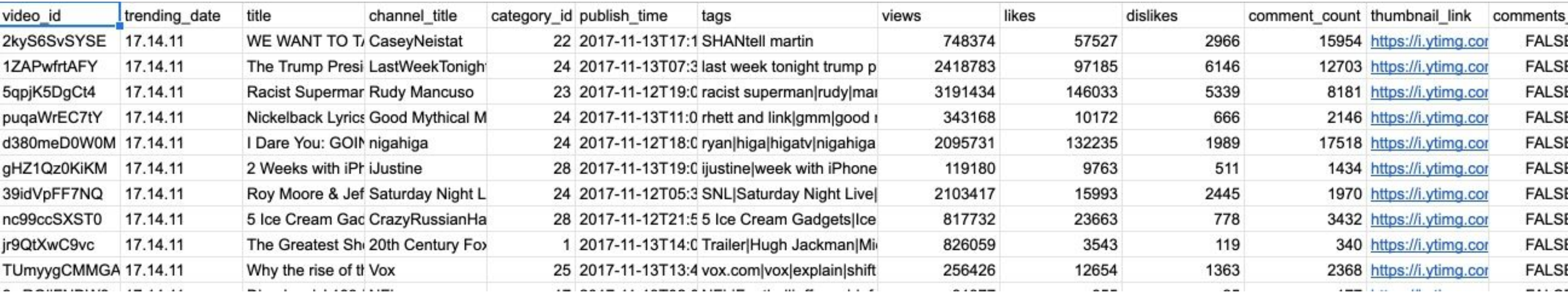
Trending Youtube Videos Analysis

Anna Atlasova

Data

Youtube trending videos in US



video_id	trending_date	title	channel_title	category_id	publish_time	tags	views	likes	dislikes	comment_count	thumbnail_link	comments_
2kyS6SvSYSE	17.14.11	WE WANT TO TA	CaseyNeistat	22	2017-11-13T17:1	SHANtell martin	748374	57527	2966	15954	https://i.ytimg.cor	FALSE
1ZAPwfrtAFY	17.14.11	The Trump Presi	LastWeekTonight	24	2017-11-13T07:3	last week tonight trump p	2418783	97185	6146	12703	https://i.ytimg.cor	FALSE
5qpjK5DgCt4	17.14.11	Racist Supermar	Rudy Mancuso	23	2017-11-12T19:0	racist superman|rudy|mar	3191434	146033	5339	8181	https://i.ytimg.cor	FALSE
puqaWrEC7tY	17.14.11	Nickelback Lyrics	Good Mythical M	24	2017-11-13T11:0	rhett and link|gmm|good r	343168	10172	666	2146	https://i.ytimg.cor	FALSE
d380meD0W0M	17.14.11	I Dare You: GOIN	nigahiga	24	2017-11-12T18:0	ryan|higa|higatv|nigahiga	2095731	132235	1989	17518	https://i.ytimg.cor	FALSE
gHZ1Qz0KiKM	17.14.11	2 Weeks with iPh	iJustine	28	2017-11-13T19:0	ijustine|week with iPhone	119180	9763	511	1434	https://i.ytimg.cor	FALSE
39idVpFF7NQ	17.14.11	Roy Moore & Jef	Saturday Night L	24	2017-11-12T05:3	SNL|Saturday Night Live|	2103417	15993	2445	1970	https://i.ytimg.cor	FALSE
nc99ccSXST0	17.14.11	5 Ice Cream Gad	CrazyRussianHa	28	2017-11-12T21:5	5 Ice Cream Gadgets|Ice	817732	23663	778	3432	https://i.ytimg.cor	FALSE
jr9QtXwC9vc	17.14.11	The Greatest Sh	20th Century Fox	1	2017-11-13T14:0	Trailer|Hugh Jackman|Mi	826059	3543	119	340	https://i.ytimg.cor	FALSE
TUmyygCMMGA	17.14.11	Why the rise of th	Vox	25	2017-11-13T13:4	vox.com|vox|explain|shift	256426	12654	1363	2368	https://i.ytimg.cor	FALSE

Problem

One video can have several rows in the dataset (several days on trend).

New data

Each row - a unique video.

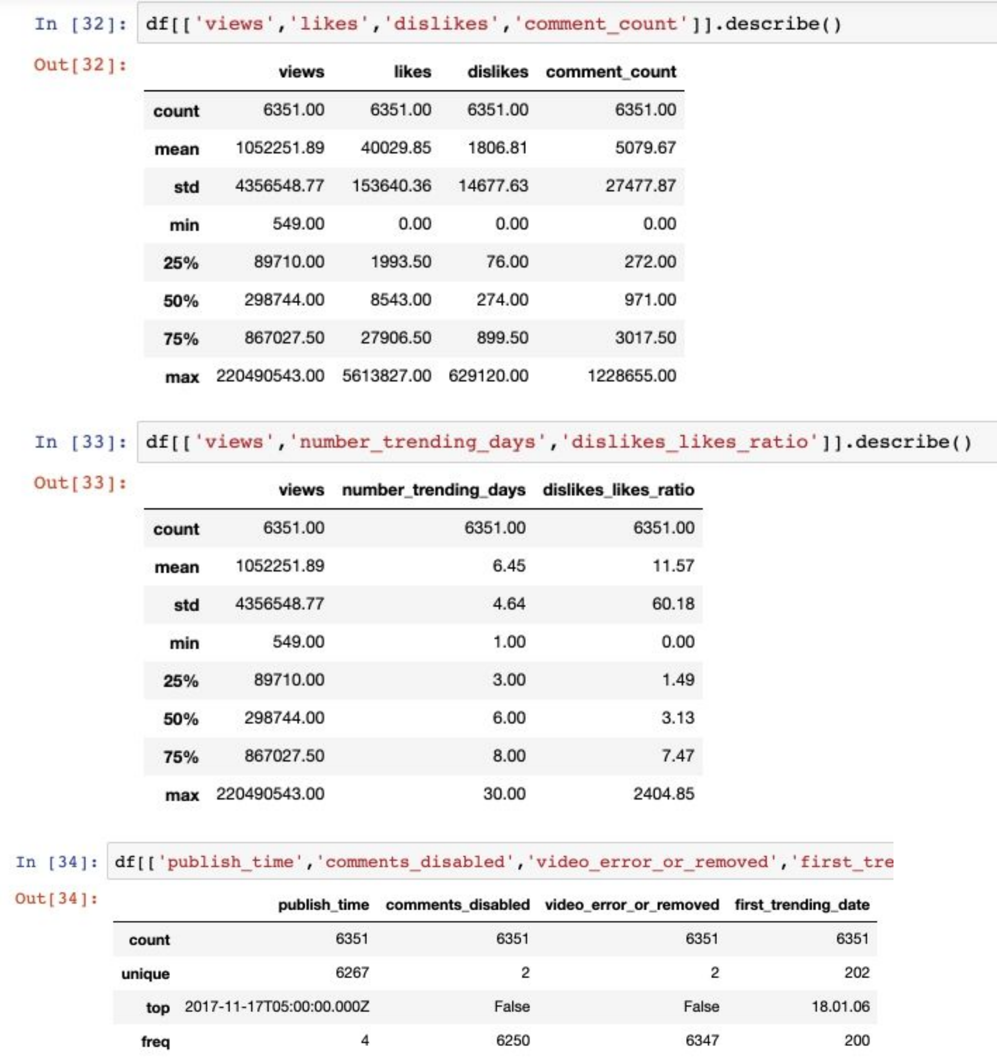
Some new columns were added.

New data

We have 6351 videos in our dataset which were on trend. Videos have data in 18 columns:

- video_id - id of a video which is on trend (unique value)
- title - title of the video
- channel_title - title of the youtube channel of this video
- category_id - category of the video
- publish_time - datetime when the video was published
- tags - tags in the video
- views - # views of the video on the first trending date
- likes - # likes of the video on the first trending date
- dislikes - # dislikes of the video on the first trending date
- comment_count - # comments of the video on the first trending date
- thumbnail_link - link to the preview image of the video
- comments_disabled - if comments of the video were disabled on the first trending date
- ratings_disabled - if likes/dislikes of the video were disabled on the first trending date
- video_error_or_removed - if video was removed on the first trending date
- description - description of the video on the first trending date
- number_trending_days - # days the video was on trend
- first_trending_date - date when the video became on trend
- dislikes_likes_ratio - percentage of dislikes vs likes

New data



In [32]:
```
df[['views','likes','dislikes','comment_count']].describe()
```

Out[32]:

	views	likes	dislikes	comment_count
count	6351.00	6351.00	6351.00	6351.00
mean	1052251.89	40029.85	1806.81	5079.67
std	4356548.77	153640.36	14677.63	27477.87
min	549.00	0.00	0.00	0.00
25%	89710.00	1993.50	76.00	272.00
50%	298744.00	8543.00	274.00	971.00
75%	867027.50	27906.50	899.50	3017.50
max	220490543.00	5613827.00	629120.00	1228655.00

In [33]:
```
df[['views','number_trending_days','dislikes_likes_ratio']].describe()
```

Out[33]:

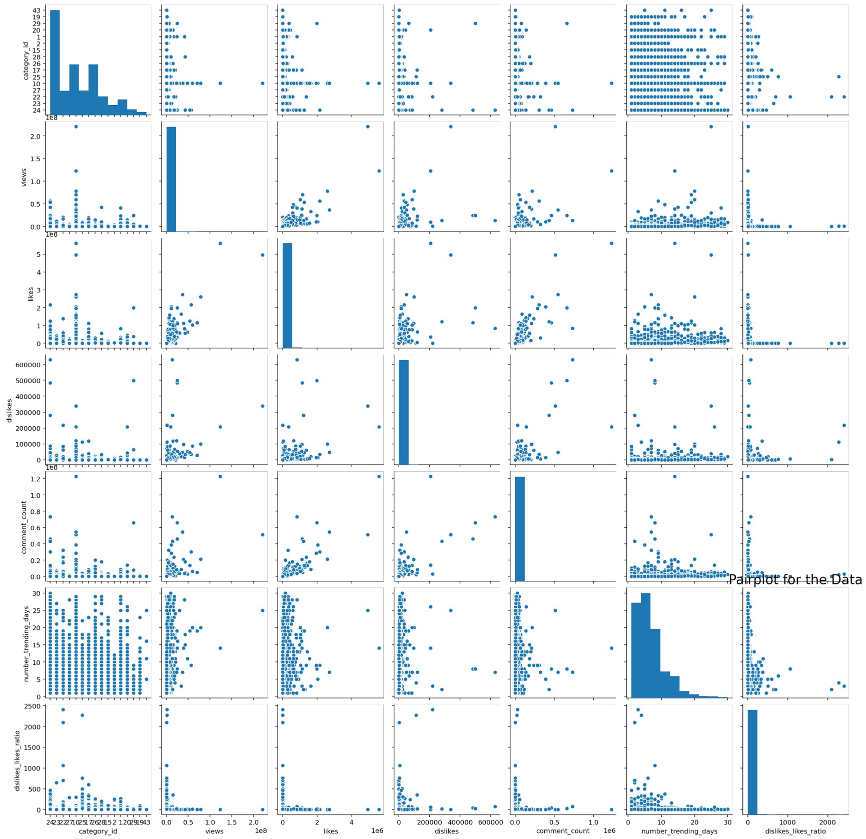
	views	number_trending_days	dislikes_likes_ratio
count	6351.00	6351.00	6351.00
mean	1052251.89	6.45	11.57
std	4356548.77	4.64	60.18
min	549.00	1.00	0.00
25%	89710.00	3.00	1.49
50%	298744.00	6.00	3.13
75%	867027.50	8.00	7.47
max	220490543.00	30.00	2404.85

In [34]:
```
df[['publish_time','comments_disabled','video_error_or_removed','first_tre
```

Out[34]:

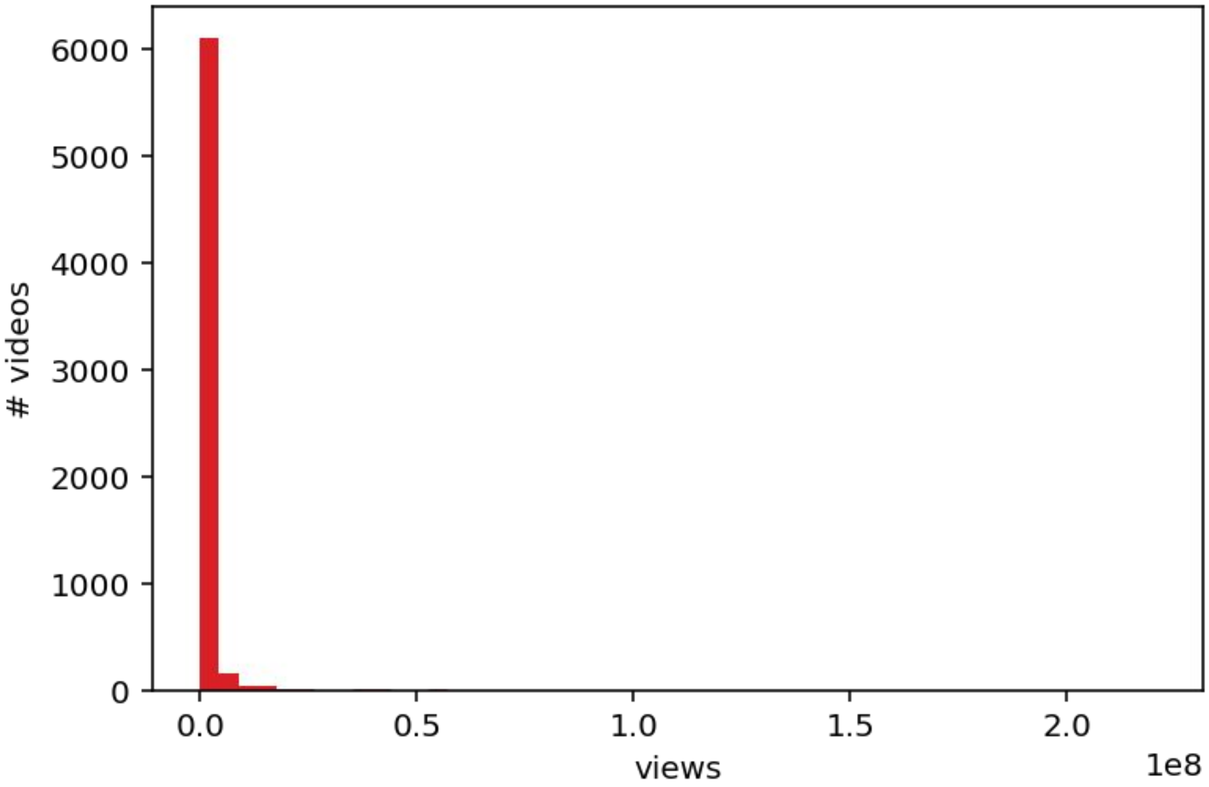
	publish_time	comments_disabled	video_error_or_removed	first_trending_date
count	6351	6351	6351	6351
unique	6267	2	2	202
top	2017-11-17T05:00:00.000Z	False	False	18.01.06
freq	4	6250	6347	200

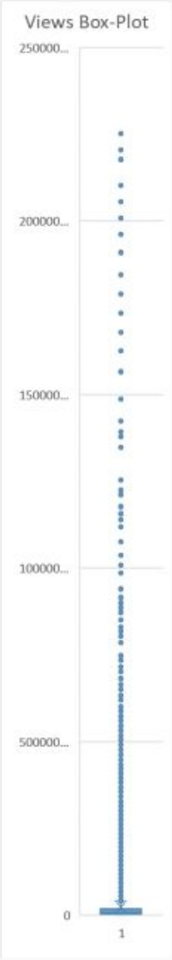
Histograms and scatter-plots of numeric vars

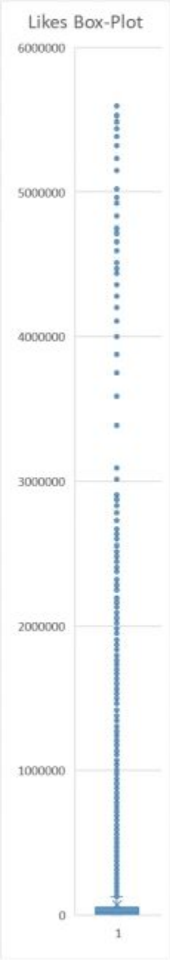


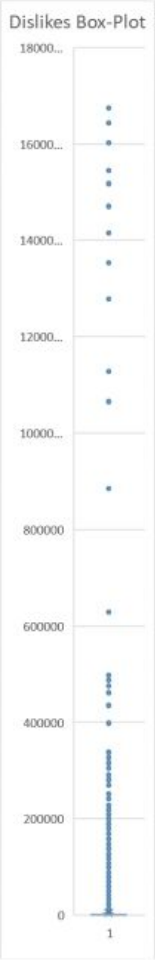
Let’s have a close look at one of the histograms

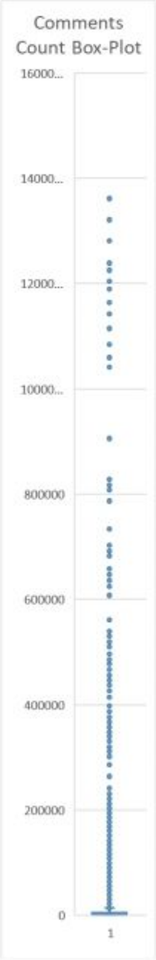
views histogram



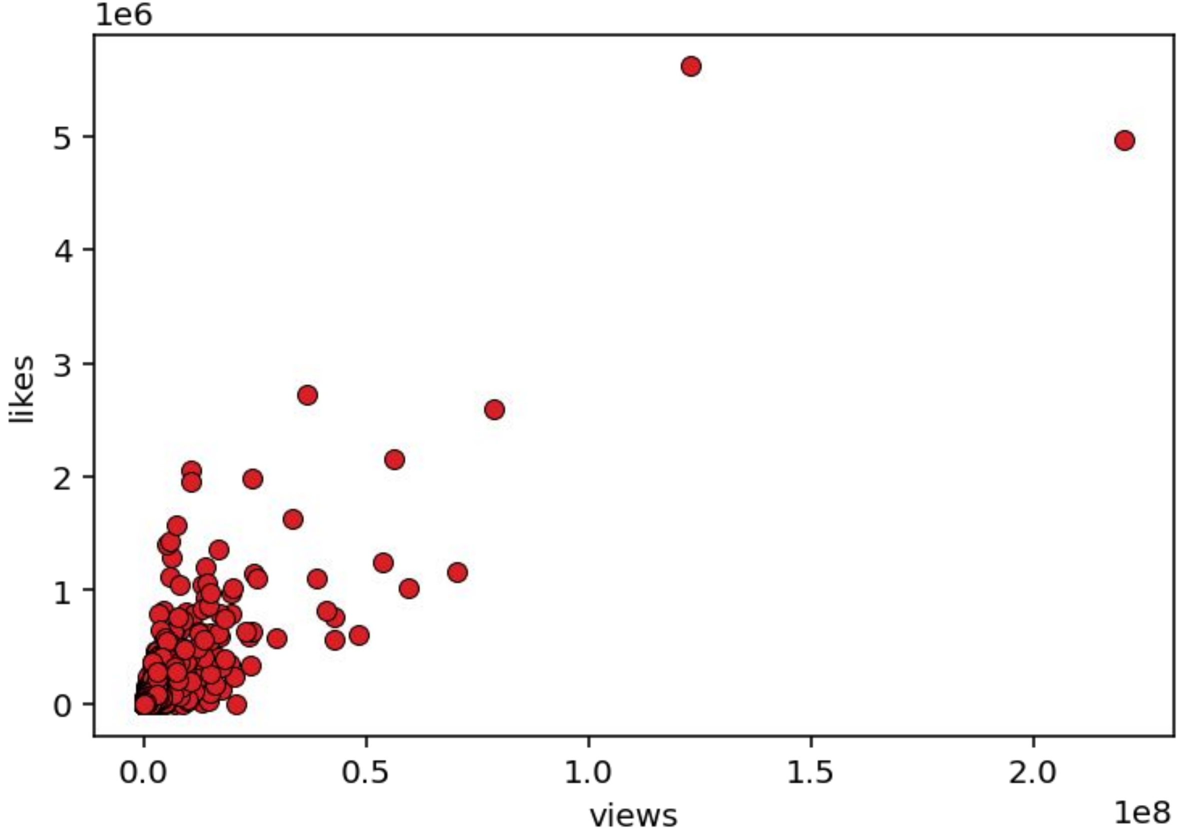

Views Box-Plot
250000...
200000...
150000...
100000...
500000...
0
1


Likes Box-Plot
6000000
5000000
4000000
3000000
2000000
1000000
0
1


Dislikes Box-Plot
18000...
16000...
14000...
12000...
10000...
800000
600000
400000
200000
0
1


Comments Count Box-Plot
16000...
14000...
12000...
10000...
800000
600000
400000
200000
0
1

views/likes scatter-plot

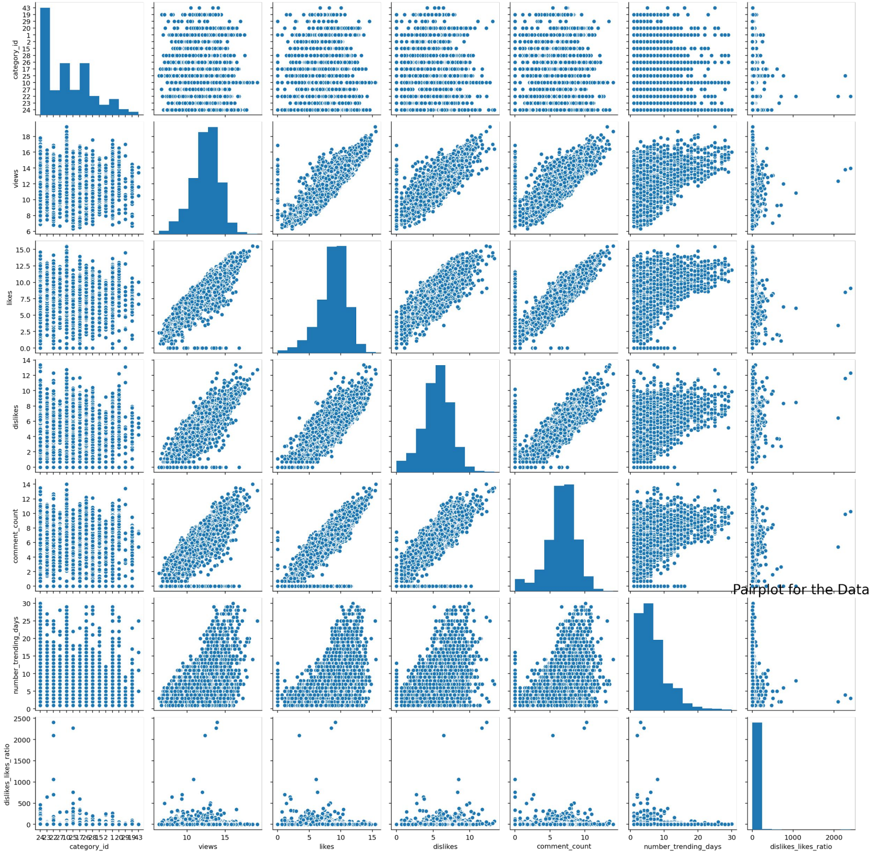


Insight

We could see that we have outliers - though, an analysis was made and without outliers the analysis was not so good still.

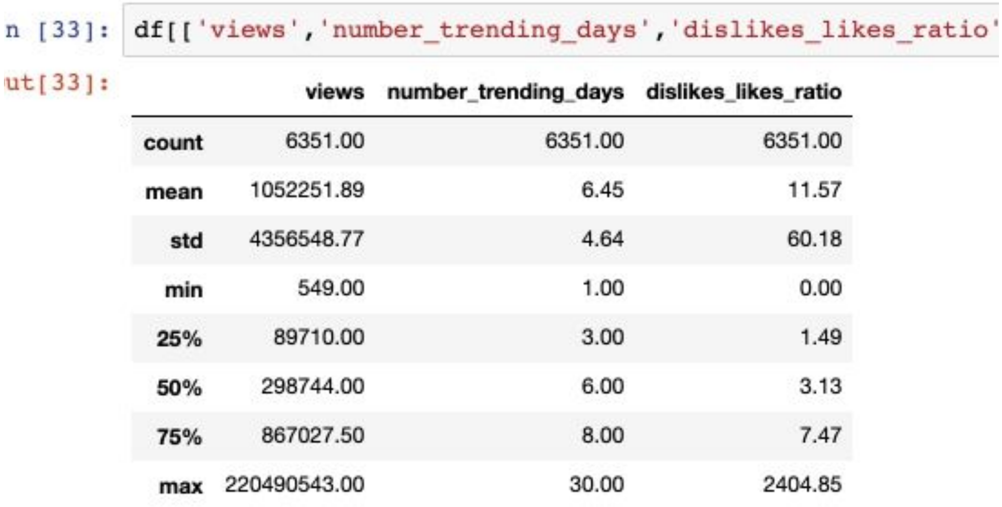
We can assume that the distribution is logarithmic. Let’s standardize it.

Standardized histograms and scatter-plots



Pairplot for the Data

Not standardized data

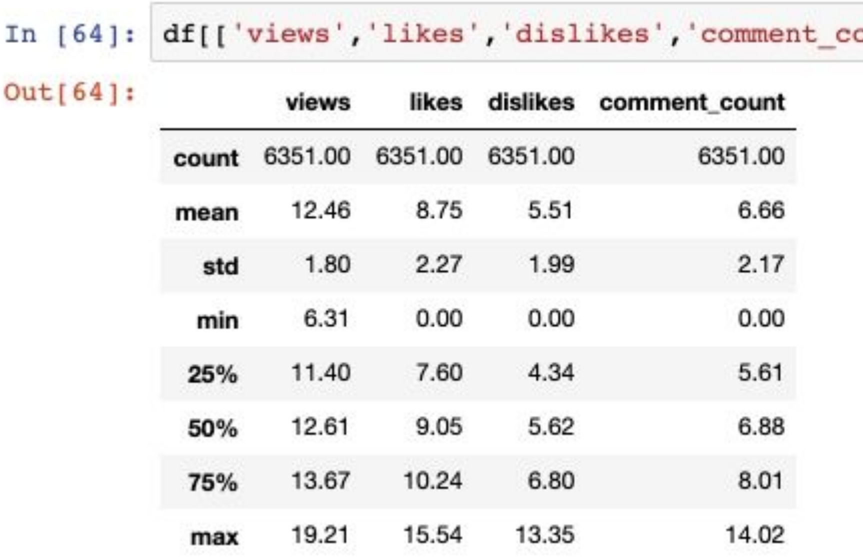


```
n [33]: df[['views','number_trending_days','dislikes_likes_ratio'
```

ut[33]:

	views	number_trending_days	dislikes_likes_ratio
count	6351.00	6351.00	6351.00
mean	1052251.89	6.45	11.57
std	4356548.77	4.64	60.18
min	549.00	1.00	0.00
25%	89710.00	3.00	1.49
50%	298744.00	6.00	3.13
75%	867027.50	8.00	7.47
max	220490543.00	30.00	2404.85

Standardized data

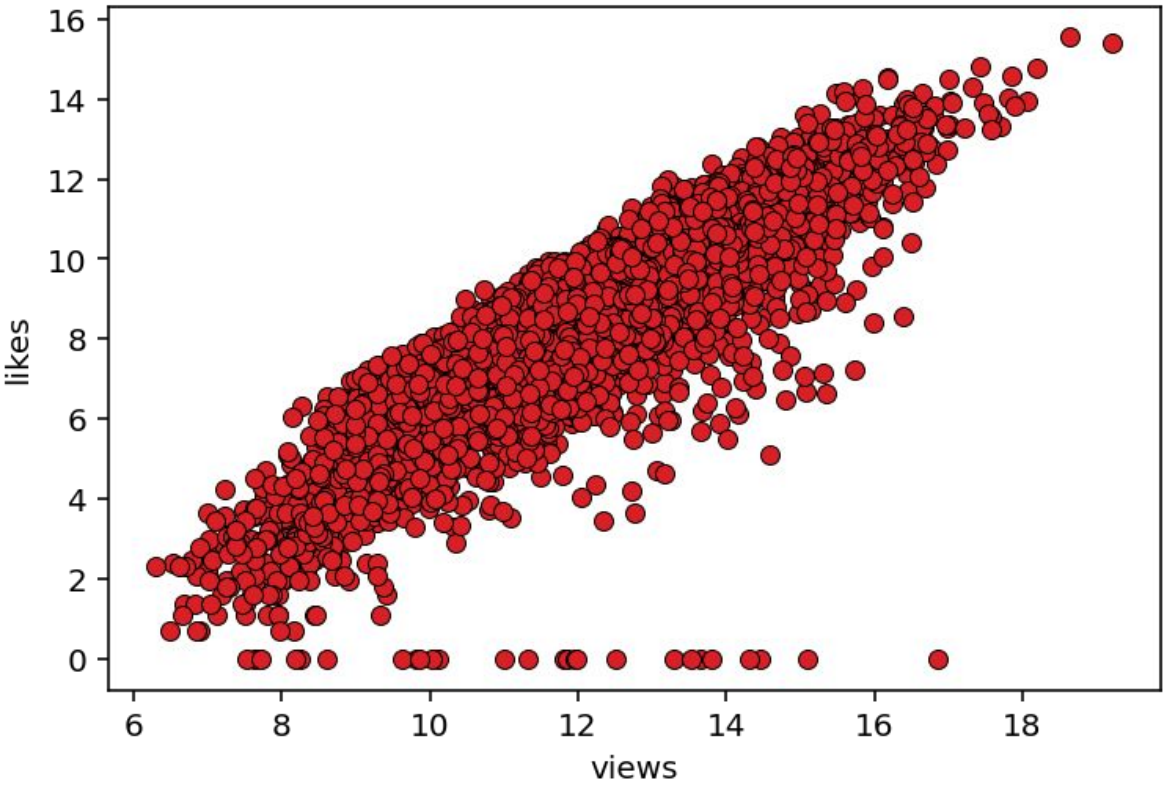


```
In [64]: df[['views','likes','dislikes','comment_co
```

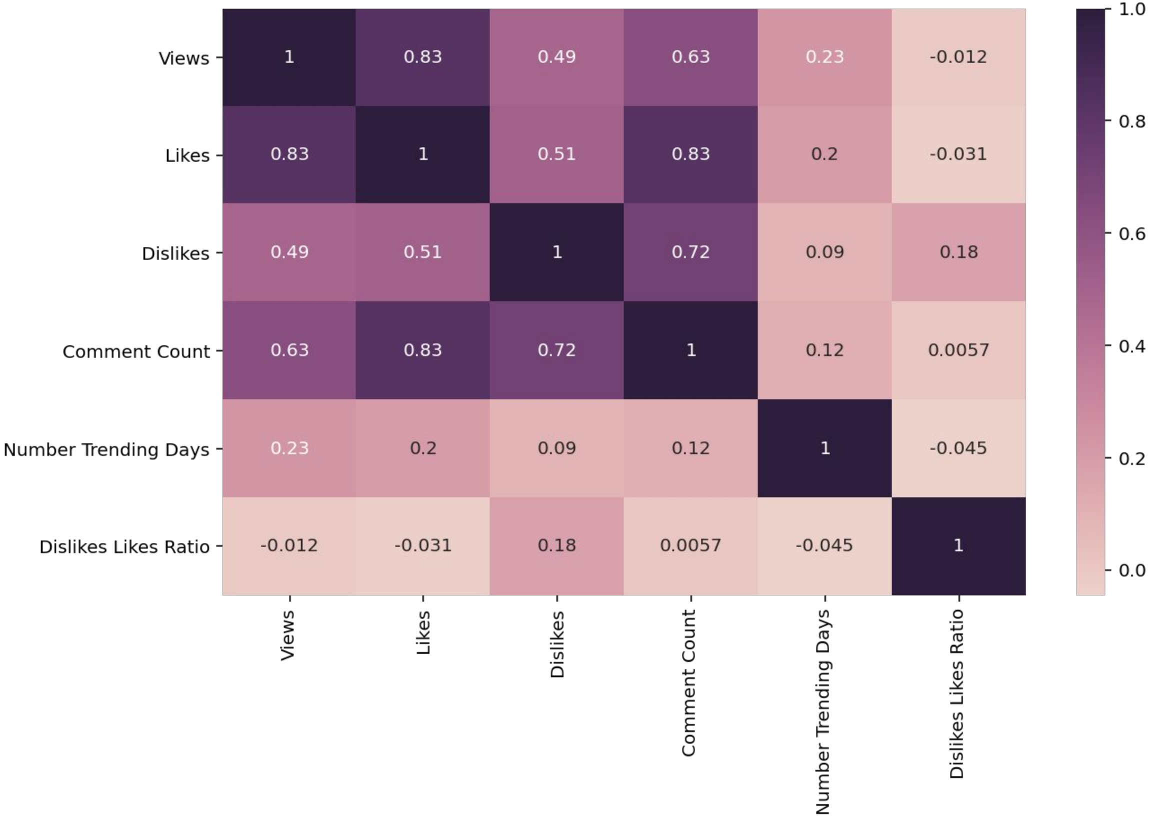
Out[64]:

	views	likes	dislikes	comment_count
count	6351.00	6351.00	6351.00	6351.00
mean	12.46	8.75	5.51	6.66
std	1.80	2.27	1.99	2.17
min	6.31	0.00	0.00	0.00
25%	11.40	7.60	4.34	5.61
50%	12.61	9.05	5.62	6.88
75%	13.67	10.24	6.80	8.01
max	19.21	15.54	13.35	14.02

views/likes st. scatter-plot

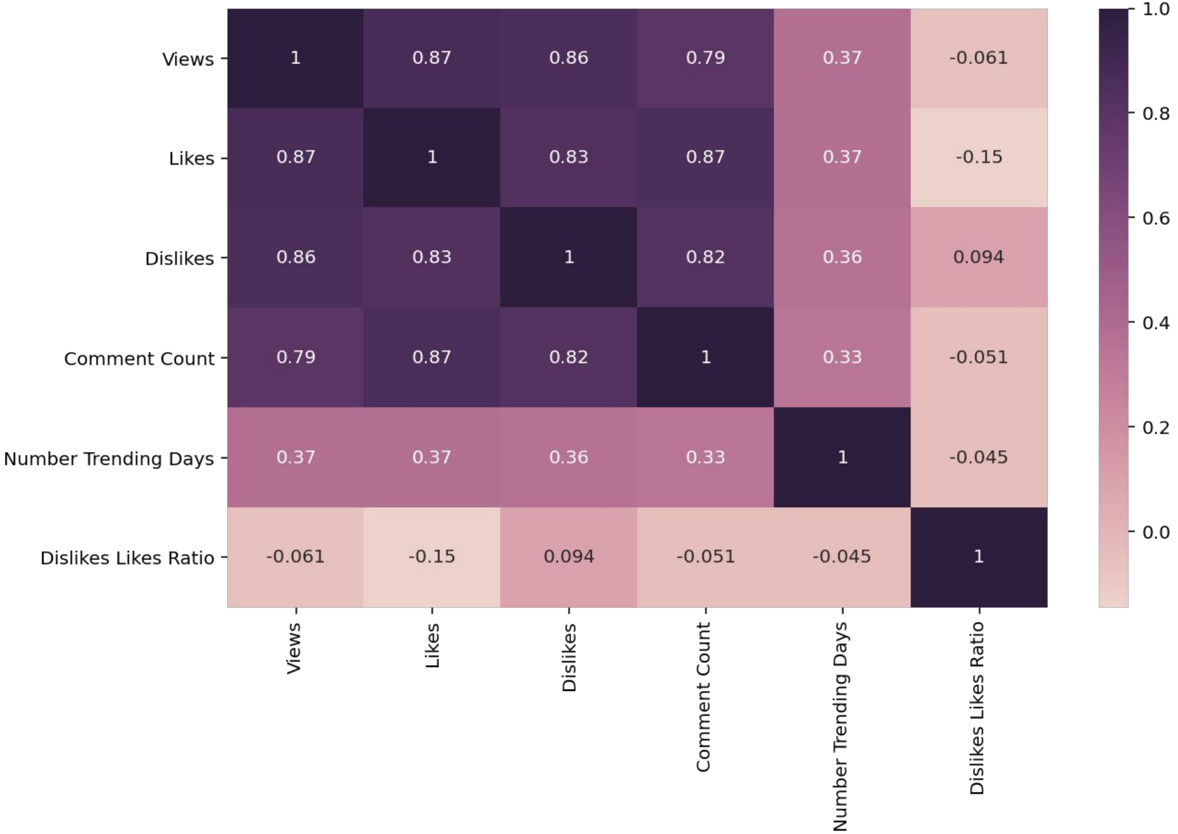


Linear correlations (without standartization)



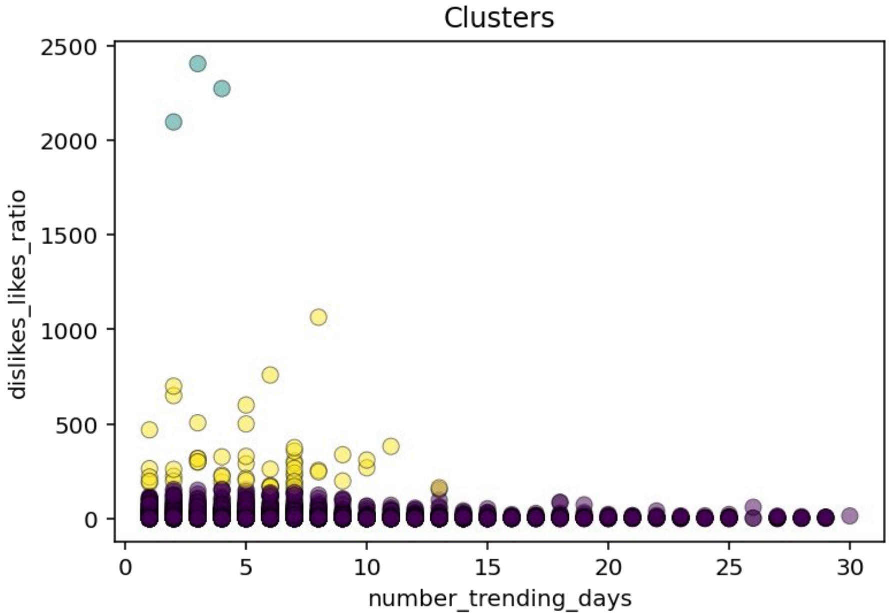
	Views	Likes	Dislikes	Comment Count	Number Trending Days	Dislikes Likes Ratio
Views	1	0.83	0.49	0.63	0.23	-0.012
Likes	0.83	1	0.51	0.83	0.2	-0.031
Dislikes	0.49	0.51	1	0.72	0.09	0.18
Comment Count	0.63	0.83	0.72	1	0.12	0.0057
Number Trending Days	0.23	0.2	0.09	0.12	1	-0.045
Dislikes Likes Ratio	-0.012	-0.031	0.18	0.0057	-0.045	1

Linear correlations (with standartization)



	Views	Likes	Dislikes	Comment Count	Number Trending Days	Dislikes Likes Ratio
Views	1	0.87	0.86	0.79	0.37	-0.061
Likes	0.87	1	0.83	0.87	0.37	-0.15
Dislikes	0.86	0.83	1	0.82	0.36	0.094
Comment Count	0.79	0.87	0.82	1	0.33	-0.051
Number Trending Days	0.37	0.37	0.36	0.33	1	-0.045
Dislikes Likes Ratio	-0.061	-0.15	0.094	-0.051	-0.045	1

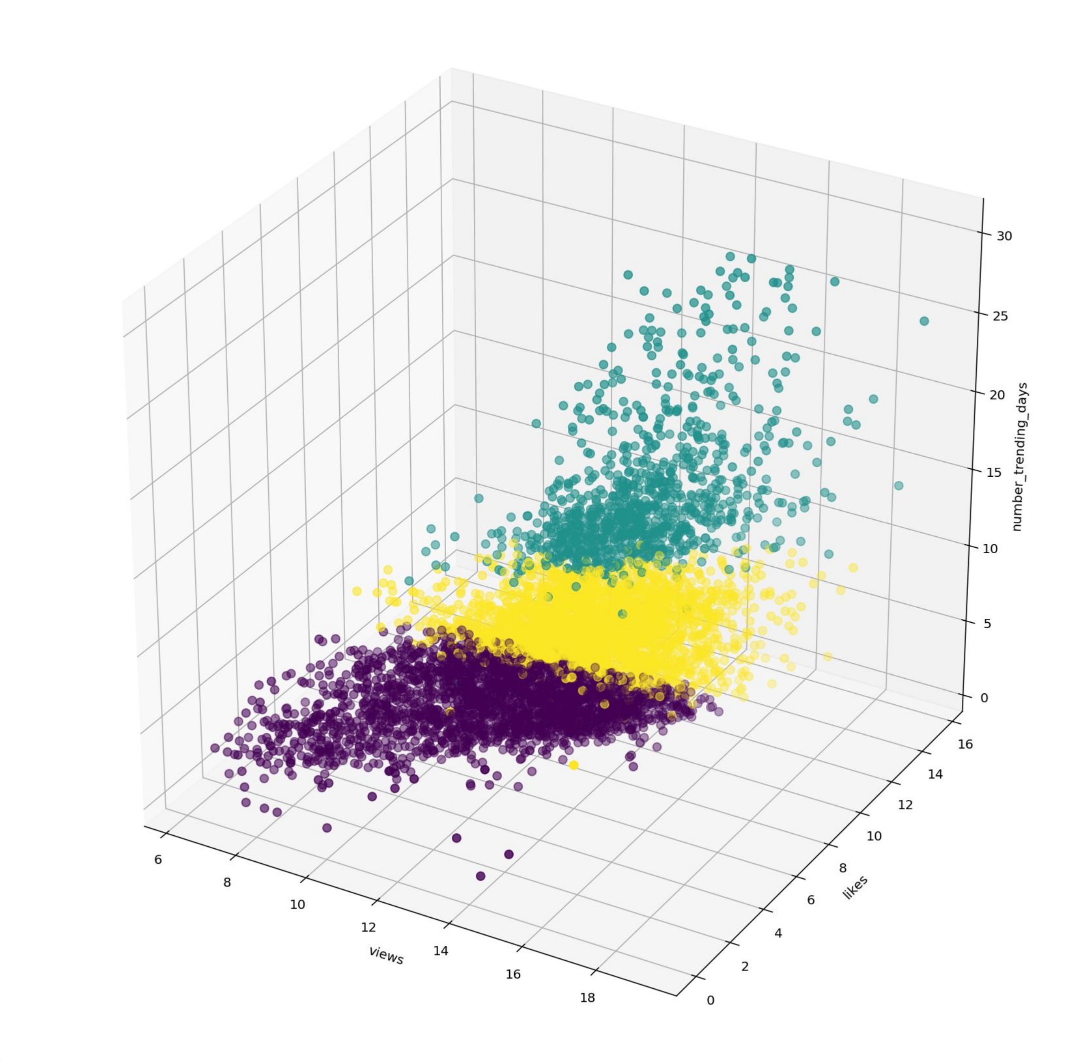
Cluster analysis



1 cluster (green) is "negative" videos which have a high dislikes/likes ratio and which are on trend not so many days

2 cluster (yellow) is "contreversial" videos which have a medium dislikes/likes ratio which can be on trend up to approximately 15 days,

3 cluster (purple) is "positive" videos which have a low dislikes/likes ratio which can have low and high number of trending days.



1 cluster (green) is "super popular" videos which have a high number of trending days, likes, views.

2 cluster (yellow) is "popular" videos which have a medium number of trending days, likes, views.

3 cluster (purple) is "not so popular" videos which have lower number of trending days, likes, views.

Results

There were “negative” videos on trend with high dislikes/likes ratio and usually these “negative” videos are on trend not so many days.

We can say that likes/views, likes/comment_count, dislikes/comment_count are correlated.