Analyzing Youtube Trends

Capstone Project by Liyan Cen

1 INTRODUCTION

As YouTuber becomes one of the most popular freelancer jobs nowadays, it's getting more challenging and competitive to achieve fame on the platform. In other words, to make money by uploading youtube videos is just becoming more difficult.

So, how does a YouTuber make money? When a YouTuber links Google AdSense to his or her channel, youtubers earn 68% of the as revenue. At the same time, YouTube charges advertisers when a viewer watches 30 seconds or more of the ad. On average, it chagres around $.18 per view. Furthermore, the youtuber earns only when viewers view the ad. The longer time viewers spend on watching ads, the more profits youtubers could earn. However, only about 15% of viewers will be counted as a “paid view” since many of them skip. In sum, average youtubers earn around $18 per 1,000 views.

Accordingly, view count is the direct and critical factor that affects a YouTuber’s income. There are many factors such as category, comment counts, likes, dislikes, publish time, trending month etc.. which can possibly affect the view counts. We use datasets from sources containing these factors to build supervised machine learning models to determine which factor has the most significant influence to the view counts.

2 Data Acquisition and Cleaning

The dataset we acquire includes several months (and counting) of data on daily trending YouTube videos for US with up to 200 listed trending videos per day.

Given the fact that the original dataset only has category ID, we add another column to have the ID translated into specific category names. The names we add in are 'Film and Animation', 'Cars and Vehicles', 'Smartphone', 'Music', 'Pets and Animals', 'Sports', 'Travel', 'Game', 'People and Blogs', 'Comedy', 'Entertainment', 'News and Politics', 'How to and Style', 'Education', 'Science and Technology', 'Nonprofits and Activism', 'News & Politics'. These names are based on the associated JSON file, “US_category_id.json”.

For our machine learning models, we have “comment counts”, “likes”, “dislikes”, “publish time”, “trending month” as feature variable, and “view counts” as our target variable. Hence, in order to make all feature variable uniform in type, we convert all the values to numeric. In specific, for “trending_date” and “public_time”, we only extract the month and hour respectively, and convert them into integers. More details on the data cleaning process be found from the IPython notebook. The cleaned dataset is then ready for explorations.

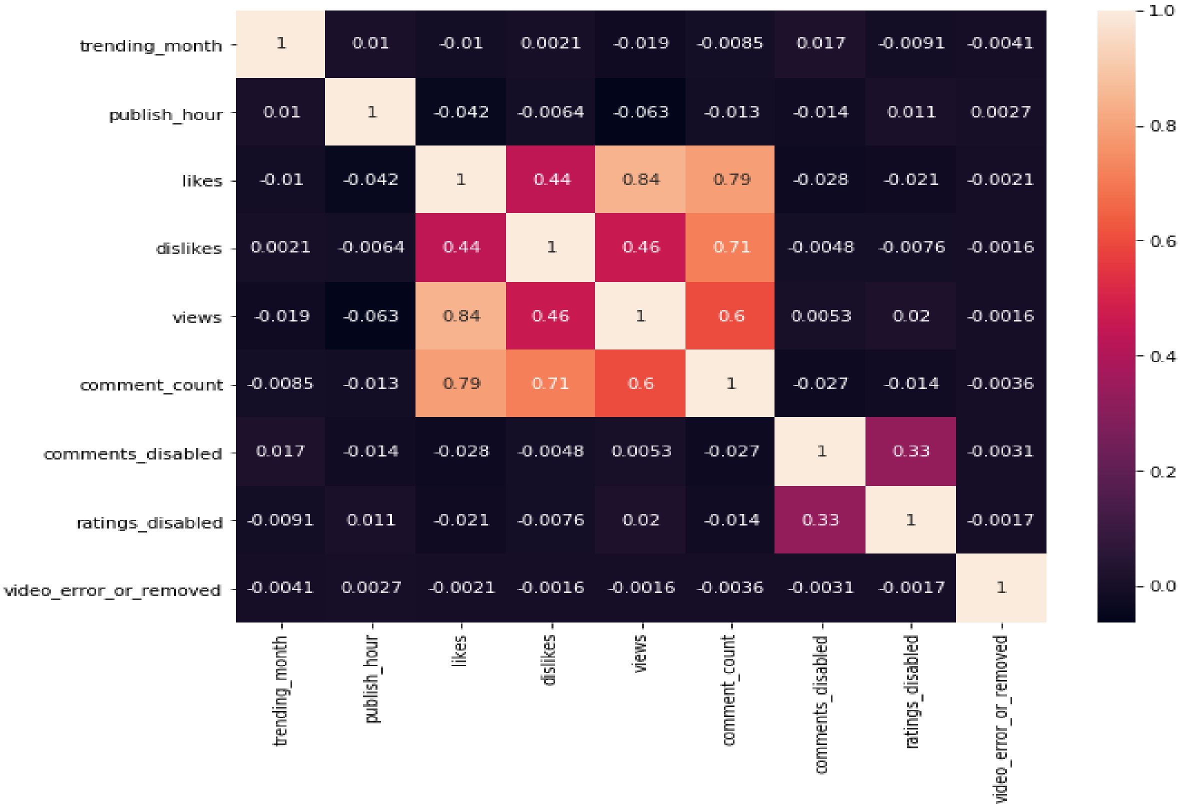
3 Data Exploration

3.1 Introduction to the cleaned data

There are 37549 records on daily trending videos in US for several months.We will go through most of the fields (or columns) in the dataset to explore their relationship with view counts. The target column for this project is called “views”.

3.2 Correlation between variables

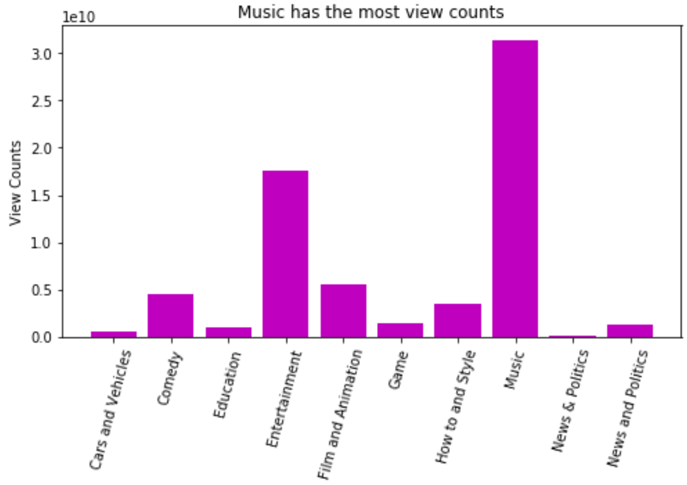
Based on the correlation matrix figure shown below, “likes”, “dislikes”, and “comment_counts” are more likely to have a correlation with the ‘views”.

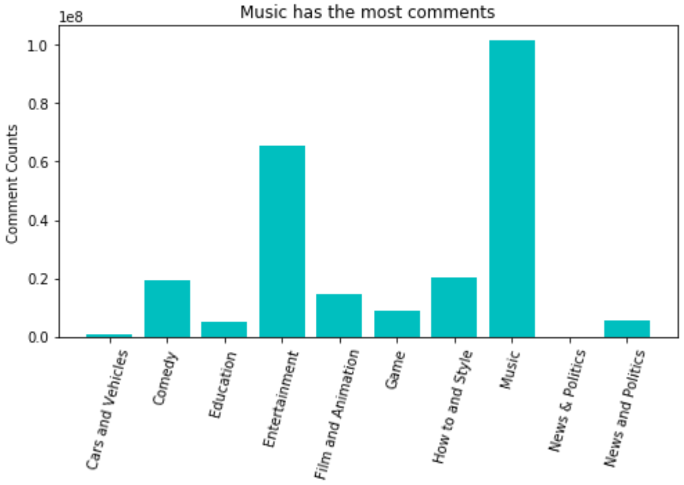


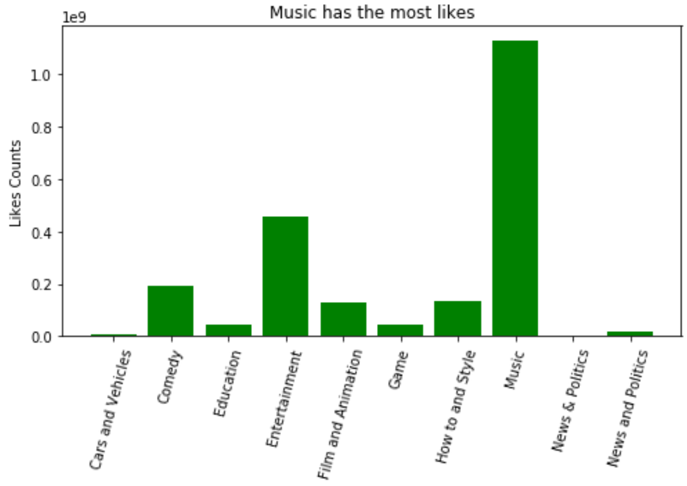
	trending_month	publish_hour	likes	dislikes	views	comment_count	comments_disabled	ratings_disabled	video_error_or_removed
trending_month	1	0.01	-0.01	0.0021	-0.019	-0.0085	0.017	-0.0091	-0.0041
publish_hour	0.01	1	-0.042	-0.0064	-0.063	-0.013	-0.014	0.011	0.0027
likes	-0.01	-0.042	1	0.44	0.84	0.79	-0.028	-0.021	-0.0021
dislikes	0.0021	-0.0064	0.44	1	0.46	0.71	-0.0048	-0.0076	-0.0016
views	-0.019	-0.063	0.84	0.46	1	0.6	0.0053	0.02	-0.0016
comment_count	-0.0085	-0.013	0.79	0.71	0.6	1	-0.027	-0.014	-0.0036
comments_disabled	0.017	-0.014	-0.028	-0.0048	0.0053	-0.027	1	0.33	-0.0031
ratings_disabled	-0.0091	0.011	-0.021	-0.0076	0.02	-0.014	0.33	1	-0.0017
video_error_or_removed	-0.0041	0.0027	-0.0021	-0.0016	-0.0016	-0.0036	-0.0031	-0.0017	1

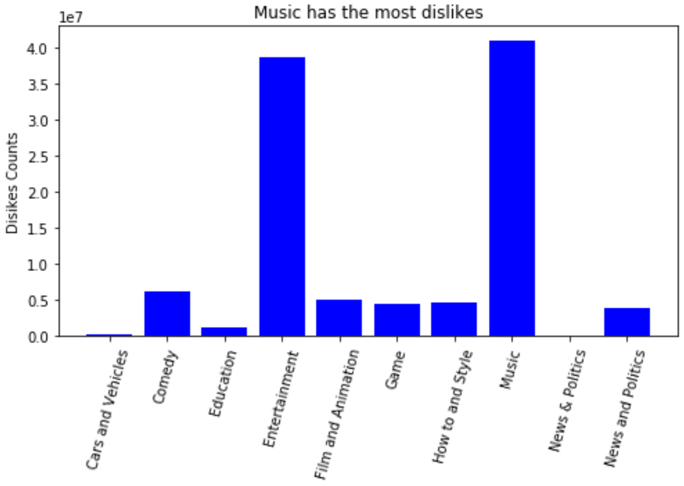
3.3 Most popular video category on youtube is “Music”

By referring to 4 graphs shown below, and by comparing the total views counts, comment counts, likes, and dislikes, we can conclude that music has the most of all the features.





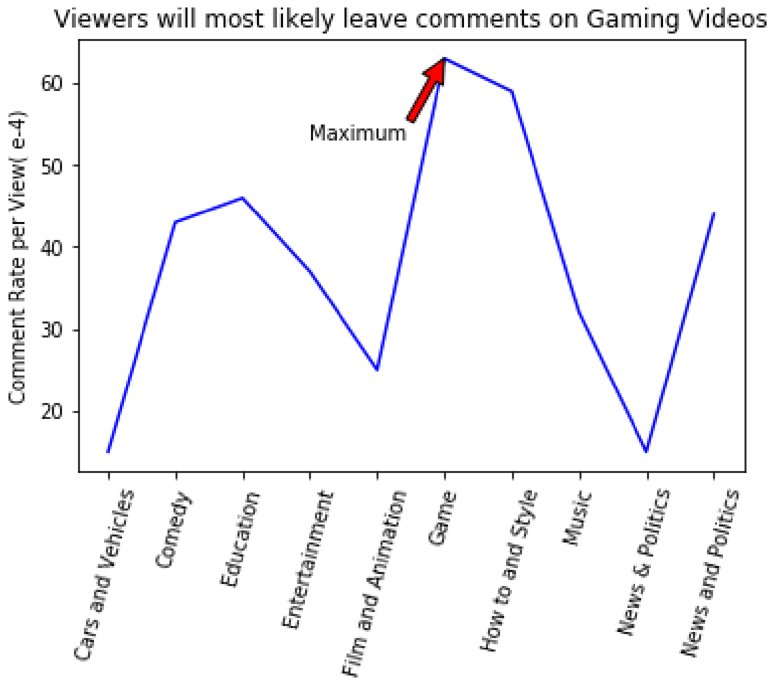




However, the results turn out differently if we compare each video category by total feature counts dividing the total view counts for each category, which are explained in details in the next 5 sections below.

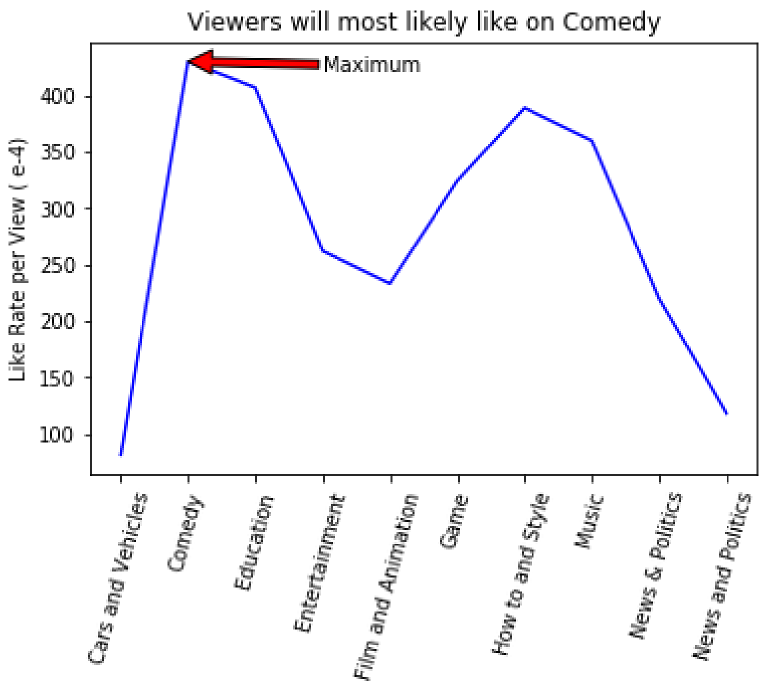
3.4 Viewers will most likely leave comments on Gaming Videos

By comparing comment counts for each video category and total view counts for each category, we get the conclusion that viewers mostly like leave comments on gaming videos.



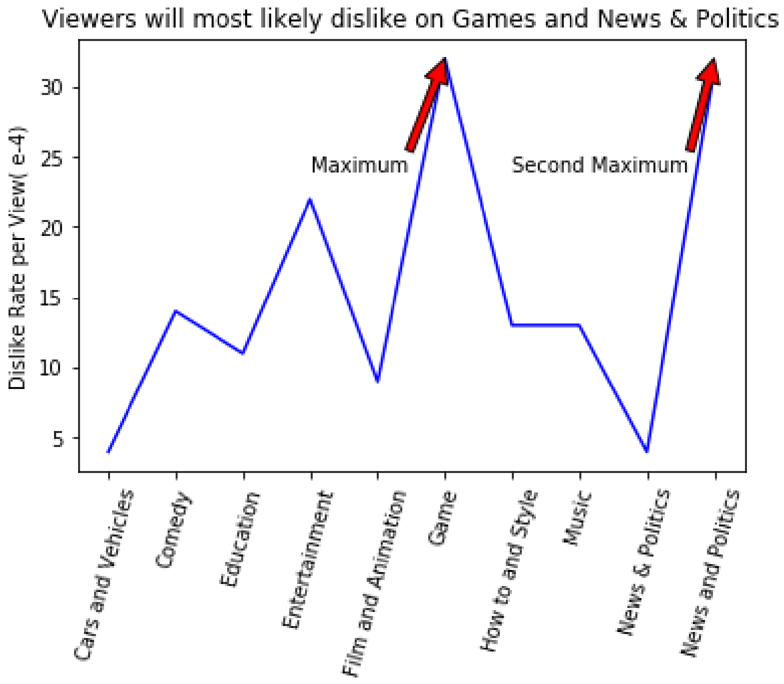
3.5 Viewers will most likely click on “like” on Comedy

By comparing like clicks for each video category and total view counts for each category, we get the conclusions that viewers mostly likely click on “like” on comedy.



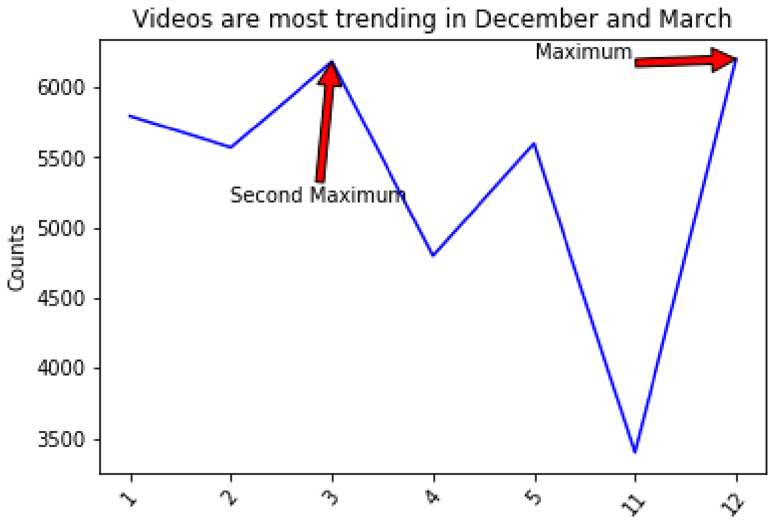
3.6 Viewers will most likely dislike on Games and News & Politics

By comparing dislike click counts for each video category and total view counts for each category, we get the conclusions that viewers mostly likely dislike on gaming videos, News and Politics.



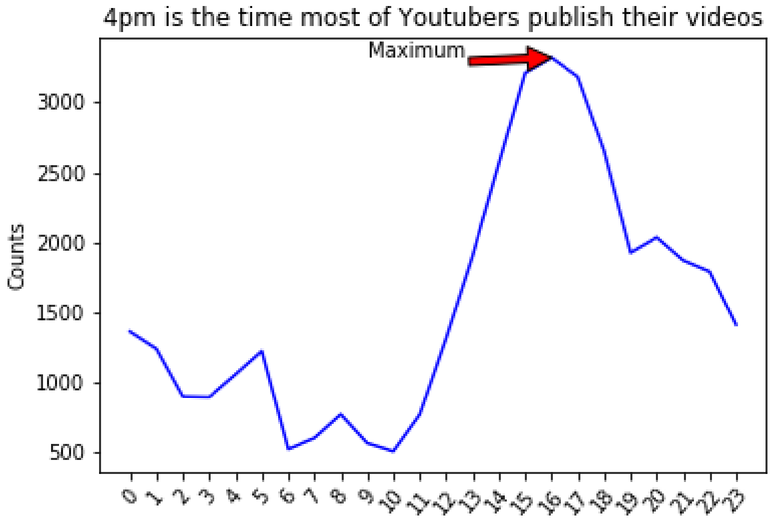
3.7 Videos are most trending in December and March

By comparing the total video counts for each month by the total trending videos overall, we get the conclusions that videos are most trending in December and March.



3.8 4pm is the time most of Youtubers publish their videos

By comparing the total video counts for each hour in day by the total trending videos overall, we get the conclusions that Youtubers most likely publish their videos at 4 pm.



7 Conclusions

7.1 Data Exploration Conclusions

Even though “Music” has the most “likes”, “dislikes”, “comment counts” and “views” based on the total counts of each feature in the dataset, but the rank differs if we compare each video category by total feature counts dividing the total view counts for each category. As a result, if we compare each category in ratio, the top performers on the YouTube trending list are music videos, but viewers will most likely leave comments on Gaming Videos, click “like” on Comedy, and click “dislike” on Games, News, and Politics.

On the other hand, videos are most trending in December and March. Also, most of the YouTubers like to publish their videos around 4 pm.