Alberina Latifaj, Kamil Solinski, Anthony Tony-Itoyah, Ryan Monaghan

CISC 4631 R03

Final Project

YouTube Trending Page Dataset

Dataset: https://www.kaggle.com/datasnaek/youtube-new

ABSTRACT

Our project will focus on determining the most important attributes of a video on the YouTube trending page. Our business purpose is to find out what a YouTube creator can do to get their video trending. The exact algorithm for the trending page is unknown, so we will be conducting our experiments without being entirely aware of what the results will be. We will be conducting three experiments on a YouTube dataset in order to figure out answers to the following questions;

1. How much can the creator control in order to get the most amount of engagement?
    a. (i.e; Disabling or enabling comments and ratings, tags)
2. Do certain categories appear on the trending page more often than others? Why?
3. Does the ratio between likes and dislikes play a role in video being considered trending? What was the average like to dislike ratio of a trending video?

The assumptions we have made so far are that videos that have the most amount of engagement (likes, comments, views) end up on the trending page more frequently. In addition to this, we assume that the trending page will have many different categories of videos on it every day, but some will consistently appear every day. Lastly, we are assuming that the characteristics of YouTube videos which appear on the trending page in this dataset, will have similar characteristics to those which appear on the trending page in the future.

INTRODUCTION

According to Webwise, “YouTube is a video sharing service where users can watch, like, share, comment and upload their own videos. The video service can be accessed on PCs, laptops, tablets, and via mobile phones.” These videos have a wide range of different content from music videos to sports highlights, to even everyday life. If you can think of it, there is a great chance that it is on Youtube. With that being said, if everything is on Youtube, what makes one video superior to another? Our dataset from Kaggle shows a daily record of the top trending YouTube videos from January 12, 2017, to May 31, 2018, in ten countries including, the United States, Japan, Russia, Canada, France, Germany, United Kingdom, South Korea, India, and Mexico. For our project, we decided to focus only on the United States as it is the most relevant to our business purpose. There are up to 200 listed trending videos per day and each video is listed in the dataset with the following attributes attached to it:

1. Video_id // string of characters created by YouTube to create URL for video
2. Trending_date // date that video appeared on the trending page
3. Title // string, what the creator chose to call their video
4. Channel_title // string, name of the creator’s channel
5. Category_id // int, used by to YouTube categorize videos (each number represents a different category)
6. Publish_time // date and time that the video was published by the creator
7. Tags // string of tags separated by “|”
8. Number of views // int, how many times the video was viewed
9. Number of likes // int, how many times the video was liked

10. Number of dislikes // int, how many times the video was disliked

In the dataset, there are no missing values for any column or row, but some of the columns will not be necessary for our purposes. The usefulness of our experiments will be appreciated by creators who are interested in growing their YouTube channels and attaining more subscribers by having their video featured on the trending page.

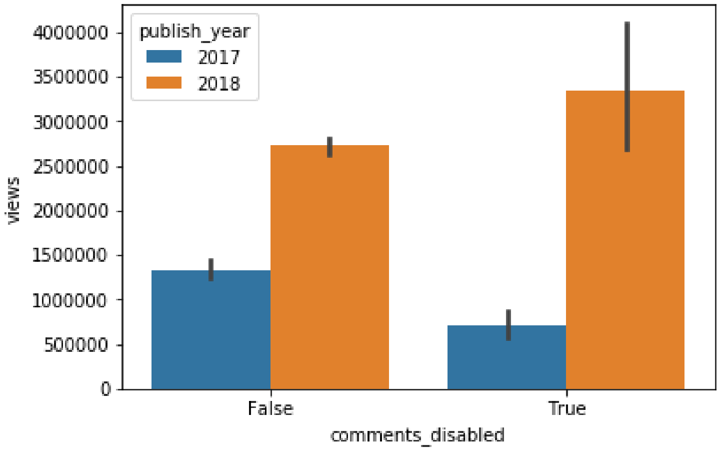
DATA PRE-PROCESSING

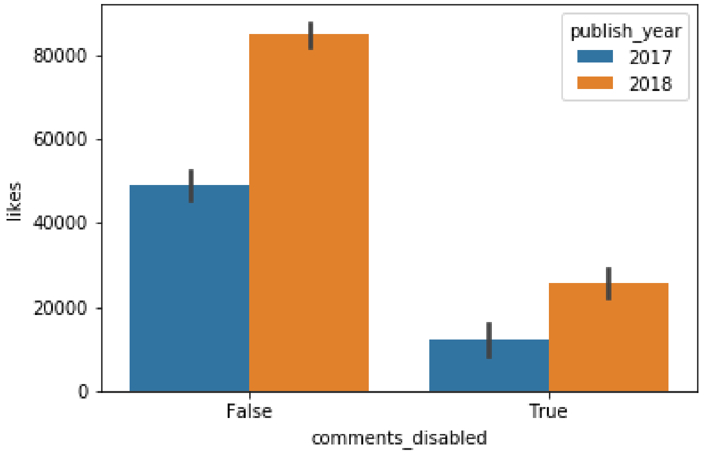
Some attributes will not be necessary in order to complete our experiments and achieve our business purpose. For example, Channel_title and Video_id will not provide us with any helpful information to conduct our experiments, so they will be dropped. In addition to that, some videos which appeared on the trending page had the same title, which we assumed meant that they are the same video (ie music videos), so the videos with the same title were also removed. For the first experiment, there are some patterns in the data that will help us make predictions for how the results will come out. The following bar graphs represent some important information pertaining to the question; How much can the creator control in order to get the most amount of engagement? The first instance of choice that a creator has is whether or not they want to enable or disable their comment section and rating section. The following bar graphs show the relationship between

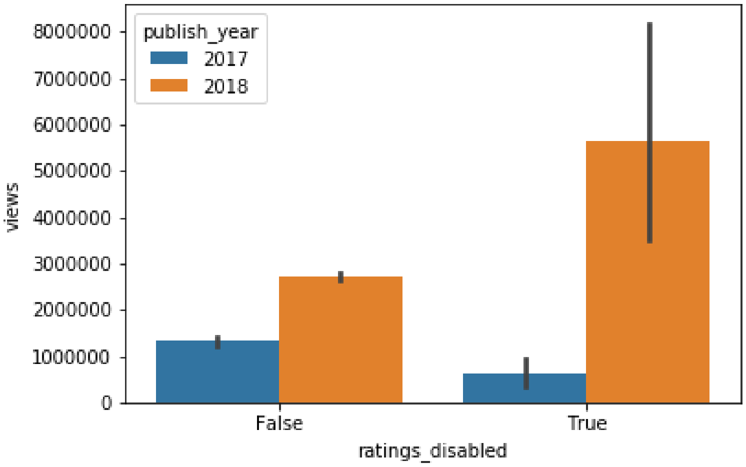
1.y =mean views of every video which appeared on the trending page in 2017 and 2018 // x= comments_disabled = true and false

2. y= mean likes of every video which appeared on the trending page in 2017 and 2018 // x = comments_disabled = true or false

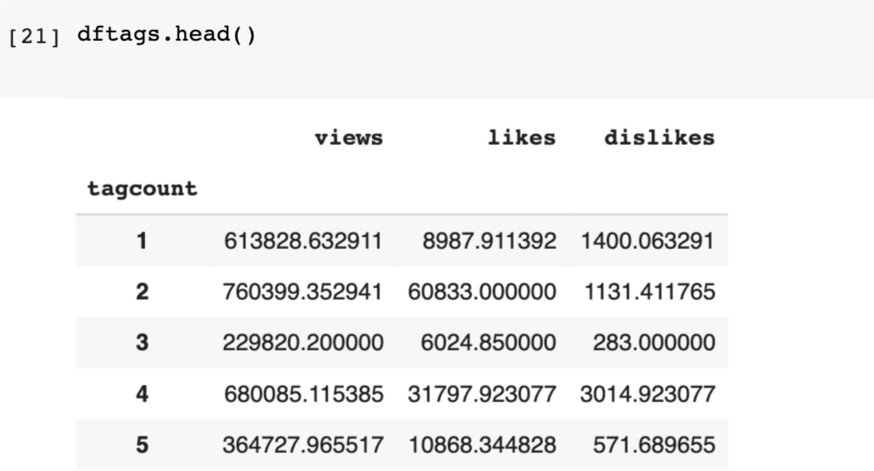
3. 1.y =mean views of every video which appeared on the trending page in 2017 and 2018 // x= ratings_disabled = true and false





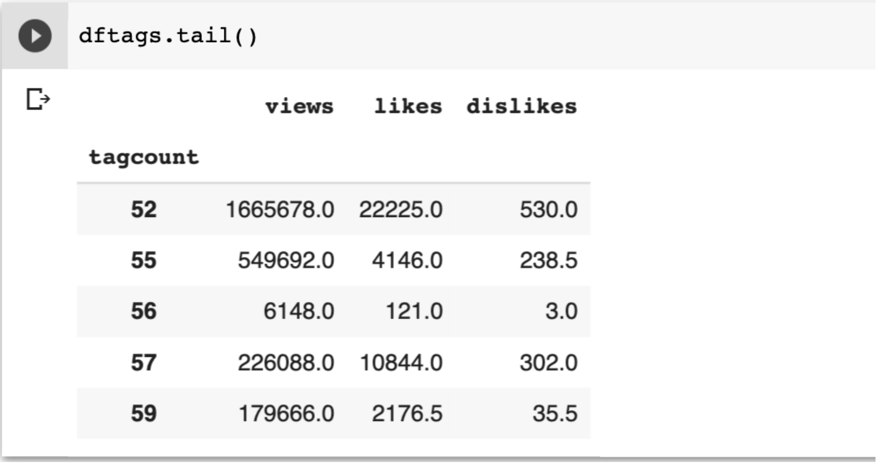


In addition to the option of disabling or enabling comments and ratings, the creator also has the option to add as many different tags as they want. The dataset shows all of the tags as strings, but because there is such a large number of different tags, we will be using the amount of tags rather than the specific string values. In order to make meaningful conclusions about tags, we have grouped the tagcount with mean views, likes and dislikes for every video in the dataset. After making the table, we found that the amount of tags ranges from 1-59 per video.



```
[21] dftags.head()
```

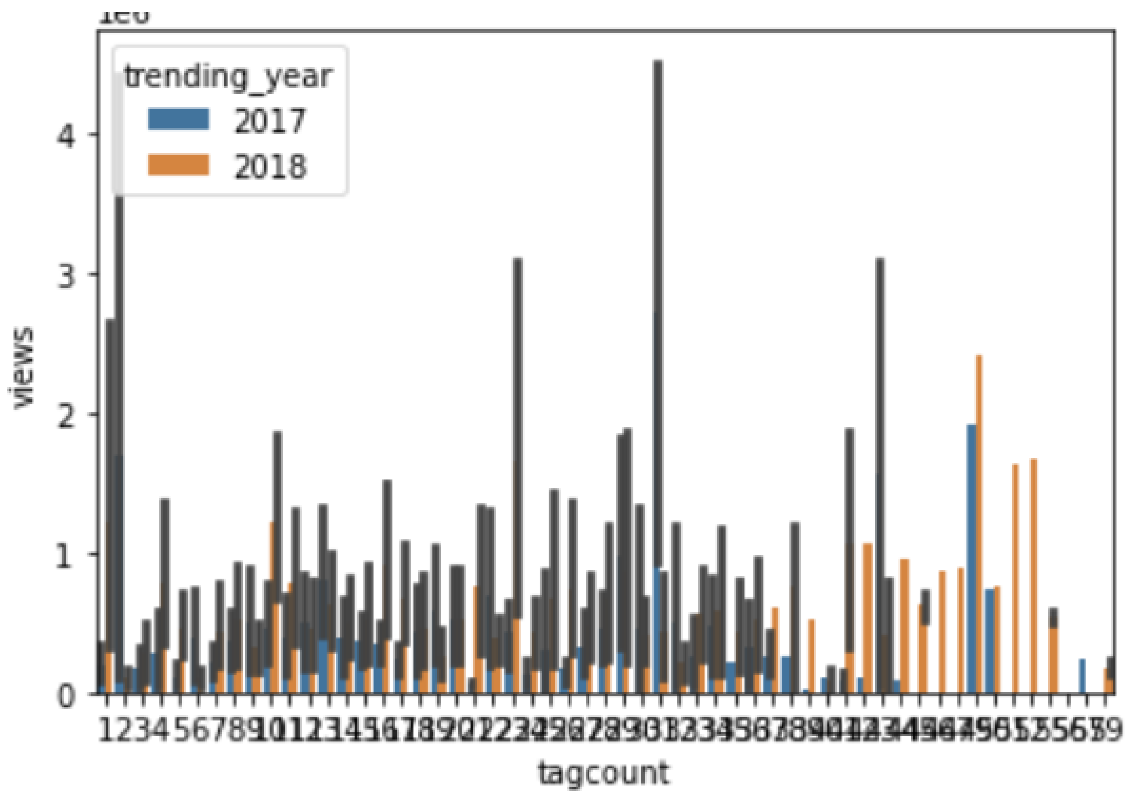
tagcount	views	likes	dislikes
1	613828.632911	8987.911392	1400.063291
2	760399.352941	60833.000000	1131.411765
3	229820.200000	6024.850000	283.000000
4	680085.115385	31797.923077	3014.923077
5	364727.965517	10868.344828	571.689655

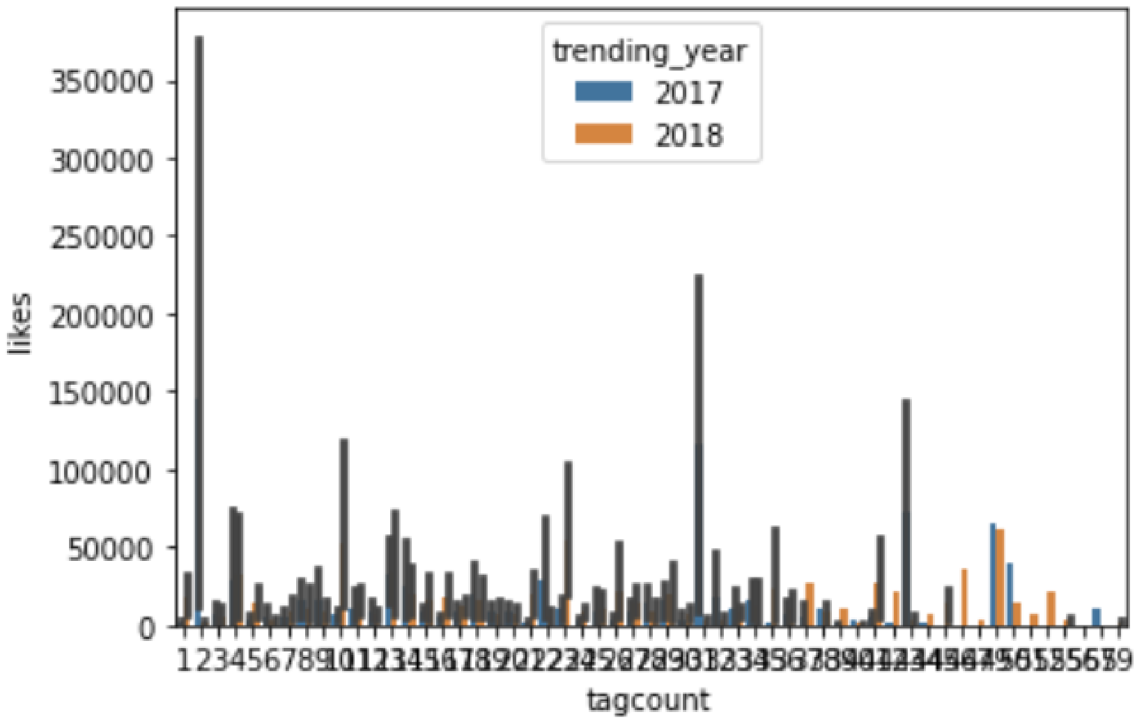


```
dftags.tail()
```

tagcount	views	likes	dislikes
52	1665678.0	22225.0	530.0
55	549692.0	4146.0	238.5
56	6148.0	121.0	3.0
57	226088.0	10844.0	302.0
59	179666.0	2176.5	35.5

There are also two more bar graphs which help to visualize the relationship between tagcount, views and likes.


1e6
trending_year
2017
2018
views
4
3
2
1
0
tagcount


trending_year
2017
2018
likes
350000
300000
250000
200000
150000
100000
50000
0
tagcount

EXPERIMENT 1: Decision Tree

After reviewing the graphs, it is clear that comments, ratings and tags do have a significant impact on the views and likes of a YouTube video, and this is something that can be further investigated using a decision tree based on information gain. To conduct this experiment, we used the 5 most viewed videos in the dataset along with the 5 most liked videos in the dataset along with the tagcount associated with each. We will be using the decision tree to predict if

Comments_enabled = 1 , true Ratings_enabled =1, true

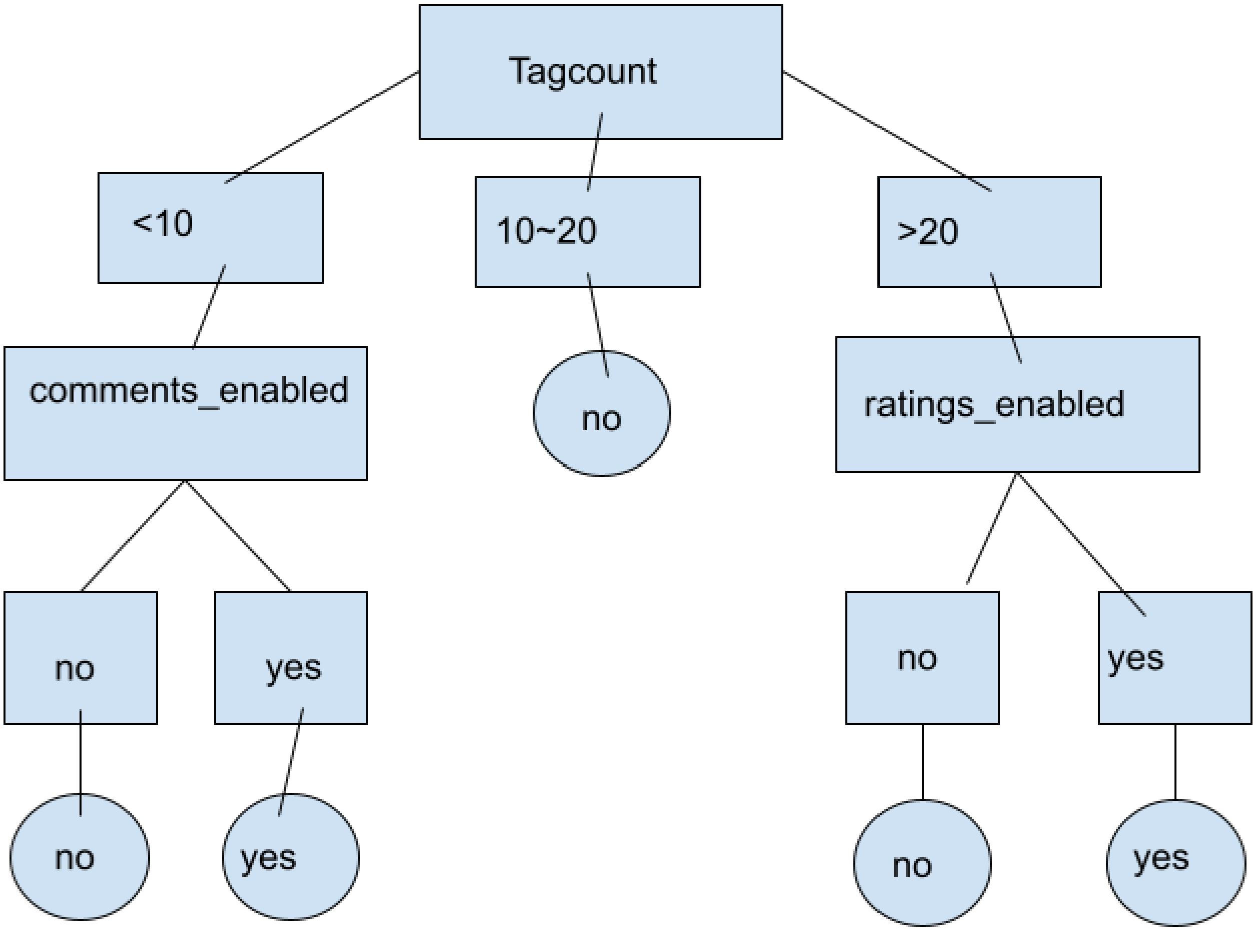
Comments_enable = 0, false (comments_disabled) Ratings _enabled=0, false(ratings_disabled)

Top_five_Views= 1 : top five

Top_five_Views =0 : bottom five

ID	Tag_count	Comments_enabled	Ratings_enabled	Top_five_Views
1	<10	1	1	1
2	<10	0	1	1
3	>20	1	0	1
4	>20	1	1	1
5	<10	1	1	1
6	<10	0	1	0
7	10~20	1	1	0
8	>20	1	0	0
9	<10	1	1	0
10	10~20	0	1	0

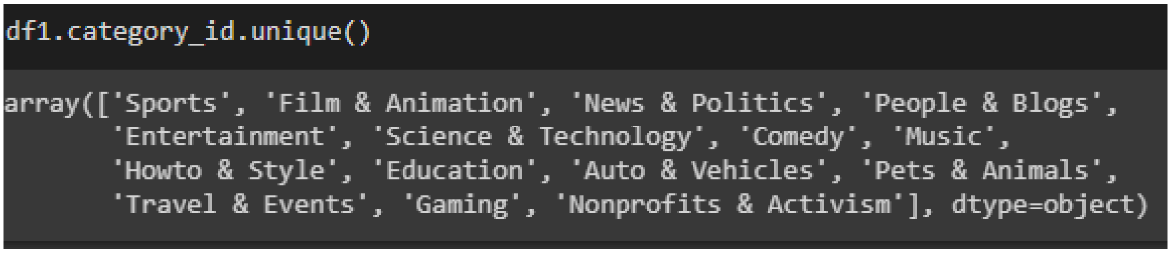
A Decision Tree for top_Five_Views



Based on this decision tree, one can decide whether or not comments or ratings need to be enabled and how many likes should be added to a video in order to achieve the most amount of views. The results show that when the tag count is less than 10 and comments are enabled the video will likely achieve a similar amount of views to a video which ended up on the top five most viewed list. Similarly, a video with over 20 tags and ratings enabled would have the best chance of ending up in the top five most viewed videos. After conducting this experiment, it is clear that the user's decision to enable or disable comments/ratings and how many tags they attach to their video will have a significant impact on the amount of views.

Experiment 2: Association Rule on Categories

For the second experiment we wanted to analyze the categories on the trending list. Some questions we wanted to see if we could answer were, Did some categories trend more at a certain time? Do categories have correlation to what time of the year they trend? Which categories get the most trending videos? Are some categories always trending? Are there categories that never show up on the list? Is there an association between specific categories?

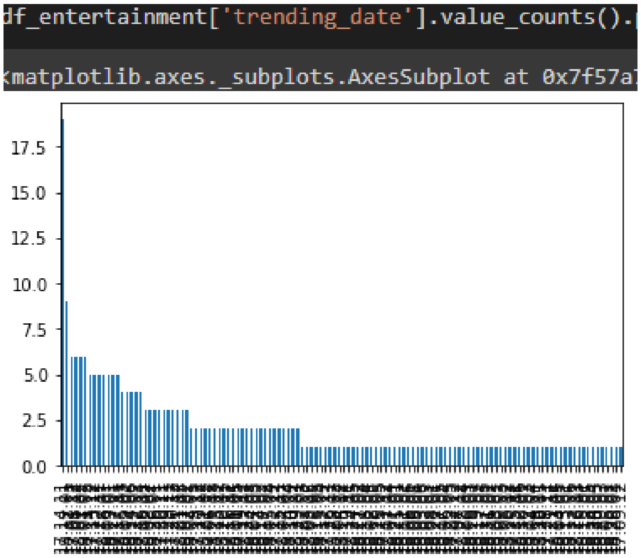


```
df1.category_id.unique()
```

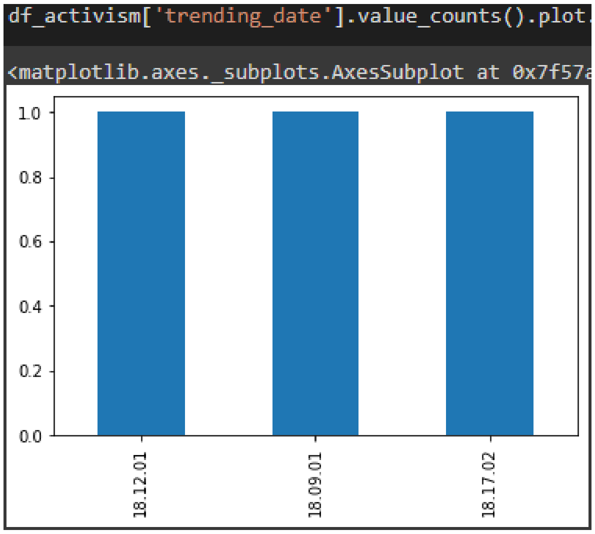
```
array(['Sports', 'Film & Animation', 'News & Politics', 'People & Blogs',
       'Entertainment', 'Science & Technology', 'Comedy', 'Music',
       'Howto & Style', 'Education', 'Auto & Vehicles', 'Pets & Animals',
       'Travel & Events', 'Gaming', 'Nonprofits & Activism'], dtype=object)
```

Examining the categories in the data set we were able to see which categories showed up on the trending list which were News & Politics, Education, Film & Animation, Music, Pets & Animals, Sports, Gaming, People & Blogs, Comedy, Entertainment, Howto & Style, Science & Technology, Nonprofits & Activism. The “US_category_id.json" file includes every category that can be listed on YouTube and the categories in the file that do not show up in the data set are Movies, Classics, Short Movies, Video Blogging, Trailers, Shorts, Foreign, Family, Anime/Animation, Action Adventure, Thriller, Sci-Fi/Fantasy, Horror, Drama, Documentary, Comedy. This answers the question of which categories never show up on the list.

In order to find which categories get the most trending videos and to find correlation between time we plotted the information in a bar graph.From the graphs we can see that entertainment has the most days of being on the trending list with sports, news, and comedy also showing up the most on the trending list. The graphs also show that activism was very rarely on the list along with pets and gaming rarely showing up.

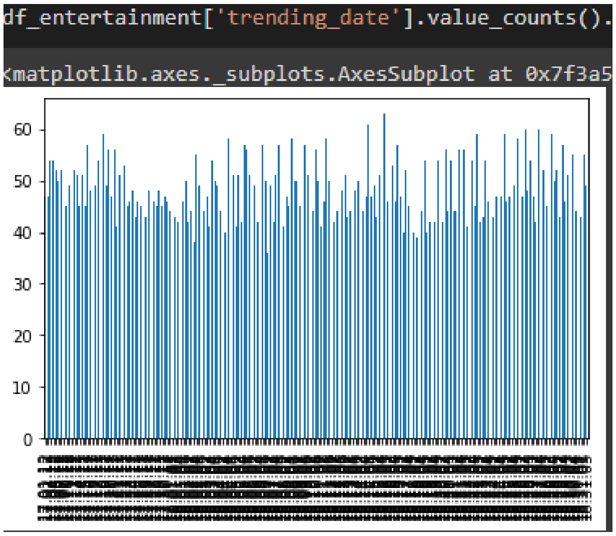


```
df_entertainment['trending_date'].value_counts().
<matplotlib.axes._subplots.AxesSubplot at 0x7f57a
```

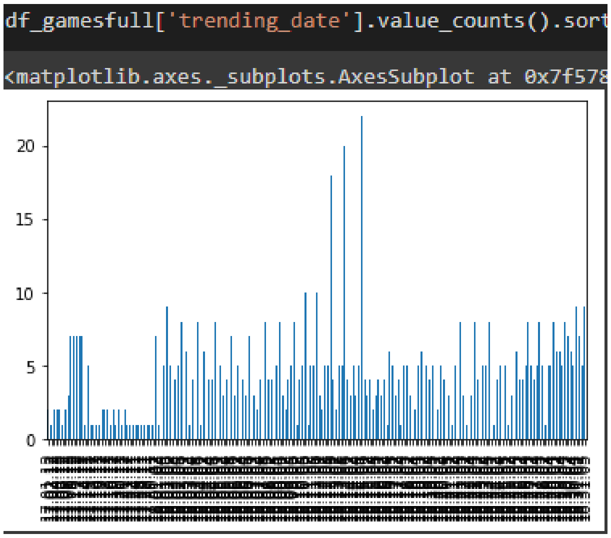


```
df_activism['trending_date'].value_counts().plot.
<matplotlib.axes._subplots.AxesSubplot at 0x7f57a
```

Using the entire data set, we examined which categories would take up most of the trending list per day. Entertainment was the most consistent and also the most trending of the categories. The gaming category had an interesting trend and it is possible it coincides with the gaming industry.

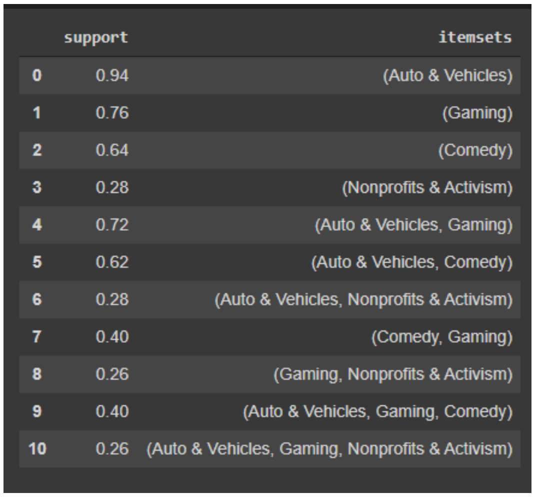


```
df_entertainment['trending_date'].value_counts().
<matplotlib.axes._subplots.AxesSubplot at 0x7f3a5
```



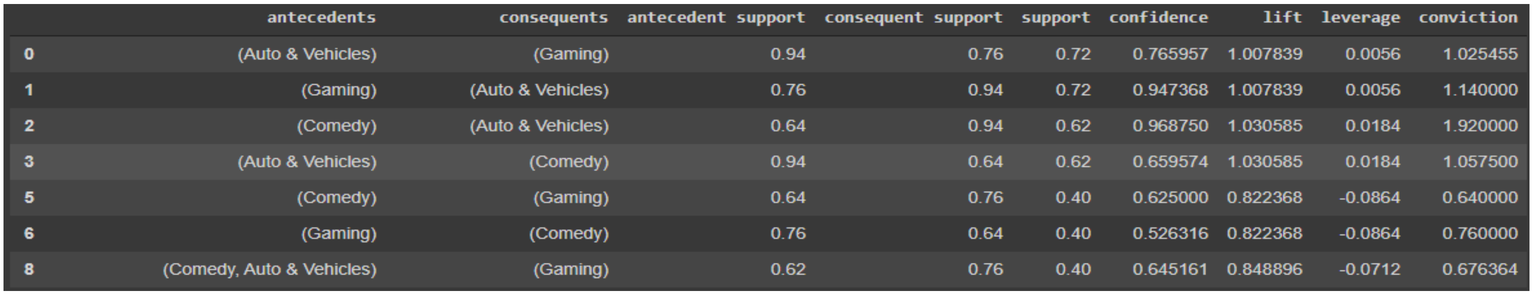
```
df_gamesfull['trending_date'].value_counts().sort
<matplotlib.axes._subplots.AxesSubplot at 0x7f578
```

Finally we ran the association rule on all the categories first run with all the categories which then showed us that the categories of Education, Film & Animation, Music, Pets & Animals, Sports, People & Blogs, Entertainment, Howto & Style, Science & Technology show up everyday in the trending list. This gave all the categories a support of 1 so we ran the association rule minus the categories showing up every day.



	support	itemsets
0	0.94	(Auto & Vehicles)
1	0.76	(Gaming)
2	0.64	(Comedy)
3	0.28	(Nonprofits & Activism)
4	0.72	(Auto & Vehicles, Gaming)
5	0.62	(Auto & Vehicles, Comedy)
6	0.28	(Auto & Vehicles, Nonprofits & Activism)
7	0.40	(Comedy, Gaming)
8	0.26	(Gaming, Nonprofits & Activism)
9	0.40	(Auto & Vehicles, Gaming, Comedy)
10	0.26	(Auto & Vehicles, Gaming, Nonprofits & Activism)

Auto & Vehicles has large support with the other categories with Comedy and Gaming also having a good amount of support.



	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(Auto & Vehicles)	(Gaming)	0.94	0.76	0.72	0.765957	1.007839	0.0056	1.025455
1	(Gaming)	(Auto & Vehicles)	0.76	0.94	0.72	0.947368	1.007839	0.0056	1.140000
2	(Comedy)	(Auto & Vehicles)	0.64	0.94	0.62	0.968750	1.030585	0.0184	1.920000
3	(Auto & Vehicles)	(Comedy)	0.94	0.64	0.62	0.659574	1.030585	0.0184	1.057500
5	(Comedy)	(Gaming)	0.64	0.76	0.40	0.625000	0.822368	-0.0864	0.640000
6	(Gaming)	(Comedy)	0.76	0.64	0.40	0.526316	0.822368	-0.0864	0.760000
8	(Comedy, Auto & Vehicles)	(Gaming)	0.62	0.76	0.40	0.645161	0.848896	-0.0712	0.676364

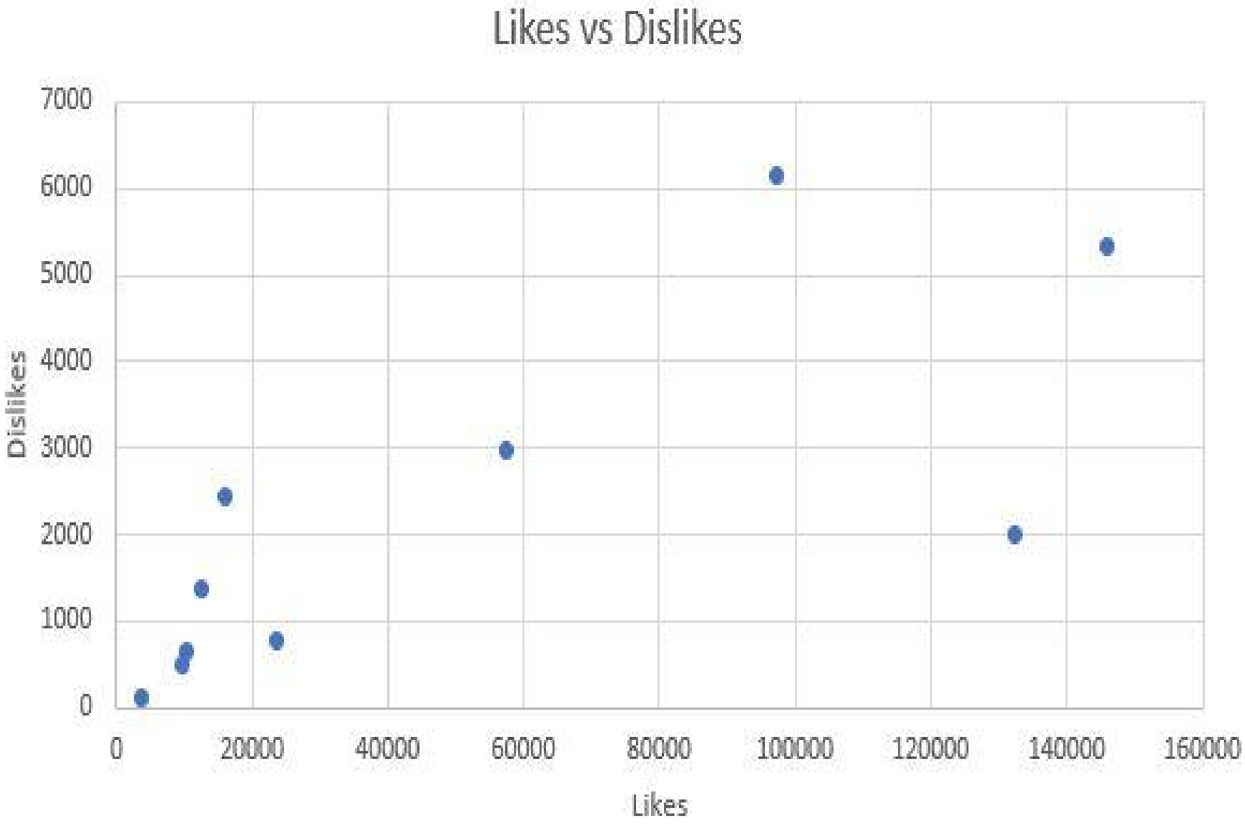
Our results from the association rule show that other than categories that trend every day Auto & Vehicles has a large association with the other categories.

Experiment 3: Clustering “Likes” and “Dislikes”

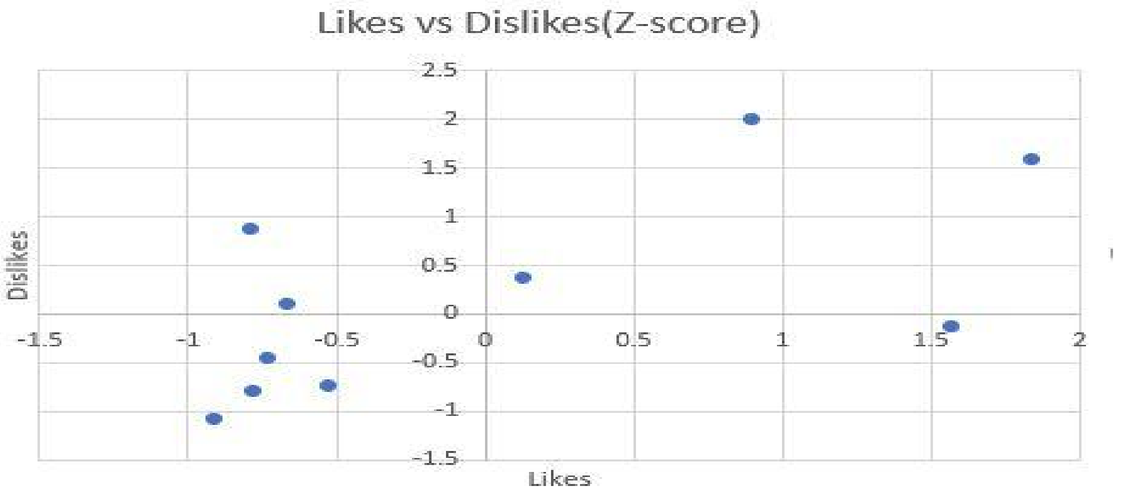
For the third experiment we wanted to cluster the likes and dislikes to further analyze their relationship with video trends. We wanted to answer the following question of if the ratio between likes and dislikes played a role in video being considered trending? What was the average like to dislike ratio of a trending video? Our hypothesis was that the average trending video had about a 10:1 like to dislike ratio. From the dataset we took the first ten likes and dislikes and plotted them, after that, we found their Z scores before we moved on to actually clustering them. We clustered them into two groups(group 1, and outliers). We did this because although most of the data was relevant, there were a few numbers that just did not agree with the others. The midpoint of the cluster on the left side of the scatter plot which is group one was about (-0.572,-0.139). The cluster on the right side of the scatter plot which is group two was (1.389, 1.174).

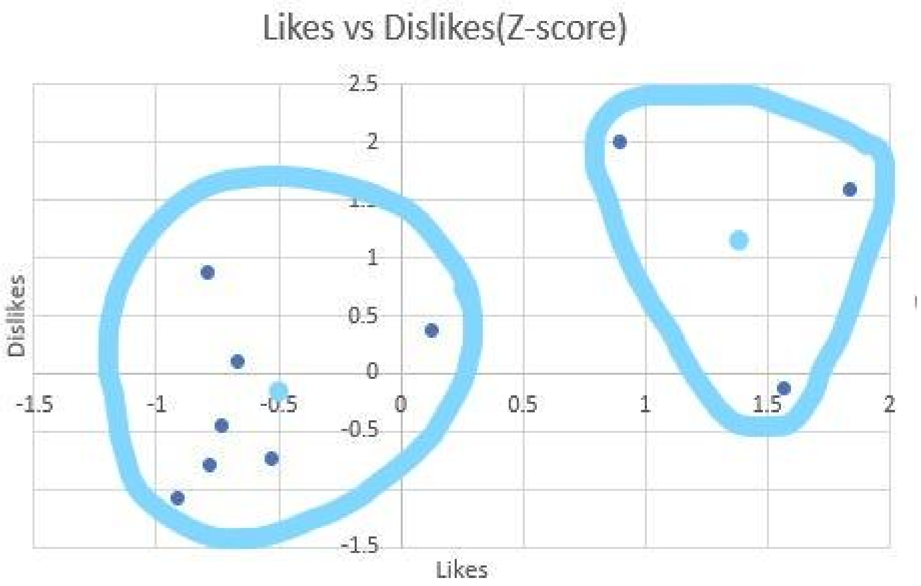
Video_id	Likes	Dislikes
2kyS6SvSYSE	57527	2966
1ZAPwfrtAFY	97185	6146
5qpjK5DgCt4	146033	5339
puqaWrEC7tY	10172	666
d380meD0W0M	132235	1989
gHZ1Qz0KiKM	9763	511
39idVpFF7NQ	15993	2445
nc99ccSXST0	23663	778

jr9QtXwC9vc	3543	119
TUmyygCMMGA	12654	1363



Video_id	Likes(Z-Score)	Dislikes(Z-Score)
2kyS6SvSYSE	0.128	0.374
1ZAPwfrtAFY	0.893	1.998
5qpjK5DgCt4	1.835	1.586
puqaWrEC7tY	-0.785	-0.799
d380meD0W0M	1.569	-0.124
gHZ1Qz0KiKM	-0.793	0.878
39idVpFF7NQ	-0.672	0.108
nc99ccSXST0	-0.525	-0.742
jr9QtXwC9vc	-0.913	-1.078
TUmyygCMMGA	-0.737	-0.443





Based on our results from the graph and dataset, we were able to come to the conclusion that the average ratio between likes and dislikes was about 23:1(there were few cases that this did not hold true). Also, even though there is not a direct correlation on the dataset between trends and like and dislikes, one can agree that based on the clustering graphs and dataset that the ratio can have an effect on if a Youtube video goes viral.

CONCLUSION:

Our goal was to get a better understanding of the term “Trending” on Youtube. We were trying to figure how exactly one achieves such reward and we did so by analyzing “The Trending Youtube Video Statistics” dataset(https://www.kaggle.com/datasnaek/youtube-new). We answered questions such as How much can the creator control in order to get the most amount of engagement? Do certain categories appear on the trending page more often than others? Why Does the ratio between likes and dislikes play a role in video being considered trending? What was the average like to dislike ratio of a trending video? We answered the following questions using a decision tree to prove that comments, ratings, and tags had a significant impact on how well a video did. After that we used the Association rule to answer the question on if timing had an affect on trends. After an experiment we came to the conclusion that timing did influence the success of video because some categories trended everyday while others were seasonal. Our final experiment was to see if there was a pattern between the ratio of likes to dislikes and the success of the video based on the dataset and clustering. The results of the experiment showed that if you had a ratio of 23:1 likes to dislikes, there was a great chance that one's video would be a success(trend). This information and the multiple experiments are useful because it helps future Youtube video developers by giving them a better understanding of how exactly they can achieve their goal of "Trending" and gives them a better chance of growing their viewership. By having more users exposed to your video through the trending page, the creator is likely to gain subscribers and have a better chance at making money off of monetization through Google AdSense.