Capstone Project
Seoul Bike Sharing Demand Prediction

Team

Abhijeet Kulkarni , Kundan Lal , pankaj Ganjare , Akshay Auti

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.

Content

- ❑ Data Pipeline
- ❑ Data Description
- ❑ Exploratory Data Analysis
- ❑ Regression plot
- ❑ Heat map
- ❑ One Hot Encoding
- ❑ ML Algorithm
- ❑ Evaluating models
- ❑ Conclusion

Data Pipeline

Data Preparation and Exploratory Data Analysis

↓

Building Predictive Model using Multiple Techniques/Algorithms

↓

Optimal Model Identified through testing and evaluation

Data Description

Dependent variable:

- Rented Bike count - Count of bikes rented at each hour

Independent variables:

- Date : year-month-day
- Hour - Hour of he day
- Temperature-Temperature in Celsius
- Humidity - %
- Windspeed - m/s
- Visibility - 10 m
- 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 Hours), Fun(Functional hours)

Attribute Information : Null Values and Dtypes

```
#check for count of missing values in each column
df2.isnull().sum()
```

```
Date                        0
Rented Bike Count           0
Hour                        0
Temperature(°C)             0
Humidity(%)                 0
Wind speed (m/s)            0
Visibility (10m)            0
Dew point temperature(°C)   0
Solar Radiation (MJ/m2)     0
Rainfall(mm)                0
Snowfall (cm)               0
Seasons                     0
Holiday                     0
Functioning Day             0
dtype: int64
```

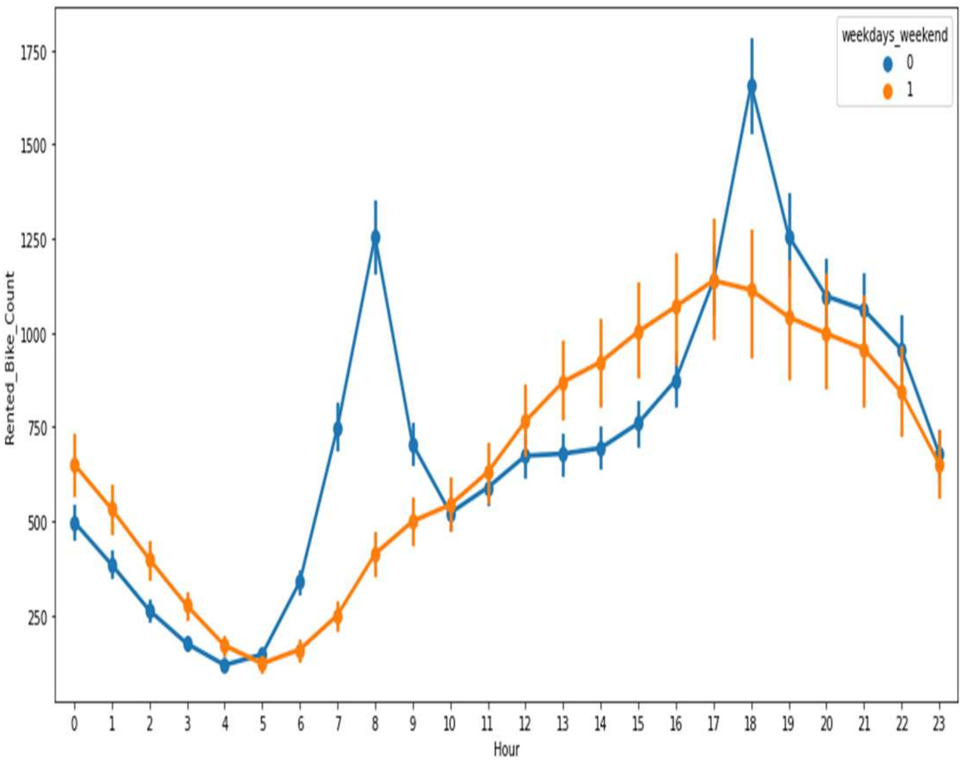
```
# Check details about the dataset
df2.info()
```

```
<class 'pandas.core.frame.DataFrame'>
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)
memory usage: 958.2+ KB
```

EDA AND DATA PROCESSING

Exploratory Data Analysis

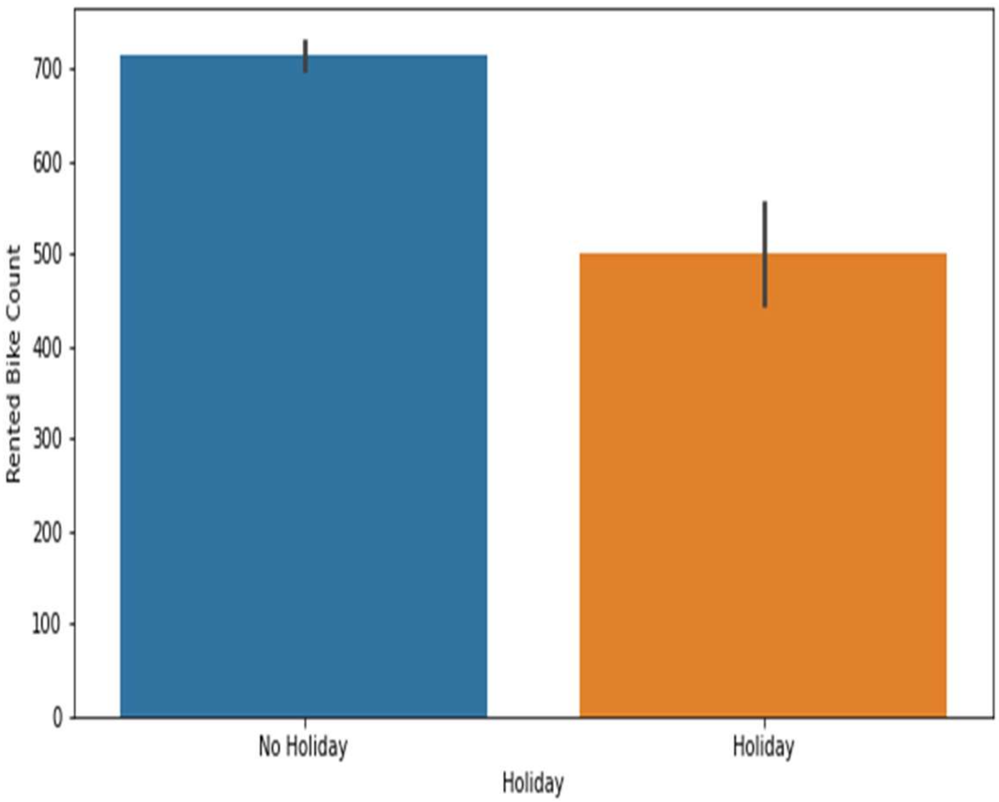
Analysing Categorical Variables (week days & weekends)



Usage of rented bikes are more during weekdays than weekends.
During weekdays(blue line) from 5 am to 10 am and evening from 4 pm to 8 pm the renting is highest
During weekends (orange line) the renting is very low during morning but gradually the rented numbers increases being maximum around 5 pm.

Exploratory Data Analysis

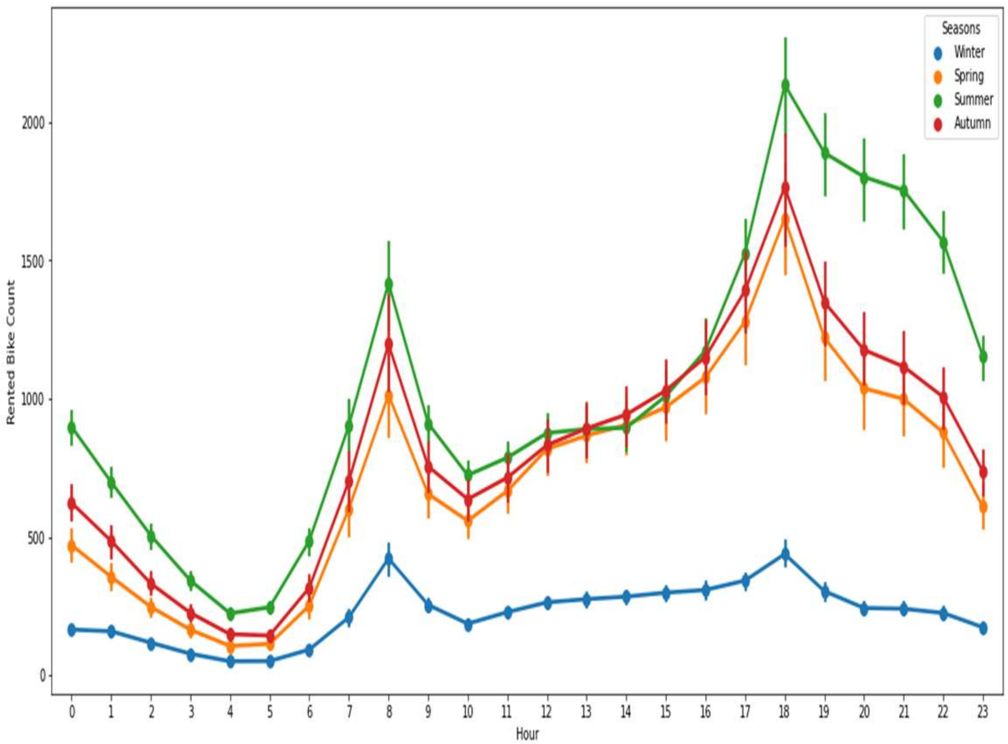
Analysing Categorical Variables (Holiday)



The higher number of Renting is done on weekdays and lower on Holidays.
It can also be inferred that a good percentage of bike are rented for office usage of people.

Exploratory Data Analysis

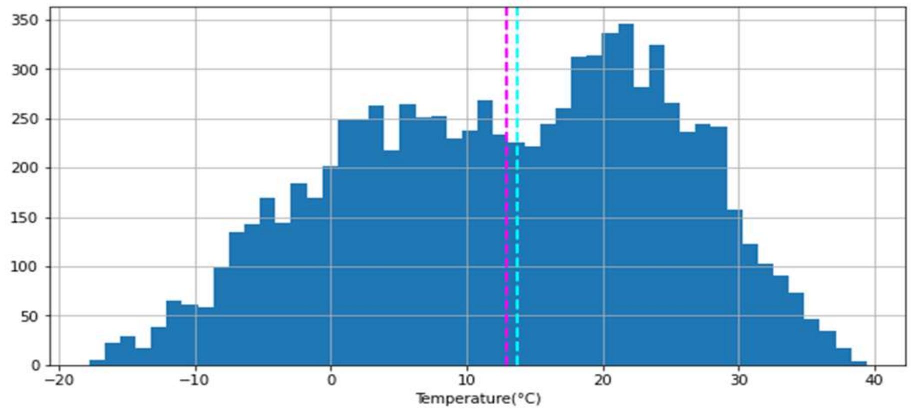
Analysing Categorical Variables (Seasons/day Trend of renting)

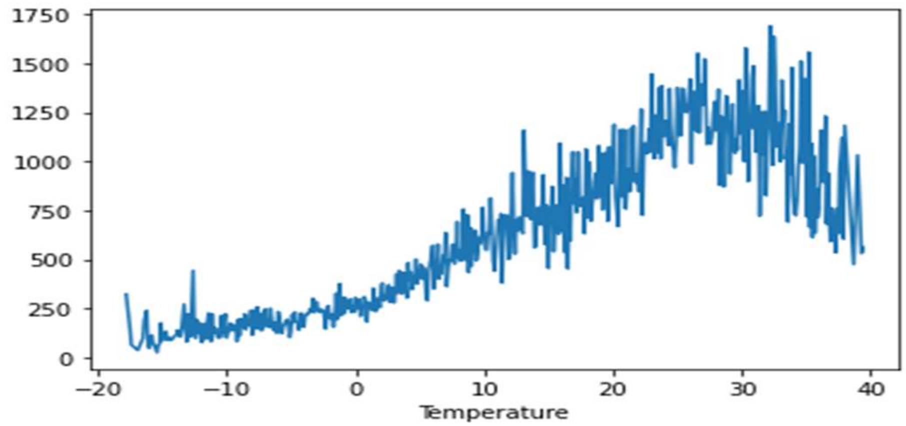


The trend of renting is similar for Summer , Autumn and Spring , which shows peak renting from 6 am to 9 am & from 4 Pm to 10 Pm.
The renting is lowest in Winter season.

Exploratory Data Analysis

Analysis on Numerical Variables (Temperature)



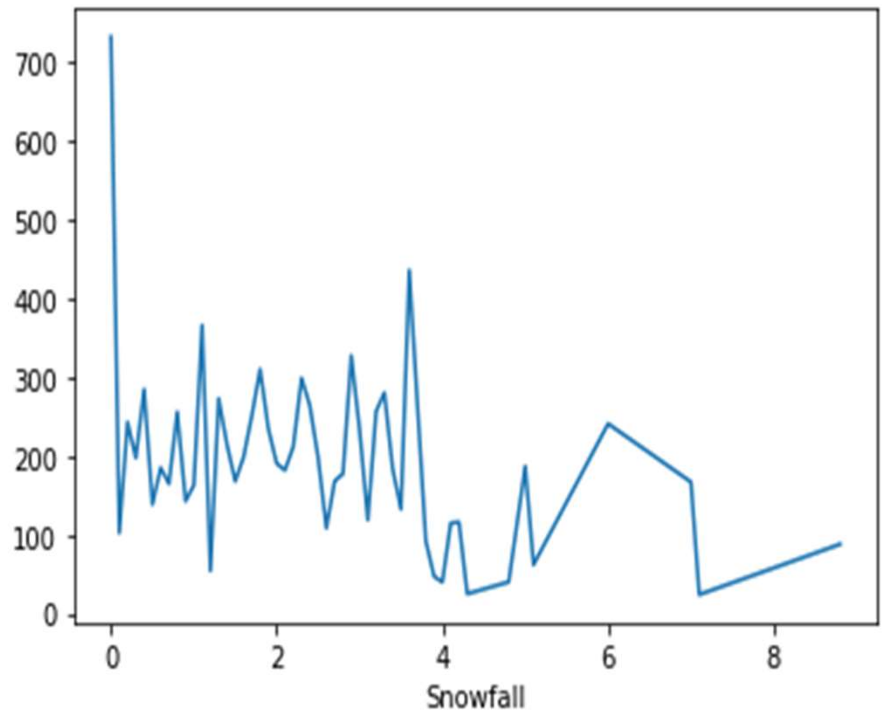


- The peak renting happens between 18 degrees to 25 degrees centigrade.

- Below 2 degrees and above 28 degrees there is a steep reduction is renting numbers.

Exploratory Data Analysis

Analysis on Numerical Variables (Snow Fall)

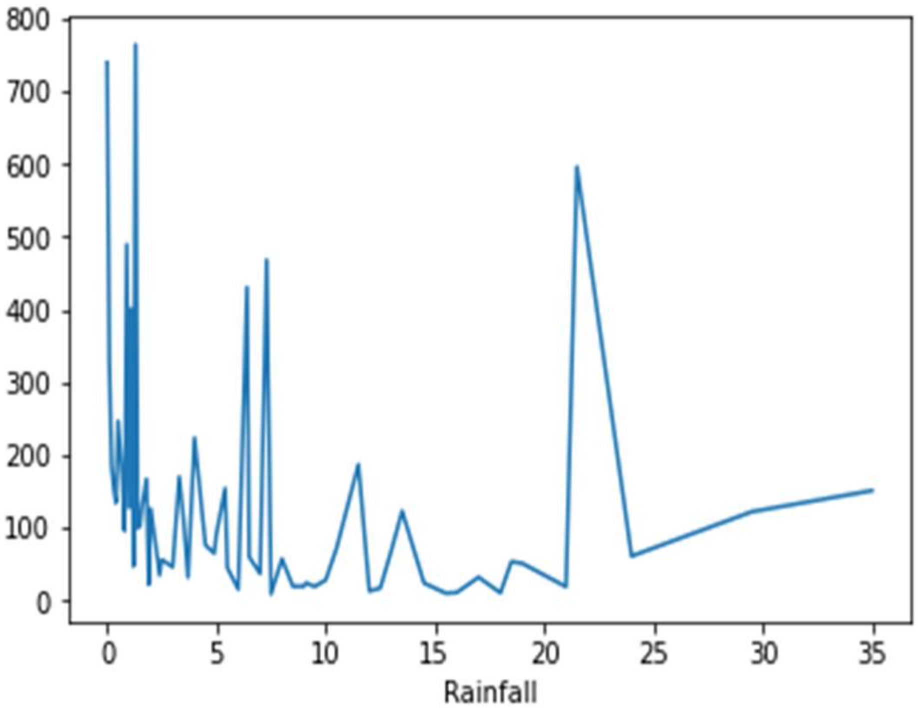


- **It can be analyzed that renting of the bikes are maximum when there is no Snow but it decreases drastically after 4 cms of snowfall.**

- **Snowfall hinders renting a lot and reduces renting by around half.**

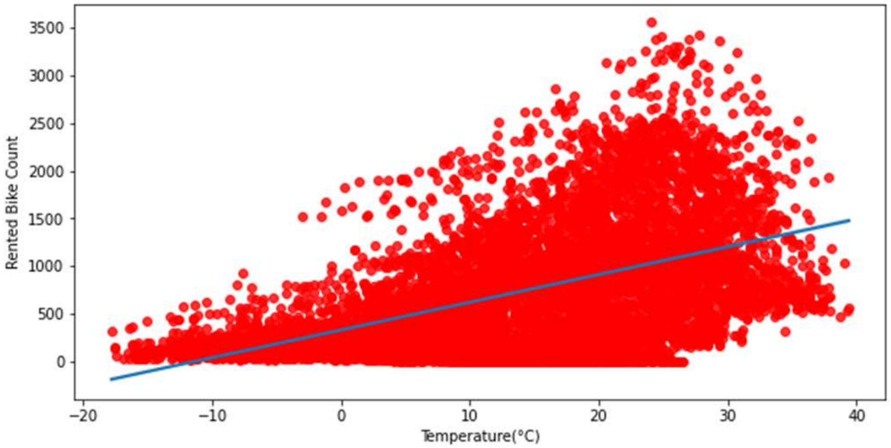
Exploratory Data Analysis

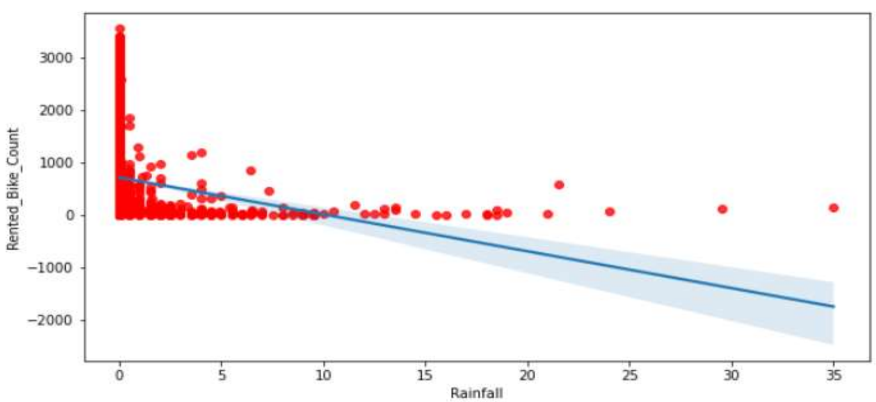
Analysis on Numerical Variables (Rain Fall)



It can be seen than opposite of expected , there is no decrease in the renting of the bikes even if its raining , intermittently there are surges in the renting numbers .

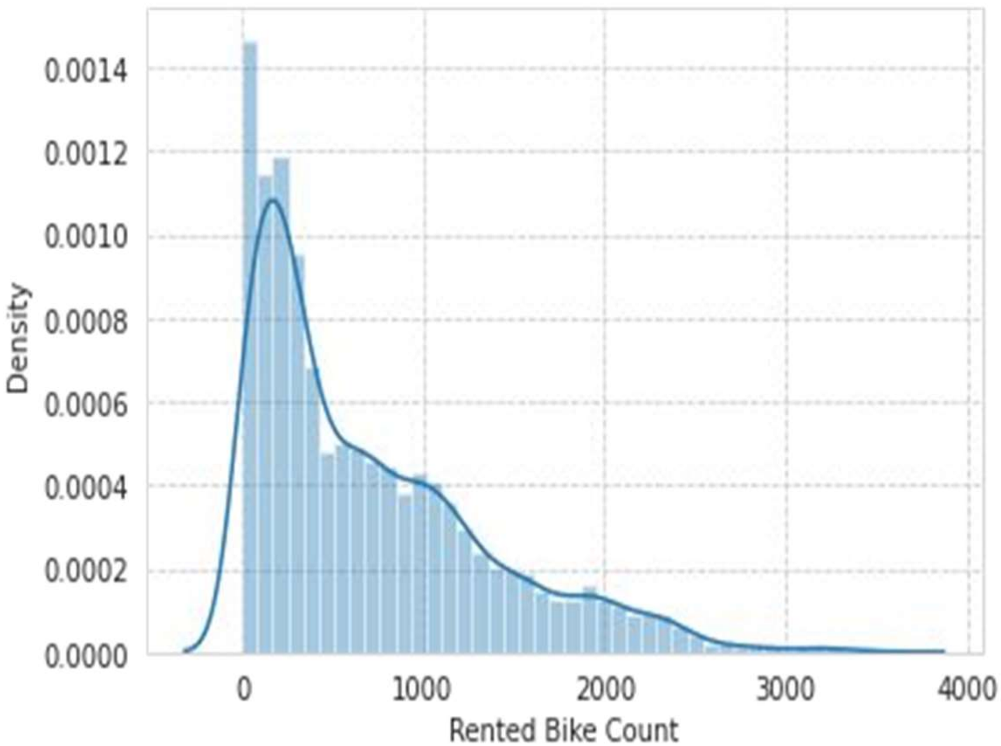
Regression Plot showing Linear Relationship with Target Variables



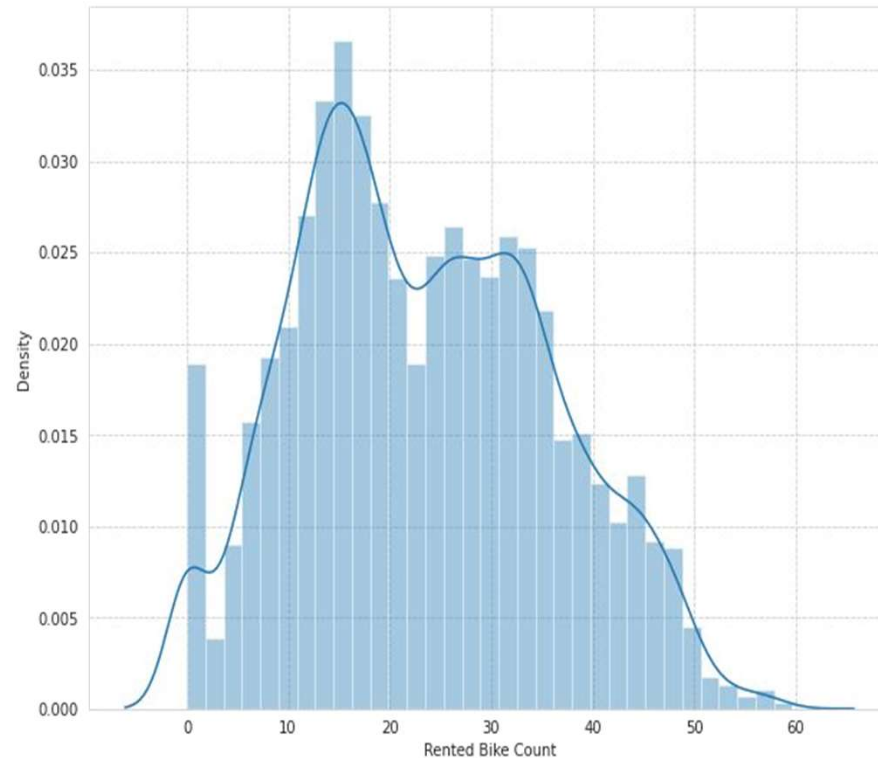


- **Variables like Temperature , Hour , Wind Speed , Visibility ,Dew point temperature & Solar Radiation are Positively correlated to our Dependent variable (Rented bike count).**

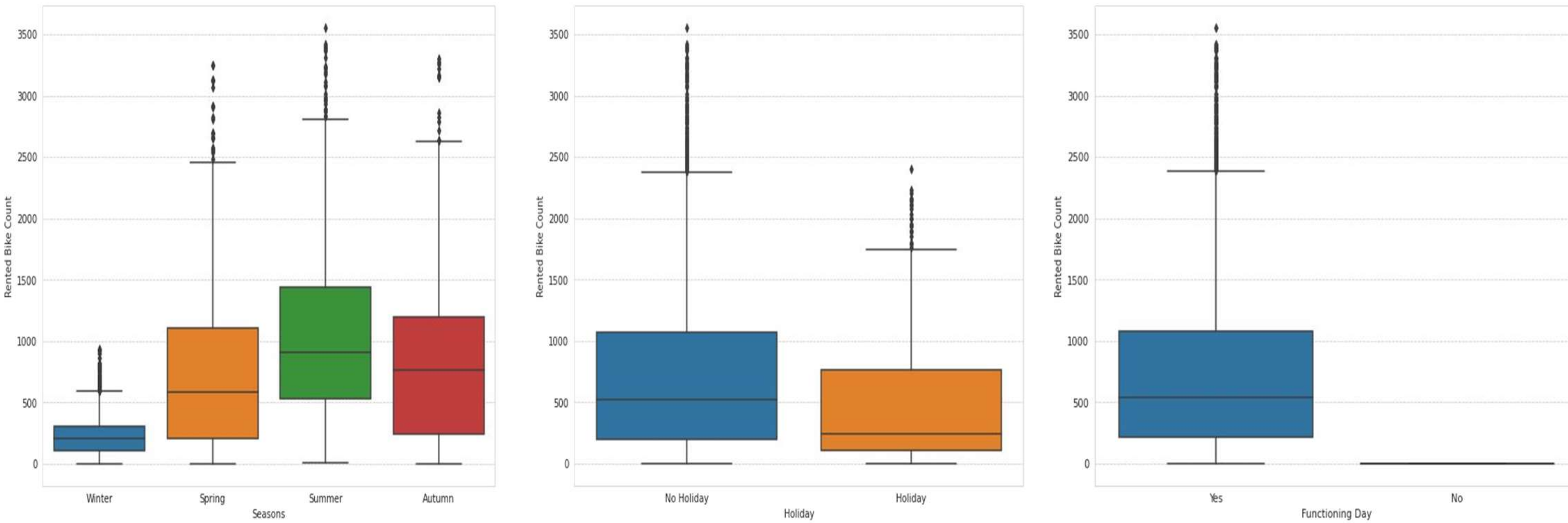
- **Variables like Snow fall , Rain fall & Humidity are Negatively correlated .**



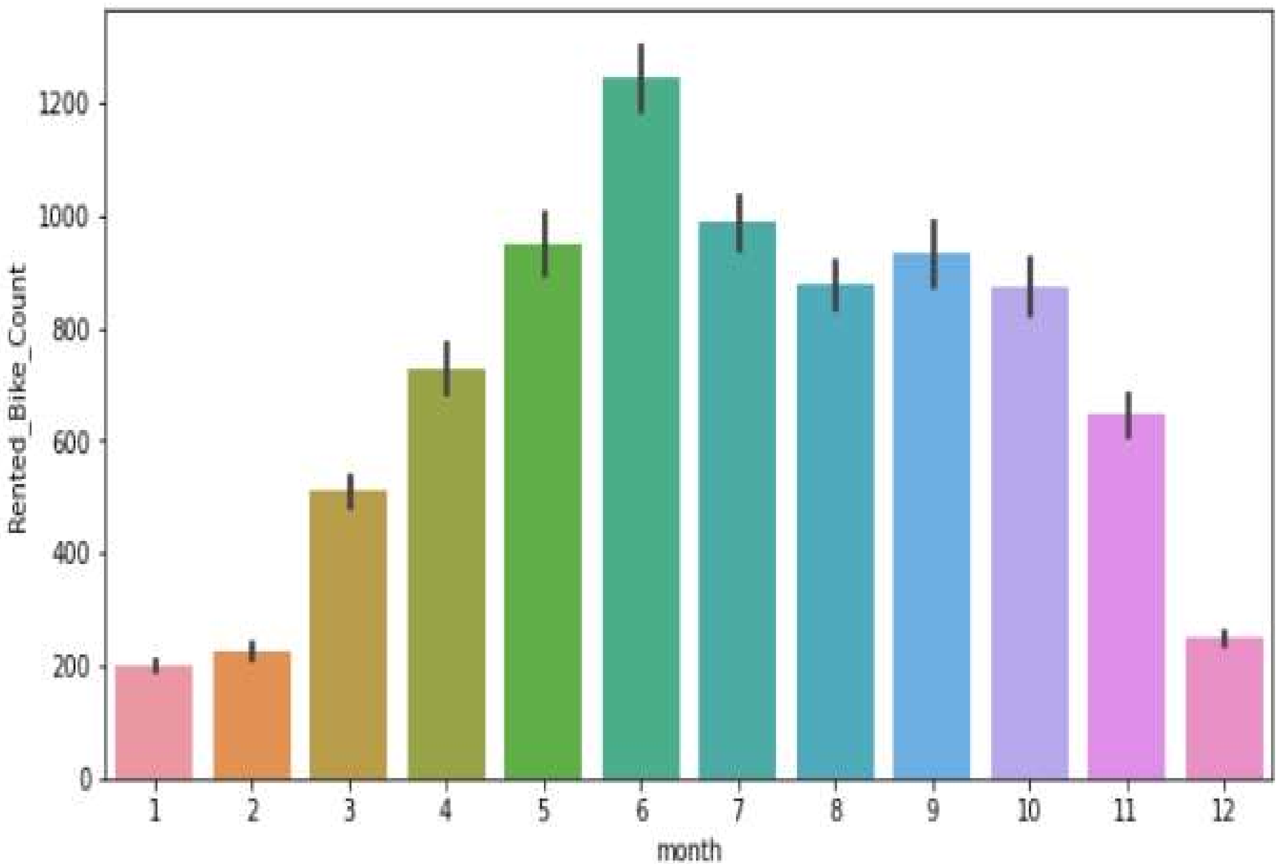
Distribution of rented bike count



Square root transformation of rented bike count

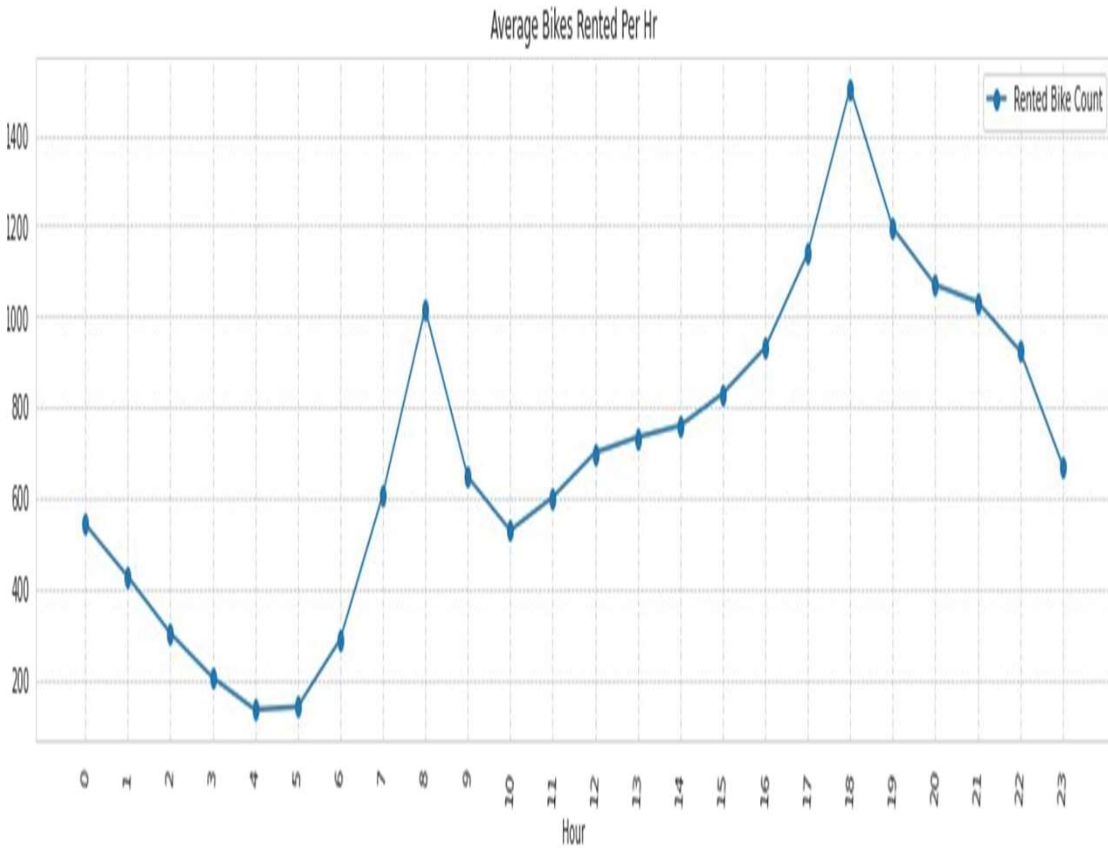


- **Less demand on winter seasons**
- **Slightly Higher demand during Non holidays**
- **Almost no demand on non functioning day**



- **We can see that there 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**



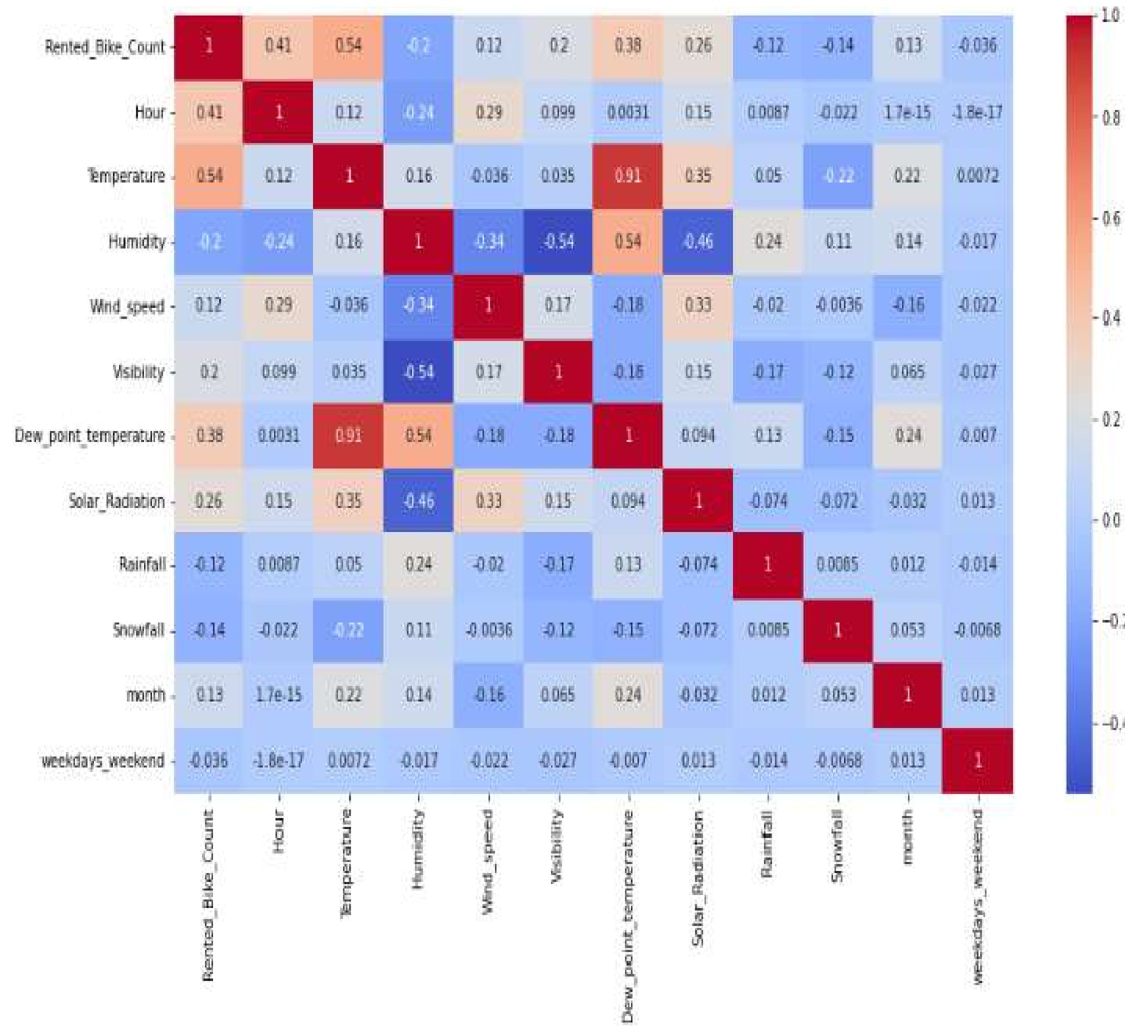
High rise of Rented Bikes from 8:00 a.m to 9:00 p.m means people prefer rented bike during rush hour.

we can clearly see that demand rises most at 8 a.m and 6:00 p.m so we can say that that during office opening and closing time there is much high demand

Analysis on : Correlation Heat map

- **From the Heatmap we can see that the temperature and Dew_point_temperature have high correlation i.e, 0.91.**

- **Humidity is moderately correlated with Solar Radiation and Visibility**



One-Hot Encoding

Due to the presence of categorical features we can't feed our data directly in ML algorithm. We need to transform categorical features that have string datatype to numerical datatype . For which we have used One-hot encoding and label encoding for categorical features.

Seasons
Summer
Winter
Autumn
Spring

One hot encoding

Summer	Winter	Autumn	Spring
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

Machine Learning Model – Regression

Applying ML Algorithms

Since we have to predict the count of rented bikes required per hour. Hence, we have to use regression algorithm.

Algorithms that we will use are:

- **Linear Regression**
- **Decision Tree**
- **Random Forest**
- **Elastic Net Regression**

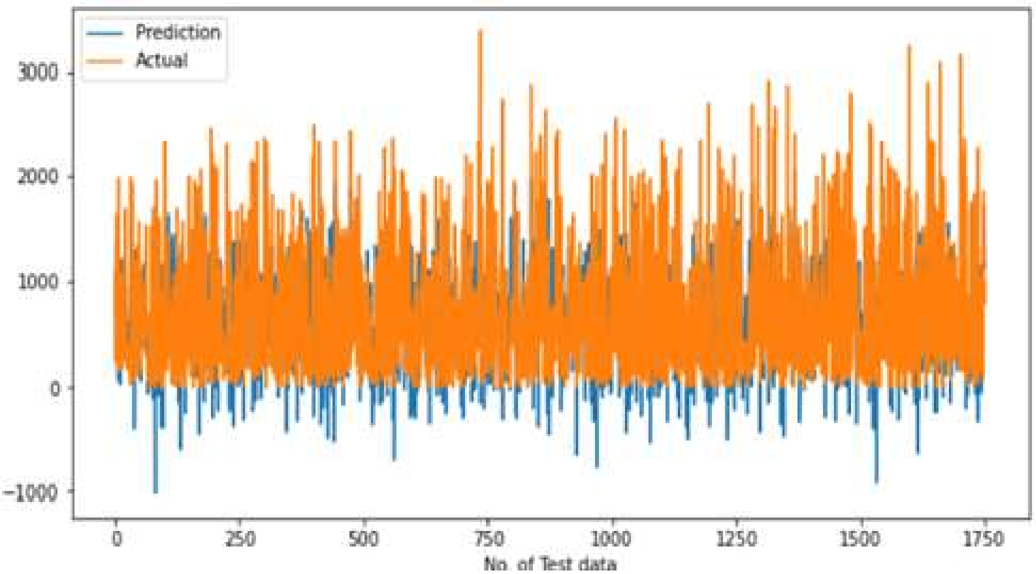
Linear Regression

Train Set Result

```
MSE: 140206.61624939015
RMSE: 374.44173945941196
MAE : 282.480522260274
R2_Score: 0.6638417466299076
```

Test Set Result

```
MSE: 136823.99994832542
RMSE: 369.8972829696447
MAE : 278.93567479799873
R2_score: 0.6673417356182685
```



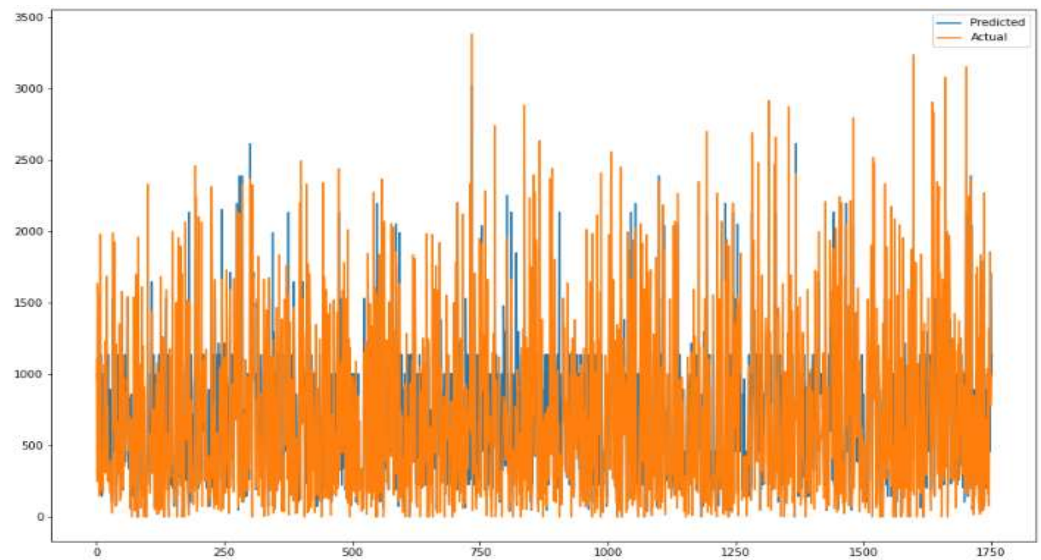
Decision Tree

Train Set Result

```
Model score: 0.5757435377609246
MSE: 176951.07109861638
RMSE: 420.6555254583213
MAE : 288.42324530629
R2_score: 0.5757435377609246
```

Test Set Result

```
MSE : 192208.7355797449
RMSE : 438.41616710580473
MAE : 304.7141588337355
R2 : 0.53268560777997
```



Random Forest

Train Set Result

Model Score: 0.9873599030046023
MSE: 5271.99677836758
RMSE: 72.60851725774036
MAE : 41.69463470319635
R2_score: 0.9873599030046023

Test Set Result

MSE : 32425.597170890414
RMSE : 180.07108921448332
MAE : 111.38534246575342
R2_Score : 0.9211641022007586

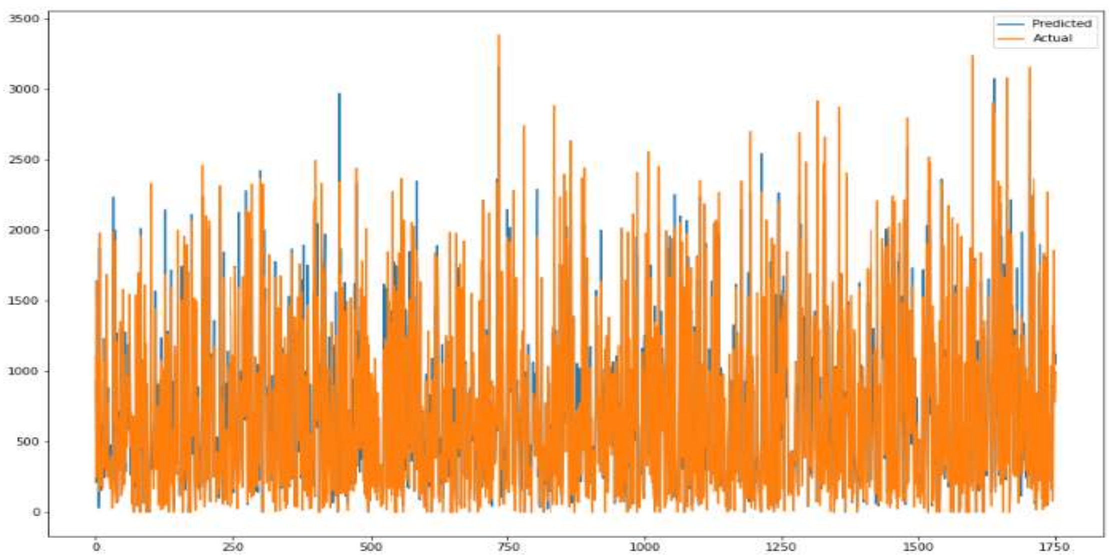
Elastic Net

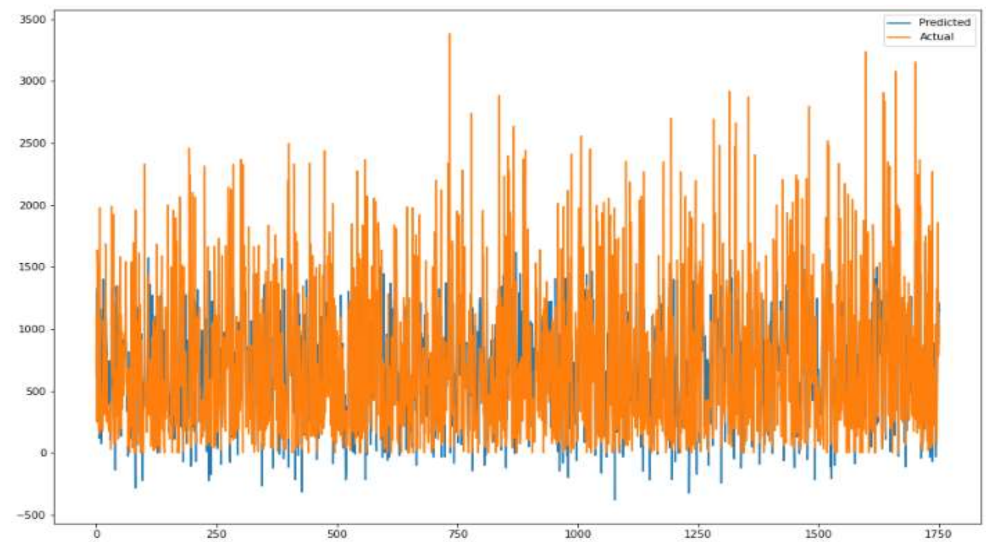
Train Set Result

MSE : 177834.94694853635
RMSE : 421.704810203223
MAE : 309.0419441515174
R2 : 0.5736243641452045

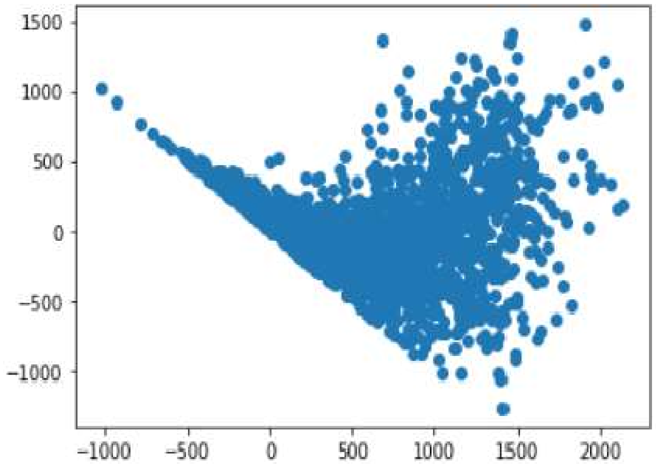
Test Set Result

MSE : 175442.3949535531
RMSE : 418.8584426194046
MAE : 309.43682474292893
R2 : 0.5734493756485335

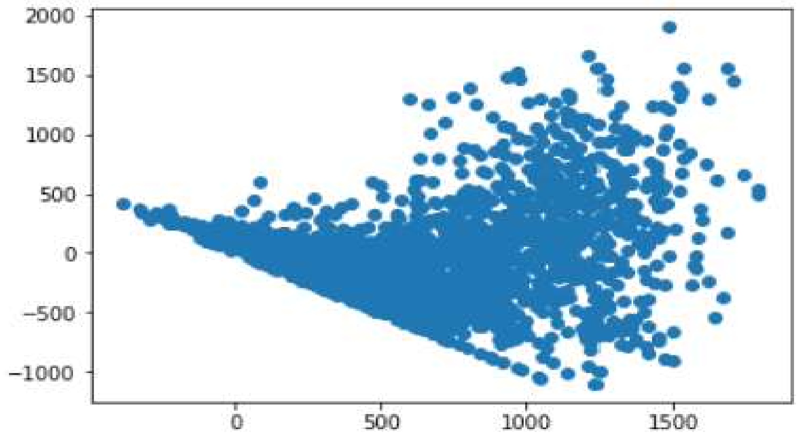




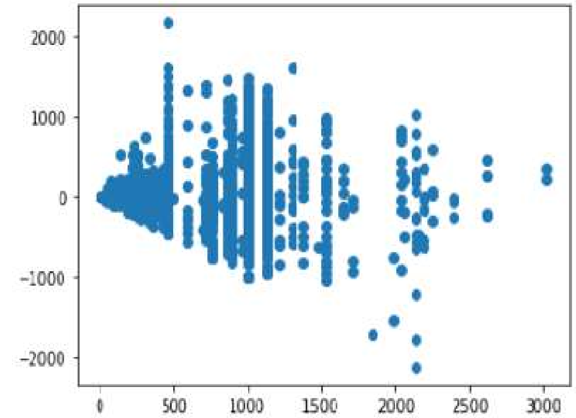
Heteroscedasticity Plot



Linear Regression

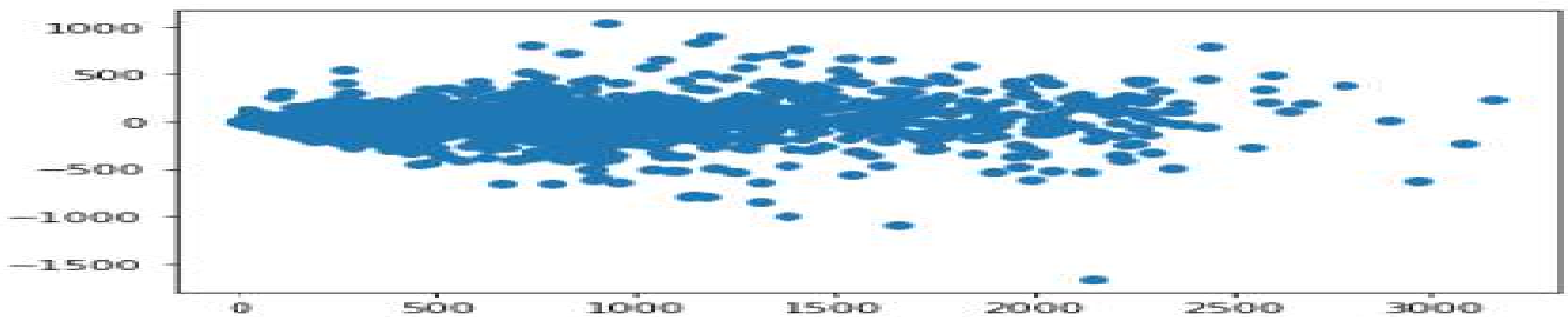


Decision Tree



Elastic Net

Homoscedasticity Plot



Random Forest

Evaluating Models

Model	Train data-MSE	Test data-MSE	Train data-R2-Score	Test data-R2-Score
Linear Regression	140206.61	136823.99	0.663	0.667
Decision Tree	176951.07	192208.73	0.575	0.53
Random Forest	5271.99	32425.59	0.987	0.921
Elastic Net	177834.94	175442.39	0.573	0.573

Conclusion

. We initially did EDA on all the features of our dataset.
. We analysed our dependent variable, 'Rented Bike Count' and also transformed it.
. Random Forest Regressor gives highest R2 Score of 98% for training set and 92% for testing set
. Decision Tree gives the lowest R2 Score of 57% for training set and 53% for testing set
. Hour of the day is most important in prediction.
. Season are also influencing the prediction.
. Variables like Temperature , Hour , Wind Speed , Visibility ,Dew point temperature & Solar Radiation are Positively correlated to our Dependent variable
. Variables like Snow fall , Rain fall & Humidity are Negatively correlated .
. Peak renting from 6 am to 9 am & from 4 Pm to 10 Pm during peak hours.

THANK YOU