Capstone Project-2

Bike Sharing demand prediction

Team Members:

Sagar Malik

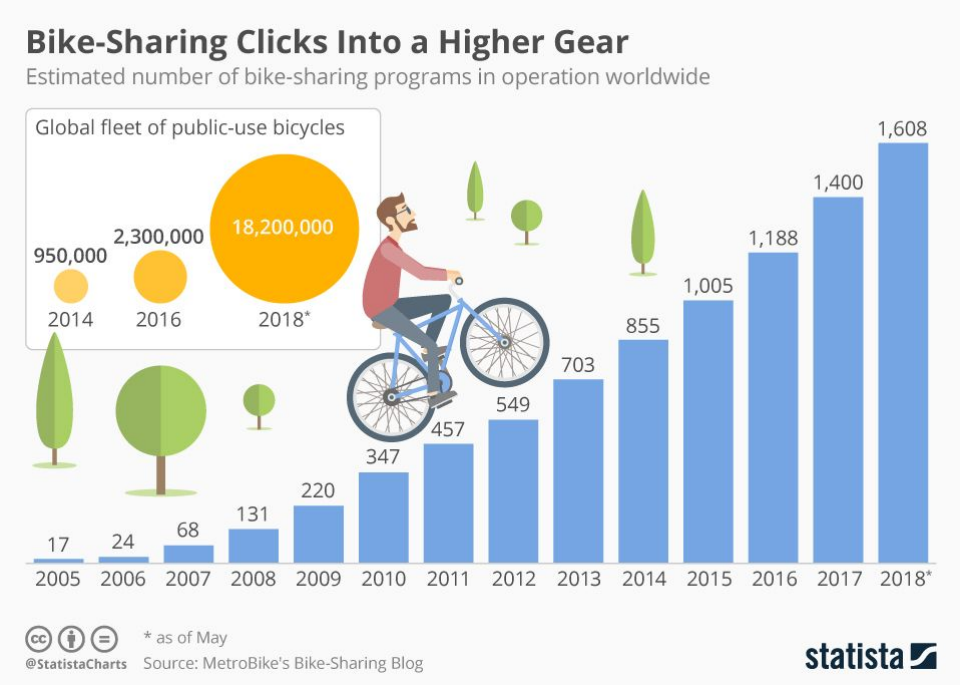
Sharad Tawade

Vinay Kumar

Yashwant Reddy

Contents:

Why bike renting ?

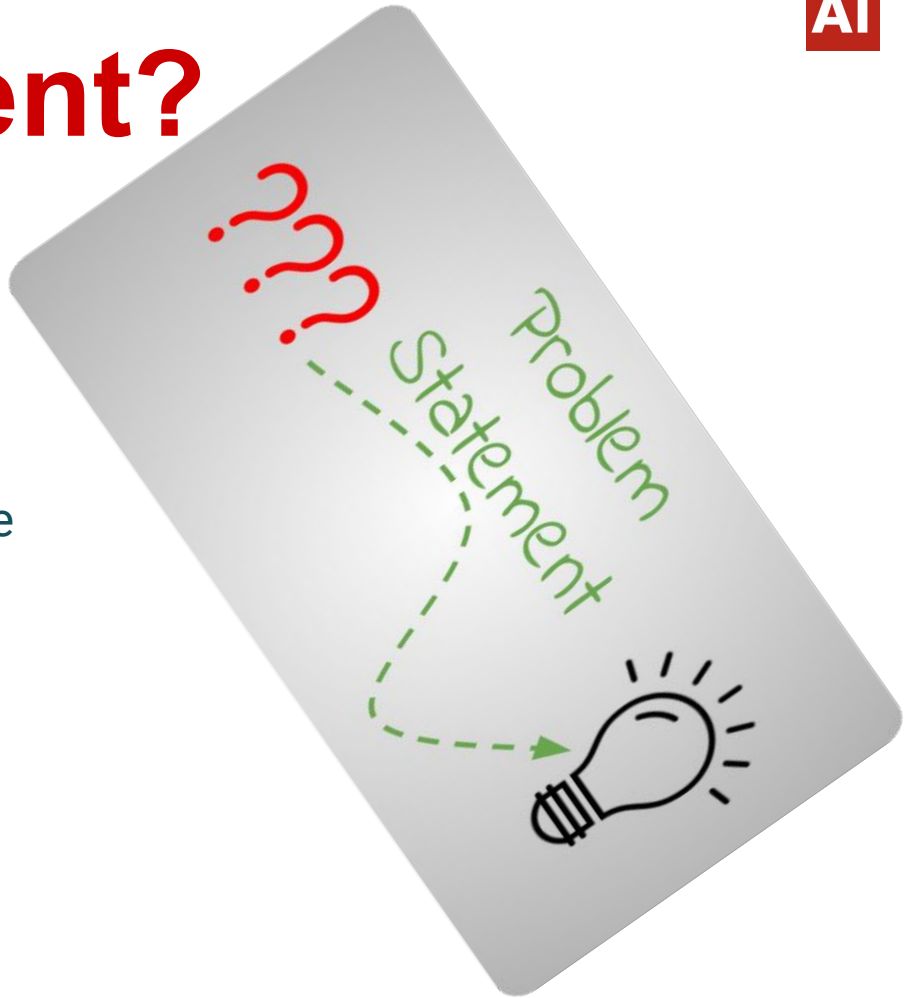


Introduction:

- Bike sharing systems are a type of bicycle rental service in which the procedure of obtaining a membership, renting a bike, and returning the bike is all done through a network of kiosks located around a city.
- People can rent a bike from one location and return it to a different location on an as-needed basis using these systems.
- The purpose of this study is to estimate bike rental demand by combining past bike usage trends with meteorological data. The data set consists of two years' worth of hourly rental data.

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.



Summary of the dataset!

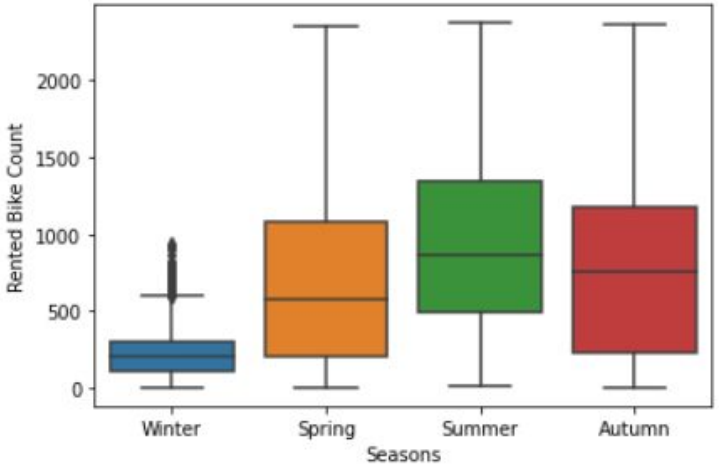
- We have 8740 rows and 14 columns, out of which 11 are numerical and 3 are categorical.
- Rented bike Count is our dependent variable.
- Description of columns:
  - Rented Bike count - Count of bikes rented at each hour - Int64
  - Date : year-month-day - Object
  - Hour - Hour of he day - int64
  - Temperature-Temperature in Celsius - float64
  - Humidity - % - float64
  - Wind-speed - m/s - float64
  - Visibility - 10m -int64
  - Dew point temperature - Celsius -float64
  - Solar radiation - MJ/m2 - float64
  - Rainfall - mm - float64
  - Snowfall - cm - float64
  - Seasons - Winter, Spring, Summer, Autumn - object
  - Holiday - Holiday/No holiday - object
  - Functional Day - NoFunc(Non Functional Day), Fun(Functional Day) - object

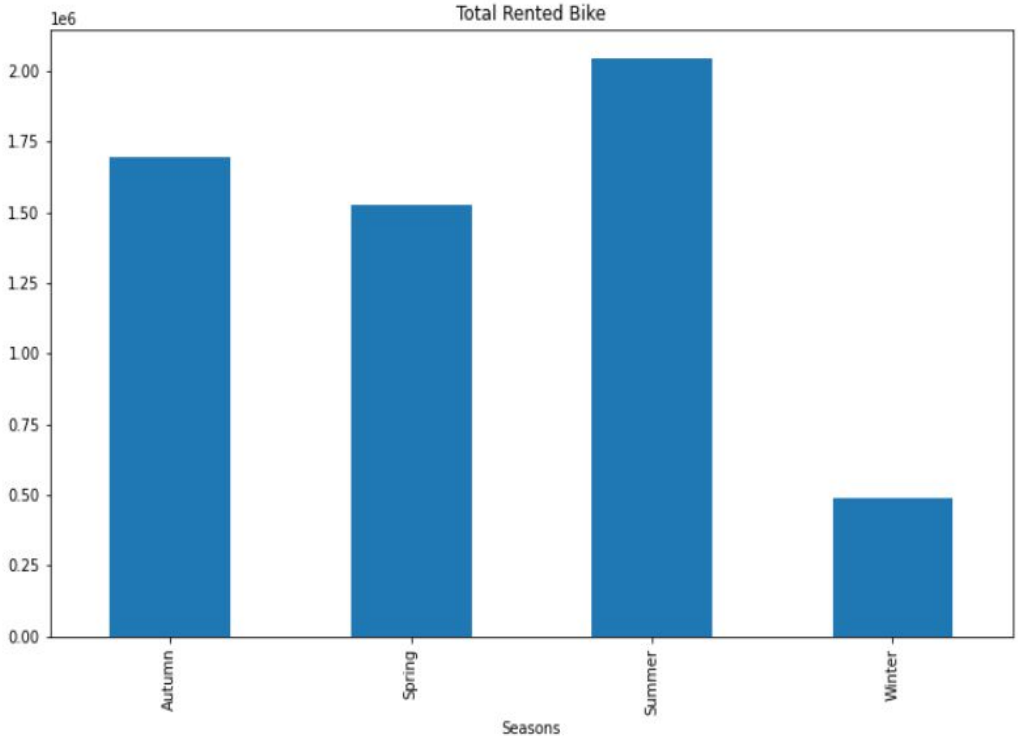
Data wrangling

- Converting date column to date-time and extracting day, month and year.
- Outlier detection
- No Null values were found.

Exploratory Data Analysis

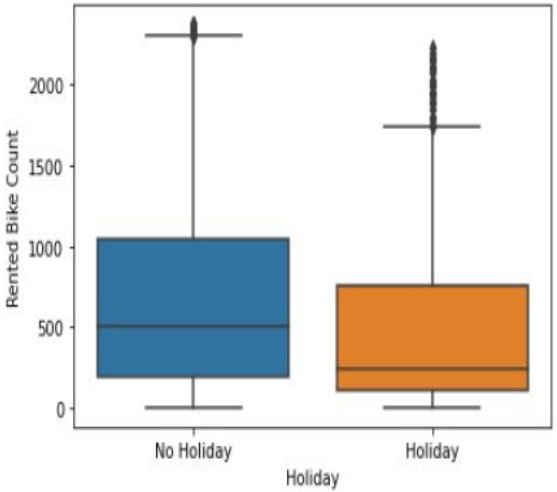
Analysing Count of Rented Bikes for different seasons.

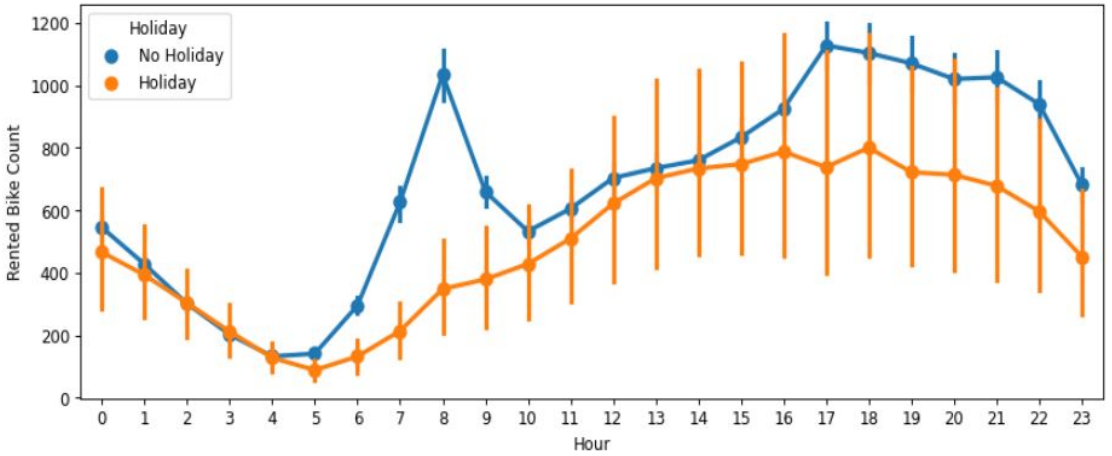




- Summer season was the peak (2208) of all the activity with the most number of Rented bike count.
- Whereas; Winter was the least (2160) popular season with the bike counts recorded.

Analysing at which hour the Rented Bike count is maximum w.r.t. Functional day

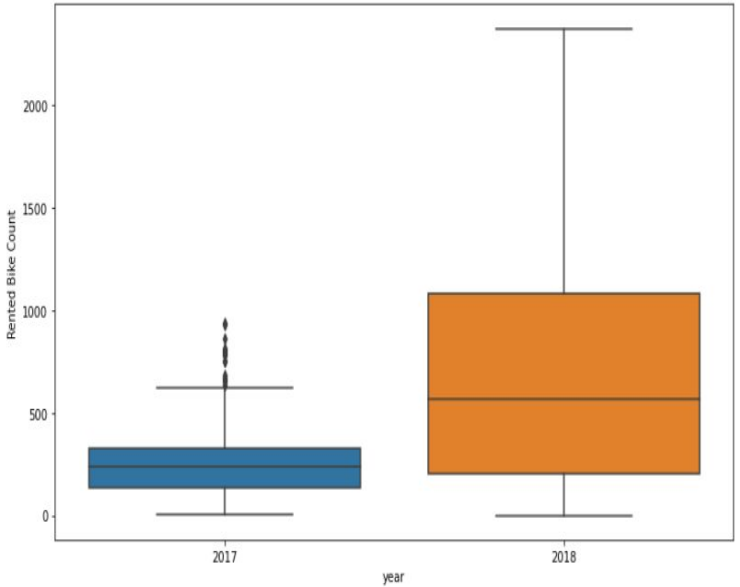


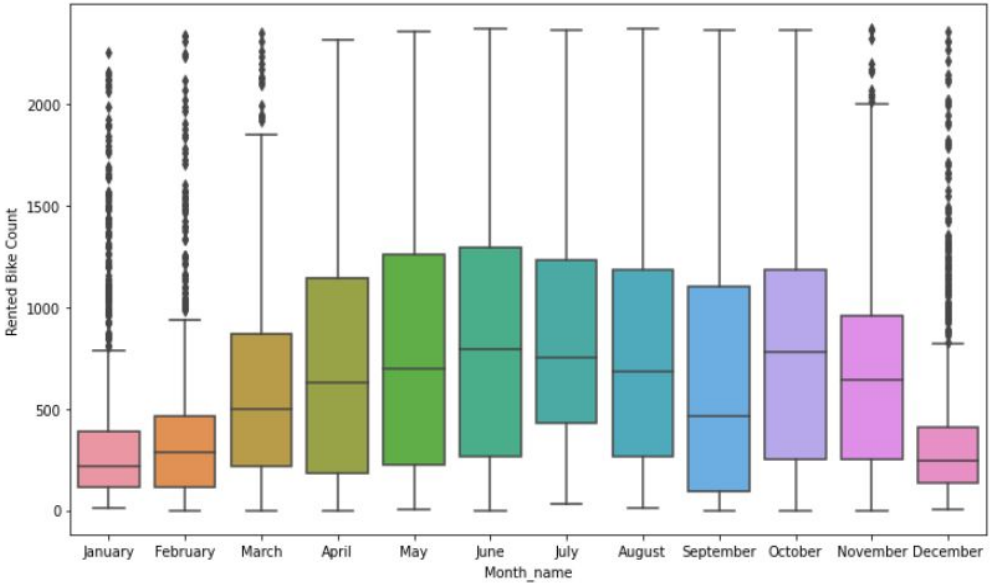


- The above trends indicate that during the holidays the demands of bikes plummet down. Maybe due to lower travel activity and people prefer to stay at homes more.

- Whereas on “No holidays” - the demand is very high around 6-9 and 18-22 hour of the day, as it maybe a convince to get home after work.

Analysing in which Year and Month the Rented Bike Count was maximum:

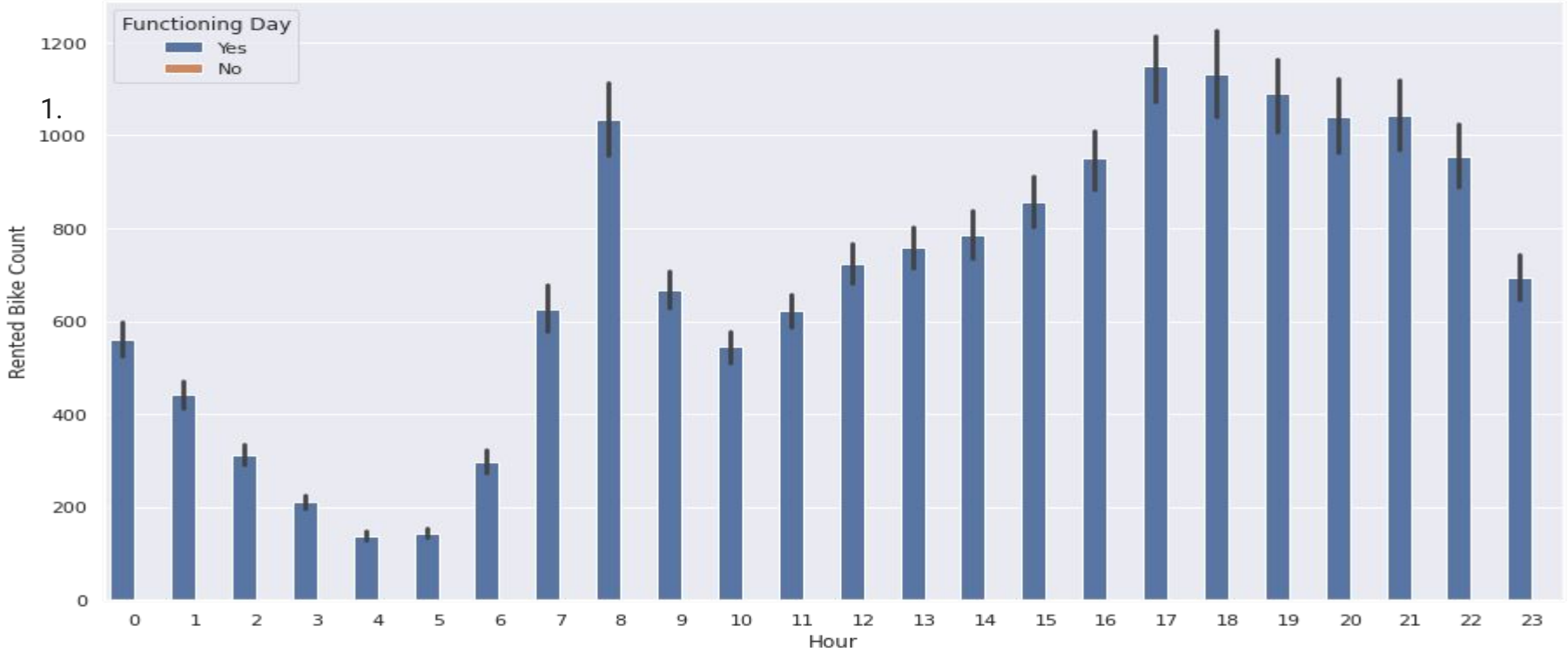




- The demand increased drastically from 2017-2018.

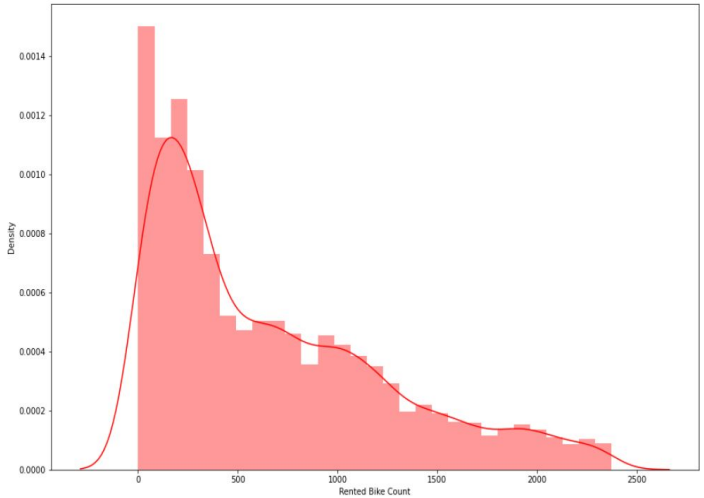
- Monthly trends of the Rented bike count
- April - August being the peaks

How demand of bike change with Functioning days

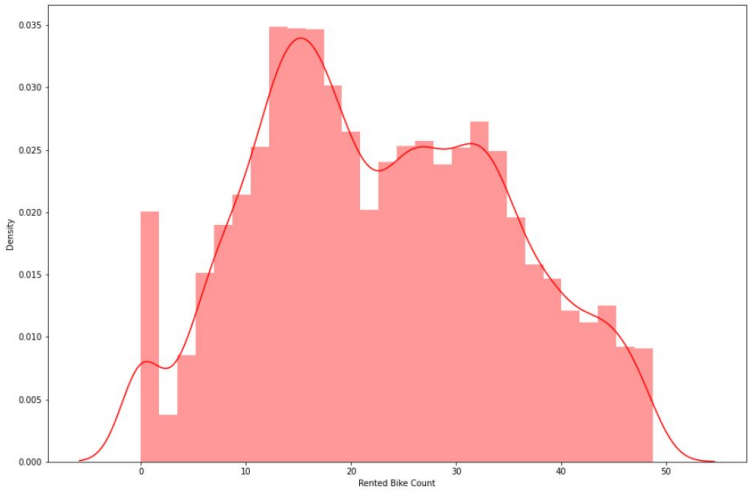


1. From above bar graph we get to know that there was no bike rent on non functioning day

Distribution of Dependent Variable

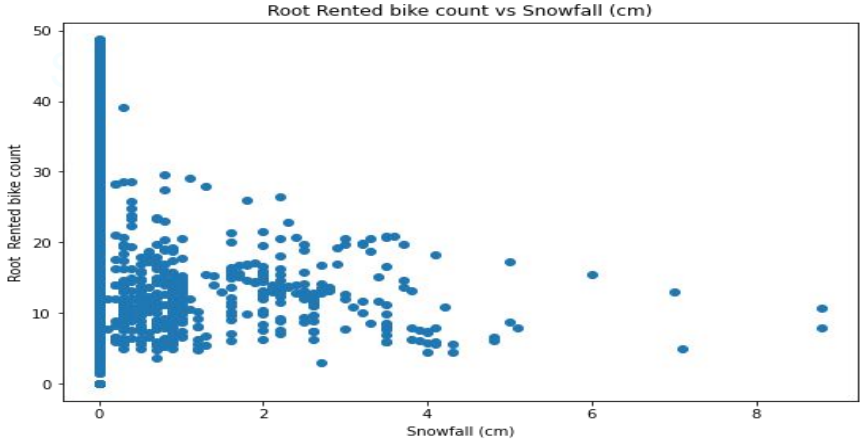


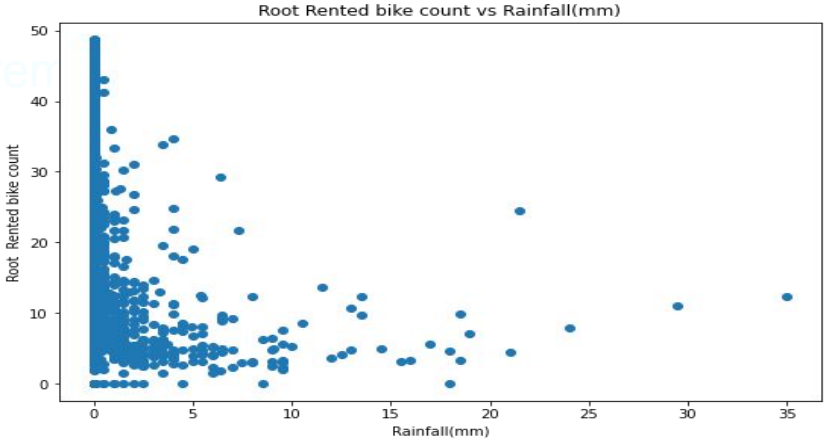
square-root transformation

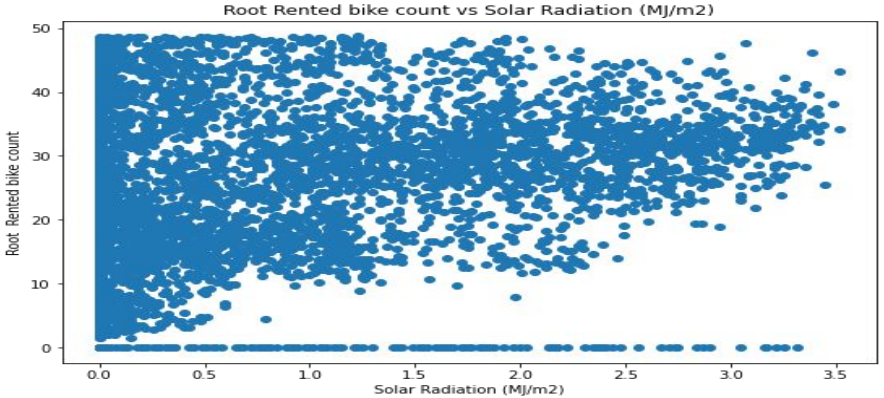


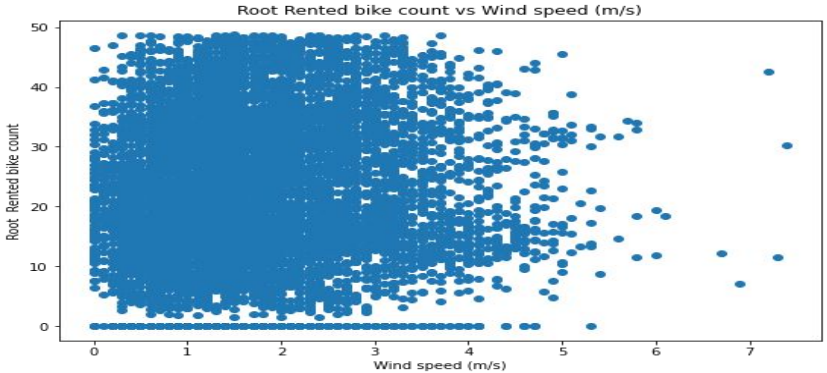
As it was right skewed, so we have taken square root of dependent variable to visualize it in a better way...

Visualizing the relationship b/w dependent & independent variable after transformation





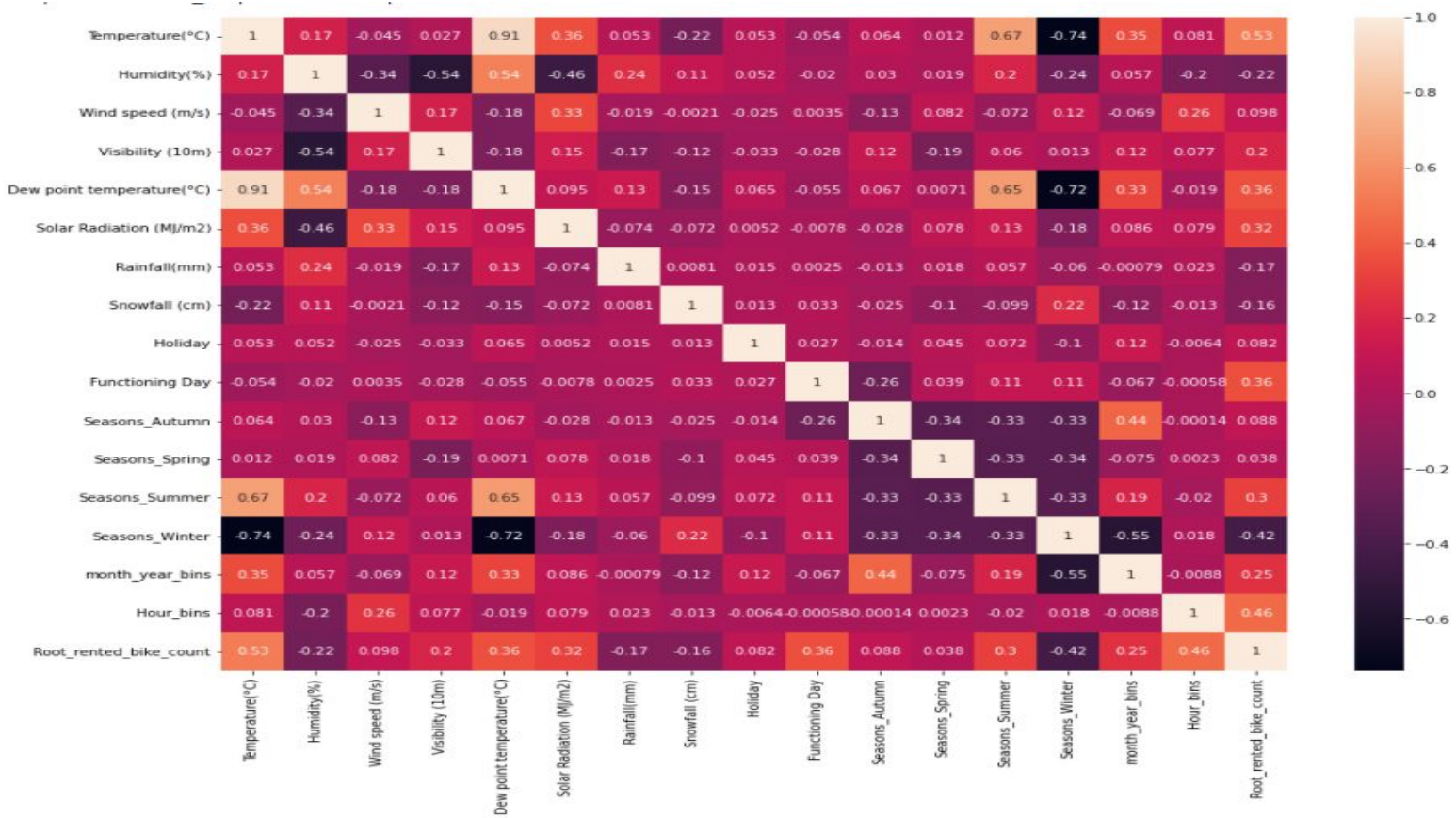




Visualizing the relationship b/w dependent & independent variable after transformation

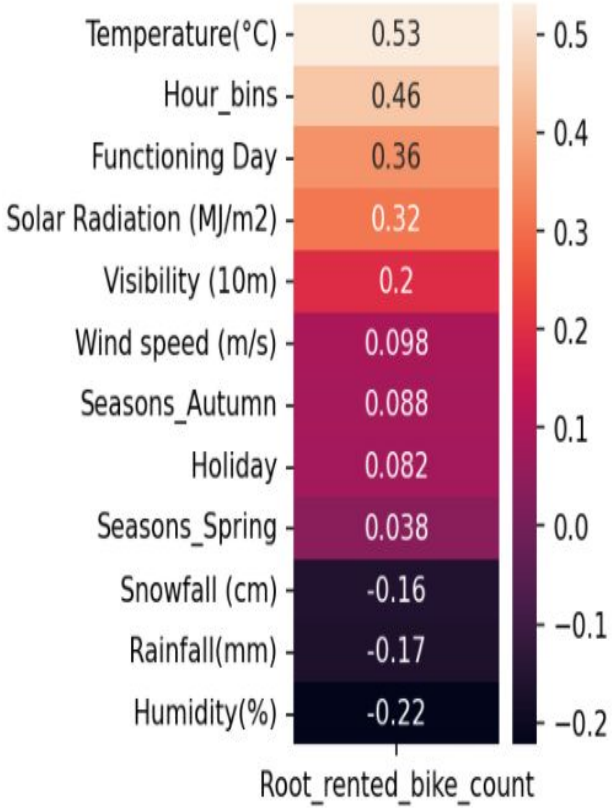
- So after visualizing these scatter plots we removed the unwanted or extra data which were making our dataset quite unwell.
- So, for windspeed – value higher than 4.5m/s,
- Solar Radiation(MJ/m2) value higher than 3MJ/m2,
- Rainfall value higher than 10mm &
- Snowfall value higher than 4cm were not taken.

Multicollinearity:



Removing collinearity:

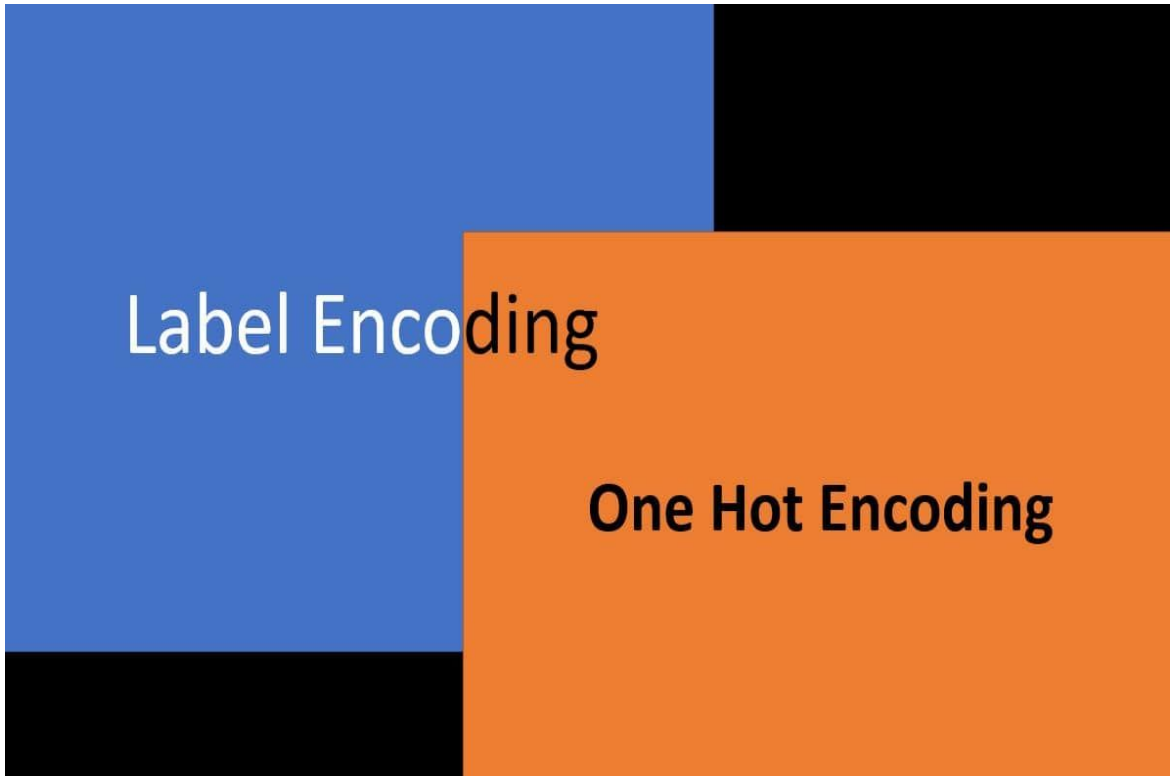
- Here Temperature's and Dew Point Temperature's VIF are highly correlated so we will focus on these two to remove collinearity.
- So we will be focusing on Dew Point Temperature.



Pre-processing of Data :-

Feature Engineering:

- Label Encoding
- One hot Encoding



Feature Engineering:

- Feature engineering is the process of selecting, manipulating, and transforming raw data into features that can be used in supervised learning.
- Feature engineering, in simple terms, is **the act of converting raw observations into desired features using statistical or machine learning approaches**.
- It is the process of designing artificial features into an algorithm. These artificial features are then used by that algorithm in order to improve its performance, or in other words reap better results.

Label Encoding:

- Label Encoding refers **to converting the labels into a numeric form so as to convert them into the machine-readable form**.
- Label Encoding is a popular encoding technique for handling categorical variables. In this technique, each label is assigned a unique integer based on alphabetical ordering.
- Machine learning algorithms can then decide in a better way how those labels must be operated. It is an important pre-processing step for the structured dataset in supervised learning.
- Here we encodethe variables like Functioning Day and Holiday in the form of 0 and 1. also convert the seasons column into dummy variables like Spring, Summer, Rainy and Winter.

One hot Encoding:-

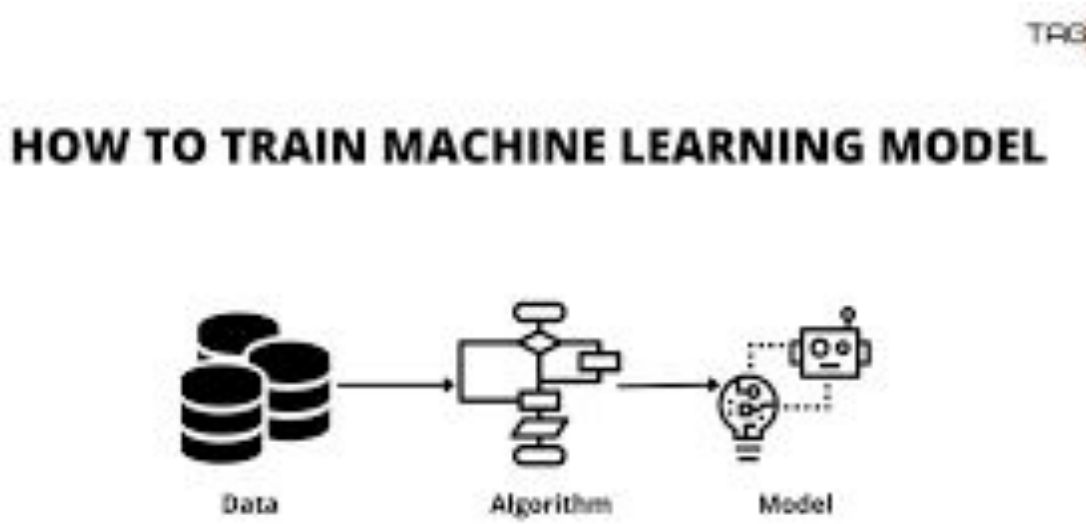
- A one hot encoding is **a representation of categorical variables as binary vectors**.
- This first requires that the categorical values be mapped to integer values. Then, each integer value is represented as a binary
- In this Feature Engineering, we apply lambda function to convert respective columns in the form of 0 and 1. Ex. We convert Visibility column in the form of 1 when it is greater than 2000, also for rainfall if the value is greater than 0.148 then it is converted into 1 otherwise 0. Same procedure follows for snowfall and solar radiation

Feature Selection:

1. In Feature selection we remove non-informative or redundant predictors from the model. At beginning we have 8760 rows and 14 columns. After label encoding and feature enginnering we get 8459 rows and 15 colu

Model’s Performed

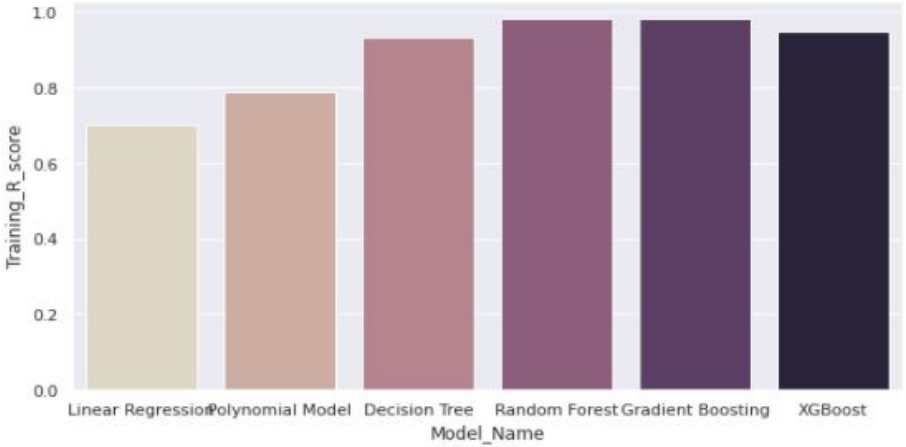
- Linear Regression with regularizations
- Polynomial Regression
- Decision tree
- Random forest
- Gradient Boosting
- eXtreme Gradient Boost

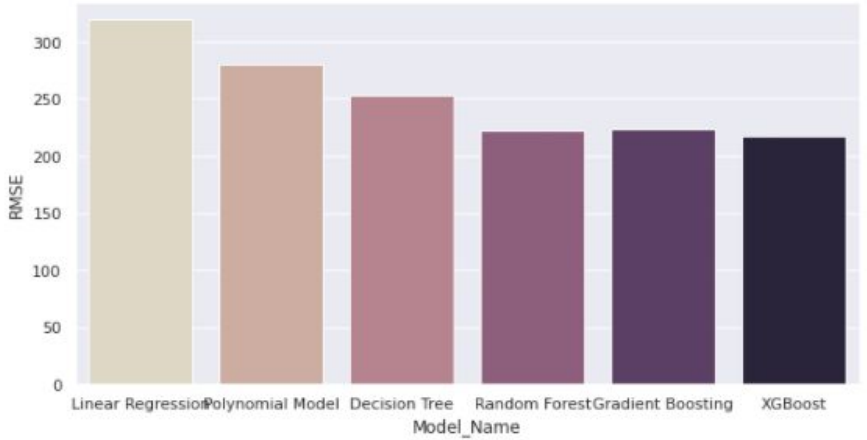


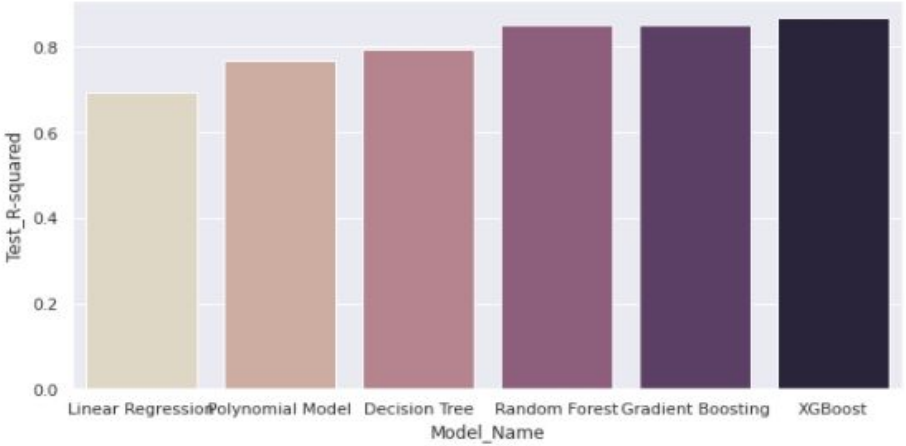
Applying Models - Validation & Selection

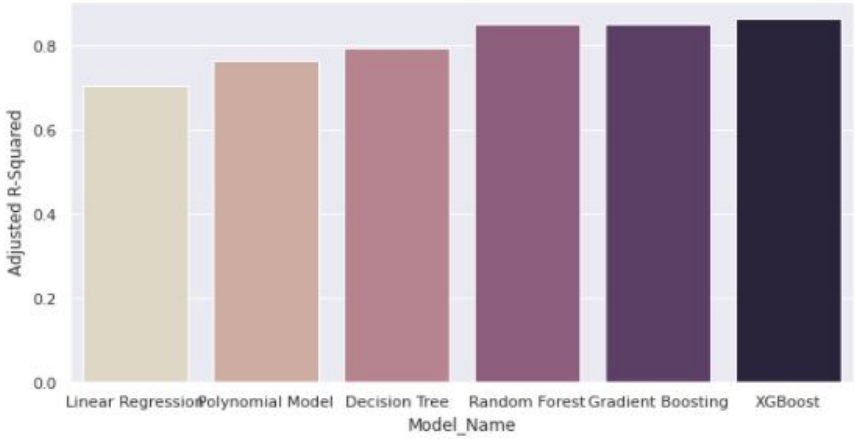
	Model_Name	Training_R_score	Test_R-squared	RMSE	Adjusted R-Squared
0	Linear Regression	0.703	0.6948	318.97	0.7020
1	Polynomial Model	0.789	0.7660	280.49	0.7650
2	Decision Tree	0.932	0.7950	253.35	0.7920
3	Random Forest	0.980	0.8520	222.06	0.8510
4	Gradient Boosting	0.980	0.8520	223.35	0.8500
5	XGBoost	0.947	0.8662	216.83	0.8619

Model - Validation & Selection









Model - Validation & Selection

Observation 1: As observed from the table above Linear Regression did not generated great results, some improvement in the results were achieved by Polynomial linear regression and Decision tree, but had lower Test_R_squared values.

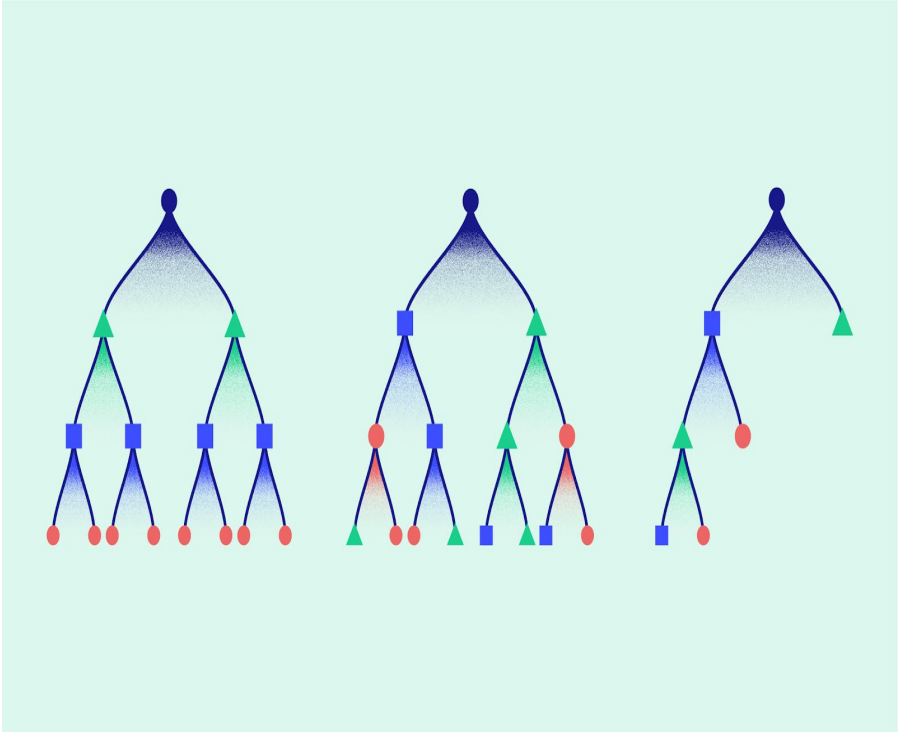
Observation 2: Random forest & Gradient Boosting have performed equally good, but XGBoosting take the best place of all.

Observation 3: From the above observation we have come to a conclusion that we would choose our regression model from XGBoost.

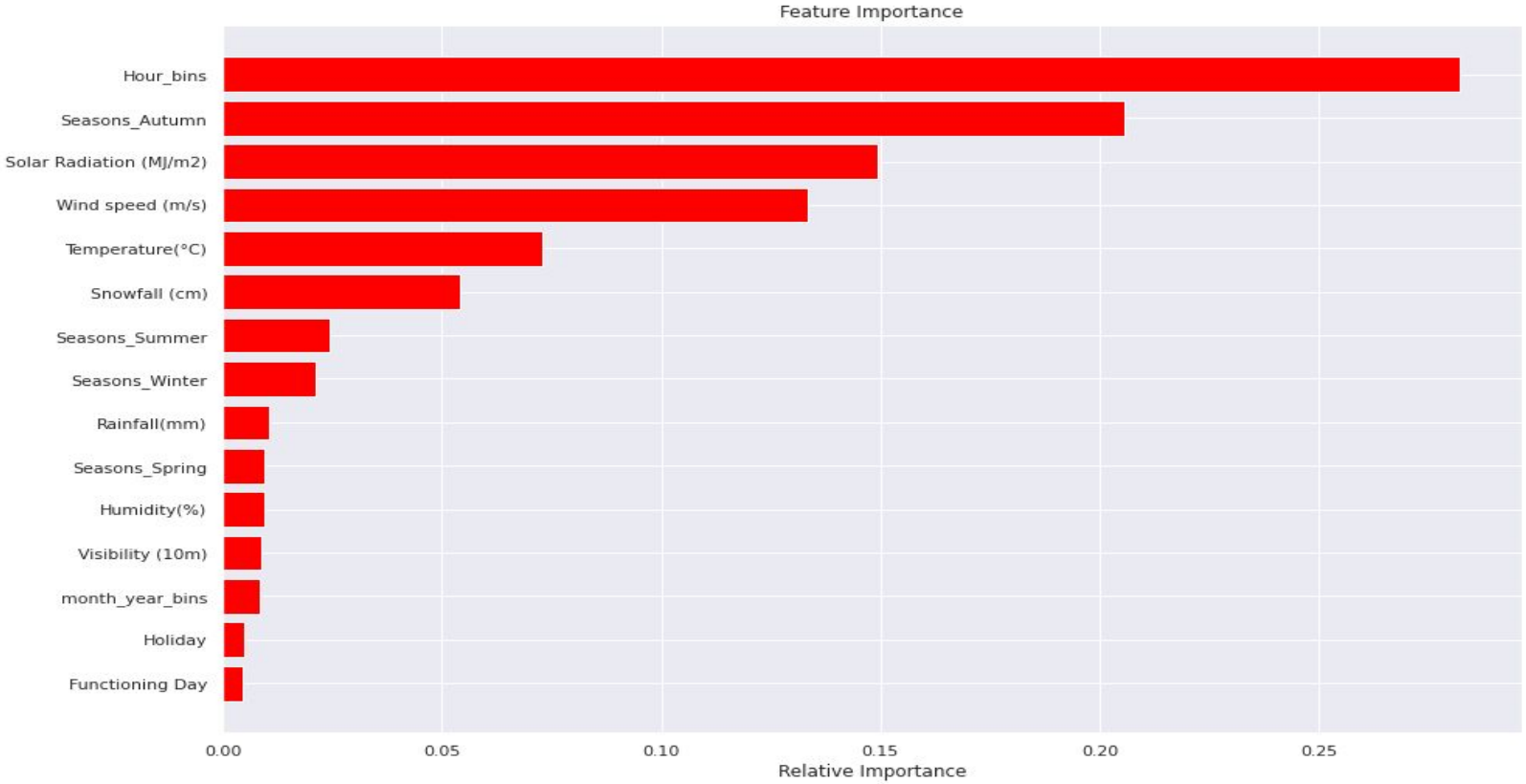
About the model

Since we have chosen XGBoost as our regression model. Below are the best hyperparameters:

```
'max_depth': 6
'min_child_weight': 12
'n_estimators': 200
```



Feature Importance



Challenges

- As dataset was quite big enough which led more computation time.
- Data Cleaning
- A huge amount of data needed to be dealt while doing the project which is quite an important task and also even small inferences need to be kept in mind.

Conclusion

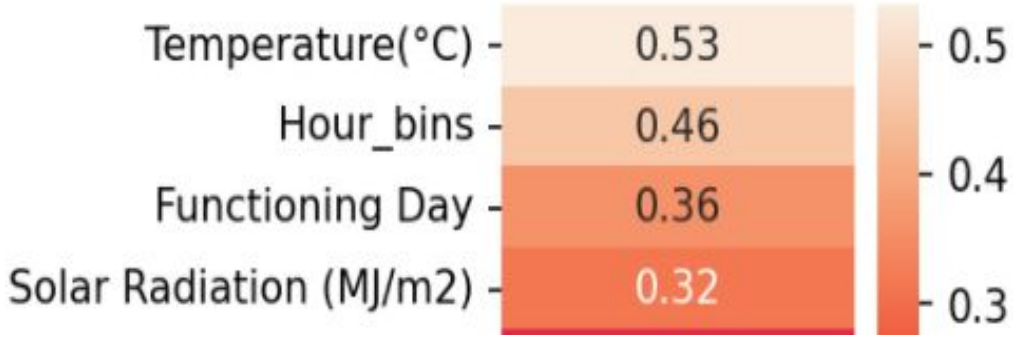
- From above model we can conclude the below points:
- Rented Bike Count is very much dependent :-

1) At what Temperature the Bike is rented.

2) On what Hour the Bike is rented.

3) How much Humidity present in the atmosphere.

4) Is it a functioning day or not.



Conclusion conti..

- In project, after trying combinations of features with linear regression the model underfit. It seemed obvious because data is spread too much. It didn't seem practical to fit a line.
- The experimental results prove that the XGBoost model predicts best the trip duration with the highest R2 and with less error rate compared to Linear Regression, Decision Tree, Random Forest, Gradient Boosting.


Thank You