Seoul Bike Sharing Demand

Tom HAVYARIMANA & Oussama EL ATRACHE

Why we choose this dataset ?

- The rental of electric scooters has experienced a phenomenal boom in recent years. It has brought back into fashion other means of transport, such as the bike.

- In some countries, the pollution is such that there are restrictions on the means of transport used depending on the air quality.

The use of these means of transport has therefore become daily.
So it is important to make the rental bike available and accessible to the public at the right time as it lessens the waiting time.

So we found the study about these means of transport interesting that's why we chose the dataset about bikes rented in Seoul.

What is the problem ?

As we have seen previously, the rental of these bikes has become common.

Providing the city with a stable supply of rental bikes becomes a major concern.

Rental companies seek to know the number of bikes to be made available to the population, ensuring that there is something for everyone. And this depending on the time and day.

So we need to predict the number of rented bike by day and hour.

How to solve the problem ?

We have 13 variables in the dataset for the prediction of the Rented Bike Count.

Date
Hour - Hour of the day
Temperature - Temperature in Celsius
Humidity - %
Windspeed - m/s
Visibility - 10m
Dew point temperature - Temperature in Celsius
Solar radiation - MJ/m2
Rainfall - mm
Snowfall - cm
Seasons - Winter, Spring, Summer, Autumn
Holiday - Holiday/No holiday
Functioning Day - Yes/No

How to solve the problem ?

The first step is to reflect on our experience of the reasons that lead us to rent a bike. This will allow us to make a first study of existing variables and possibly create new ones.

- We know that when it's cold, it is not very pleasant to take the bike, for example.
- It is also evident that when it rains or snows, there are a lot less bike rentals.
- We can also think that visibility plays a role in the safety of the user.
- The seasons also seem to play a role in choosing to ride a bike. A good weather like in summer will motivate much more than a bad weather like in winter.
- The Functioning Day variable indicates whether or not there are rented bikes. It therefore gives us clear information depending on the day.

How to solve the problem ?

We can also find other interesting variables that could influence the number of bikes rented.

We can think about air pollution. In France and in several countries, when air pollution is too high in certain cities, there are restrictions on the use of cars and citizens are therefore forced to find other means of transport. After few inquiries, we can see that this is also the case in Korea and more precisely in Seoul.
Logically, these pollution peaks will therefore increase the number of bicycles rented.

We can also think about public holidays which can also influence the rental of bicycles.

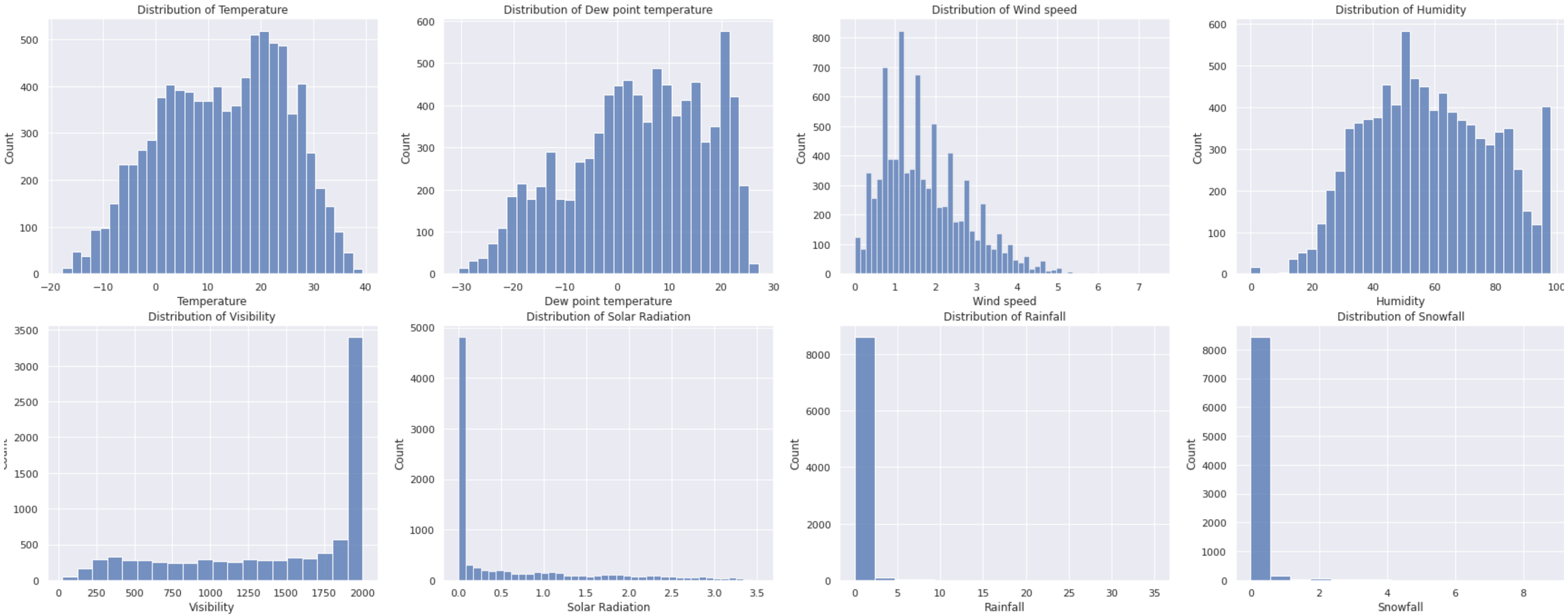
Data analysis

First, we look at the distribution of data for each of the numerical variables.

	Hour	Temperature	Dew point temperature	Wind speed	Humidity	Visibility	Solar Radiation	Rainfall	Snowfall	Rented Bike Count
count	8760.000000	8760.000000	8760.000000	8760.000000	8760.000000	8760.000000	8760.000000	8760.000000	8760.000000	8760.000000
mean	11.500000	12.882922	4.073813	1.724909	58.226256	1436.825799	0.569111	0.148687	0.075068	704.602055
std	6.922582	11.944825	13.060369	1.036300	20.362413	608.298712	0.868746	1.128193	0.436746	644.997468
min	0.000000	-17.800000	-30.600000	0.000000	0.000000	27.000000	0.000000	0.000000	0.000000	0.000000
25%	5.750000	3.500000	-4.700000	0.900000	42.000000	940.000000	0.000000	0.000000	0.000000	191.000000
50%	11.500000	13.700000	5.100000	1.500000	57.000000	1698.000000	0.010000	0.000000	0.000000	504.500000
75%	17.250000	22.500000	14.800000	2.300000	74.000000	2000.000000	0.930000	0.000000	0.000000	1065.250000
max	23.000000	39.400000	27.200000	7.400000	98.000000	2000.000000	3.520000	35.000000	8.800000	3556.000000

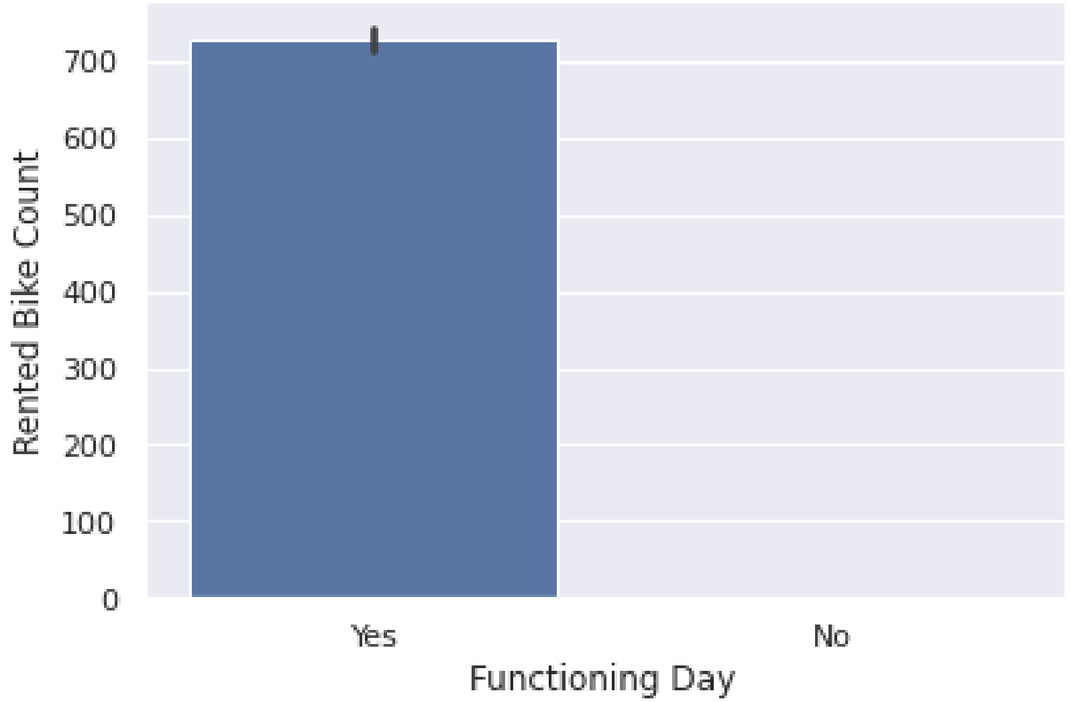
Data analysis

We also plot each distribution for a better overview.



Data analysis

After a few manipulations, we can see that Functioning Day = "No" indicate that there were no rental on that day.

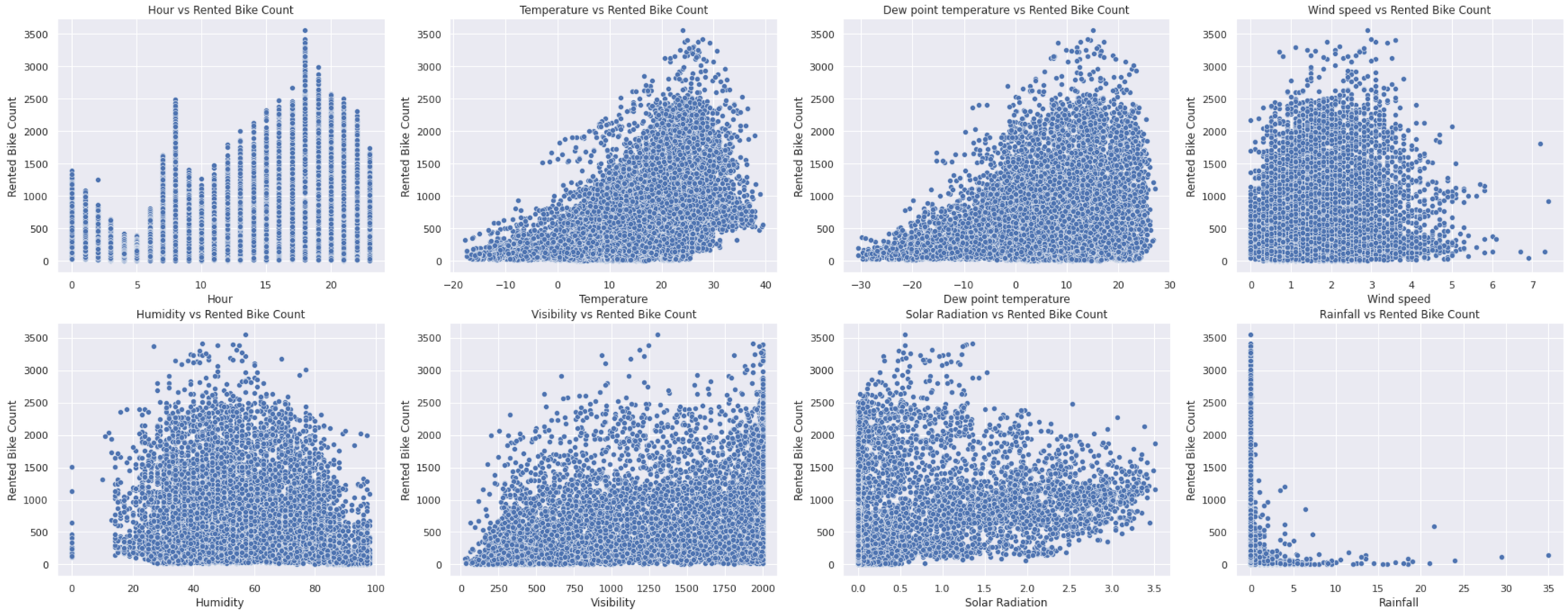


Functioning Day	Rented Bike Count
No	0
Yes	6172314

Here we have calculated the sum of the bikes rented for each of the Functioning Day cases.

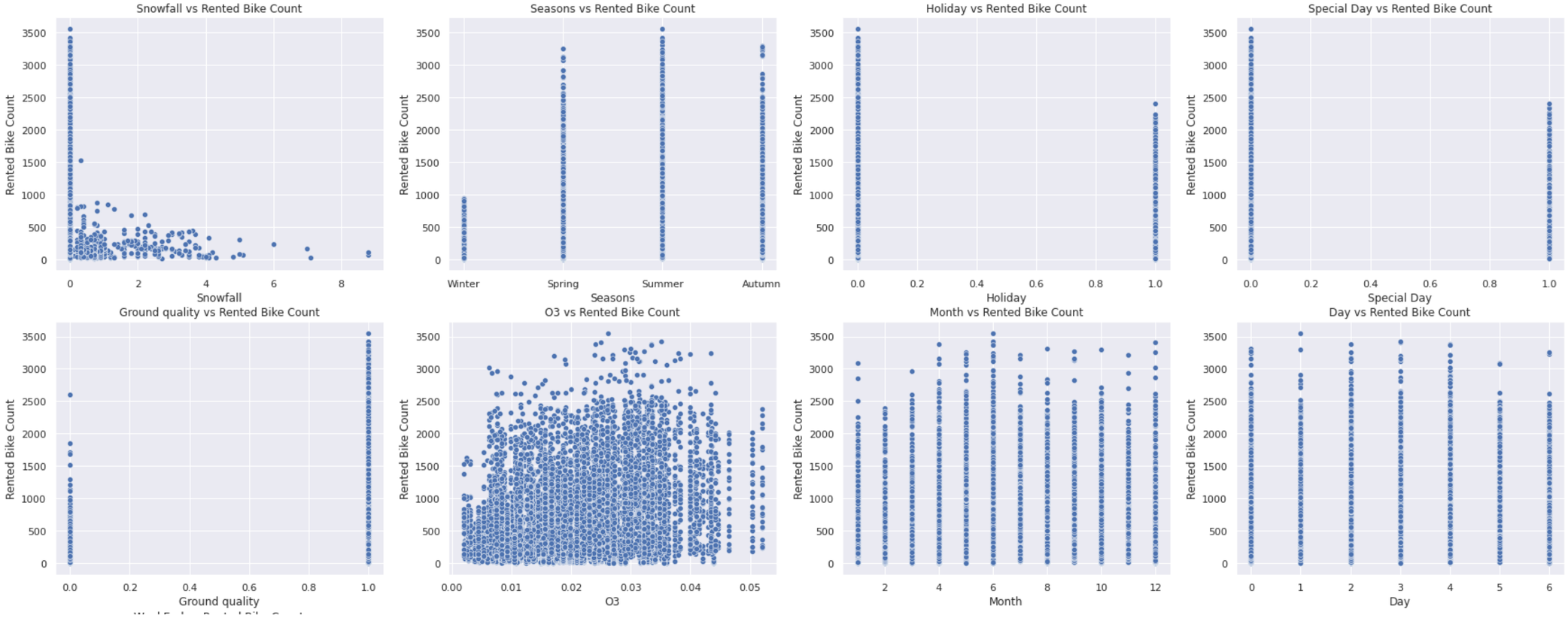
Data analysis

After adding some variables, including pollution and public holidays that we mentioned previously, we carry out some visualizations to see the correlations between each variable and the target variable (Rented Bike Count).



Data analysis

We can see on the graphs the different variables that have been added, they will be explained.

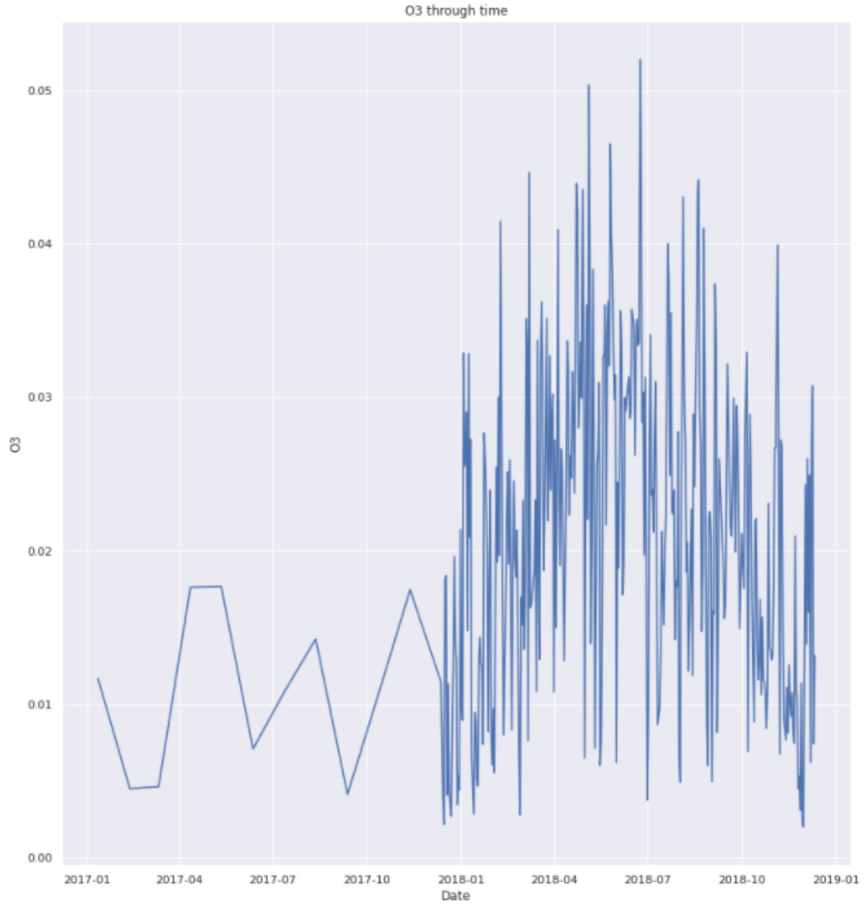


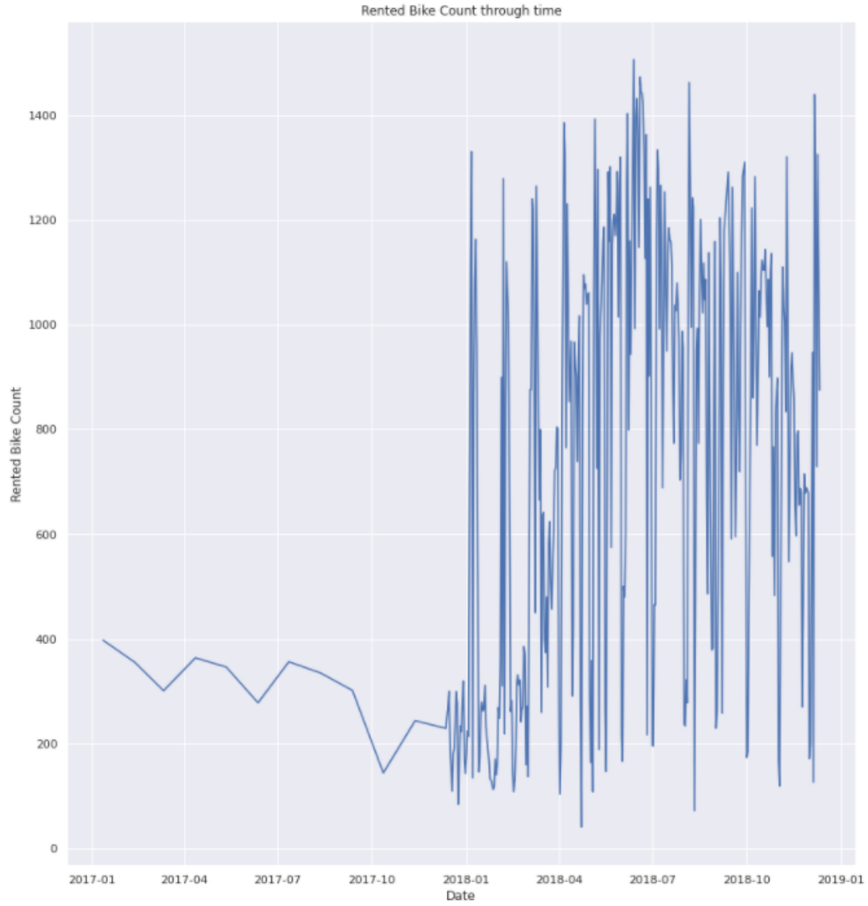
Data analysis

- O3 : The daily amount of O3 pollution in Seoul. We got this information from a dataset we found on Kaggle.
- Special Day : Public holidays in Seoul.
- Ground quality : Gives information on the quality of the ground (slippery or not). The quality of the ground is good when it has not rained or snowed.
- Month : We obtained this information from the date.
- Day : We obtained this information from the date.
- Week-end : Yes/No

Data analysis

Here, we carry out an analysis according to the date between the O3 and the number of bikes rented. We can see that the two variables evolve in the same way.





Data analysis

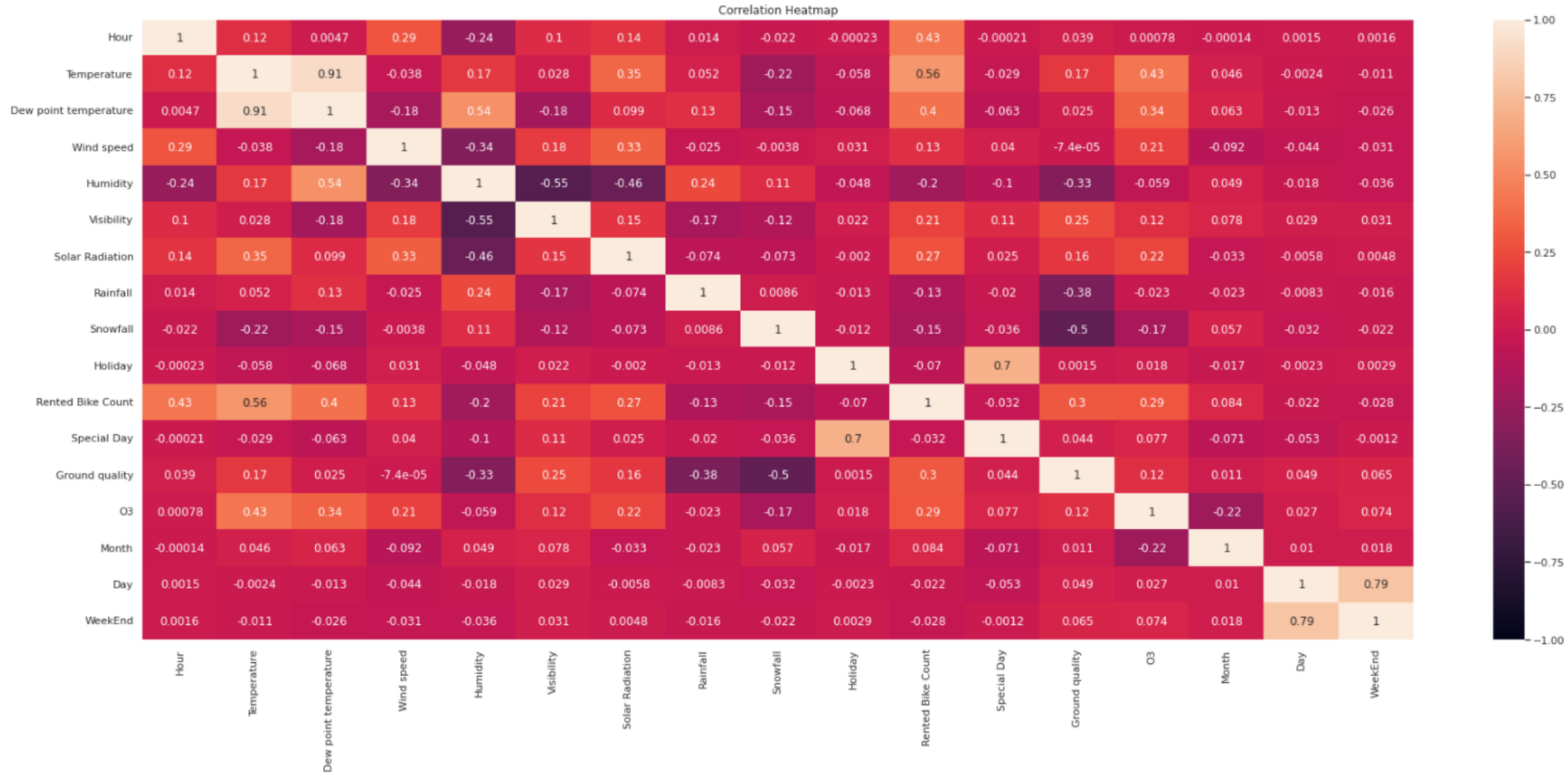
Thanks to previous visualizations, we can already verify some of our hypotheses made previously.

- We can see that there are more bikes rented in the evening than in the morning, for example. There also seem to be peak hours (8am in the morning and 6pm in the evening)
- There seems to be a good correlation between temperature and the number of bikes rented.
- There are generally more bikes rented in summer than in winter.
- There are usually more bikes rented when the ground is not slippery.
- There are a lot less rental bikes when it rains or snows.

We will now look at the correlations in more detail in order to identify the most important variables. For this, we will use a correlation heatmap.

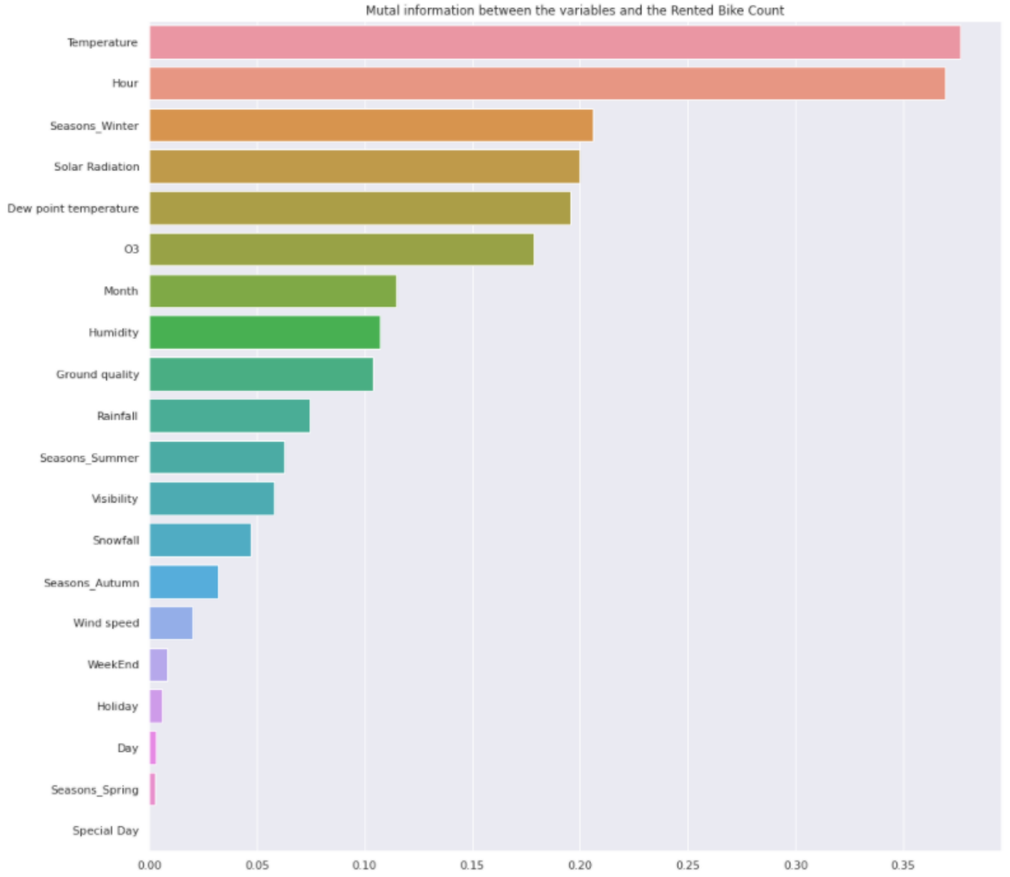
Data analysis

We can read this correlation heatmap in row or column. The correlations that interest us are mainly those between Rented Bike Count and the other variables.



Data analysis

Here, we have a plot illustrating the mutual information. It is a widely used tool for the selection of interesting variables. The mutual information (MI) quantifies the "amount of information" obtained about one random variable by observing the other. We can compare it to the correlation but it is not linear.



The higher the MI score, the more the variable "provides information" to predict the target variable.

Data analysis

Thanks to the two previous plots, we can finally see clearly which variables play the greatest role in the prediction of the target variable.
Unsurprisingly, temperature has a great correlation with the number of bikes rented and the best Ml score.

Now, our goal is to select variables that we consider unnecessary and remove them in order to have an optimal machine learning model.
Thanks to the study of the data that we have done so far, it seems that the less important variables are "Holiday", "Special Day", "Day", "WeekEnd", "Wind speed" and "Month "(because of its bad correlation).

Data analysis

To find which variables to delete, we can create a system allowing to test a Machine Learning model and to give a score to the dataset after deleting each variable separately (in our case, we will use the mean squared error).

The program that we use will therefore indicate the variable that we deleted from the dataset and the mean squared error of the dataset without this variable. The goal is to minimize the error.

The model that we will use for this test is the Decision Tree.

We launch the program on all the selected variables.

Month	92629.249410
Day	87205.170997
Special Day	85505.954653
Holiday	83105.350496
WeekEnd	83026.638167
Wind speed	80968.380255

=> It can be seen that removing "Wind speed" minimizes the error.

We relaunch the program on all the selected variables with the "Wind speed" already dropped.

Month	87420.399622
Day	86869.103448
Special Day	84971.863958
WeekEnd	83714.701937
~~Wind speed~~	~~80968.380255~~
Holiday	79922.571564

=> The error is minimal when removing the "Holiday" and "Wind speed" variables. So we drop them.

Machine Learning model selection

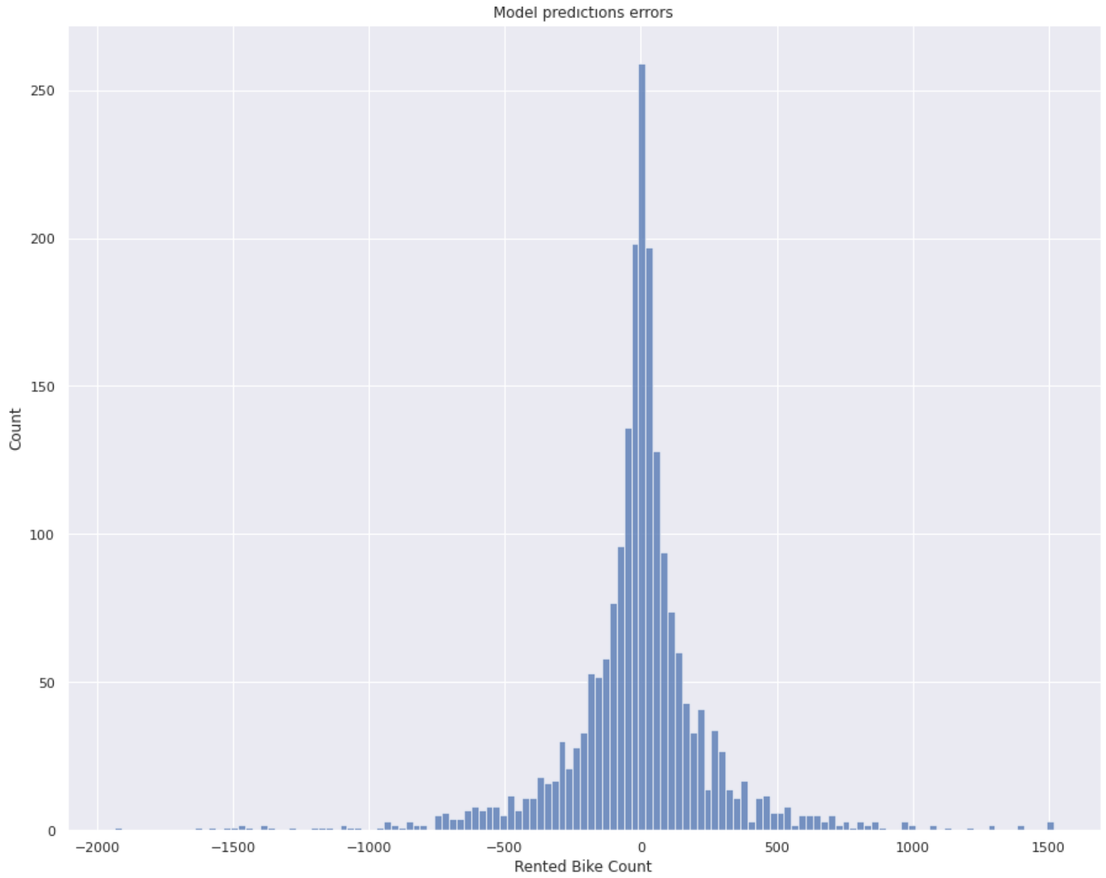
The criterion for selecting an ML model is how well it makes good predictions. Or how much the model minimizes errors. To perform these tests, we separate the dataset into train and test set.

Here, as we have to predict the number of rented bikes, it is a regression problem.
We can try several models like the Decision Tree, the Random Forest and the Xtreme Gradient Boosting.

We carry out an optimization of the hyperparameters of each of the models in order to be able to compare them on a solid basis.
Hyperparameters being information on which models depend to be efficient.

Machine Learning model selection

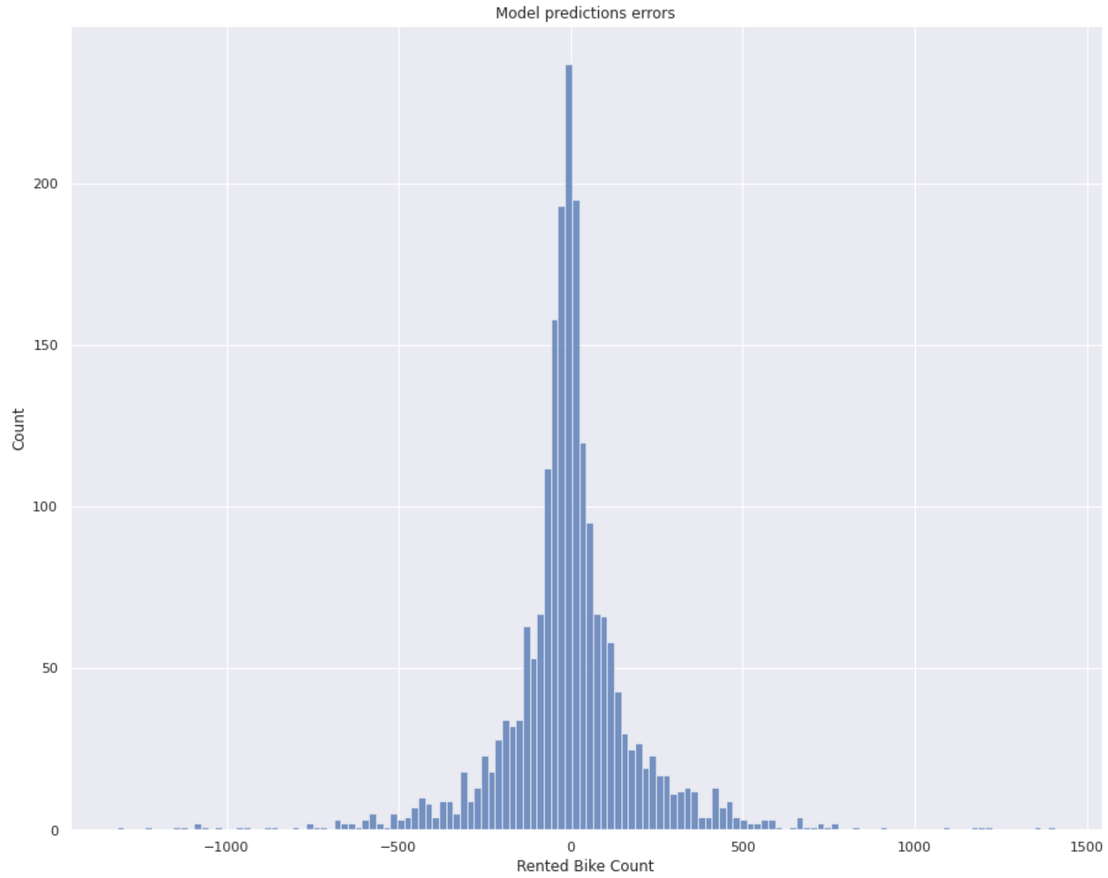
For the Decision Tree model:




```
count     2117.000000
mean       -11.873406
std        283.403735
min      -1938.000000
25%        -93.000000
50%         -1.000000
75%         80.000000
max       1518.000000
Name: Prediction error stats
```

Machine Learning model selection

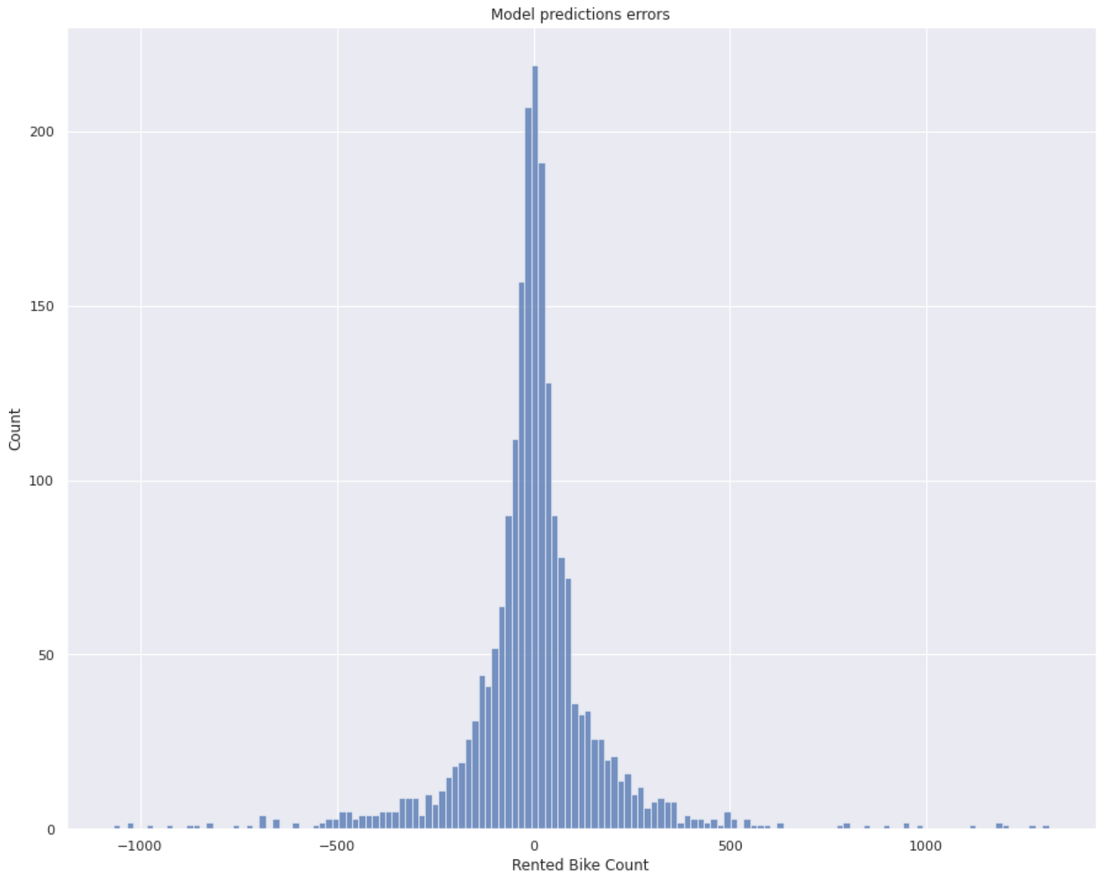
For the Random Forest model:




```
count    2117.000000
mean       -6.431271
std       208.654775
min     -1319.000000
25%       -70.000000
50%        -8.000000
75%        61.000000
max      1407.000000
Name: Prediction error stats
```

Machine Learning model selection

For the Xtreme Gradient Boosting model:




```
count    2117.000000
mean       -0.944261
std       181.826050
min     -1069.000000
25%       -56.000000
50%        -1.000000
75%        53.000000
max      1311.000000
Name: Prediction error stats
```

Machine Learning model selection

To compare the performance of our models, we can use 2 statistics from each table, the mean and the standard deviation (std).
Our goal is to have a distribution of the errors which is centered (average = 0) and the least "spread" possible (minimum std).

When we compare our models, we see that it is the Xtreme Gradient Boosting that best meets these conditions. It is therefore the most efficient model (The others statistics allow to confirm that).

In addition, the R^2 score obtained with the Xtreme Gradient Boosting model is 0.9216. This means that 92.16% of the variation in the number of rented bikes is explained by our data, which is excellent.

Conclusion

The datas provided by the dataset are very interesting and relevant. They allow to efficiently predict the number of bikes rented in Seoul by date and hour.

To improve the performance of the model, we think it would be interesting to collect data on the profile of users of this type of bike.