Youtube

One of the sources we need to scrap was the Youtube website. We retrieve the comments related to plane and airlines videos. In this report, we're going to explain in a first time which data we scrap and the format of these data. In a second time, we're going to develop our scraping method and how we automate the code with a robot and we're going to finish with some statistics.

I Format of scraped data :

The table below summary all data we scraped and give their type and a description of what is the column.

Variable name	Type	Description
Data_Source	string	Source of scrapping
Date_Review	date	Date of comment
Review	string	Comment in english
Title	string	Vidéo title
Author	string	Video channel
Description	string	video description
Date_publication	date	Date of publication
View_Count	int	Number of view
Likes	int	Number of likes
Dislikes	int	Number of dislikes
Nb_subscribers	int	Number of subscriber
Nb_comments	int	Number of comment
hashtags	string	Hashtags of video

II Scraping method:

We scrape the data in several steps :

- Definition of the search equations
- For each of these equations, select videos with a publication date of "this week" if you want only the one week old videos. You can select all videos for each of these equations.
- For each comment of each video, we will retrieve and save the different variables in a dataframe.
- The dataframe will be exported in a json file.

In order to retrieve the comments, you have to go through phantonJs, so that you can "scroll" the web page. Indeed phantomJS will simulate the opening of a web page and the actions of a human being which consists in displaying the bottom of the page and update the youtube comments.

In order to retrieve the data we go through the htlm code of the page.

III Statistics:

All the statistics below are made on one week old videos.

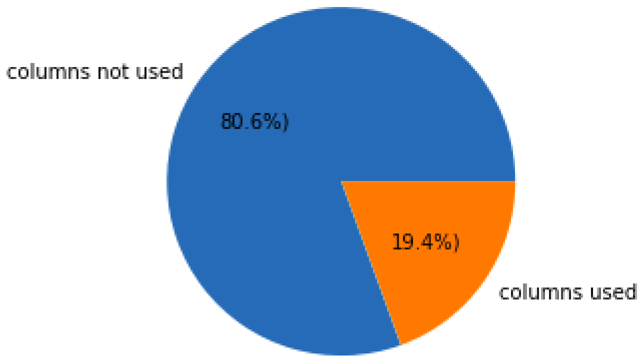
Being on youtube, all variables are normally filled except hashtags

we have 3,431 search equations, which gives us about 16 283 comments.

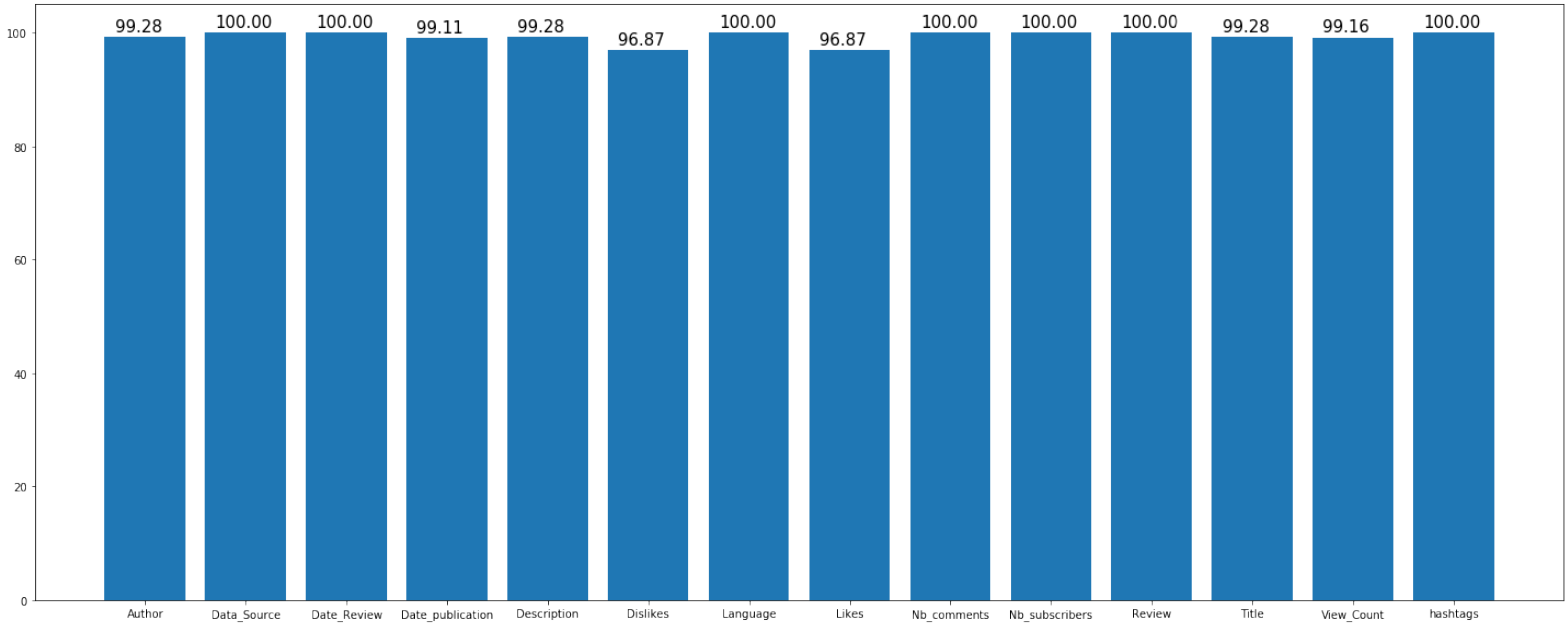
There is 3888 videos during the last week.

Statistics on every videos in the last 7 days

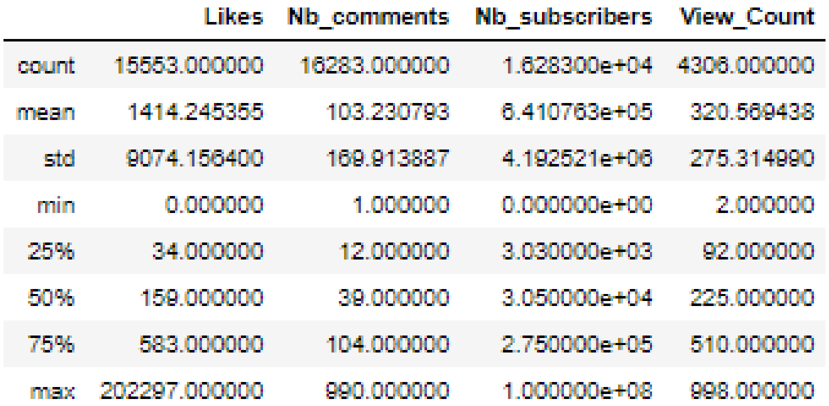
Rate columns used and no used



Rate of non-zero value in the columns used, in pourcentage



Description of the quantitative variable



	Likes	Nb_comments	Nb_subscribers	View_Count
count	15553.000000	16283.000000	1.628300e+04	4306.000000
mean	1414.245355	103.230793	6.410763e+05	320.569438
std	9074.156400	169.913887	4.192521e+06	275.314990
min	0.000000	1.000000	0.000000e+00	2.000000
25%	34.000000	12.000000	3.030000e+03	92.000000
50%	159.000000	39.000000	3.050000e+04	225.000000
75%	583.000000	104.000000	2.750000e+05	510.000000
max	202297.000000	990.000000	1.000000e+08	998.000000