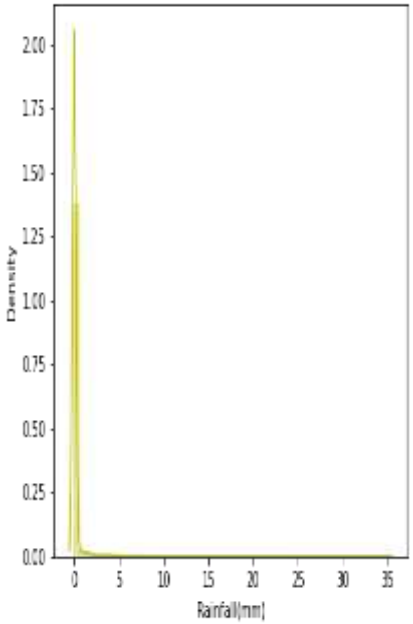



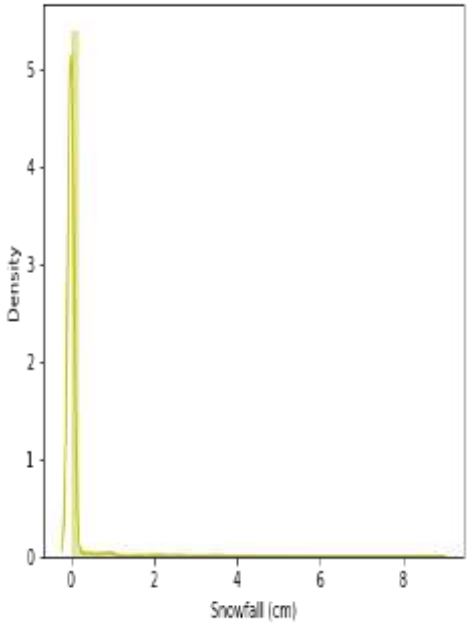



# Distribution of independent Variables



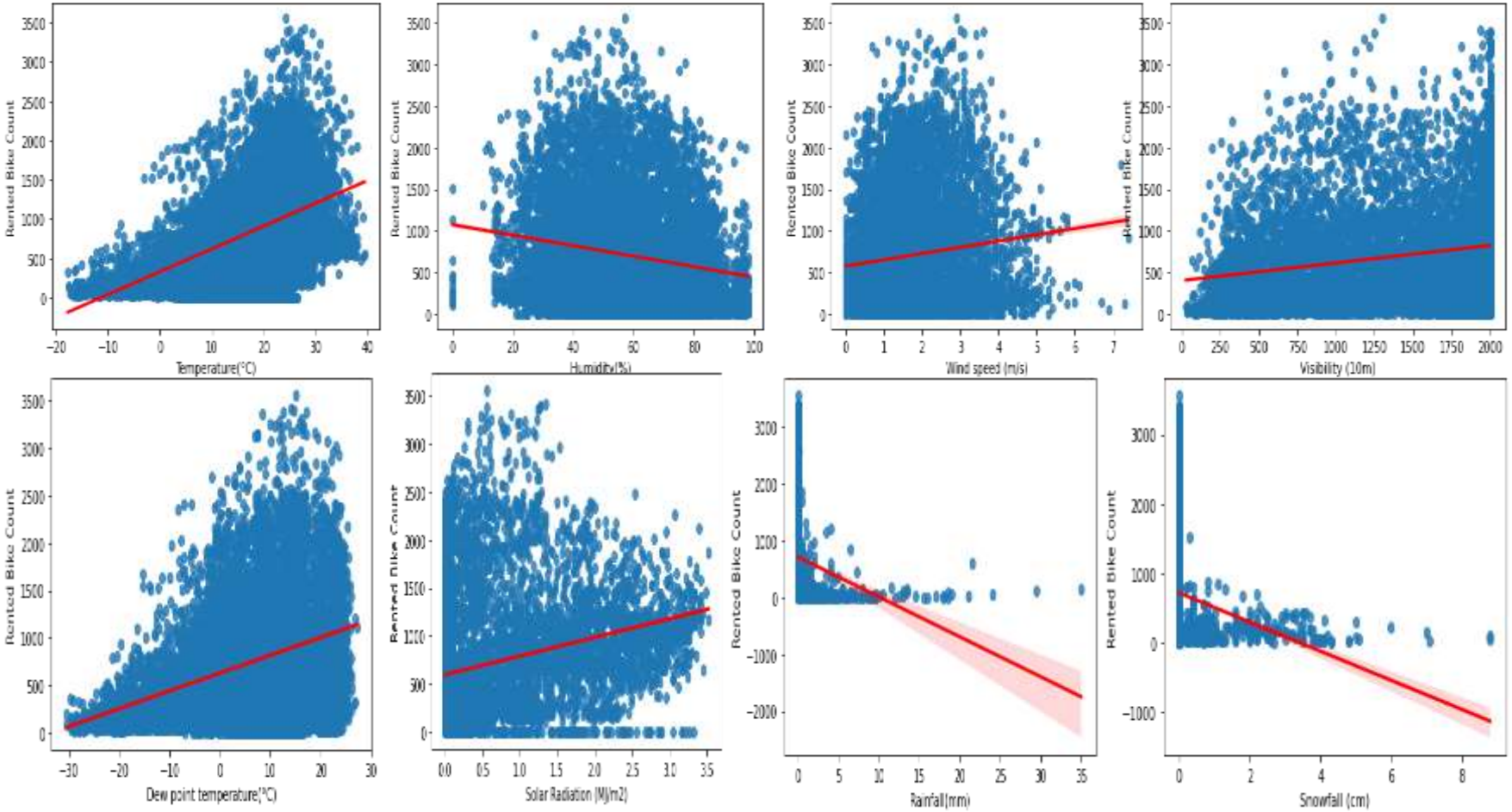



# CORRELATION HEATMAP


Temperature and dew point temperature are highly correlated. ie (0.91)

# Relation between dependent and independent variable

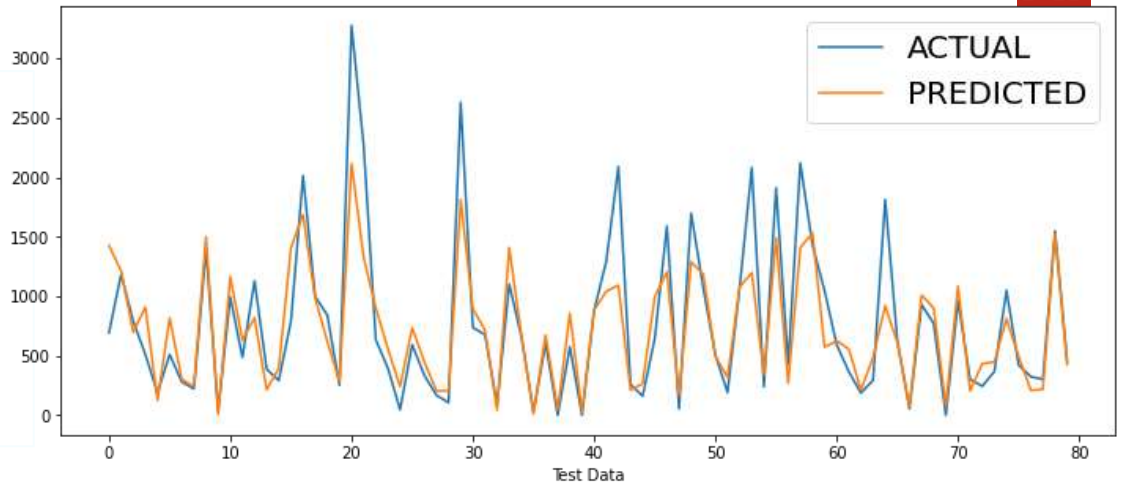


# Pre-processing

- Dew point temperature is highly correlated with temperature and on checking VIF multicollinearity is confirmed and the feature is dropped.
- Categorical features are nominal so one hot encoding is done on them.
- Except tree algorithms ,other models are transformed to accommodate the assumptions and since independent variables are having outliers and are not normally distributed , we use yeo- Johnson power transformation before fitting it to the said models.
- Square root transformation is done for the dependent variable for non tree based algorithm like Linear and polynomial regression to satisfy their assumption before training them.

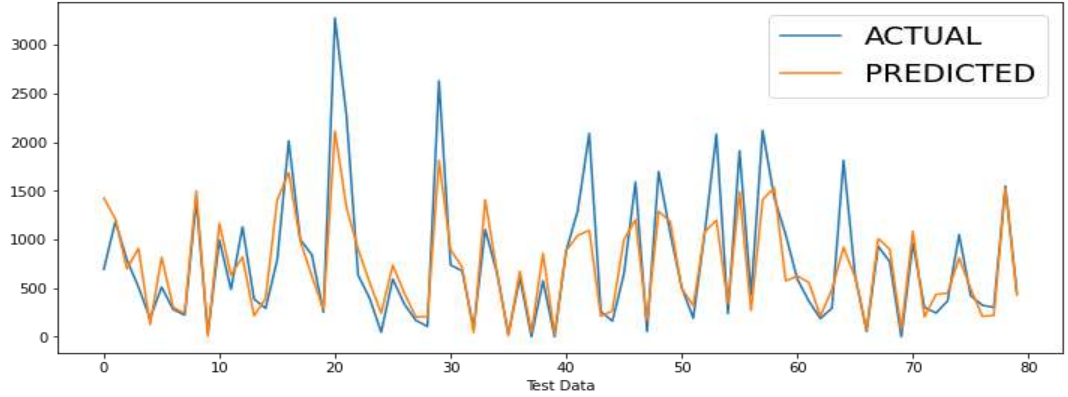
# LINEAR REGRESSION

Training score: 0.811
MAE:203.56
MSE: 91189.56
RMSE: 301.976
R2: 0.77935
Adjusted R2 :  0.7731



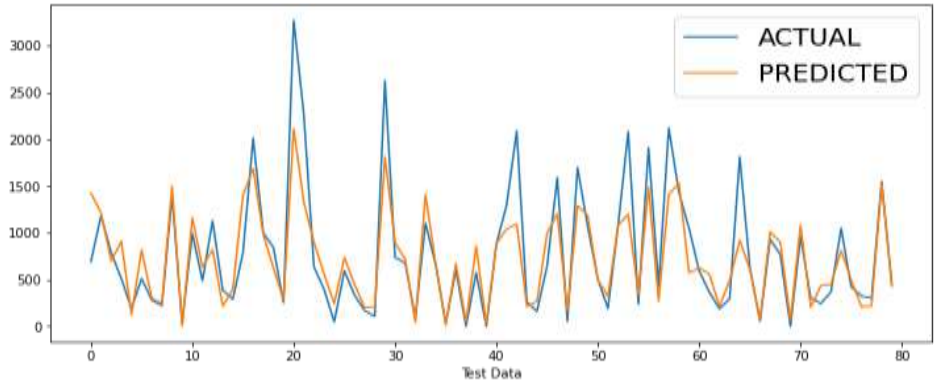

# LASSO REGREGRESSION(HYPER PARAMETER TUNING)

Training score:0.81065
MAE: 203.54
MSE: 91199.11
RMSE: 301.99
R2: 0.779
Adjusted R2 :  0.7731
Alpha(Hyper tuning parameter)=0.001



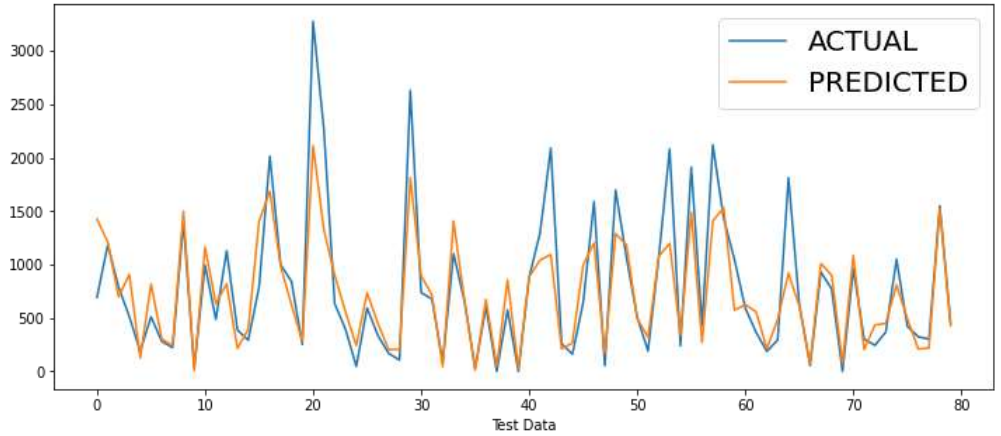
# RIDGE REGRESSION(HYPER PARAMETER TUNING)

- Training Score:0.81065
- MAE: 203.547
- MSE: 91191.198
- RMSE: 301.9788
- R2: 0.7793
- Adjusted R2 :  0.7731
- Ridge Hyper tuning parameter(alpha=5)



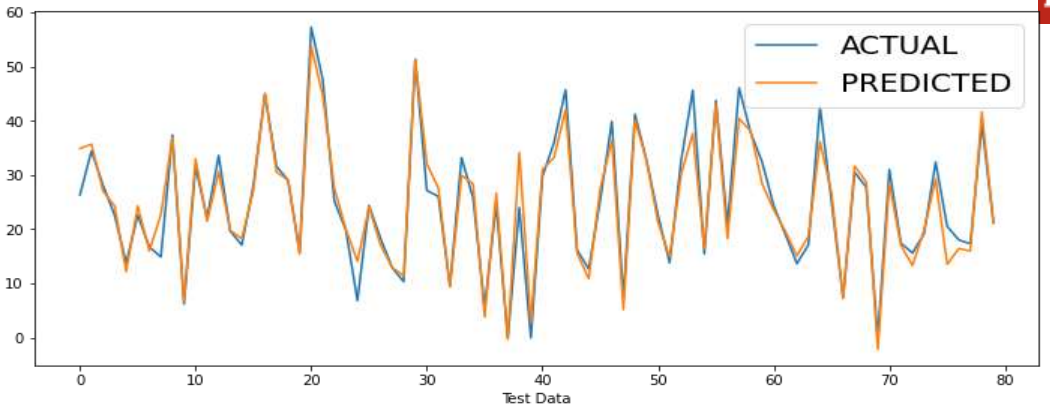

# ELASTICNET REGRESSION(HYPER PARAMETER TUNING)

- Training Score:0.8106
- MAE: 203.5481
- MSE: 91199.8513
- RMSE: 301.9931
- R2: 0.7793
- Adjusted R2 :  0.77311
- Hyper tuning parameter{'alpha': 0.001, 'l1_ratio': 0.5}



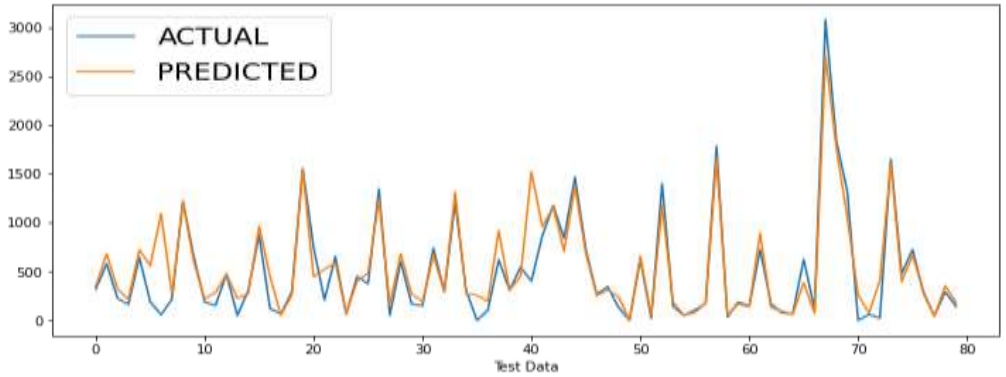
# Polynomial regression

- Training score:0.9408
- MAE: 104.8597
- MSE: 31088.18
- RMSE: 176.318
- R2: 0.924779
- Adjusted R2 :  0.92266



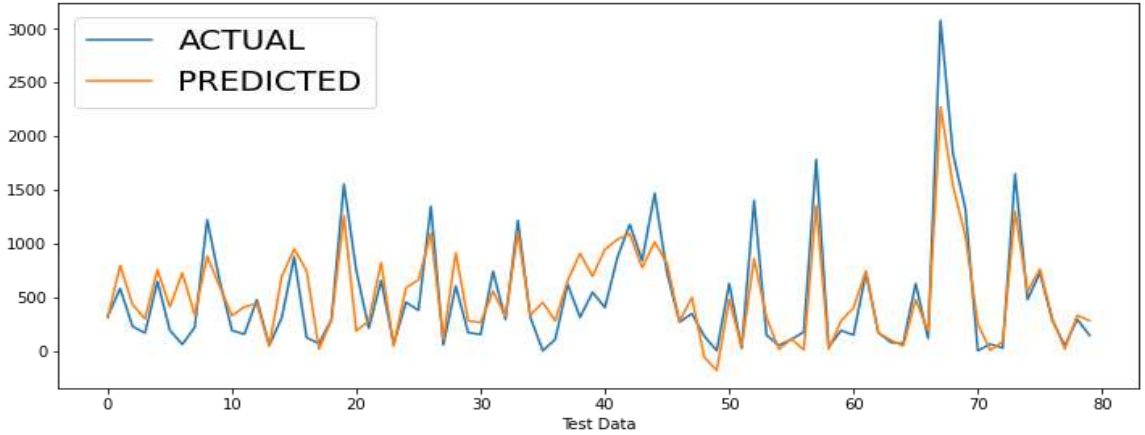

# K Nearest Neighbour

- Training score:0.8877
- MAE: 160.43687
- MSE: 77615.02596
- RMSE: 278.5947
- R2: 0.8146
- Adjusted R2 :  0.8092



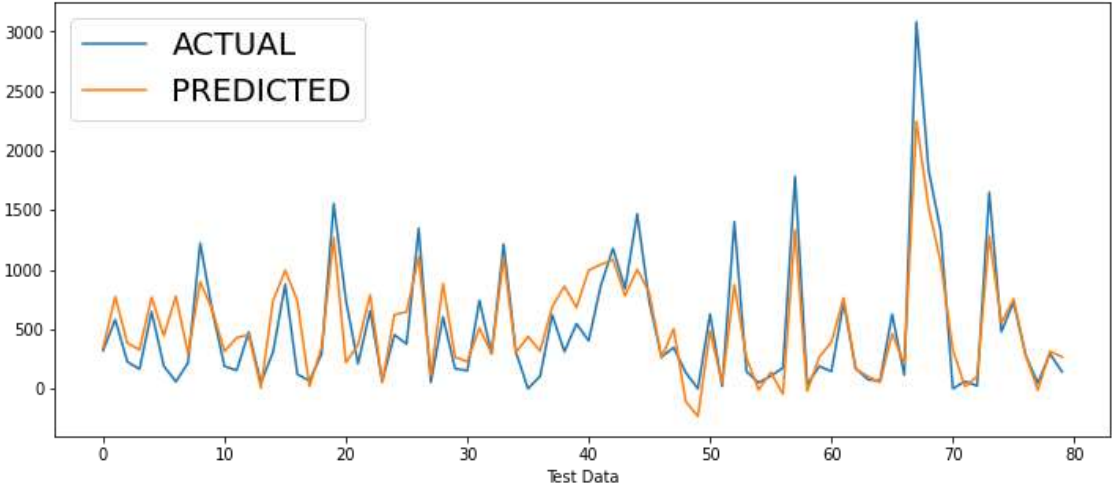
# Gradient Boosting

- Training score :0.8507
- MAE: 189.151
- MSE: 68698.756
- RMSE: 262.104
- R2: 0.83585
- Adjusted R2 : 0.8311



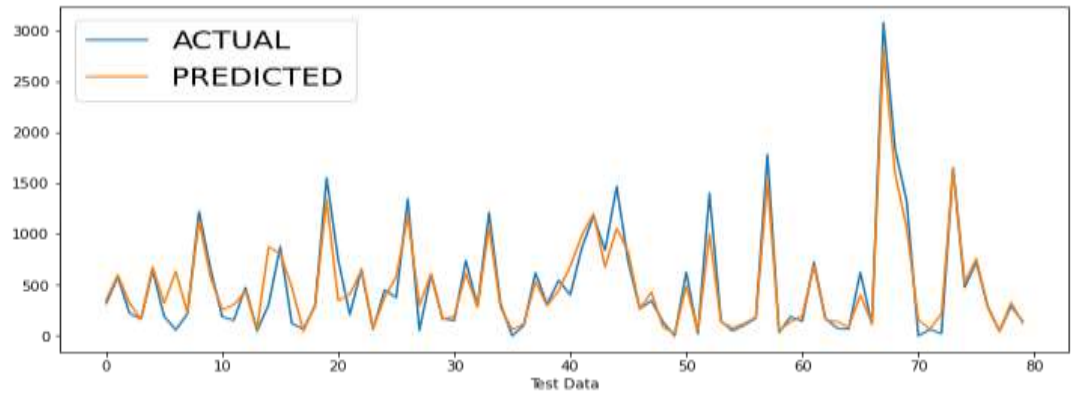

# XG Boost

- Training score:0.8490
- MAE: 188.9805
- MSE: 69274.8417
- RMSE: 263.20114
- R2: 0.8344776514
- Adjusted R2 : 0.8297



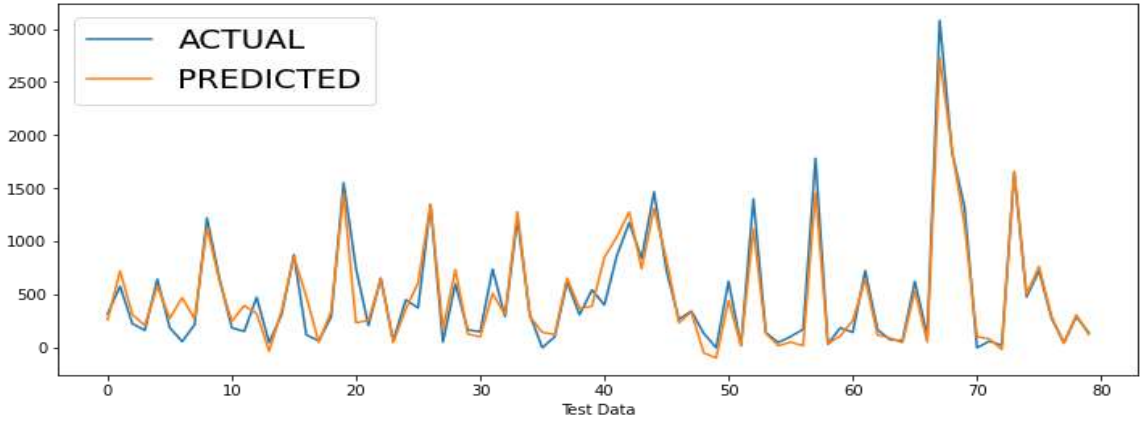
# Random Forest

- Training score:0.9878
- MAE: 114.336
- MSE: 38449.628
- RMSE: 196.086
- R2: 0.9081
- Adjusted R2 : 0.9055



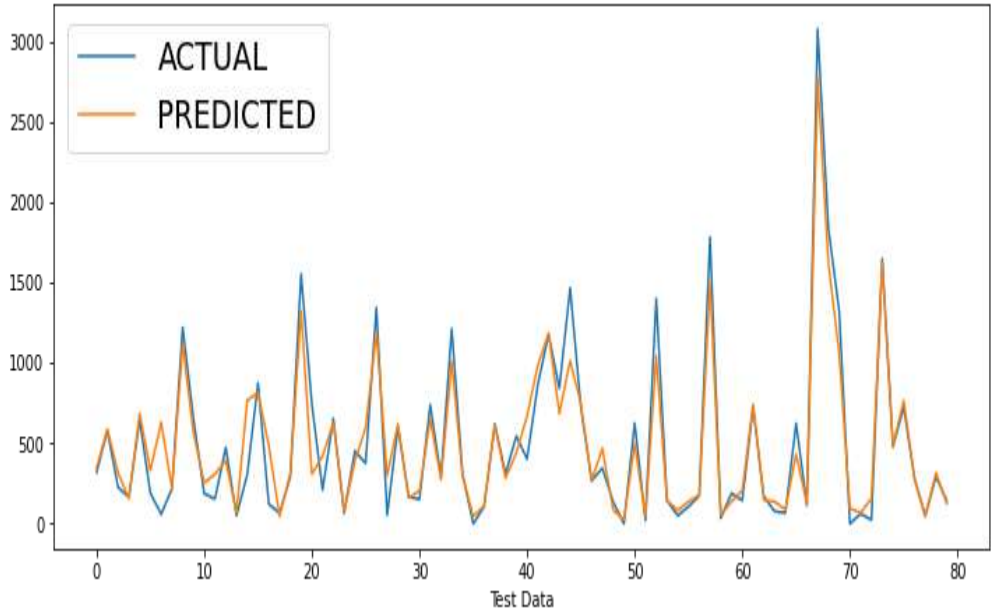

# CATBOOST

- Training score:0.9673
- MAE: 110.9749
- MSE: 29883.2216
- RMSE: 172.8676
- R2: 0.9286
- Adjusted R2 : 0.9265



# Random forest (Hyper parameter tuning)

- Training score:0.9789
- MAE: 118.27608678453713
- MSE: 40473.119781284186
- RMSE: 201.17932244961008
- R2: 0.9032952559842582
- Adjusted R2 :  0.9005111593586581.




**Hyper tuning parameters:**

max depth: 40

min samples leaf:2

min samples split: 2

n estimators: 125}

# XG BOOST grid search CV(Hyper parameter tuning)

- Training score:0.991
- MAE: 111.009
- MSE: 35680.697
- RMSE: 188.893
- R2: 0.9147
- Adjusted R2 : 0.9123

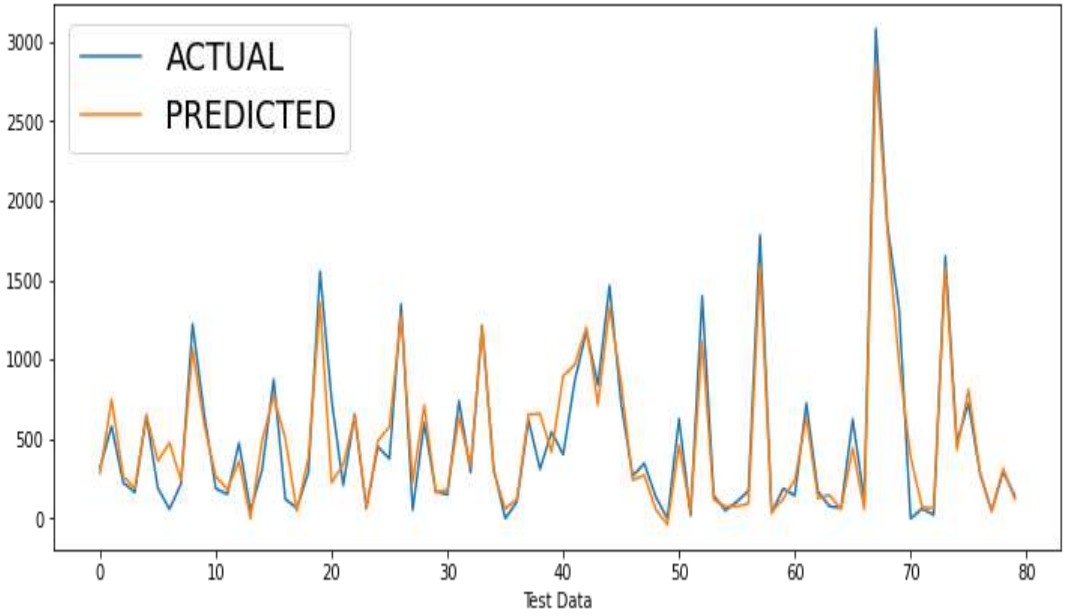
**Hyper tuning parameters**

max_depth: 10

min_samples_leaf: 30

min_samples_split: 10

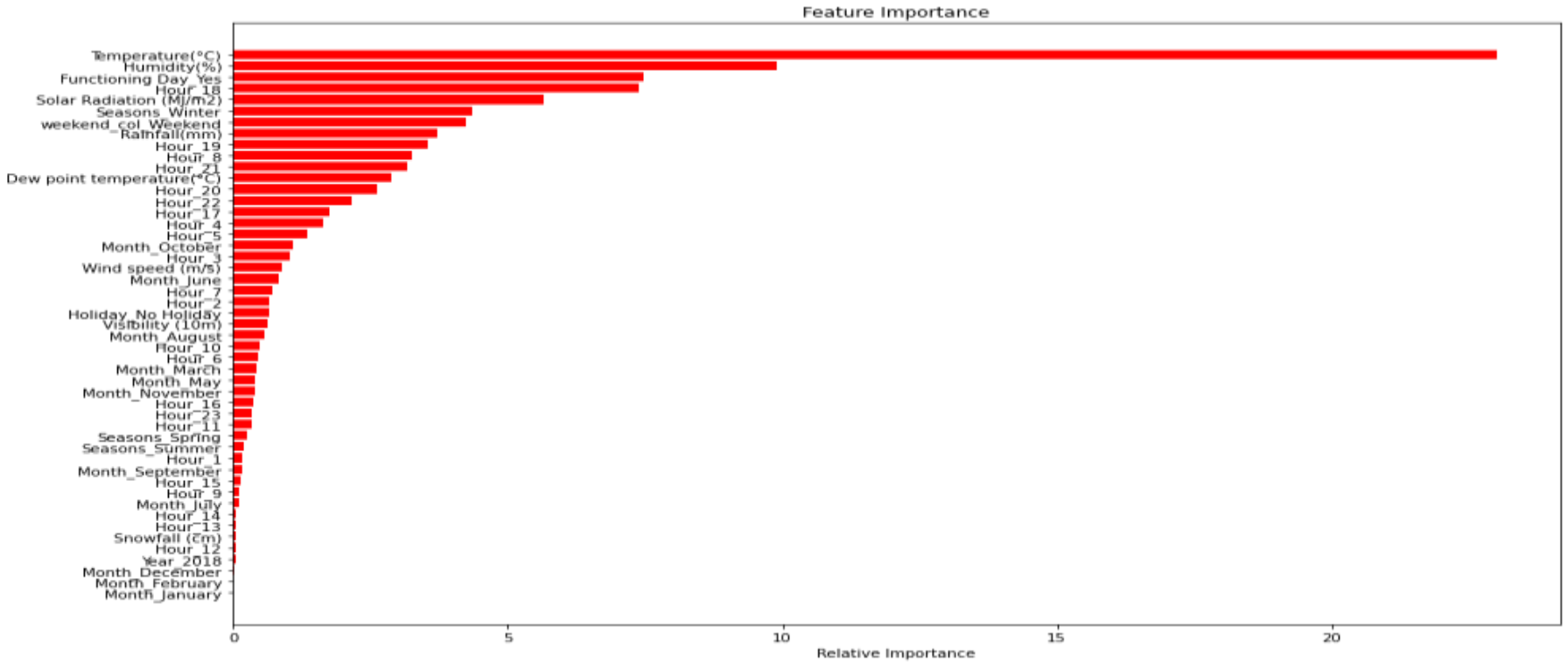
n_estimators: 100



# Metrics for Model Evaluation

| | MODEL NAME | MAE | MSE | RMSE | Training_Accuracy | R2 | Adjusted_R2 |
|---|---|---|---|---|---|---|---|
| 0 | Linear Regression | 203.561096 | 91189.557175 | 301.976087 | 0.810658 | 0.779357 | 0.773138 |
| 1 | Lasso Regression | 203.536653 | 91199.115199 | 301.991912 | 0.810657 | 0.779334 | 0.773115 |
| 2 | Ridge Regression | 203.547532 | 91191.198682 | 301.978805 | 0.810656 | 0.779353 | 0.773134 |
| 3 | Elasticnet Regression | 203.548072 | 91199.851266 | 301.993131 | 0.810656 | 0.779332 | 0.773113 |
| 4 | polynomial Regression | 104.859774 | 31088.182794 | 176.318413 | 0.940829 | 0.924779 | 0.922659 |
| 5 | KNN | 160.436872 | 77615.025959 | 278.594734 | 0.887709 | 0.814550 | 0.809211 |
| 6 | Gradient Boosting | 189.151015 | 68698.756761 | 262.104477 | 0.850676 | 0.835854 | 0.831128 |
| 7 | XGBoost | 188.980488 | 69274.841750 | 263.201143 | 0.849023 | 0.834478 | 0.829712 |
| 8 | Random Forest | 114.336250 | 38449.628269 | 196.085768 | 0.987814 | 0.908130 | 0.905485 |
| 9 | CATBOOST | 110.974926 | 29883.221697 | 172.867642 | 0.967306 | 0.928598 | 0.926543 |
| 10 | RandomForest- GridSearchCV | 118.276087 | 40473.119781 | 201.179322 | 0.978946 | 0.903295 | 0.900511 |
| 11 | XGBoost- GridSearchCV | 111.009659 | 35680.697432 | 188.893349 | 0.991599 | 0.914746 | 0.912292 |

# Feature Importance of Catboost Model



# Conclusion

- **CATBOOST** model delivers the best model that can be used for the Bike Sharing Demand Prediction since the performance metrics (mse, rmse) shows lower and (r2,adjusted r2) shows a higher value.
- CATBOOST has an R2 score of 0.929 and Adjusted R2 score of 0.926.
- This is followed by Polynomial Regression with R2 score of 0.925 and Adjusted R2 score of 0.923.

# THANK YOU