Report : Streamroot Test
Hamza Jeddad

The project exists in the github :
`git clone https://github.com/Jeddadh/StreamRootProjectTest.git`

The repo contains:
- data_analysis.py : by executing this file you obtain different tables used in this document
- data_visualisation : by executing this file you obtain different data vizualisation used in this document

I. Tech choices & Installation requirements:

Tech choices:

To analyse and process the given data, I choose to use the **Python** library **Pandas** for many reasons:

- The **pandas** library is **open-source**
- It has a large set of statistical functions and methods.
- It is combined with python:
  - Gives us the possibility to visualize the data using **matplotlib**
  - Python has other advanced libraries to do mathematical and statistical calculus like **numpy** and **scipy**.
- I am used to work with Python.

Installation Requirements:

Install Python3
Install pandas (you can use `pip3 install panda` in the command line)
Install matplotlib (if it is not already installed with python3)

II. Method & data analysis

The goal is to increase the data downloaded using peer-to-peer (p2p).

NB: the first line is a comment. Pandas interprets the first line as a header (by default). I left the "#" so in this document and in the code, I use "#stream" instead of "stream"

1) Import the data and transform it to a pandas dataframe
   `stream_data = pd.DataFrame(pd.read_csv("data (1).csv"))`
2) Have a brief look at the data using :
   `stream_data.head()`
   which gives us a look at the first 5 rows of the data.

	#stream	isp	browser	connected	p2p	cdn
0	1	Fro	Iron	True	195910.840977	109025.960619
1	1	Fro	EarthWolf	True	186711.522041	113744.856814
2	1	Arange	Iron	True	189428.293434	115944.246844
3	1	Arange	Iron	True	0.000000	307577.191067
4	1	BTP	EarthWolf	True	207246.640473	107010.608093

3) Have a global look at the data using the function describe
`stream_data.describe()`

	#stream	isp	browser	connected	p2p	cdn
count	534954.0	534954	534954	534954	5.349540e+05	5.349530e+05
unique	9.0	5	4	2	NaN	NaN
top	3.0	Arange	EarthWolf	True	NaN	NaN
freq	100000.0	165341	283311	485753	NaN	NaN
mean	NaN	NaN	NaN	NaN	3.540061e+06	1.532365e+07
std	NaN	NaN	NaN	NaN	1.243739e+07	3.328645e+07
min	NaN	NaN	NaN	NaN	0.000000e+00	2.851538e+03
25%	NaN	NaN	NaN	NaN	0.000000e+00	7.726825e+04
50%	NaN	NaN	NaN	NaN	1.247840e+05	1.901855e+05
75%	NaN	NaN	NaN	NaN	1.941179e+05	2.071052e+05
max	NaN	NaN	NaN	NaN	5.249998e+07	1.049986e+08

- we notice that the column "cdn" has not the same number of components as the other columns, there is one missing data (534953 vs 534954) .To deal with this we delete all the rows that have missing data. We use the function : `stream_data.dropna(axis=0, how='any')`
- we can see the number of values that takes every variable.
    - 9 for #stream, 5 for isp, 4 for browser and 2 for connected
- We can notice also that:
    - The 3rd video is the most viewed
    - Arange is the most used isp
    - EarthWolf is the most used browser

4) Add 3 columns:
    a. data_downloaded = p2p + cdn represents the global data downloaded for every video content
    b. p2p_percentage = p2p/ data_downloaded : represents the percentage of the video downloaded using p2p (because video size is different from one video to an other)
    c. cdn_percentage = cdn/data_downloaded : represents the percentage of the video downloaded using cdn.

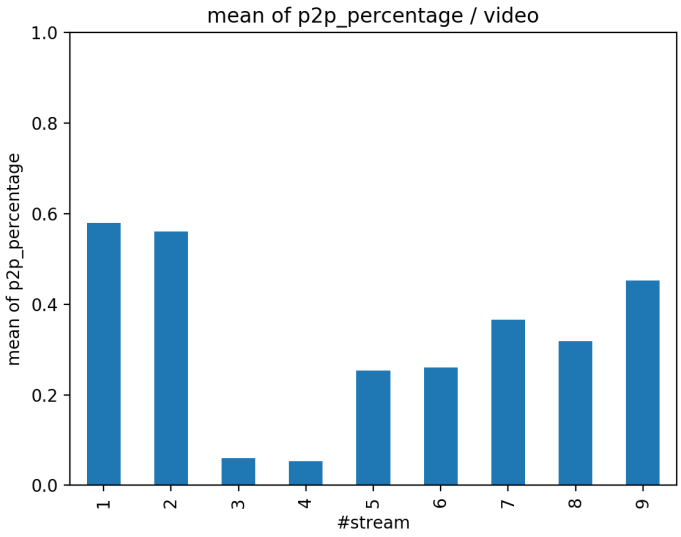
These 3 columns allow us to compare well and have a clearer view of the data.

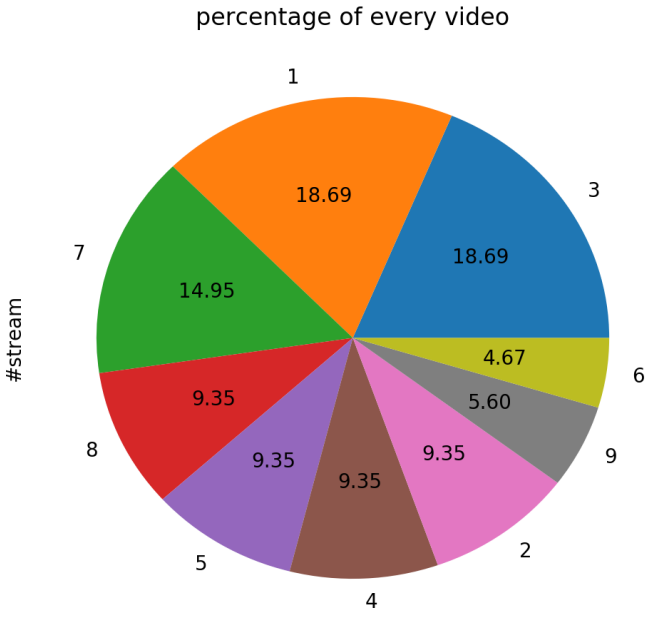
5) Analysis using every variable:
    a. **#stream**

	p2p_percentage	cdn_percentage
#stream		
1	0.578957	0.421043
2	0.560373	0.439627
3	**0.059510**	0.940490
4	**0.053937**	0.946063
5	0.253002	0.746998
6	0.260744	0.739256
7	0.366369	0.633631
8	0.318345	0.681655
9	0.452496	0.547504

The mean of p2p_percentage by video

- We notice that the The video 3 and 4 have the lower percentage of data downloaded using p2p. even if the video 3 is the most viewed (maybe the users are not connected in the same time or are not geographically near to others)





b. Connected

	p2p_percentage	cdn_percentage
connected		
False	**0.000000**	1.000000
True	0.355143	0.644857

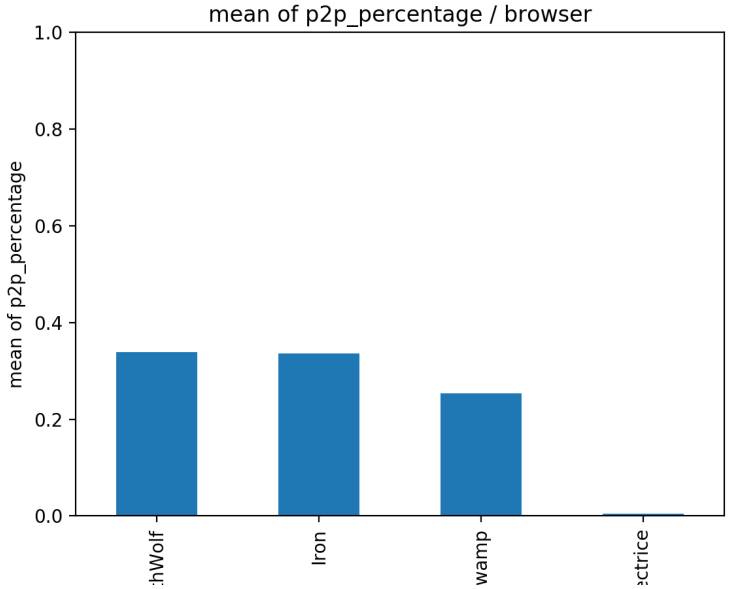
Mean of p2p_percentage and cdn_percentage by connected

- We notice that if the user is not connected to the backend, the data can't be downloaded using p2p.

c. **Browser :**

	p2p_percentage	cdn_percentage
browser		
EarthWolf	0.339323	0.660677
Iron	0.335398	0.664602
Swamp	0.254125	0.745875
Vectrice	**0.004848**	0.995152

The mean of p2p_percentage and cdn_percentage by browser

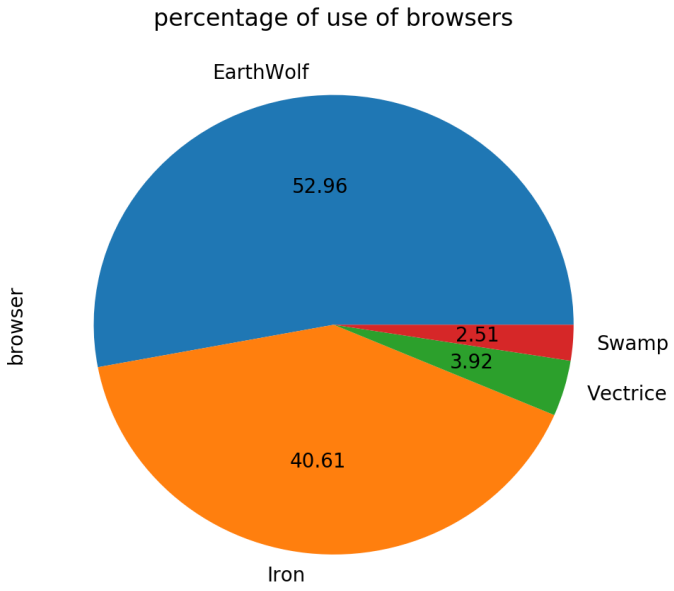


- The mean of p2p_percentage for Vectrice browser is very low compared to the tree other browsers even if the number of people who are connected to the backend is higher using this browser.

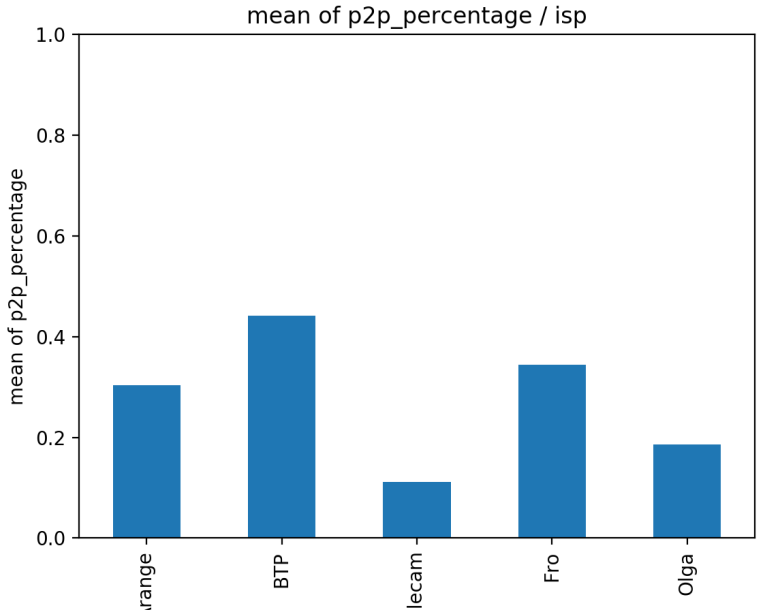
browser	EarthWolf	Iron	Swamp	Vectrice
connected				
False	0.084981	0.105162	0.095926	0.047258
True	0.915019	0.894838	0.904074	**0.952742**

The mean of p2p_percentage by connected by browser

- Fortunately, vectrice represents only 3,92% of used browsers :



d. Isp:



	p2p_percentage	cdn_percentage
isp		
Arange	0.303336	0.696664
BTP	0.441997	0.558003
Datch Telecam	**0.112094**	0.887906
Fro	0.343874	0.656126
Olga	**0.185440**	0.814560

The mean of p2p_percentage and cdn_percentage by isp

- The mean of p2p_percentage for Datch Telecam and olga isp is low compared to the tree other isps.

browser	EarthWolf	Iron	Swamp	Vectrice
isp				
Arange	0.316500	0.326638	0.263050	**0.004846**
BTP	0.470267	0.452252	0.299305	**0.006098**
Datch Telecam	**0.118046**	**0.112034**	**0.067333**	**0.000000**
Fro	0.356667	0.366827	0.252700	**0.004858**
Olga	**0.191660**	**0.170468**	**0.259887**	**0.001481**

The mean of p2p_percentage by browser and by isp

- Using this table we can see also that it seems that there is no relation between the browser and the isp (the percentage for Vertice is always lower independently of the isp / the percentage for Datch Telecam and Olga is always lower independently of the browser)

III. Data-driven Recommendations

By analyzing the given data, I recommend the next:

- Try to find why the p2p_percentage of the 3rd video is not very high even if it is the most viewed.
- Explore the technology used in vectrice browser to find out what hinders the downloading using p2p, and try to adapt your technology to that of vectrice to cover as many users as possible.
- We notice that p2p technology may not be supported by Datch Telecom, for that it serves its clients directly using cdn network. Thus, we can suppose that the clients of this isp are not satisfied due to the problems of cdn network(low debit , interruption of service...) which maybe justify the low use of this isp. Maybe, you can propose your service to this isp to increase the number of its users.