Yitong Li, Alec Linse, Ohmar Myint, Areeb Ahsan

IS 451: Business Data Analytics

9 March 2020

How Do Trending YouTube Videos Receive More Likes?

Background and Objective

YouTube, the world-famous video sharing website, maintains a list of the top trending videos on its platform. According to *Variety* magazine, "YouTube uses a combination of factors including measuring users' interactions (number of views, shares, comments, and likes) to determine the year's top-trending videos". This does not mean that all are videos with the highest number of views, but rather a combination of views, shares, comments, and likes to generate the top-trending videos.

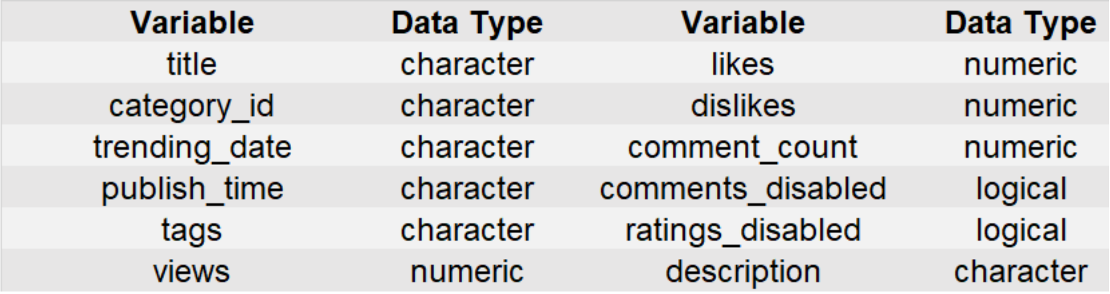
Compared to the number of views, the number of likes is a more accurate predictor of how successful videos are because it measures both the ability of the video to spread and its popularity. Thus, our objective is to identify the key factors which lead to successful YouTube videos (i.e. the large number of likes), which will be helpful for content creators, as they will be able to efficiently focus their efforts on certain aspects of the process, such as choosing the right titles, tags, when to post, etc.

Description of Data

Although the original data set includes records from multiple regions, our project is focused on trending videos in the United States. In turn, the dataset we used explores YouTube trends based on 16 different variables including, but not limited to, title, category, publish date, trending date, views, likes, dislikes, and the number of comments per video. It contains more than 20,000 valid records from 2006 to 2018, with most having concentration between November 2017 and June 2018.

Data Exploration

As mentioned before, the dataset has 20,263 valid records based on 16 different variables. Excluding some variables that have no real meaning, the data type of all variables are as follows:



Variable	Data Type	Variable	Data Type
title	character	likes	numeric
category_id	character	dislikes	numeric
trending_date	character	comment_count	numeric
publish_time	character	comments_disabled	logical
tags	character	ratings_disabled	logical
views	numeric	description	character

Titles and Description

To see the keyword of trending videos' titles and description, we created a word cloud based on text mining. Please refer to graphic in Appendix A. The word cloud on the left shows the titles and the many words relating to the most popular category (which we found to be entertainment), while the word cloud on the right shows words relating to the promotion of the video through social channels.

Categories

We can see that the Entertainment category contains the largest number of trending videos with around 5,000, followed by the Music category with over 3,000. The remaining significant categories contain anywhere from 1,000 to 2,000 trending videos, while the categories of News & Politics and Non-Profits & Activism each contain about 30 videos. These two represent the smallest categories. (See graphic in Appendix B).

Trending Date and Publish Time

The data summarized in this text was collected between November 2017 and June 2018, with even distribution among each day of the week. As for the videos publish time, we can see a significant increase in the number of videos posted during the hours of 10am to 3pm, and this number peaks from 3pm to 5pm. To the contrary, there are also large numbers of videos posted between the hours of 7pm and 10pm. Appendix C graphics represent these findings.

Distribution of Numeric Variables

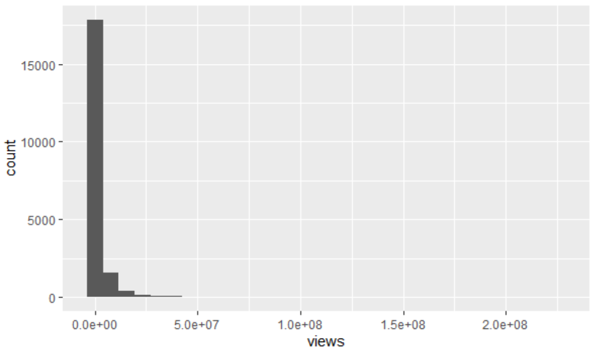
Views

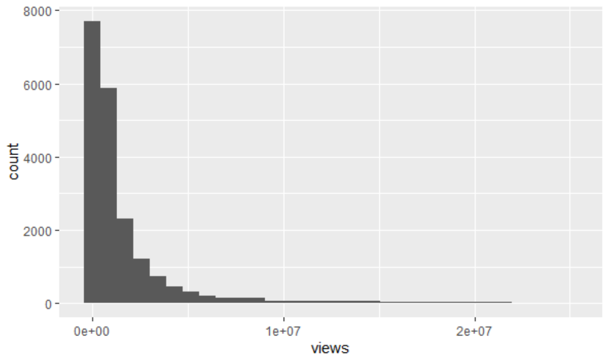
To assist in understanding the following data, please see Appendix B. According to the histogram displaying the number of views on the left, most trending videos have

12.5 million views or less. We get this figure by calculating:

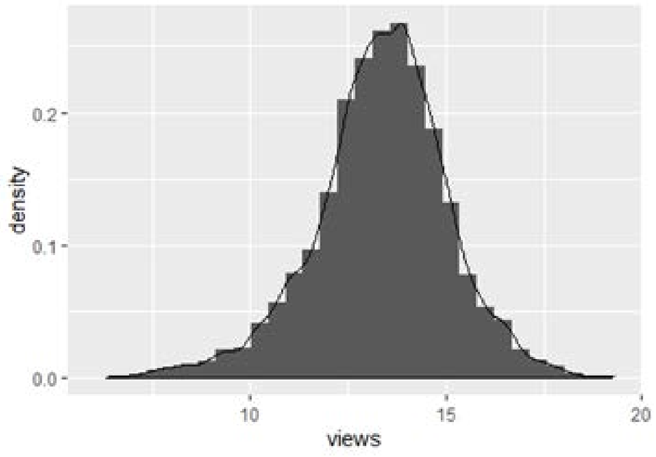
$$\mathbf{\frac{2.5\times 10^{7}}{2}=12.5\times 10^{6}}$$

We then isolated those videos (shown on the right) with 12.5 million views or less, which gives us a more precise look at the distribution of the data. Based on these distributions, it is safe to say that most of the data is confined within a small range even though the distribution has a long tail covering an expansive range. What's more is that this variable has a high order of magnitude. All of this indicates that it is not a good idea to directly use the number of views as a sole predictor.



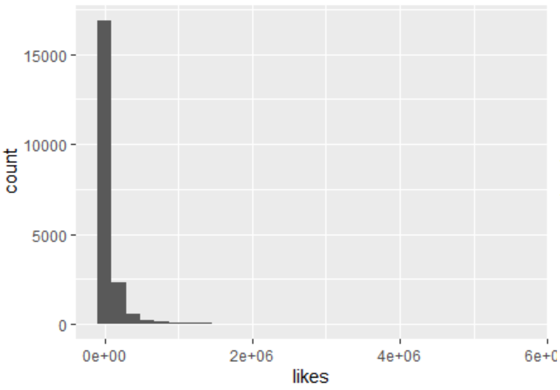


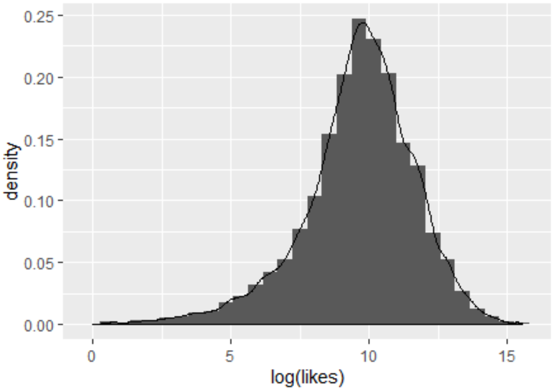
Compared to the set number of views found in our data, the percentage change in number of views is another meaningful variable to focus on. The distribution of the density function, log(views), is shown below, and gives a more readable trend.



Likes

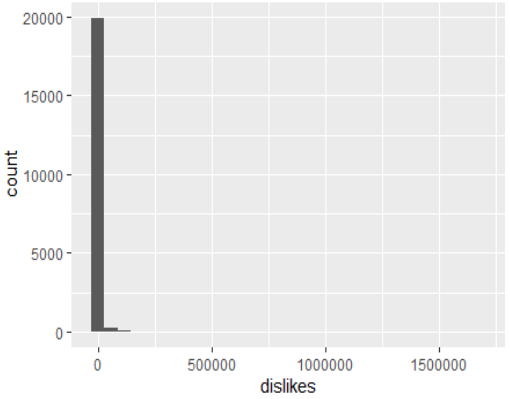
Because of the aforementioned reason, we decided to use log(likes) as our target variable, which indicates the percentage change in the number of likes.

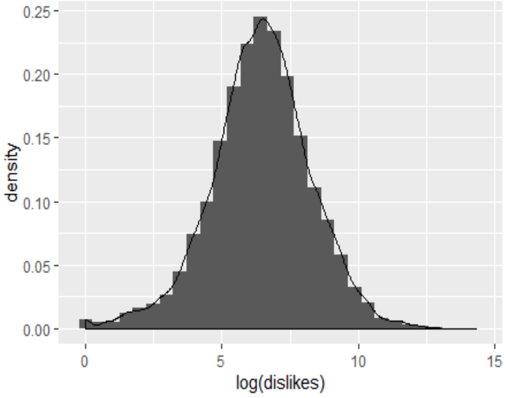


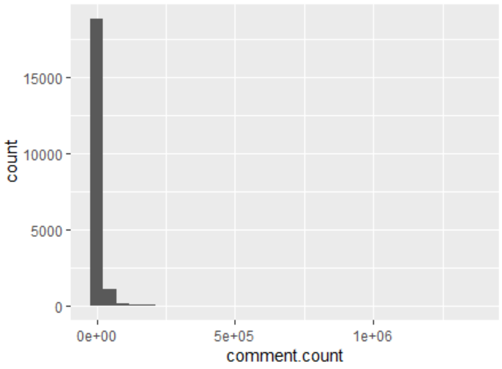


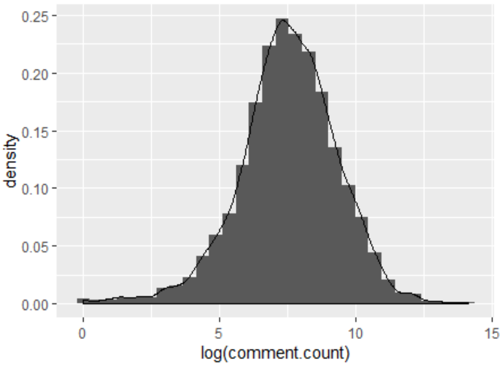
Dislikes and Comment Count

We can see the distribution of dislikes and comment count from the histograms below.



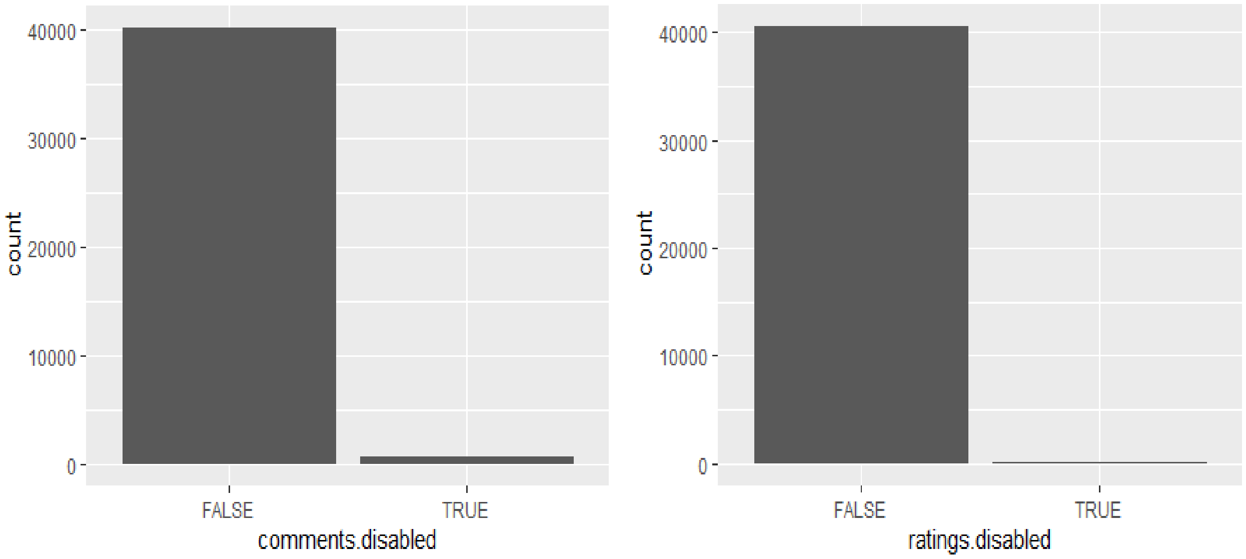






Although these two variables, especially comment count, may have strong correlation with likes, we will not use them as predictors. The reasons for this are as follows. First, we only want to find effective factors which lead to high likes, yet the causality between comment count and number of likes is very complex. These two variables influence one another after a video is posted. Second, the content creators cannot control the number of comment counts once they upload the videos. Even though it is known that high comment counts often relate to high likes, we cannot exploit this factor to get more likes.

Comments Disabled and Ratings Disabled

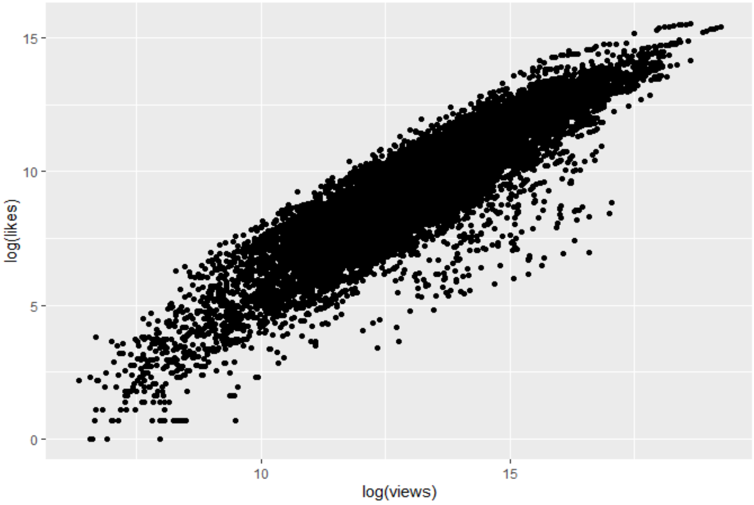


Most of the trending videos can be commented on and rated by users, which means these two variables are not distinguishable for most of the records. Therefore, these two variables are not suitable predictors.

Exploratory Analysis

Views

The following scatter plot indicates a strong positive correlation between log(views) and log(likes), which means that any percentage increase in views will lead to an increase in likes.

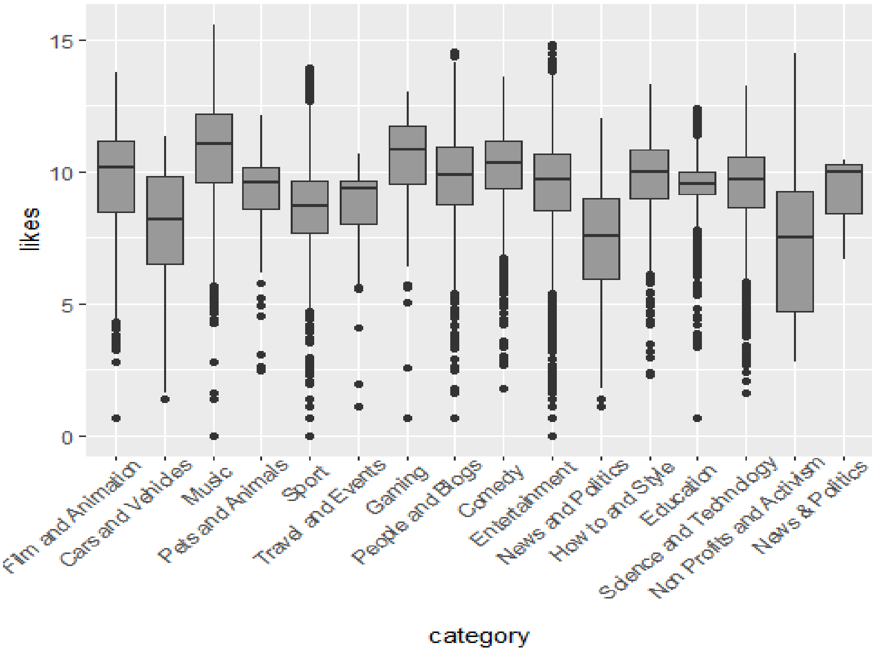


Publish Time

Compared to trending date, publish time is more easily controlled by content creators themselves. We then use boxplots to see the distribution of log(likes) grouping by publish day or hour. As shown in Appendix D, the distribution of log(likes) is relatively the same, publish day set aside. The same can be said for publish hour, also referenced in Appendix D. As stated before, we know that most videos are posted between the hours of 1pm and 10pm, so it seems that choosing these hours to publish would increase the competition among content creators of getting likes.

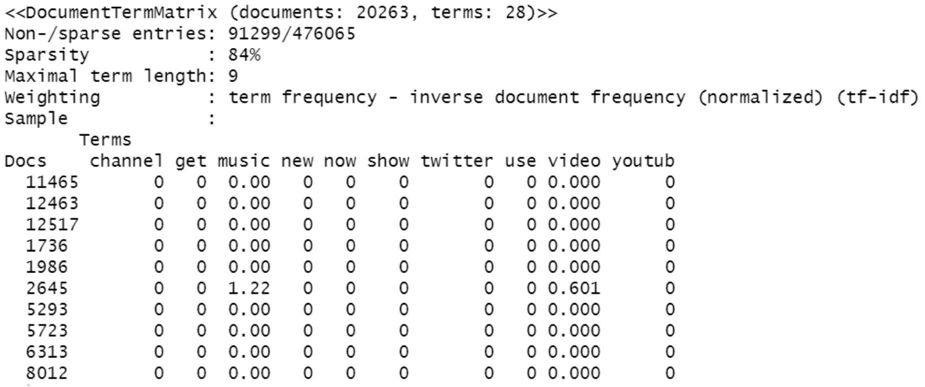
Category

As shown in the figure below, the distribution of log(likes) varies significantly according to different category. Videos in the Music and Gaming category usually get more likes, while those in News & Politics, Cars & Vehicles, and Non-profits & Activism usually receive fewer likes. We know from previous calculations that the Entertainment category has the most trending video, however, the mean of log(likes) is not significantly higher than others. Videos in the category of Education has the most concentrated distribution of likes.



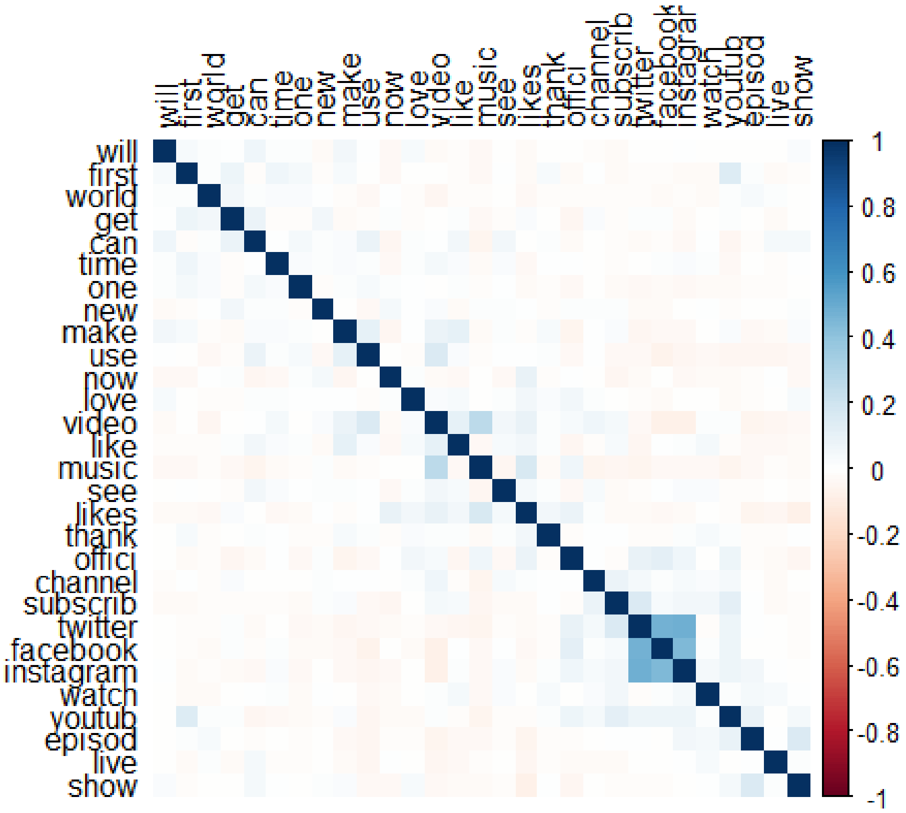
Description

As part of this project, we’ve implemented text mining techniques in order to explore the characteristics of description. To reduce term dimension, we have removed the less frequent terms such that the sparsity is lower than 0.9. Out of this, we get 28 term.



```
<<DocumentTermMatrix (documents: 20263, terms: 28)>>
Non-/sparse entries: 91299/476065
Sparsity           : 84%
Maximal term length: 9
Weighting          : term frequency - inverse document frequency (normalized) (tf-idf)
Sample             :
       Terms
Docs    channel get music new now show twitter use video youtub
  11465       0   0  0.00   0   0    0       0   0 0.000      0
  12463       0   0  0.00   0   0    0       0   0 0.000      0
  12517       0   0  0.00   0   0    0       0   0 0.000      0
  1736        0   0  0.00   0   0    0       0   0 0.000      0
  1986        0   0  0.00   0   0    0       0   0 0.000      0
  2645        0   0  1.22   0   0    0       0   0 0.601      0
  5293        0   0  0.00   0   0    0       0   0 0.000      0
  5723        0   0  0.00   0   0    0       0   0 0.000      0
  6313        0   0  0.00   0   0    0       0   0 0.000      0
  8012        0   0  0.00   0   0    0       0   0 0.000      0
```

The heatmap below shows the correlation of terms and log(likes). We can see that the words ‘twitter’, ‘facebook’ and ‘instagram’ are significantly correlate. As for log(likes), we do not conclude strong correlations with it.



Data Modeling

Multiple Linear Regression

Based on our data exploration, we have chosen log(views), category, publish day and the 28 terms from TFIDF matrix as predictors, with log(likes) as our target variable. After partitioning the data into 50% training data and 50% validation data, we fitted a MLR model. Then, after running stepwise regression, we received the following results (found in Appendix E). From this, we can see that most coefficients are statistically significant.

Interpretation

As for the description, the use of ‘get’, ‘like’, ‘love’, ‘see’, ‘thank’, ‘video’, ‘channel’ has a significant *positive* influence on number of likes to the degree of 0.1%. However, the use of ‘facebook’, ‘youtub’, ‘offici’, ‘show’, ‘world’ has a significant *negative* influence on number of likes to the degree of 0.1%. The coefficient of log(views) is 1.00 which is statistically significant on the level of 0.1%. This means a 10% increase in the number of views is predicted to lead to 10% increase in the number of likes.

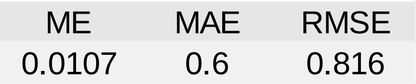
As for category, the base case is ‘Film & Animation.’ When compared to it, the videos in the categories of Cars & Vehicles, Sports, and News & Politics are predicted to receive less likes, while videos in other categories are predicted to receive more likes. Travel & Events is the only category not significant, even to the degree of 10%.

As for publish day, we set the base case as Friday. When compared, videos published on Tuesday are predicted to receive 11% more likes than on Friday. We found other coefficients to not be statistically significant on the level of 1%.

The MLR model has an R-squared of 83.9% and adjusted R-squared of 83.9%, which may indicate that this model has a good explanatory ability of the percentage change in number of likes.

Evaluation of Validation Data

After making predictions based on our validation dataset, we calculated accuracy measures of our model to be:



ME	MAE	RMSE
0.0107	0.6	0.816

Considering these results, we believe this model lacks proper accuracy, and that other extraneous factors are influencing the change in likes.

Result and Insights

To conclude this report, we find that the number of views, the use of certain words in video descriptions, varying categories and publish days are related to greater number of likes. With this information, content creators as well as influencers can and should focus more of their efforts on the following aspects in order to get more likes:

- ✧ **Video promotion:** Try to attract a greater number of views because it relates to a greater number of likes.
- ✧ **Description:** Use certain words such as ‘like’, ‘love’, ‘see’, ‘thank’ to get more likes. Avoid words which may lead to lower likes such as ‘offici’, ‘show’, ‘world’.
- ✧ **Category:** Having not decided which category to post in, choose categories with higher likes. For instance, Film & Animation is better than Car & Vehicles.
- ✧ **Publish day:** Publish videos on the days that generate higher numbers of likes. Tuesday is better than Friday.

Improvements

As mentioned in the evaluation section, our model is not accurate enough to make objective statements, though relationships among data points lead us to believe that soft relationships do in fact exist. And yet there are still many other important factors that significantly influence the generation of likes. For example, we are missing notable variables that represent re-watching of videos, such as playback or repeat.

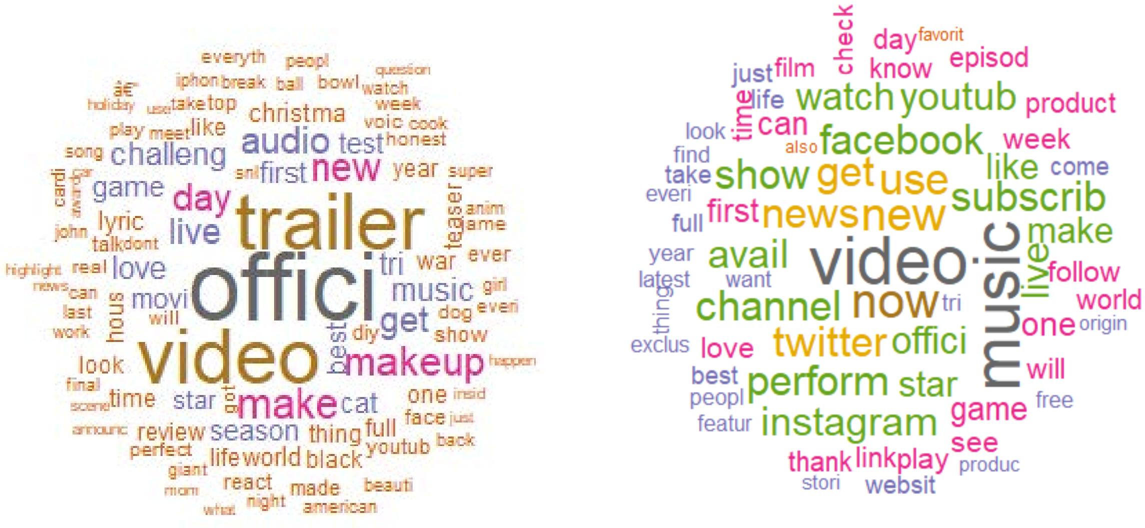
What’s more, we do not have data of non-trending videos. If we had that data, we might be able to tell if there’s something different the content creators should do to generate more likes for non-trending videos. Also, discerning trending videos from non-trending videos would be useful.

Obstacles

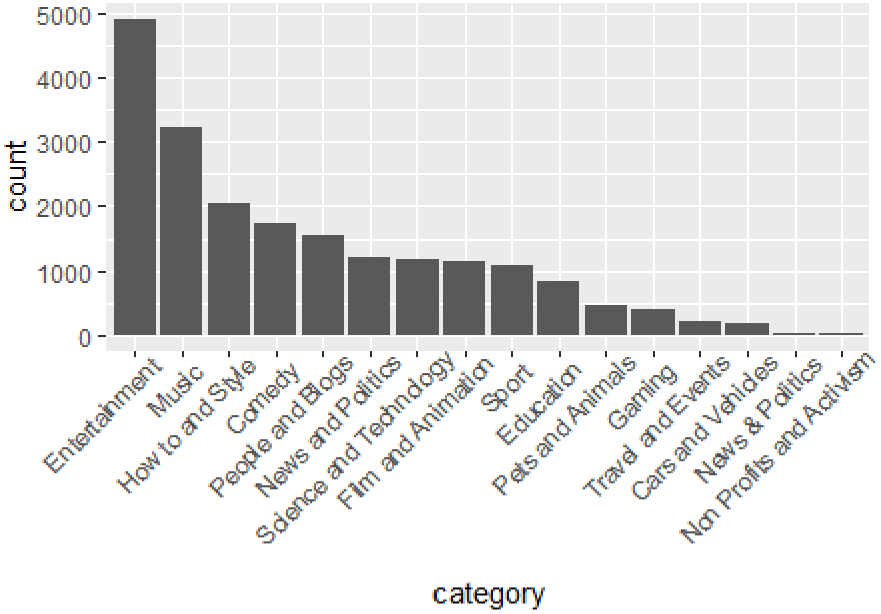
Throughout this project, we found it difficult to focus on a meaningful objective since this dataset contains only information regarding trending videos. From the very beginning, we wanted to find out what makes a trending video different from a non-trending video. It wasn’t until later that we realized that we only had information on trending videos, and not on both types—trending and non-trending. We then tried to predict the number of views but found that there was a causal inversion when using likes to predict views. At last, we decided to predict the percent change in number of likes. We also had trouble when running regression models. Our computers tended to crash due to the very large nature of the dataset. We ended up having to wait for the program to process the data since the storage/memory proves insufficient at times.

Appendices

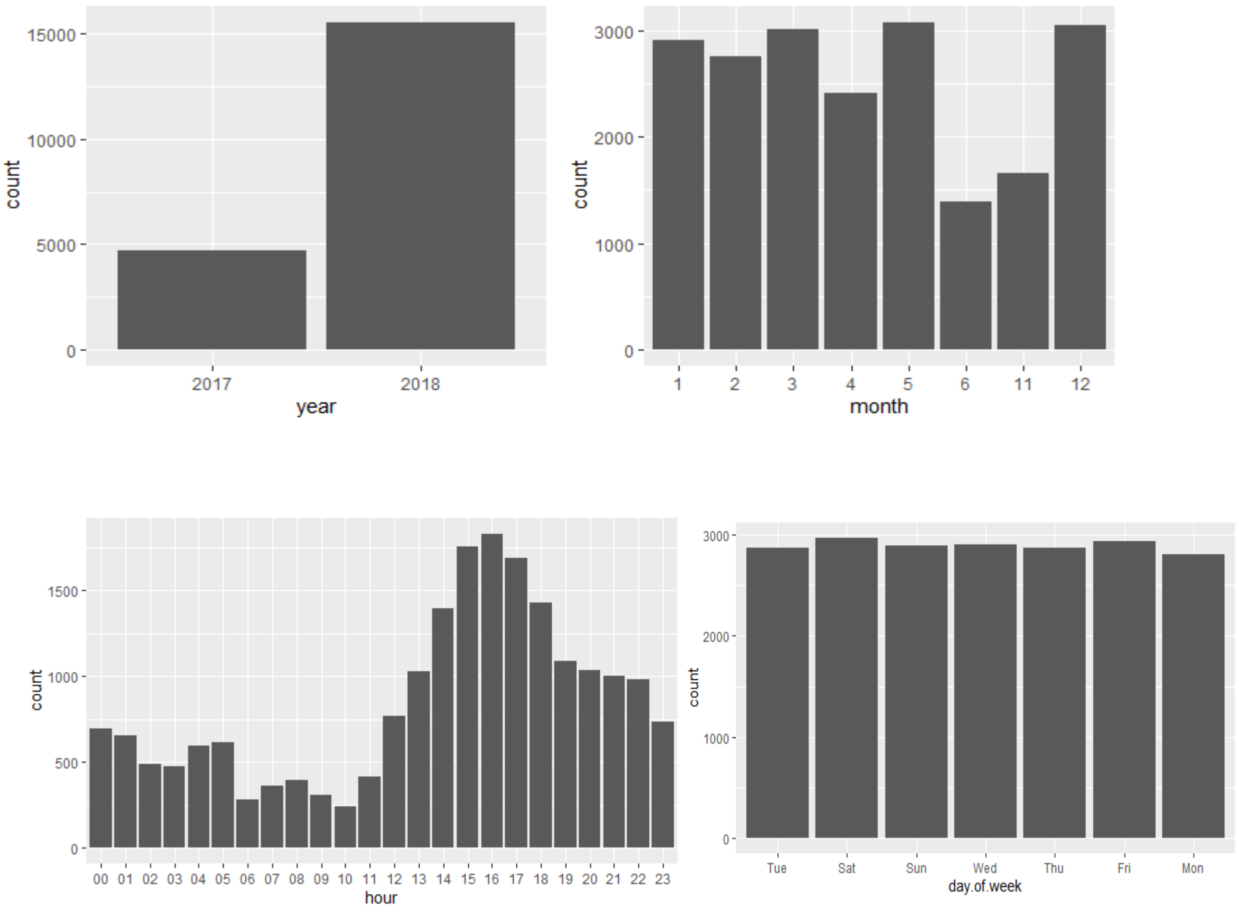
Appendix A



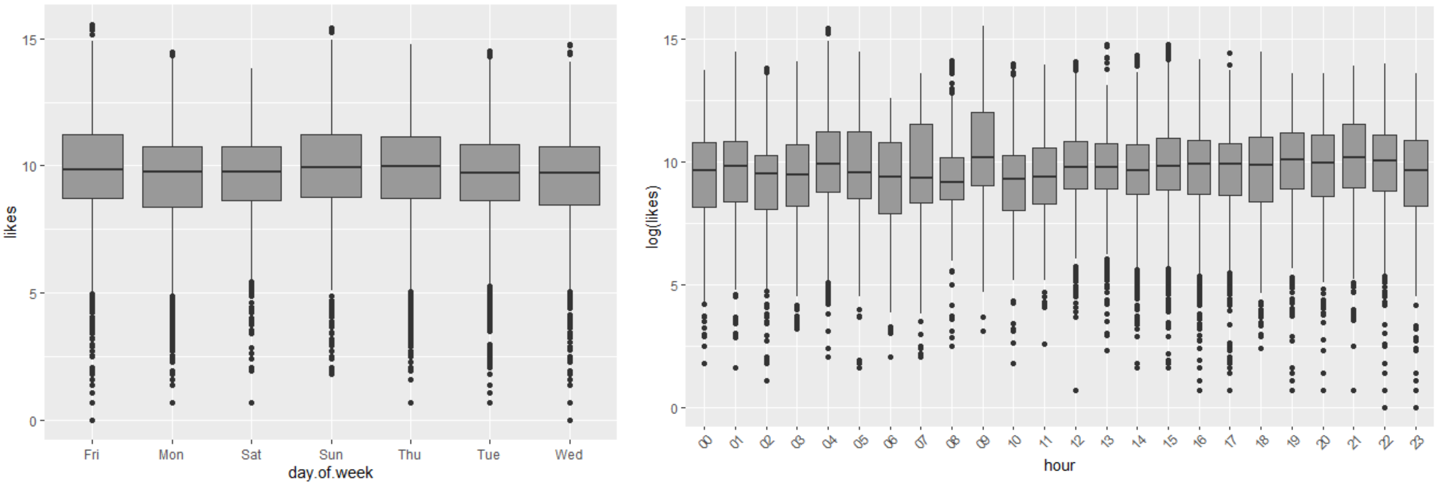
Appendix B



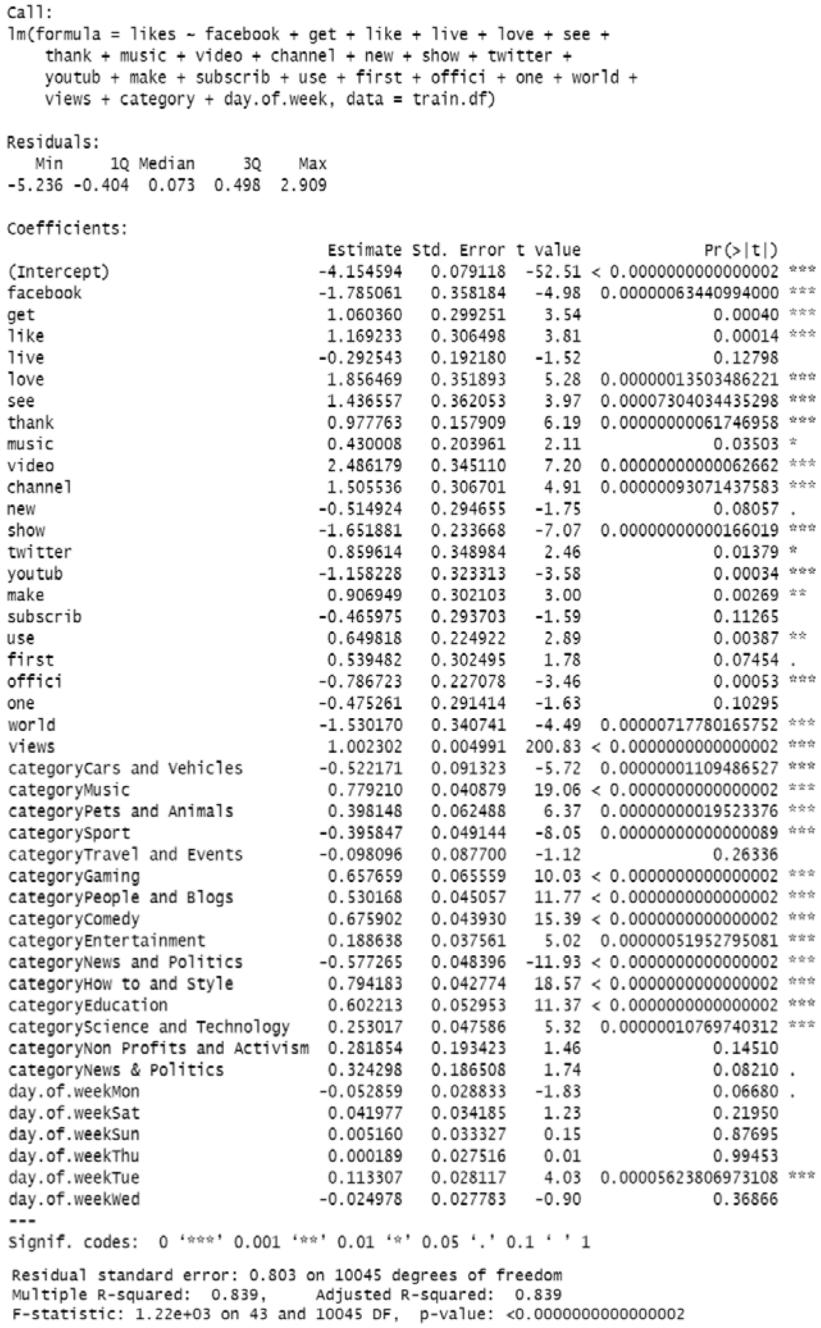
Appendix C



Appendix D



Appendix E



```
Call:
lm(formula = likes ~ facebook + get + like + live + love + see + 
    thank + music + video + channel + new + show + twitter + 
    youtub + make + subscrib + use + first + offici + one + world + 
    views + category + day.of.week, data = train.df)

Residuals:
   Min     1Q Median     3Q    Max 
-5.236 -0.404  0.073  0.498  2.909 

Coefficients:
                                  Estimate Std. Error t value             Pr(>|t|)    
(Intercept)                      -4.154594   0.079118  -52.51 < 0.0000000000000002 ***
facebook                         -1.785061   0.358184   -4.98  0.00000063440994000 ***
get                               1.060360   0.299251    3.54              0.00040 ***
like                              1.169233   0.306498    3.81              0.00014 ***
live                             -0.292543   0.192180   -1.52              0.12798    
love                              1.856469   0.351893    5.28  0.00000013503486221 ***
see                               1.436557   0.362053    3.97  0.00007304034435298 ***
thank                             0.977763   0.157909    6.19  0.00000000061746958 ***
music                             0.430008   0.203961    2.11              0.03503 *  
video                             2.486179   0.345110    7.20  0.00000000000062662 ***
channel                           1.505536   0.306701    4.91  0.00000093071437583 ***
new                              -0.514924   0.294655   -1.75              0.08057 .  
show                             -1.651881   0.233668   -7.07  0.00000000000166019 ***
twitter                           0.859614   0.348984    2.46              0.01379 *  
youtub                           -1.158228   0.323313   -3.58              0.00034 ***
make                              0.906949   0.302103    3.00              0.00269 ** 
subscrib                         -0.465975   0.293703   -1.59              0.11265    
use                               0.649818   0.224922    2.89              0.00387 ** 
first                             0.539482   0.302495    1.78              0.07454 .  
offici                           -0.786723   0.227078   -3.46              0.00053 ***
one                              -0.475261   0.291414   -1.63              0.10295    
world                            -1.530170   0.340741   -4.49  0.00000717780165752 ***
views                             1.002302   0.004991  200.83 < 0.0000000000000002 ***
categoryCars and Vehicles        -0.522171   0.091323   -5.72  0.00000001109486527 ***
categoryMusic                     0.779210   0.040879   19.06 < 0.0000000000000002 ***
categoryPets and Animals          0.398148   0.062488    6.37  0.00000000019523376 ***
categorySport                    -0.395847   0.049144   -8.05  0.00000000000000089 ***
categoryTravel and Events        -0.098096   0.087700   -1.12              0.26336    
categoryGaming                    0.657659   0.065559   10.03 < 0.0000000000000002 ***
categoryPeople and Blogs          0.530168   0.045057   11.77 < 0.0000000000000002 ***
categoryComedy                    0.675902   0.043930   15.39 < 0.0000000000000002 ***
categoryEntertainment             0.188638   0.037561    5.02  0.00000051952795081 ***
categoryNews and Politics        -0.577265   0.048396  -11.93 < 0.0000000000000002 ***
categoryHow to and Style          0.794183   0.042774   18.57 < 0.0000000000000002 ***
categoryEducation                 0.602213   0.052953   11.37 < 0.0000000000000002 ***
categoryScience and Technology    0.253017   0.047586    5.32  0.00000010769740312 ***
categoryNon Profits and Activism  0.281854   0.193423    1.46              0.14510    
categoryNews & Politics           0.324298   0.186508    1.74              0.08210 .  
day.of.weekMon                   -0.052859   0.028833   -1.83              0.06680 .  
day.of.weekSat                    0.041977   0.034185    1.23              0.21950    
day.of.weekSun                    0.005160   0.033327    0.15              0.87695    
day.of.weekThu                    0.000189   0.027516    0.01              0.99453    
day.of.weekTue                    0.113307   0.028117    4.03  0.00005623806973108 ***
day.of.weekWed                   -0.024978   0.027783   -0.90              0.36866    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.803 on 10045 degrees of freedom
Multiple R-squared:  0.839,	Adjusted R-squared:  0.839 
F-statistic: 1.22e+03 on 43 and 10045 DF,  p-value: <0.0000000000000002
```