Capstone Project-2

Project Title

Seoul Bike Sharing Demand Prediction

By-Bhushan Patil

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.

(a)

(b)

Data Summary

- Seoul Bike Sharing Demand Prediction
  - Categorical
    - Date
    - Seasons
    - Holiday
    - Functioning Day
  - Numerical
    - Rented Bike Count
    - Hour
    - Temperature(°C)
    - Humidity(%)
    - Wind speed (m/s)
    - Visibility (10m)
    - Dew point temperature(°C)
    - Solar Radiation (MJ/m2)
    - Rainfall(mm)
    - Snowfall (cm)

Dataset

- There are 8760 entries and 14 columns

- 10 out of 14 are numeric:
(Rented Bike Count, Hour, Temperature(°C), Humidity(%), Wind speed (m/s), Visibility (10m), Dew point temperature(°C), Solar Radiation (MJ/m2), Rainfall(mm), Snowfall (cm))

- 4 out of 14 are categorical:
(Date, Seasons, Holiday, Functioning Day)

```
RangeIndex: 8760 entries, 0 to 8759
Data columns (total 14 columns):
 #   Column                     Non-Null Count  Dtype  
---  ------                     --------------  -----  
 0   Date                       8760 non-null   object 
 1   Rented Bike Count          8760 non-null   int64  
 2   Hour                       8760 non-null   int64  
 3   Temperature(°C)            8760 non-null   float64
 4   Humidity(%)                8760 non-null   int64  
 5   Wind speed (m/s)           8760 non-null   float64
 6   Visibility (10m)           8760 non-null   int64  
 7   Dew point temperature(°C)  8760 non-null   float64
 8   Solar Radiation (MJ/m2)    8760 non-null   float64
 9   Rainfall(mm)               8760 non-null   float64
 10  Snowfall (cm)              8760 non-null   float64
 11  Seasons                    8760 non-null   object 
 12  Holiday                    8760 non-null   object 
 13  Functioning Day            8760 non-null   object 
dtypes: float64(6), int64(4), object(4)
```

Data Description

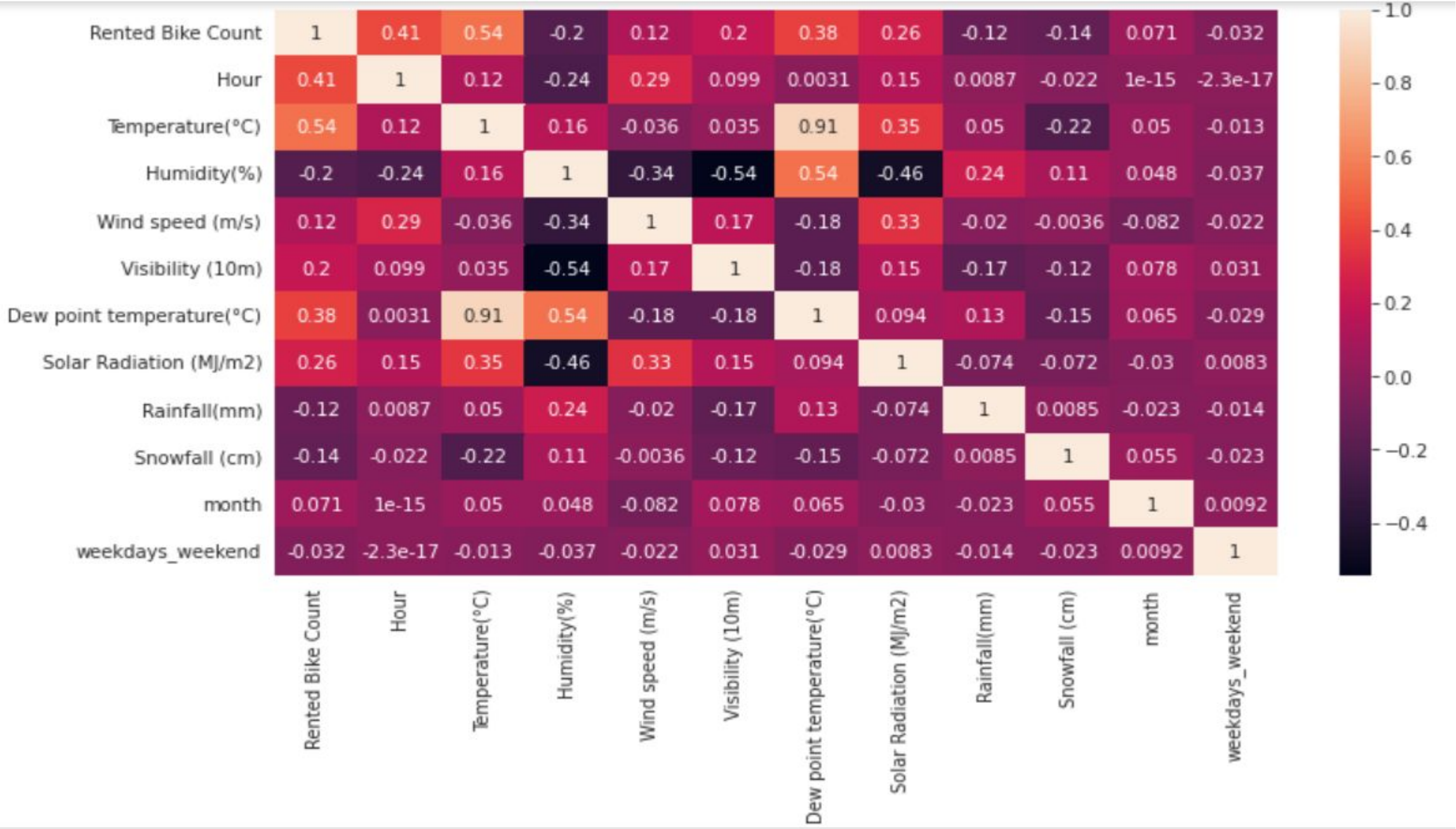
Dependent variable:

- Rented Bike count

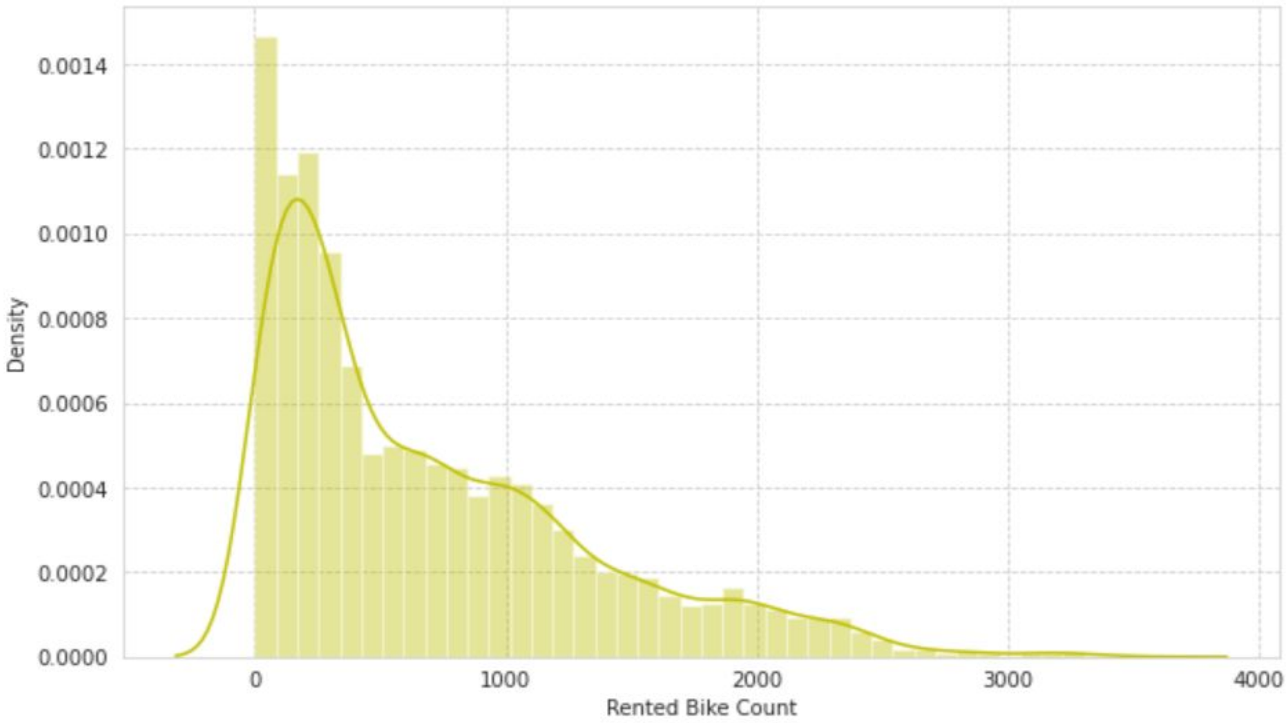
Independent variables:

- Date
- Hour
- Temperature(°C)
- Humidity(%)
- Wind speed (m/s)
- Visibility (10m)
- Functioning Day
- Dew point temperature(° C)
- Solar Radiation (MJ/m2)
- Rainfall(mm)
- Snowfall (cm)
- Seasons
- Holiday

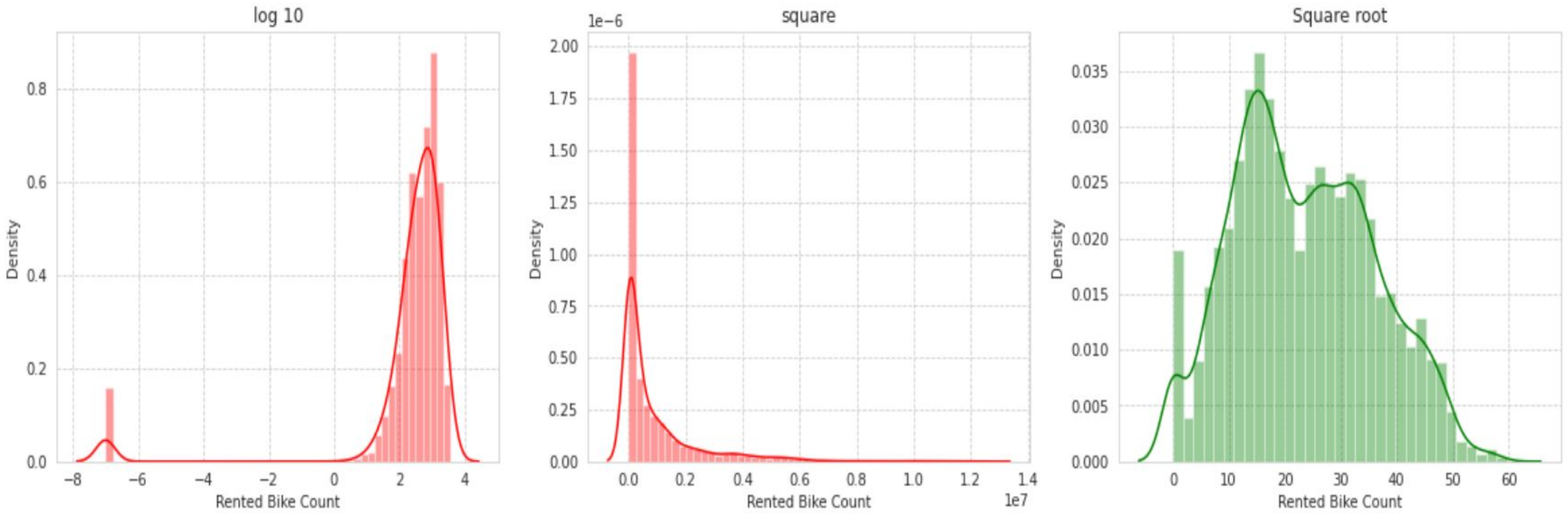
EDA Correlation matrix



EDA Distribution of rented bike count

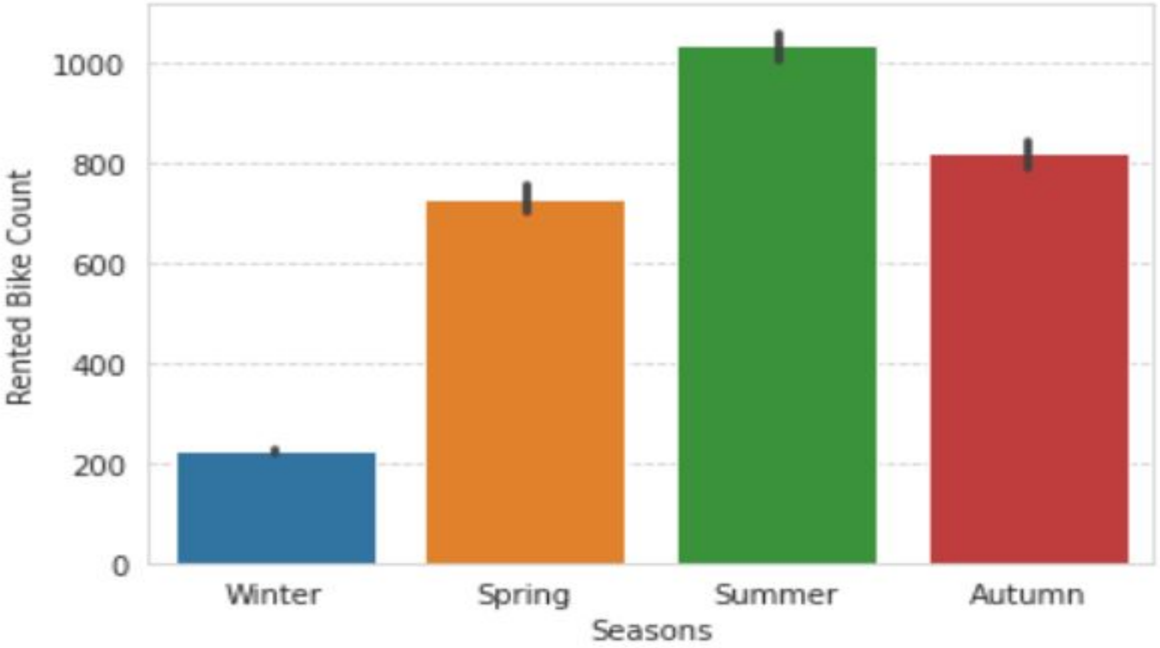


EDA Transformation of distribution



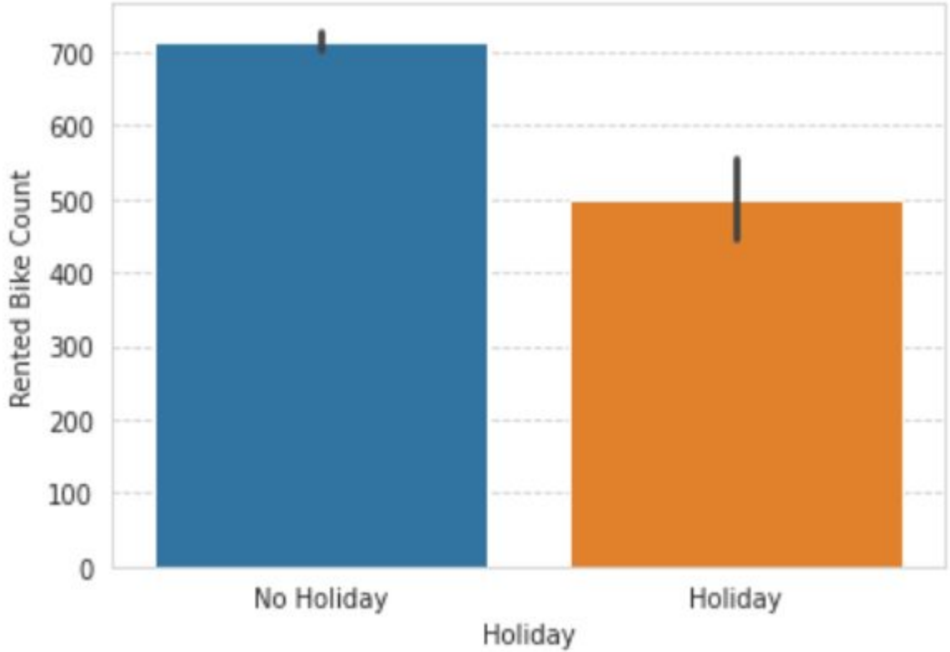
EDA Rented bike count and Seasons

- Higher bike demand on Summer season and
- Less bike demand on Winter season



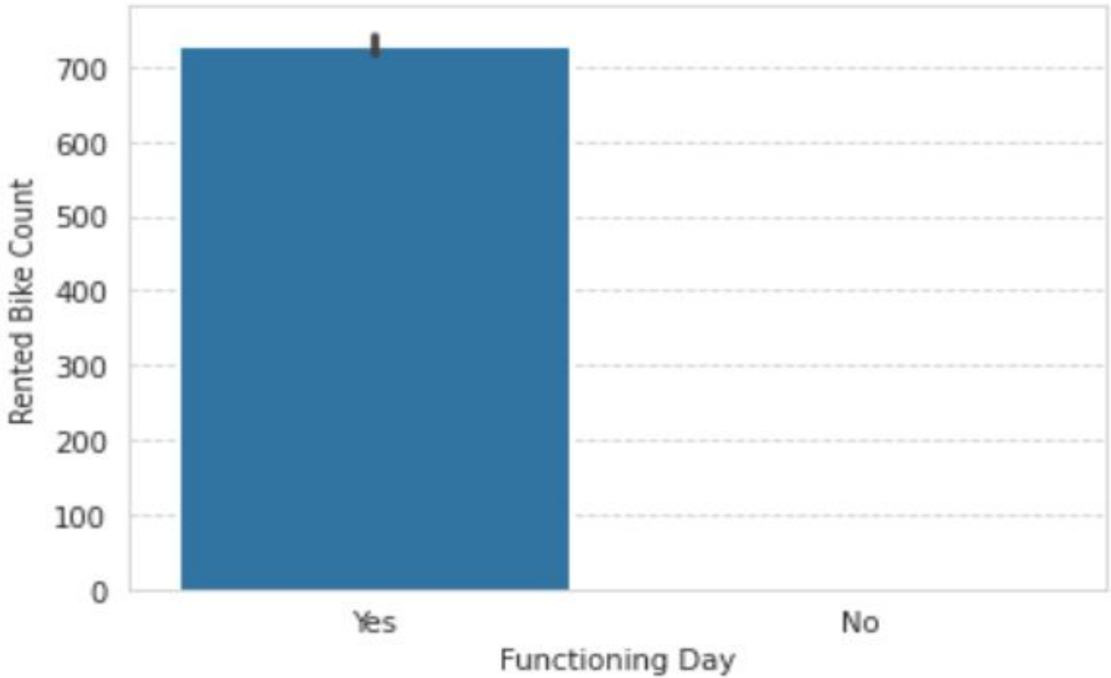
EDA Rented bike count and Holiday

- Slightly Higher demand of bike during Non holidays



EDA Rented bike count and Functioning day

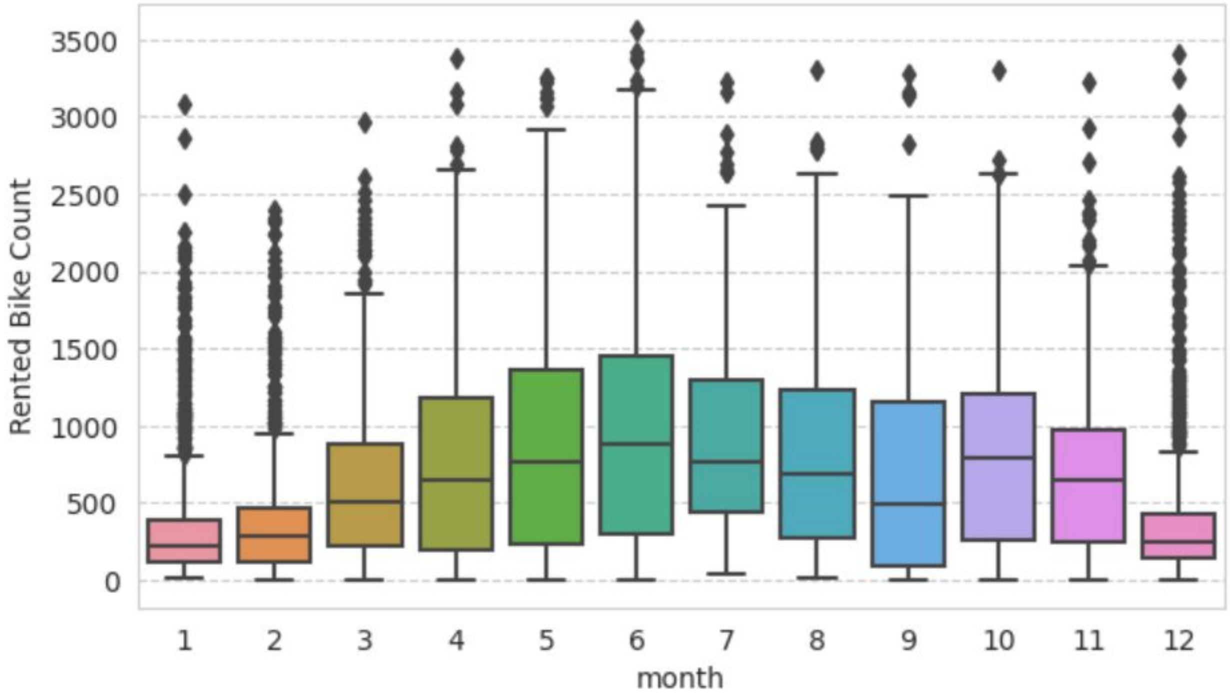
- Almost no demand of bike on non functioning day



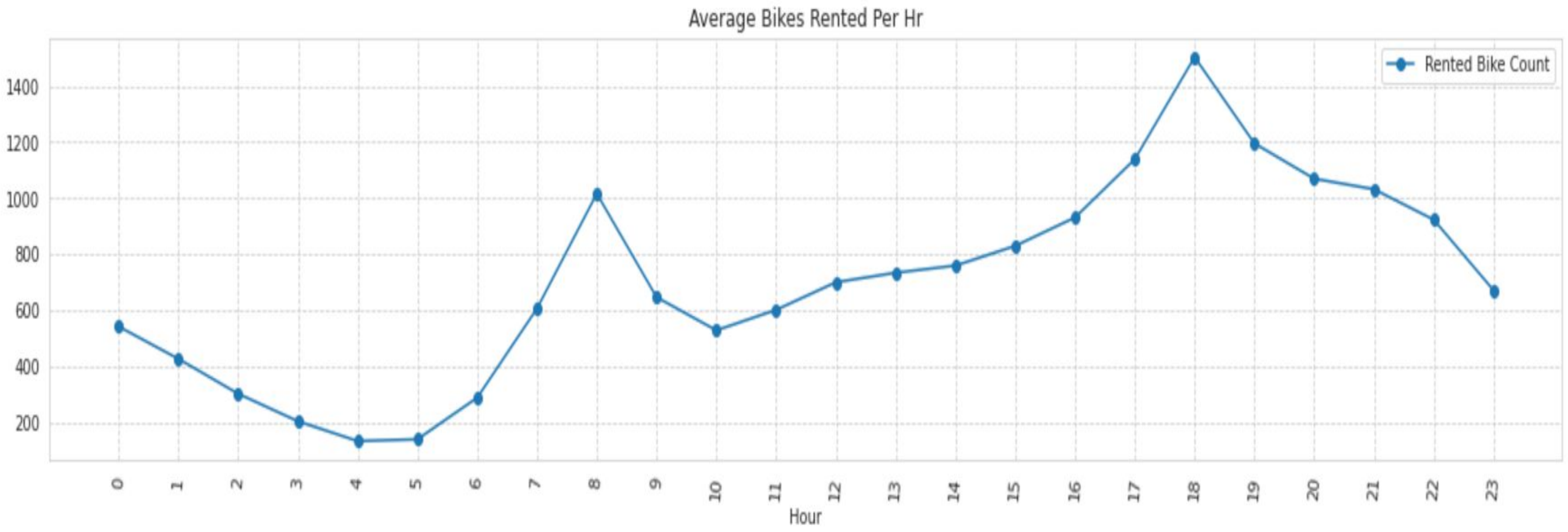
EDA Rented bike count and Month

- We can see that there is less demand of Rented bike in the month of December, January, February i.e. during winter seasons.

- Also demand of bike is maximum during May, June, July i.e. Summer seasons.

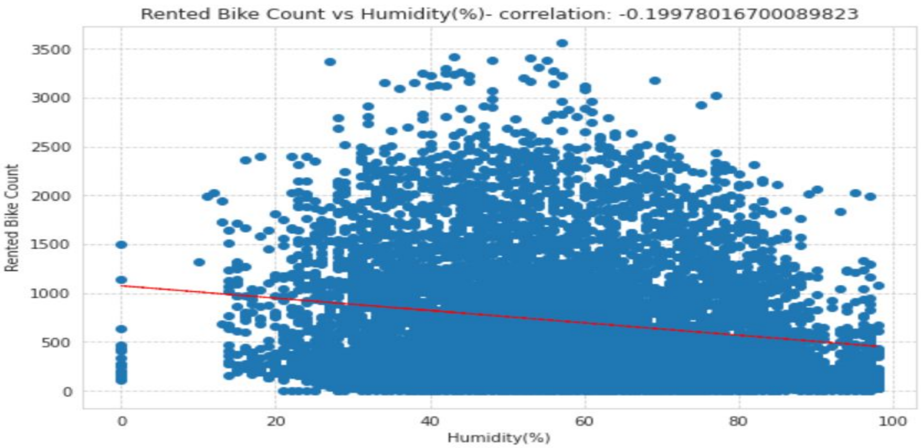


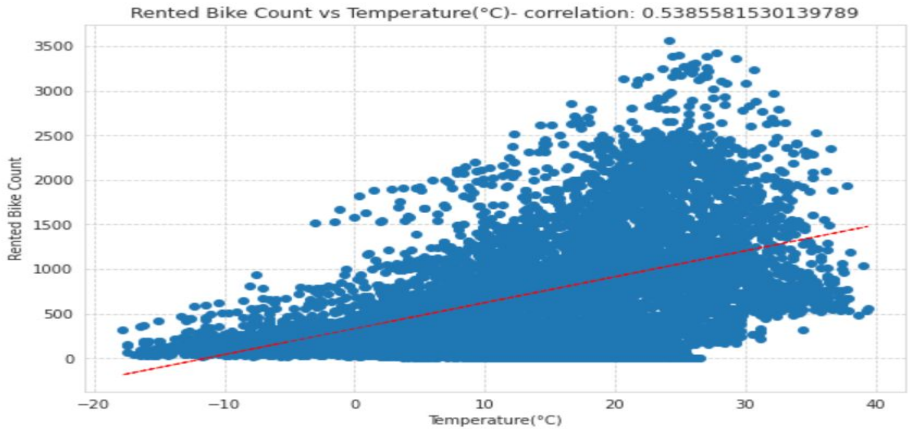
EDA Average bike rented over time(hr)



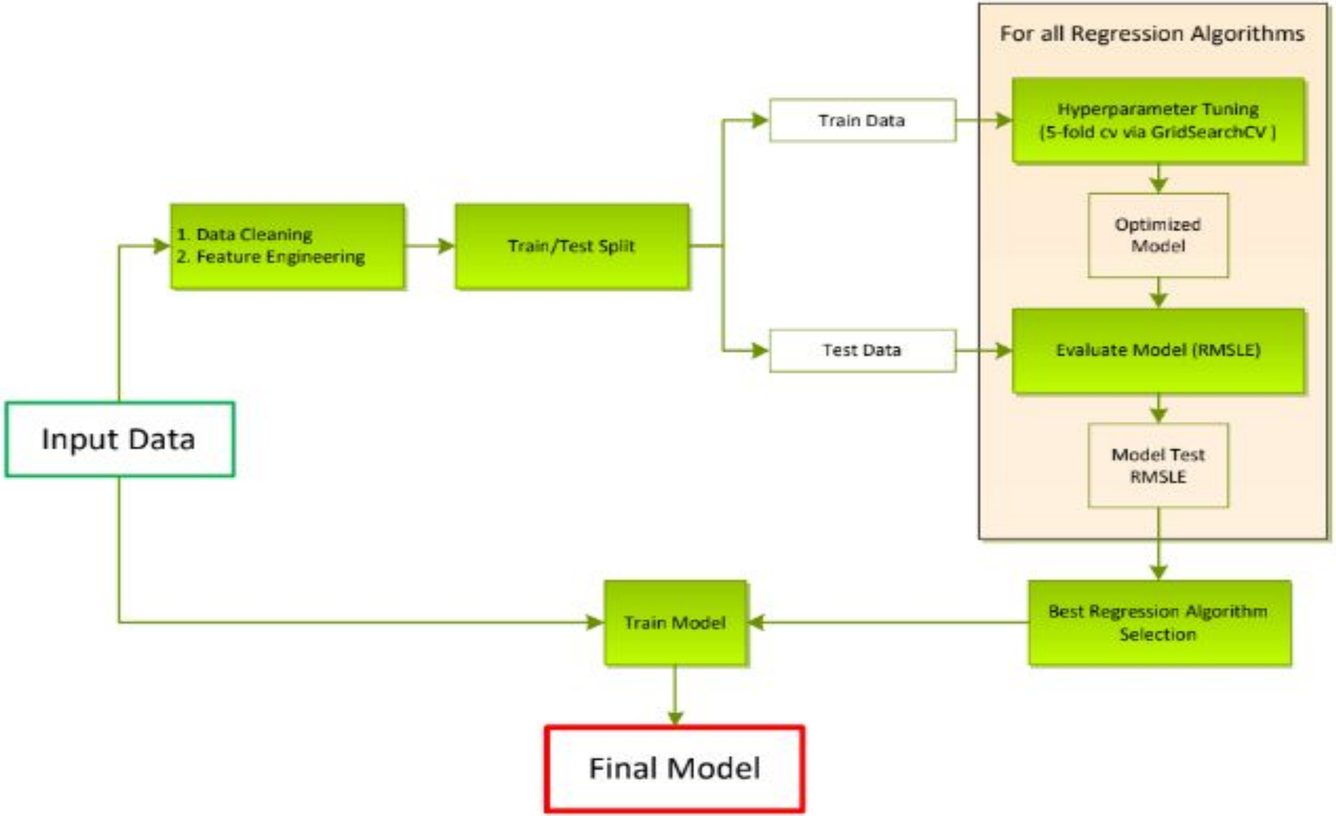
EDA Regression Plots

- We see a strong positive correlation between Rented bike count and temperature
- We see a strong negative correlation between Rented bike count and humidity





Supervised Learning Regression Problem



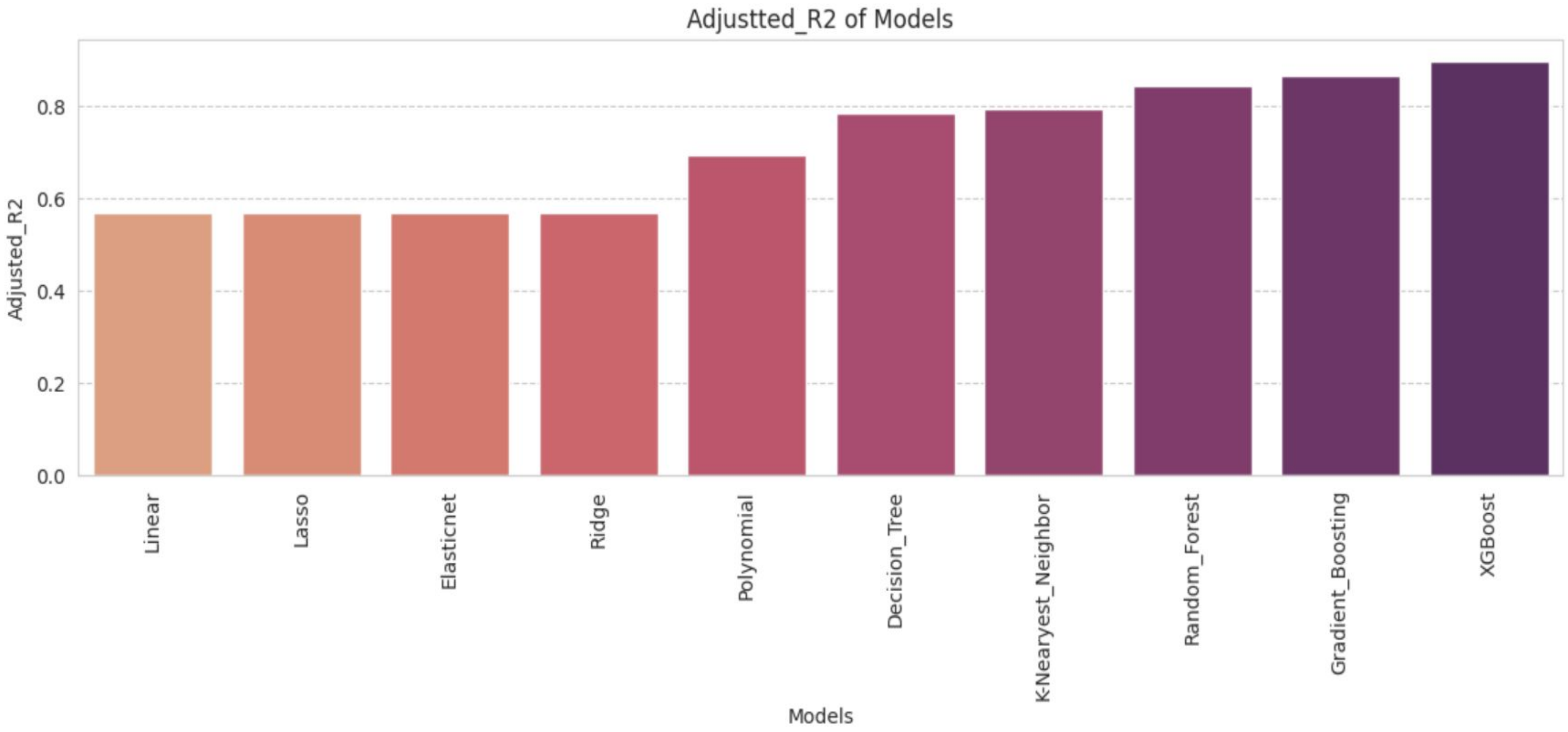
Model's Used

- Linear Regression
- Lasso regression
- Ridge Regression
- Elastic net Regression
- Polynomial Regression
- KNN Regression
- Decision Tree Regression
- Random Forest Regression
- Gradient Boosting Model
- Extreme Gradient Boosting (XGBoost)

Combined Evaluation Matrix of All the models

	Models	Mean_square_error	Root_Mean_square_error	R2	Adjusted_R2
0	Linear	175590.552873	419.035264	0.572911	0.569766
1	Lasso	175560.907118	418.999889	0.572983	0.569839
2	Ridge	175248.935066	418.627442	0.573742	0.570603
3	Elasticnet	175479.947047	418.903267	0.573180	0.570037
4	Polynomial	123952.860328	352.069397	0.698509	0.696289
5	K-Nearyest_Neighbor	83411.759209	288.810940	0.796159	0.794659
6	Decision_Tree	86944.836073	294.864098	0.787525	0.785961
7	Random_Forest	62948.565985	250.895528	0.846167	0.845034
8	Gradient_Boosting	54511.256233	233.476458	0.866786	0.865805
9	XGBoost	40812.801816	202.021785	0.900262	0.899528

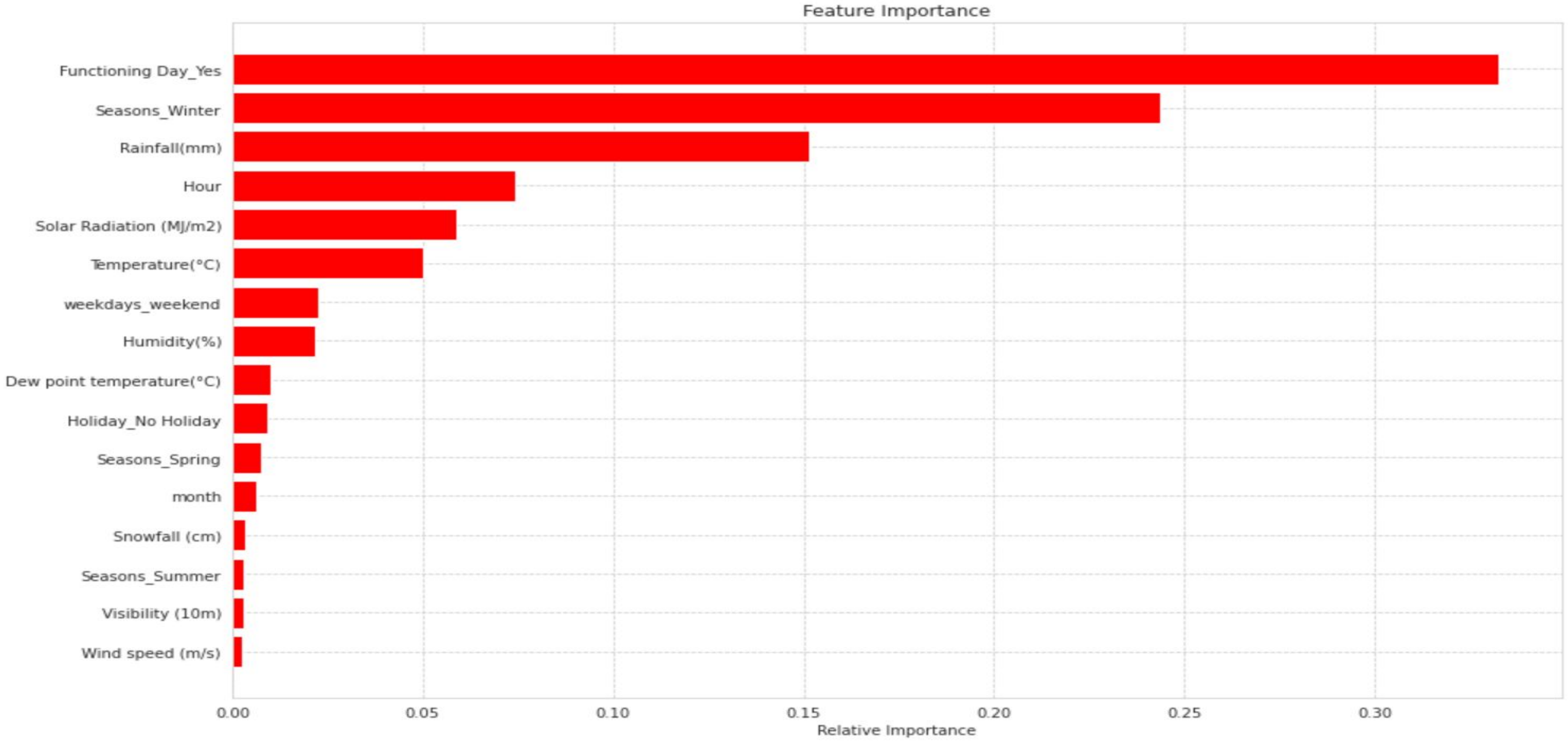
Adjusted R2 of model



Model Validation and Selection

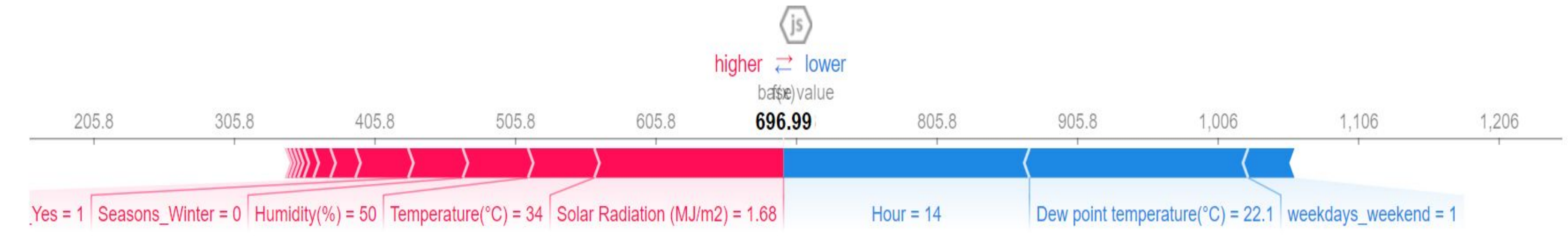
- From R2 and Adjusted_R2 it is clearly seen that linear regression and KNN not giving good results.
- Random forest and Gradient Boosting giving good result in terms of R2.
- But we are getting best result from XGBoost.

Feature Importance



XGBoost

Model Explainability - SHAP



XGBoost

Model Explainability – ELI5

y (score **696.485**) top features

Contribution?	Feature	Value
+705.290	<BIAS>	1.000
+221.971	Solar Radiation (MJ/m2)	1.680
+142.423	Temperature(°C)	34.000
+54.145	Functioning Day_Yes	1.000
+43.052	Humidity(%)	50.000
+11.099	weekdays_weekend	1.000
+5.692	Rainfall(mm)	0.000
+3.001	Visibility (10m)	1744.000
+2.295	month	7.000
+1.151	Wind speed (m/s)	1.200
+0.123	Seasons_Summer	1.000
+0.017	Seasons_Spring	0.000
-0.025	Holiday_No Holiday	1.000
-192.930	Dew point temperature(°C)	22.100
-300.820	Hour	14.000

Conclusion

It is quite evident from the results that XGBoost is 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) show a higher value for the XGBoost. So, we can use XGBoost model for the above problem.


THANK
YOU