Python for data analysis

FINAL PROJECT :
SEOUL BIKE SHARING DEMAND DATASET

https://archive.ics.uci.edu/ml/datasets/Seoul+Bike+Sharing+Demand

Léo DUJOURD'HUI
William GAINNIER
Mélanie GAMBIEZ

Data Exploration

COLUMNS NAME AND TYPES

```
Date                          object
Rented Bike Count              int64
Hour                           int64
Temperature(°C)              float64
Humidity(%)                    int64
Wind speed (m/s)             float64
Visibility (10m)               int64
Dew point temperature(°C)    float64
Solar Radiation (MJ/m2)      float64
Rainfall(mm)                 float64
Snowfall (cm)                float64
Seasons                       object
Holiday                       object
Functioning Day               object
dtype: object
```

NON NUMERIC DATA

Date (object) : each day date from 31/12/2017 to 30/11/2018

Seasons (object) : Winter, Spring, Summer, Autumn

Holiday (object) : No Holiday, Holiday

Functioning Day (object) : No, Yes

THERE ARE 8760 ROWS AND 14 COLUMNS

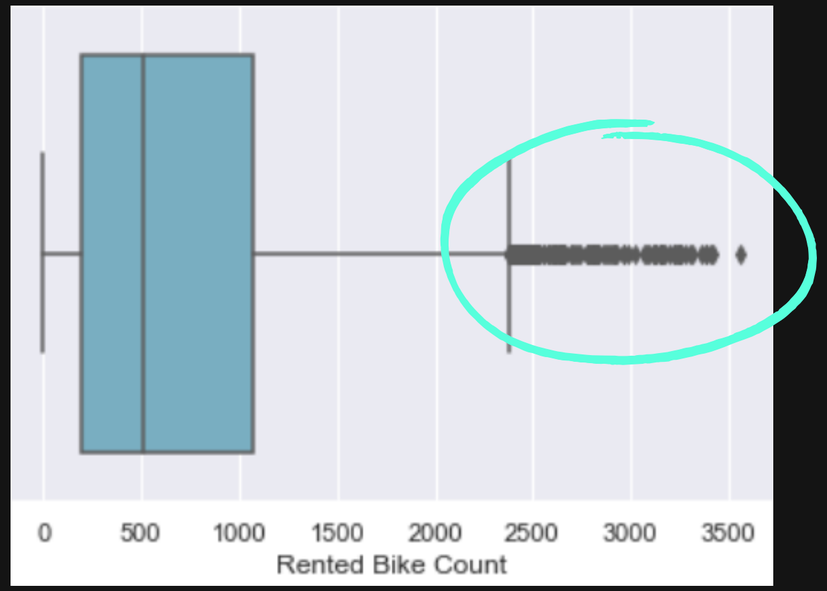
Data Cleaning

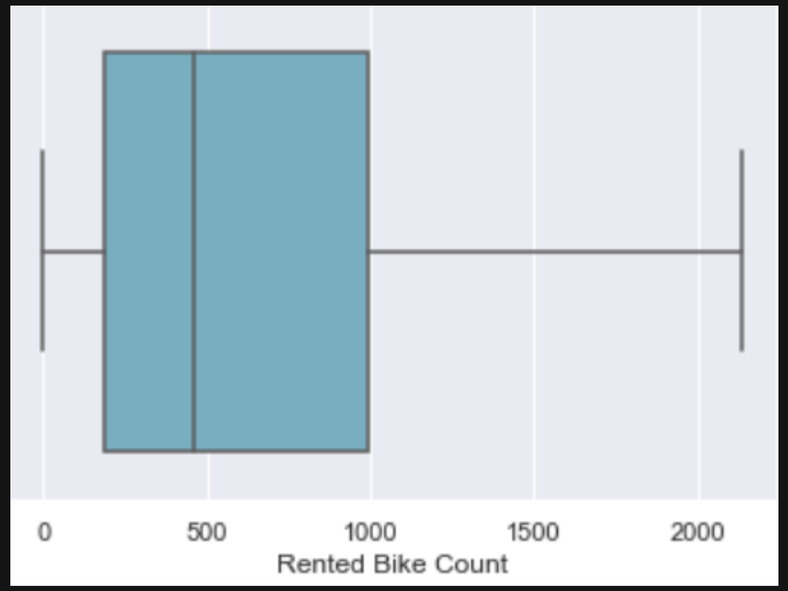
NO NULL VALUES

NO DUPLICATED VALUES

SOME EXTREME VALUES

```
quant96  = df["Rented Bike Count"].quantile(0.96)
df = df[(df["Rented Bike Count"] < quant96)]
```





Data Preparation

COLUMNS CREATION

Year : 2017
2018

Months : 1 to 12

Day : Monday > 0,
Tuesday > 1,
Wednesday > 2,
Thursday > 3,
Friday > 4,
Saturday > 5,
Sunday > 6

COLUMNS CHANGES : OBJECT TO NUMERICAL

Seasons : Winter > 0,
Spring > 1,
Summer > 2,
Autumn > 3

Holiday : No Holiday > 0,
Holiday > 1

Functioning Day : No > 0,
Yes > 1

COLUMNS DELETION

Date_Format (*temporary created for the Year and Months columns creation*)

Date

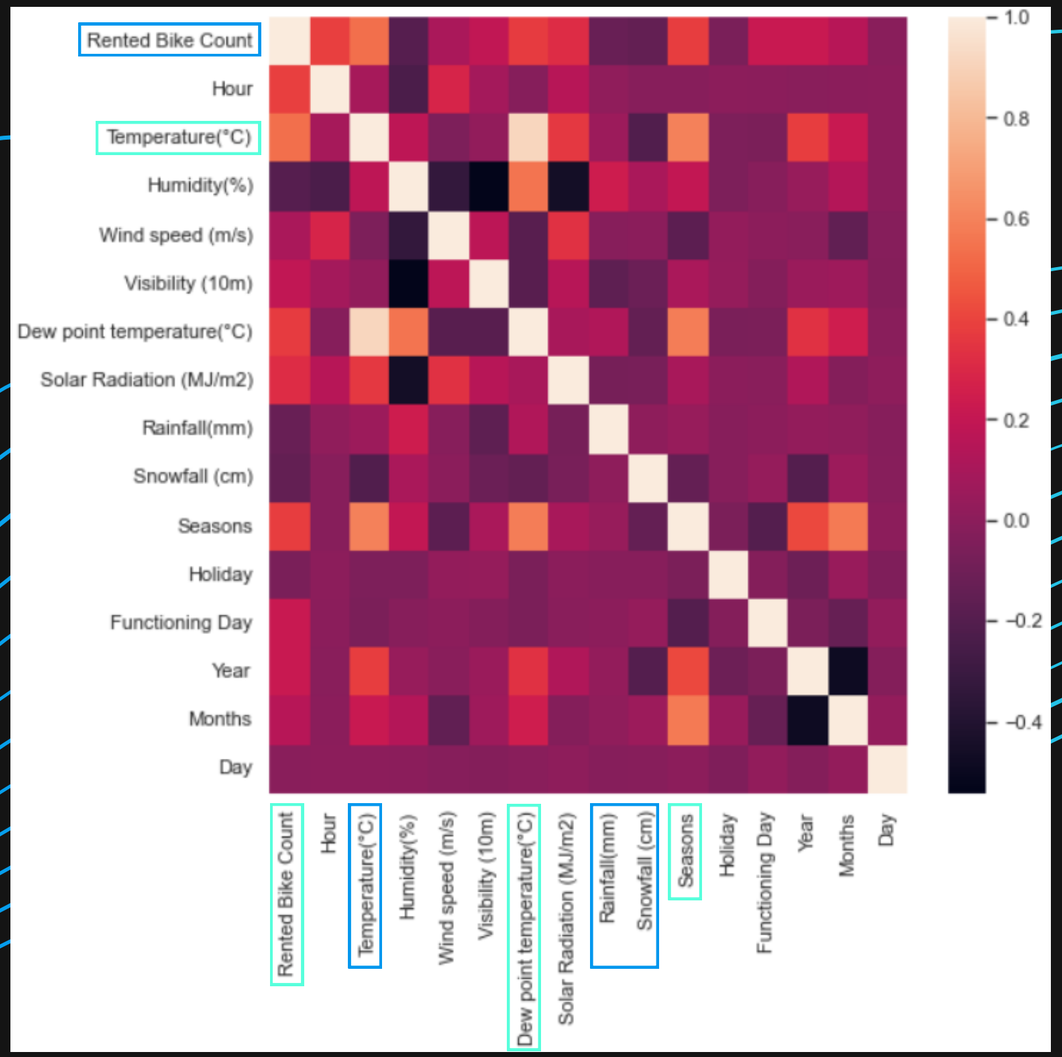
NEW COLUMN NAME AND TYPES

```
Rented Bike Count              int64
Hour                           int64
Temperature(°C)              float64
Humidity(%)                    int64
Wind speed (m/s)             float64
Visibility (10m)               int64
Dew point temperature(°C)    float64
Solar Radiation (MJ/m2)      float64
Rainfall(mm)                 float64
Snowfall (cm)                float64
Seasons                        int64
Holiday                        int64
Functioning Day                int64
Year                           int64
Months                         int64
Day                            int64
dtype: object
```

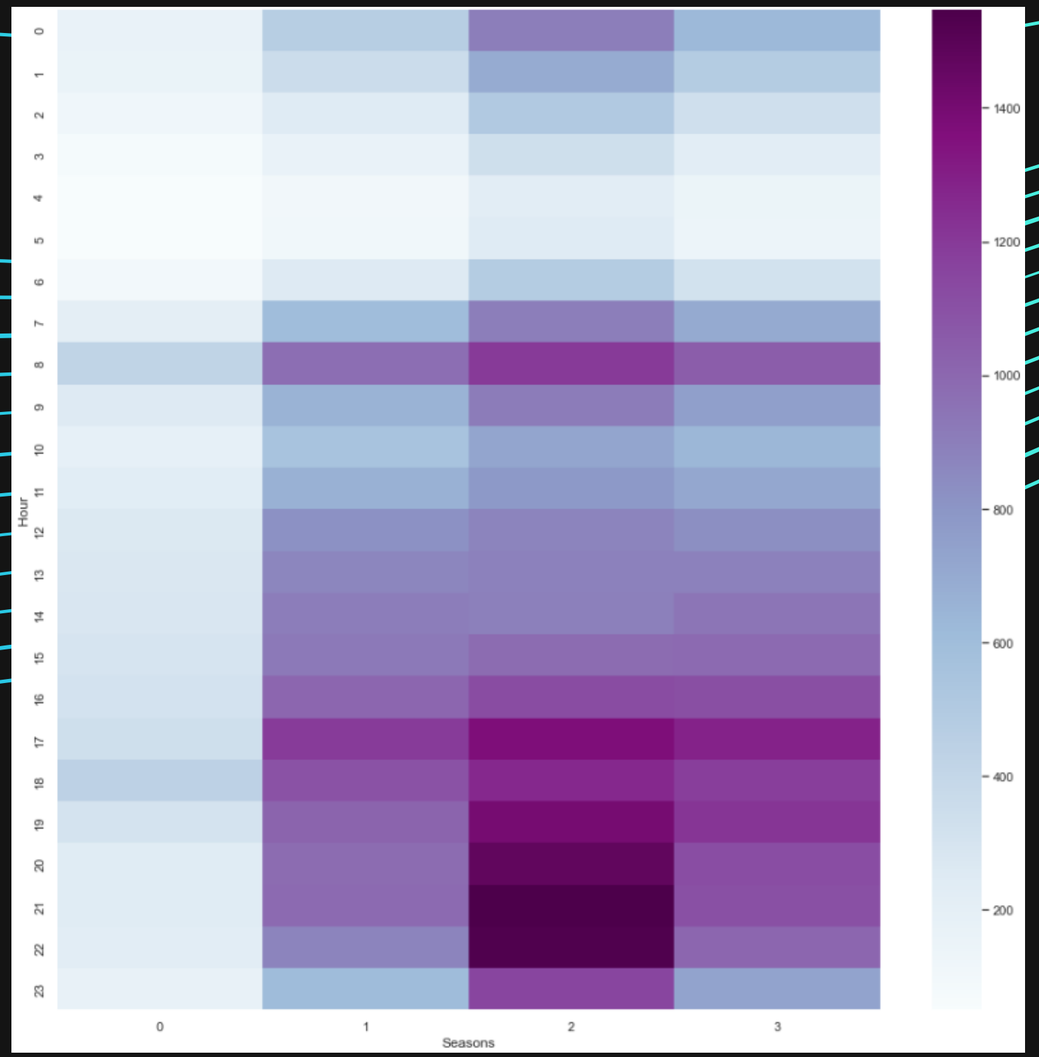
THERE ARE 8409 ROWS AND 16 COLUMNS

Data Analysis

CORRELATION BETWEEN DATASET COLUMNS

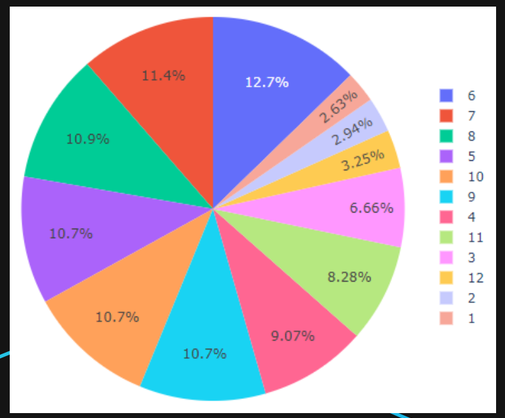


RENTED BIKE COUNT CONSIDERING HOUR AND SEASON



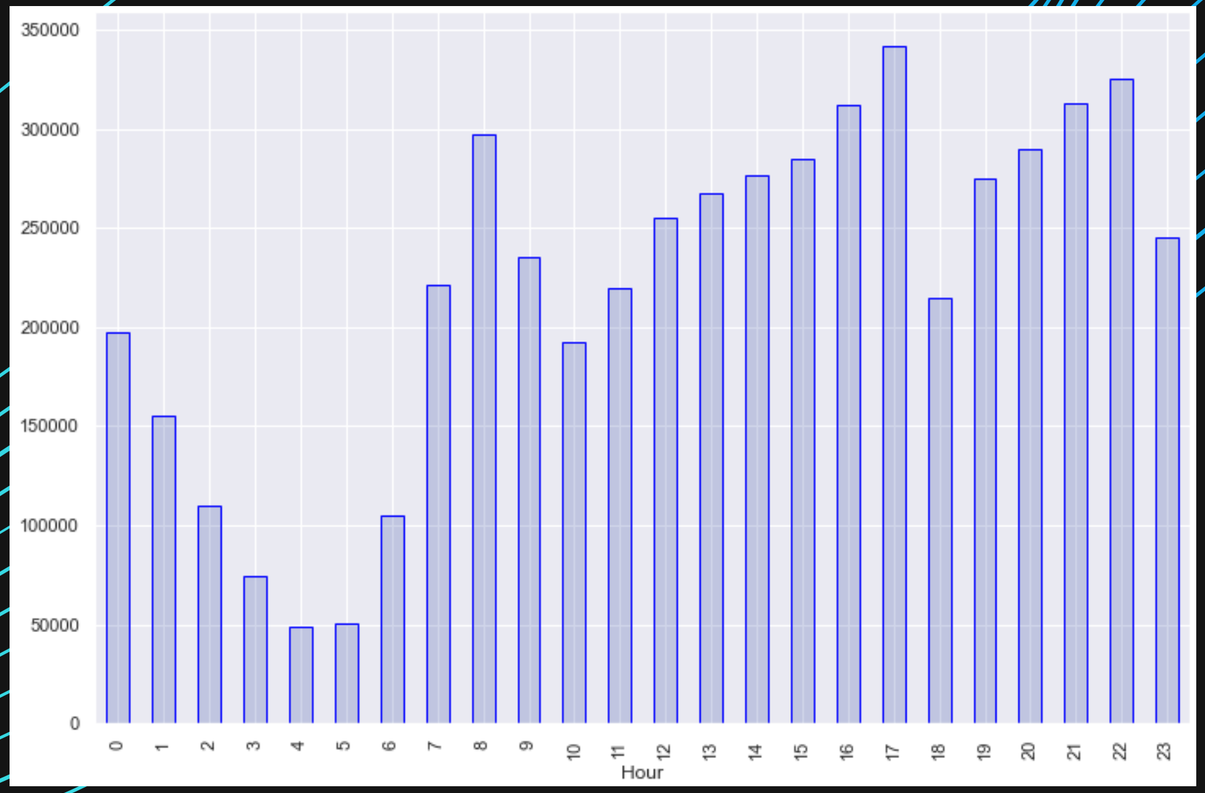
Data Analysis

% OF RENTED BIKE PER MONTH



RENTED BIKE COUNT PER HOUR AND SEASON

TOTAL RENTED BIKE COUNT PER HOUR



Data Modeling :

PREDICT THE NUMBER OF RENTED BIKES AT GIVEN HOUR

DATAFRAME DIVIDED INTO TARGET AND DATA SETS

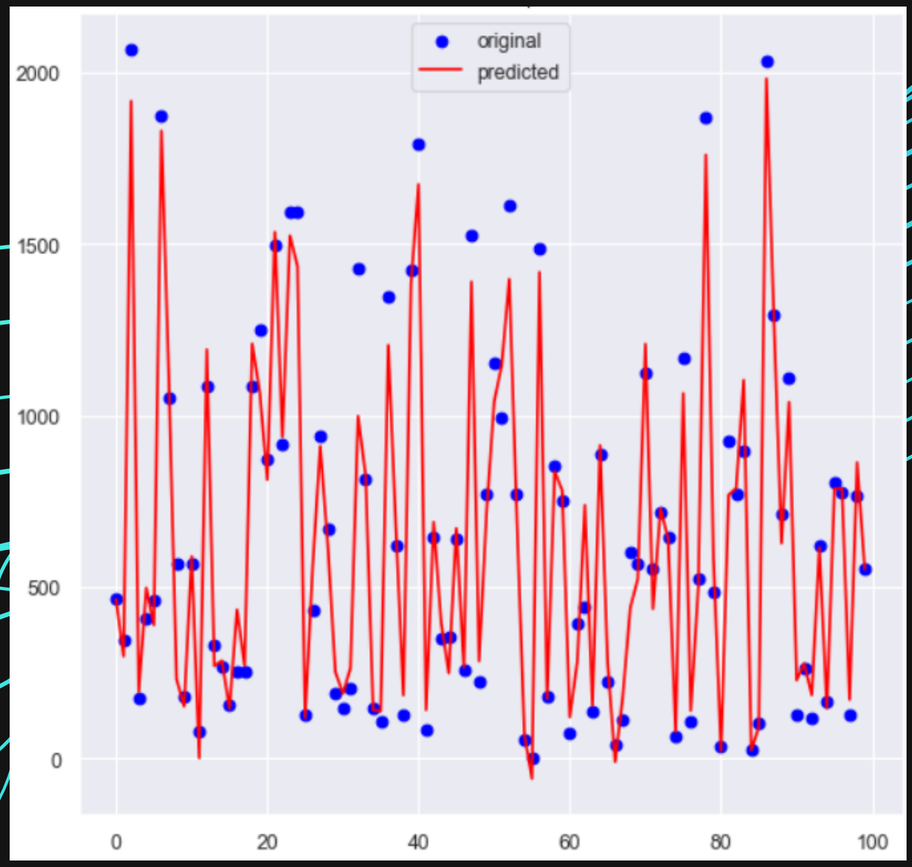
Target = Rented Bike Count

BEST MODEL

GradientBoostingRegressor
(with GridSearchCV)
> Mean error = 77.2523

TEST : On the 06/12/22 in Seoul, considering all conditions (9am, 1°C, ...), our model predicted the rent of 457 bikes.

RENTED BIKE NUMBER PREDICTION FOR 100 HOURS



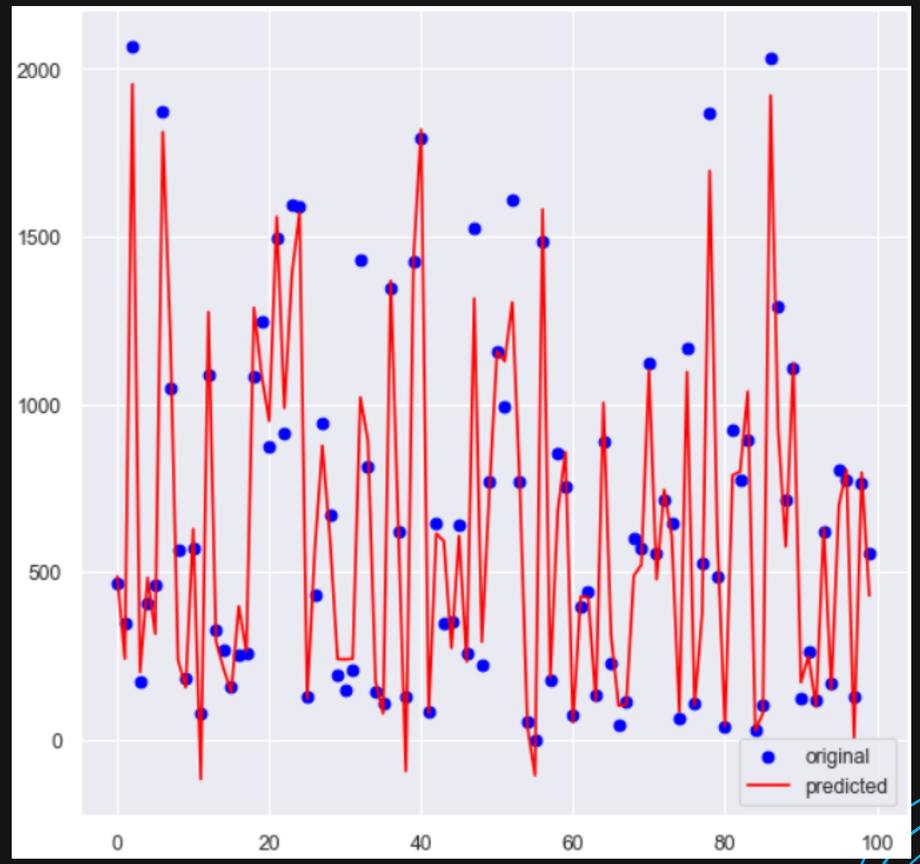
Data Modeling :

PREDICT THE NUMBER OF RENTED BIKES AT GIVEN HOUR

TESTED MODELS

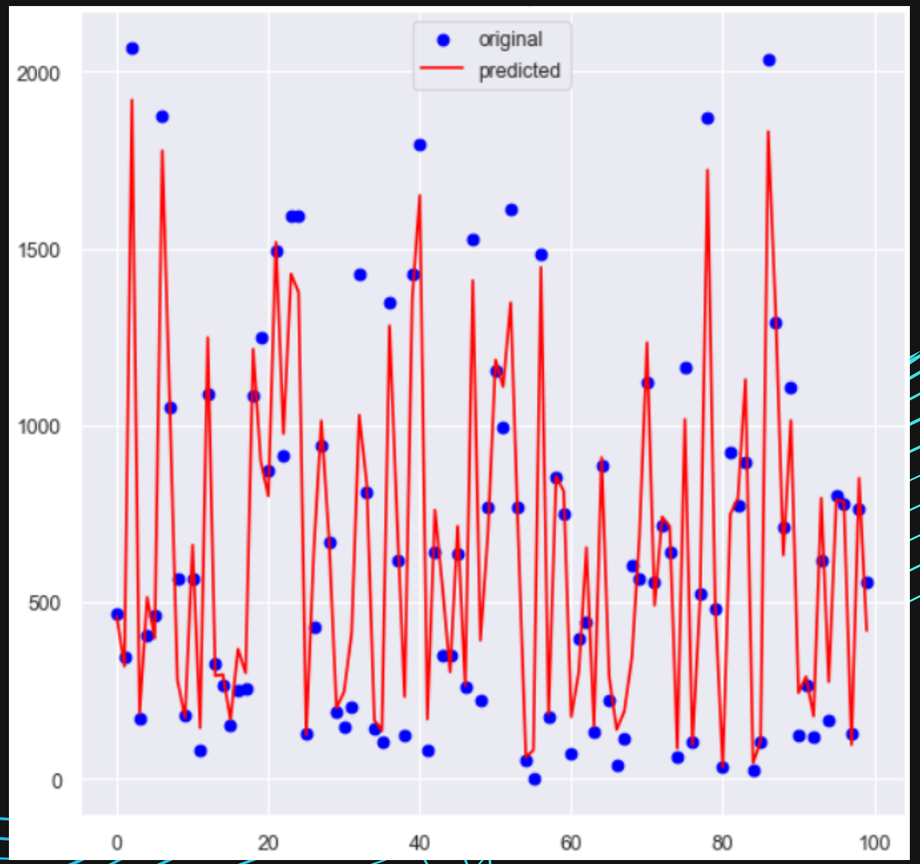
HistGradientBoostingRegressor (without GridSearchCV)

> Mean error = 88.6770



RandomForestRegressor

> Mean error = 97.6920



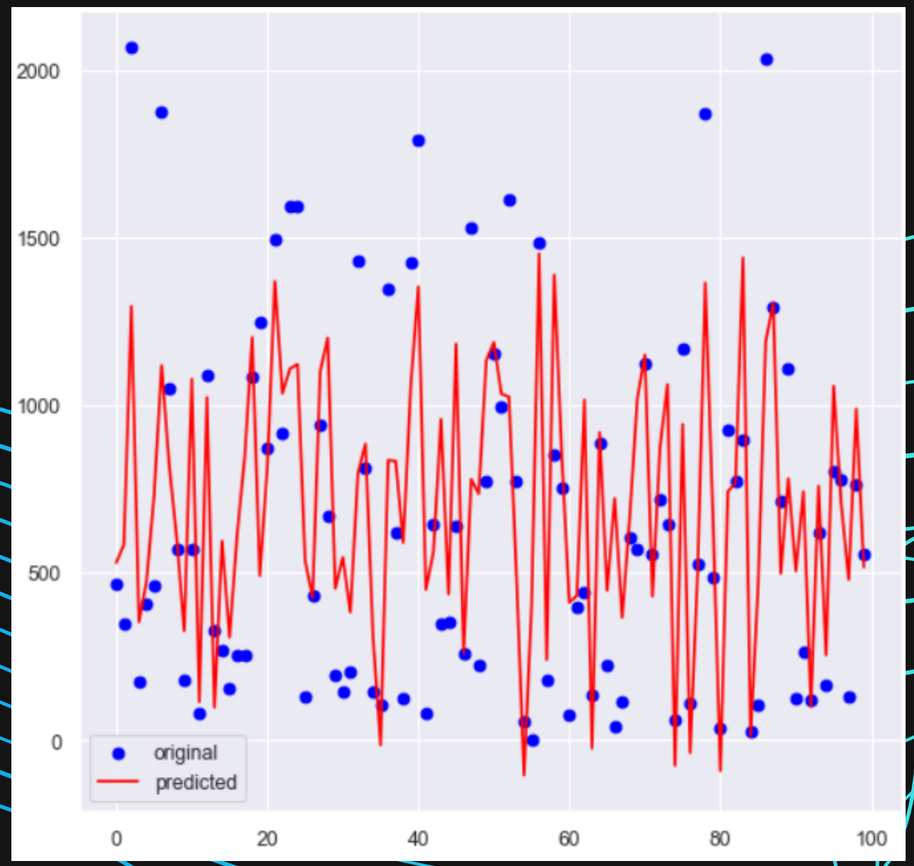
Data Modeling :

PREDICT THE NUMBER OF RENTED BIKES AT GIVEN HOUR

TESTED MODELS

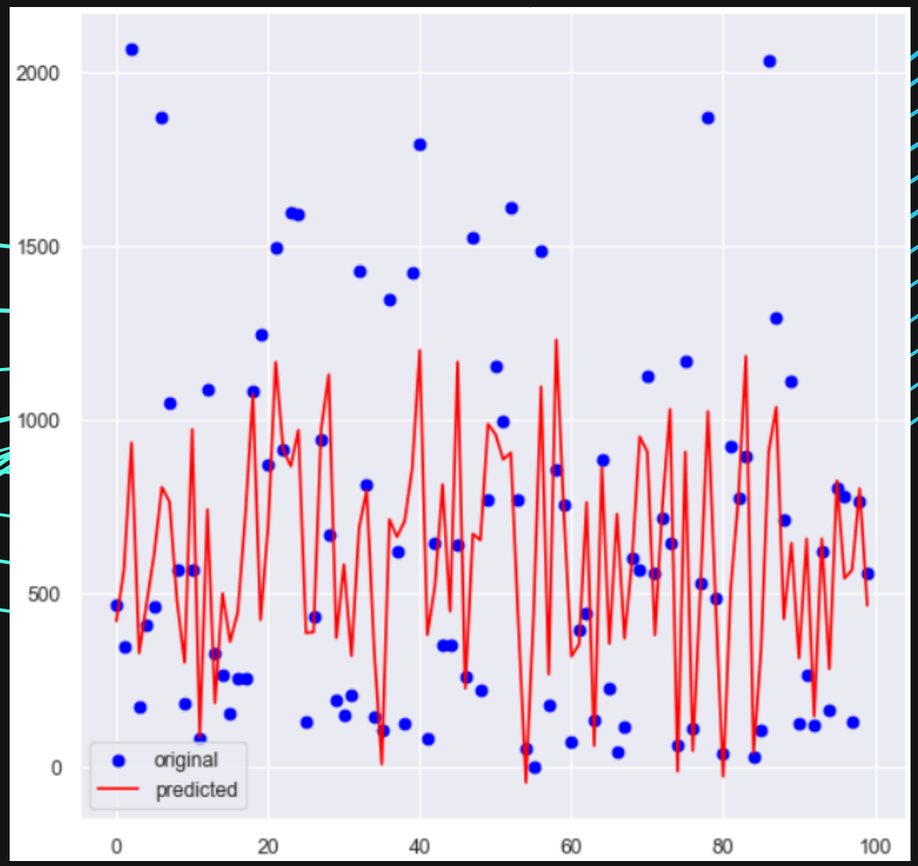
Bayesian Ridge Model

> Mean error = 270,1722



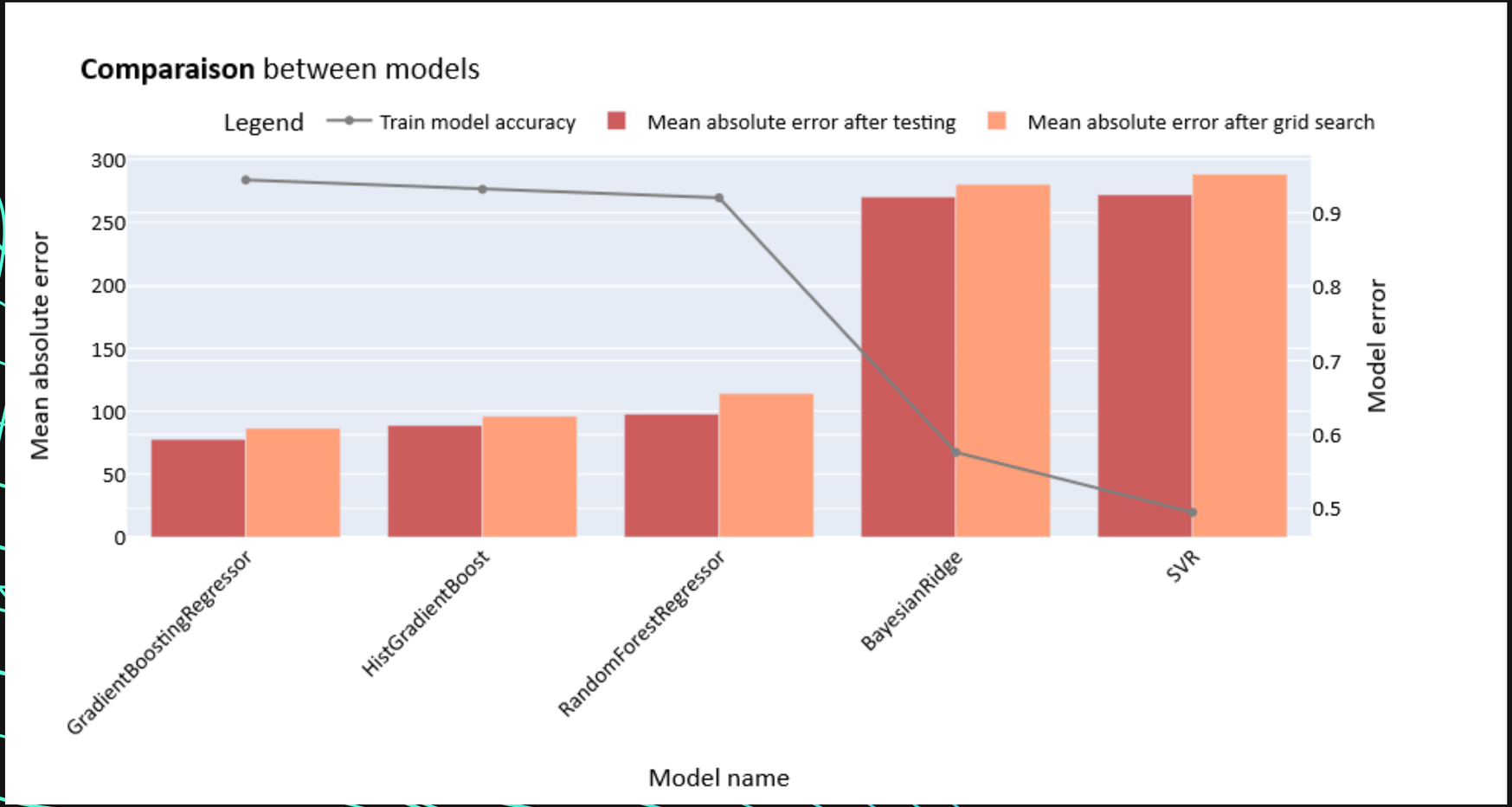
SVR Model

> Mean error = 271.8103



Data Modeling :

PREDICT THE NUMBER OF RENTED BIKES AT GIVEN HOUR



API Python

Model accuracy choose model

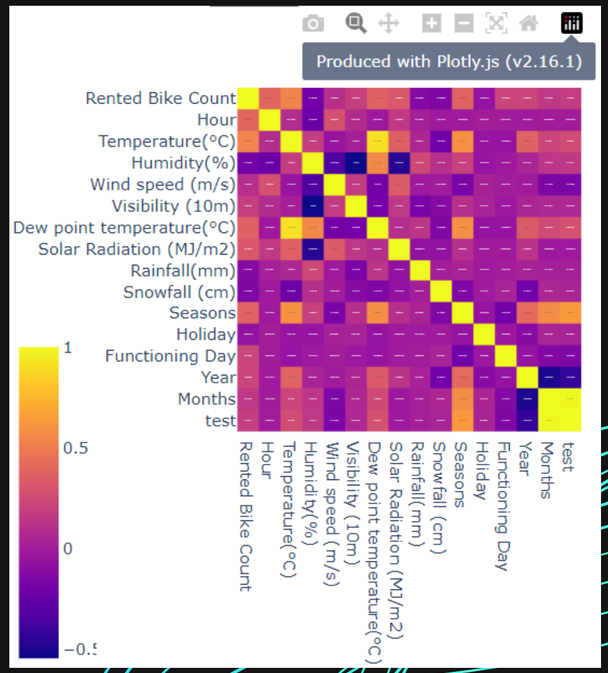
choose model
HistgradientBoosting
BayesianRidgeModel
DecisionTreeRegressorModel

Model predict on average with an error of : b'127.27075008106952' bikes

Explore the dataset throught graphics

Choose the graphic you want choose model

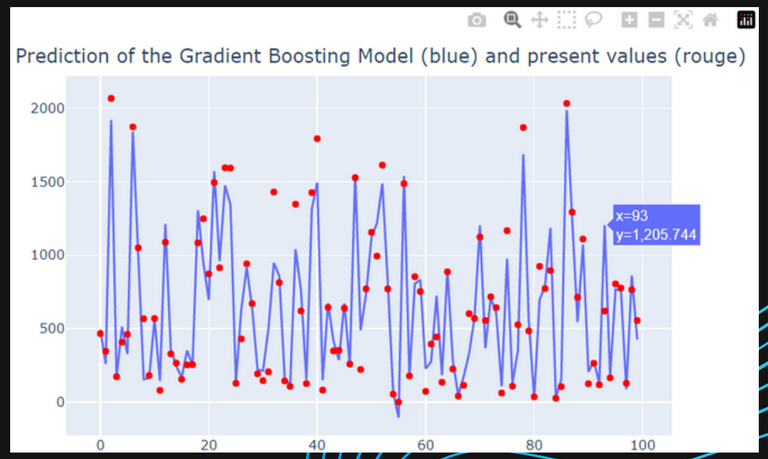
choose model
Meanrentedbike/months
mean rented bike for each hour 2017-2018
CorrelationMatrix



Models of Machine Learning

choose model

choose model
GradientBoosting
BayesianRidge
RandomForestRegressor



For more information on this dataset analysis, please check our GitHub : https://github.com/le-cmyk/Seoul_Bike