PYTHON FOR DATA ANALYSIS

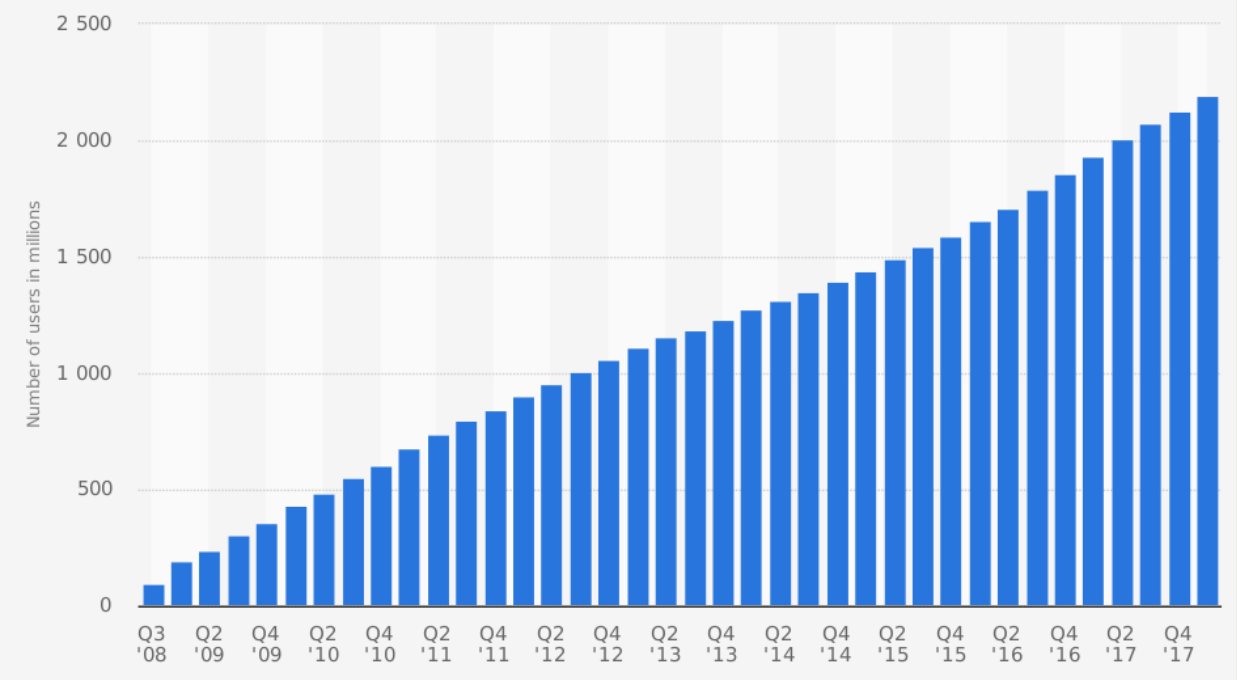
Hugo STEPHAN

SUMMARY

- Context
- Dataset
- Data discovery
- Data Visualization
- Models
- API

CONTEXT

- Facebook is an ever-growing social network
  - It accounts for almost 2.5 billion accounts across the world



DATASET

This is the “Facebook Comment Volume Dataset”
(https://archive.ics.uci.edu/ml/datasets/Facebook+Comment+Volume+Dataset)

- The dataset comes in 5 different variants. Each of these variants represents a base time for collection. I chose to only work with the first one : "Features_Variant_1“
- There are 54 attributes, one being the target output
- As the columns have no tags, let's add names to the attributes, as given on the UCI website

DATA DISCOVERY

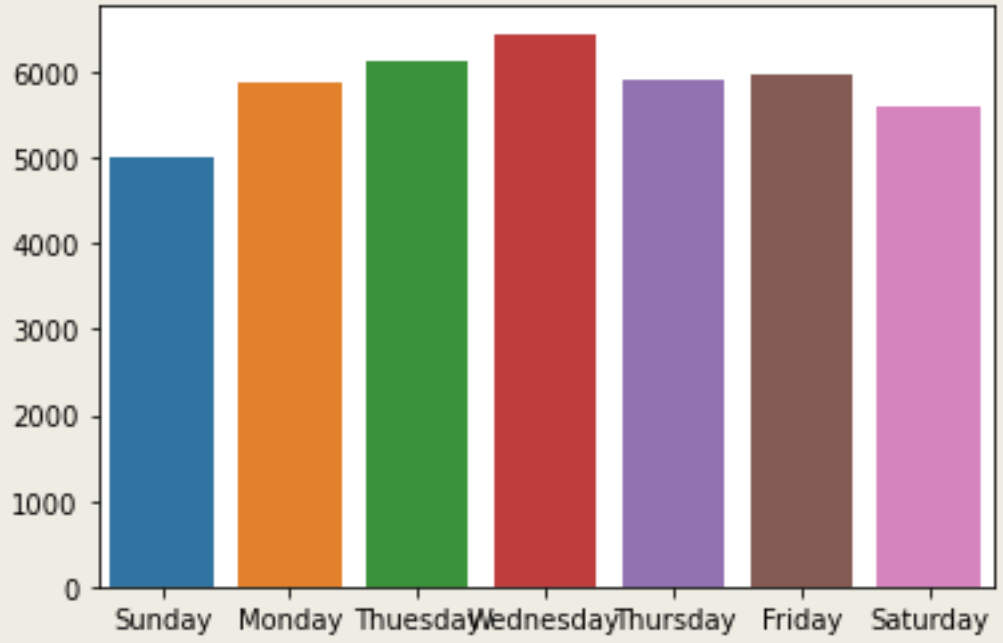
- We seem to be dealing with only numerical values
    - 15 features are binary
        - post_pub_sunday/…/post_pub_Saturday
        - base_date_sunday/…/base_date_Saturday
    - One is categorical
        - page_category
    - The others are continuous
- There are no missing values

DATA DISCOVERY

- The 'post_promotion_status' feature doesn't seem useful, as it is always zero, so I got rid of it
- A lot of boxplots were used to better understand the data
- It was concluded that our data comes in many different forms. It is quite visible to notice the disparity and variety in scale, distribution and density between all the different variables

DATA VISUALIZATION

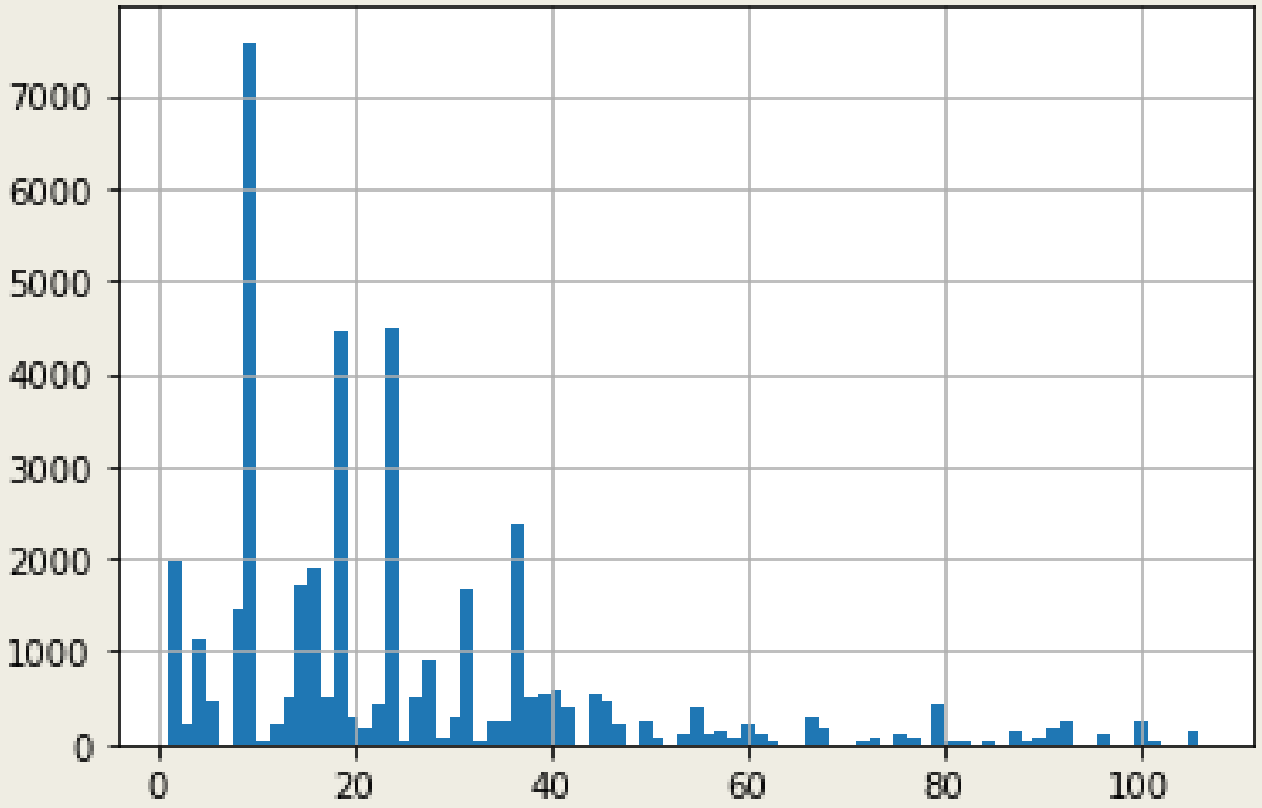
Number of posts per day



The difference between the different days of the week is relatively small, even though there seem to be low and high peaks, respectively on Sundays and Wednesdays.

DATA VISUALIZATION

How many categories are there ? Which are the most and less active ones ?

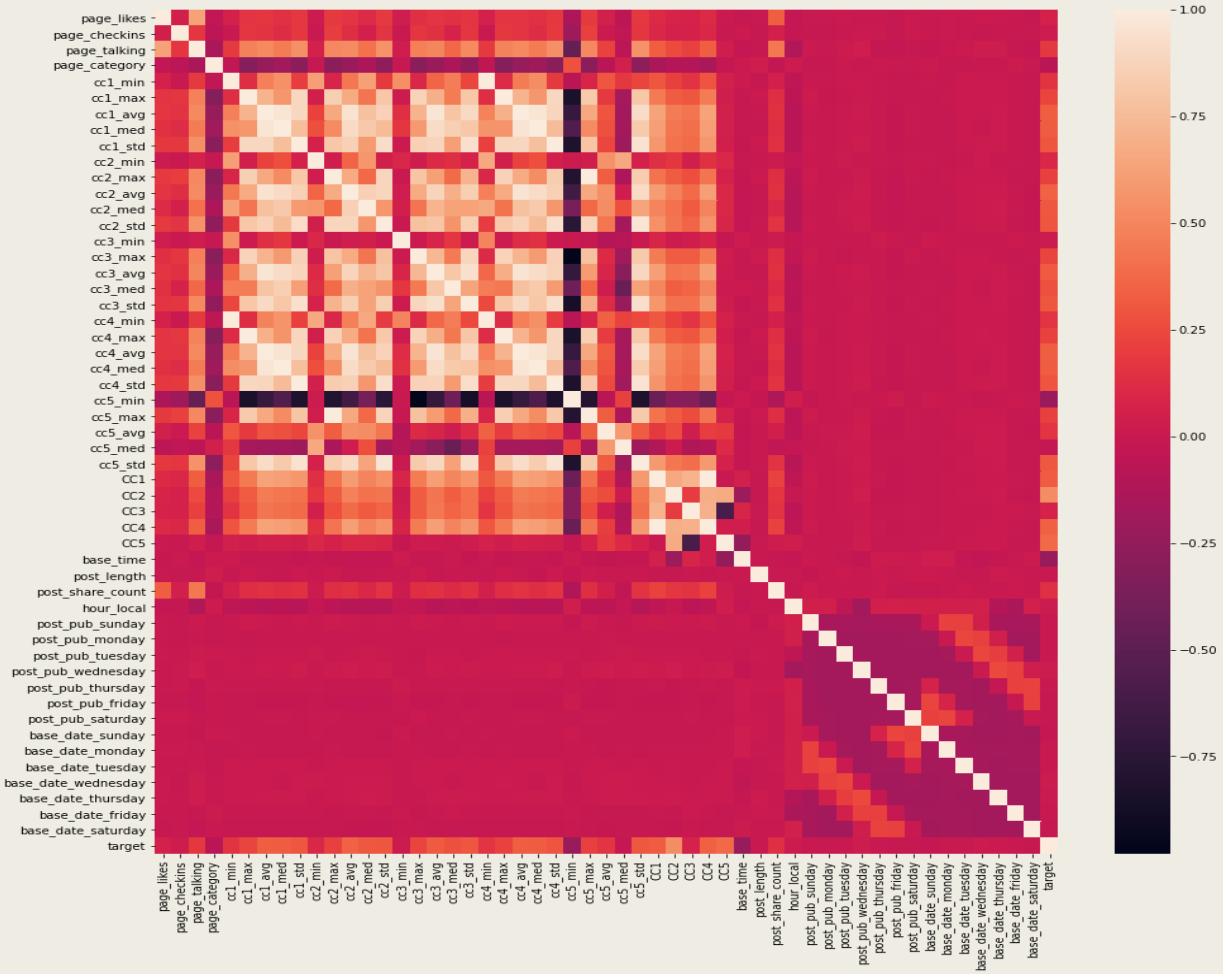


9	7494
24	4511
18	4301
36	2387
16	1890
	...
62	16
63	4
58	2
93	1
83	1

There are 81 different categories.
The ninth is the most frequent.

DATA VISUALIZATION

Correlation between the variables and the target



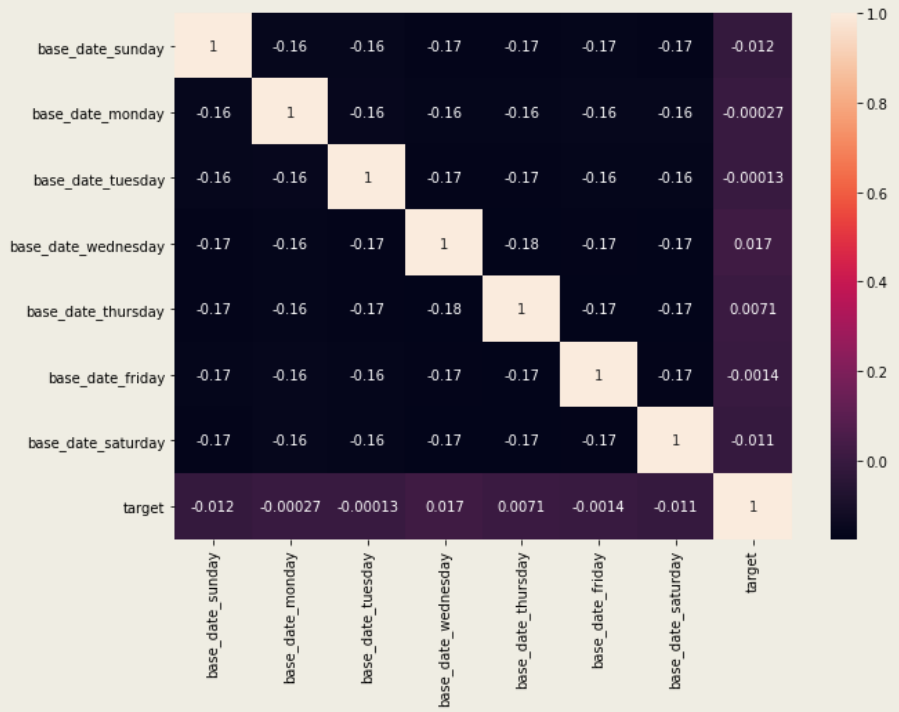
All the derived variables create zones of rich correlation both with the others variables, but also with the target.

It is poorly perceptible on this slide, the graph is more visible in the python file.

DATA VISUALIZATION

Correlation between the days of the week of publishing and the target

Correlation between the base_time days of the week and the target

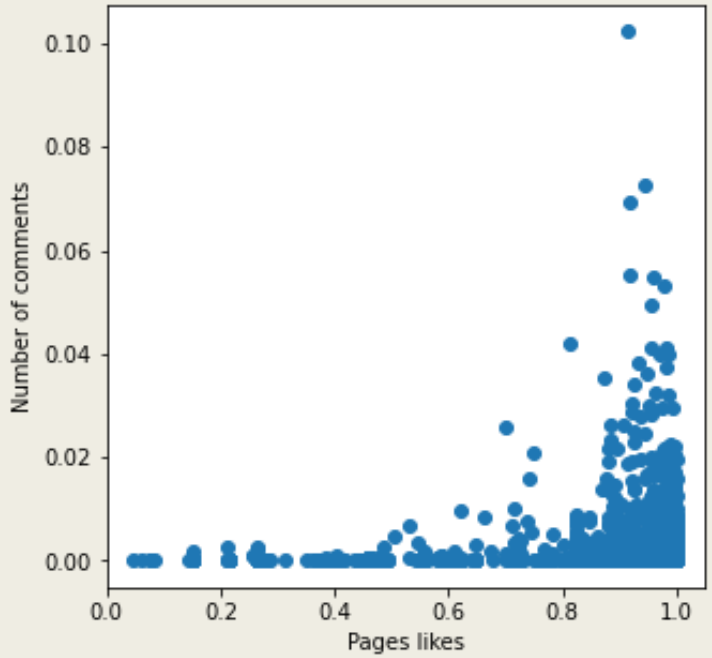


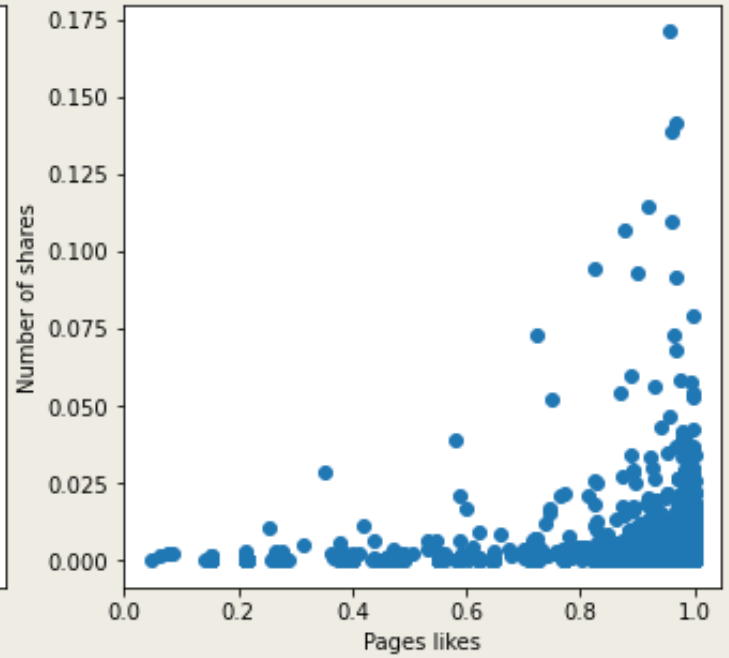
It doesn't seem to make much of a difference if a post is published on any specific day. The results are the same for the base_time.

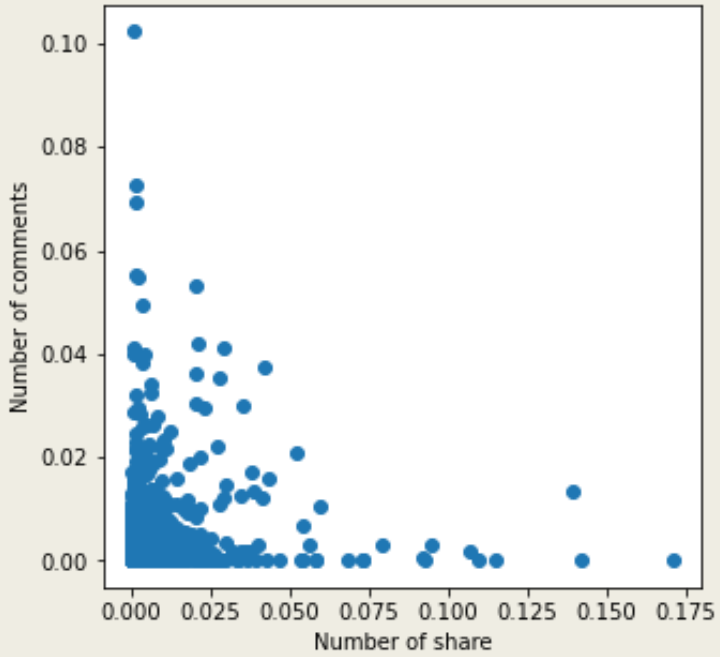
DATA VISUALIZATION

After normalizing, some scattering :

- Number of comments depending on the page likes
- Number of shares depending on the page likes
- Number of comments depending on the number of shares



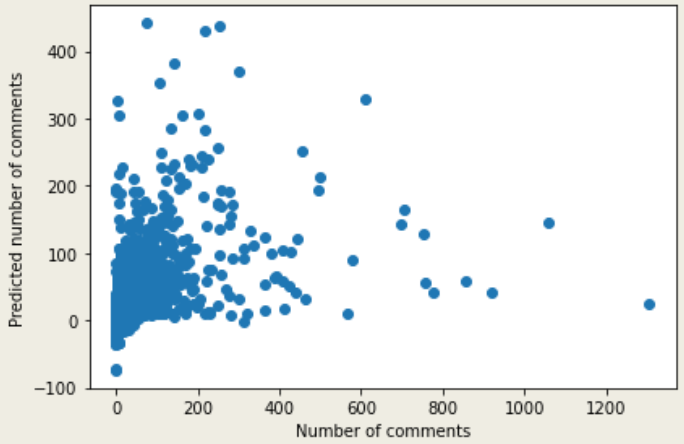




These results are quite obvious, but we can see that likes on a page influence both shares and comments. Also, comments and shares have an impact on one another.

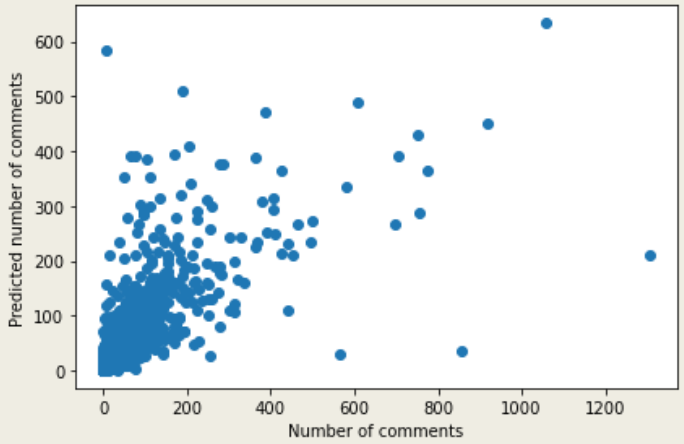
MODELS

Linear Regression



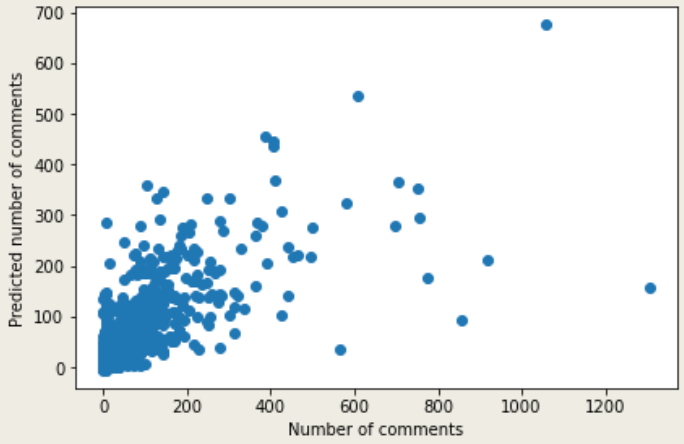
- Mean absolute error : 8.386173210429627
- Root mean square error : 30.11697659099272
- R2 Score : 0.27298656832423984

Random Forest



- Mean absolute error : 4.118192890983581
- Root mean square error : 22.32798947261331
- R2 Score : 0.600406026983124 6

Gradient Boosting



- Mean absolute error : 4.384396545587227
- Root mean square error : 22.102497302726626
- R2 Score : 0.6084363354661635

API

The API was done by using pickle and flask.

Just launch first the api.py, then the request.py.