Bike sharing in Seoul

Nicolas BONNET

DIA3

Topic

- 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**.
- The dataset contains weather information (Temperature, Humidity, Windspeed, Visibility, Dewpoint, Solar radiation, Snowfall, Rainfall), the number of bikes rented per hour and date information.

Dataset composition

General composition

- Shape : 8760 rows & 14 columns
- Features : 10 float & 4 categorical
- No NaN
- No empty columns/rows

Attribute information

- Date : year-month-day
- Rented Bike count - Count of bikes rented at each hour
- Hour - Hour of the day
- Temperature-Temperature in Celsius
- Humidity - %
- Windspeed - m/s
- Visibility - 10m
- 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)

General idea of data

Thanks to the data.describe().T function, we can pin up a general idea of the dataset. Thus, we can scale each parameters by knowing the mean, min or max value.

We can also know if values are missing or wrong.

This is an important phase to prepare data in order to execute visualization and modeling

	count	mean	std	min	25%	50%	75%	max
Rented Bike Count	8760.0	704.602055	644.997468	0.0	191.00	504.50	1065.25	3556.00
Hour	8760.0	11.500000	6.922582	0.0	5.75	11.50	17.25	23.00
Temperature	8760.0	12.882922	11.944825	-17.8	3.50	13.70	22.50	39.40
Humidity(%)	8760.0	58.226256	20.362413	0.0	42.00	57.00	74.00	98.00
Wind speed (m/s)	8760.0	1.724909	1.036300	0.0	0.90	1.50	2.30	7.40
Visibility (10m)	8760.0	1436.825799	608.298712	27.0	940.00	1698.00	2000.00	2000.00
Dew point temperature	8760.0	4.073813	13.060369	-30.6	-4.70	5.10	14.80	27.20
Solar Radiation (MJ/m2)	8760.0	0.569111	0.868746	0.0	0.00	0.01	0.93	3.52
Rainfall(mm)	8760.0	0.148687	1.128193	0.0	0.00	0.00	0.00	35.00
Snowfall (cm)	8760.0	0.075068	0.436746	0.0	0.00	0.00	0.00	8.80

Data preparation

Verification of NaN values

```
#No nan nor null values
nulls = data.isnull().sum()
nans = data.isna().sum()

print(nulls[nulls>0])
print(nans[nans>0])
```

✓ 0.6s

We have to check if all parameters and information are coherent. In order to manipulate data the best way, we check the NaN (Not a Number) or Null values. Those are all the values that could be wrong in the dataset and that could distort our work.

We perform One Hot Encoding...

This helps us to manipulate data. It creates new distinct columns for each parameter that has less than 12 values. We can therefore compare data according to seasons or month for example.

```
#We perform one hot encoding for categorical features
distinct = data.nunique()
print(distinct)
cat_col = list(distinct[distinct<=12].index)
non_cat_feat = list(distinct[distinct>12].index)
data = pd.get_dummies(data, columns = cat_col, drop_first = True)
```

✓ 0.1s

```
Rented Bike Count            2166
Hour                           24
Temperature                   546
Humidity(%)                    90
Wind speed (m/s)               65
Visibility (10m)             1789
Dew point temperature         556
Solar Radiation (MJ/m2)       345
Rainfall(mm)                   61
Snowfall (cm)                  51
Seasons                         4
Holiday                         2
Functioning Day                 2
day                             7
month                          12
year                            2
dtype: int64
```

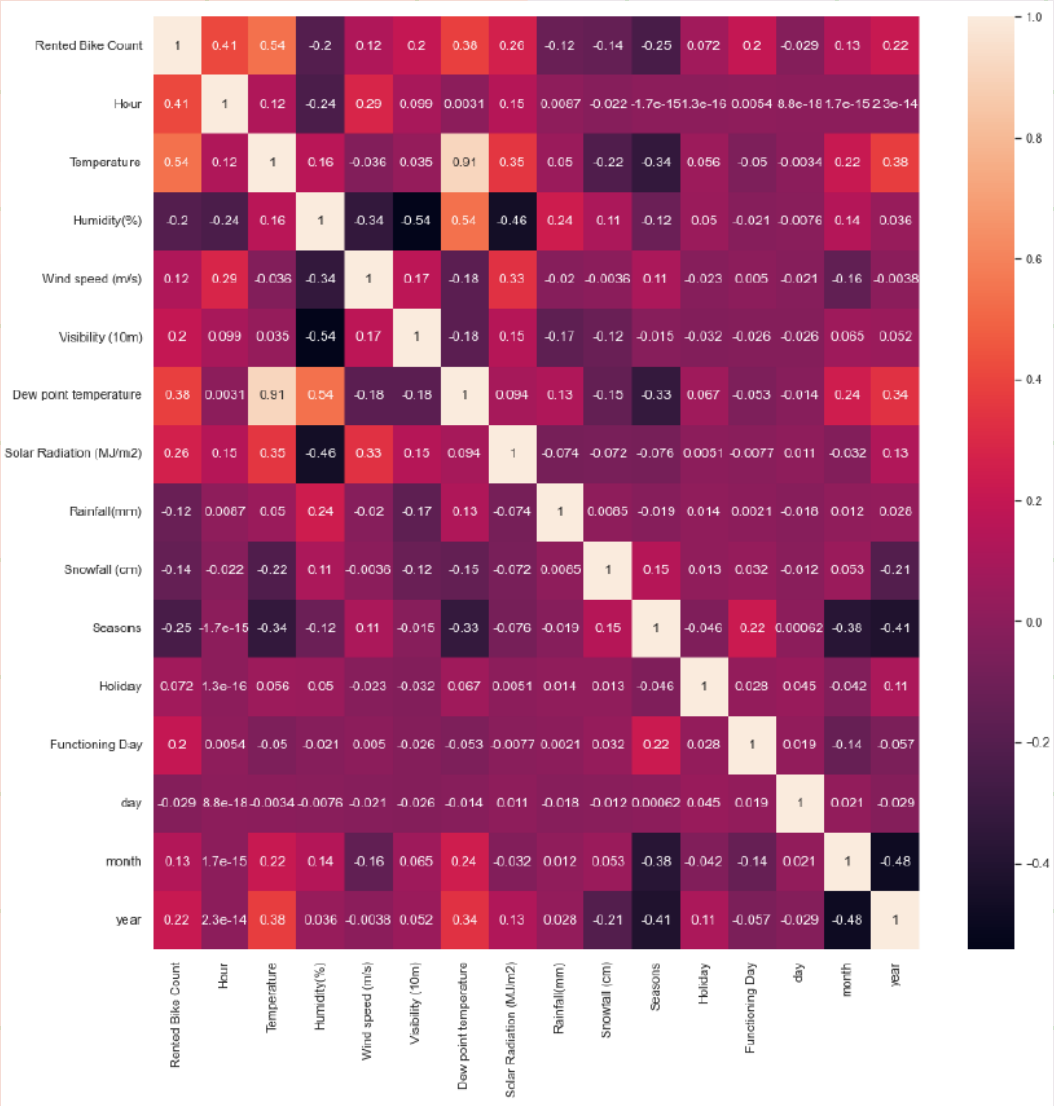
...to create more valuable parameters

THUS, WE HAVE CREATED SEVERAL PARAMETERS TO USE THEM IN OUR MODELS.

THIS PARAMETERS GIVE US A BETTER VISUALISATION AND EXPLOITATION OF THE DATASET. THIS IS DUE TO A BETTER SEPARATION OF THE INFORMATION THAT ARE IN THE DATASET.

FOR EXAMPLE, WE SEPARATED THE SEASONS OR THE WEEKDAY, WHICH IS MORE RELEVANT THAN A COMPARISON BY MONTH.

Data visualisation

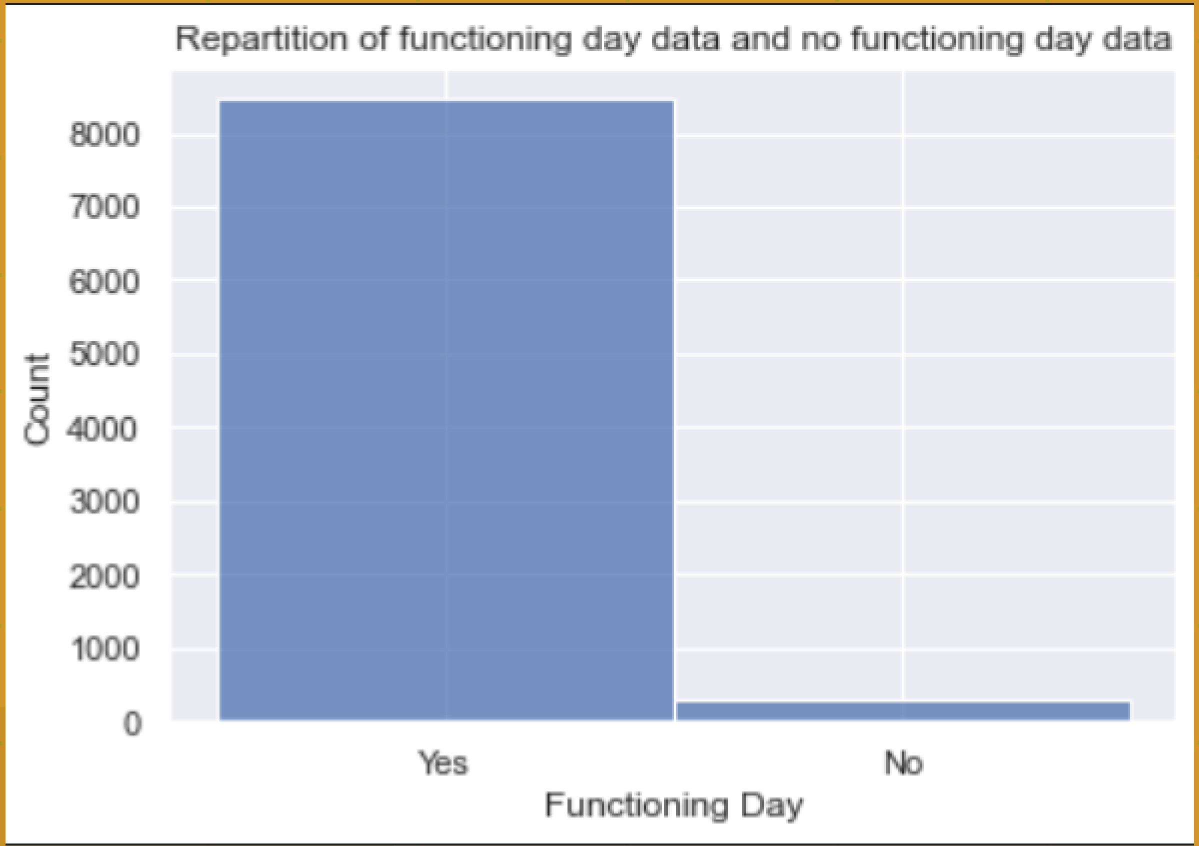


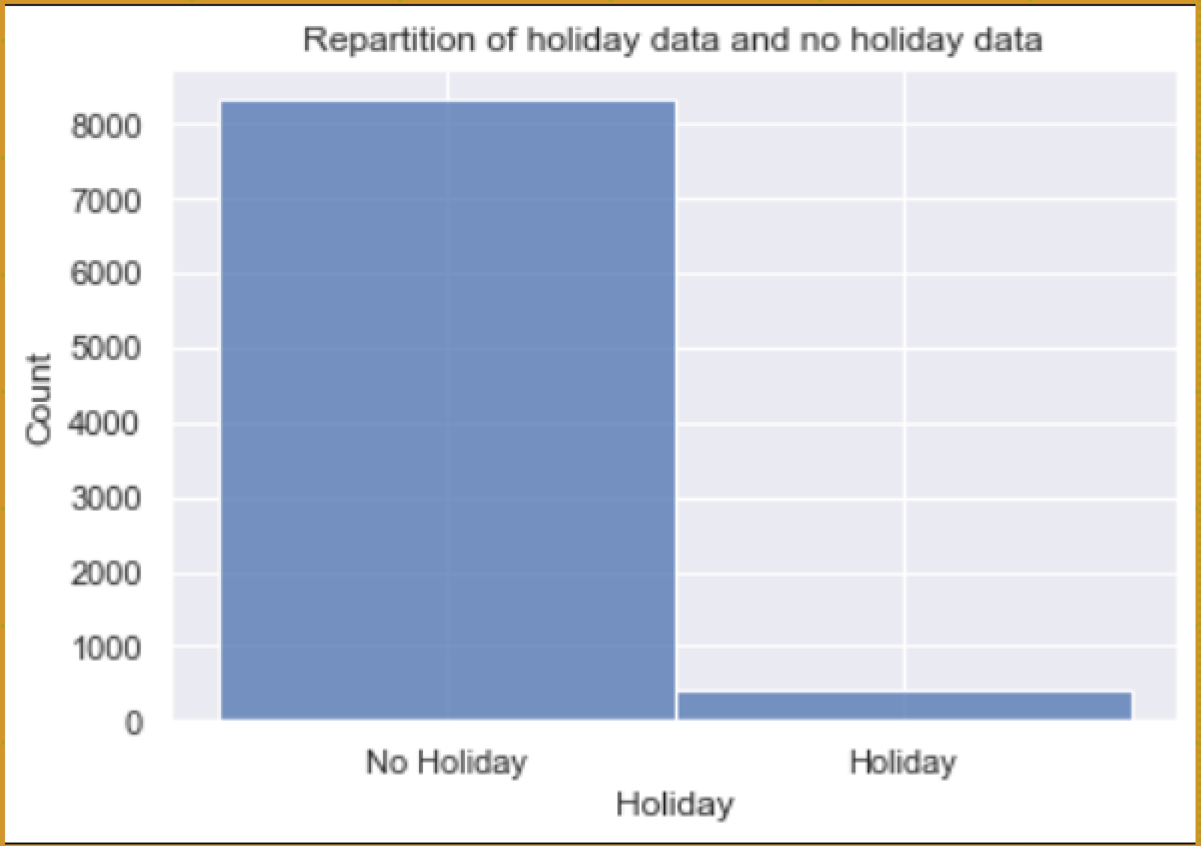
Correlation Matrix

The more the correlation number is close to 1, the more correlated is the Rented Bike Count to the Parameter.

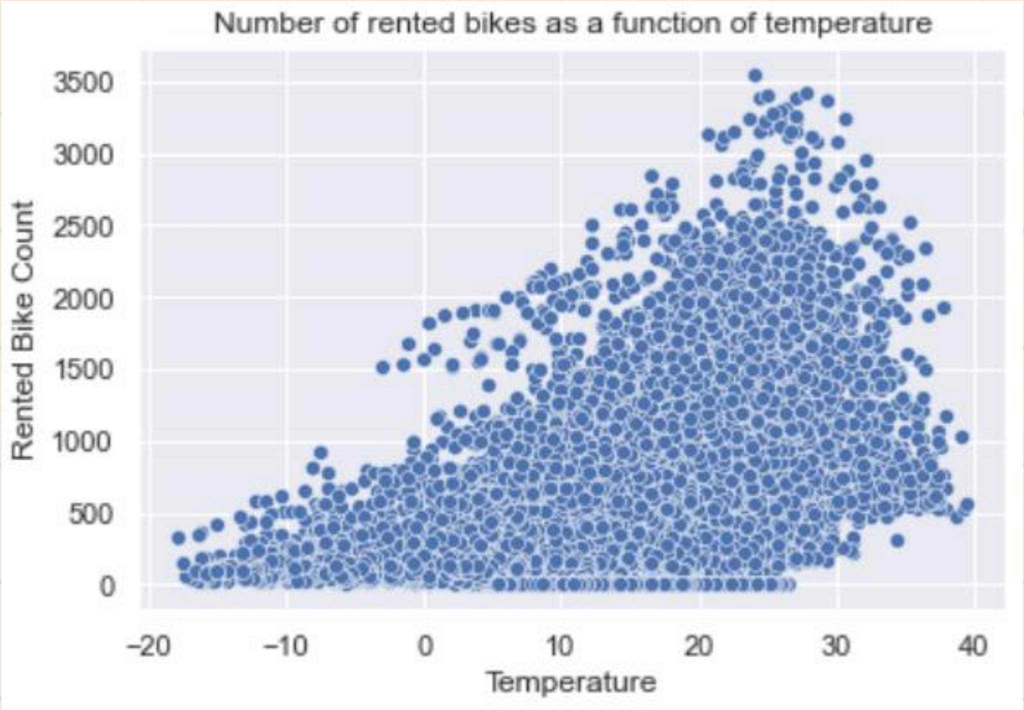
For example, here, we can see that the number of bikes rented is highly correlated to the temperature, the hour or the dew point temperature.

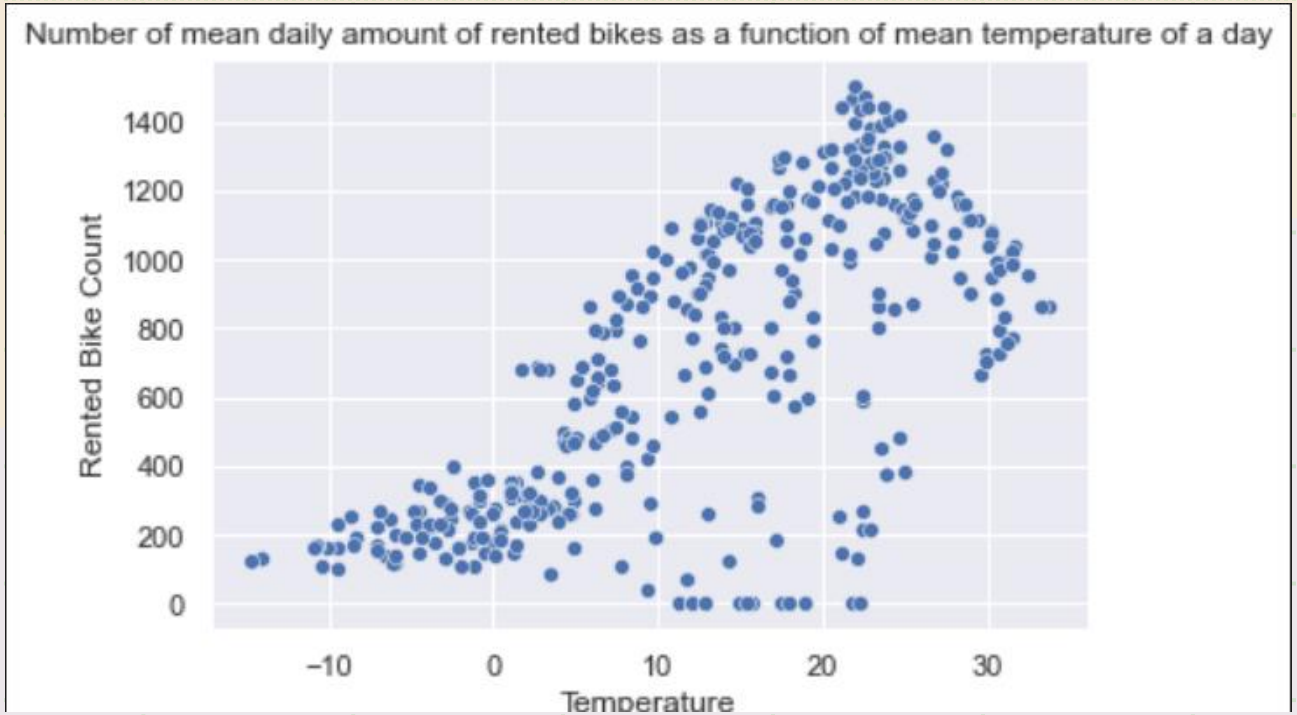
Repartition Functioning/Holiday Day data





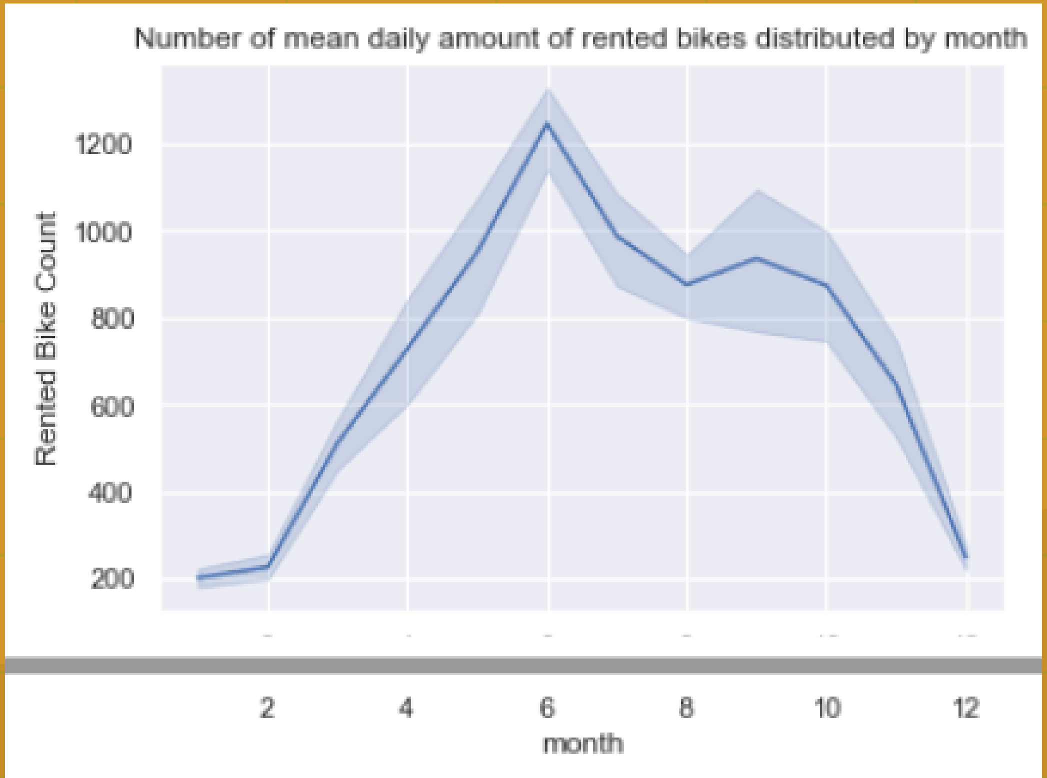
This shows us that we have to be careful because the model will learn more on the periods out of holidays and will be maybe less precise on vacations

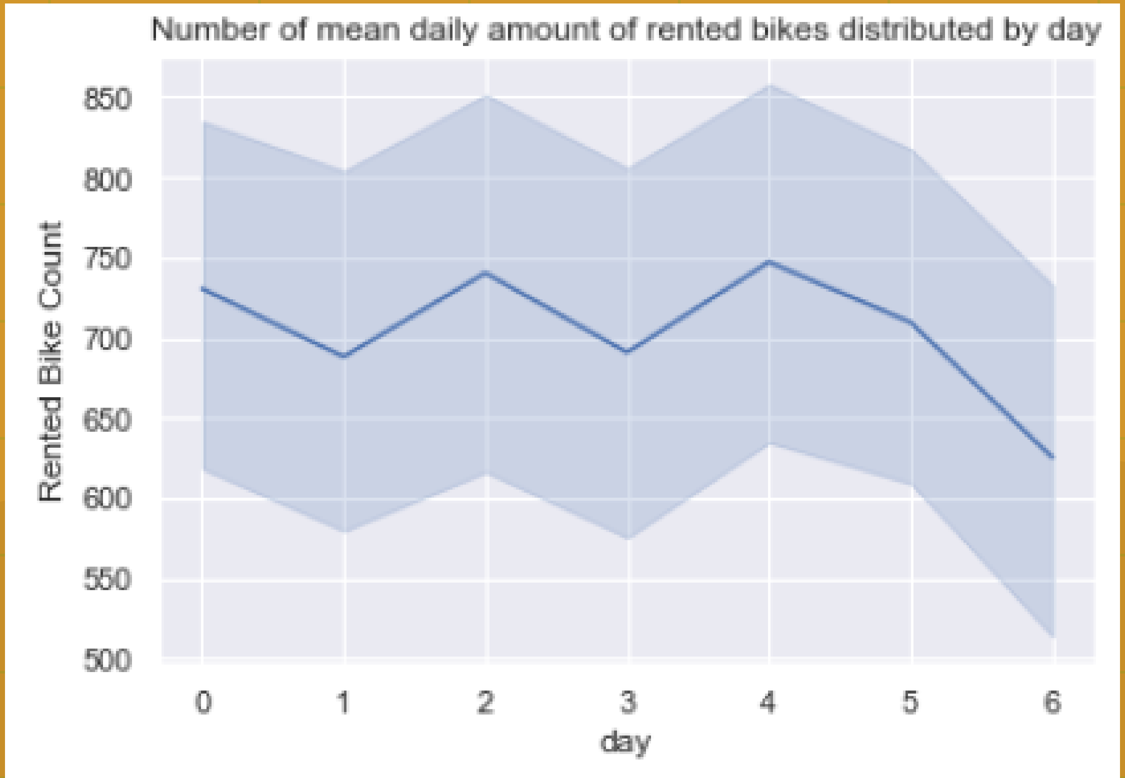




Rented Bike Count Visualisation

Rented Bike Count Visualisation





Data modeling

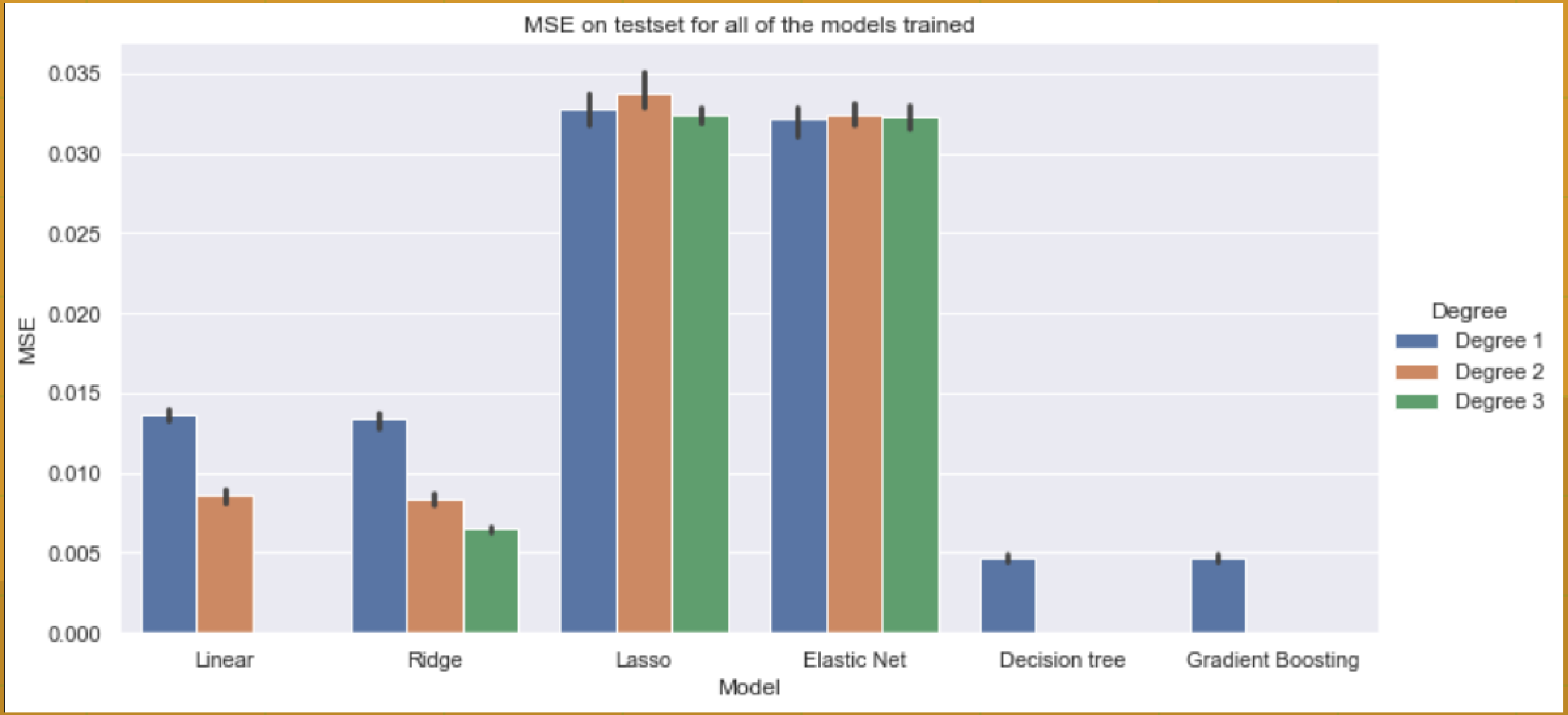
Search for the best model

We scale data

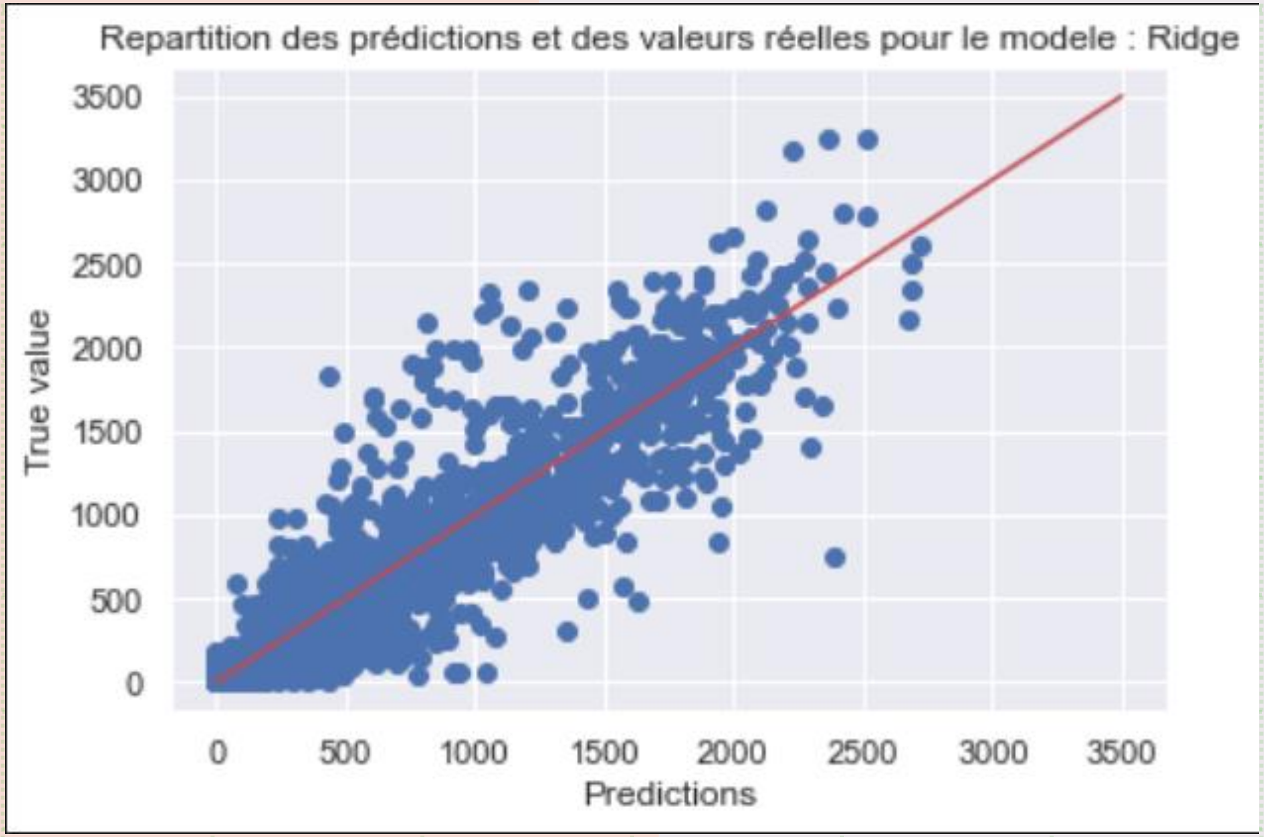
We prepare date for future Regression

We perform a grid search to search for the best model and best hyper parameters

Visualisation of the best model



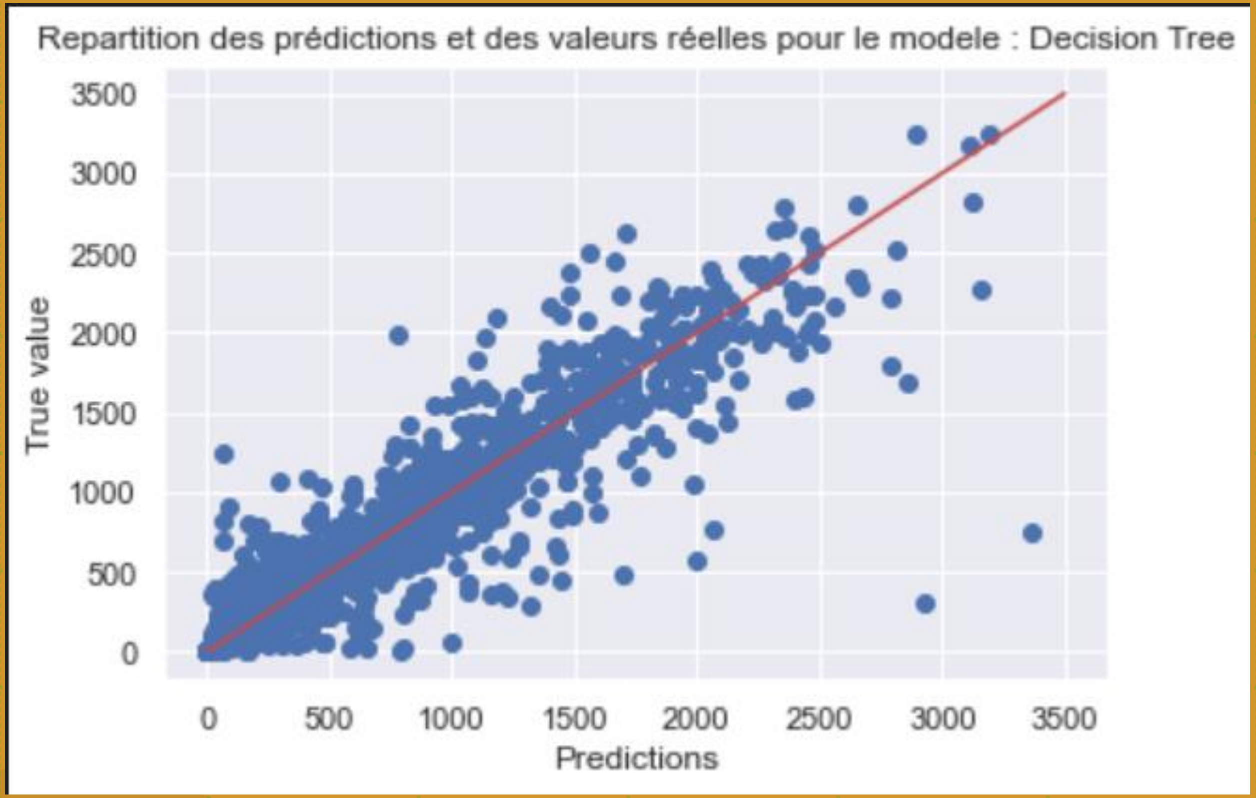
We study more in depth each model



Ridge Model

MAE : 186.22029

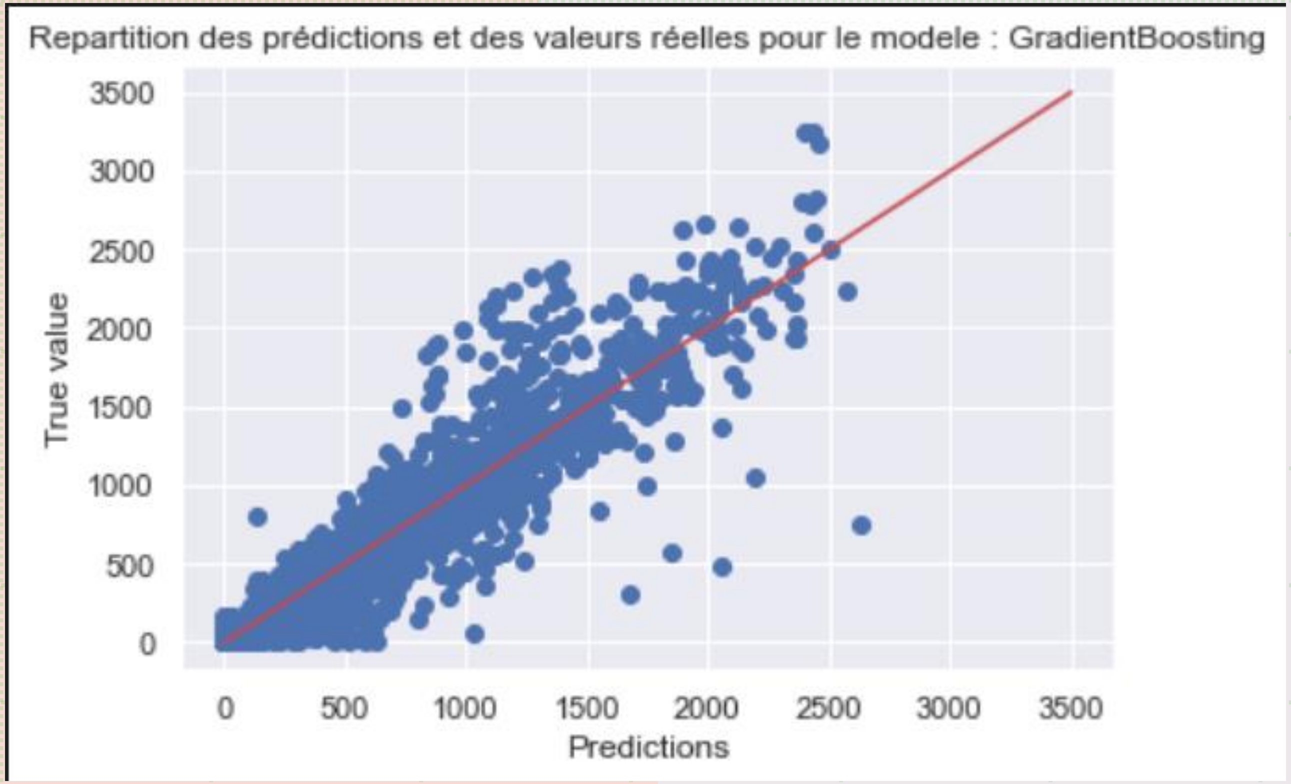
MSE : 77083.25



Decision Tree

MAE : 129.98744…

MSE : 54158.654065…

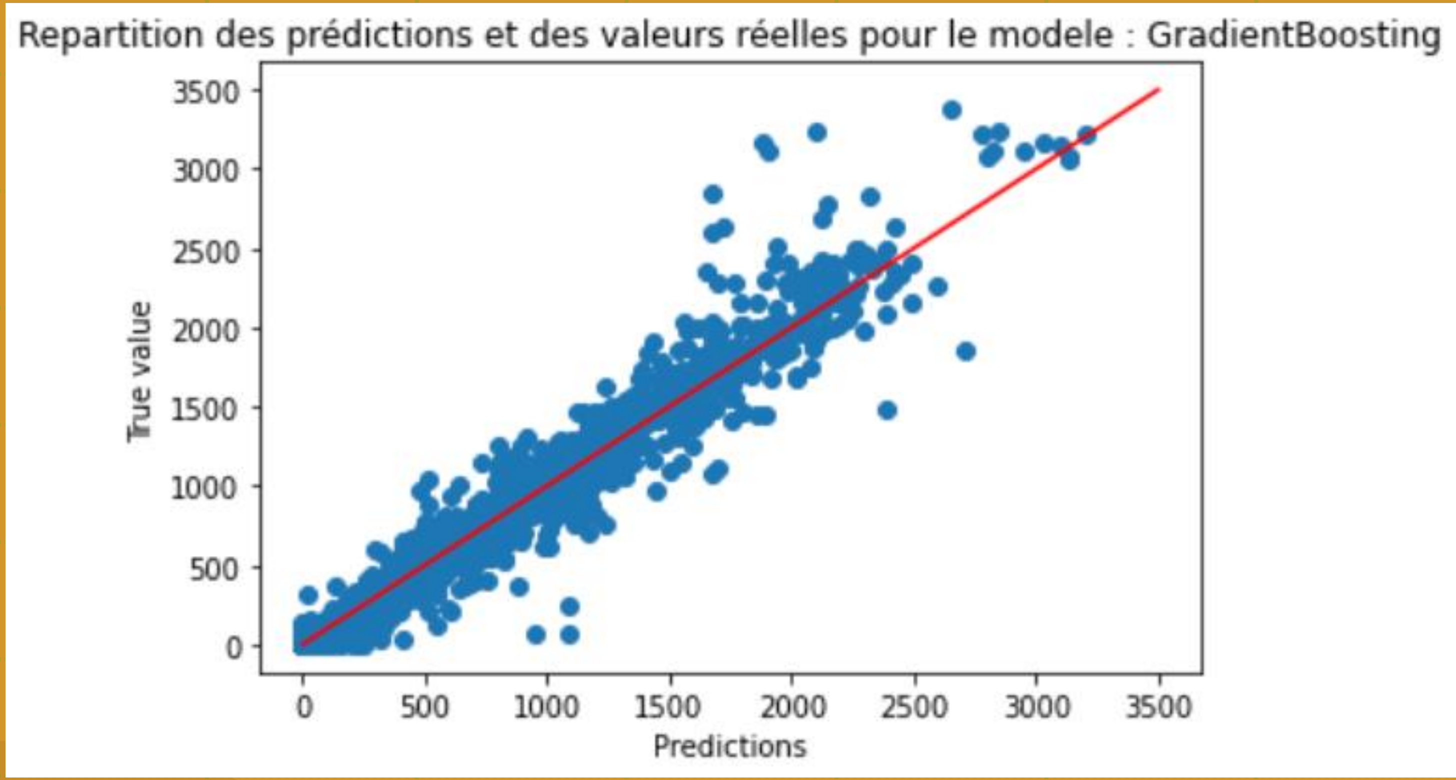


Gradient Boosting

MAE : 158.185571…

MSE : 58809.840246…

Bonus : We test a real time prediction



We add a column with the number of bikes rented during the last hour. The company might not always have this information, but this could lead to a strategy of live gestion of bikes in the city.

MAE : 91.39186990…

MSE : 23484.216761…