

# YouTube Trending Study

Berkeley Willis

# Introduction

## YouTube Trending

YouTube Trending is a tab that shows popular videos that are recommended due to their popularity in a certain area or country.

## Question

My initial question was, is it possible to predict what videos go into the trending tab? But this had limitations, and so instead I found a different analysis possible, can we predict number of views for a trending video.

## Hypothesis

I think that it is possible to predict the number of views for a trending video based on the number of likes, dislikes, comment counts, and age of the video.

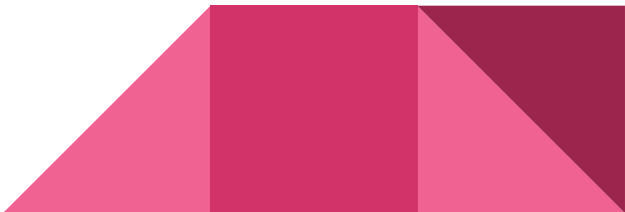
# Data Used

The data used in this project originated from kaggle: <https://www.kaggle.com/datasnaek/youtube-new>

What is contained is some basic statistical data on videos that appeared in a country's trending tab everyday for about 6 months.



# Variables Outline

- Likes - the number of registered YouTube users that have clicked the like button for a trending video.
  - Dislikes - the number of registered YouTube users that have clicked the dislike button for a trending video.
  - Comments Count - the number of registered YouTube users that have written a comment for a trending video.
  - Video Age - the age of a video from the date of upload to a day of it being in the trending tab.
  - Trending date - the date a video appears in trending
- 

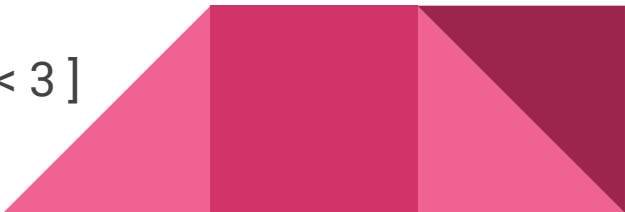
# Data Cleaning

The YouTube trending data is fairly clean but there are some issues of duplicates, and very large outliers. The primary reason for these outliers would be certain viral videos that get large number of views and other indicators of those views such as likes and comment counts.

The outliers will be dealt with by calculating the z-score for each of the columns and filtering those that have an absolute value of greater than 3.

$$z\text{-score} = \frac{y - \text{mean}(y)}{\text{std}(y)}$$

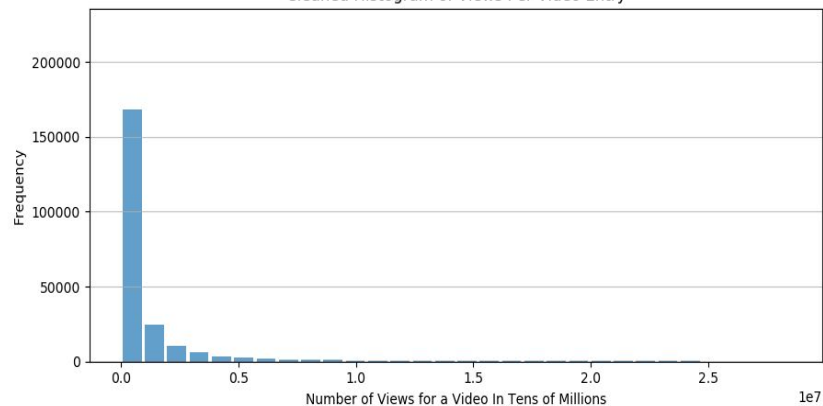
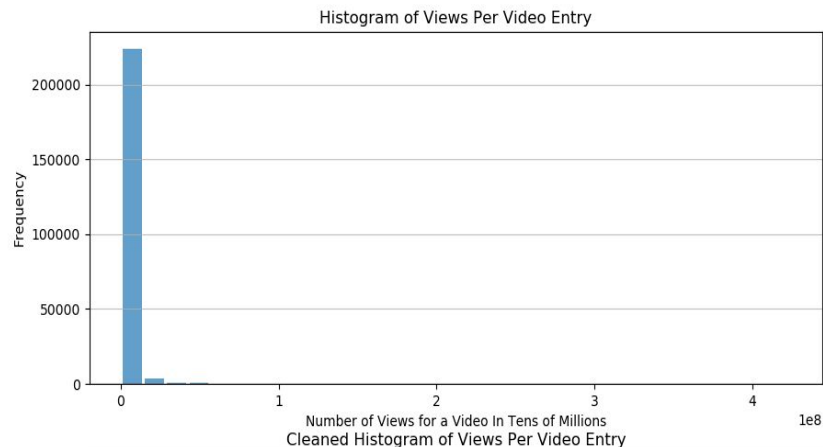
```
clean_pd = clean_pd [ abs(z-score) < 3 ]
```



# Views

Number of times that a video was viewed by a user, including anonymous users

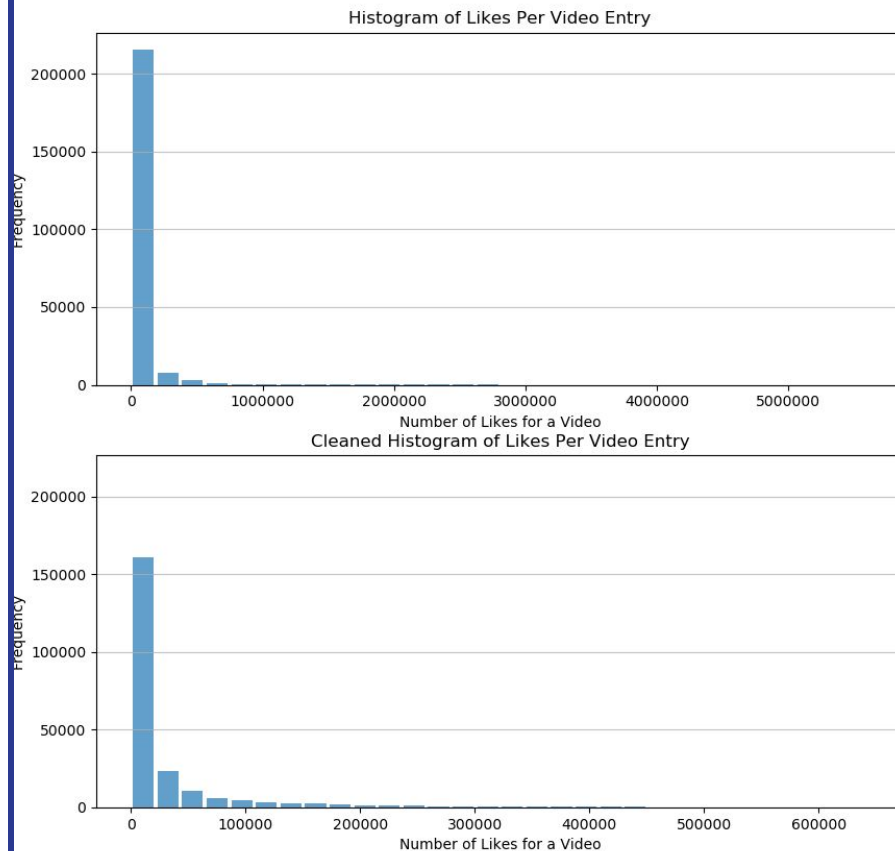
Mean	1.897883e+06
Mode	6573
Std. Dev.	8.876277e+06
Spread	This is not evenly distributed and still varies quite widely for trending videos.
Outliers	Outliers will far more views, likely due to viral videos.



# Likes

The number of registered YouTube users that clicked the like button on a video

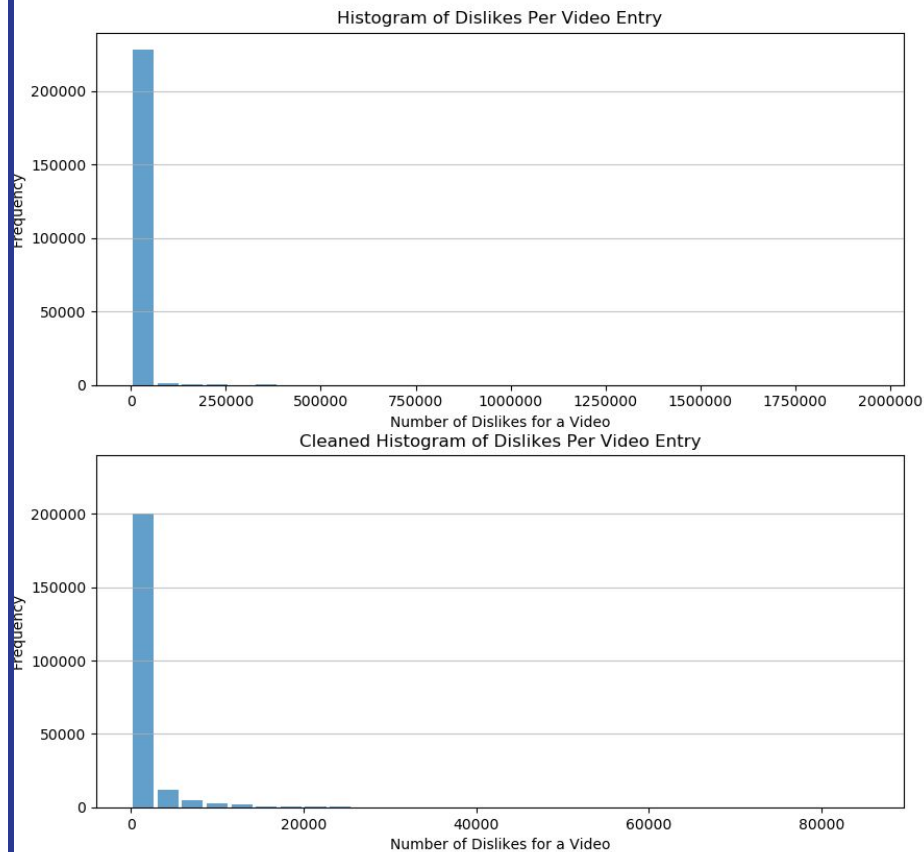
Mean	5.225978e+04
Mode	0
Std. Dev.	1.971921e+05
Spread	This is not evenly distributed and still varies quite widely for trending videos.
Outliers	Outliers will far more likes, likely due to popular viral videos.



# Dislikes

The number of registered YouTube users that clicked the dislike button on a video

Mean	2.297060e+05
Mode	0
Std. Dev.	2.756118e+04
Spread	This is not evenly distributed and still varies quite widely for trending videos.
Outliers	Outliers will far more dislikes, likely due to viral videos with a lot of hate.

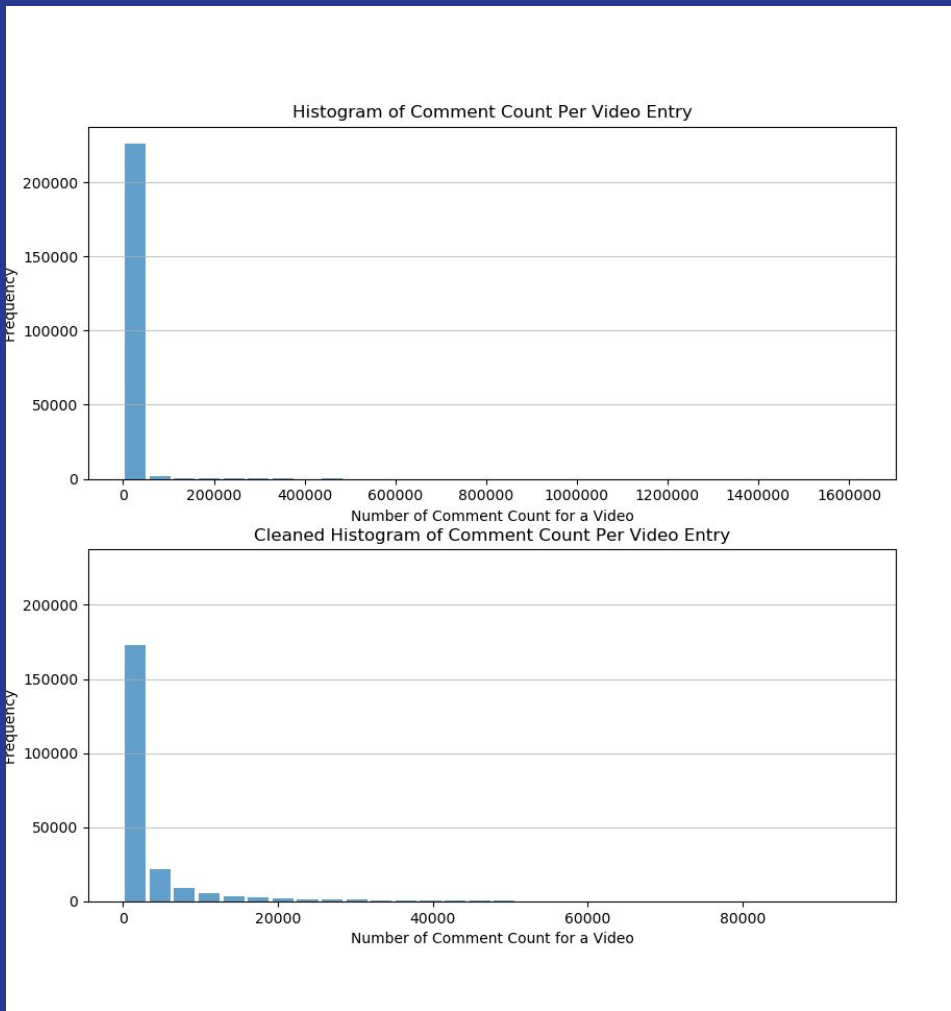




# Comment Count

The number of registered YouTube users that left a comment on a video

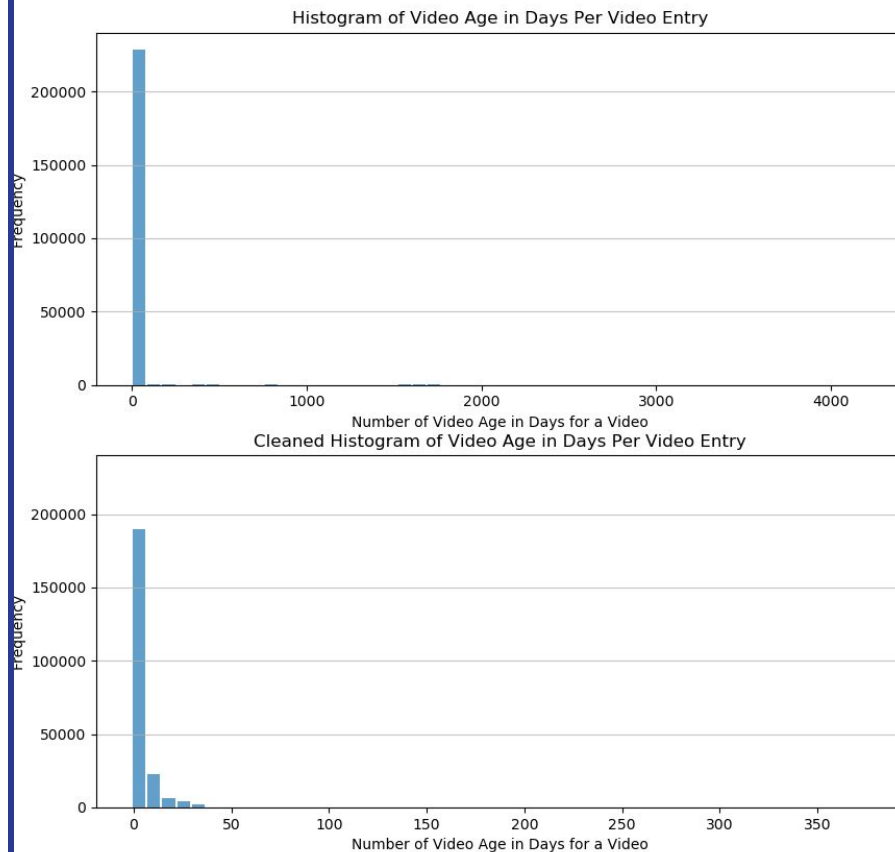
Mean	5.664155e+03
Mode	0
Std. Dev.	2.992399e+04
Spread	This is not evenly distributed and still varies quite widely for trending videos.
Outliers	Outliers will far more views, likely due to viral videos that garner comments.



# Video Age

The number of days that a trending video is on the platform at the time of trending

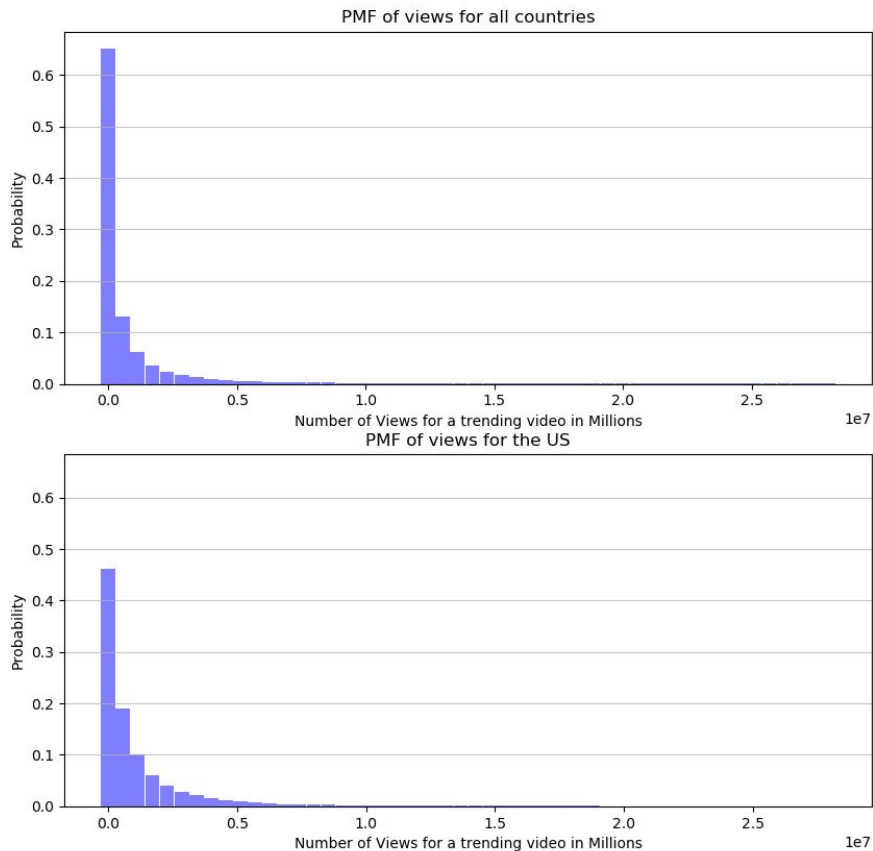
Mean	9.915418
Mode	0
Std. Dev.	122.153945
Spread	This is not evenly distributed and still varies quite widely for trending videos.
Outliers	Outliers will far more views, likely due to viral videos.



# PMF of Views from World vs US Views

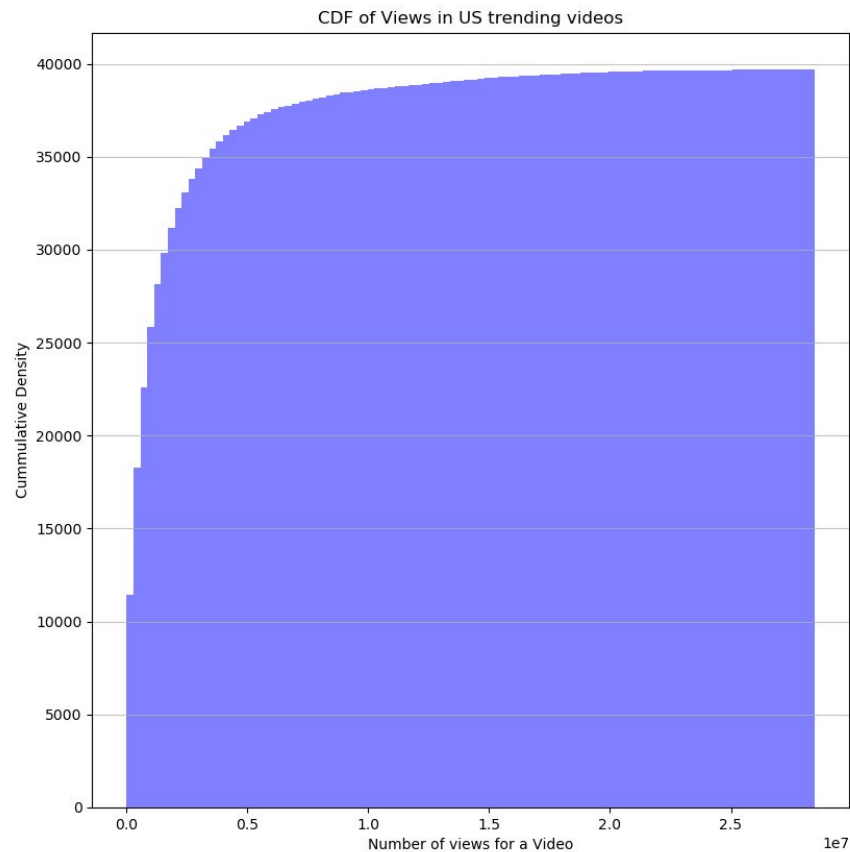
PMF of the number of views a video has for all countries, and then narrowed down to the US.

What we can see here is interesting, across the world there is likely a country that is pushing the distribution further to the left, because we see a somewhat more even distribution of views for trending videos in the US.



# CDF for Views in the US

What does a CDF of likes look like and what does it reveal?

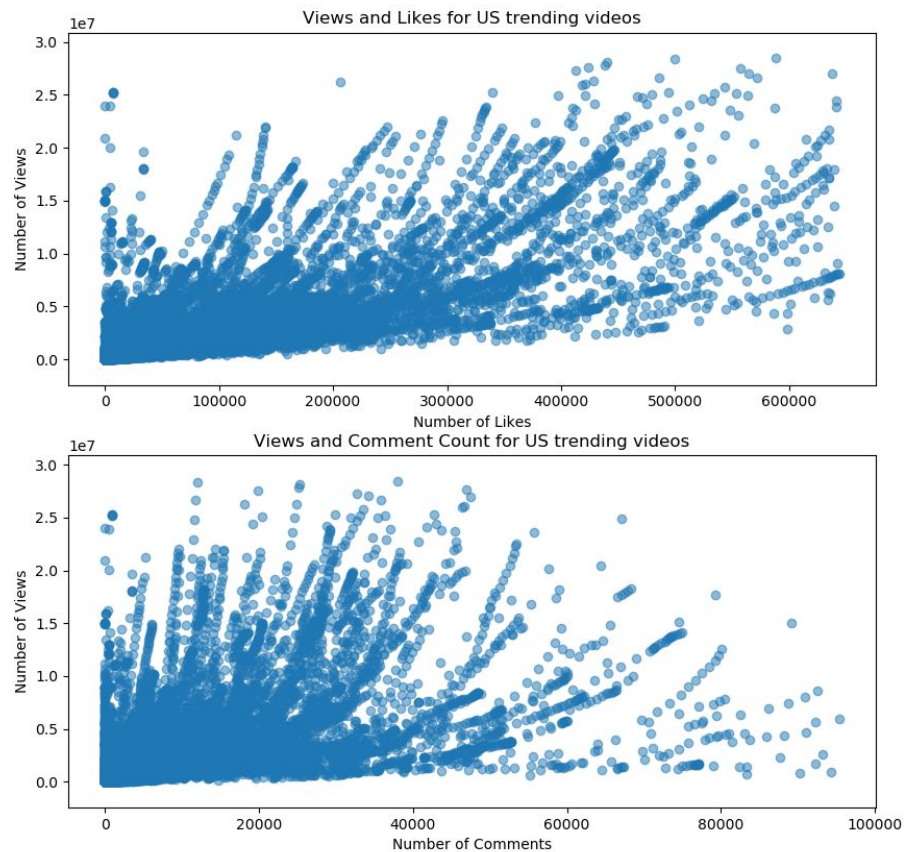


# Views and Likes

The scatter plots on the right illustrate two relationships US trending videos:

- Views and Likes
- Views and Comment Count

What we can see there seems to be very strong correlative relationship with views and likes, and that there might be some relationship between views and the number of comments. However the relationship for views and comments seems a little more skewed with some oddities, due to some outliers I may not have caught.



# Relationship between Views and Likes

I ran a correlation test between the three variables of interest thus far, and it does seem to illustrate a fairly strong and correlative relationships between views and the number of likes or comments.

Some thing of possible concern here is that likes and comment do have a very high relationship, even possibly showing some causality which in turn may interfere with the model.

	views	likes	comments
views	1.0	0.789743	0.620483
likes	0.789743	1.0	0.810042
comments	0.620483	0.810042	1.0

---

# Modeling for Views

A model was then created to multivariate linear regression to predict how many views that a trending video would get based on several factors.

*Formula:  $views \sim likes + dislikes + comment\_count + video\_age$*

From the R-squared values given, it seems the model isn't very the most accurate, with the model explaining the variance in ~68% for the actual view counts.

## Model Summary

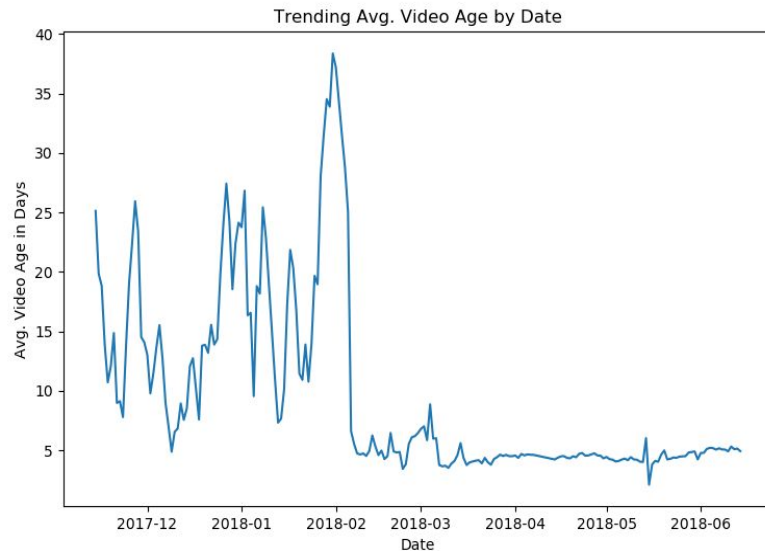
Dep.Variable	views
Model	OLS
Method	Least Squares
No. Observations	224788
Df Residuals	224783
R-squared	0.681
Adj. R-squared	0.681
F-statistic	1.197e+05

Variable	coef	std_err	P >  t
Intercept	1.126e+05	3544.719	0.000
likes	26.2963	0.073	0.000
dislikes	159.8984	0.914	0.000
comment_count	-60.1904	0.720	0.000
video_age	3.206e+04	365.585	0.000

# Other Exploration

This led me to want to explore some time series views of this and found some very interesting patterns that I explore in the code and figures I created in the GitHub project.

An example of this would be the mean in the change in video age between days. There is the fascinating drop off in video age that would indicate to me a change in either the trending algorithm or user behavior.





# Conclusions

There is certainly a relationship between the number of views and the number of likes, dislikes, and number of comments. I think maybe limited, and that there are likely other data points or more complex models that maybe better suited to predict the number of views. However, using only those in an analysis and creation of a model.

There is also much more that can be done with this dataset, I ended up spending a considerable amount of time doing some basic time series exploration but was limited by the constraints of the data.

