The Views of YouTube Trending Video Analysis Report

Yifeng Luo

Abstract

Most of us have been ever watched videos in YouTube. When people do not know what video they want to watch, they could look through the trending tab to watch the current hot videos and know what is happening in the rest of the world. Some trends are predictable, like a new song from a current popular artist or a new movie trailer. Others are surprising, like a viral video. Trending aims to surface videos that a wide range of viewers will appreciate, so YouTube users always can find the videos they interested in from the trending list. However, do you know what are the factors has effect on a trending video's view? And how to estimate a trending video's view relied on its initial records and channel features? This report will select the rational predictors based on the exploratory data analysis (EDA) and use them to fit an appropriate linear regression model or multilevel model. In the end, this report provides a detailed diagnosis and implication of the model.

Introduction

YouTube (the world-famous video sharing website) maintains a list of the top trending videos on the platform. According to Variety magazine, "To determine the trending videos, YouTube considers a combination of factors including measuring videos' records (number of views, likes, comments and video category)". The YouTube trending system selects videos from massive videos based on a mature algorithm and specific criteria to predict a video will popular or not in the following days, then recommend them with users in trending list. The list of trending videos is updated roughly every 15 minutes. With each update, videos may move up, down, or stay in the same position in the list. If the time of a video stays in trending list is longer, generally speaking, its views is more likely to increase quickly during trending period. As a video creator, he/she tends to know what are the factors impact on the views of a trending video in order to improve the next video's popularity and views. This report will analyze what the factors are and fit a model to predict the views for new trending videos.

Some people conducted YouTube related researches in Kaggle, for example, some of them applied Knn, decision tree and random forest method to predict the category of a video by analyzing the features of data. Most of them are working for machine learning and deep leaning filed. Only a few people analyzed the views of recommendation video in YouTube, which means there are more probabilities for this topic.

Data Source

This first dataset this research used is a daily record of the top trending YouTube videos in United States from Nov-14-2017 to Jun-14-2018. It was downloaded from Kaggle. Some people scraped the data by YouTube's API and shared them in Kaggle. The dataset records the number of views, likes and comments of trending videos in YouTube. Meanwhile, it included other video information as well, like its title, brief description, categoryID, trending date, publish date and the tags attached.

There are many videos in trending list more than 1 day, but the data were collected daily, so it was multiple recorded. In order to analyze the factors that have effect on the views during the trending period, this research only keep the first day and last day record. And, setting the views in the last day record (called final views) for each video as the response variable and the amount of view (called initial views), like, dislike and comment count in the first day as explanatory variables. Meanwhile, extracting some useful variables from the original dataset, such as the number of day in trending list, the number of tags, the month of trending is an important step.

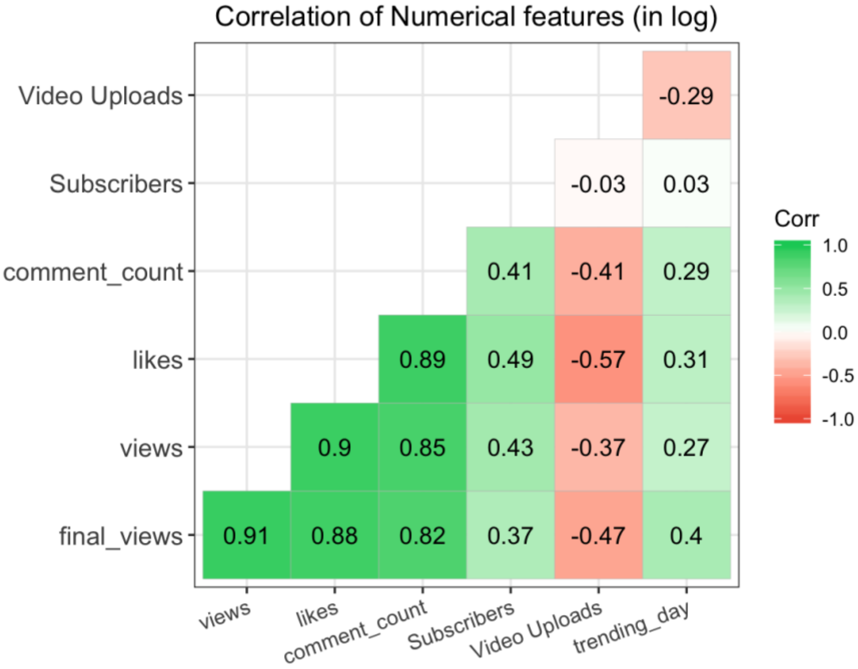
From the figure 1 in appendix, it is easily to notice that the average views differ among channels, it probably results from the features of channel. Socialblade is a well-known company which maintains statistics of YouTube channels, their website features a page which shows Top 5000 YouTube channels and some basic information about them, such as: the number of videos uploaded by the channel, total number of subscribers on the channel and the total number of views on all the video content by the channel. These metrics are useful for finding insights and the revealing possible correlations between the features of the channels and their video's view. This dataset records the channel information in the past half year. Although the record periods are not complete overlap with the first dataset (2017-11-14 to 2018-06-14), the information still reflect the differences among channels to some extent.

```
##      views               likes             dislikes      
##  Min.   :-0.458703   Min.   :-0.36580   Min.   :-0.15079  
##  1st Qu.:-0.374212   1st Qu.:-0.33707   1st Qu.:-0.13974  
##  Median :-0.244471   Median :-0.26561   Median :-0.11473  
##  Mean   : 0.000000   Mean   : 0.00000   Mean   : 0.00000  
##  3rd Qu.: 0.002663   3rd Qu.:-0.06559   3rd Qu.:-0.05289  
##  Max.   :16.618554   Max.   :20.60990   Max.   :39.00019  
##  comment_count       trending_day      final_views        Subscribers     
##  Min.   :-0.26153   Min.   :-1.1070   Min.   :-0.34051   Min.   :-1.1305  
##  1st Qu.:-0.24180   1st Qu.:-0.6902   1st Qu.:-0.29555   1st Qu.:-0.6862  
##  Median :-0.19791   Median :-0.2734   Median :-0.22877   Median :-0.2814  
##  Mean   : 0.00000   Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.0000  
##  3rd Qu.:-0.07134   3rd Qu.: 0.3518   3rd Qu.:-0.08042   3rd Qu.: 0.3785  
##  Max.   :21.44591   Max.   : 4.5198   Max.   :20.69943   Max.   : 8.2872  
```

After combining the two datasets as a big one, from the R output above, it is clear to see that the range is too large for one dataset as the 3rd quartile for most of the numerical features is less than the mean. And, there is a small portion of videos with extremely high number of views. The boxplot shows there exactly exist outliers in the dataset, so moving forward, the data will be divided into two subsets, one with final views less than its median and the other one greater than median and less or more than 3 σ videos with other indicators (the number of likes, dislikes and comment, etc.) will be removed in both subsets.

Correlation among Numerical Variables

The histogram shows the distribution of outcome *(“final_view”)* is approximate normal distribution after logarithmic transformation. Then, plotting graphs to check the relationship between log of final view and other potential predictors. If the predictors are in standardized scale, the relationship with log of final views is not linear. After transforming the predictors in logarithm scale, the figure 4 shows the linear correlation between outcome and predictors.

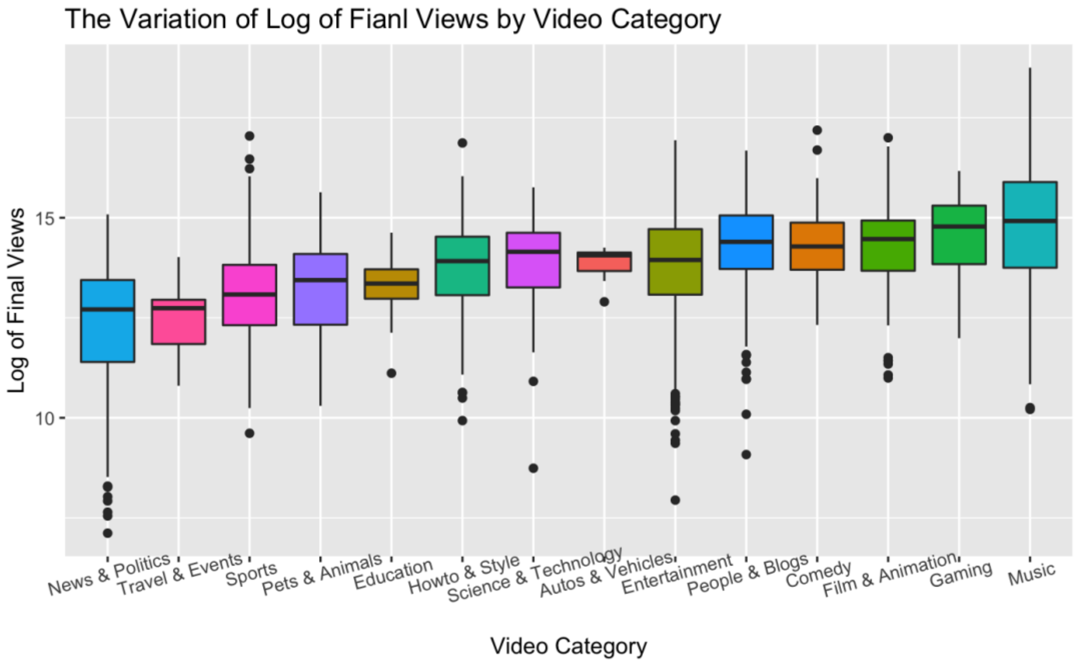


This heat map demonstrates many interesting things, overall, the log of views of trending videos in the last day *(“final_view”)* has highly positive correlation with the log of views, likes and comment count in the first day *(“views”, “likes”, “comment_count”)*, the number of trending day *(“trending_day”)* and the log of subscriber of channel *(“Subscribers”).* On the contrary, the log of the total number of video uploaded *(“Video Uploads”)* by channel has highly negative correlation with the final views, which means the video views increase slowly due to the channel uploaded video frequently. Meanwhile, it is easily to notice that the number of views, likes, dislikes and the comment count in log scale are highly correlated. It tends to result in the multicollinearity problem.

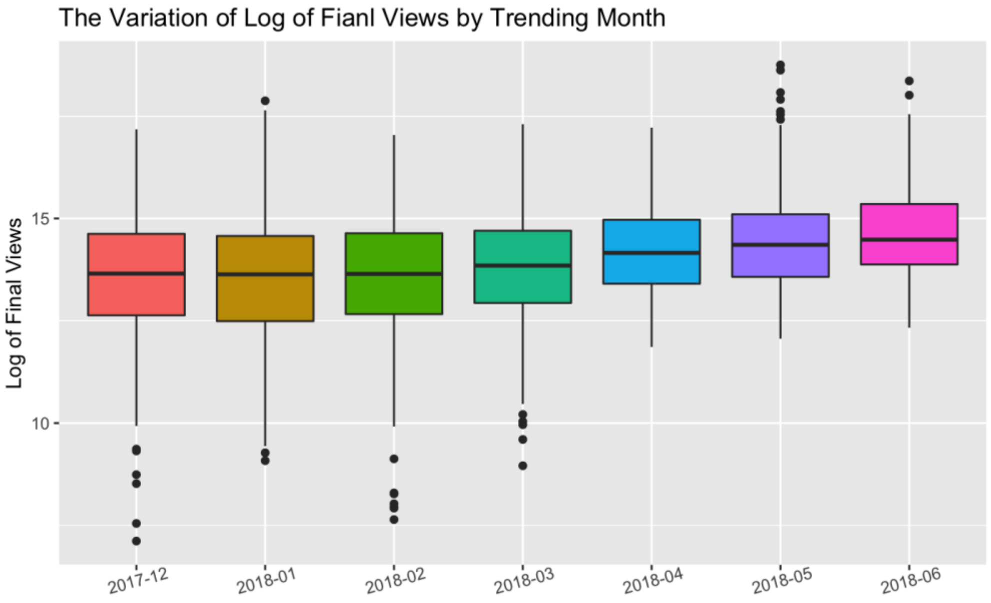
Categorical Factors Analysis

In the dataset, there are 4 categorical factors--video categories, the month of trending, the trending days and the tag count. In the following analysis, the research will analyze these variables one by one.

Video Category: The boxplot shows the log of final views vary in different video categories, the avarage of view in Music category was higher than other. Followed by Film & Animation and Gaming category. The category with lowest avarage views is News & Politics, which means only a few users would like to watch a news in YouTube, they prefer to have fun with entertainment videos, like music, film and comedy.

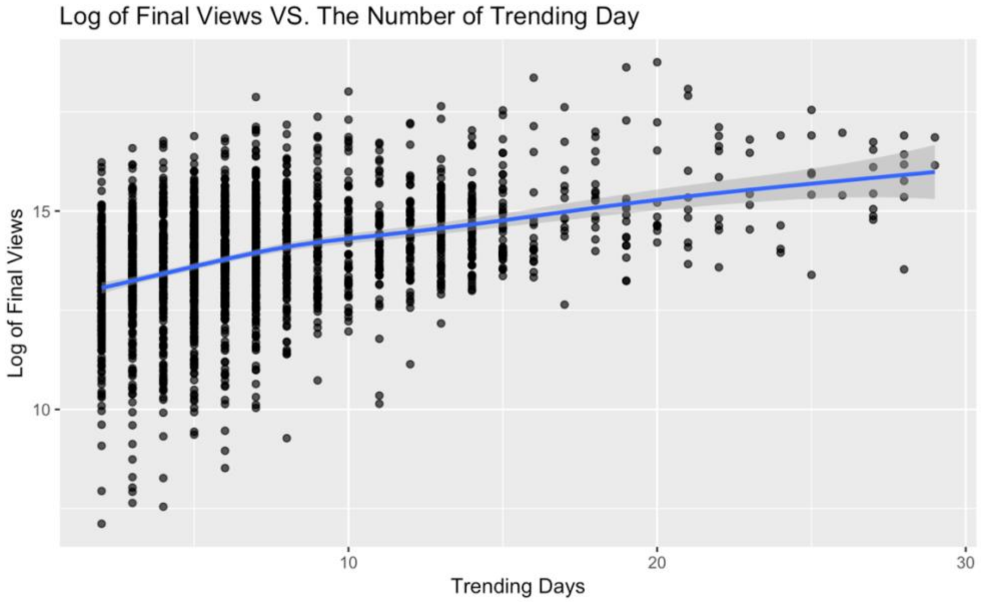


Trending Month: The boxplot shows the variation of final views by months in log scale, there is a slight upward trend from Dec 2017 to June 2018.



The Interaction of Trending Month and Video Category: The staked bar plot indicates the proportion of video category in trending tab did not vary too much by month, except the proportion of News & Politic category increased in Feb and March 2018, then decreased in the following two months. It probably happened some special events or political changes during these two months and appeal more people to know current news via YouTube. However, the average of views for news in these two months did not higher than April and May that had relatively less news video (figure 6).

Trending Days: From the dot plot, it is clear to see there is a positive relationship between final views and the number of trending day. If a video remains in trending tab for a long time, its views will rapidly increase during the trending period.



Tag Count: People commonly think the number of tag affect the views, the more of tags attached, the video will be searched by any key words in tags, which means YouTube users are more likely to watch this video. However, the boxplot shows there is no obvious pattern to proof the positive correlation between tag count and video views.

Model Used

There are too many potential predictors, so the first step is to reduce them and select relative important variables as model predictors. Relying on the EDA result and backward stepwise regression, the number of view, like, dislike and comment count in log scale, trending day, the log of subscriber and video uploaded by channels and video category can explain the outcome to some extent. Because there is linear correlation between outcome and predictors, this research tried to fit a log-log linear regression model and multilevel linear regression models.

Log-Log Linear Regression Model

The model 1 includes all of the predictors and the model 2 reduces two predictors-the number of video uploaded by channel and comment count in log scale.

```
model_1<-lm(log(final_views)~log(views)+log(likes)+log(dislikes+1)+log(commen
t_count)+log(Subscribers)+log(`Video Uploads`)+factor(trending_day)+factor(ca
tegory)+factor(month_c),trending_start)


model_2<-lm(log(final_views)~log(views)+log(likes)+log(dislikes+1)+log(Subscr
ibers)+factor(trending_day)+factor(category)+factor(month_c),trending_start)
```

Multilevel Regression Model

Because this is a hierarchical structure dataset, the category of video and the trending days can seem as the group level, and each specific video is in the individual level. In this part, due to the constraint of dataset, this research tried to fit two partial polling models without group level predictor. Model 3 and 4 are partial polling model by setting video categories and trending days as group level respectively.

```
#varying-intercept model with no predictor
#set category as group
model_3<-
lmer(log(final_views)~log(views)+log(likes)+log(dislikes+1)+log(Subscribers)+
factor(trending_day)+factor(month_c)+(1|category),trending_start)

#set trending days as group
model_4<-
lmer(log(final_views)~log(views)+log(likes)+log(dislikes+1)+log(Subscribers)+
factor(category)+factor(month_c)+(1|trending_day),trending_start)
```

Model Choice

Comparing model 1 and model 2 with ANOVA table and AIC indicator, both of them show there is no significant difference after removing the two predictors. Model 2 seems more appropriate with less predictors.

```
anova(model_1,model_2)
## Analysis of Variance Table
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1   1903 502.80
## 2   1905 503.41 -2  -0.61737 1.1683 0.3111

AIC(model_1,model_2)

##         df      AIC
## model_1 54 3001.719
## model_2 52 3000.119
```

On the other hand, the group variances are too small (both of them are less than 0.1) in model 3 and 4, which means there is no obvious discrepancy by video category and trending day. By comparing the AIC indictors of models, the model 2 with the least AIC is better than two partial polling models.

```
AIC(model_2,model_3,model_4)

##         df      AIC
## model_2 52 3000.119
## model_3 40 3157.044
## model_4 26 3147.575
```

In order to make the model comparison result more persuasive, the figure 8 in appendix shows there is no significant difference among these three models in Residuals vs Fitted plot. Therefore, this research selects model 2 as the best fitted model after considering all of the comparison method.

Interpretation

$$log(views_{end}) = \beta_0+\beta_1 log(views_{start})+\beta_2 log(likes)+\beta_3 log(dislikes+1)+ \beta_4 log(Subscribers)+factor(days_{trending})+factor(category)+factor(month)$$

```
coef(model_1)

##                          (Intercept)                            log(views)
##                          3.265181894                           0.681683031
##                           log(likes)                     log(dislikes + 1)
##                          0.133125027                           0.108083913
##                     log(Subscribers)                 factor(trending_day)3
##                         -0.073801392                           0.092685466
##                factor(trending_day)4                 factor(trending_day)5
##                          0.201235907                           0.142213514
##                factor(trending_day)6                 factor(trending_day)7
##                          0.347509599                           0.425164633
##                factor(trending_day)8                 factor(trending_day)9
##                          0.480191168                           0.488527494
##               factor(trending_day)10                factor(trending_day)11
##                          0.690767810                           0.478213824
##               factor(trending_day)12                factor(trending_day)13
##                          0.475229140                           0.658990731
##               factor(trending_day)14                factor(trending_day)15
##                          0.526881921                           0.658368425
##               factor(trending_day)16                factor(trending_day)17
##                          0.611897459                           0.581790990
##               factor(trending_day)18                factor(trending_day)19
##                          0.696626155                           0.740340468
##               factor(trending_day)20               factor(trending_day)21
##                          0.872232116                          1.046266587
##               factor(trending_day)22               factor(trending_day)23
##                          0.895202440                          0.620033170
##               factor(trending_day)24               factor(trending_day)25
##                          0.861162210                          0.479523545
##               factor(trending_day)26               factor(trending_day)27
##                          0.845444249                          0.797482588
##               factor(trending_day)28               factor(trending_day)29
##                          1.055274300                          1.883719163
##               factor(category)Comedy            factor(category)Education
##                          0.345857884                          0.236573610
##        factor(category)Entertainment     factor(category)Film & Animation
##                          0.505519483                          0.532908653
##               factor(category)Gaming        factor(category)Howto & Style
##                          0.224845540                          0.295168902
##                factor(category)Music      factor(category)News & Politics
##                          0.945503418                          0.286297309
##       factor(category)People & Blogs       factor(category)Pets & Animals
##                          0.391231630                          0.548722464
## factor(category)Science & Technology               factor(category)Sports
##                          0.344970441                          0.427533146
##      factor(category)Travel & Events               factor(month_c)2018-01
##                          0.655299481                          0.120023713
##               factor(month_c)2018-02               factor(month_c)2018-03
##                          0.098585787                          0.049406547
##               factor(month_c)2018-04               factor(month_c)2018-05
##                         -0.036316014                         -0.041469174
##               factor(month_c)2018-06
##                          0.003899566
```

This model can be explained by two kinds of variables-continuous variable and categorical variable. The log of views, likes, dislikes and subscribers are continuous variables; video category, trending day and month are categorical variables. The changes of continuous variable are affect model slope, but the changes of categorical variable only affect the model intercept, which is same with vary intercept model. Because there are a lot of estimate coefficients, it is not easy to interpret one by one. This part picks some typical coefficients and explain what is their meaning in the real world.

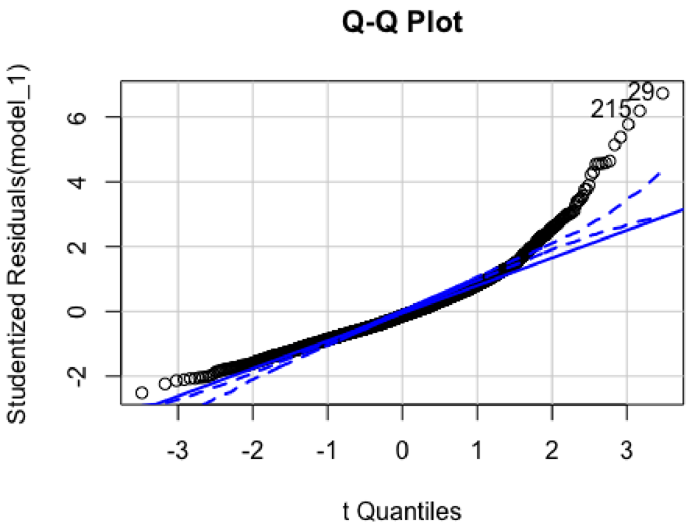
The estimate coefficient of log of views is 0.68, which means each 1% difference in initial views, the predicted difference in final views is 0.68%. Moreover, the coefficient of initial views is higher than the coefficient of likes and dislikes, so the changes of initial views has a larger positive effect on the views at the end of trending. it is not surprise that the coefficient of likes and dislikes are positive. For each 1% increment in like and dislike, the expected final view increases by 0.13% and 0.1% respectively. Because the number of like and dislike reflect a video's popularity, like and dislike just two kinds of attitude to the video. However, the coefficient of subscriber is negative, for each 1% increment in subscriber, the predicted final views decreases by 0.07%.

Overall, the coefficients of trending day increase with the day of trending goes up. So if other variables do not change, a video stay in trending list longer, its final views will increase largely. And the coefficient of music video is 0.94, which is highest among video categories.

Model Diagnosis

A linear regression model bases on some assumptions, such as the normality and independence of response variable and the normality and homoscedasticity of error. Meanwhile, this part will check whether the model is affected by collinearity, outliers and influential points. Finally, using cross validation method to check the prediction accuracy of model.

Check Normality



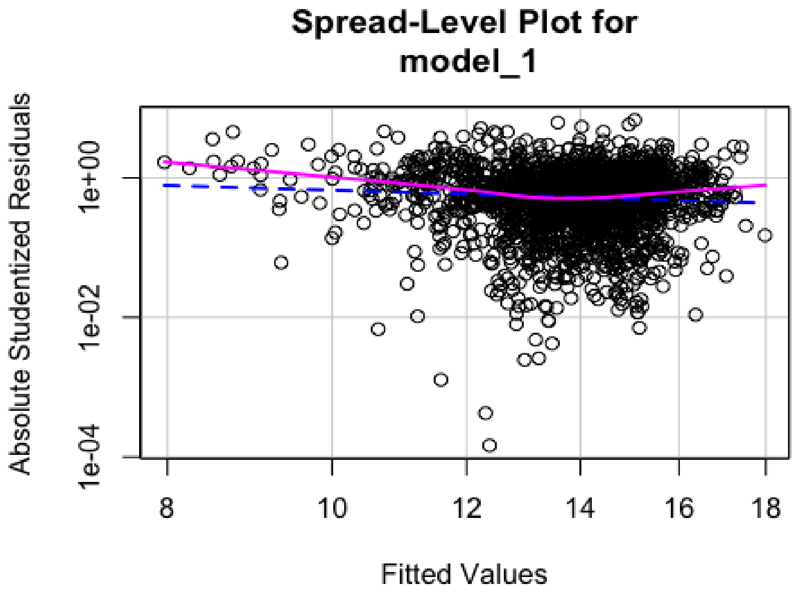
From the Q-Q plot in the left side, the dots in the plot present a curve shaped, so the normality of residual seems not valid in the 95% confidence interval. The dots from -2 to +2 t-quantiles are complete fit with the blue solid line. However, the tail and head are a little far away with the line. Figure 3 in appendix shows the distribution of outcome in bell-shaped. The assumption about outcome normality seems to be held.

The Independence of Error

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.1097742      1.780131       0
##  Alternative hypothesis: rho != 0
```

The Durbin Watson test result is not good, the p-value is significant and the independent hypothesis should be rejected. It may result from the competitive relationship among trending videos and in the same day, the views of a video are affected by the performance of the similar category videos. Therefore, the response variable is not independent by time.

Homoscedasticity of Error



The spread-level plot shows the points are evenly distributed on the both sides of the horizontal line, the same variance assumption seems to be held.

Check Multicollinearity

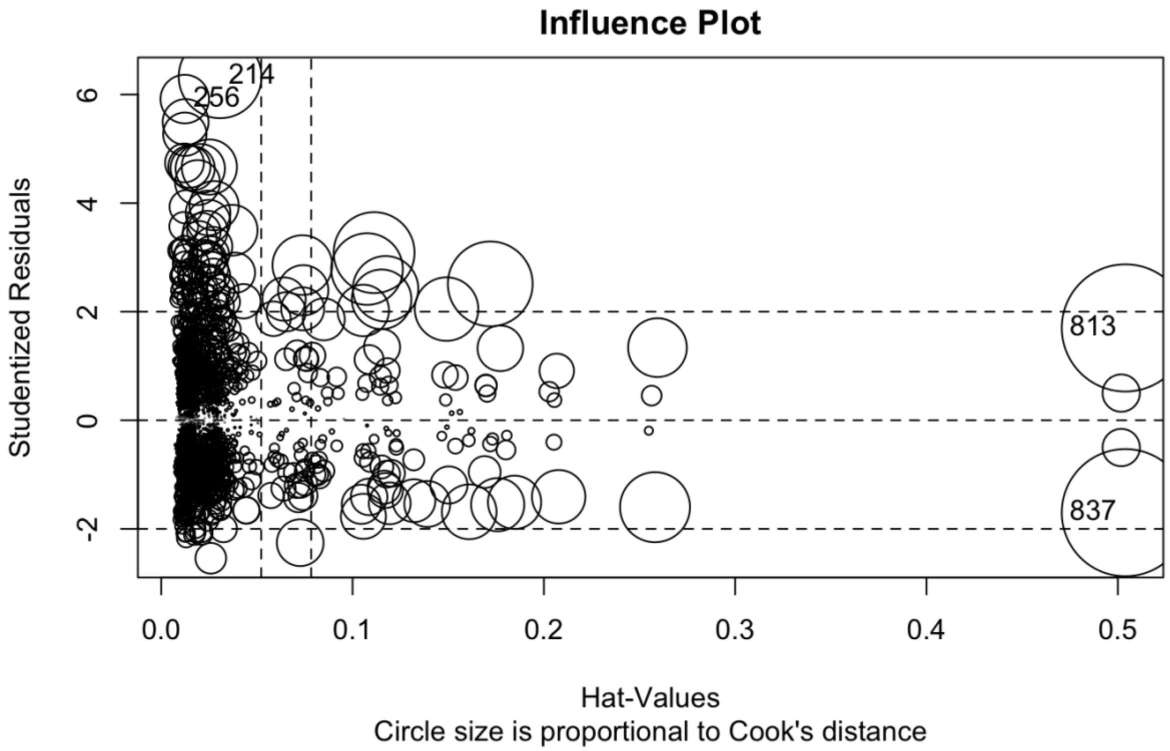
```
##                             GVIF Df GVIF^(1/(2*Df))
## log(views)              9.478187  1        3.078666
## log(likes)              9.278186  1        3.046011
## log(dislikes + 1)       4.172150  1        2.042584
## log(Subscribers)        1.524592  1        1.234744
## factor(trending_day) 2.530252 27        1.017340
## factor(category)     2.753852 13        1.039731
## factor(month_c)      2.138574  6        1.065394
```

```
##                          GVIF    Df GVIF^(1/(2*Df))
## log(views)               TRUE FALSE           FALSE
## log(likes)               TRUE FALSE           FALSE
## log(dislikes + 1)        TRUE FALSE           FALSE
## log(Subscribers)        FALSE FALSE           FALSE
## factor(trending_day) FALSE  TRUE           FALSE
## factor(category)     FALSE  TRUE           FALSE
## factor(month_c)      FALSE  TRUE           FALSE
```

The square of VIF (Variance Inflation Factor) are less than 2, which means there is no multicollinearity problem in the model.

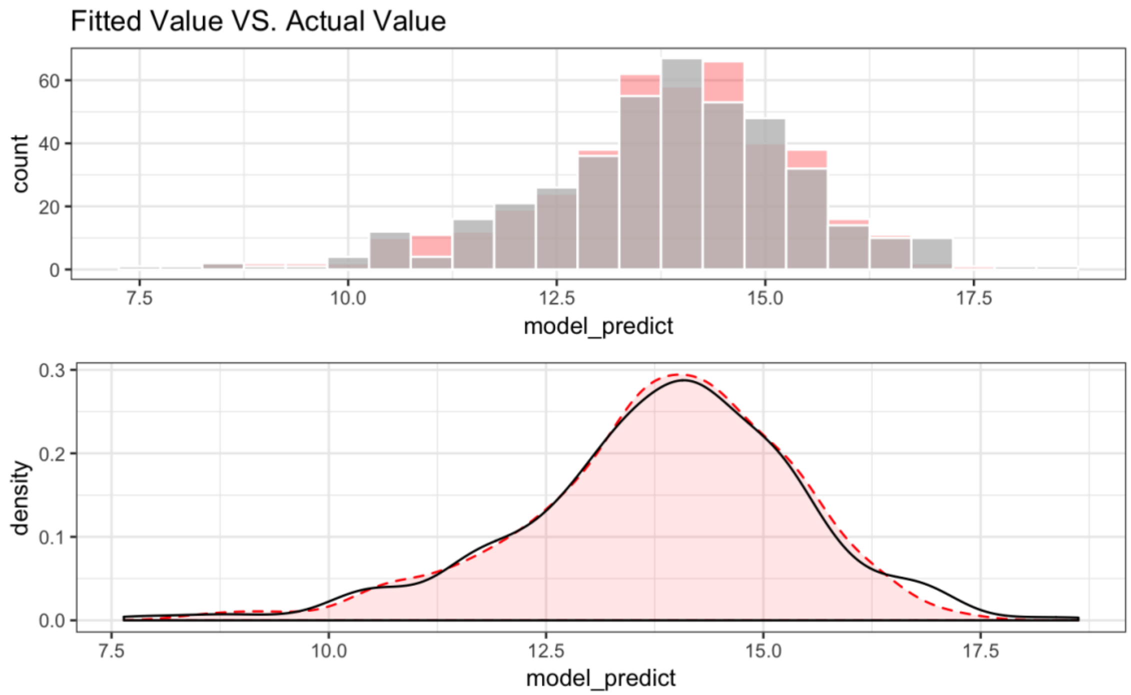
Outlier and Influential Observation

From the influence plot, it is clear to see there are some outliers (studentized residuals bigger than the range of -2 to +2) and strong influential observations (the radius of circles is bigger than others), for example, row 813 and 837 with a relative large circle are influential observations. These points have effect on the estimation of model coefficients.



Cross-Validation

Dividing the dataset in two subsets according to 20/80 rule (Pareto principle), the big one as training set, another one as test set. Fitting a model relied on training set, then use the model to predict the final views based on the variables in test set. The bar chart and density plot below show the fitted value versus actual value, it is clear to see their distribution are mostly overlap, which means there is no big difference between fitted views and actual views.



```
##  Welch Two Sample t-test
##
## data:  gg_model$model_predict and gg_model$V2
## t = -0.2653, df = 826.46, p-value = 0.7908
## alternative hypothesis: true difference in means is not equal to 0
```

Meanwhile, in order to test whether fitted value differs with actual value in test set, applying two sample t-test to check it. The p-value is 0.79, so it is not significant, which means there is no evidence to reject the null hypothesis, the model prediction looks accurate.

Implication

This research aims to find the factors have effect on views of trending videos. It is useful for video creators to improve video's views. The model coefficients explain many things, the amount of initial views, likes, dislikes, video category and how many days in trending list determine a trending video's views at the end of trending. That reason why bloggers always remind viewer to thumb up and subscribe at the end of each video. However, the fact is the number of dislike increases, the views increase as well, that the reason why some bloggers like to share controversial videos or say extreme remarks in the video to attractive audiences' attention. Moreover, there are inherent differences among the video category. The entertainment videos are really popular in U.S. viewers, especially, music and gaming video, that is reasonable due to people always watch videos to release the stress from work and study. Although the views of news and politics related videos is not high, generally that is not the major motivation of creator to share them.

One thing should be noticed is that the negative correlation between views and the number of video uploaded. If a blogger wants to improve their video views, except by improve the quality of video, he/she has better to control the frequency of video update. On the other hand, the model can be used to estimate the final views based on video initial records and channel characters.

Limitation

This research has some limitations, which is difficult to be solved in current dataset. Actually, the length of a video impacts the views mostly. The length of a movie or a documentary are usually more than one hour, but most viewers watch videos in their scattered time. Thus, the interaction effect of video category and video length should be considered in the model. Moreover, this model does not consider the time factor, so the assumptions of residual-normality and independence-are not met according to the diagnosis result. On the other hand, the amount of view can be manually manipulated by spending money to buy artificial followers. It is possible the views do not reflect the popularity of videos.

Future Direction

In the next step, this research will do sentiment analysis for video description to check the views variation by the sentiments (positive, natural and negative). Meanwhile, using LDA method in video tags to extract some topic keywords. It is a good way to split the category in more specific subsets, for example, a music video can be distinguished by Jazz, Pop and Rock, etc. On the other hand, I will try to fit some time series models to predict how long a video can stay in trending tab based on the video features.

Acknowledge

The data could not have been created without the hard work of the person who grasped the data from YouTube and Socialblade. They actually did a lot of work of collecting all the necessary metrics of the video records and channels. And thanks to the everyone who shared their great ideas and EDA process in Kaggle, which inspires this study to dig in deeper.

References

Donyoe, *"Youtube new Trending Statistics",* Kaggle, March 2018, https://www.kaggle.com/donyoe/exploring-youtube-trending-statistics-eda

Garrett Grolemund & Hadley Wickham, *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data,* O'REILLY, https://r4ds.had.co.nz

Joseph M. Juran, *"Pareto Principle",* Wikipedia, https://en.wikipedia.org/wiki/Pareto_principle

Mohit, *"US Youtube EDA & Category Predictions"*, Kaggle, June 2018, https://www.kaggle.com/bansmohit/us-youtube-eda-category-predictions?scriptVersionId=3814849

Mitchell J, *"Trending YouTube Video Statistics",* Kaggle, June 2018, https://www.kaggle.com/datasnaek/youtube-new/home

Robert I. Kabacoff, *R in Action: Data analysis and graphics with R*, Manning Publications Co., Shelter Island, NY 11964, http://kek.ksu.ru/eos/datamining/1379968983.pdf

Todd Spangler, *"YouTube Reveals 2017 Top Viral and Music Videos"*, Variety Magazine, December 6, 2017, https://variety.com/2017/digital/news/youtube-2017-top-trending-videos-music-videos-1202631416/

"TOP 5000 INFLUENTIAL YOUTUBE CHANNELS (SORTED BY SB RANK)", Socialblade, https://socialblade.com/youtube/top/5000

"The Pareto Principle applied to betting", PINNACLE, Oct 31, 2018, https://www.pinnacle.com/en/betting-articles/Betting-Strategy/The-Pareto-Principle-of-Prediction/FS52BE6XD4ZJMSBQ

"Trending on YouTube", YouTube Help, https://support.google.com/youtube/answer/7239739?hl=en

W. Holmes Finch, Jocelyn E. Bolin & Ken Kelley, *Multilevel Modeling Using R*, CRC Press, June 13, 2014, 230 Pages, https://www.crcpress.com/Multilevel-Modeling-Using-R/Finch-Bolin-Kelley/p/book/9781466515857

Appendix

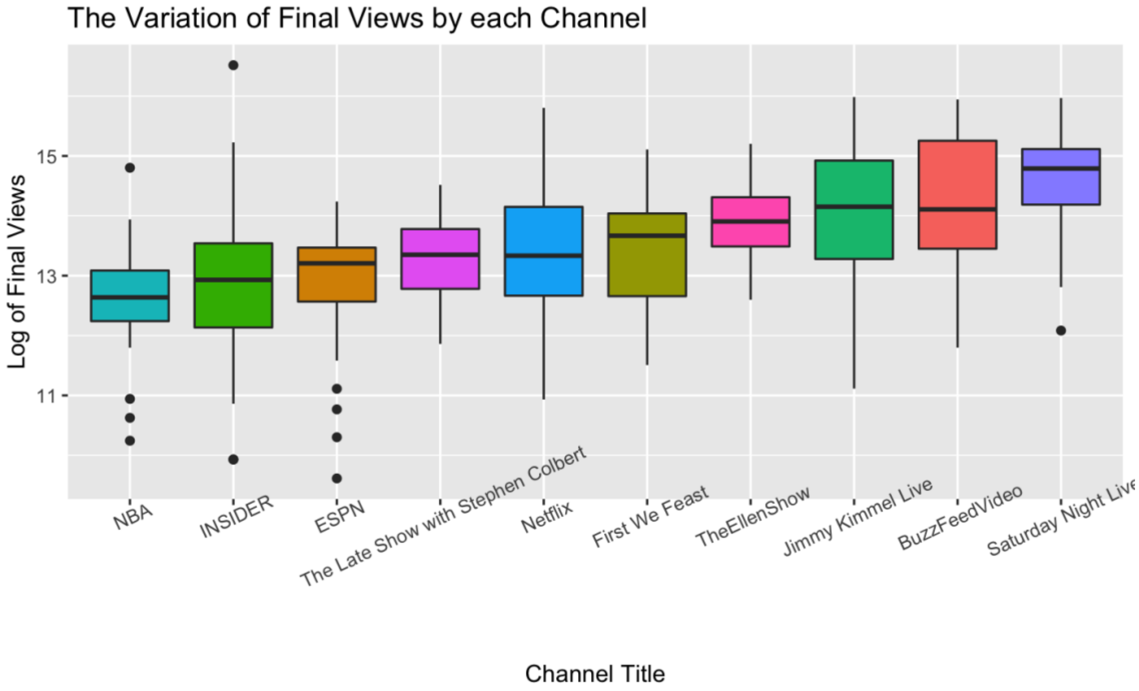


Figure1

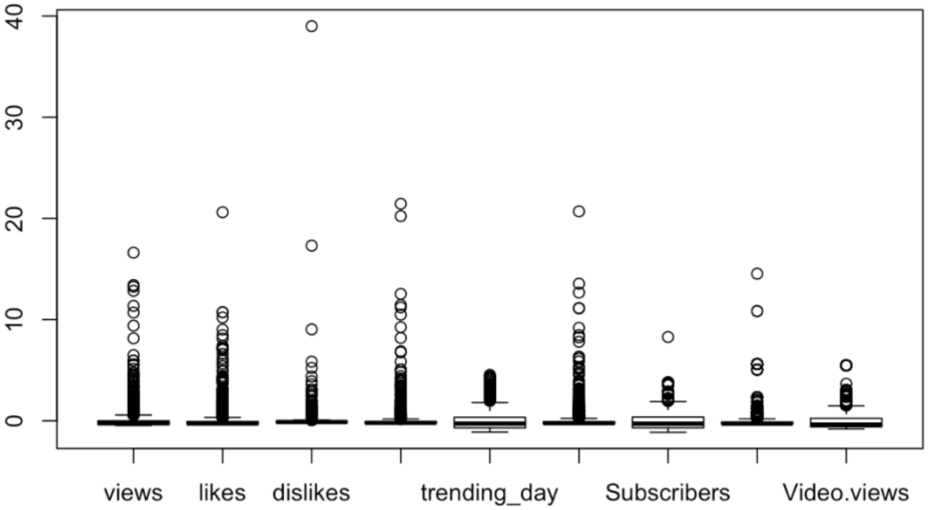


Figure 2

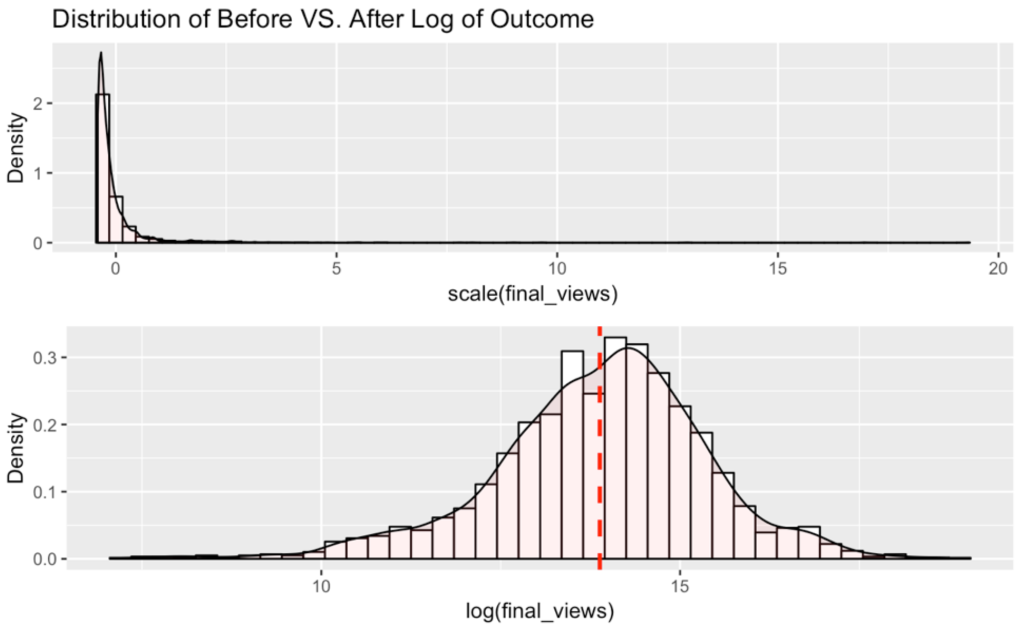


Figure 3

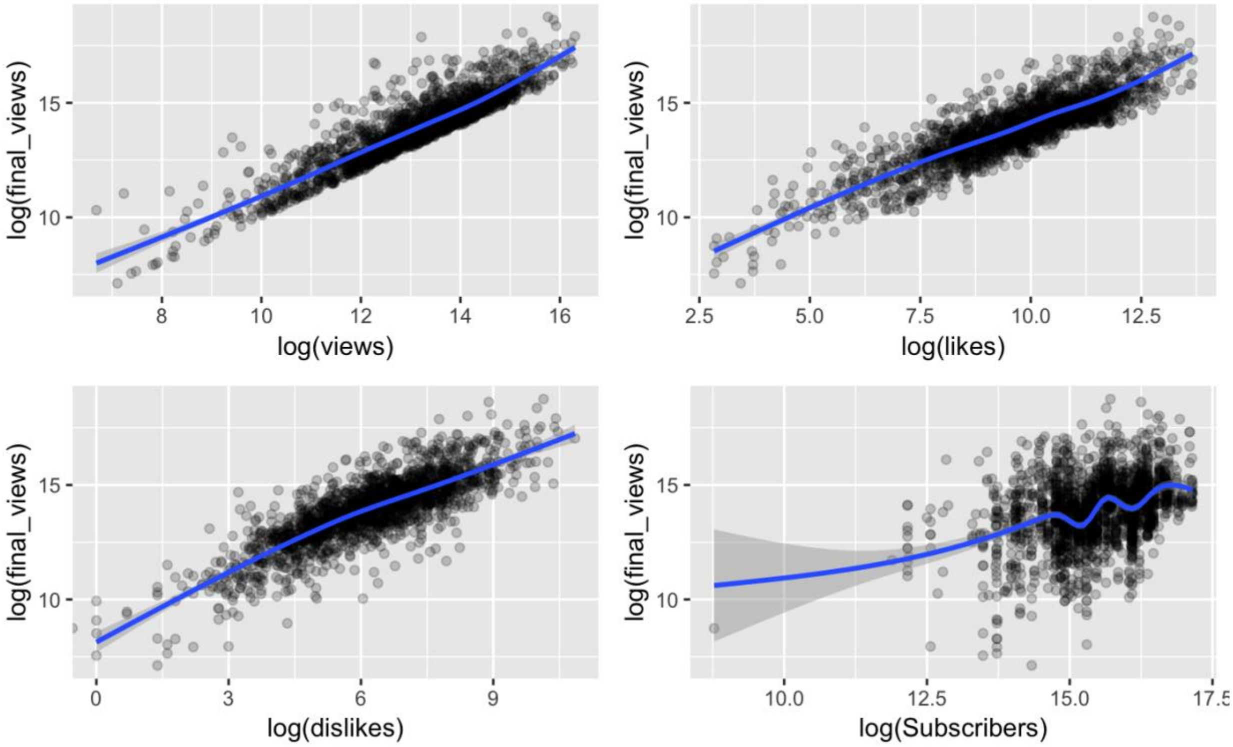


Figure 4

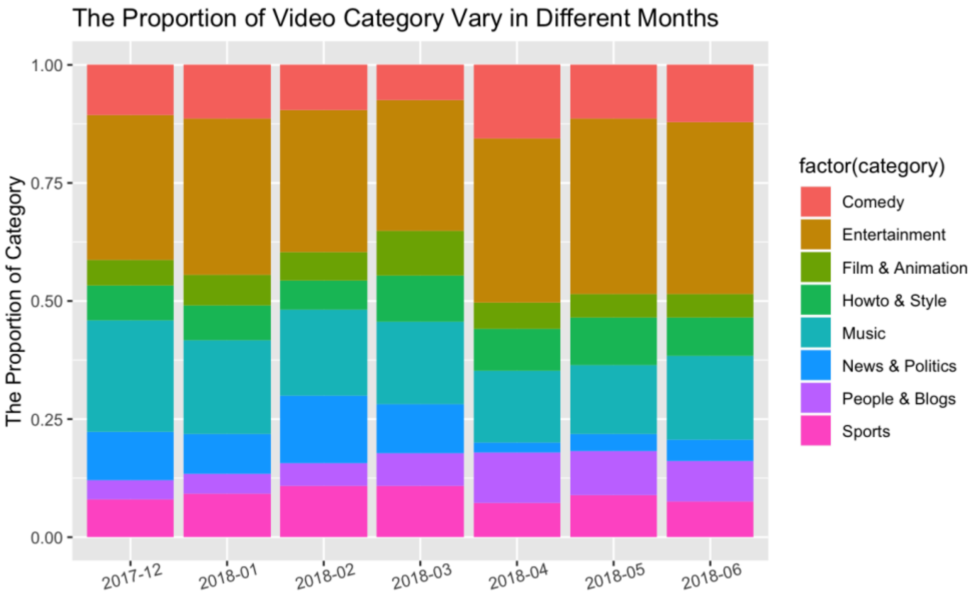


Figure 5

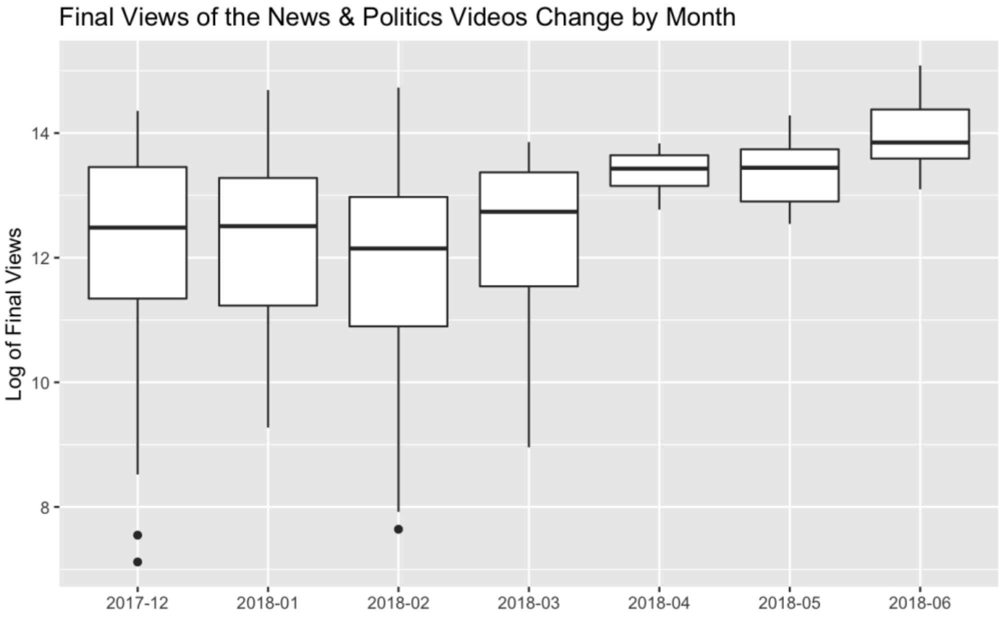


Figure 6

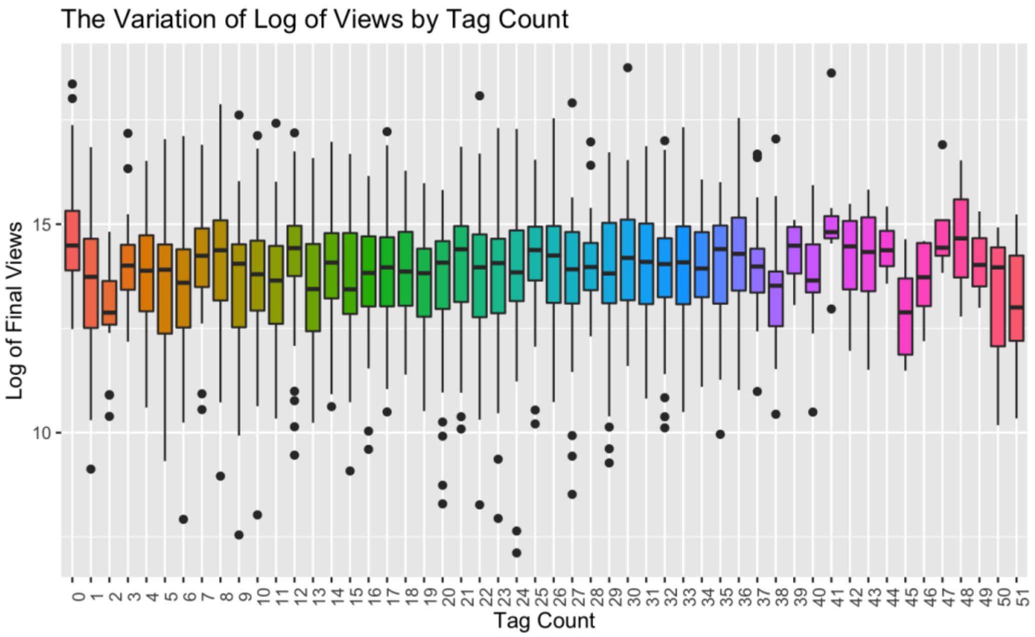
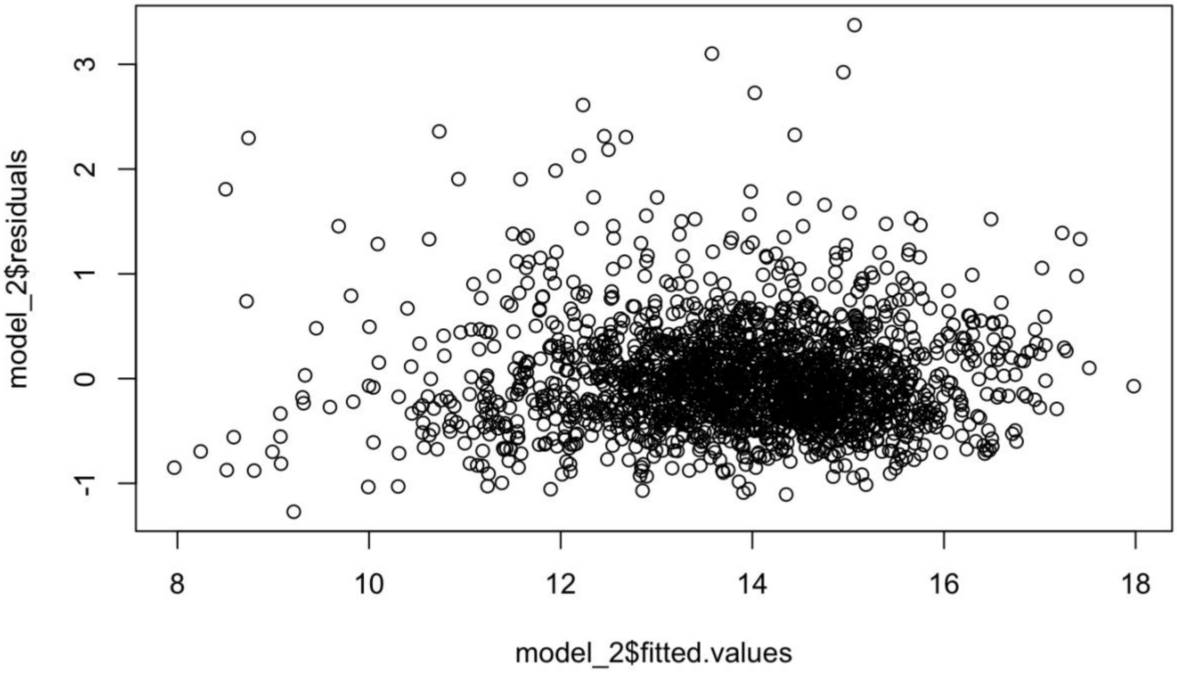

The Variation of Log of Views by Tag Count
Log of Final Views
15
10
Tag Count
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51

Figure 7


model_2$residuals
3
2
1
0
-1
model_2$fitted.values
8
10
12
14
16
18

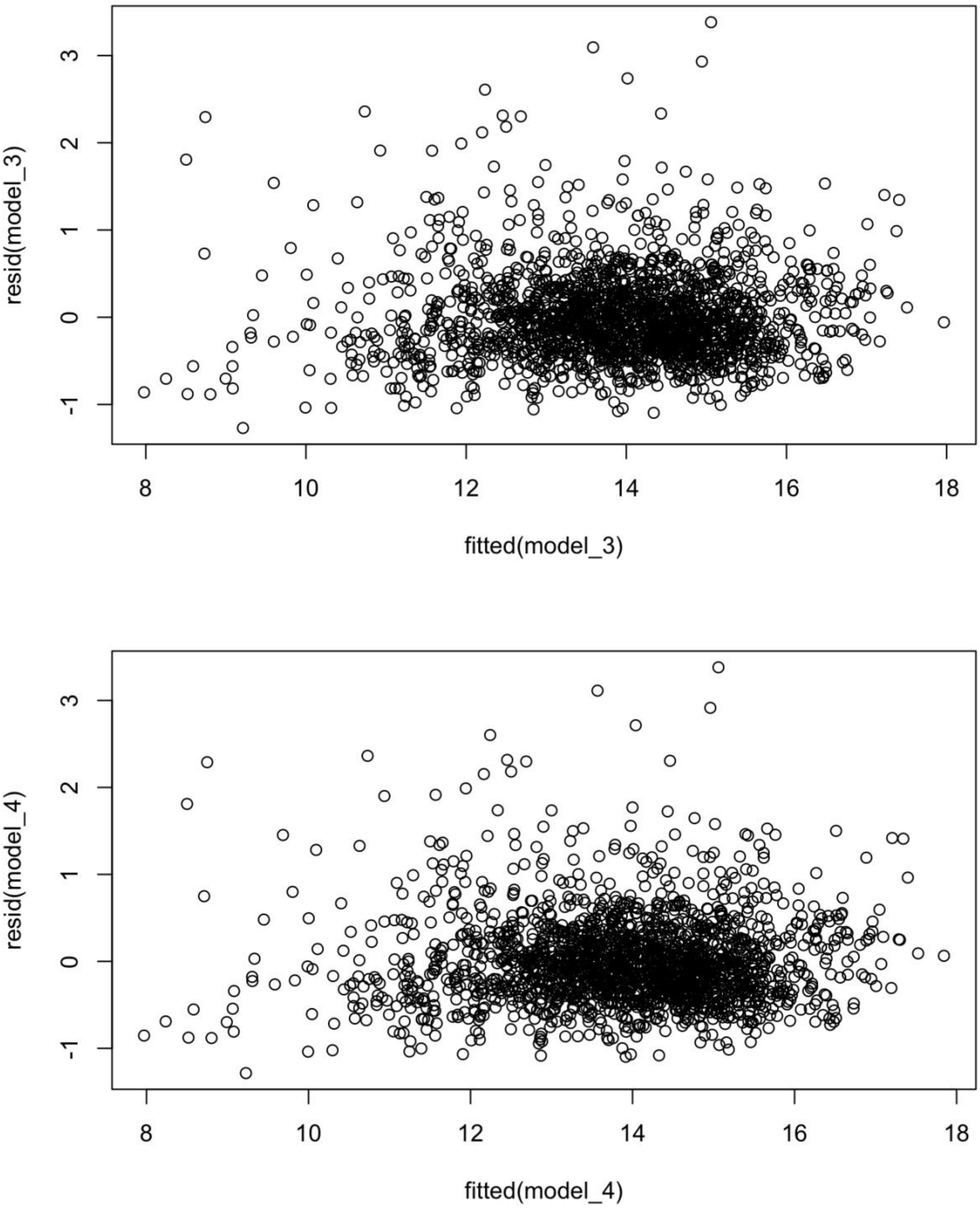

resid(model_3)
3
2
1
0
-1
8
10
12
14
16
18
fitted(model_3)
resid(model_4)
3
2
1
0
-1
8
10
12
14
16
18
fitted(model_4)

Figure 8