STA 9750 Final Project

Analysis of the YouTube Videos Trending in the United States

Uros Trifunovic, Mehdi Lahlou Charki, Udit Bhandari

18 December, 2020

Contents

Introduction

Project description

YouTube uses factors including but not limited to the number of views, comments, and likes to determine which videos a wide range of viewers would find interesting. These videos are listed as "Trending" and pushed to the platform users. The list gets updated roughly every 15 minutes.
The analysis looks into the trending YouTube videos in the United States to determine the factors that influence a video's popularity the most. Furthermore, it attempts to construct a model for predicting the number of views for the videos trending in the United States. We begin by exploring the relationship between the number of views and continuous variables like likes, dislikes, and comments. Next, we analyze the categorical variables such as channel, category, and tags to determine if they increase the chances of getting more views. Lastly, we look into the publishing times to see if posting a video at a particular time of the day is favorable for getting more views on a video. The data was obtained from kaggle.com.

Dataset overview

Initially, a total of 0.08% of the dataset had missing values, all of which were in the "Description" column. Due to a relatively low proportion of missing values and the irrelevance of the "Description" column for the analysis, we chose to ignore them. Additionally, the "Category ID" and "Category Title" columns contained the same information. Hence, we dropped the former to avoid redundancy as the latter was descriptive. Lastly, we removed the "Thumbnail Link" column as we were not interested in it, either. Upon cleaning the data, we are left with a dataset consisting of 40,949 rows, 14 columns, and no missing values.

The dataset provides information on, including but not limited to, the number of views and likes videos had, video creators, tags, and the date they went trending. The overview of the Trending YouTube videos in the United States dataset is below:

Continuous variables analysis

Attribute frequency distribution

We begin the analysis by exploring the relationship views have with likes, dislikes, and comments, as these variables are likely to influence the number of times a video gets seen. Looking at the distribution of the views, dislikes, likes, and comments, we observe that the values in the 99th percentile are significantly lower compared to the ones in the 100th percentile. The observation indicates that user engagement is concentrated on the small number of videos, meaning that our dataset contains outliers.

The frequency distribution plots confirm that most videos have a low number of likes, dislikes, and commentaries. It appears that there are trending videos that do extremely well while the majority get a fraction of their performance.

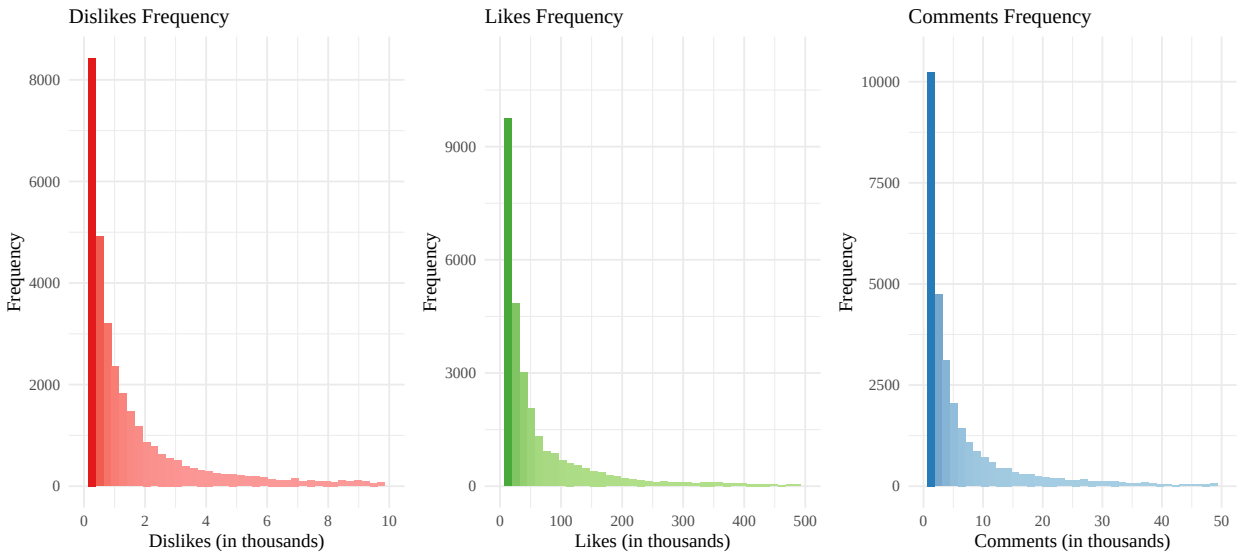
Table 1: Sample data from select columns

Title	Channel	Category	Views	Likes	Dislikes	Comment count	Trending date
WE WANT TO TALK ABOUT OUR MAR-RIAGE	CaseyNeistat	People & Blogs	748,374	57,527	2,966	15,954	11/14/17
The Trump Presi-dency: Last Week Tonight with John Oliver (HBO)	LastWeekTonight	Entertainment	2,418,783	97,185	6,146	12,703	11/14/17
Racist Superman | Rudy Mancuso, King Bach & Lele Pons	Rudy Mancuso	Comedy	3,191,434	146,033	5,339	8,181	11/14/17
Nickelback Lyrics: Real or Fake?	Good Mythical Morning	Entertainment	343,168	10,172	666	2,146	11/14/17
I Dare You: GOING BALD!?	nigahiga	Entertainment	2,095,731	132,235	1,989	17,518	11/14/17

Table 2: Views, dislikes, likes, and comment count distribution

	25%	50%	75%	90%	95%	99%	100%
Views	242,329	681,861	1,823,157	4,602,002	9,017,287	29,917,344	225,211,923
Dislikes	202	631	1,938	6,033	11,808	43,562	1,674,420
Likes	5,424	18,091	55,417	160,315	307,403	923,016	5,613,827
Comment Count	614	1,856	5,755	16,959	30,783	100,604	1,361,580

Dislikes, likes, and comments frequency distributions



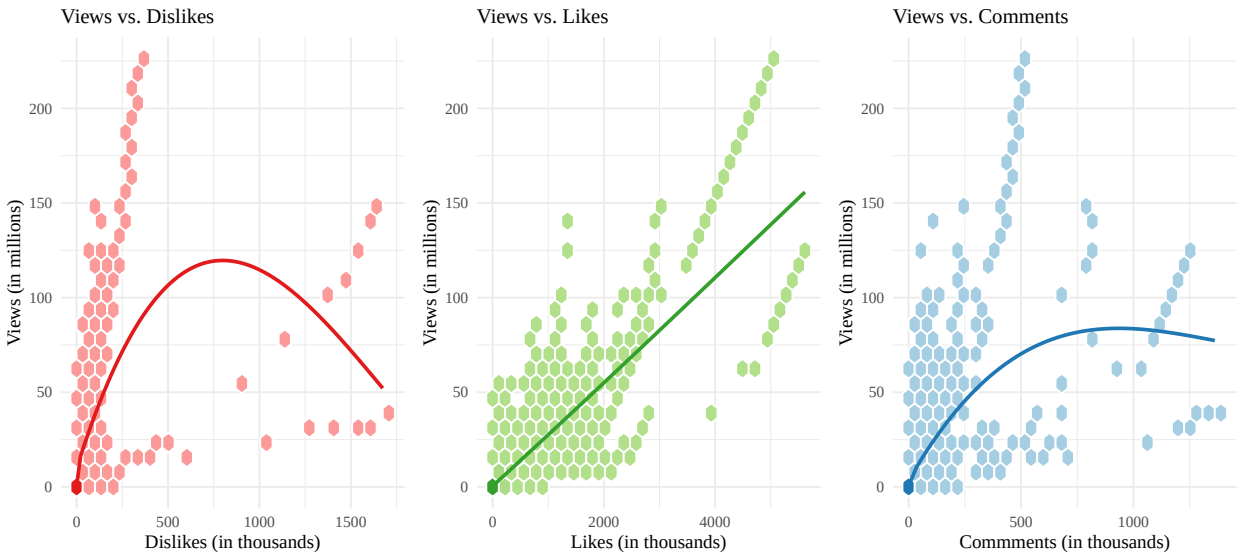
Plotting the relationships

We plot the top three most important variables against views. The plots reveal a few interesting points:

- Getting dislikes is very beneficial for increasing the number of views a video gets. Up to ~400,000 dislikes, the slope of the linear relationship between dislikes and view count is very steep. However, once a video reaches ~750,000 dislikes, the relationship becomes negative. It is possible that such videos get out of Trending or that YouTube stops pushing them to the platform users.
- The number of likes has a strong linear relationship with the number of views. Generally, a user can expect the views count to increase as the likes count rises throughout the life of a video.

- Receiving comments on videos increases the view count up until a video reaches ~500,000 commentaries. Beyond that point, the increase in view count slows down. The relationship eventually inverses.
- The majority of videos have a low number of likes, dislikes, and commentaries confirming the presence of outliers that the user engagement revolves around.

Views relationship with dislikes, likes, and comments



Fitting a linear model

Considering the linear relationship views have with dislikes, likes, and comments up to a certain point, we fit a linear model using these attributes as predictors. We split the dataset into training and test components using an 80/20 split. Then, we fit the model on the training set using the following formula:

$$Views \sim Likes + Dislikes + CommentCount$$

Linear model output

The summary table of the linear model output shows the following:

- p-value for all three variables is below 0.05, indicating their importance for predicting the number of views.
- Each dislike is expected to generate ~87 additional views.
- Each like results in ~35 more views.
- The coefficient for the comment count predictor is negative. As previously shown, the relationship between views and the number of comments starts positively but inverses eventually.
- Likes have the lowest standard error and the highest t-value, indicating that likes could be the most reliable predictor for the number of views out of the three variables used.

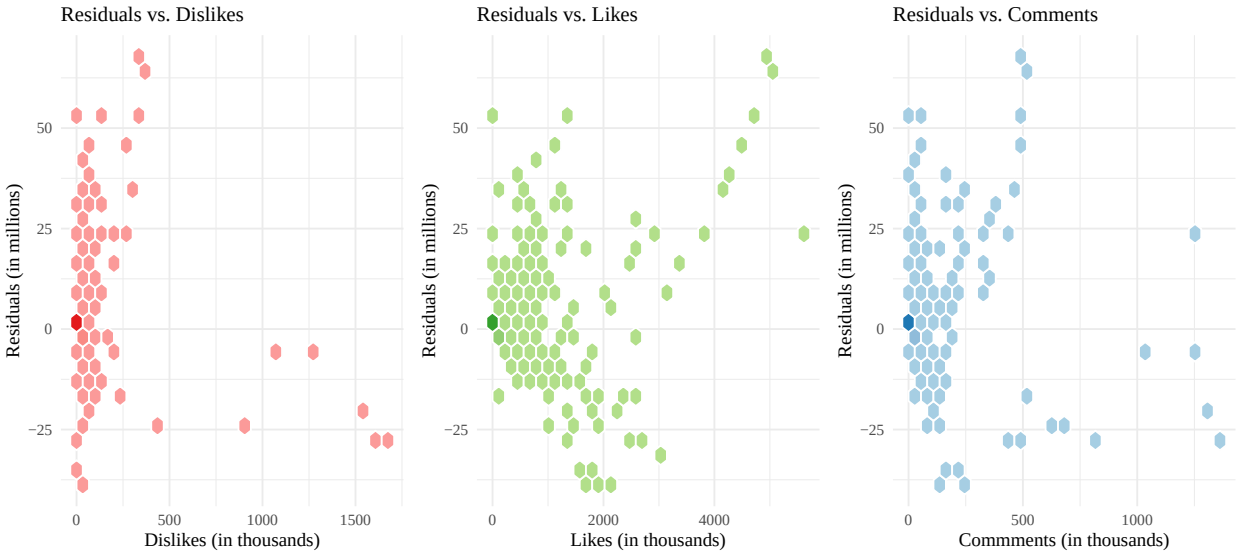
The R-squared value of 0.78 means that variation in the number of dislikes, likes, and comments can explain ~77.87% of the variation in the view count. Additionally, the Root Mean Squared Error of the linear model on the test set is 3,544,193.

The residuals plots show some large residuals among all three predictors. Furthermore, there is a pattern within "Likes" that our model fails to capture.

Table 3: Linear model summary

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	246,040	20,301.62	12	0
Likes	35	0.15	237	0
Dislikes	87	1.06	82	0
Comment Count	-96	1.11	-86	0

Dislikes, likes, and comments residuals



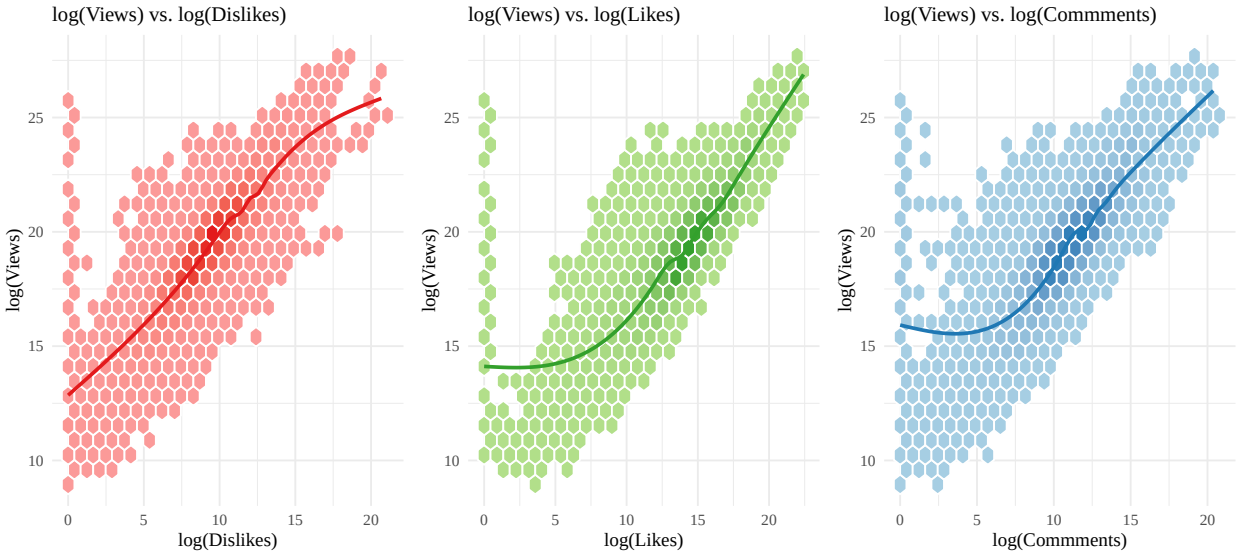
Plotting the normalized relationships

As the plots show, the relationship views have with dislikes, likes, and comment count is skewed. To transform skewed data to approximately conform to normality, we use the log transformation. The plots reveal that the log-transformation makes the patterns more linear.

Normalized views relationship with dislikes, likes, and comments

Table 4: Linear model summary

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	10.17	0.03	354.6	0.00
Likes	0.38	0.00	88.9	0.00
Dislikes	0.43	0.00	101.9	0.00
Comment Count	-0.01	0.00	-2.1	0.03



Fitting the normalized linear model

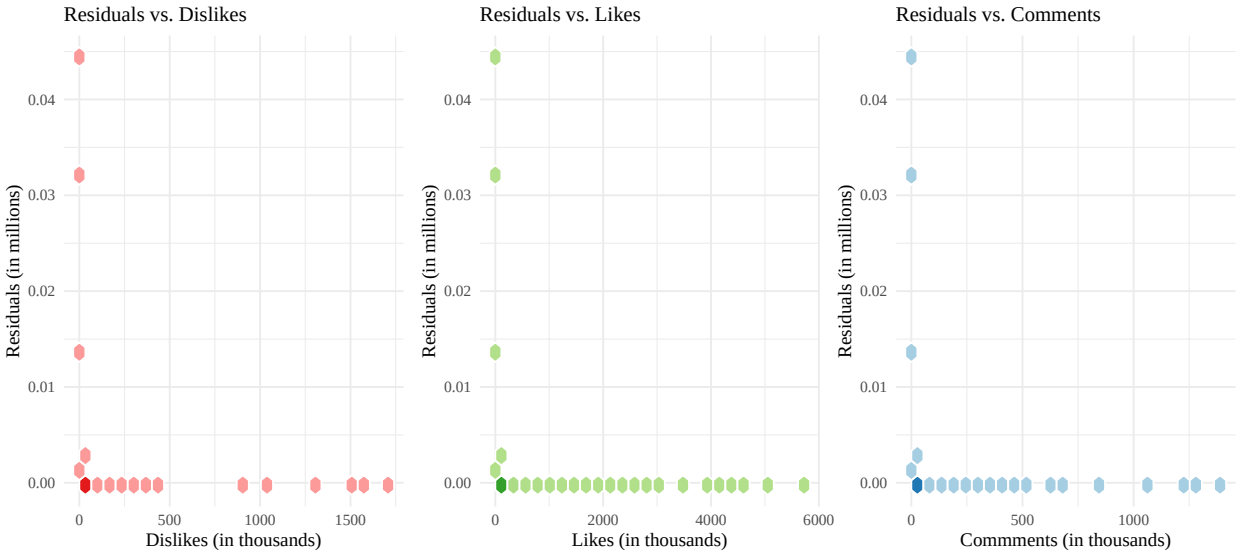
Next, we fit a linear model on the training set using the log-transformed predictors. The model is fitted with the following formula:

$$log(Views) \sim log(Likes) + log(Dislikes) + log(CommentCount)$$

Normalized linear model output

The R-squared value of 0.8 shows that the log-transformed linear model is slightly more accurate than the initial one. The RMSE of the log-transformed model is 5,241,643 compared to 3,544,193 of the earlier model.

Normalized dislikes, likes, and comments residuals The initial linear model residuals' plots have earlier revealed that there are some patterns the model fails to capture. We look at the log model residuals to see if the log-transformation took care of the problem. We back transform the residuals and plot them to see how their distribution.



With most of the residuals concentrated around 0, we confirm that the normalized linear model is a better predictor of video view count than the initial one. However, there still are some values with large residuals. To gain a better understanding of other factors impacting video popularity we turn our focus to tags next.

Categorical variables analysis

Tags analysis

YouTube tags are words and phrases used to give YouTube context about a video. They are an important ranking factor in YouTube's search algorithm. The word cloud below summarizes the tags most frequently used among the videos trending in the United States in 2017 and 2018.

Content creators repeatedly used tags like "funny" and "comedy", suggesting dominance of the entertainment videos among the trending ones. The observation makes sense as videos from this category are 24.33% of the dataset. The "how to", "music", and "pop" videos get a relatively large number of views, too.

Fitting the Random Forest model

To gain insight into which tags are the most important for a video to generate more views, we fit a Random Forest model. We focus on the channel, category, and the top 500 most frequently used tags for which we add a logical column to the dataset for each. The values in the columns are “TRUE” if the phrase appears within the video tags, otherwise “FALSE”. Then, we fit the Random Forest model on the training set using the following formula:

$$Views \sim Channel + Category + Top500MostFrequentTags$$

Random Forest model output

The Random Forest model slightly underperforms the linear one when it comes to accuracy with the R-squared value of 0.76. The model was fit on 500 with 167 predictors sampled for splitting at each node. Additionally, the Root Mean Squared Error on the test set is 6,251,189.

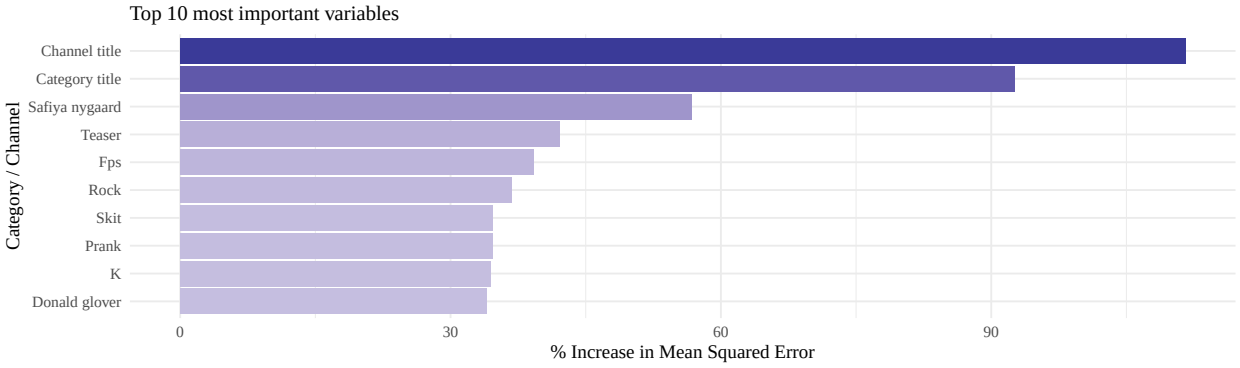
The importance chart points out a few observations:

- Certain channels are likely to get more views on their videos, probably because a large number of subscribers.

- The above is strengthened by the presence of “Safiya Nygaard”, “Liza Koshy”, “Selena Gomez”, and “Nikkietutorials” among the top 10 most important categories/channels/tags. Although the owners of the listed channels are not all strictly YouTubers, they split ~67 million subscribers among themselves.

- Videos in particular categories get seen more frequently relative to the others.
- Frames per second (fps) is one of the determinants of video quality, hinting that viewers are more likely to re-watch the better quality videos.

Top 10 most important variables among channel, category, and tags



Publishing time analysis

Most popular time of day for publishing

Next, we turn our focus to the time the creators published their videos. We begin by splitting a day into four parts and classify videos based on the publishing time as follows:

- Videos uploaded between midnight and 6 am we categorize as “Night”.
- The ones published between 6:01 am and 12 pm fall into the “Morning” group.
- Uploads 12:01 pm and 6 pm puts those videos in the “Afternoon” category.
- Lastly, videos published 6:01 pm and 11:59 pm are categorized as “Evening”.

All the times are expressed in the Coordinated Universal Time (UTC). Per the table, over two-thirds of the trending videos were uploaded in the afternoon and evening hours. Moreover, around 45% of all the videos were published in the afternoon, between 12:01 pm and 6 pm UTC.

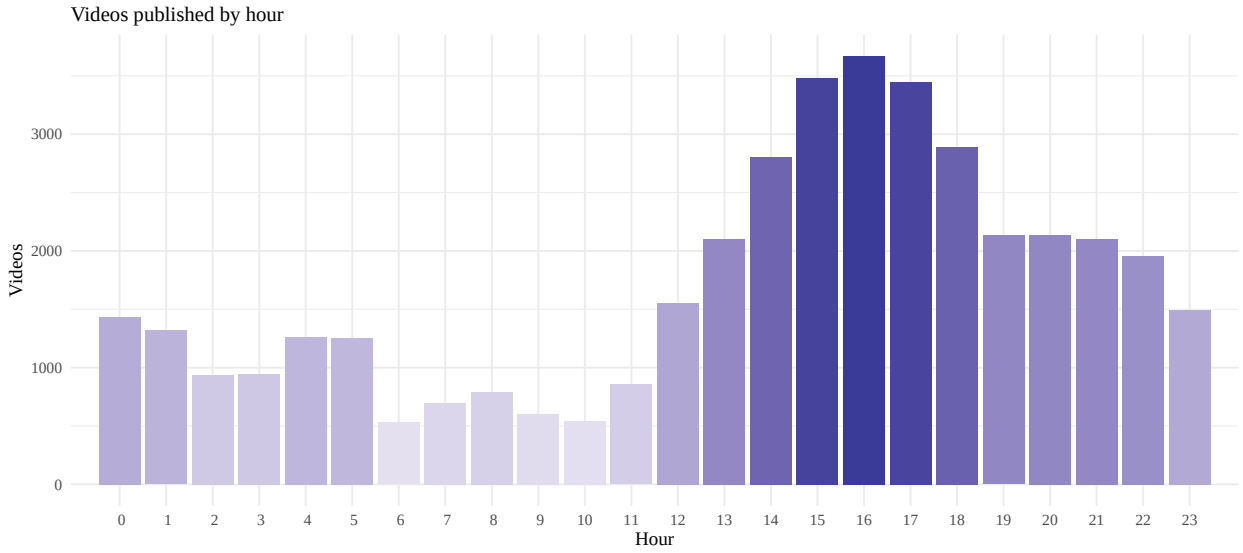
Table 5: Videos published by time of day

Time of day	Videos	Percentage
Afternoon	18,400	45
Evening	9,826	24
Night	7,680	19
Morning	5,043	12

Most popular hours for publishing

The chart below confirms the claim the afternoon hours are the most popular ones for publishing a video as most videos were uploaded to YouTube between 3-5 pm UTC. Upload frequency decreases after 6 pm UTC and remains low before picking up around noon the following day.

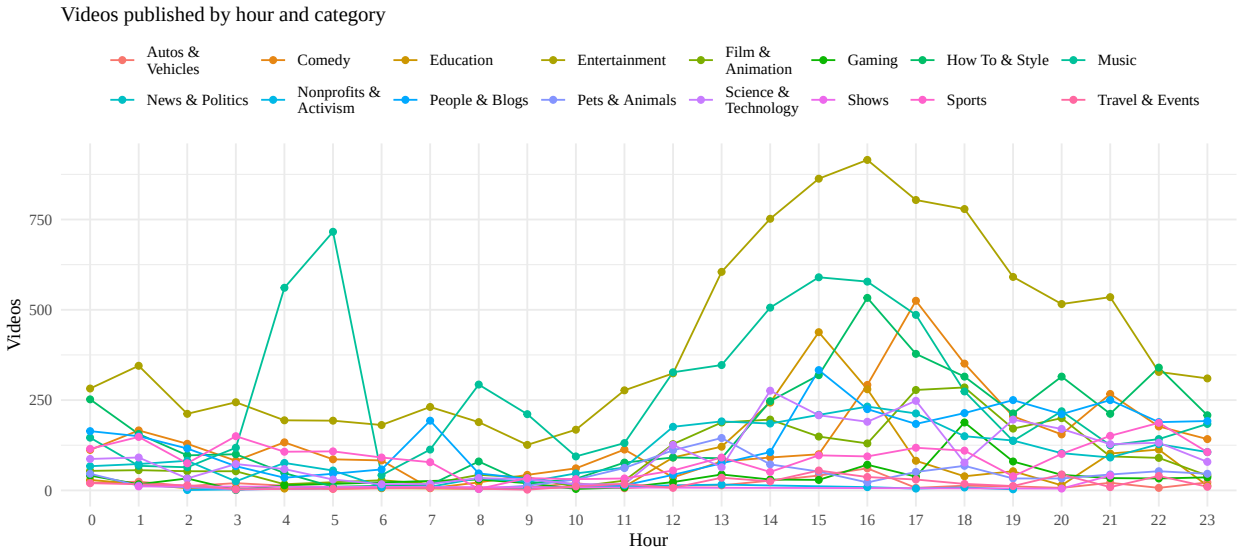
Number of videos published by hour



Most popular hours for publishing, by category

We look into the most popular time of day to publish a video for each category. As expected, most video categories have their peak publishing time between 3 pm and 5 pm UTC. However, Music, as the second most popular category, has seen most videos published at 5 am UTC.

Number of videos published by hour and video category



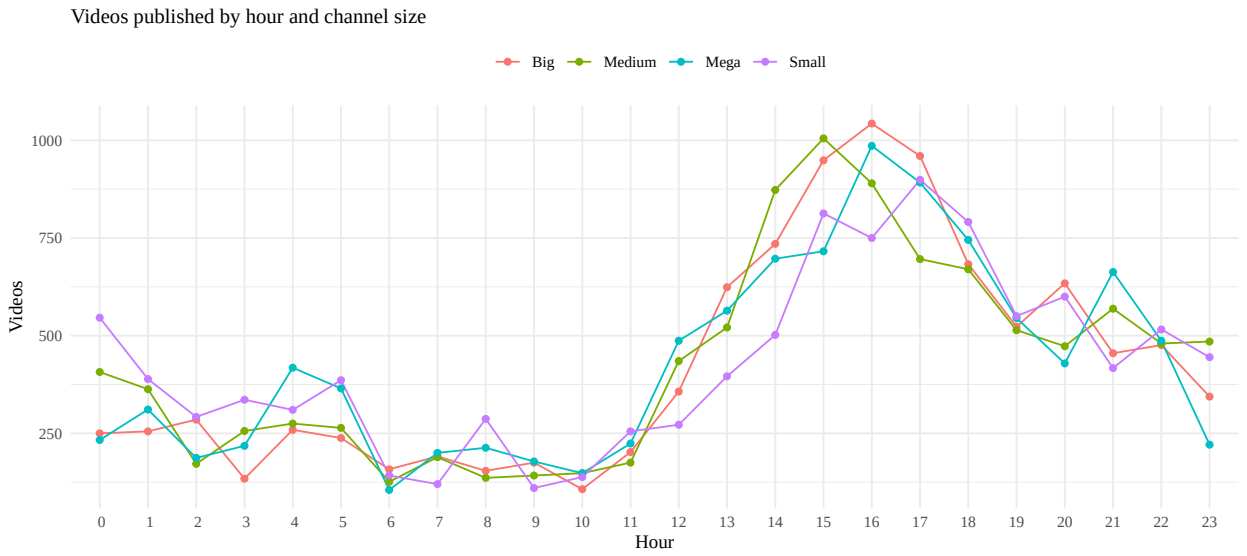
Most popular hours for publishing, by channel size

We split the channels by size to see whether more popular channels have different peak publishing time relative to the rest. We split the YouTube channels into four groups based on the total number of views channels have in the following way:

- "Mega" channels' view count is larger than the 99th percentile.
- "Big" channels' have view count between the 75h and 99th percentile.
- "Medium" channels' total views are between the 25th and 75th percentile.
- "Small" channels' view count is below the 25th percentile.

It seems that peak posting time doesn't change much based on the channel size as channels of all sizes have their peak publishing time between 3 pm and 5 pm UTC.

Number of videos published by hour and channel size



Putting everything together

Fitting the aggregate Random Forest model

As the final component of our analysis, we fit another Random Forest model on the training set based on the previous observations regarding variable importance. We use likes, dislikes, comment count, top 200 most important variables among category, channel, and tags, as well as publish time of day as determinants to predict the view count for a video. We ignore the hours at which videos were uploaded as those ultimately determine the time of day variable. Therefore, the formula for the final Random Forest model is as follows:

$$Views \sim Likes + Dislikes + CommentCount + Top200MostImportantCategoricalVars + TimeOfDay$$

Aggregate Random Forest model output

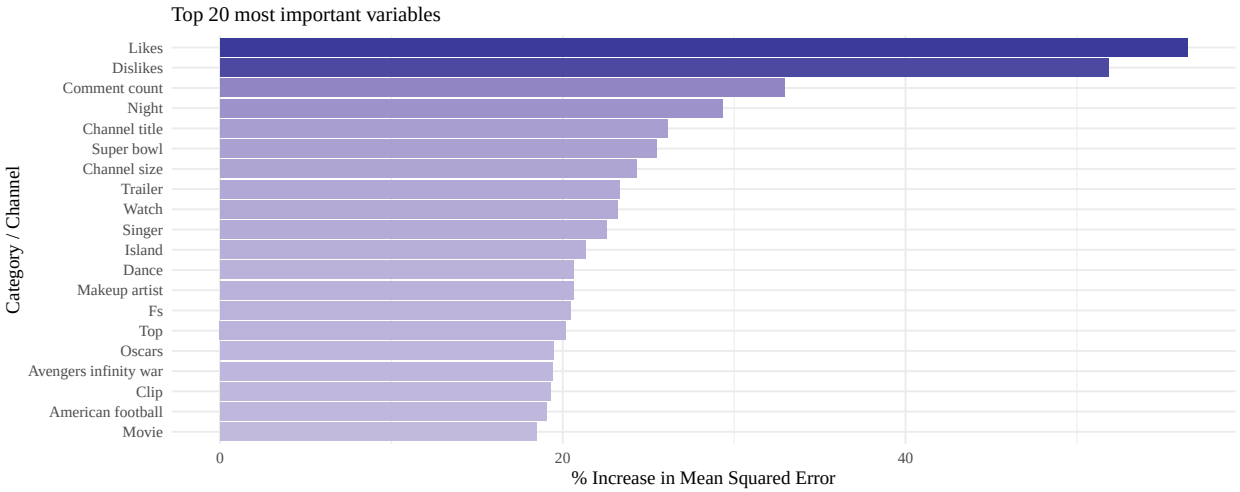
The final model noticeably outperforms both the linear model and the earlier Random Forest model. With the R-squared value of 0.98, the model explains ~97.61% of the variation in the view count. Like the previous model, this one was fit on 500 but was sampling 70 predictors for splitting at each node. The RMSE on the test set is 1,315,841.

The importance chart highlights the following:

- Likes, dislikes, and comment count are the most important for determining the video view count.
- Entertainment, music, and sports videos get more views than the others .
- Although the majority of the creators of the trending videos publish to YouTube in the afternoon hours, it appears that the ones uploaded between midnight and 6 am UTC get the most views. As previously identified, these are most likely Music videos. Viewers are more likely to repeatedly watch music videos when they fall in love with a song.

- Viewers enjoy watching trailers for the movies .
- YouTube users care about the video quality expressed in the number of frames per second and the slow-motion functionality.

Top 20 most important variables among channel, category, and tags



Model comparison

In summary, we compare the four models to see which one is best for predicting the view count for YouTube Trending videos. We look into the accuracy of the linear model fitted on the continuous variables ("Linear (cont)"), the log-transformed linear model ("Linear (log)"), the Random Forest model fitted on the select categorical variables ("Random Forest (cat)"), and the Random Forest model fitted on the most important continuous and categorical variables ("Random Forest (agg)").

The results show the following:

- Out of the four models, Random Forest with a channel, category, and top 500 most frequently used tags as determinants is the worst predictor of a video view count.
- The linear model halved the previously-mentioned model's RMSE and has a slightly better R-squared value.
- The model with log-transformed variables further improves the R-squared value. However, its RMSE is worse than the initial linear model's.
- All three models fall far behind the aggregate Random Forest model. Its R-squared value of 0.98 is substantially higher than those of the other two models. Furthermore, the model's RMSE is ~2.7x smaller than that of the linear model, ~4x smaller than that of the normalized linear model, and ~4.8x smaller than that of the other Random Forest model.

Table 6: Linear vs. Linear log vs. RF Cat vs. RF Agg

Model	RMSE	R-squared
Linear (cont)	3,544,193	0.78
Linear (log)	5,241,643	0.80
Random Forest (cat)	6,251,189	0.76
Random Forest (agg)	1,315,841	0.98

Conclusion and the next steps

The analysis shows that even getting dislikes on a video is beneficial for increasing the views count up to a certain point. As expected, more likes lead to more views while the rising comment count eventually stops contributing to getting views. Additionally, there are videos that do exceptionally well. Those are most likely entertainment and music videos. Although the majority of the videos are published in the afternoon, the ones uploaded overnight are more important for predicting the view count. The next step of the analysis could be to narrow its focus to these videos that outperform the others and try to understand the driving forces behind their performance.