YouTube Trending List:
An Analysis of the Factors That Influence Video Trending Duration

Brendan Sigale
Brianna Mueller
McKenzie Fuller
Sung Chul Hong

Data and Decisions MSCI:9100
Fall 2019

Contents

Executive Summary

Opportunity

As the prevalence of social media has increased, companies have taken advantage of various social media platforms to reach target customers through pay-per-click and pay-per-view advertising. Marketers are tasked with analyzing a multitude of factors when deciding where to invest in advertising. As a high traffic platform, YouTube is an online video streaming site where advertisers can pay to place their ads on popular videos. The purpose of this report is to examine factors that may predict the number of days a video remains trending after appearing on YouTube's trending list - knowledge that could be valuable to marketers to identify relevant videos for effective ad placement.

Findings

The following independent variables were examined in this analysis: count of first day views, count of first day likes, count of first day dislikes, count of first day comments, time of the week the video appeared on the trending list, and video category. Due to reasons explained in the methodology section of this report, video category was not a variable examined in our main analysis using regression modeling. A separate regression was performed with video category included as a variable of interest. The findings are as follows:

- In the primary regression analysis, first day likes, first day views, and first day dislikes were found to be significant variables at the p=.05 level.
- In our secondary analysis including the top 5 YouTube video categories, we found that the categories "Music" and "How-to & Style" were significant predictors of total days trending, in addition to first day views, first day likes, and first day comments.

Recommendations

Based on the results of our analysis, the following recommendations have been made with marketers in mind:

- Take in consideration the number of first day views, first day likes, first day dislikes, and video category while making decisions regarding video ad placement.
- Explore other factors that may explain variation in total video trending time. Due to the low adjusted R-squared value of our regression, our model should only be used as part of a larger method for predicting total trend time.

Please see the remainder of the report for a more detailed description of the variables of interest, our methodology, findings from our analysis, and a discussion regarding the value of the model.

Opportunity Statement

As advertising and marketing have evolved, companies have adjusted their strategies for reaching a target audience. Traditional marketing channels have decreased in effectiveness while modern marketing tactics such as email marketing, social media advertising, and pay-per-click advertising have proven to be increasingly effective. Marketing departments are presented with the challenge of deciding which channels they should take advantage of to reach their target audience and how to optimize their presence on each platform.

YouTube is the most popular online video streaming site today with 27% of worldwide internet users visiting the site at least daily (Clement, 2019). With the high volume of users, advertising on YouTube presents ample opportunities for companies to attract customers. Our goal is to create a model to predict the number of days a video will remain trending after it appears on the YouTube trending list. This information would be valuable to marketing teams to guide their selection of which videos to place their advertisements on in order to maximize their reach.

Dependent Variable

The dependent variable of interest in this analysis is the number of days a single YouTube video is on the trending list. This variable was defined as a count of the total days between the first day the video appeared on the trending list and the last day it was reported to appear on the trending list. Since videos can trend over non-consecutive days, we decided to disregard any days within the first and last day of appearing on the trending list that the video did not make it on the trending list.

Since the YouTube trending list is not personalized to each user, YouTube aims to present videos that a vast range of viewers will find interesting. The trending list displays videos from diverse channel categories, with the goal of highlighting a variety of video creators. Ultimately, the videos shown on the trending list at any given moment are meant to provide an accurate representation of the current state of YouTube and what's happening globally. Every 15 minutes, the list is updated and videos may move up and down in the rankings, or be removed from the trending list entirely. Each country can have a different trending list, but for the sake of our analysis we are focused on U.S. trending videos from 2017 - 2018.

YouTube has released a few pieces of information on what determines if a video appears on the trending list. According to the platform, the following signals play a role in a video's likelihood of appearing on the Trending list: view count, how quickly the video is generating views, where the views are coming from outside of YouTube, and the age of the video (n.d.). These factors help predict which videos are likely to make its first appearance on the trending list. However, in our analysis we examine factors that may influence a video's total days trending *after* it appears on the list and is classified as a trending video.

Independent Variables

To conduct our analysis, our group obtained a dataset published on Kaggle that holds information pertaining to all the US videos that were trending on YouTube from 2017 - 2018. This dataset was collected using the YouTube KPI. After an initial examination of the data, our group was able to identify independent variables of interest and derive new variables from existing values. The overall purpose of our model is to predict how long a video will remain trending after the first day. Although data was available for each day a video was trending, we only analyzed the video's statistics on the first day it appeared on the trending list. Below is a description of each independent variable used in our analysis, along with our initial hypothesis for each variable.

Number of Views On the First Day

The first independent variable we looked at was the total number of views a video had on the day it started trending. We hypothesized that a greater number of views on the first day a video appears on the trending list would result in a longer video trending time.

Number of Likes On the First Day

The second independent variable of interest is the number of likes accumulated by the first day a video appears on the trending list. We hypothesized a greater number of likes on the first day a video appears on the trending list is expected to result in a longer video trending time.

Number of Comments On the First Day

The next independent variable of interest was the number of existing comments the first day a video appears on the trending list. We hypothesized a greater number of comments observed the first day a video appears on the trending list is expected to result in a longer video trending time.

Number of Dislikes On the First Day

We also examined the number of dislikes a video receives by the first day it appears on the trending list. We hypothesized a greater number of dislikes a video has accumulated by the first day it appears on the trending list is associated with a shorter video trending time.

Weekday Or Weekend

One categorical variable incorporated in our initial model is the time of the week. This variable was split into two groupings: weekday or weekend. To make these groupings, we examined the date a video first appeared on the trending list and noted if this day was on a weekend (Saturday, Sunday) or weekday (Monday, Tuesday, Wednesday, Thursday, Friday). We hypothesized videos that first appear on the trending list on weekends are associated with longer video trending times.

Video Category

The second categorical variable examined in our analysis was video category. The original column in our data set for video category contained the numerical values 1-44. A separate text file contained information specifying the video category associated with each number. Due to reasons further explained in the methodology section of this report, we decided to perform our main analysis without incorporating this variable. We then created an additional model incorporating the top 5 video categories. These categories were selected by examining a Pivot table of the average number of days videos trend in each category. We hypothesized video category plays a role in predicting the number of days a video will remain trending. This hypothesis was based on the idea that certain video categories are more popular than others.

Description of Methodology Used

As previously mentioned, the data used in our analysis was readily available from Kaggle. Once the CSV file was uploaded to Excel, we performed various steps to clean and prepare the data set for regression. Unnecessary columns were deleted so that only variables of interest remained. After rudimentary cleaning tasks were performed in Excel, we transported the data to R for more advanced data manipulation. Since the dataset contained separate rows for each day a video appeared on the trending list, we condensed the data to only include statistics for the first day a video appeared on the trending list.

In order to overcome this obstacle, we executed commands in R to group rows by video title, channel title, video category ID, and publish date. We then examined the following columns for the grouped data: total trend time, minimum view count, minimum like count, minimum dislike count, and minimum comment count. The total trend time was calculated by subtracting the maximum trending date from the minimum trending date. Minimum values for views, likes, dislikes, and comments were chosen because these variables would have their lowest count on the first day a video appears on the trending list. The date was used to derive a new column with factor levels ‘weekend’ and ‘weekday’.

Once the dataset was in the desired format, we exported the file back to Excel to perform a multiple linear regression. Due to limitations in Excel, we were unable to perform a regression with all 26 video categories. We made the decision to first create a model without including video category as an independent variable. While we were unable to perform a regression with all 26 video categories, we were still interested in examining the influence of this categorical variable on total trending time. Therefore, we created an additional model by condensing the dataset to only include videos from the top 5 video categories. The findings for this model are discussed in a separate section of the report (Extended Analysis - Video Category). In both models, backwards regression was performed to eliminate insignificant variables (defined by having a P-value less than .05)

Final Model

The summary output for our final Regression Model (not including video category) can be found in Appendix 1. The significant independent variables included in the final model were first day views, first day likes, and first day dislikes. The insignificant independent variables eliminated through backwards regression were first day comments and time of the week. The following equation defines the final model:

Days Trending = 5.98 + 2.72E-07(First Day Views) + 3.80E-06(First Day Likes) -1.11E-05(First Day Dislikes)

Discussion

While the final Regression Model contained significant independent variables (determined by having coefficients significantly different than 0), we examined additional factors to evaluate the model's power to predict the number of days a video will trend. In this model, the intercept 5.91 makes sense in the context of our data. With no additional first day likes, views, or dislikes, a video is predicted to remain on the trending list for 5.98 days. The coefficient values also make sense and match our original hypotheses:

- All else being equal, an additional 10 million views on the first day a video appears on the trending is predicted to increase total trend time by 2.72 days.
- All else being equal, an additional million likes on the first day a video appears on the trending list is predicted to increase the total trend time by 8.35 days.
- All else being equal, an additional 100,000 dislikes on the first day a video appears on the trending list is predicted to decrease the total trend time by 1.62 days.

We also examined the residuals of our model. To investigate if the errors are independent of the values of each significant variable, we plotted the residuals against the individual data points of each independent variable included in our final model. These residual plots can be found in Appendix 3. In general, the residuals appear to be randomly dispersed. In all three significant variables, there is a higher density of residuals on the lower end of the X axis but this may be attributed to the data set containing fewer videos with views, likes, and dislikes havings counts in the upper range. Additionally, we created a histogram to examine the distribution of residuals. The residuals did not appear to be uniformly distributed as the shape of the histogram is skewed right.

Since a relatively small standard error is an indicator of a quality model, we compared the model's standard error to 10% of the average number of days a video remains on the trending list. The standard error, 4.55, was significantly larger than 10% of the average number of days a video remains trending (.63) - not lending support to our model. The last value considered while

evaluating the fit of the model, was the adjusted R-squared value. The model's low adjusted R-squared value of .035 indicates a low percentage of variation in the dependent variable is explained by the independent variables. This may suggest there are a multitude of factors not included in our model that influence the amount of days a YouTube video remains on the trending list.

Extended Analysis - Video Category

As previously noted, the high number of video categories made it difficult to perform multiple regression using the tools available to us at this time. However, we were able to examine the influence of this variable by condensing our data to only include videos from the top 5 video categories.

The summary output for our final Regression Model when video category was included as an independent variable can be found in Appendix 2. The significant independent variables included in the final model were first day views, first day likes, first day comments, video category 10 (Music), and video category 26 (How-to & Style). The insignificant variables eliminated through parsimony were time of the week, first day dislikes, video category 22, video category 23, and video category 24. The following equation defines this model:

Days Trending = 5.91 + 1.30(Category 10) + 8.35E-01(Category 26) + 1.79E-07(First Day Views) + 5.03E-06(First Day Likes) -1.62E-05(First Day Comments)

While we did not perform a full evaluation of the fit of this model, we examined a few key indicators of an appropriate model to examine the possible role that video category plays in predicting the amount of days a video remains trending. First, the intercept 5.91 makes sense in the context of our data. When there is no input from the independent variables included, a video is predicted to remain on the trending list for 5.91 days. With the exception of independent variable first day comments, the coefficients also make sense:

- A video in category 10 (Music) is predicted to remain on the trending list 1.3 days longer than videos in other categories.
- A video in category 26 (How-to & Style) is predicted to remain on the trending list .835 days longer than videos in other categories.
- All else being equal, an additional 10 million views on the first day a video appears on the trending list is predicted to increase total trend time by 1.79 days.
- All else being equal, an additional 1 million likes on the first day a video appears on the trending list is predicted to increase total trend time by 5.03 days.
- All else being equal, an additional 100,000 comments on the first day a video appears on the trending list is predicted to decrease total trend time by 1.62 days (This goes against our intuition for how comments would influence the dependent variable).

As was the case with our main regression model, this model did not have a standard error less than 10% of the average of our dependent variable values. Furthermore, the resulting adjusted R-squared value of .043 is much lower than a desired value close to 1.

Recommendations

Taking into account all of the factors examined in evaluating our main regression model - context of the intercept and coefficients, residual plots, residual distribution, size of the standard error, and the adjusted R-squared value - we conclude that our significant independent variables can be used to predict the duration a YouTube video remains trending to an extent. In a business context, marketers can analyze the number of likes, views, and dislikes a video acquires on the first day it appears on the trending list to predict which videos will trend the longest. For example, marketers can maximize their companies reach by placing advertisements on videos with greater first day likes and views, and few first day dislikes.

The additional regression model created with the top 5 video categories also indicates some video categories are predicted to trend longer than other video categories. Therefore, marketers should also take video category into consideration as they seek to maximize their return on investment with YouTube advertising. Videos in the categories “Music” and “How-to & Style” were determined to be significant predictors of how long a video would stay trending. Due to their significance, we recommend putting ads on videos trending in the Music and How-to & Style categories to gain the most views possible.

Limitations and Future Considerations

As discussed in the Findings section of this report, we noted a few limitations of both models while evaluating the summary output. While we believe these models hold value, we would have liked to observe lower standard errors and higher adjusted R-squared values. The observed adjusted R-squared values suggest there are other factors not examined in this analysis that may help explain variation in total video trending time. Considerations for future analysis include the following:

- Examine additional independent variables that may increase the model’s power to predict total trending time:
    - Time of day a video was posted
    - Duration of the video
    - Text analysis of the video title, description, and video tags
    - History of the video channel performance, number of channel subscribers
- Gather additional information YouTube has released regarding their algorithms for selecting videos on the trending list
- Seek out the most recent data as video statistics are continuously generated through YouTube’s KPI

APPENDIX 1: Final Regression (Video Category Not Included)

SUMMARY OUTPUT

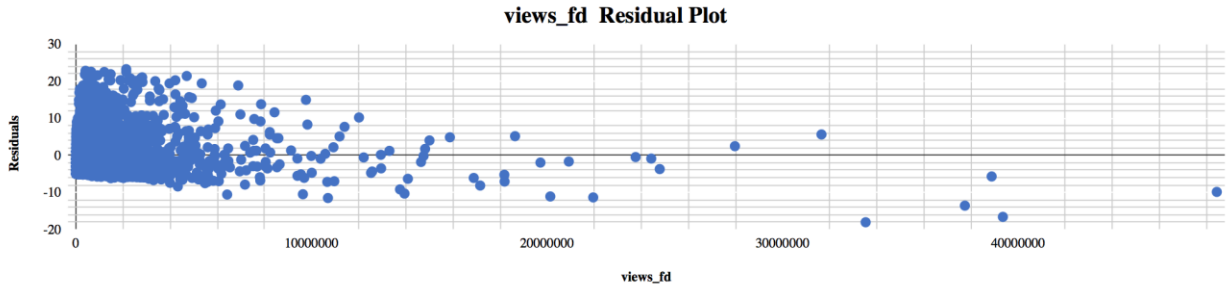
Regression Statistics	
Multiple R	0.18706
R Square	0.034992
Adjusted F	0.034545
Standard I	4.547838
Observatic	6490

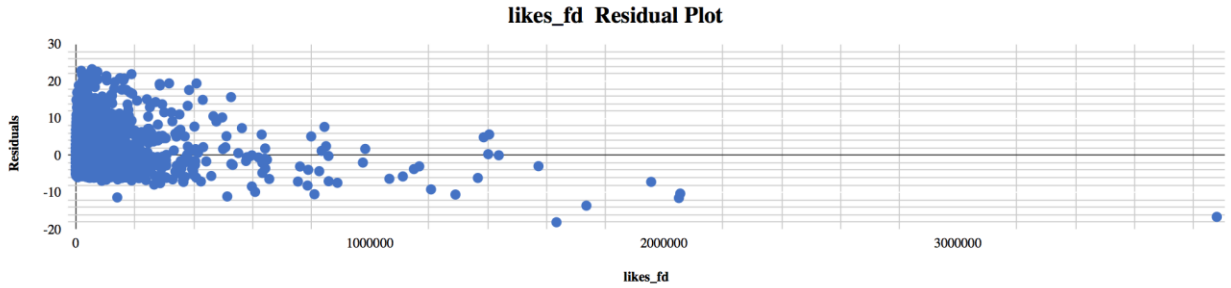
ANOVA

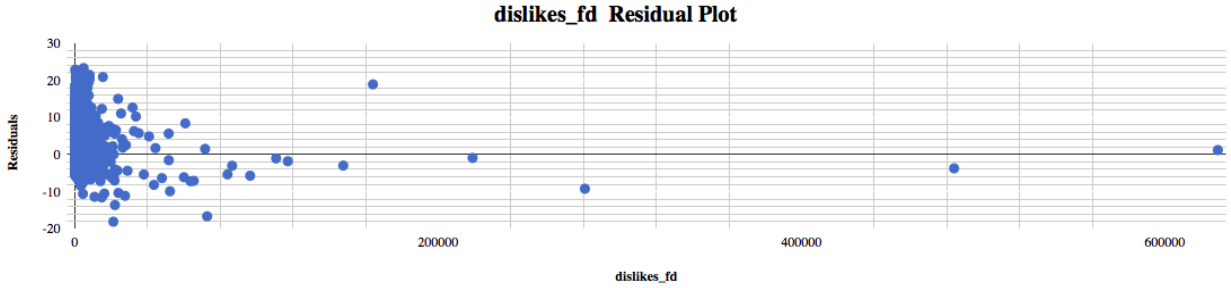
	df	*SS*	*MS*	*F*	*gnificance F*
Regressior	3	4864.278	1621.426	78.3948	8.25E-50
Residual	6486	134148.8	20.68283		
Total	6489	139013.1			

	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	5.984257	0.060708	98.57516	0	5.86525	6.103264	5.86525	6.103264
views_fd	2.72E-07	4.43E-08	6.12769	9.44E-10	1.85E-07	3.58E-07	1.85E-07	3.58E-07
likes_fd	3.8E-06	7.45E-07	5.105442	3.39E-07	2.34E-06	5.26E-06	2.34E-06	5.26E-06
dislikes_fc	-1.1E-05	5.16E-06	-2.15195	0.031438	-2.1E-05	-9.9E-07	-2.1E-05	-9.9E-07

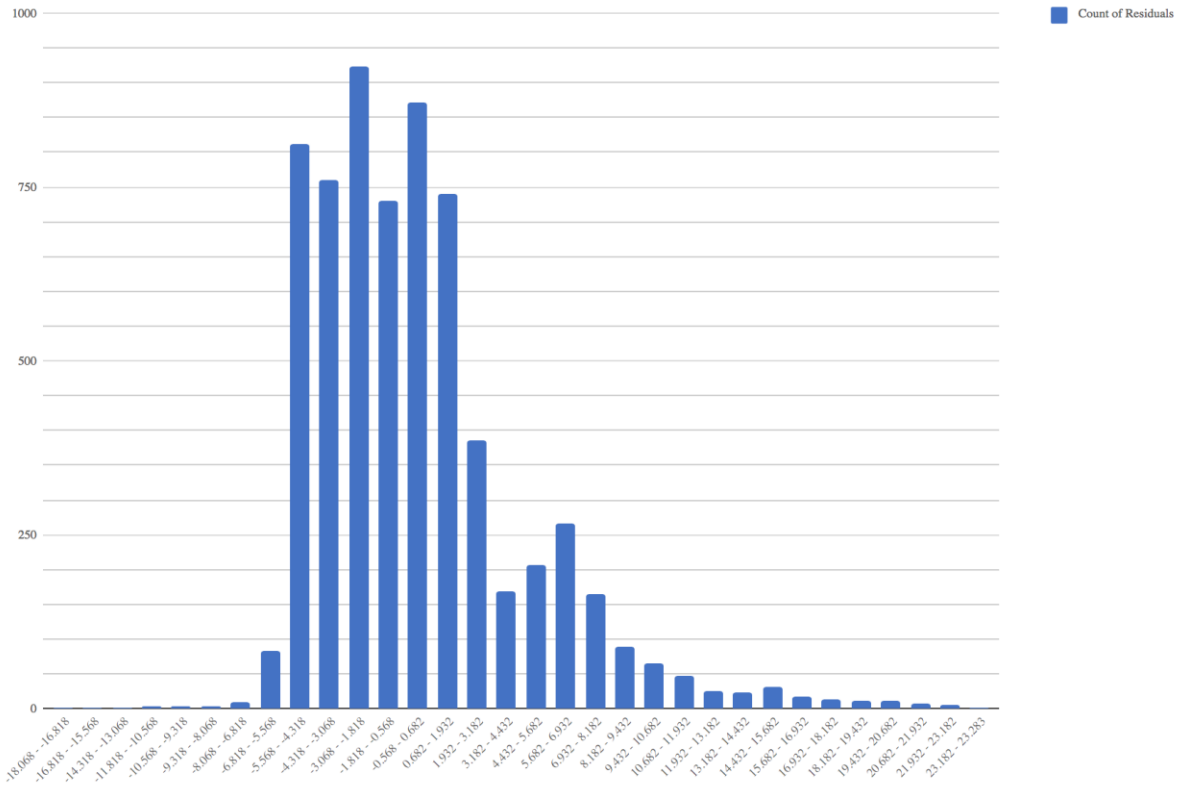
APPENDIX 2: Significant Variable Residual Plots







APPENDIX 3: Residual Histogram



APPENDIX 4: Final Regression (Video Category Included)

Regression Statistics	
Multiple R	0.209669
R Square	0.043961
Adjusted R	0.042805
Standard Er	4.689599
Observatio	4139

ANOVA

	df	*SS*	*MS*	*F*	*ignificance F*
Regression	5	4179.572	835.9144	38.00935	2.98E-38
Residual	4133	90894.34	21.99234		
Total	4138	95073.91			

	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	5.910414	0.09379	63.01742	0	5.726535	6.094293	5.726535	6.094293
cat_10	1.295661	0.192504	6.730578	1.92E-11	0.91825	1.673071	0.91825	1.673071
cat_26	0.834554	0.21167	3.942704	8.19E-05	0.419566	1.249541	0.419566	1.249541
views_fd	1.79E-07	4.93E-08	3.632826	0.000284	8.24E-08	2.76E-07	8.24E-08	2.76E-07
likes_fd	5.03E-06	1.13E-06	4.460343	8.4E-06	2.82E-06	7.24E-06	2.82E-06	7.24E-06
comment_c	-1.6E-05	4.91E-06	-3.29155	0.001005	-2.6E-05	-6.5E-06	-2.6E-05	-6.5E-06

Works Cited

Clement, J. (2019, June 25). Topic: YouTube. Retrieved from https://www.statista.com/topics/2019/youtube/.

Trending on YouTube - YouTube Help. (n.d.). Retrieved from https://support.google.com/youtube/answer/7239739?hl=en.