Capstone Project - 2

Seoul Bike Sharing Demand Prediction

ML Supervised Regression

Individual Project:
Soumya Ranjan Mishra

Contents

Problem Statements

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

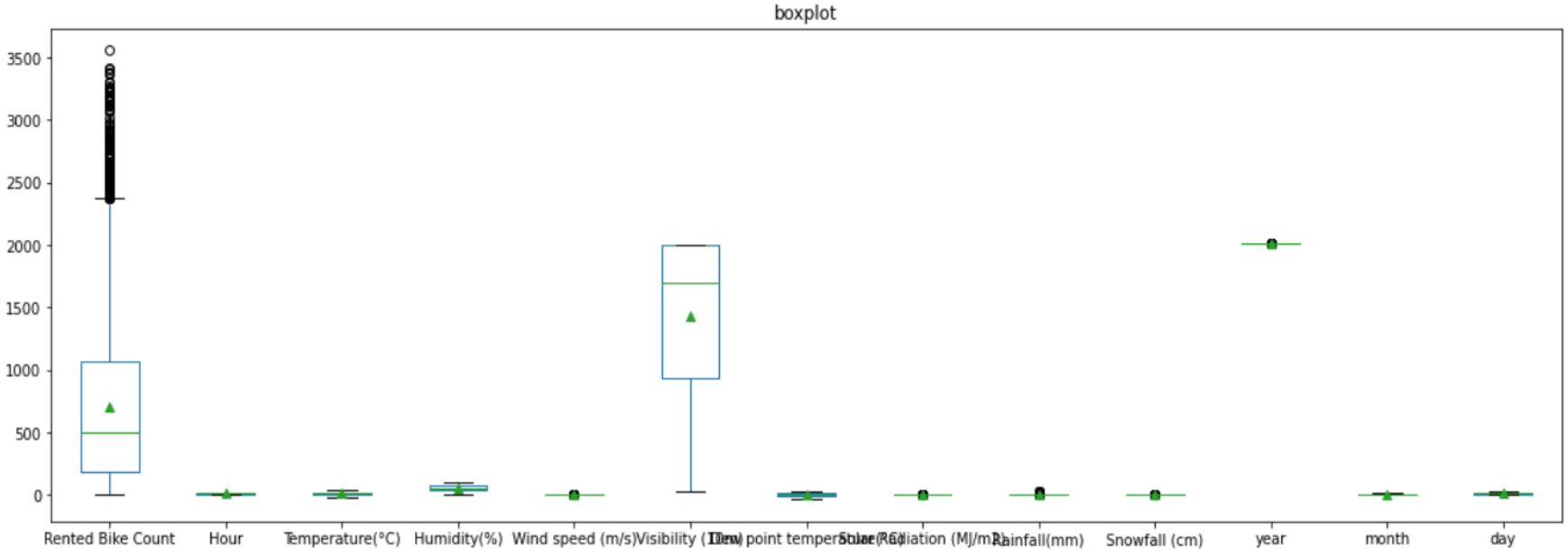
Data Summary

- 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)

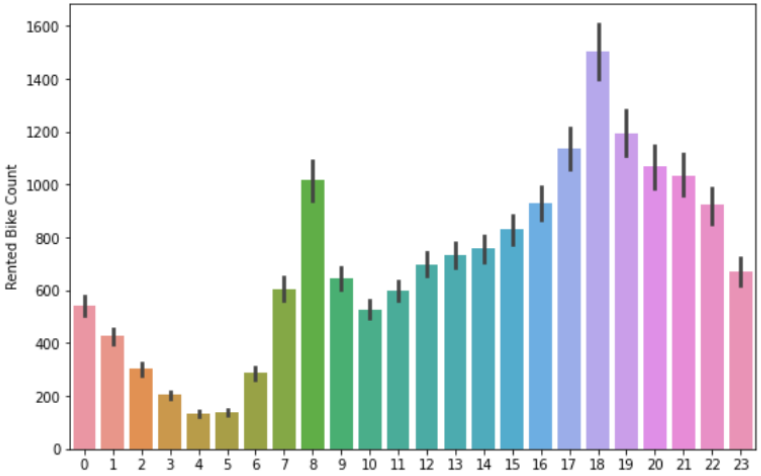
Basic Data Exploration

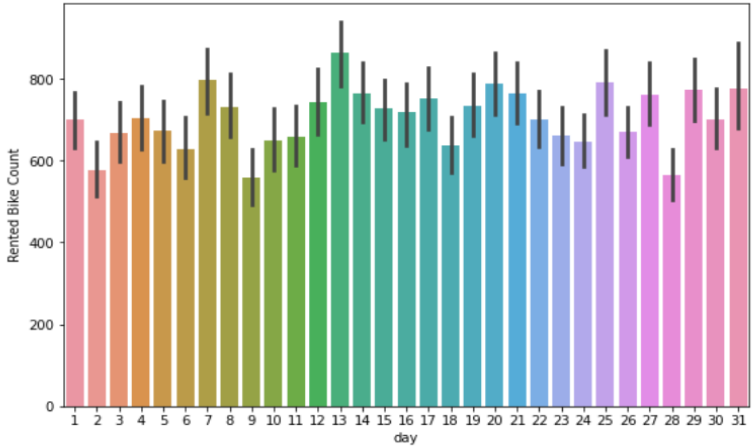
- The dataset has 8760 rows and 14 features(columns).
- Three categorical features ‘Seasons’, ‘Holiday’, & ‘Functioning Day’.
- One Datetime[ns] features ‘Date’.
- Outliers present only in dependent variable.
- No Missing Values.
- No Duplicated values.
- No null values.

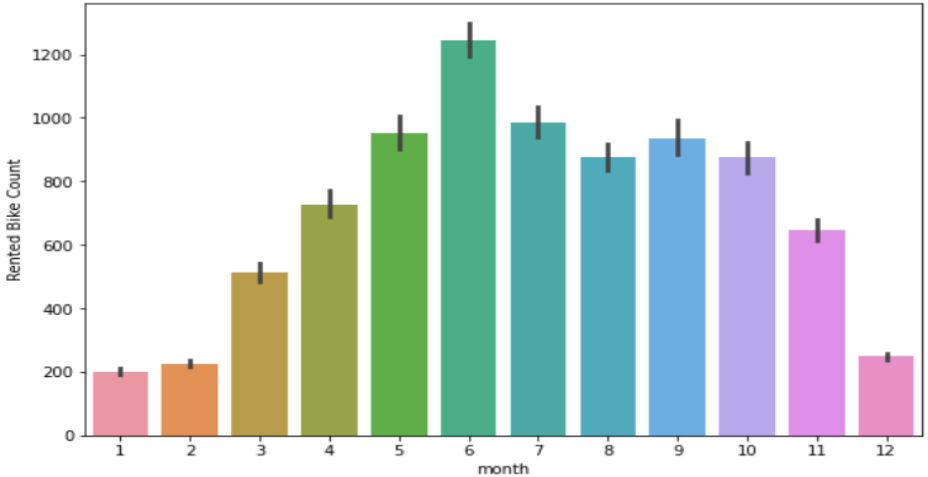
Outliers in the features

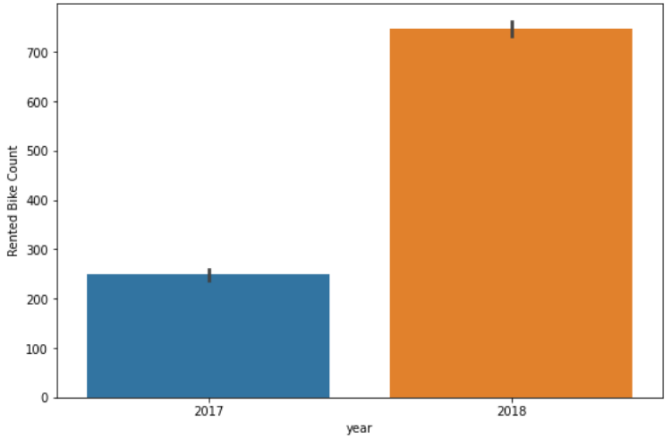


Mean Distribution of Rent Count

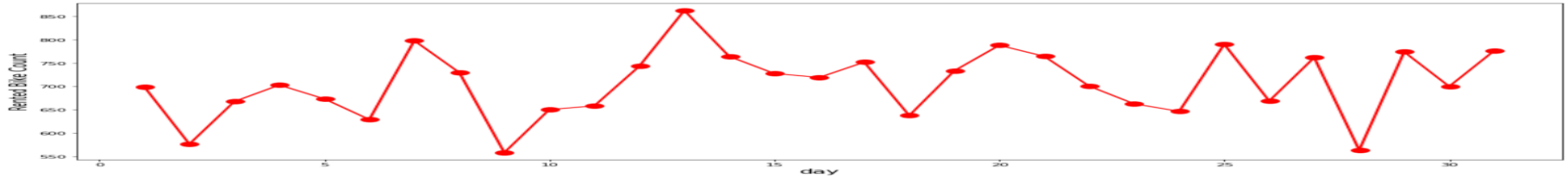


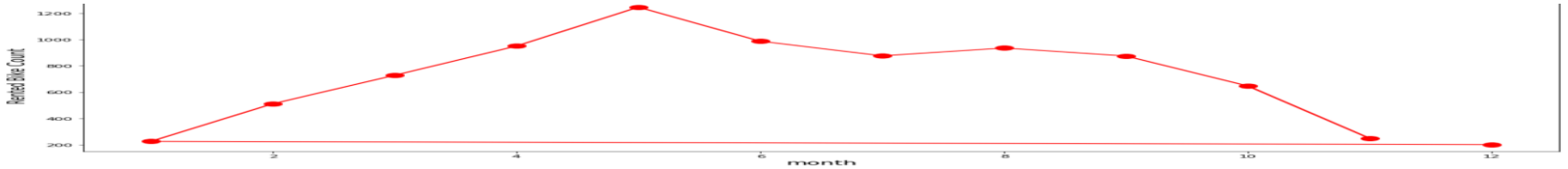


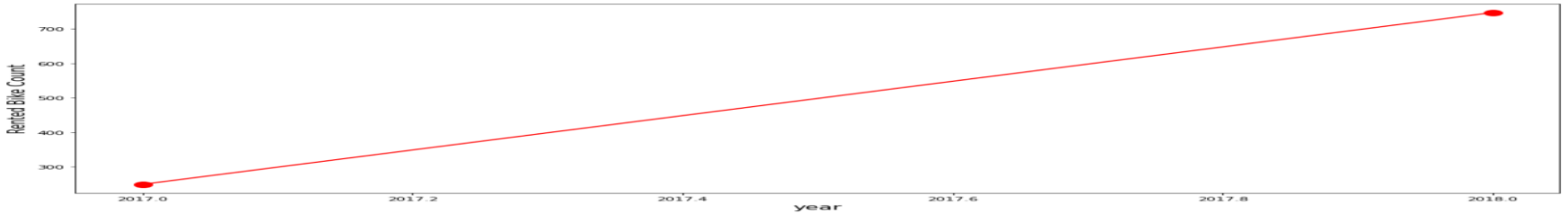


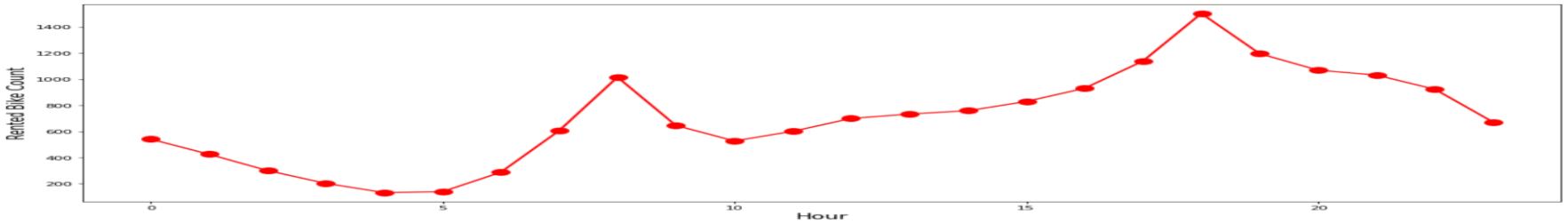


Rented Bike Spread over time

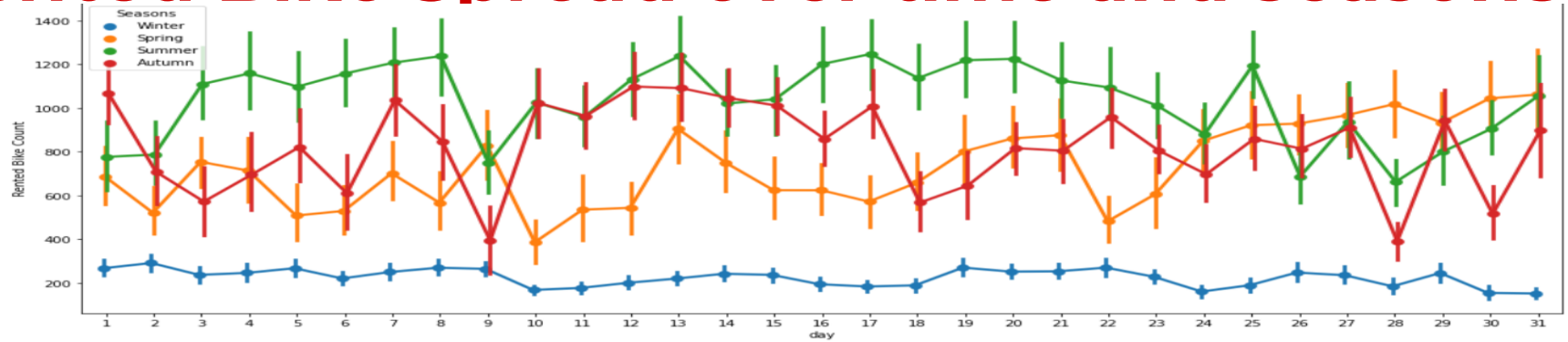


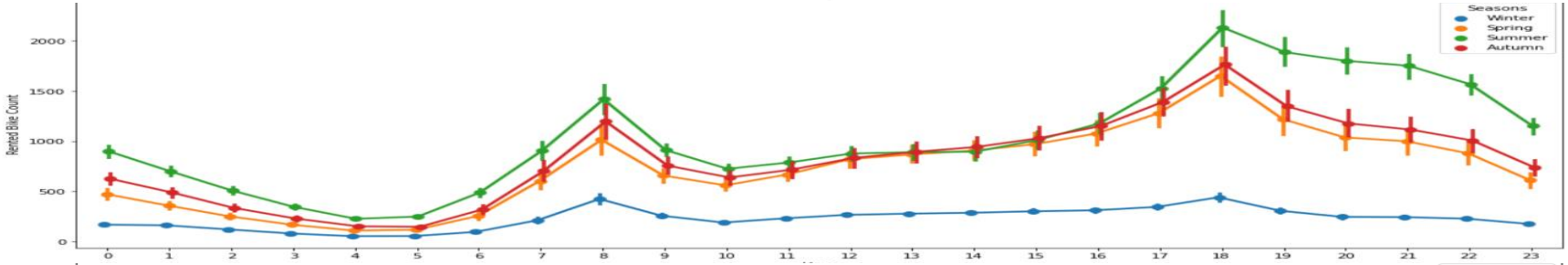


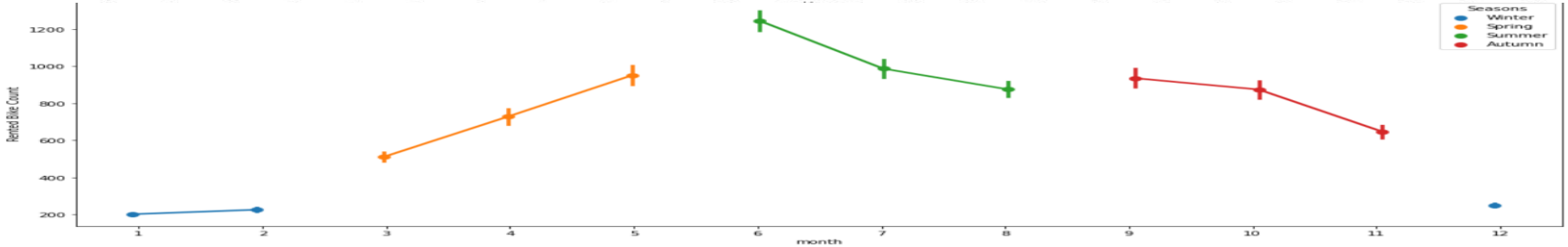




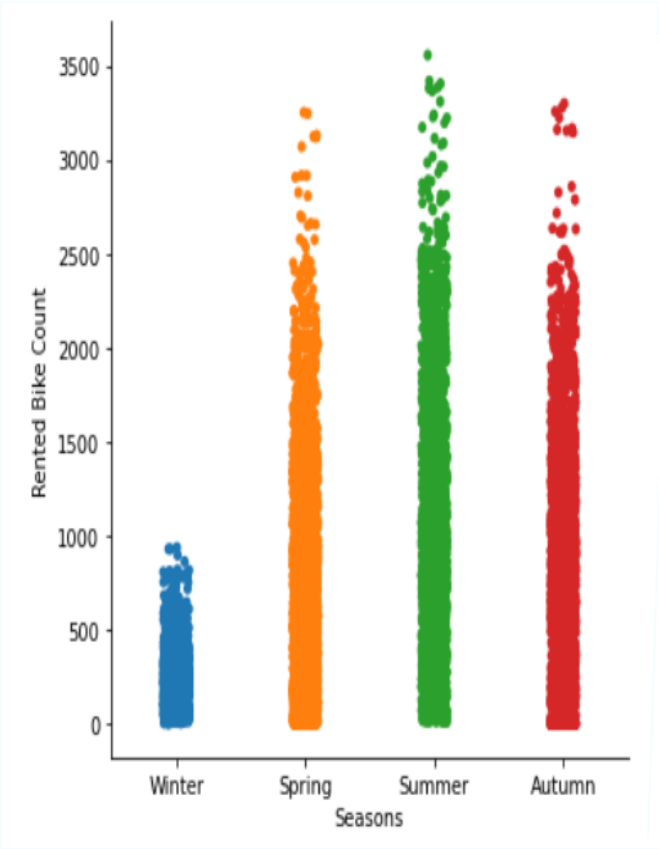
Rented Bike Spread over time and seasons

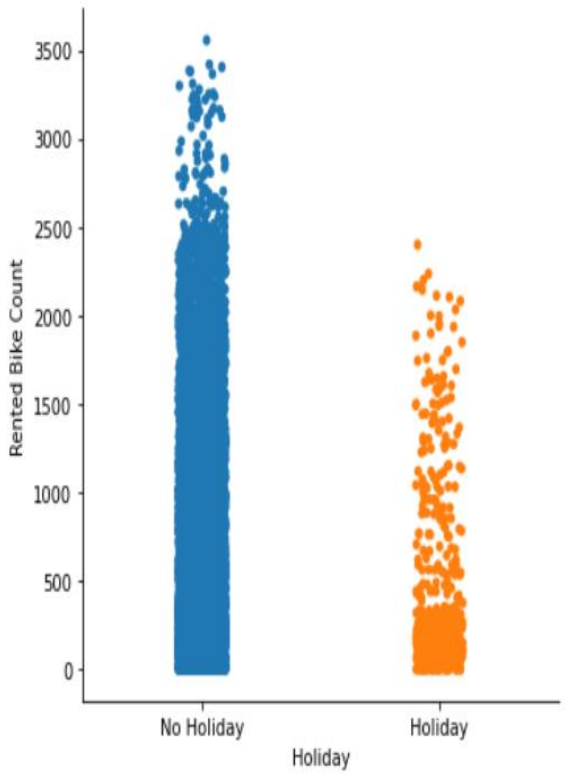


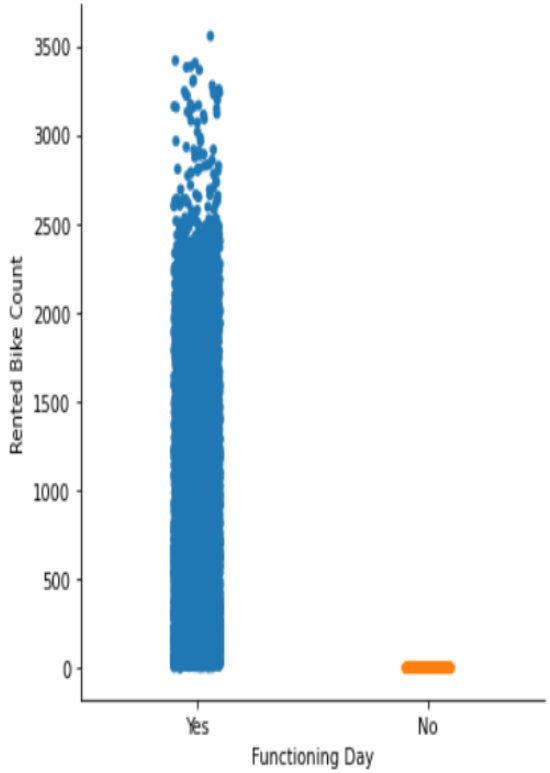




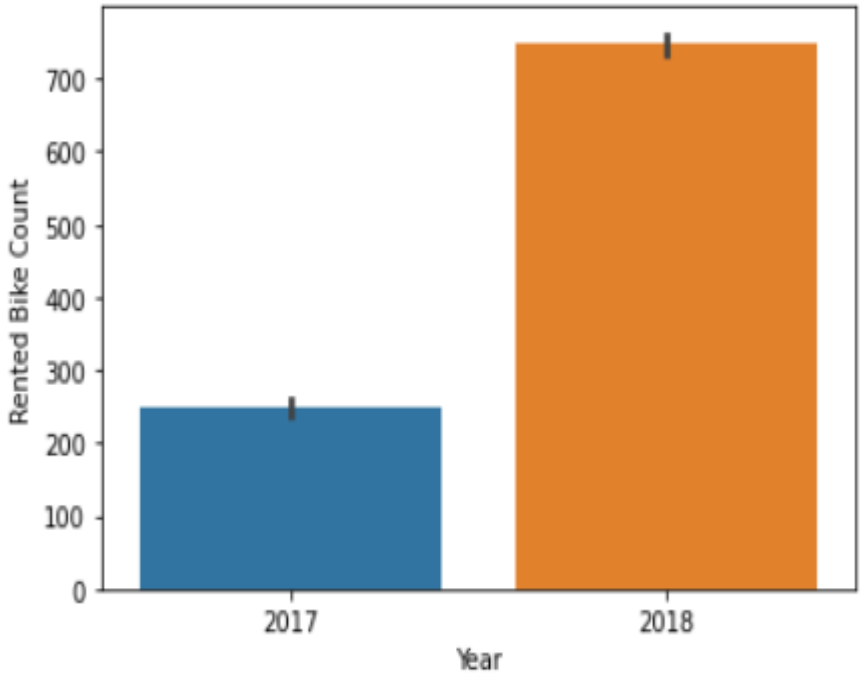
Spread of Categorical Variables



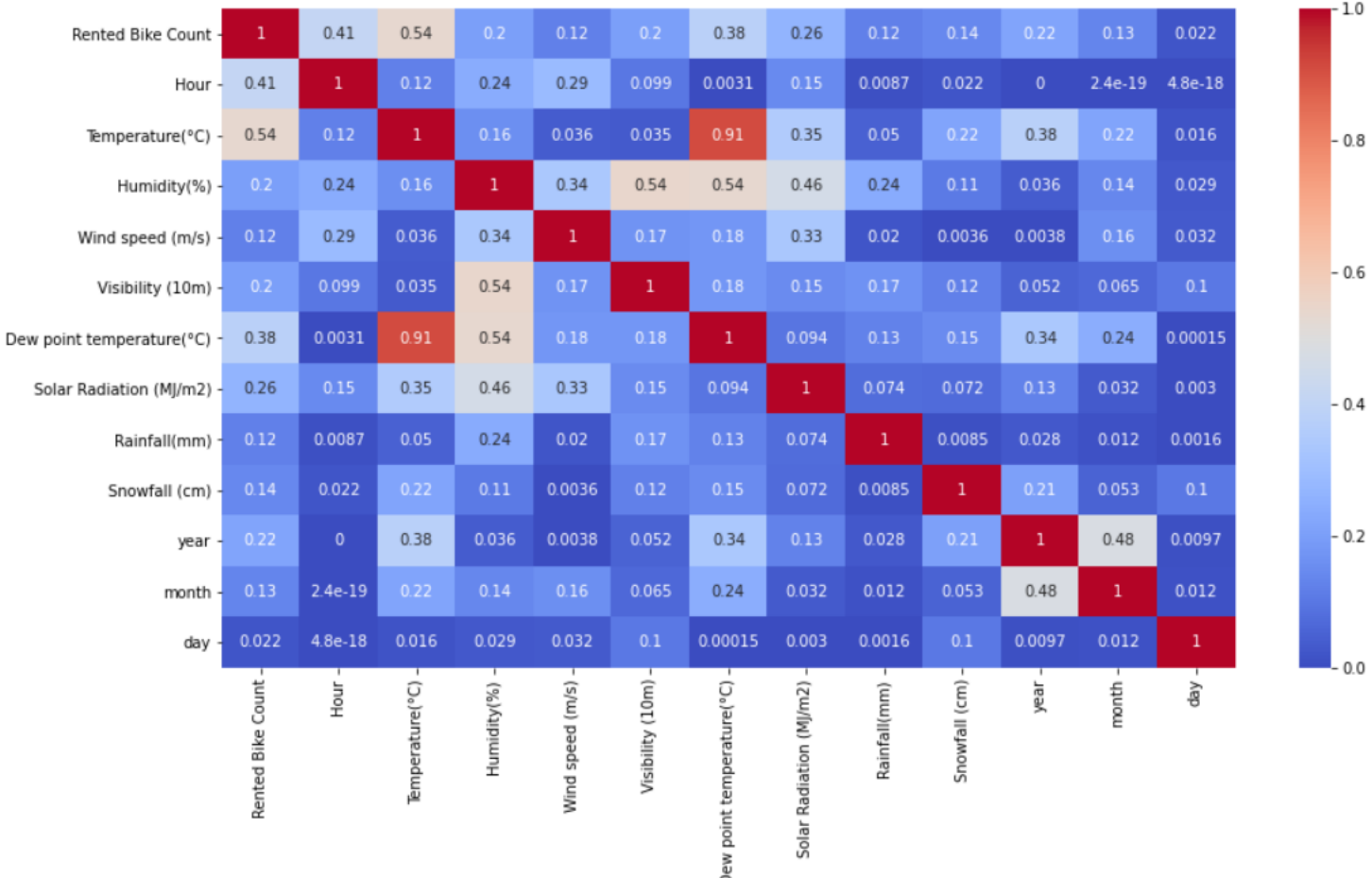




Distribution of Number of reviews



Correlation Matrix



Feature Selection

- Dropping Constant Feature Using Variance Threshold

- Feature Selection with Pearson Correlation

Data Preparation

- **Label Encoding**
- **One Hot Encoding**
- **Train Test Split** **(**test_size=0.3, random_state=0**)**

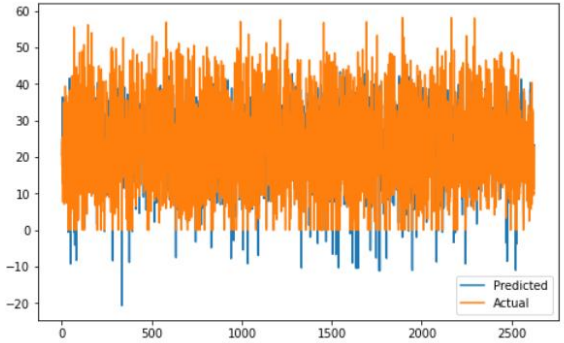
Linear Regression

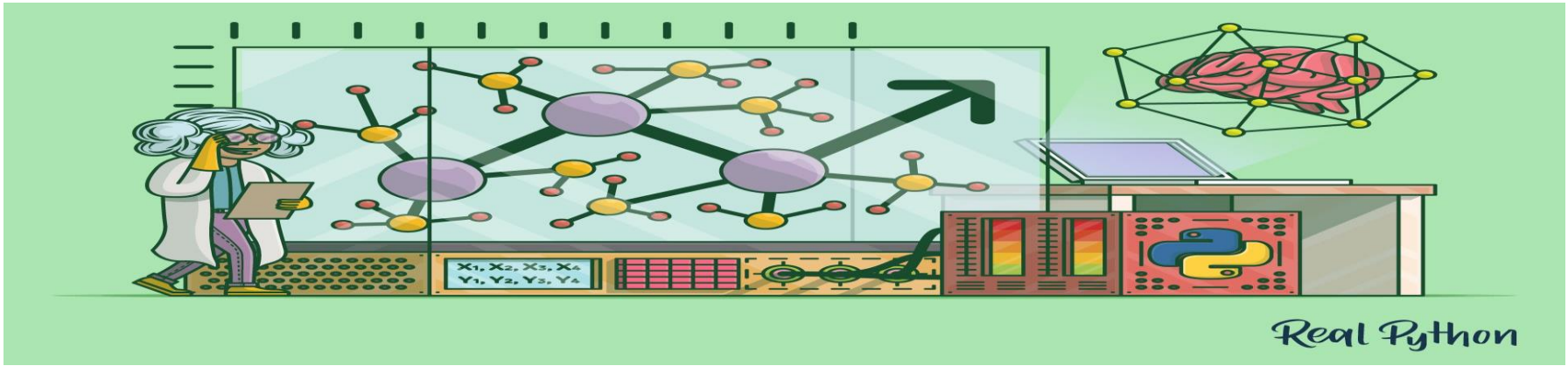
Train Set Metrics

MSE : 57.01388485097972
RMSE : 7.550753925998365
MAE : 5.82426896613067
R2 : 0.6334749885336819
Adjusted R2 : 0.63221497130557

Test Set Metrics

MSE : 57.79693607750042
RMSE : 7.602429616740981
MAE : 5.880703788151969
R2 : 0.6237131747589355
Adjusted R2 : 0.622419598965517





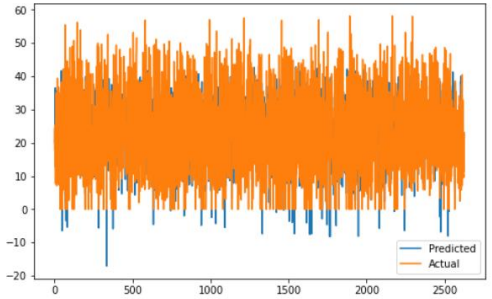
Lasso Regression

Train Set Metrics

MSE : 57.594114200783366
RMSE : 7.589078613427546
MAE : 5.86112981519809
R2 : 0.6297448696399117
Adjusted R2 : 0.6284720292376043

Test Set Metrics

MSE : 57.86197448168322
RMSE : 7.606705888995789
MAE : 5.903113482365391
R2 : 0.6232897423715185
Adjusted R2 : 0.6219947109281814



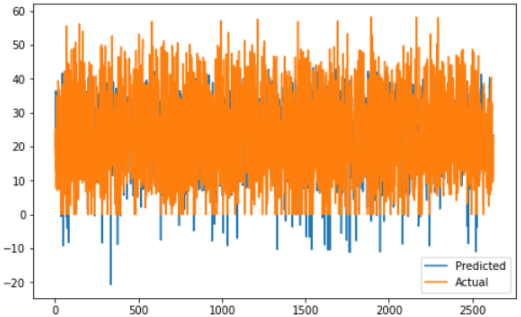
Ridge Regression

Train Set Metrics

MSE : 57.01389125173358
RMSE : 7.550754349847012
MAE : 5.82427110752822
R2 : 0.6334749473851724
Adjusted R2 : 0.6322149300156027

Test Set Metrics

MSE : 57.79594966755792
RMSE : 7.602364741812768
MAE : 5.880681422085658
R2 : 0.6237195967786973
Adjusted R2 : 0.6224260430625048



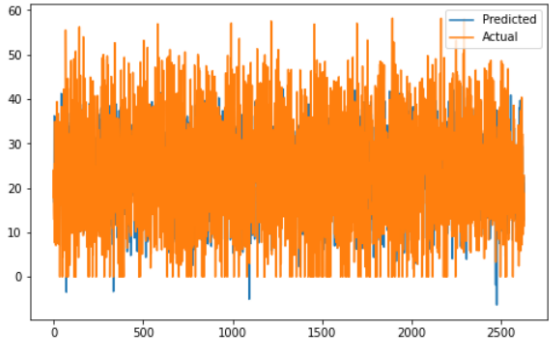
ElasticNet Regression

Train Set Metrics

MSE : 66.62165643815331
RMSE : 8.162209041561807
MAE : 6.228634404920104
R2 : 0.5717095326213393
Adjusted R2 : 0.5702371818931468

Test Set MetricS

MSE : 67.17767602505332
RMSE : 8.196198876616728
MAE : 6.2836301989098615
R2 : 0.5626398879579968
Adjusted R2 : 0.5611363581610609



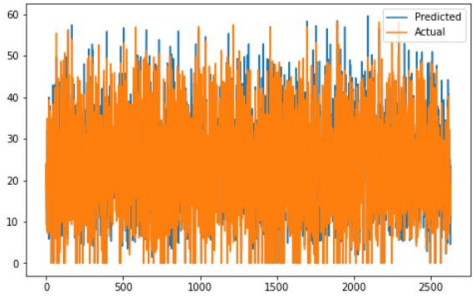
Decision Tree

Train Set Metrics

Model Score: 0.8668051374836275
MSE : 20.71879494354339
RMSE : 4.5517903009193414
MAE : 3.2163583379184484
R2 : 0.8668051374836275
Adjusted R2 : 0.8663472483458707

Test Set MetricS

MSE : 26.072031506670058
RMSE : 5.106077898609661
MAE : 3.5518925524274905
R2 : 0.8302580962064353
Adjusted R2 : 0.8296745678893451



Hyper parameter

```
`(ccp_alpha=0.0, criterion='mse', max_depth=8,
  max_features=9, max_leaf_nodes=100,
  min_impurity_decrease=0.0, min_impurity_split=None,
  min_samples_leaf=1, min_samples_split=2,
  min_weight_fraction_leaf=0.0, presort='deprecated',
  random_state=None, splitter='best')
```

Random Forest

Train Set Metrics

```
Model Score: 0.9853671557886385
MSE : 2.2761756191515756
RMSE : 1.5086999765200422
MAE : 1.008277323907675
R2 : 0.9853671557886385
Adjusted R2 : 0.985316851893336
```

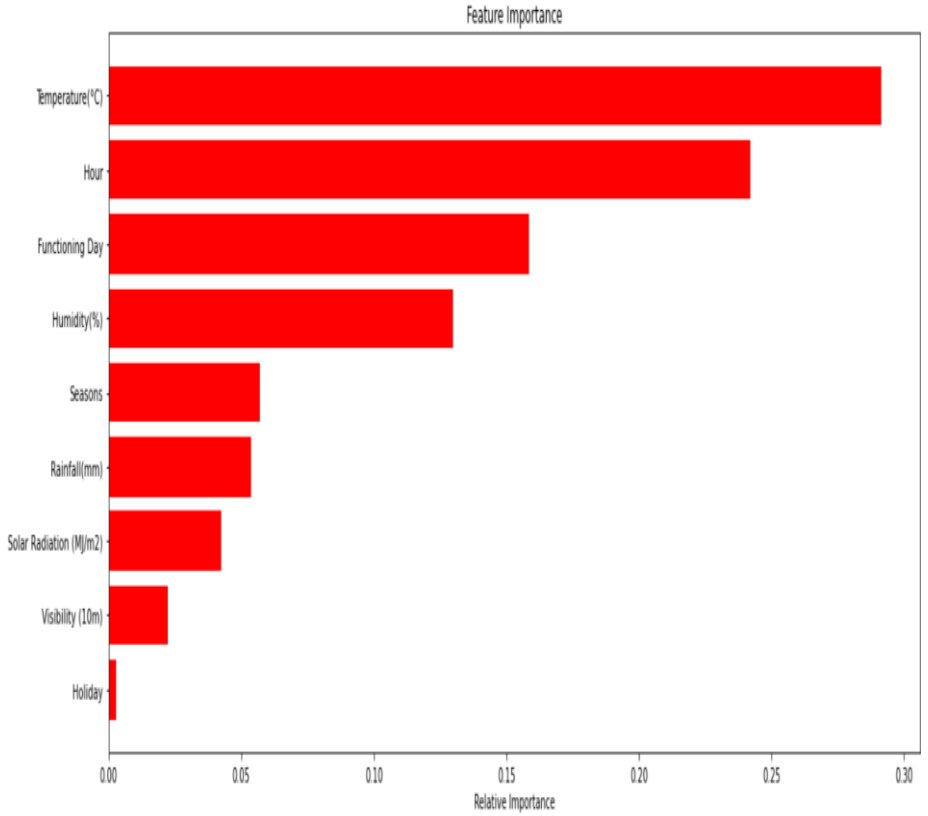
Test Set Metrics

```
MSE : 16.464919522413645
RMSE : 4.057698796413264
MAE : 2.748216698995485
R2 : 0.8928051776545558
Adjusted R2 : 0.8924366698619244
```

Hyper parameter

```
{'max_depth': 8,
 'min_samples_leaf': 40,
 'min_samples_split': 50,
 'n_estimators': 100}
```

Feature Importance



Gradient Boosting Machine

Train Set Metrics

```
Model Score: 0.8854484120234438
MSE : 17.81878682784633
RMSE : 4.221230487410789
MAE : 3.0407216817837943
R2 : 0.8854484120234437
Adjusted R2 : 0.885054613592661
```

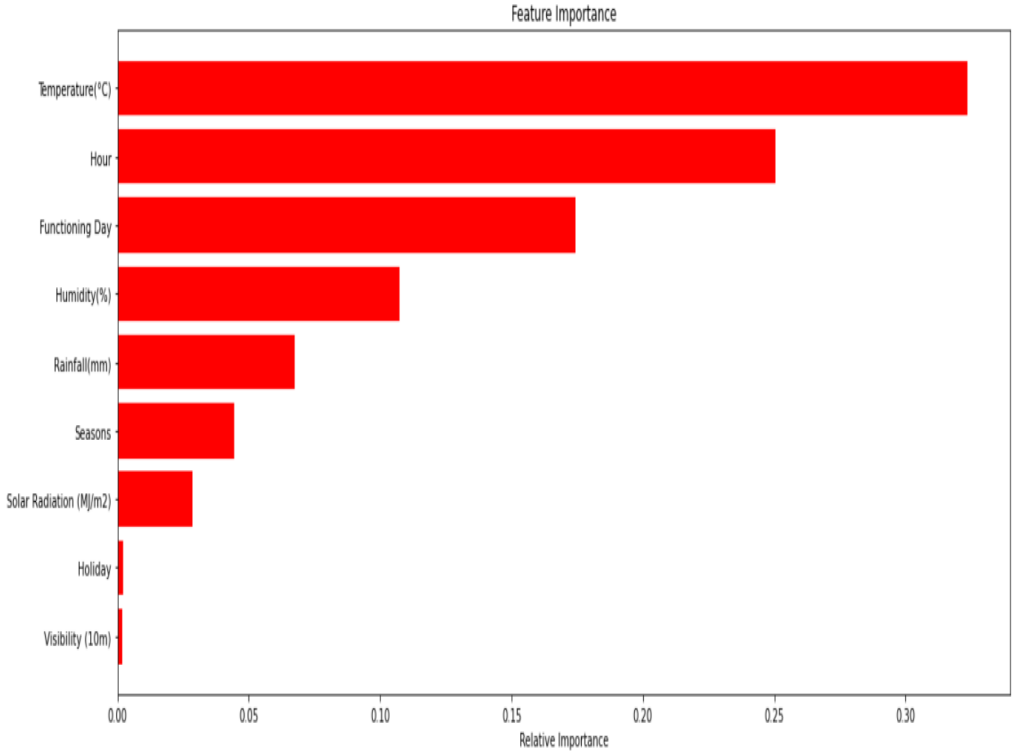
Test Set Metrics

```
MSE : 20.166419581951853
RMSE : 4.4907036844966575
MAE : 3.223205052853635
R2 : 0.8687065696562772
Adjusted R2 : 0.8682552171455463
```

Hyper parameter

```
{'max_depth': 8,
 'min_samples_leaf': 40,
 'min_samples_split': 100,
 'n_estimators': 100}
```

Feature Importance



XGBoost

Train Set Metrics

```
Model Score: 0.9668007464593723
MSE : 5.164227158554409
RMSE : 2.27249359923288
MAE : 1.5992926315221427
R2 : 0.9668007464593723
Adjusted R2 : 0.9666866160996069
```

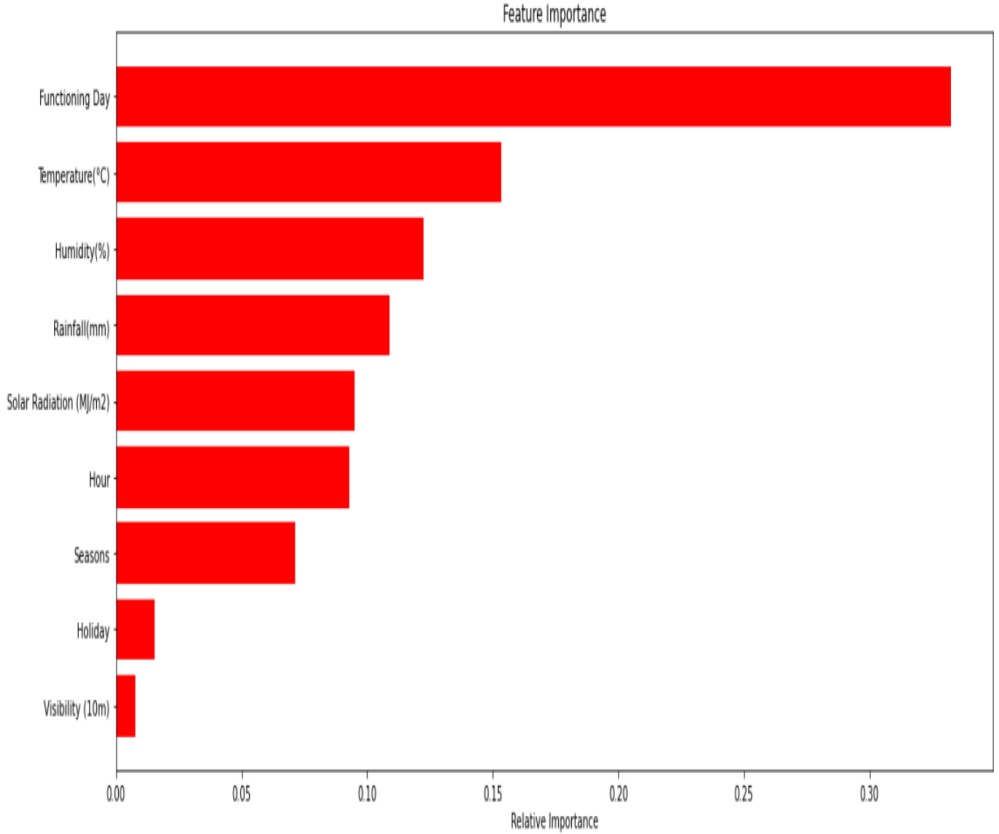
Test Set Metrics

```
MSE : 16.26765784833295
RMSE : 4.033318466019383
MAE : 2.727763409019645
R2 : 0.8940894493498932
Adjusted R2 : 0.893725356547811
```

Hyper parameter

```
{'max_depth': 8,
 'min_samples_leaf': 40,
 'min_samples_split': 50,
 'n_estimators': 80}
```

Feature Importance



Linear Regression using Statsmodels

OLS Regression Results

Dep. Variable:	Rented Bike Count	R-squared:	0.631
Model:	OLS	Adj. R-squared:	0.630
Method:	Least Squares	F-statistic:	1661.
Date:	Wed, 14 Apr 2021	Prob (F-statistic):	0.00
Time:	15:22:08	Log-Likelihood:	-30156.
No. Observations:	8760	AIC:	6.033e+04
Df Residuals:	8750	BIC:	6.040e+04
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-14.8846	0.864	-17.230	0.000	-16.578	-13.191
Hour	0.4756	0.012	38.653	0.000	0.451	0.500
Temperature(°C)	0.6406	0.008	75.773	0.000	0.624	0.657
Humidity(%)	-0.1371	0.006	-21.936	0.000	-0.149	-0.125
Visibility (10m)	0.0010	0.000	5.777	0.000	0.001	0.001
Solar Radiation (MJ/m2)	-1.0179	0.124	-8.177	0.000	-1.262	-0.774
Rainfall(mm)	-1.5598	0.074	-21.057	0.000	-1.705	-1.415
Seasons	1.2266	0.076	16.182	0.000	1.078	1.375
Holiday	3.2368	0.375	8.639	0.000	2.502	3.971
Functioning Day	26.7665	0.453	59.023	0.000	25.878	27.655

Omnibus:	230.274	Durbin-Watson:	0.480
Prob(Omnibus):	0.000	Jarque-Bera (JB):	315.896
Skew:	0.300	Prob(JB):	2.53e-69
Kurtosis:	3.711	Cond. No.	1.77e+04

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

Q & A