Micheal Peterson
William Wang
Sam Lu

Stat 489: Data Science Final Project

What makes a youtube video trending?

Youtube, a video sharing platform, is the second largest social media platform in the world with an estimated 1.5 billion monthly active users, only behind Facebook. Videos are a powerful form of media and cover almost every topic you can think of. Videos can range from instructional to humorous and everything in between. Many people now and more and more everyday use youtube as their main form of entertainment. Additionally, being a “youtuber” is a viable career option with some big channel names making millions of dollars (sometimes even close to a billion). Youtube is easily accessible for everybody to watch and upload videos. Youtube is a valuable source of information to people around the globe. As the most popular video database in the world, it's interesting to know how Youtube determines what videos qualify as “trending”. The trending tab is seen by every user and for many is a useful tool to discover videos. It is not publicly disclosed how these videos are chosen and we would like to find out.

To begin we must be able to grab videos from the trending page. We started by using BeautifulSoup because the link is consistent and will be able to grab a snapshot of the videos and channels on that page. From this page we are able to gather various channels. One problem we encountered was each channel has a unique ID.However some of the links lead the user page. The issue here is that the API needs the channel ID to be able to grab videos from that channel. So we devised a way with the API to have it return Channel ID. From the we are able to pull as many videos as we need from the channel. In this case we chose the most recent 15 videos if available. Once the API returns the video ID for the past 15 channels, We are able to use the video ID to grab view count, likes, dislikes,date uploaded and comments. The program loops this gathering data from every channel on the main page, with the past 15 videos from each channel and then saves the data collected to a .csv file

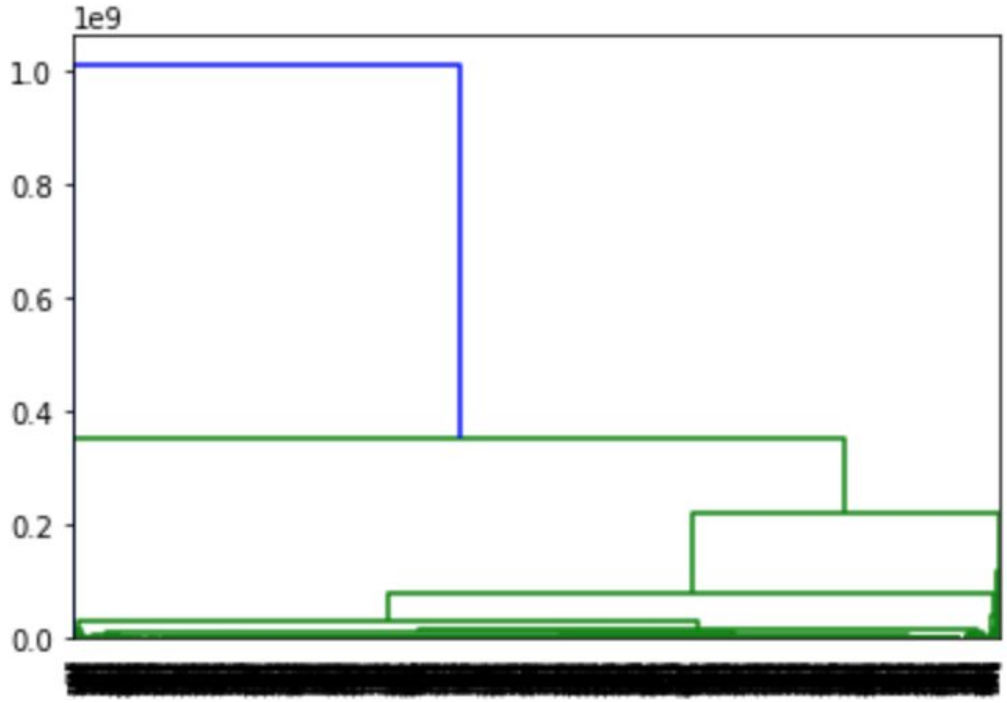
Some videos do not report some information such as views, and others do not allow liking, disliking, or comments. For these videos we chose to report those metrics as zero. Initially we created a univariate summary with the results shown below.

Micheal Peterson
William Wang
Sam Lu

	Metric	Mean	Median	Max	Standard Dev
0	Views	2873694.769	408212.000	1011252279	25292376.481
1	Likes	46933.564	11247.000	6434089	196857.754
2	Dislikes	1389.597	252.000	238537	6724.587
3	Comments	4318.440	1124.000	394721	13559.543

From this it can be seen that each metric is right skewed. This makes sense since no metric can be less than zero and can reach up to infinity. From this we can also draw a couple of easy conclusions. Some videos have an extreme amount of views compared to the average in our sample. Views is a much higher metric than the others. Most people do not engage in a video but simply watch it. Finally, people like videos at a much higher rate than the dislike them.

After this we performed a cluster analysis to see if any obvious clusters exists. Below is the complete clustering linkage report using the scipy library.



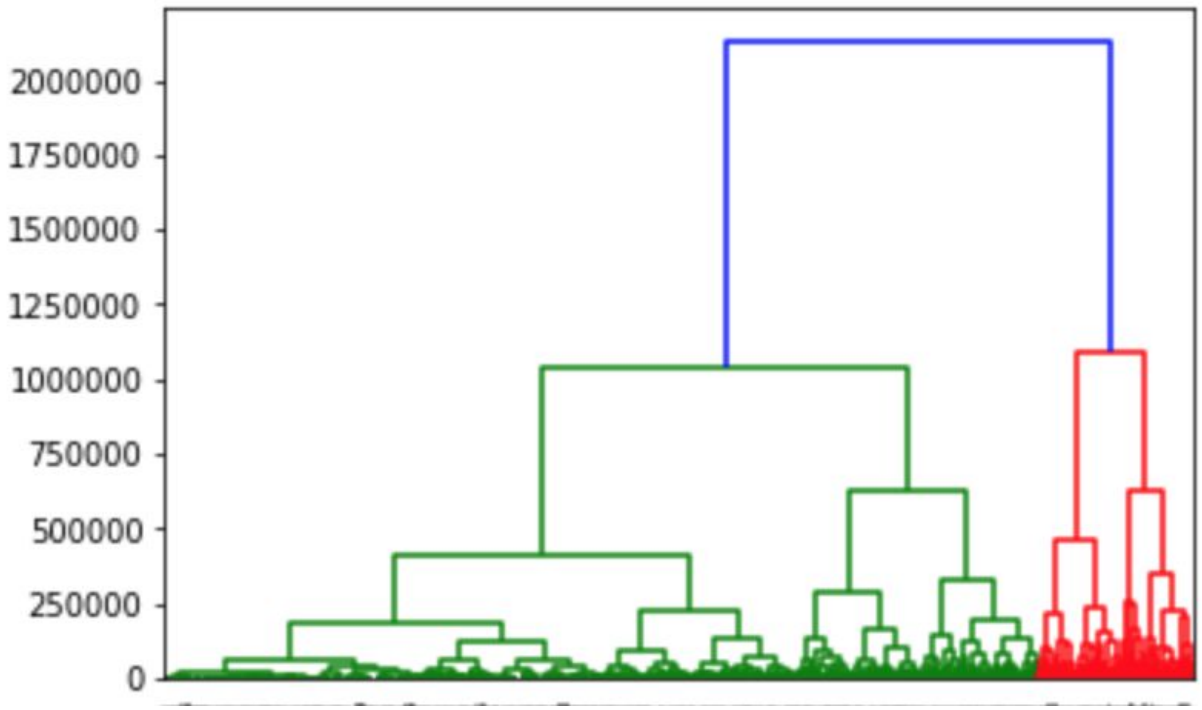
Micheal Peterson
William Wang
Sam Lu

As seen there does not exist any significant clusters. There does exist however some significant outliers. Due to this we chose to remove the outliers based on view counts using the interquartile range method. We accomplished this using the code below.

```
In [68]: views = videos["# of Views"].astype(np.int64)
         q3, q1 = np.percentile(views, [75 ,25])
         iqr = q3 - q1
         high_out = videos[views > q3 + 1.5*iqr]
         videos = videos.drop(high_out.index)
         metrics = videos.loc[:,videos.columns.isin(metric_variables)].as_matrix()
```

There were roughly 250 videos out of 2000 that were considered outliers. The max views was roughly 3 billion views. This method also removed many of the outliers of other metrics since these metrics are inherently positively correlated. Removing these outliers is not the most optimal method for analysis, especially considering the amount of “outliers” removed. These are not really outliers since they include videos inside the population we are trying to study (all youtube videos). There are other more optimal methods we could have used such as modifying the data to a logarithmic scale.

After removing these outliers we ran clustering again. This clustering resulted in two distinct groupings. One roughly about 5 times the size of the other, roughly the ratio of trending to non-trending. The linkage plot is reported below.



Micheal Peterson
William Wang
Sam Lu

After removing the outliers we performed a bivariate analysis of the metrics. By fitting a linear model between each of the variables we seek to get a sense of the relationship between each of the variables. Below is the result of using the numpy library least squares regression.

```
Linear Regression of # of Views and # of likes
Coeffecient 0.0318155093087 ; Bias -168.601825372
MSE 223318198.55

Linear Regression of # of Views and # of dislikes
Coeffecient 0.000912124708818 ; Bias 1.59258960683
MSE 590314.989733

Linear Regression of # of likes and # of dislikes
Coeffecient 0.0231116206599 ; Bias 88.8544900681
MSE 550180.564789

Linear Regression of # of Views and # of comments
Coeffecient 0.00441366689011 ; Bias -104.688704584
MSE 13783686.7322

Linear Regression of # of likes and # of comments
Coeffecient 0.134488522757 ; Bias -18.4367194112
MSE 10062990.1108

Linear Regression of # of dislikes and # of comments
Coeffecient 2.54666615198 ; Bias 877.017381089
MSE 13782288.3279
```

From this we can see a few things. Contrary to what we might think views does not highly correlate with any of the metrics. The largest correlation is between dislikes and comments. This could be explained by the concept of controversial videos. Users who dislike a video will then comment explaining why they disliked it.

Reviewing our exploratory findings, users like videos at a much higher rate than they dislike. Contrary to intuition views does not have a significant effect on whether user will like, dislike, or comment. There seems to be a correlation between dislikes and comments. There are significant outliers, most likely due to the exponential nature of viral videos. There are 2 significant clusters of videos. One being roughly 5 times larger than the other. Could these be trending and non-trending…

Micheal Peterson
William Wang
Sam Lu

In order to create a model to predict what videos are trending or not we need to pick an appropriate technique. Logistic Regression is the best model to predict trending vs. non trending for the following reasons.

- Well suited for binary classification problems
- Assumes the features are not highly correlated
- Does not require the data be linear
- Does not require Homoscedasticity

The final requirement is important to our situation since we do not know how these metrics are distributed. Most likely the metrics something similar to the gamma distribution but we did not believe we could get a good fit due to the number of of samples we got being small compared to the possible value space.

One drawback of Logistic Regression is that it can be hard to interpret the coefficients of the model in a strict sense. We are okay with this since we are trying to get a general sense of the relationship. Our interruption would not be well-suited for a risk analysis type scenario.

The following is the results from using sklearn's logistic regression and a ⅔ training ⅓ testing split.

Training	Testing
Misclassification	Misclassification
0.0512589928058	0.054844606947
Sensitivity	Sensitivity
0.972222222222	1.0
Specificity	Specificity
0.0204460966543	0.013358778626

The specificity recorded is actually (1 - specificity) due to an error in the code.

```
Coefficients:
# of Views : -5.63762966003e-06
# of likes : 6.10167563579e-05
# of dislikes : 8.72341355499e-05
# of comments : -0.000272688064317
Bias: [-0.00079543]
```

Based on our model there is a positive correlation with voting engagement and being on the trending page. The negative correlation between views maybe due to the fact that trending videos are often newer videos compared to the videos we collected

Micheal Peterson
William Wang
Sam Lu

After completing this project we identified several paths we could take in the future. Sampling videos over a time period to see if different attributes such as the derivatives of metrics or the release date of video play a role (which they most likely do). Looking at the content of comments, What are there sentiments? Is it true that comments are an indicator of controversy? Comparing trending videos on a channel basis rather than all trending videos to all non-trending videos. Or in general using sampling techniques based on statistical theory. Does video content play a role? Are certain genres of videos tend to make it to trending? Are certain content creators favored? Do certain users make it to the trending page more often? Are some users more likely than the average to make it?