Capstone project-2 (LINEAR REGRESSION)

Done by: Sharath S (Individual)

Dataset used: Seoul Bike Data

MAIN Topics to be covered:

1. Introduction

2. Exploratory Data Analysis

3. ML model creation.

Introduction

- **Dataset information:** This dataset contains data from a bike sharing company. This dataset gives us information about the number of bikes rented each hour and also gives us information about different characteristics of each hour. This data can be used for forecasting the bike demand in the future.

- **Problem statement:** From the perspective of the bike sharing companies, one of the main goals is to make sure that there are enough bikes available as per the demand. Predicting the demand for the bikes is crucial for the companies as it would prepare them ahead of time and also make sure that the supply does not fall short.

Variables Description

| Fields	Description
Date	Date
Hour	Hour of the day (0-23)
Temperature	Temperature of the day
Humidity	Humidity measure
Windspeed	Windspeed
Visibility	Visibility measure
Dew Point Temperature	Dew Point Temperature Measure
Solar Radiation	Solar Radiation Measure
Rainfall	Rainfall in mm
Snowfall	Snowfall measure
Seasons	1 = spring, 2 = summer, 3 = fall, 4 = winter
Holiday	Whether a holiday or not
Functional Day	Whether a functional day or not

Exploratory Data Analysis
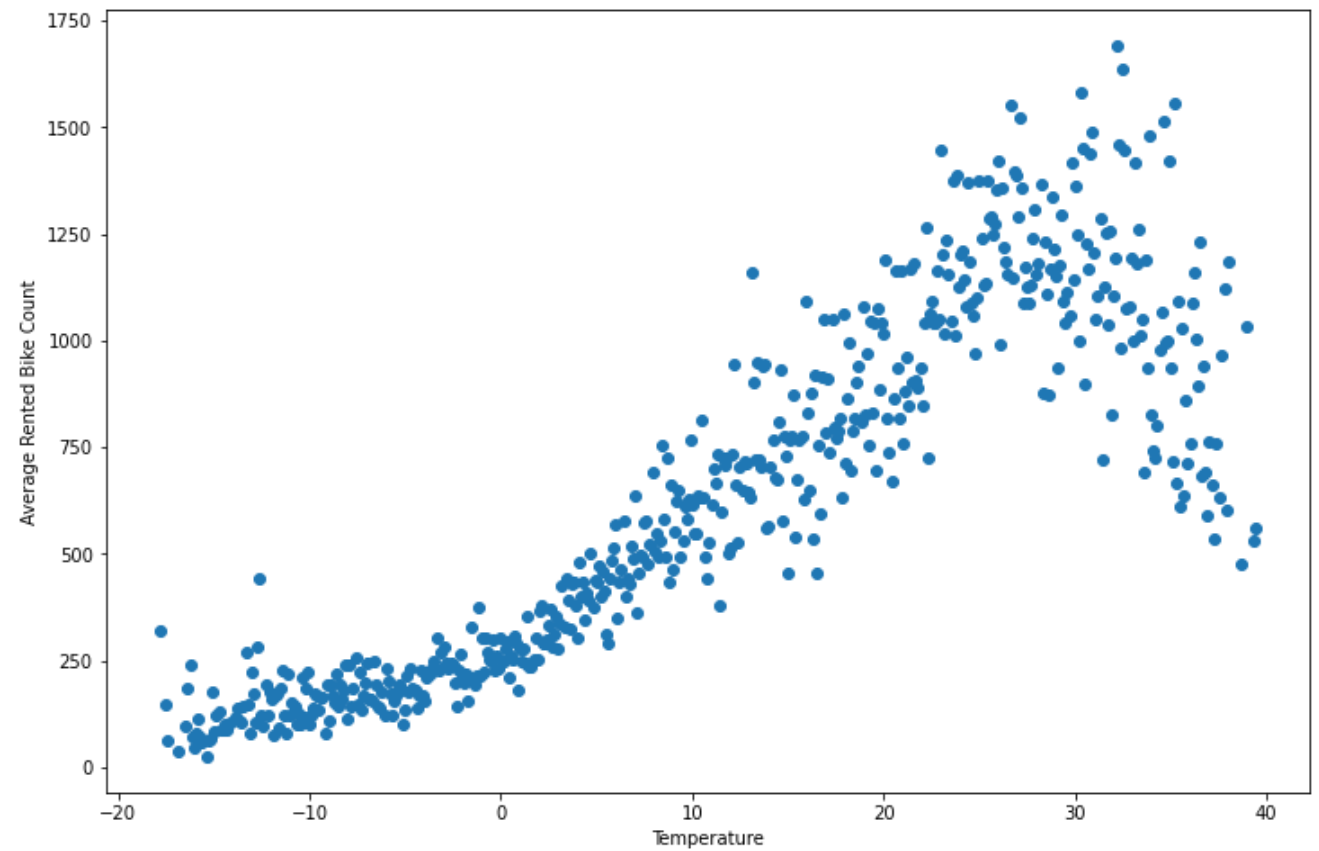
Null values
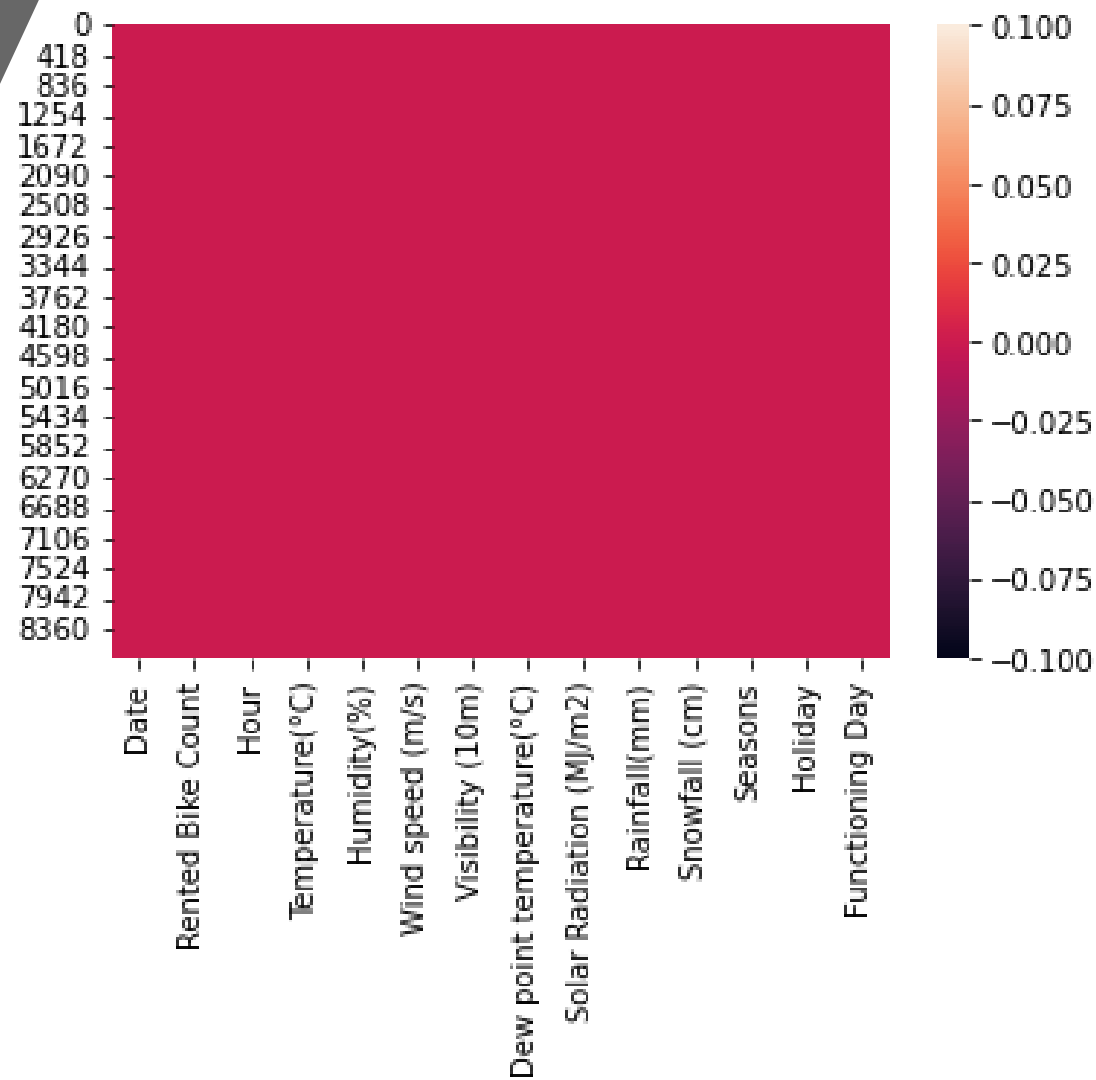
- Histogram of Rented Bike Count
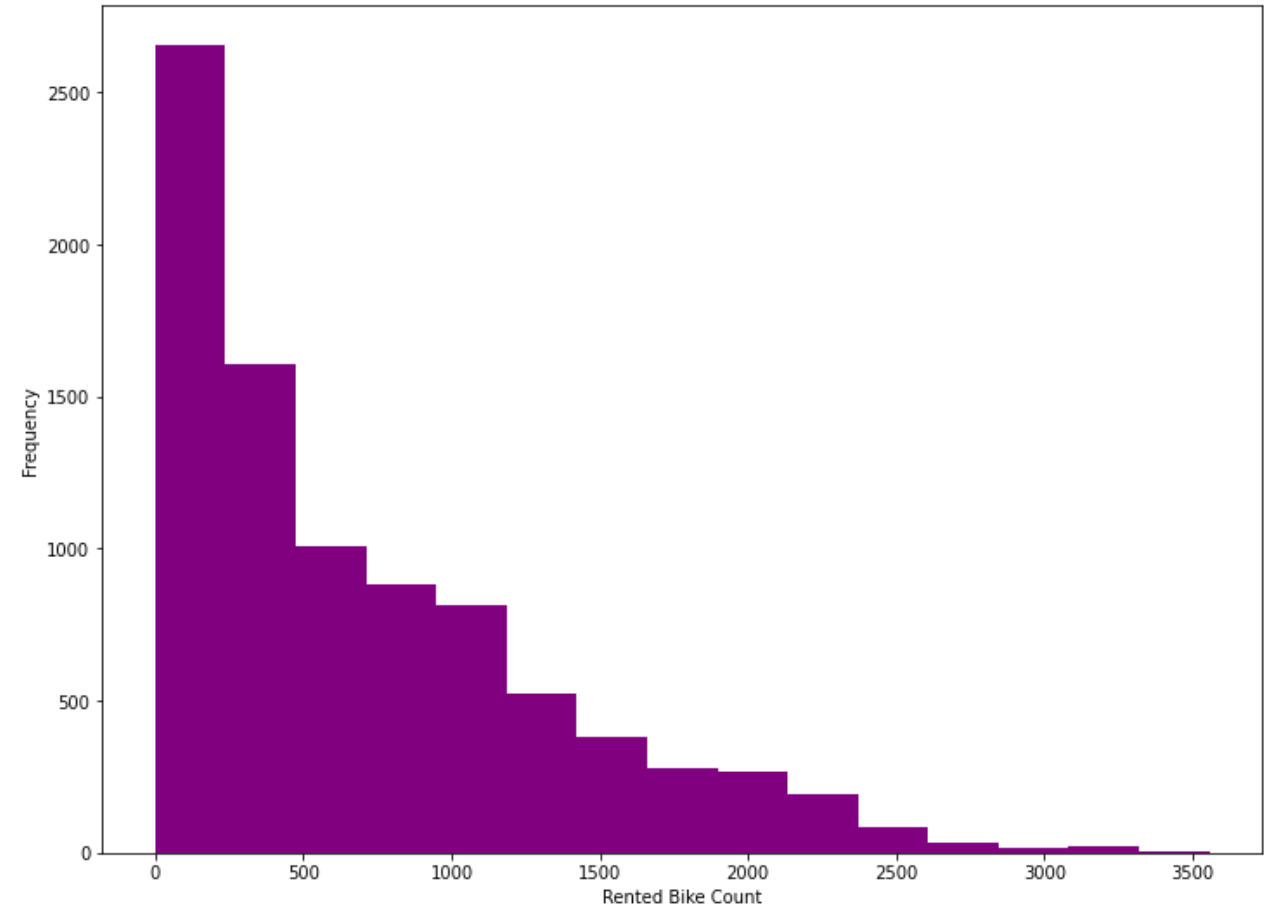
- Bar chart of Rented bike count by Hour
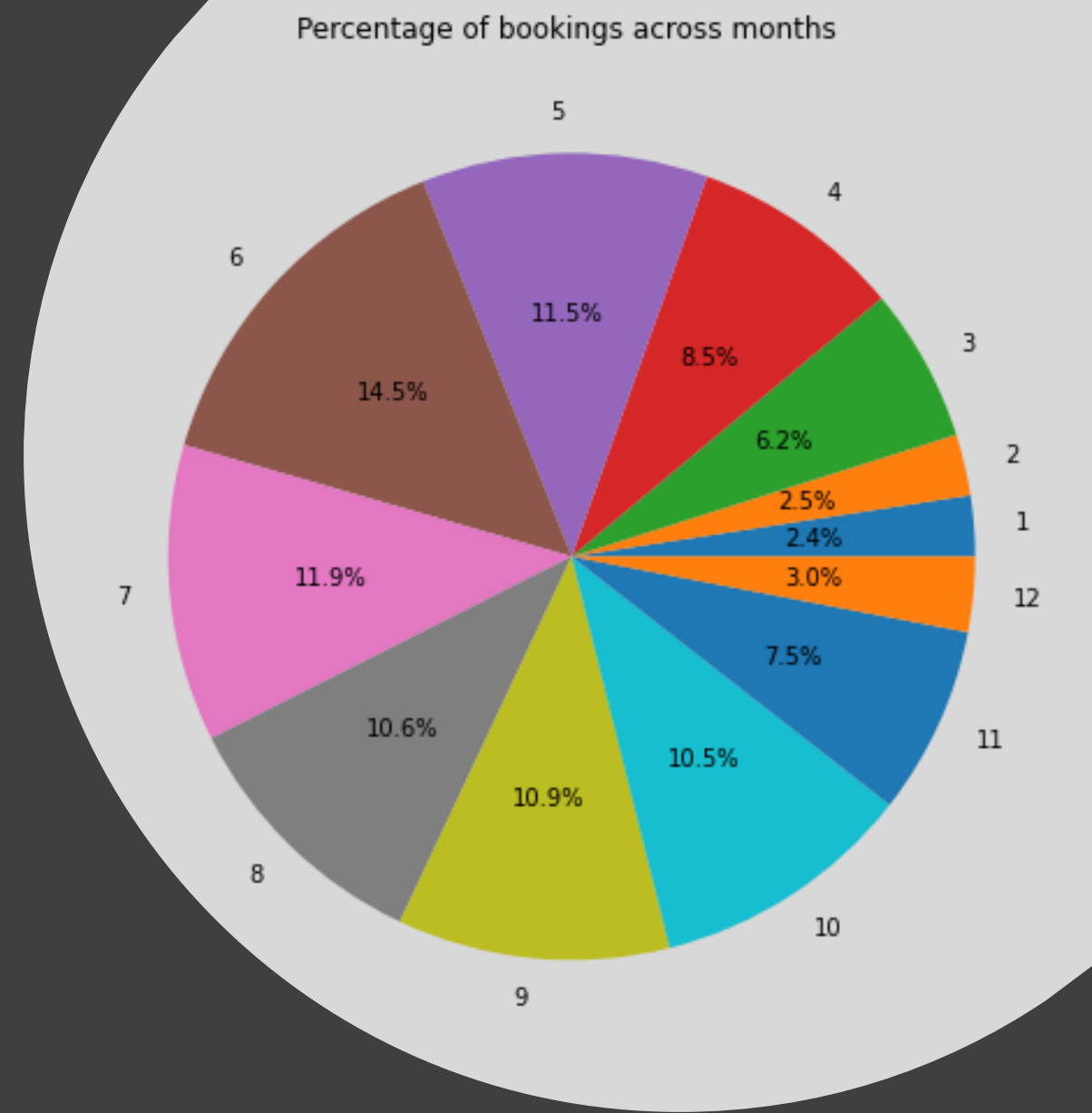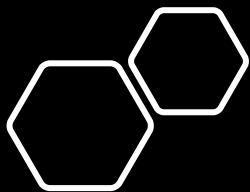

Rented bike count by hour
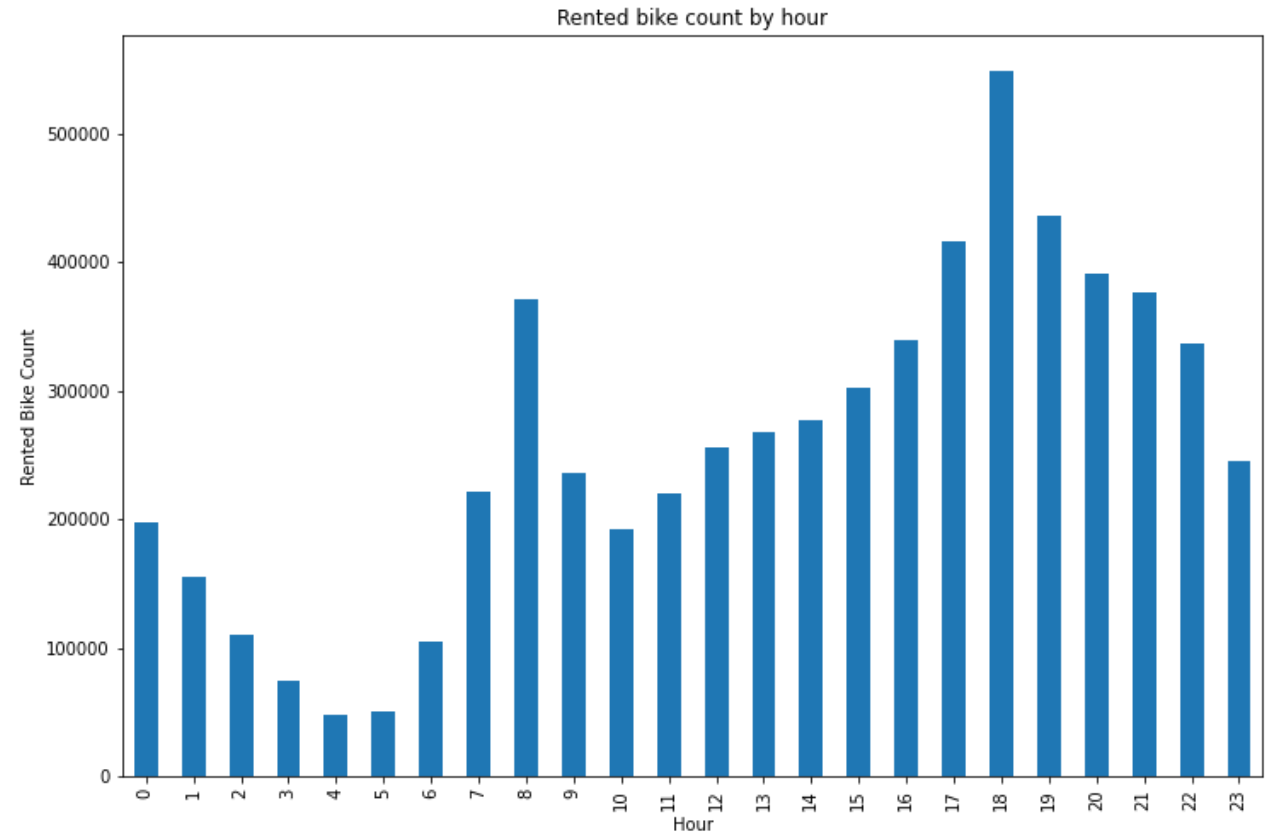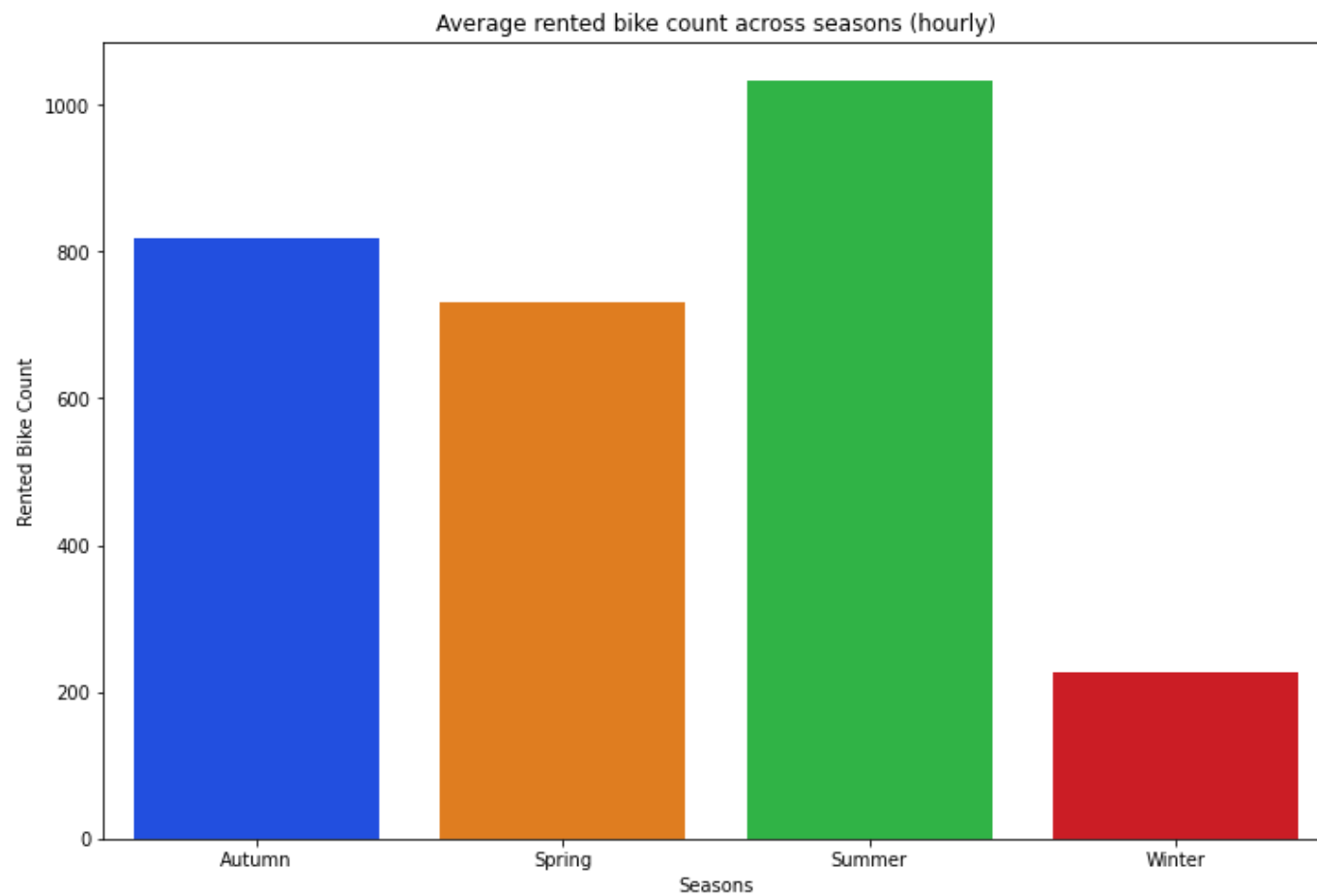
- Scatter plot of Temperature vs Rented Bike Count

- Pie chart of percentage of bookings across months

Percentage of bookings across months

- Barplot of Seasons vs Rented Bike Count

Average rented bike count across seasons (hourly)

- Scatterplot of Visibility vs Rented Bike Count

Average rented bike count (hour) vs visibility
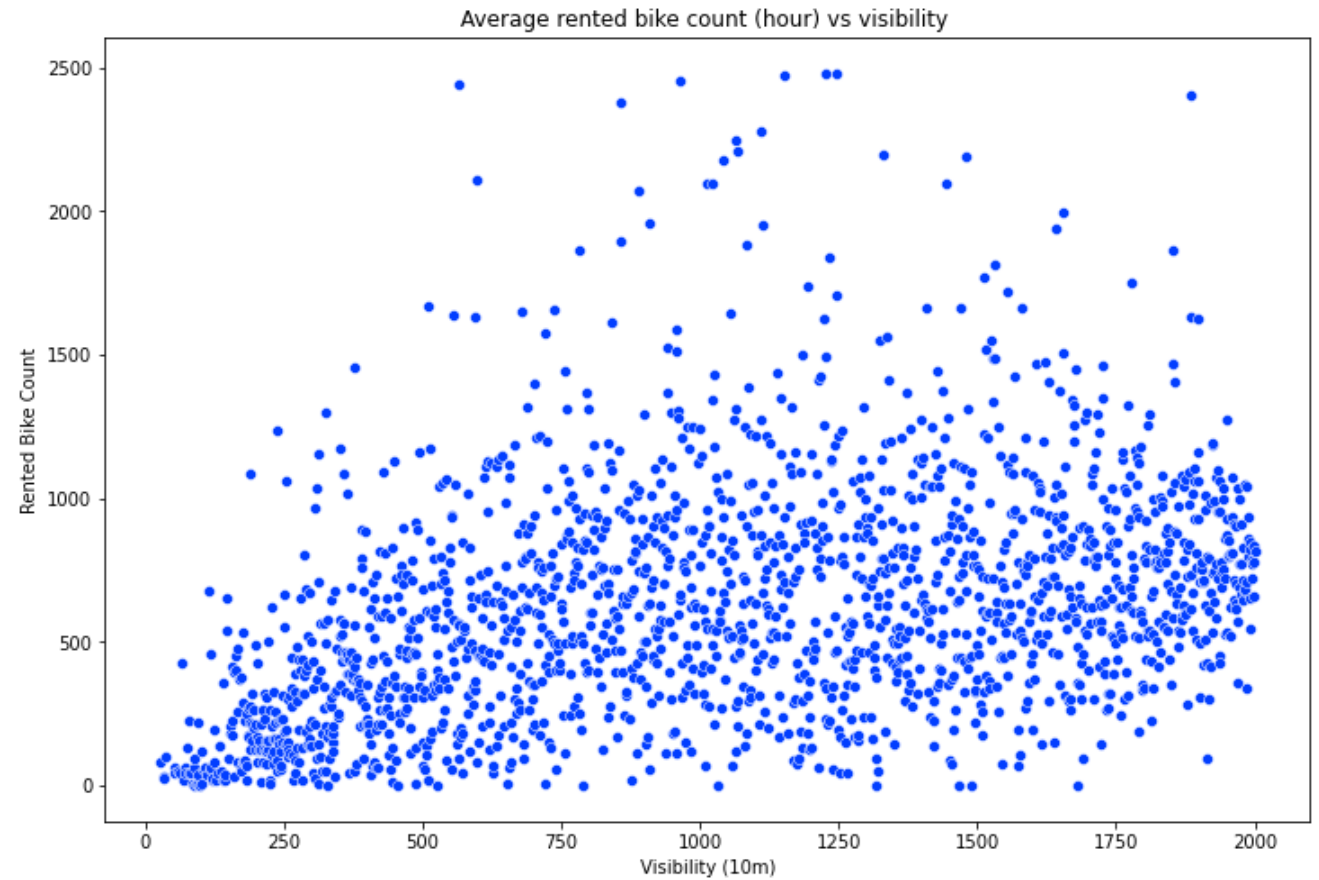
- Donut Chart of Holidays with mean of Rented Bike Count.
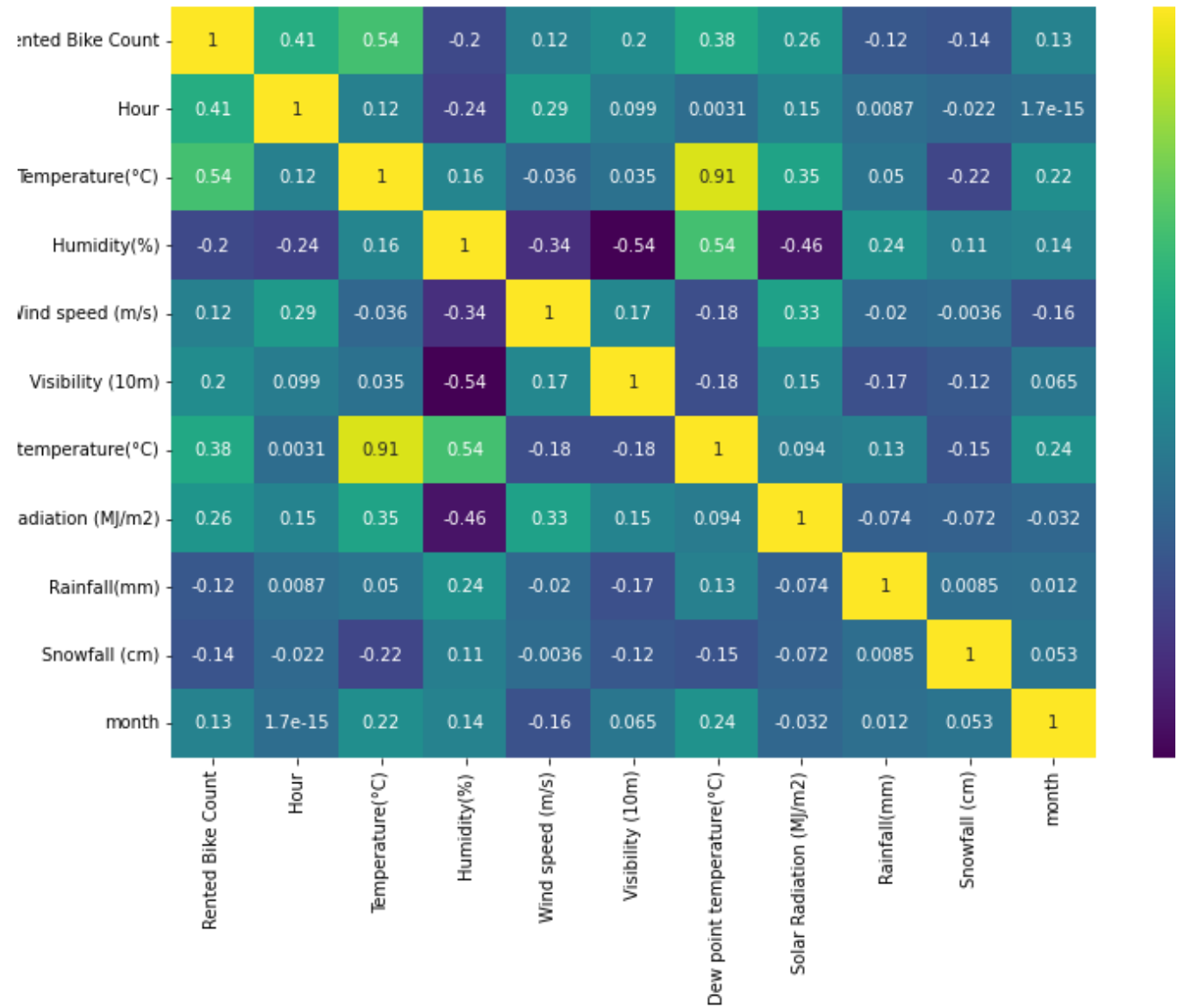
- Pointplot

Number of bikes rented per hour by season
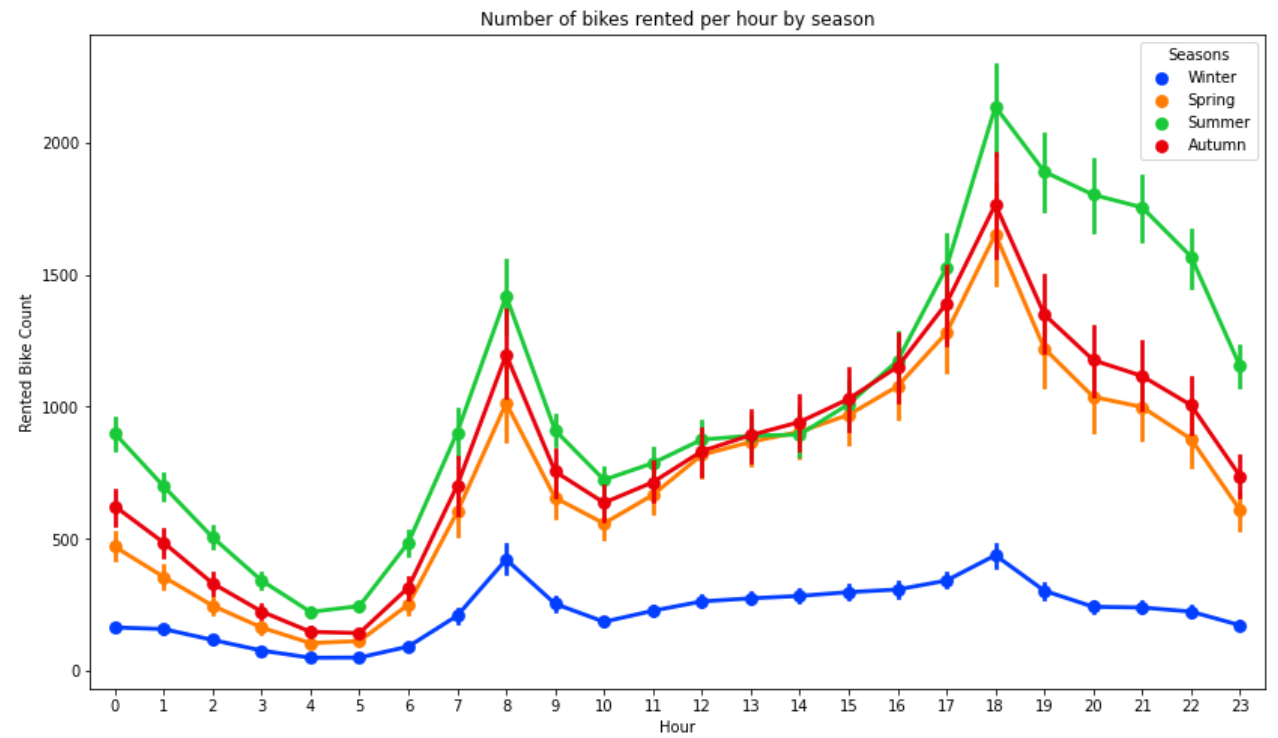
- Correlation Heatmap

ML Model Creation

Variable Selection: Variables with their corresponding VIF values:

Feature	VIF
Hour	3.840063
Temperature(°C)	3.283674
Humidity(%)	6.128147
Wind speed (m/s)	4.763637
Visibility (10m)	5.219203
Solar Radiation (MJ/m2)	2.265781
Rainfall(mm)	1.080624
Snowfall (cm)	1.131918
month	4.983588

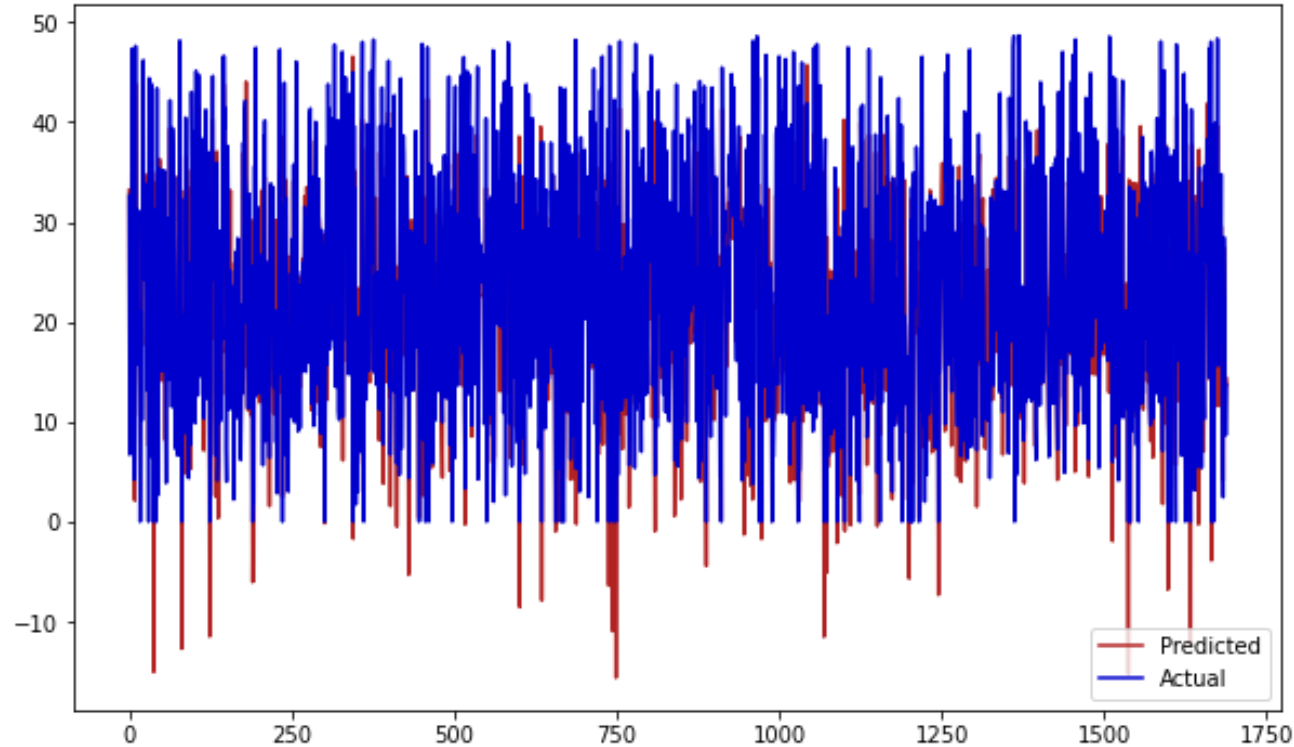- Square root transformation was done on the dependent variable.
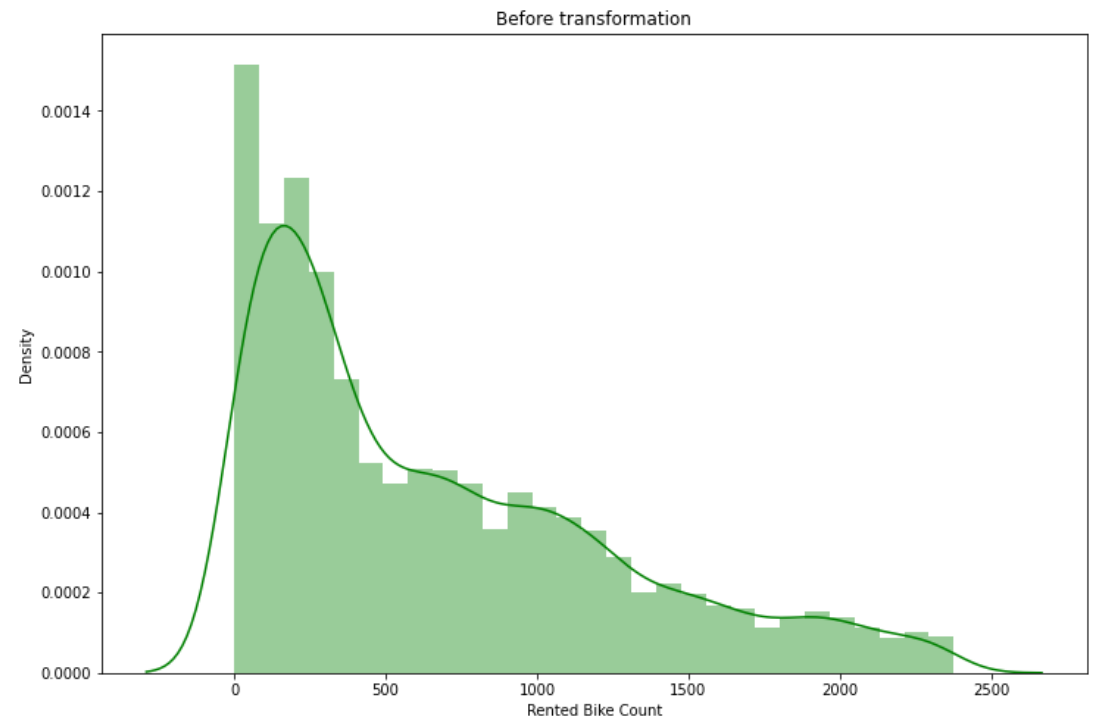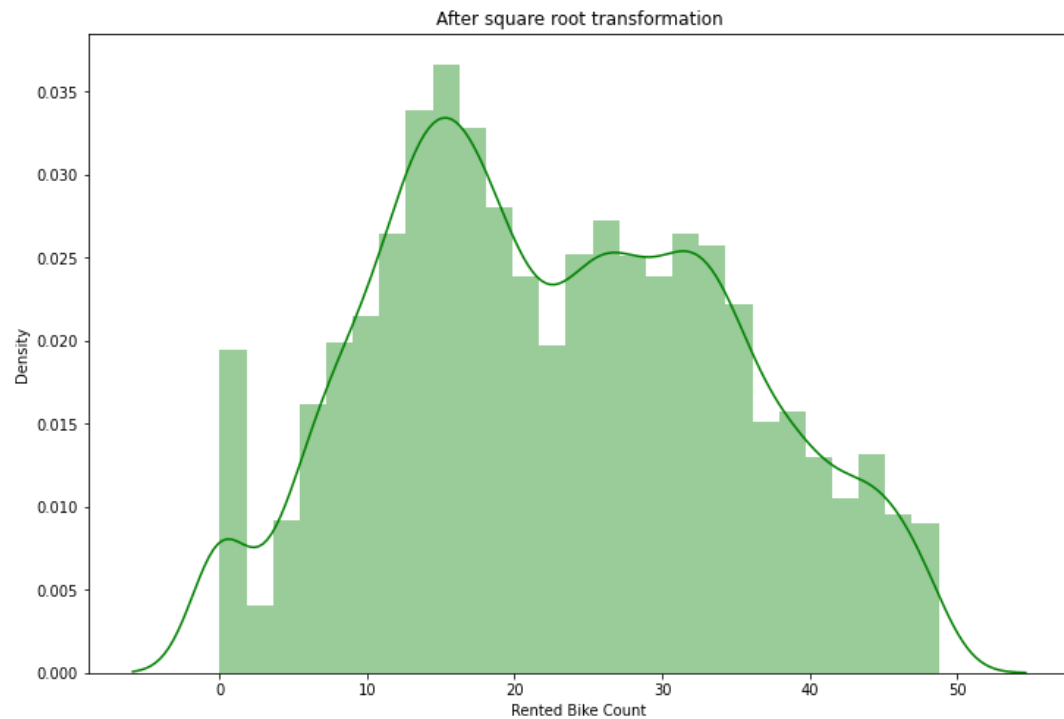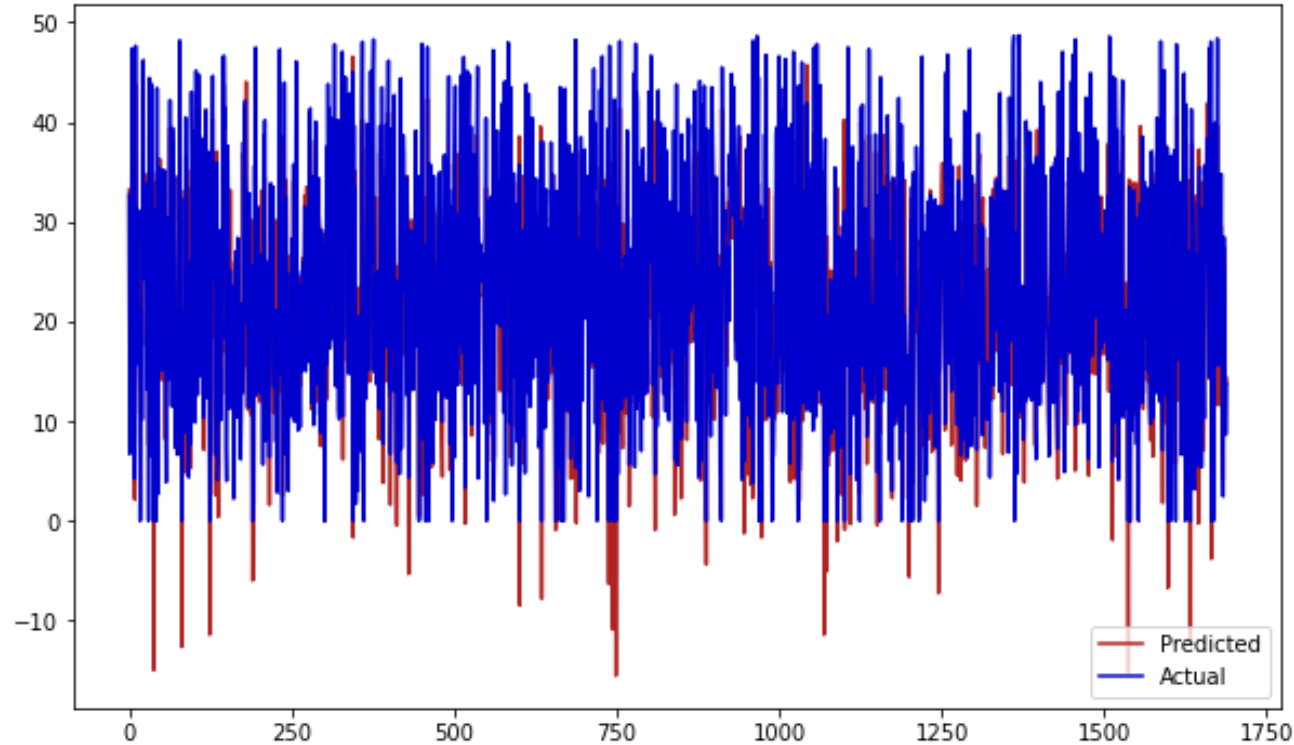
1. Linear Regression Model
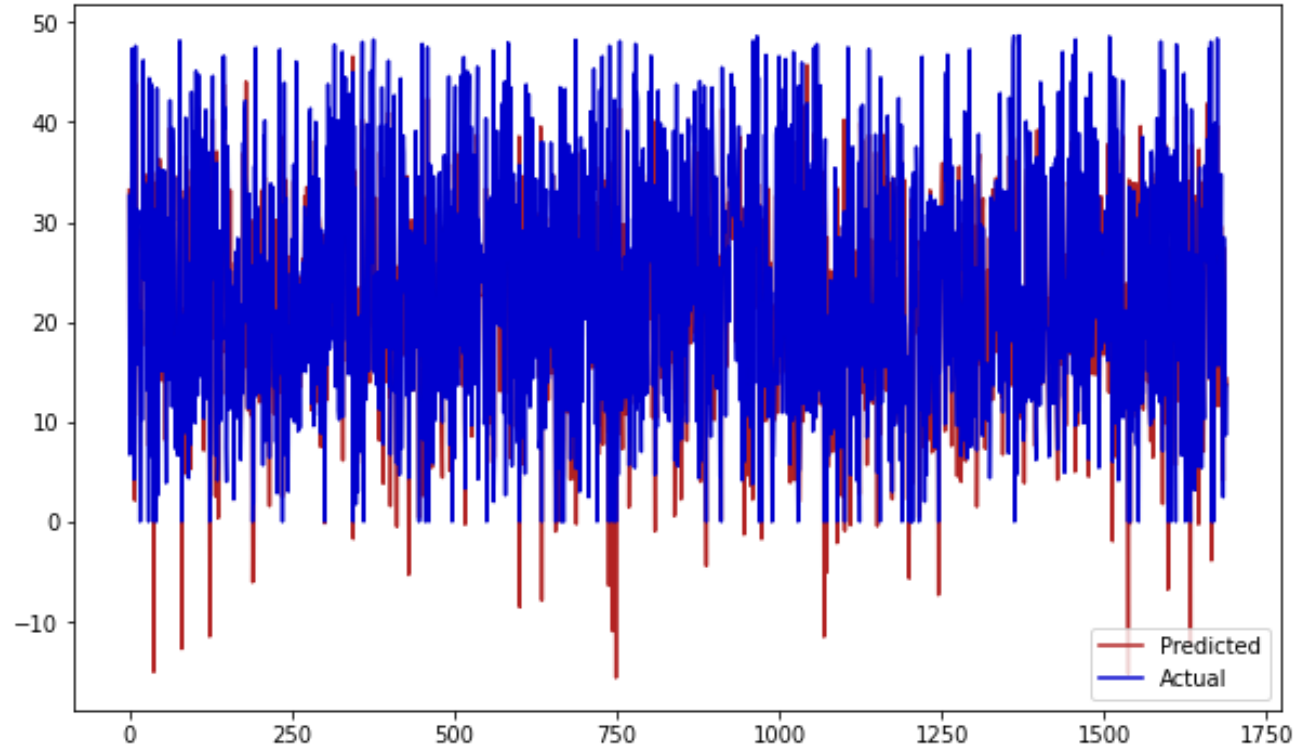
- R2: 0.7626412266265183

- Adjusted R2: 0.7552464206142718

- MSE : 33.88747574312986

- RMSE : 5.821295022856157

- MAE: 4.398558055895357

2. LASSO Regression model with CV

- R2: 0.7626469891350846

- Adjusted R2 : 0.7552523626512051

- MSE : 33.88665303551944

- RMSE : 5.821224358802832

- MAE: 4.398482074048917

3. Ridge Regression with CV

- R2: 0.762625478563748
- Adjusted R2 : 0.755230181929276765
- MSE : 33.88972407837169
- RMSE : 5.821488132631698
- MAE: 4.399340115609134

Thank you