Videos Analysis

Assumptions:

- We consider the data as one batch, I didn't commit cohort analysis. For example: I didn't look at the 'video language' column as a proxy for geographical location but as a feature.
- There is no meaning for time as time, that's mean that if a video in the last month (extract from the data) didn't achieve any views it shouldn't matter for our rule of choice.

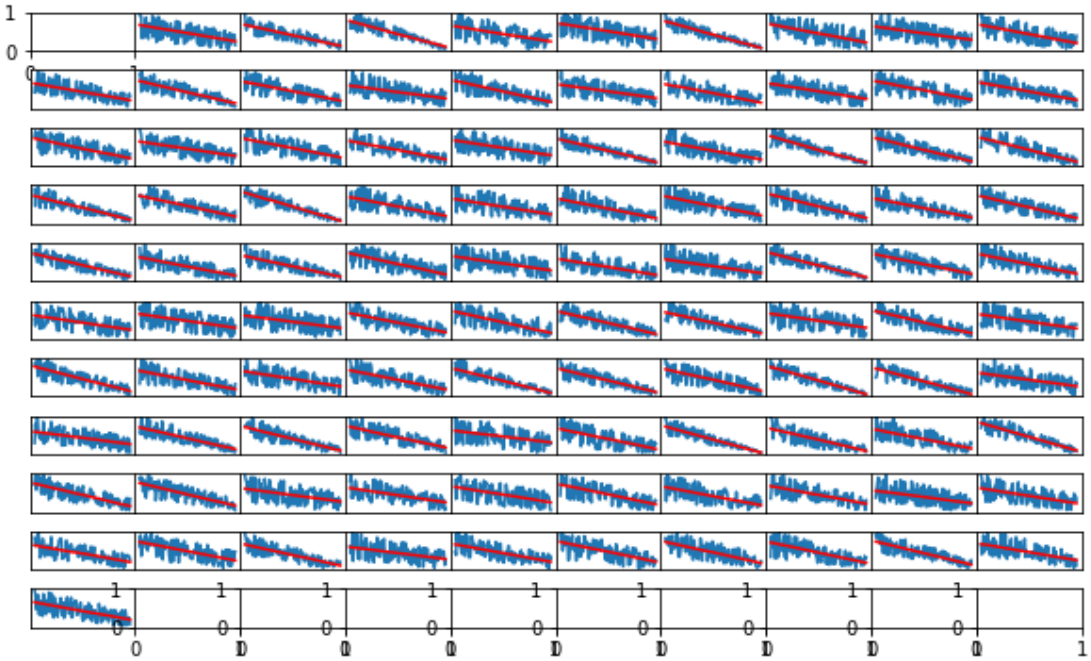
The analysis:

I looked at the data from two point of views:

- more toward general analysis – it can help you to understand how the videos behave, which feature is getting more views.
- Then I tried to cluster the videos in regard the information I gathered.

General analysis:

Let's see how views over time behave for each video:

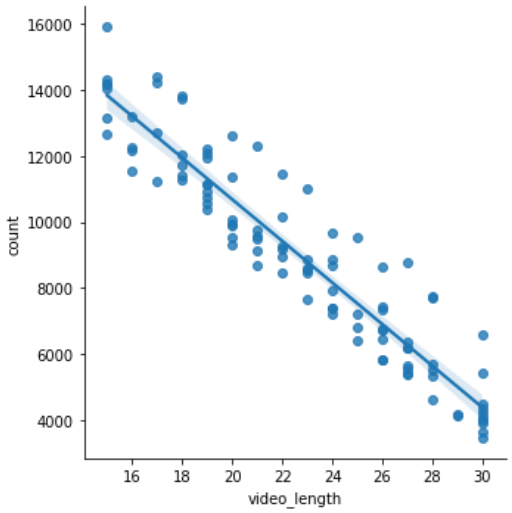


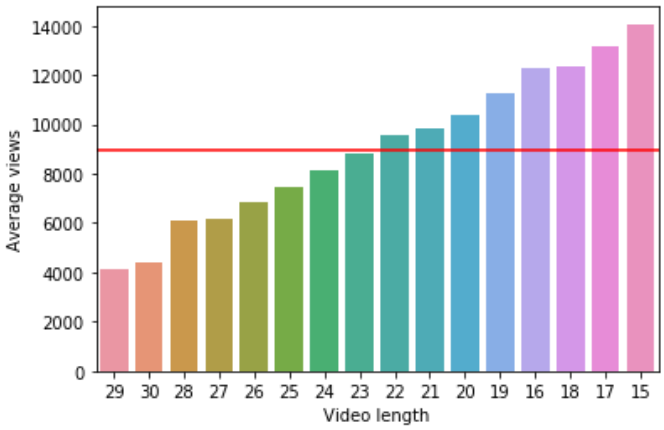
In this graph we see each video (1-100) daily views over time with linear line trend.

We can see a negative trend for all the videos – views drop over time.

We can also see that there is a lot of "spikes" in views, which is mean that views can be very unstable. We will need to find a way to take it under consideration.

Views by video length:



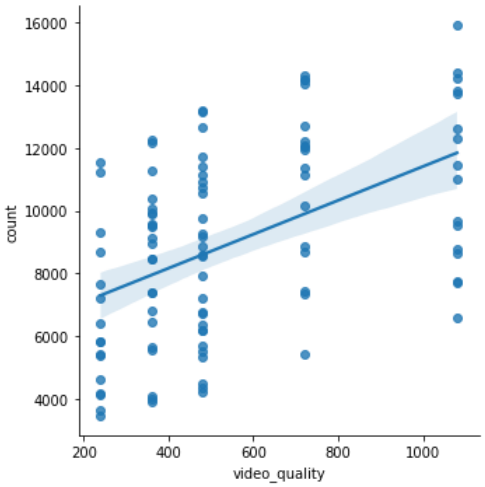


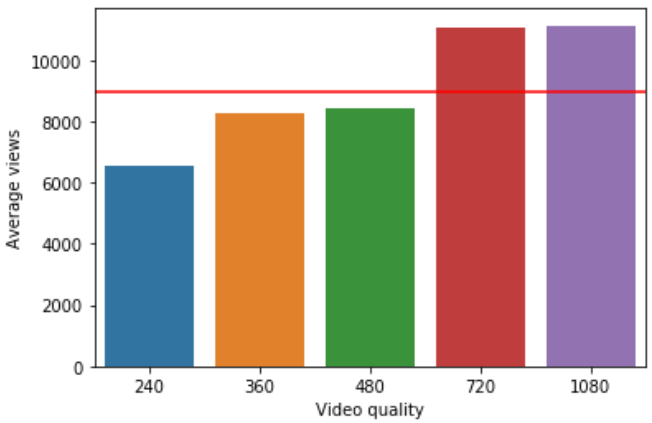
We can see negative correlation between video length and views.

The red line represents the mean views by video length, 15-22 seconds videos are above the mean.

Shorter videos have more views on average.

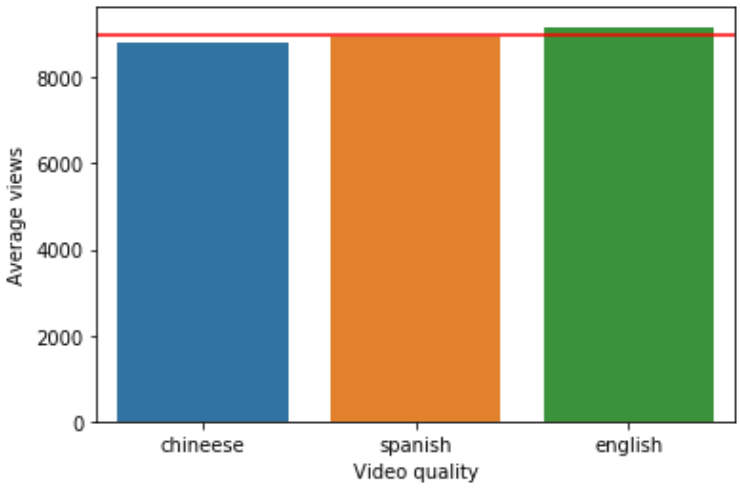
Views by video quality:





We can see positive correlation between video quality and views. The red line represents the mean views by video quality, 720P and 1080P videos are above the mean. High quality videos have more views on average.

Views by video language:



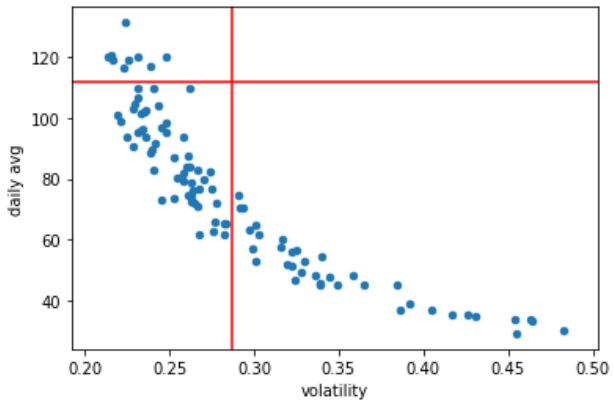
There isn't big difference in average views between different language.

Cluster:

as we seen earlier the views for each video unstable and have lots of "spikes", its mean that some videos have larger STD from others regard average views.

In the next method I tried to take this under consideration, and I tried to normalize the data. I used volatility and average daily views as my measures.

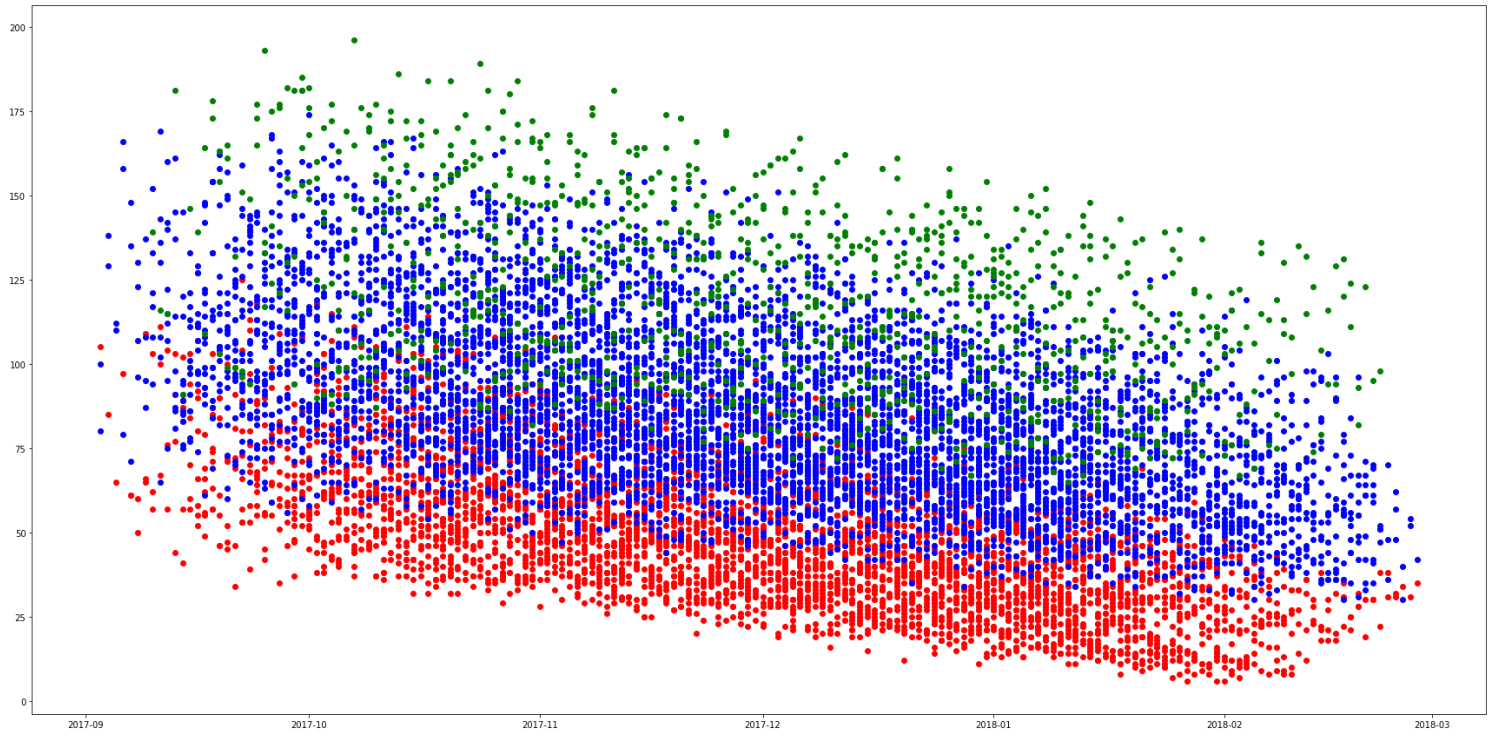
I defied volatility as: STD/ daily avg views for same video. It's a method that spread the risk over the average daily views.



- The videos in the top left considering videos with good performance and good stability.

- The videos in the bottom right considering videos with bad performance and bad stability.
- The videos in the bottom left considering videos with medium performance and good stability.

Here is a graph of the observations by daily views over time split into the three groups according to volatility and average daily views that I suggested earlier.



Next, after I cluster each video to one of our three groups, I want to see if the classification stands together with the general analysis and if it make sense.

Results:

Else:

video_id	video_length	video_quality
2	27	480
3	30	240
6	30	360
11	26	480
14	23	480
25	30	480
27	30	360
28	27	360
29	27	360
30	28	240
31	26	480
32	29	240
37	27	480
38	25	240
39	24	240
40	28	480
42	26	240
47	29	240
55	25	360
56	27	480
58	24	360
60	26	360
64	30	480
65	30	720
66	24	720
67	30	360
68	27	240
71	28	480
72	27	240
73	25	480
76	30	240
77	30	1080
79	30	480
80	23	240
81	23	360
92	26	240
98	28	480

stable:

video_id	video_length	video_quality
1	16	480
5	19	720
7	15	480
8	19	480
9	18	480
10	22	360
12	22	360
13	18	720
15	23	1080
16	24	1080
17	22	1080
18	16	360
19	24	480
20	20	360
21	18	360
22	21	360
23	26	720
24	20	720
26	22	480
33	21	360
34	17	720
35	24	720
36	21	480
41	24	360
43	20	360
45	19	720
46	19	480
48	27	1080
49	28	1080
50	19	480
53	25	1080
54	15	480
59	16	240
61	23	480
62	19	720
63	28	1080
69	16	360
70	18	480
75	22	720
78	26	720

Hot:

video_id	video_length	video_quality
4	15	720
44	17	1080
51	18	1080
52	17	1080
57	15	1080
74	15	720
82	15	720
84	15	720
93	18	1080

Do the results make sense?

Hot videos group:

We can see that all the videos have high quality feature and they are short – it makes sense from the general analysis we done earlier. This is what we are expecting from this group to look like.

Stable videos group:

We can see that in most of the group if a video has feature of high quality, the same video has also medium- long duration. This is also what we are expecting to see in this group.

Else videos group:

Mostly low quality and very long duration videos, this is looking right.

Overall, I can say that the classification is valid!