Predict amount of rented bikes in Seoul

Can we predict for a given hour, how many bikes will be rented in Seoul ?

Here is our parameters in order to predict that (with one exemple) :

We estimate the parameter « Rented Bike Count »

Rented Bike Count
725

Hour	Temperature(°C)	Humidity(%)	Wind speed(m/s)	Visibility(10m)
0	-5,2	42	0,8	2000

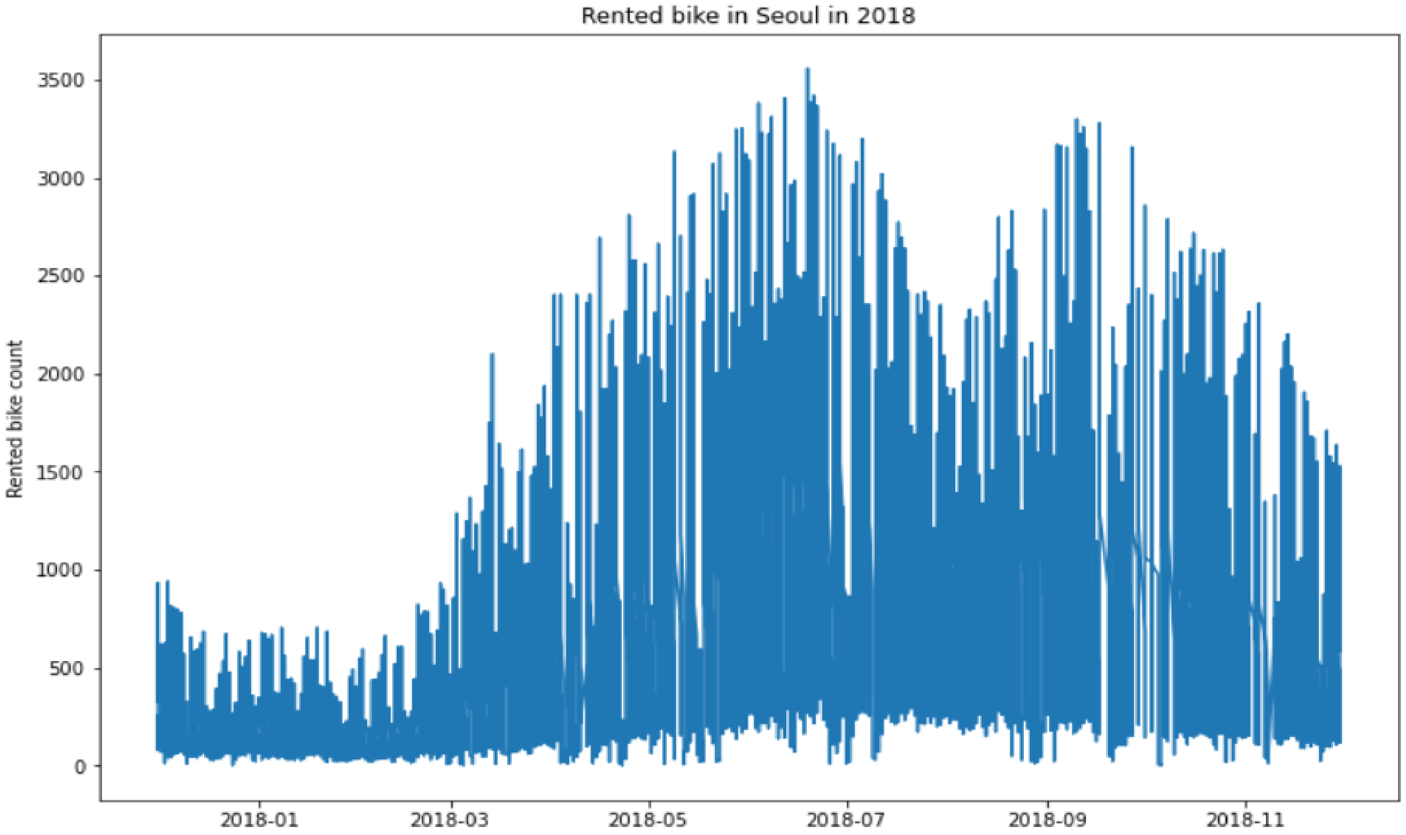
Solar Radiation(MJ/m2)	Rainfall(mm)	Snowfall(cm)	Season	Holiday
0	0	0	Winter	No

How to visualize our data ?

- After cleaning our data by removing non functioning days of bicycle rent we started to think about how to visualize relevant data.

- We create subsets in order to compare and show tendancies like the four seasons.

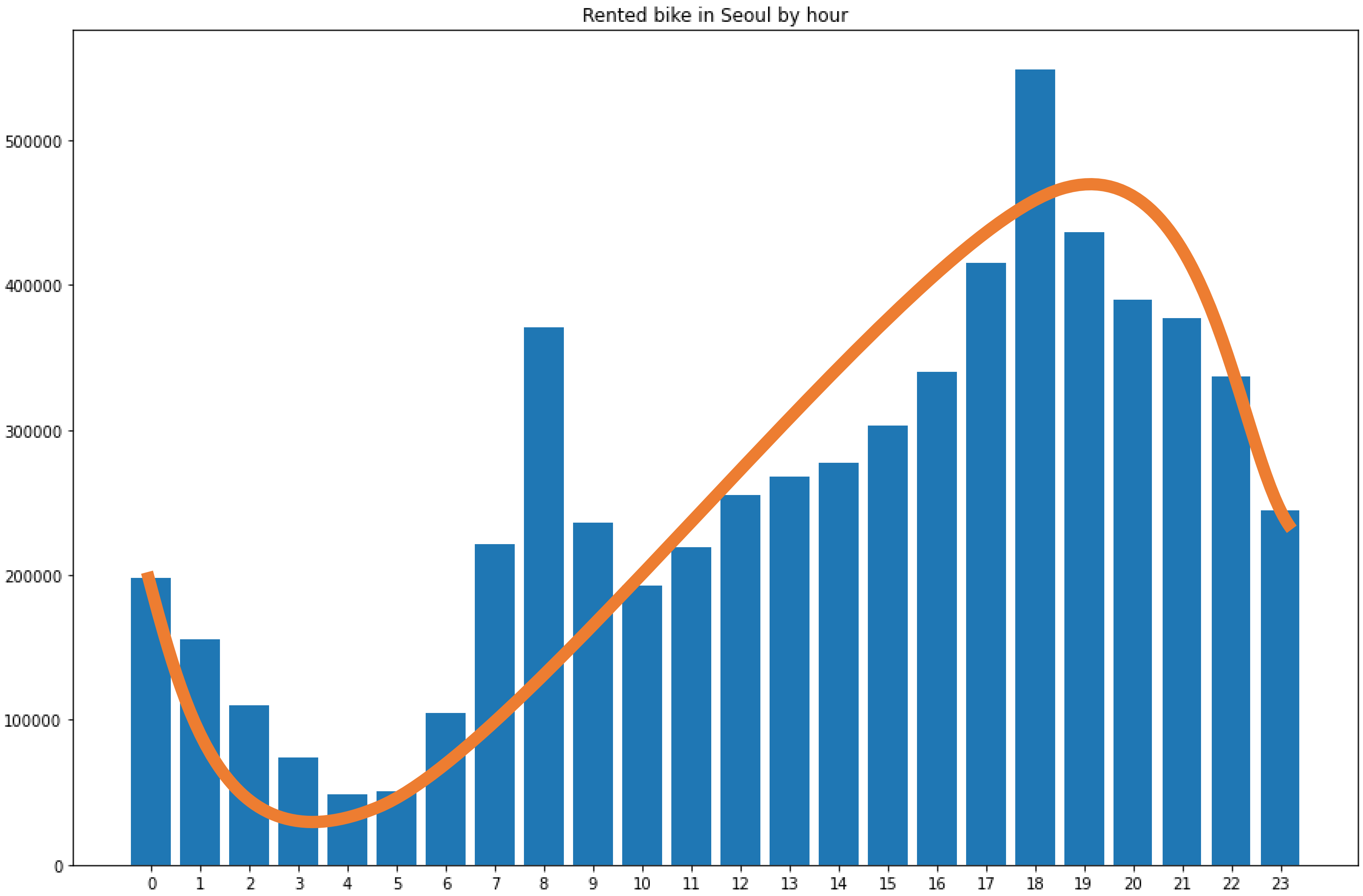
- We create variables corresponding to each hour in order to see the relation between rented bikes and the hour of the day. Then we make a barplot which represent that nicely.



Global data

Seoul 2018

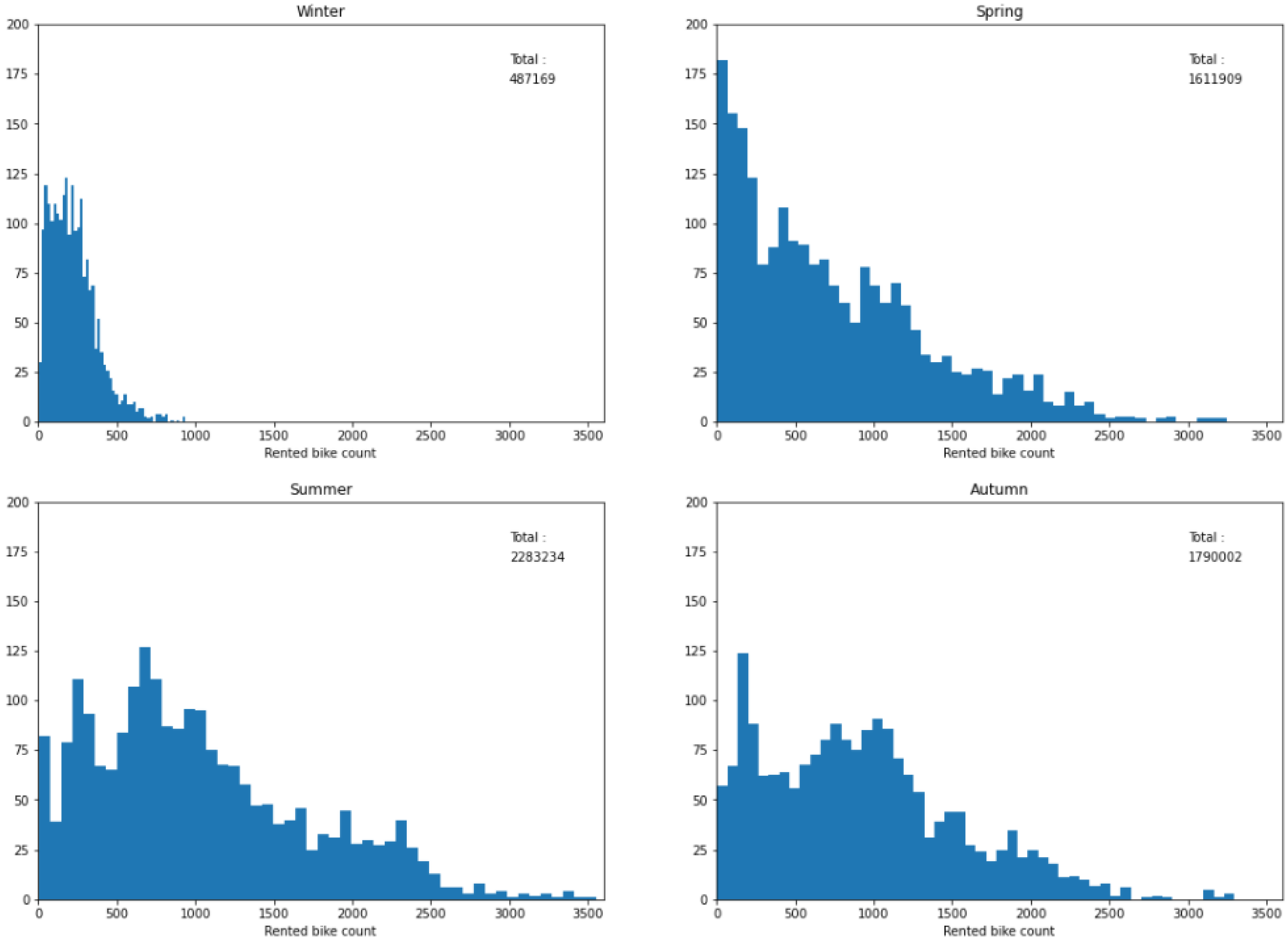
Data by hour

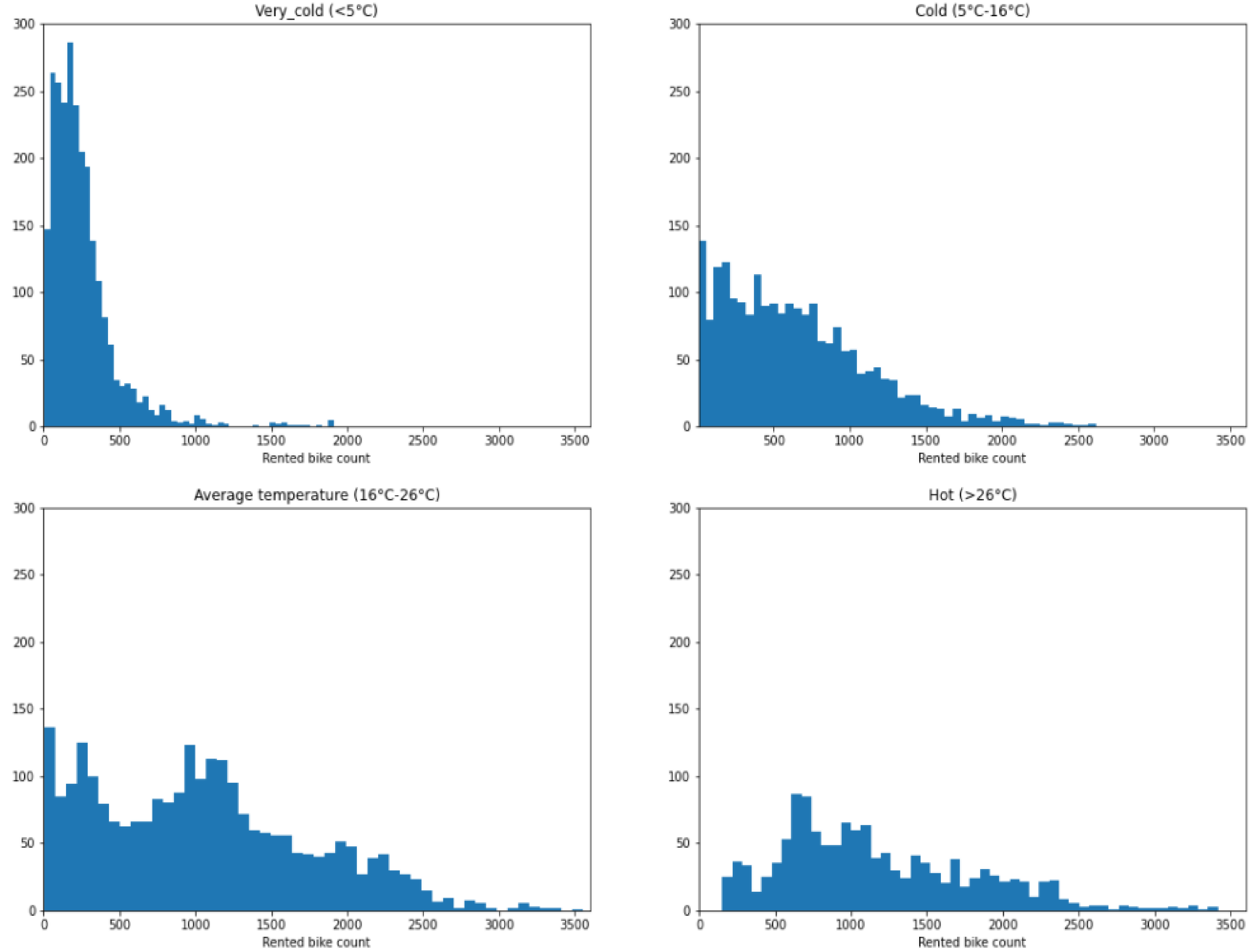


Thanks to this barplot we can see some kind of function and draw it like in this graph. But because of rush hours we have also two big peaks between 7h-9h and 17-19h

Seasons

By classificate by season and show histograms we learned that Spring and Autumn were at the same level while Winter is a lot lower and summer higher.



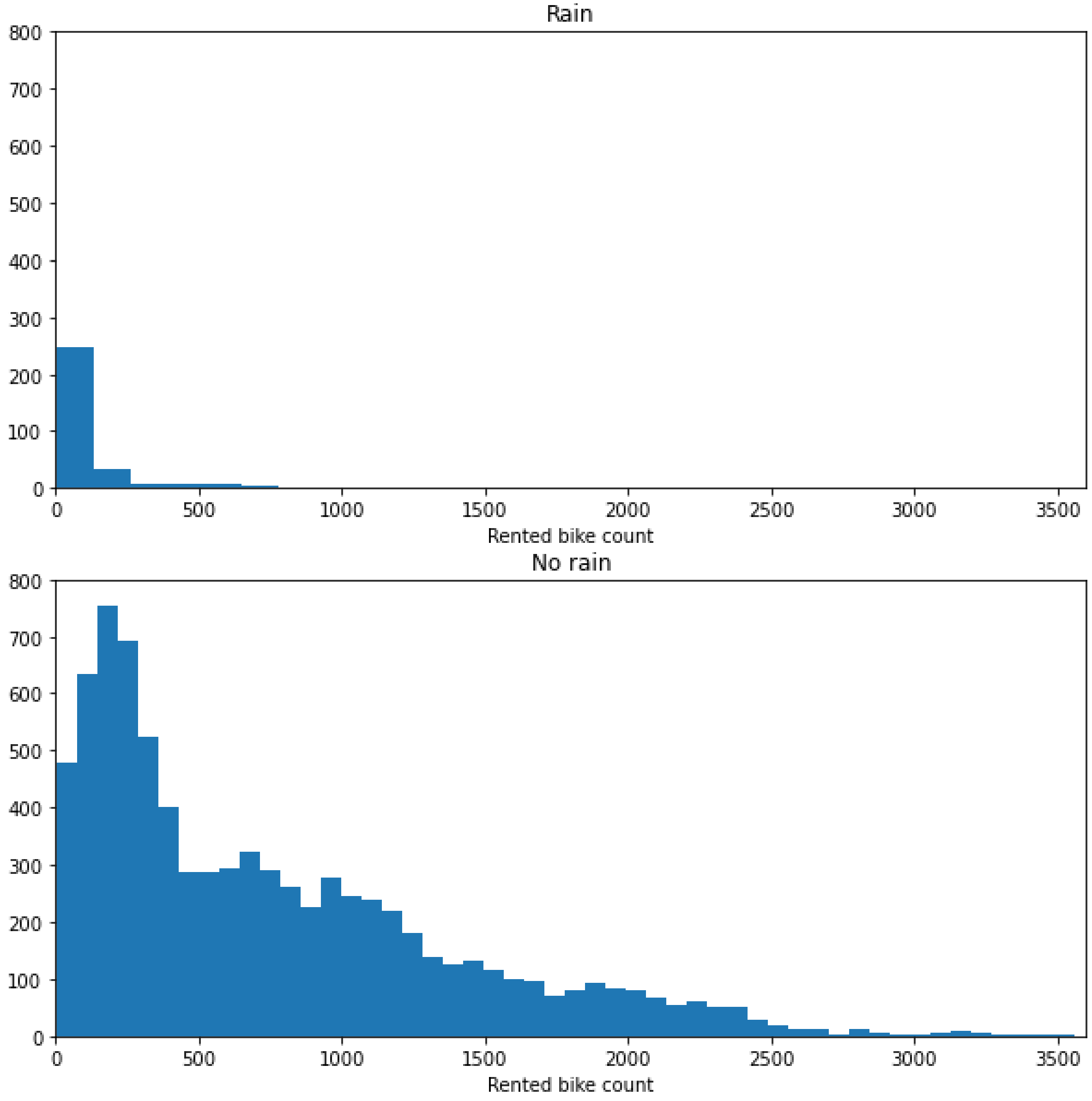


Temperature

We did the same thing with temperature. We can see that when the air is warm enough there is a lot more of bicycle rented.

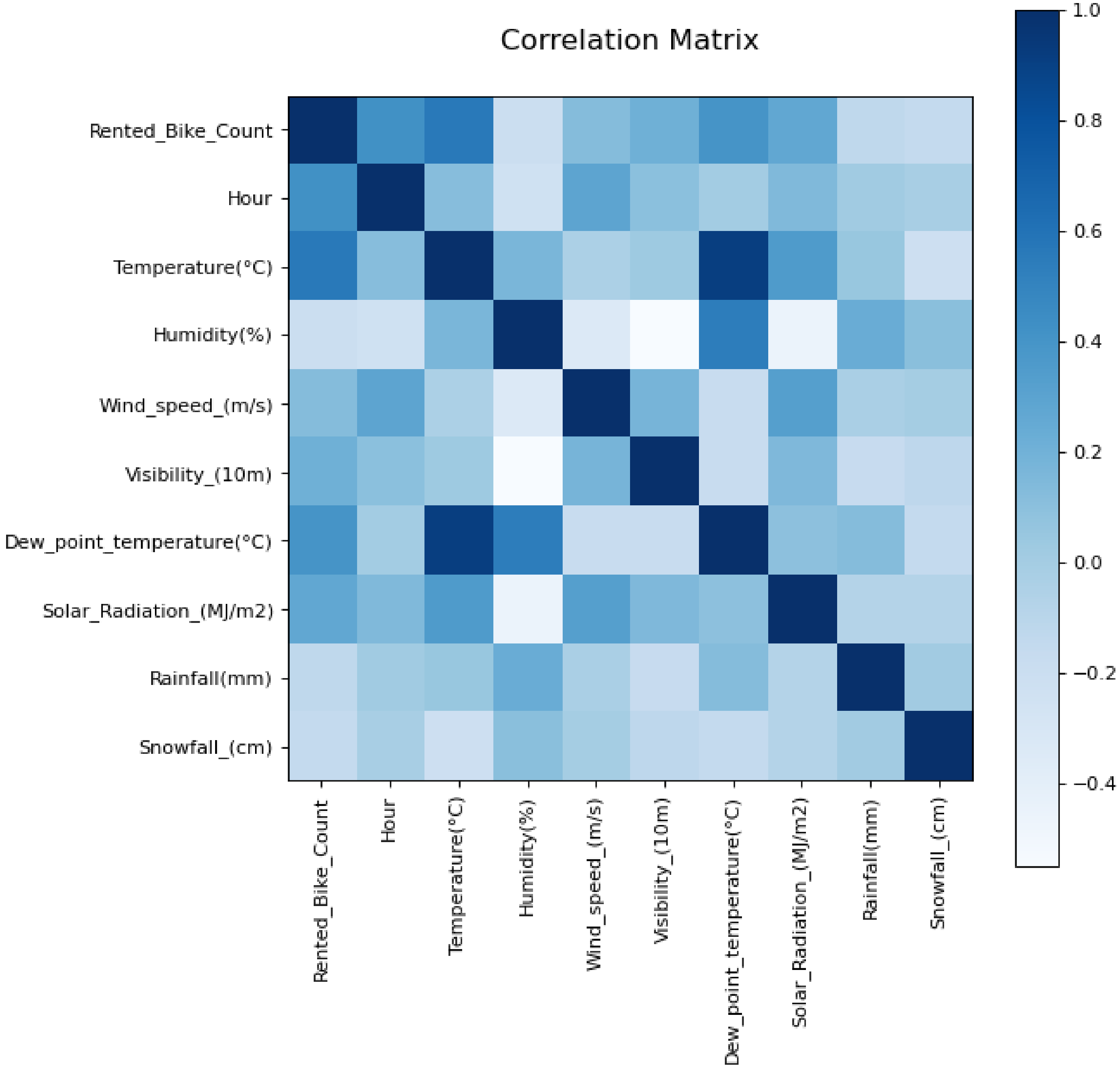
Rain

Rain (and snow as well) are eradicating the amount of bikers but this it is not happening often. As a consequence it is not really a great feature to find our estimation.



Correlation matrix

We made a correlation matrix which shows that the two more relevant features are the **hour and the temperature.**

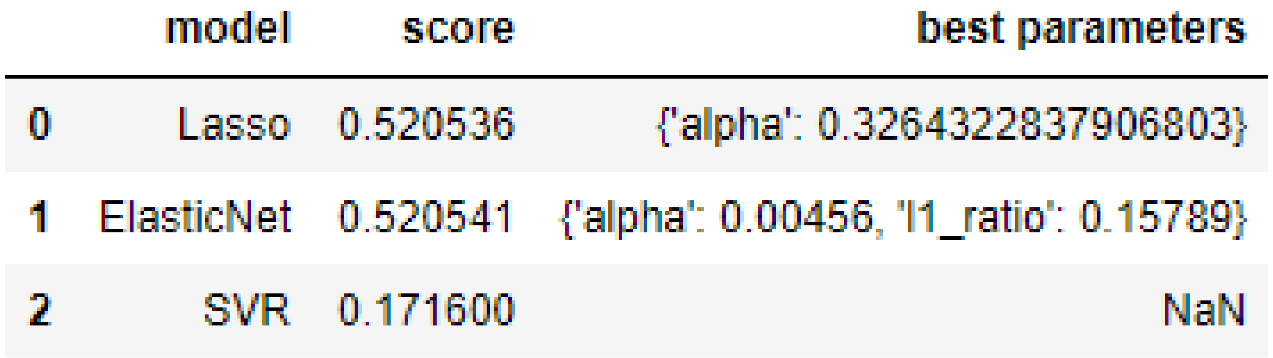


Modelisation

- We have a regression with few important features so our model will more likely be :

  - LASSO
  - ElasticNet

- With train_test_split and GridSearchCV from sci-kit learn we evaluated the best model possible and found :

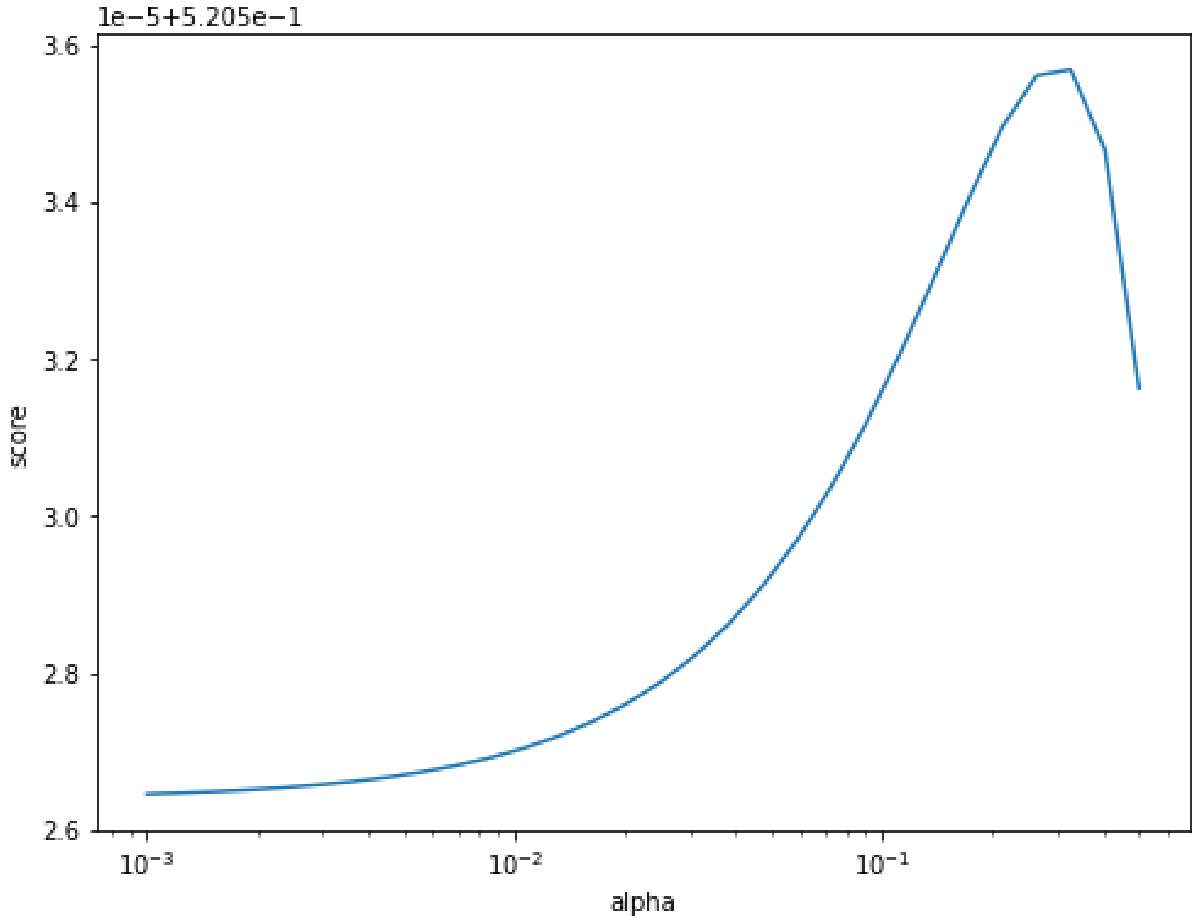


	model	score	best parameters
0	Lasso	0.520536	{'alpha': 0.3264322837906803}
1	ElasticNet	0.520541	{'alpha': 0.00456, 'l1_ratio': 0.15789}
2	SVR	0.171600	NaN

- We also made an SVR model to compare and find a score of 0.1716 which confirm our feeling about relevance of LASSO and ElasticNet due to few important features.

GridSearchCV for LASSO

LASSO model has only one hyper-parameter which is alpha so we can plot the score obtained depending on this parameter. The highest score is obtained with alpha = 0,32643



Conclusion

- With respectively 0.520536 and 0.520541 Lasso and ElasticNet have very similar results. Those scores are the coefficient of determination R^2 of our predictions. $R^2 = (1-u/v)$ where u is the RSS and v the TSS (Residual and Total Sum of Squares).

- Our final score is superior to 0,5 that means that our model can explain more than half data. It is difficult to predict the number of rented bikes because it is influenced a lot by working days and we did not have this information. Meteo and time are not enough but we prove that it was not negligeable.