Capstone Project

Bike Sharing Demand Prediction

Team members

Mujahid Sayyed
Sarthak Gupta
Lova Kumar Poluparti

Contents

Introduction

A bike rental or bike hire business rents out motorcycles for short periods of time, Usually for a few hours. Most rentals are provided by bike shops as a sideline to their main businesses of sales and service, but some shops specialize in rentals.

As with car rental, bicycle rental shops primarily serve people who do not have access to vehicles, typically travelers and particularly tourists.

Bike rental shops rent by the day or week as well as by the hour, and these provide an excellent opportunity for those who would like to avoid shipping their own bikes but would like to do a multi-day bike tour of a particular area.

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 the bike count required at each hour for the stable supply of rental bikes.

Discussion Topics

- Bike booking in each season, on functioning days, holidays, and months.
- Comparing Rented Bike Count against Numerical data columns.
- Checking for Linear relation between the Rented bike count and the Numerical data columns.
- Climate Effect in Different seasons on Bike Sharing.
- Heat Map(OR) Correlation Map.
- Linear Regression analysis, Lasso Regression Analysis, Grid Search CV for Hyperparameter tuning,
- Decision Tree Analysis, Cat Boost, XG boost, and Random Forest Analysis
- Feature Importance.

Data Analysis Steps

Imported Libraries

In this part, we imported the required libraries NumPy, Pandas, matplotlib, and seaborn, to perform Exploratory Data Analysis and for prediction, we imported the Scikit learn library.

Descriptive Statistics

In this part, we start by looking at descriptive statistic parameters for the dataset. We will use describe() this told mean, median, standard deviation

Missing Value Imputation

We will now check for missing values in our dataset. after checking not existed any missing values, In case there are any missing entries, we will impute them with appropriate values.

Graphical Representation

We will start with Univariate Analysis, bivariate Analysis and conclude with various prediction models driving the Demand for bikes.

Attributes of each variable

Date: Date in year-month-day format

Rented Bike Count: Count of bikes rented at each hour

Hour: Hour of the Day

Temperature: Temperature in Celsius

Humidity: Humidity in %

Windspeed: Speed of wind in m/s

Visibility (10m): Visibility

Dew point temperature: Dew Point Temp (Celsius)

Solar radiation: Radiation in MJ/m2

Rainfall: Rainfall (mm)

Snowfall: Snowfall (cm)

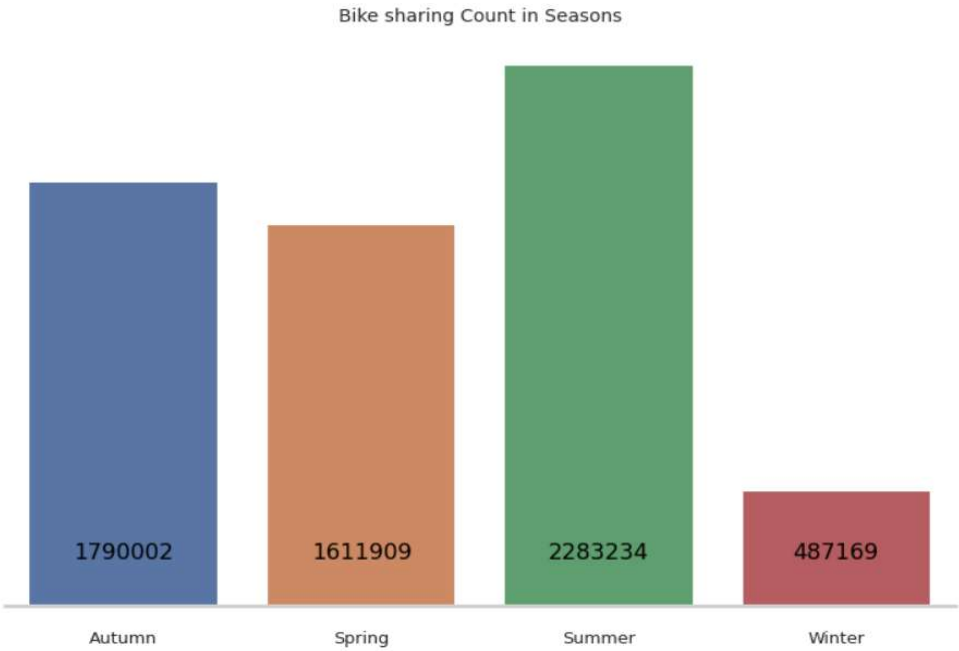
Seasons: Winter, Spring, Summer, Autumn

Holiday: Holiday/No holiday

Functioning Day: if the day is neither weekend, holiday than 1 else 0

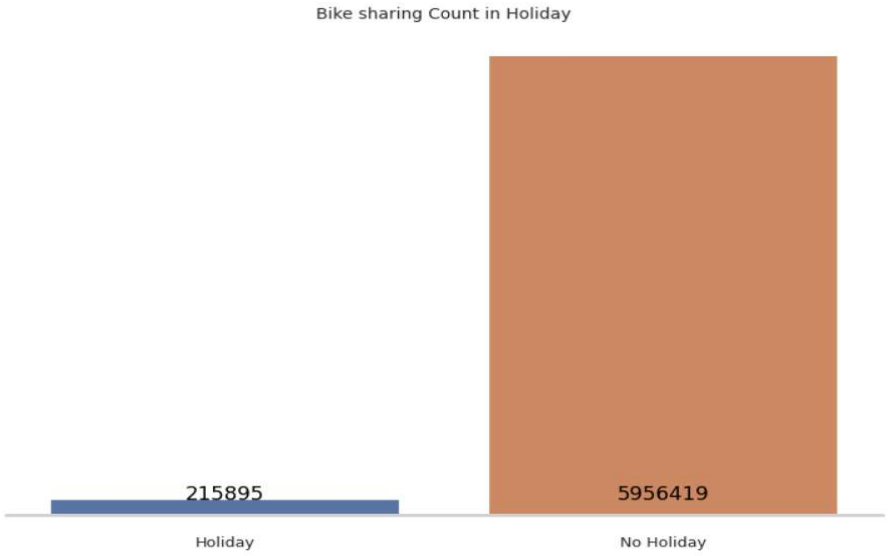
Bikes Rented per Season

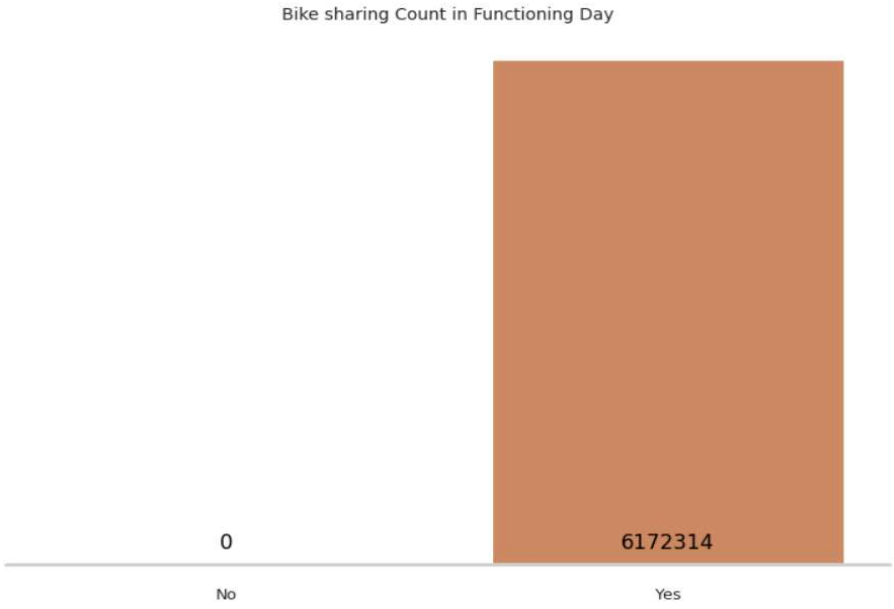
- Highest number of bikes were rented in **Summer.** The total count of bikes rented in summer was 2.28 million
- Second highest Bikes were rented in **Autumn** around **1.79 million** followed by **Spring** in which **1.6 million** bikes are rented.
- **Winter** appears to be the least popular season for bike rentals. In the winter, just 487K bikes were rented.
- The **extreme temperatures** in Seoul in the **winter** might be a factor in the **low demand** for bikes in the winter.



Bike Renting trend on holidays, Functioning days

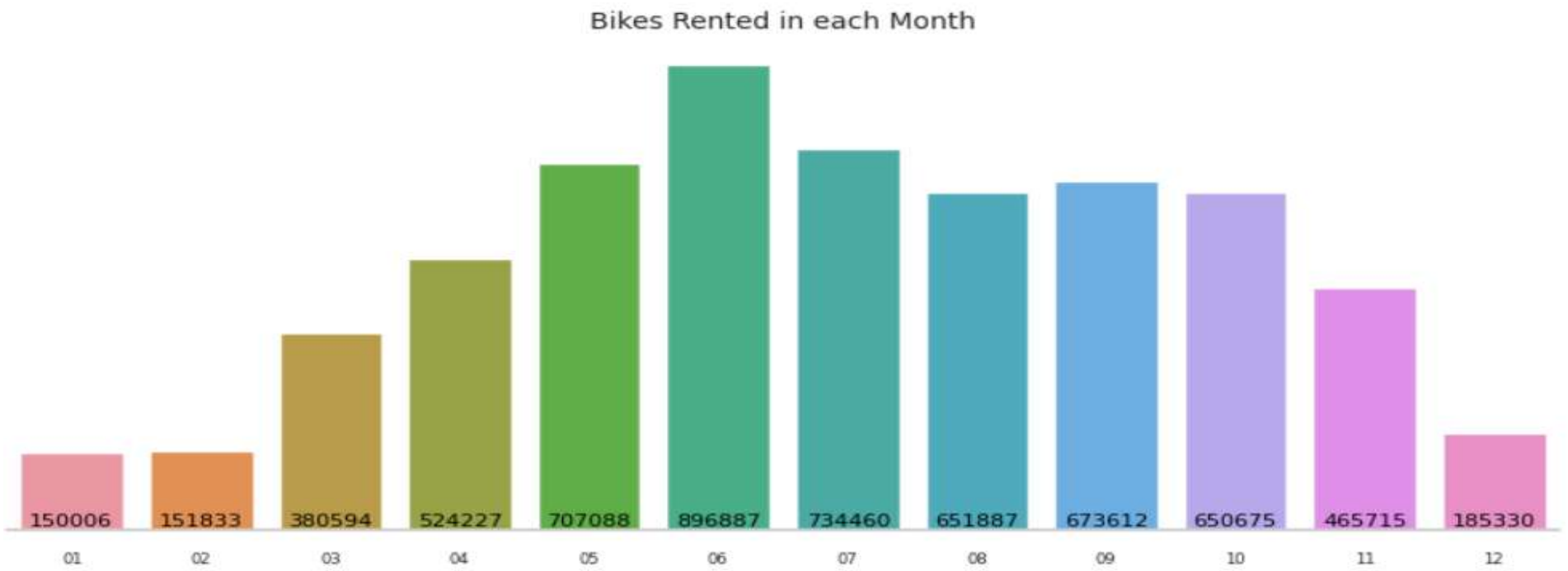
- **People prefer** to use the bike on **Non-holiday more** compared to **Holidays**.
- **5.9 million** bikes are rented on **Non-holidays**, only a meager **215K** bikes were rented on **holidays**.
- It's reasonable to conclude that the **majority of clients** in the **bike rental sector** are from **Seoul's working class**.
- All the bikes rented were on the functioning days.





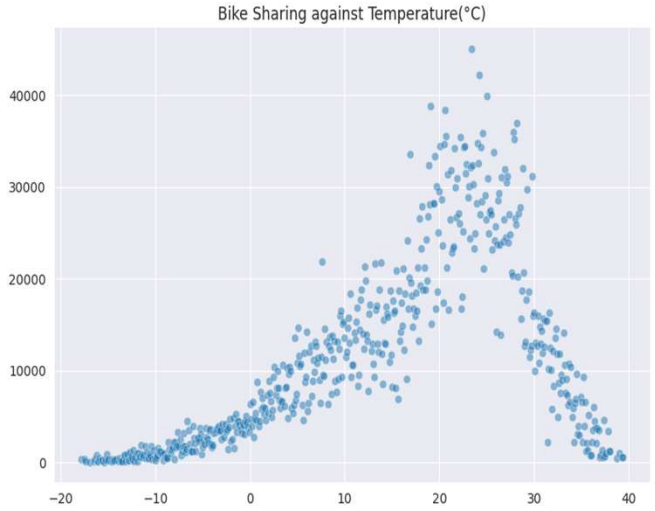
Bike Booking Monthly Trend

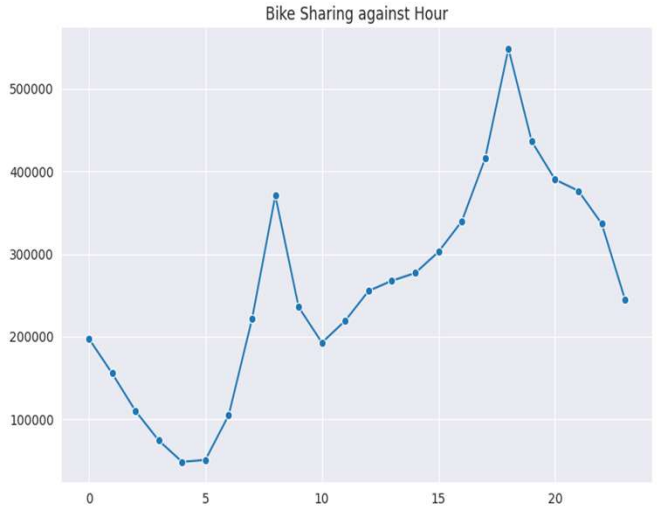
- **June** is the most preferred month for bike booking around **896K** bikes were rented in June.
- **July** and **May** are the second and third best.**734K** bikes were booked in **July**, and **707K** were booked in **May**.
- Demand for bikes was **least** in **Jan**, followed by **Feb** and **Dec**. **150K** bikes were rented in **Jan**, **151k** in **Feb**, and **185K** in **Dec.**



Rented Bike Count Against Numerical Data

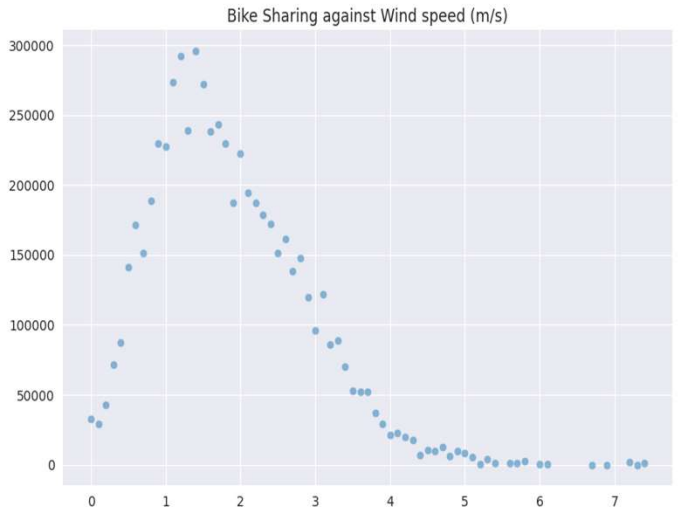
- **Most preferred** bike-sharing **temperature** is **20- 30** degrees Celsius. Bike renting is **minimal** when the **temperature** is **>35 or <5** degrees Celsius.
- Bike sharing is at its **peak between 4 pm to 8 pm**. Bike-sharing is at **least between 2 am to 6 am**, it **increases from 6 am onwards until 8 am**.
- **Snowfall** is **least favorable** for the bike renting Business.

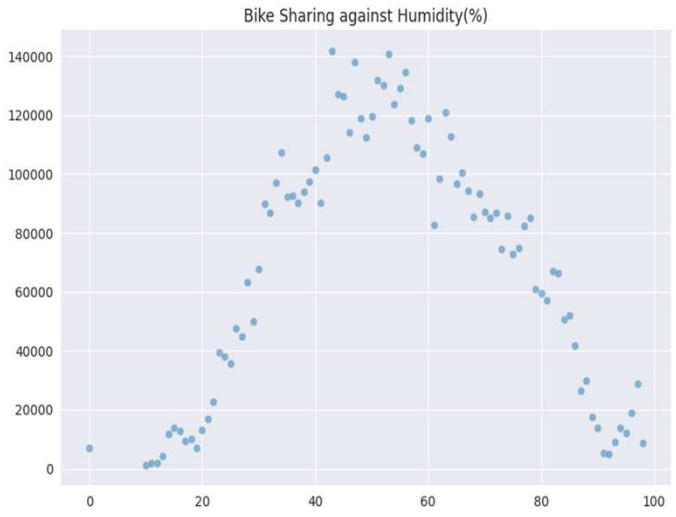


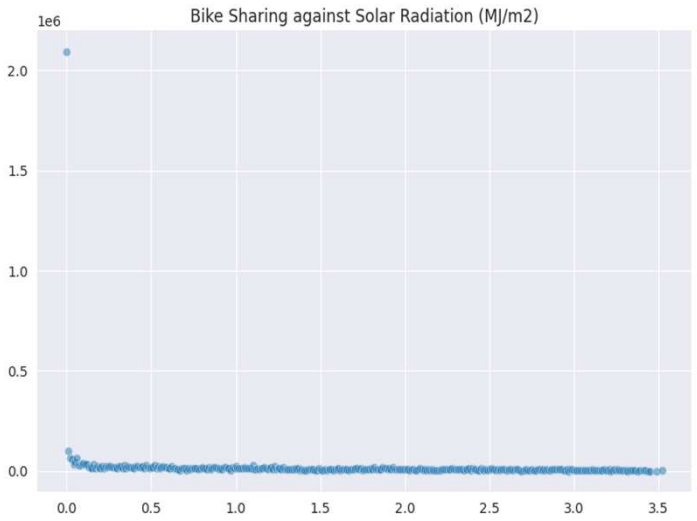


Rented Bike Count Against Numerical Data

- Bike renting is at its **peak** when the **humidity is 40%- 60%**. People avoid bikes when the climate is too humid or too dry.
- Favorable wind speed for Bike sharing is 1m/s -2 m/s as wind speed goes beyond 2m/s the count of bike-sharing starts dropping reaching minimal when the **speed > 5m/s.**
- Bike sharing is at its **peak** when the **radiation is minimal**.

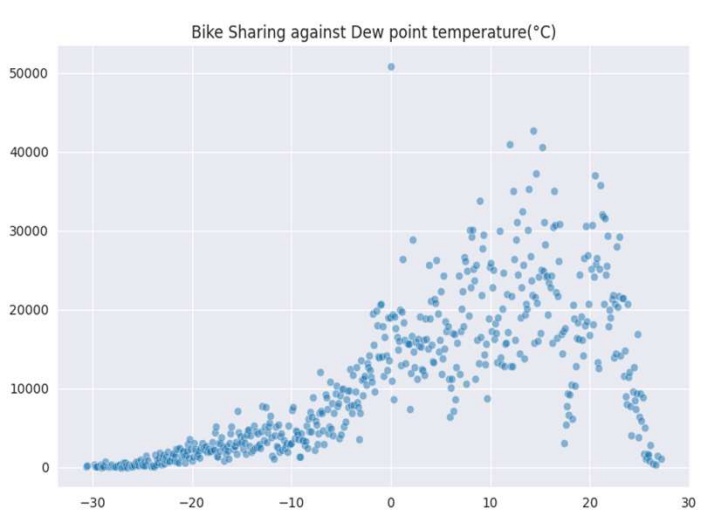


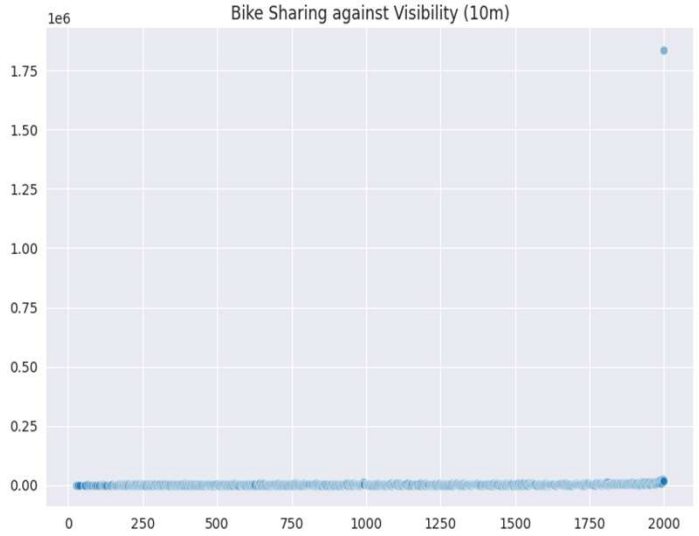


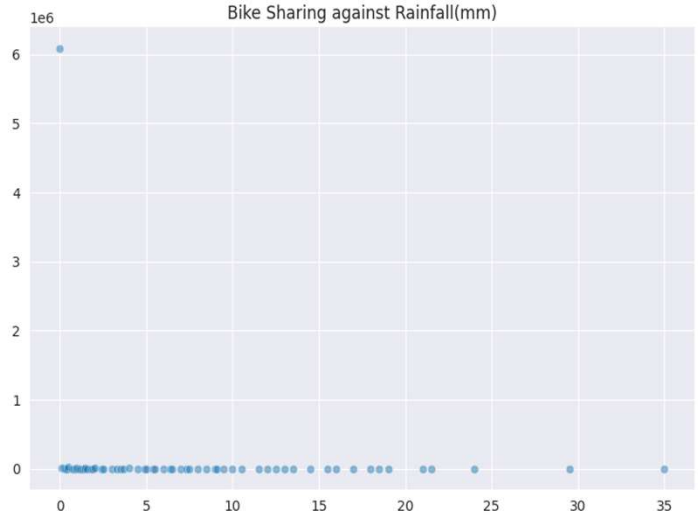


Rented Bike Count Against Numerical Data

- Dew point temperature between **5-25 Degrees** is **most favorable** for Bike sharing.
- Demand for bikes **dwindles** in case of **rainfall**.
- **Visibility** is an important factor for bike riders, bike sharing is at its **peak** when the **visibility is maximum**

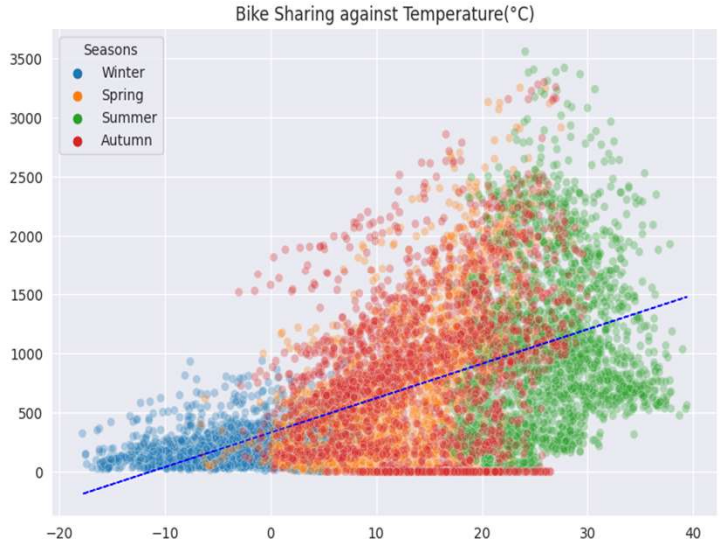


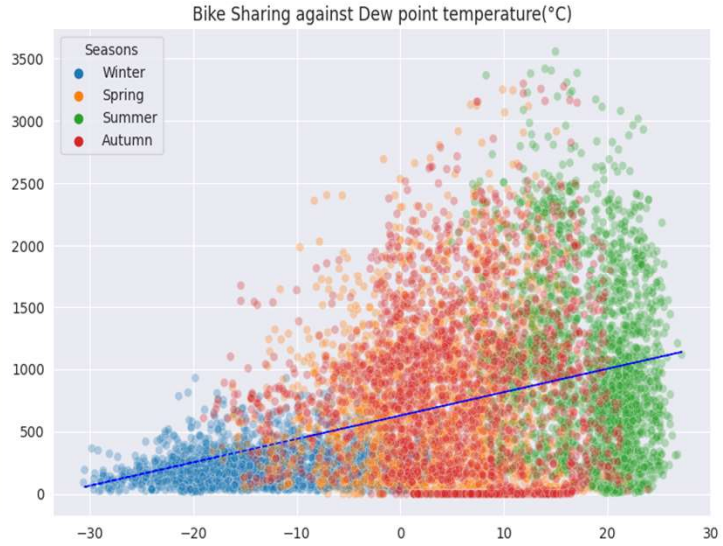


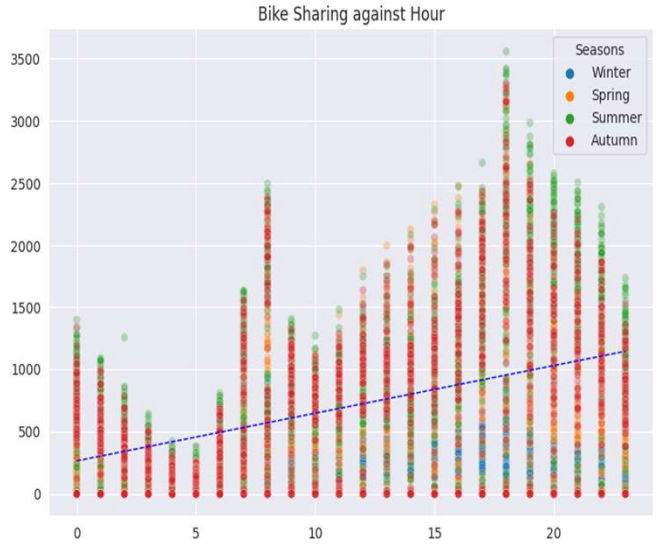


Co-relation: Rented bike count vs Temp, Dew point Temp, Hour

- Bike sharing is positively co-related to temperature and Dew point Temperature as the temperature approaches **30 degrees**.
- Though one thing to notice the positive co-relation is applicable only because the temperature in Seoul rarely crosses **40 Degrees**.
- Bike sharing count is positively co-related to hours as the Hours Progress from 0 (12 am) to 20 (8 pm) the bike-sharing count increases.

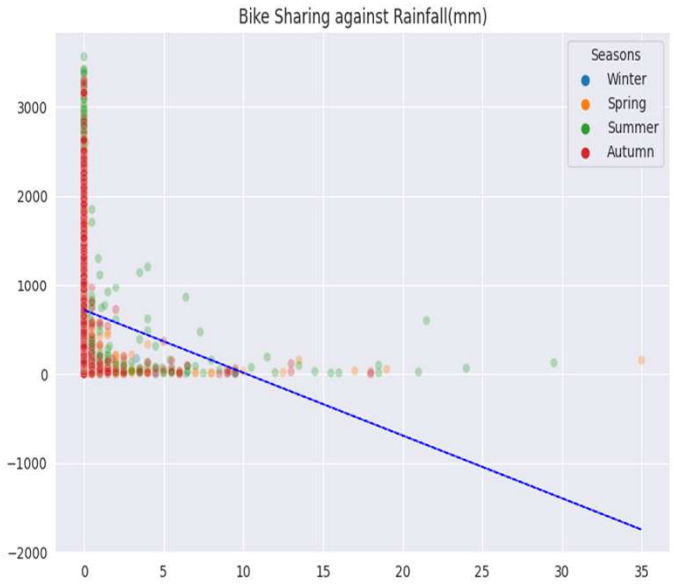






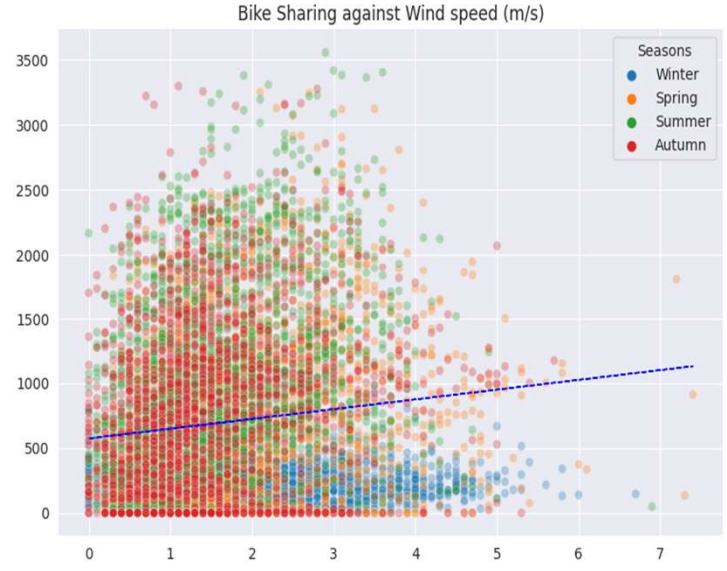
Co-relation: Rented bike count vs Visibility, Rainfall, Snowfall

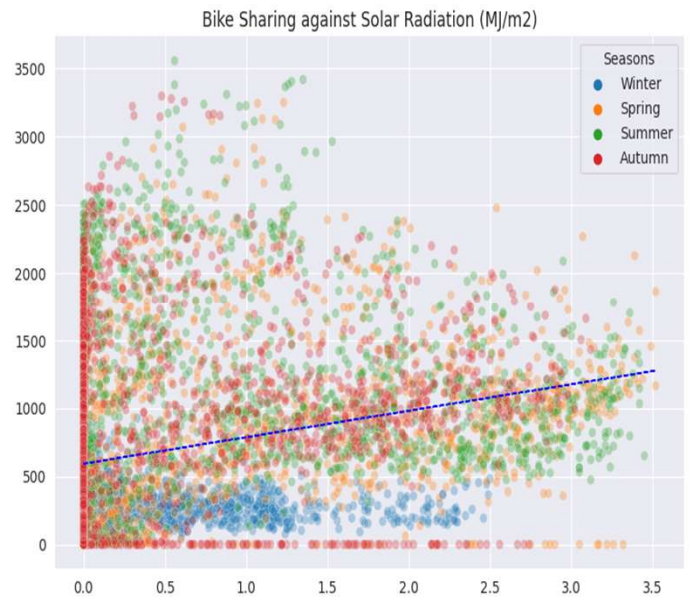
- Visibility is Also slightly positively co-related with Bike Bookings.
- Snowfall, Rainfall are negatively co-related to Bike rented count.



Co-relation: Rented bike count vs Humidity, Wind Speed, Radiation

- The bike-sharing count is slightly negatively correlated to Humidity.
- Wind speed and Solar radiation are slightly positively related to Bike-sharing count.



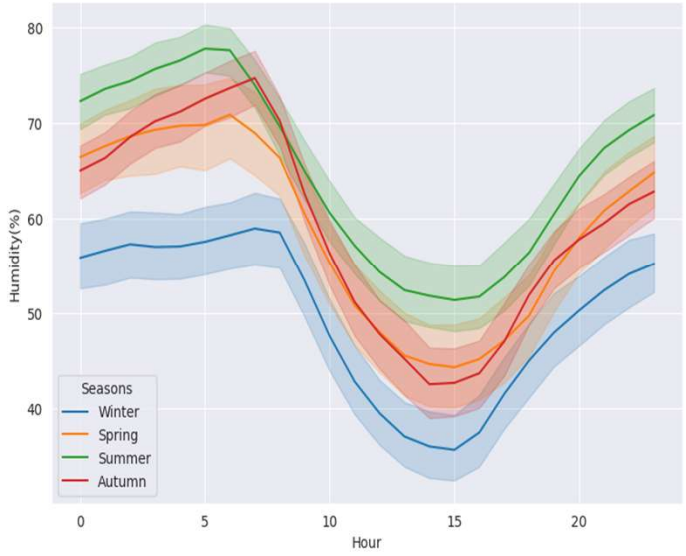


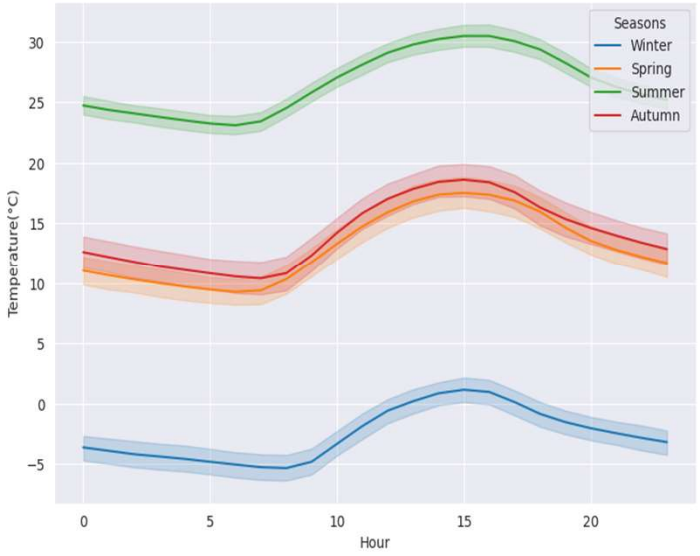
Climate Effect in Different seasons on Bike Sharing

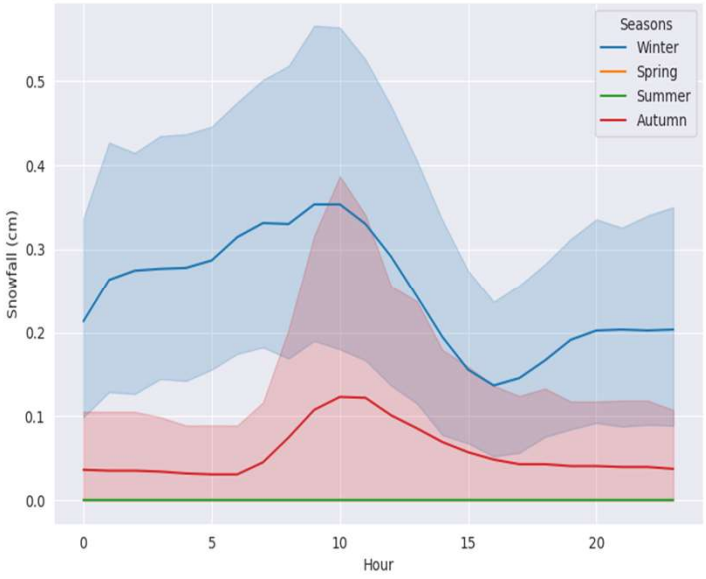
- Temperature in Summer varies between **24-31 Degrees**, in Autumn it varies between**10-18 Degrees.** Spring temperatures are between **9-16 Degrees**, for winter the range is **-5-3 Degrees.**

- Humidity in Summer ranges from **52% - 77%**, in Autumn its **43% -74%**, for spring it varies from **41%-71%**, for winter the humidity is least from **36%-59%**

- Seoul experiences Heavy snowfall in winter followed by Autumn. Snowfall is one of the factors affecting Bike bookings.

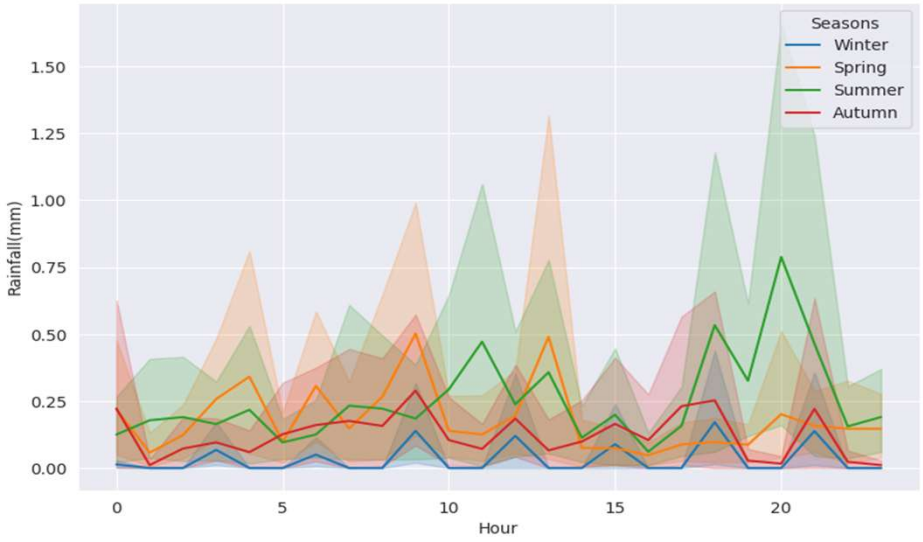


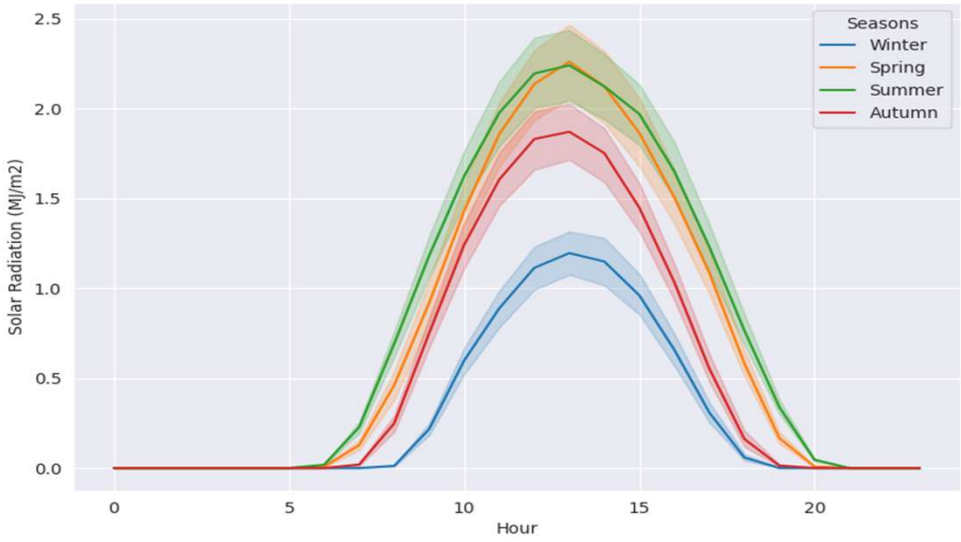




Climate Effect in Different seasons on Bike Sharing

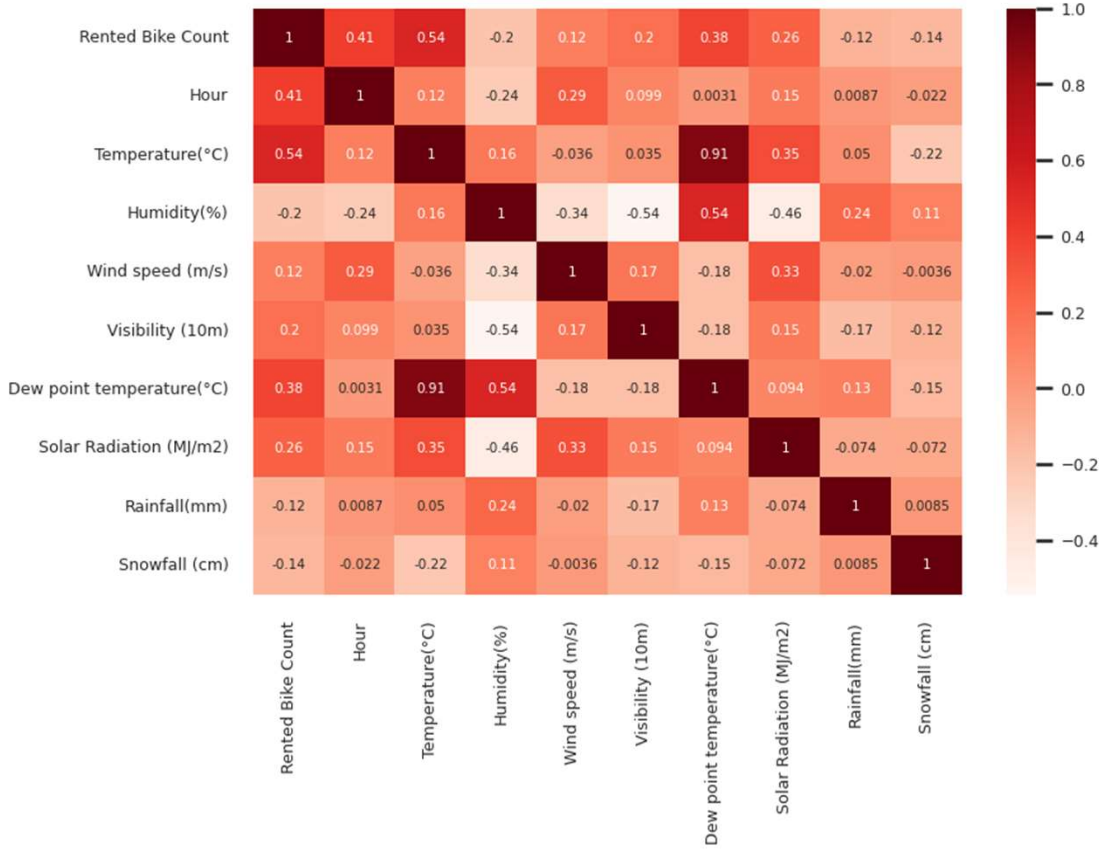
- Solar Radiation is at its peak between **13:30pm -2:30 pm** in Seoul across all the seasons.
- **Summer and Spring** have the highest recorded solar Radiation of **2.25 MJ/m2**. Peak radiation in Autumn and Winter is around **1.8 MJ/m2** and **1.2 MJ/m2**.
- **Rainfall** is most likely to occur in **summer** and Spring in Seoul.





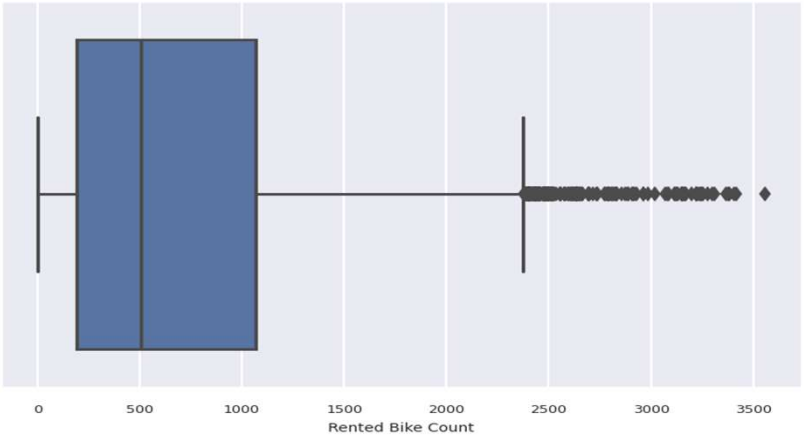
Correlation map

- Heat map shows slightly positive relation of Rented bike count with **Hour, Temperature, Dew point Temperature, Solar Radiation.**
- Bike sharing count is negatively co-related to **Humidity, Snowfall, Rainfall.**
- Temperature and Dew point temperature are positively co-related.



Dealing with Outliers

- Boxplot of Rented bike count depicts the presence of a high range of outliers.



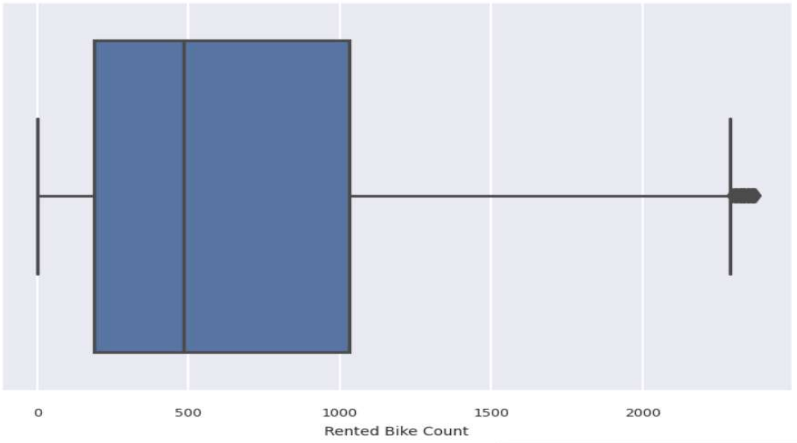
Using IQR process to remove outliers

Rented bike column indicating negligible or 0 outlier values.

Q1 = df.quantile(0.25)
Q3 = df.quantile(0.75)
IQR(Inter-quartile range) = Q3 - Q1

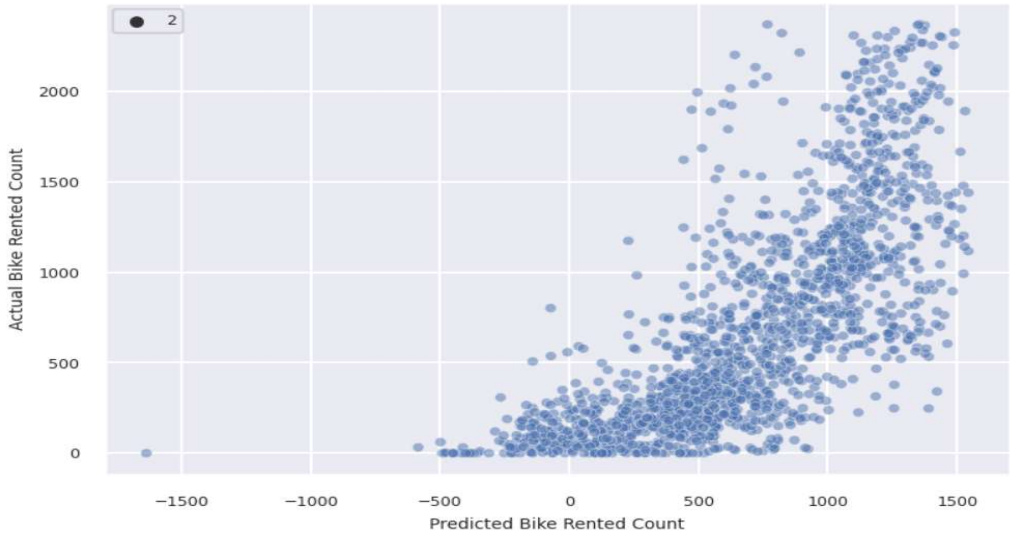
Lower range = Q1 – 1.5*IQR

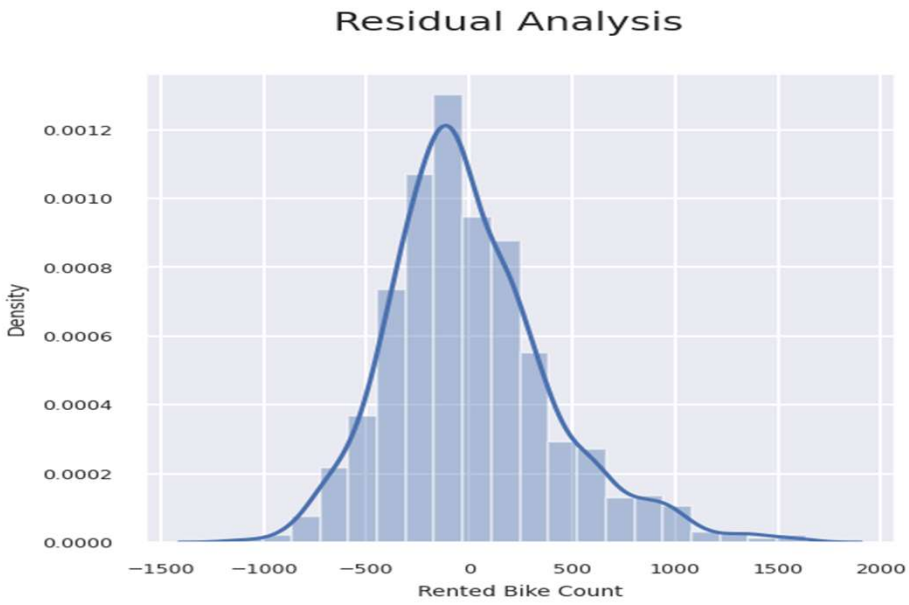
Higher range = Q2 + 1.5*IQR



Linear Regression Analysis

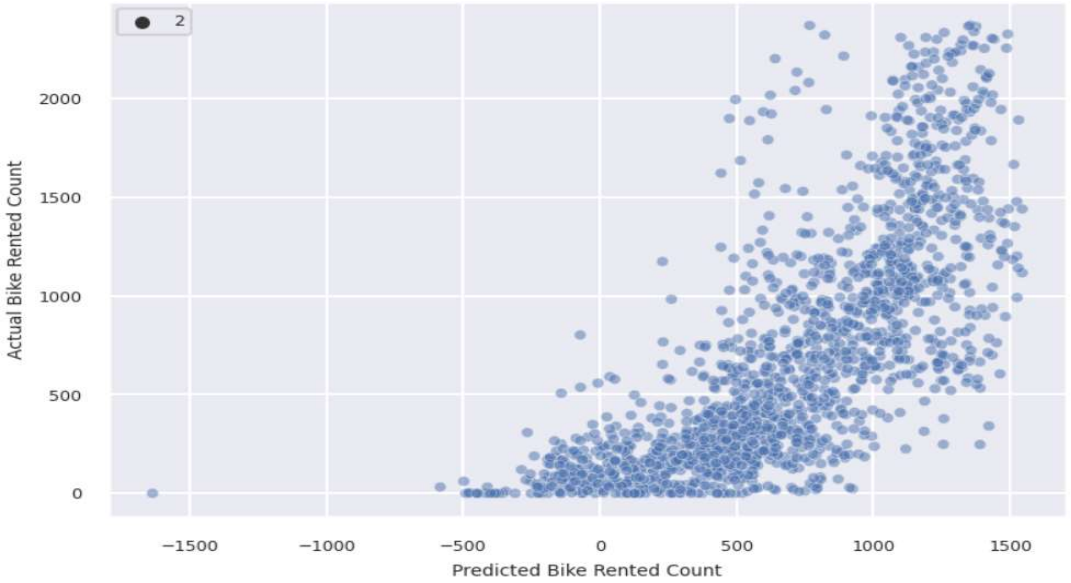
- It’s evident that a simple linear model is not giving us much accuracy.
- R-score is **0.55** and Adj R-score is **0.54.** MSE : **155676.03** MAE : **301.067** RMSE : **394.558**
- Linear Regression model residual values ranging from -1500 to 2000.

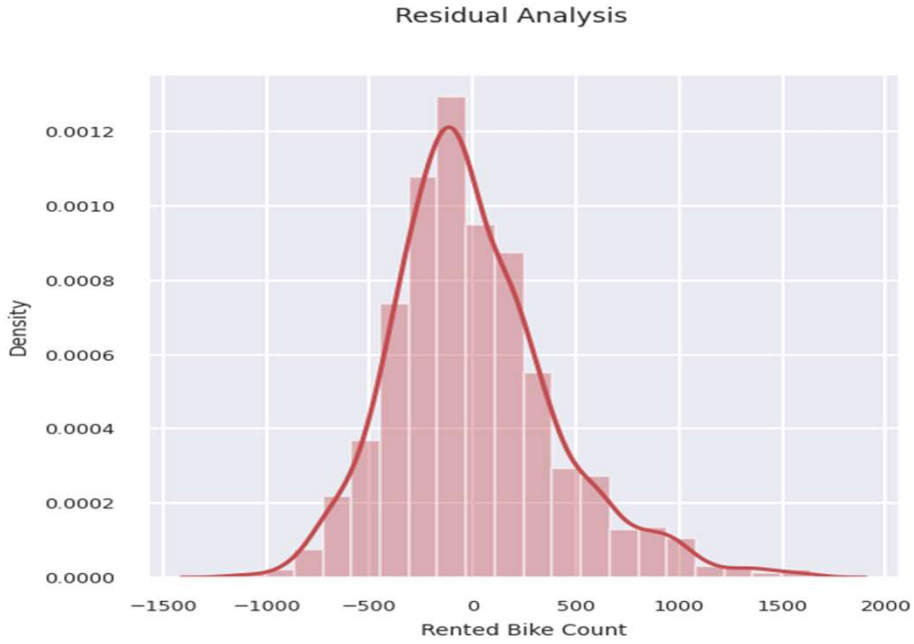




Lasso Regression Analysis

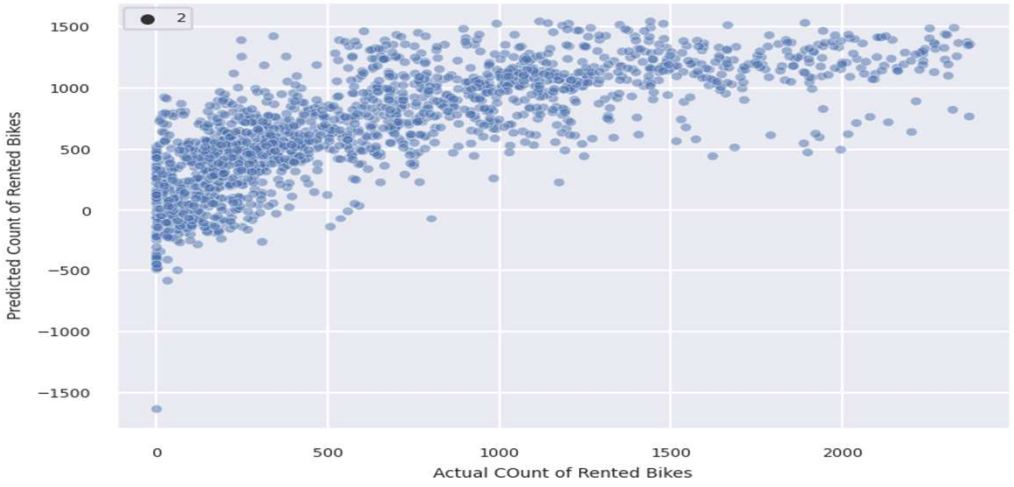
- Using Lasso regression, the model accuracy has been further reduced slightly.
- R-Score is **0.55** and Adj R-score is **0.54** MSE : **155673.688** MAE : **301.06** RMSE : **394.555**
- The Residual plot more or less remains the Same even with the Lasso Regression.

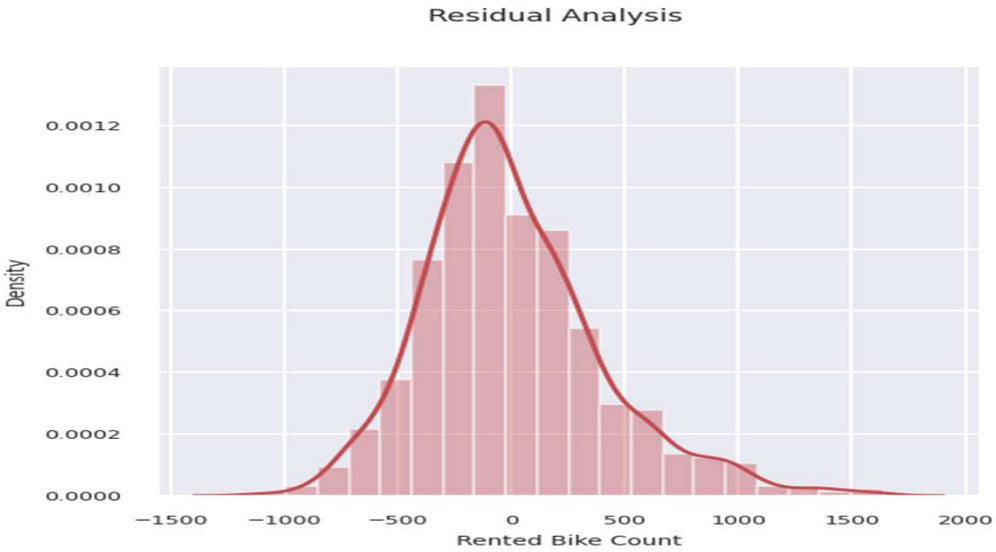




Hyperparameter Tuning Lasso Regression

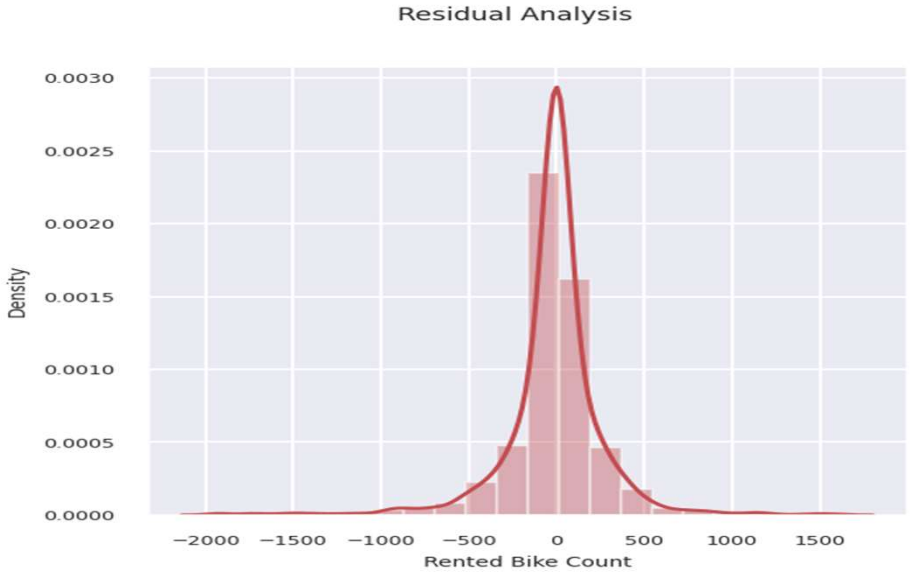
- Accuracy using Hyperparameter as well is low.
- R-Score **0.55** and Adjusted R-score **0.54** MSE : **155675.8669** MAE : **301.0676** RMSE : **394.5578**
- Residual plot remains the same using Hyperparameters as well.
- Possible **reasons** of such **low accuracy** using the Linear model is **low linear relationship**, **low correlation** between dependent and independent variables.





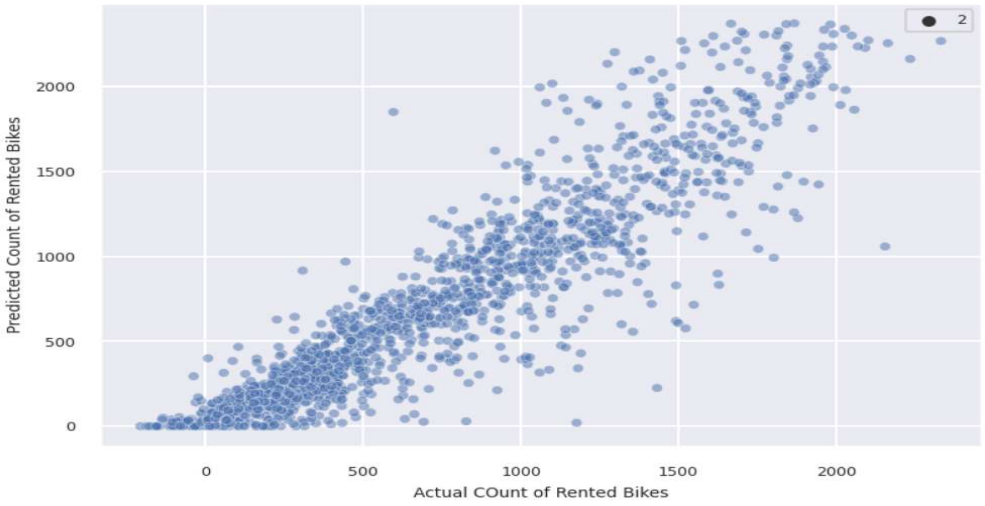
Decision Tree Analysis

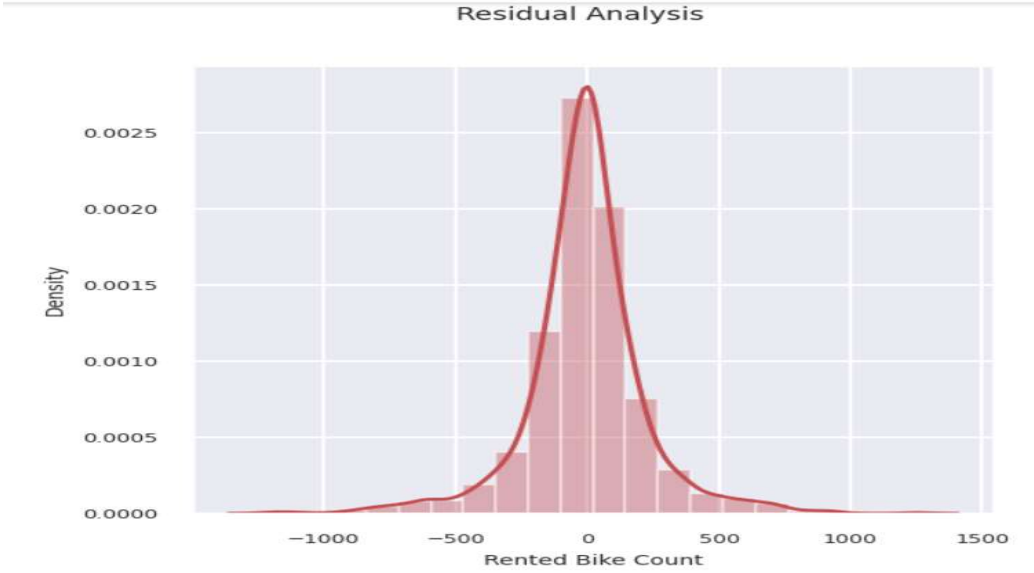
- Decision tree improves the accuracy significantly on the Test set. its evident from the below plot the Predicted and actual values are much closer compared to Linear Model
- R-score of **0.754** and Adjusted R-score of **0.751**, MSE : **83584.7431**, MAE : **164.4561**, RMSE : **289.1102**
- Residual values are reduced remarkably for the Decision tree. The KDE plot is much leaner and most of the Residual values are closer to zero.



Cat Boost Analysis

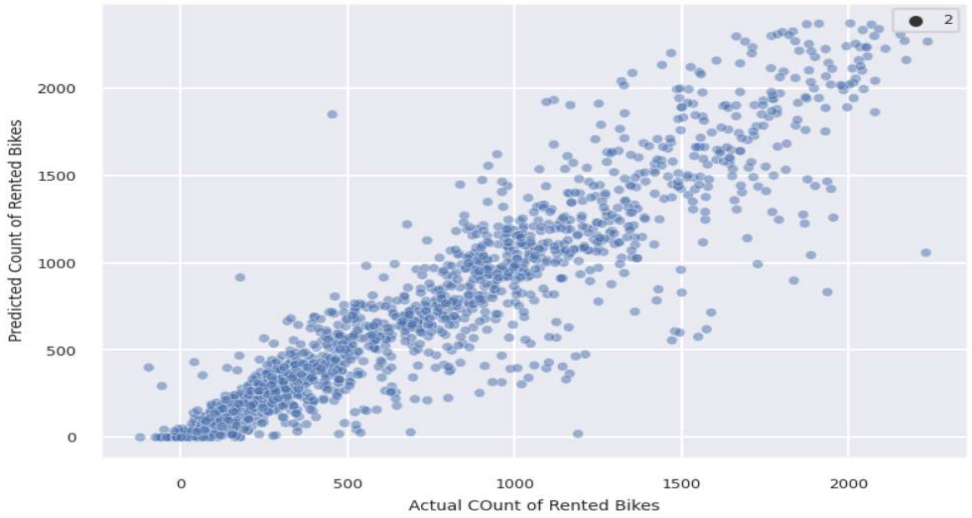
- R-score of **0.853** and Adjusted R-score of **0.852**, MSE : **50554.194**, MAE : **149.230**, RMSE : **224.842**
- Residual values are condensed remarkably towards zero for the Cat Boost. The KDE plot is much leaner and most of the Residual values are closer to zero.

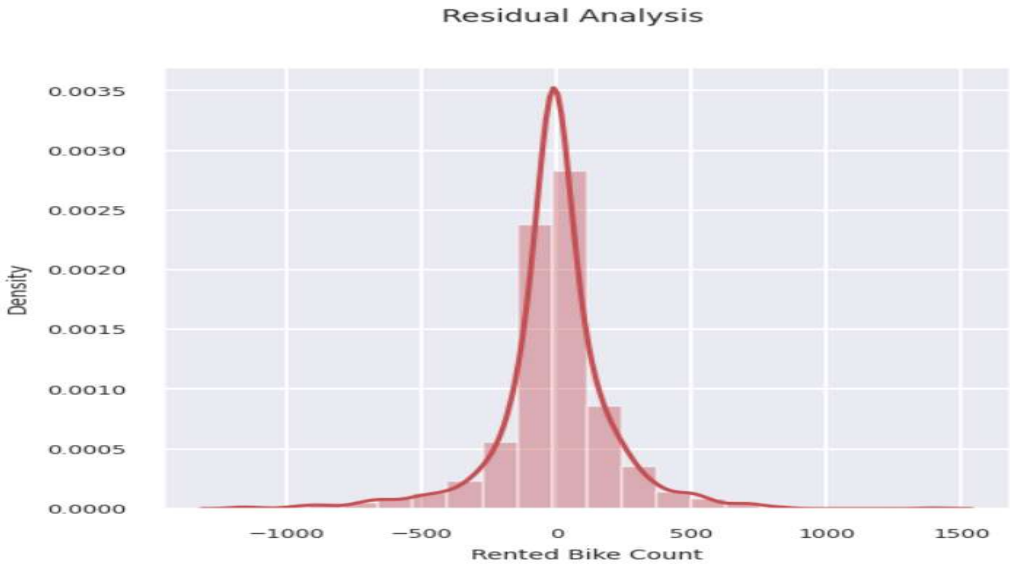




XG Boost Analysis

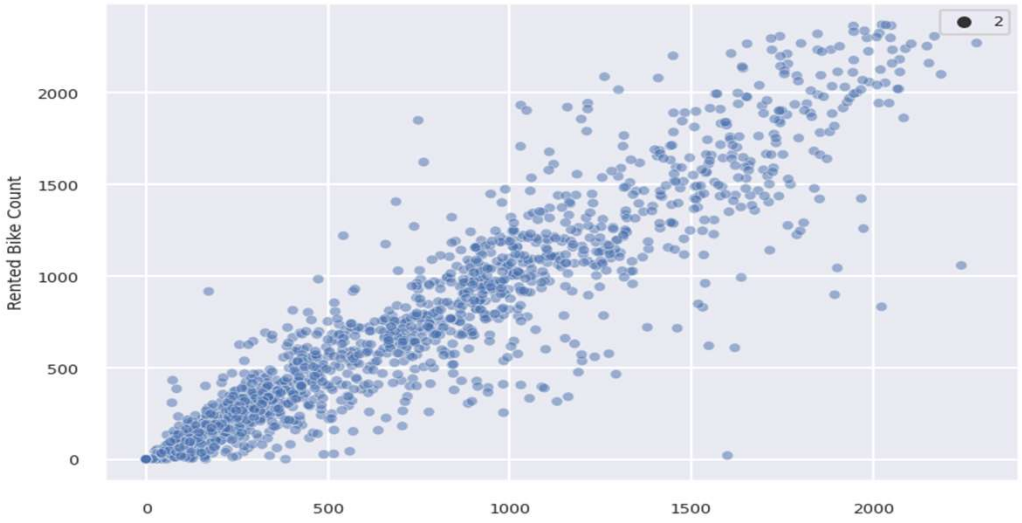
- R-score of **0.872** and Adjusted R-score of **0.871**, MSE : **44091.547** , MAE : **131.533**, RMSE : **209.979**
- Residual values are condensed remarkably towards zero for the XG Boost. The KDE plot is much leaner and most of the Residual values are closer to zero.

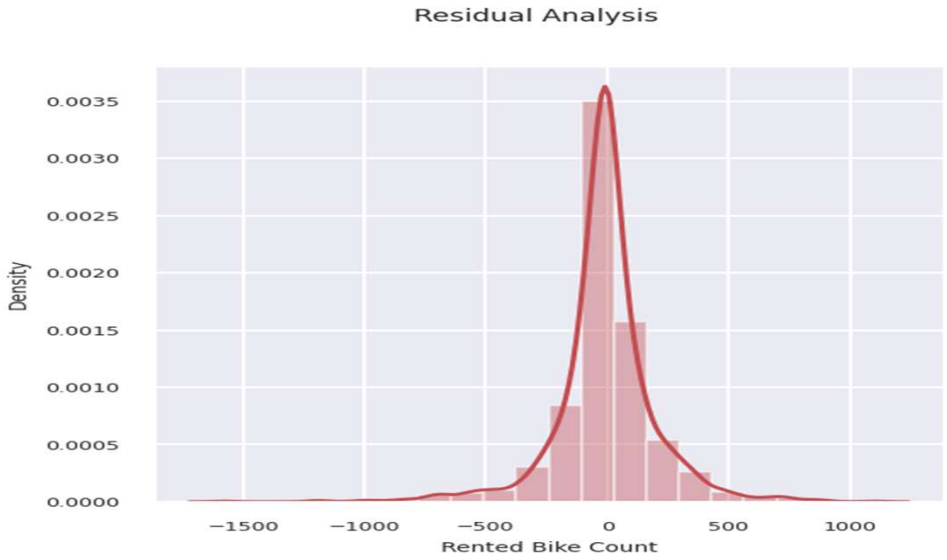




Random Forest Analysis

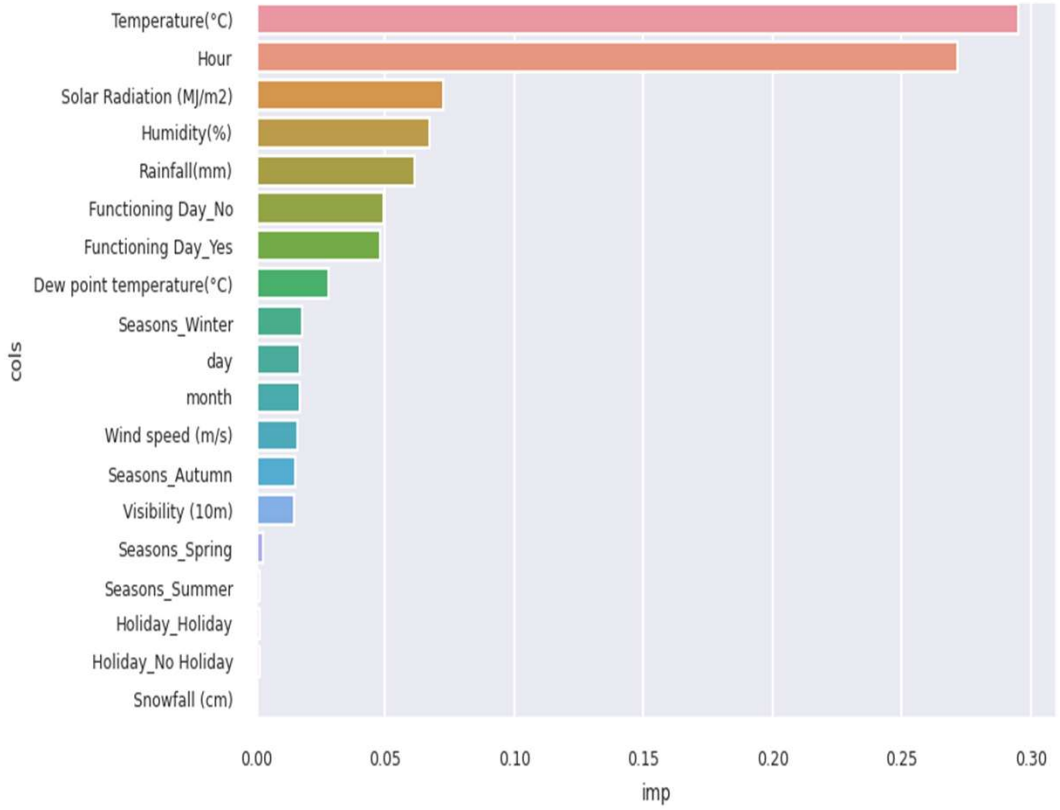
- R-score of **0.876** and an Adjusted R-score of **0.875**. MSE : **42747.1446** MAE : **128.21** RMSE : **206.753**
- Predicted and actual values of the test set follow an ideal line relation here.
- The Residual error is reduced significantly using Random forest. The majority of the Residual error are condensed around zero.





Feature Importance

- The adjacent Graph shows the importance of the features on our Rented bike count.
- **Temperature** and **Hour** of the day is a major factors **driving** the **demand for bikes**.
- **Solar Radiation, Humidity, Rainfall**, where its working day or not are other variables driving the demand for bikes.



Prediction Summary

- Predictions of the Linear Model are low due to the weak linear relationship between the dependent and independent variables.
- The nonlinear model appears to give better accuracy. Best predictions are obtained from Random forests with a test data accuracy of **0.875**.

SL NO	MODEL_NAME	MSE	RMSE	MAE	R^2	Adjusted R^2
1	Linear Regression	155676.033	394.558	301.067	0.550	0.545
2	Lasso Regression	155673.688	394.555	301.060	0.550	0.545
3	Gride Search CV	155675.866	394.557	301.067	0.550	0.545
4	Decision Tree	83584.743	289.110	164.456	0.754	0.751
5	CAT BOOST	50554.194	224.842	149.230	0.853	0.852
6	Default_XG boost	44091.547	209.979	131.553	0.872	0.871
7	Random Forest	42747.144	206.753	128.211	0.876	0.875

Conclusion

- Most numbers of Bikes were rented in **Summer**, followed by **Autumn**, **Spring**, and **Winter**. **May-July** is the peak Bike renting Season, and **Dec-Feb** is the least preferred month for bike renting.
- **Majority** of the **client** in the bike rental sector belongs to the **Working class**. This is evident from EDA analysis where bike demand is more on weekdays, working days in Seoul.
- **Temperature** of **20-30 Degrees**, **evening time 4 pm- 8 pm, Humidity** between **40%-60%** are the most favorable parameters where the Bike demand is at its peak.
- **Temperature, Hour** of the day, **Solar radiation**, and **Humidity** are major driving factors for the Bike rent demand.
- Feature and Labels had a weak linear relationship, hence the prediction from the linear model was very low. Best predictions are obtained with a **Random forest** model with an accuracy of **0.875**

Q & A