MA678 Midterm Project

Analysis of U.S. YouTube Trending Video
Chi Zhang

Abstract

As a world-famous video website, YouTube has a large number of audiences and high-quality video content from all over the world. YouTube's trending video recommendation aroused my curiosity. As we all know, many factors contribute a video to be a trending video. There are user interactions such as views, likes, dislikes, and the number of comments; Tags, categories and other factors which attract users to click the video also indirectly affects whether a video can be a trending video or not. In this project, my goal is to figure out what are the factors affect the view of trending video in different categories by using a multilevel model. The model result gives that the dislikes have most positive effect on views, and the comments have a great negative effect on views. The model has some flaws, and they are needed to fix by improving data.

Introduction

My career goal has always been to become a data scientist and working in an outstanding technology company. I really want to go to the parent company of TikTok of China, ByteDance to take up a data-related position. I think the core competitiveness of TikTok is the popular video recommendation algorithms, which can capture users' preferences and make users addicted for a long time. Not as private as TikTok's data, there are many datasets of YouTube available online. Therefore, YouTube's trending video dataset is suitable for me to study and it will help me get closer to my career goals and lay a foundation for me to experience similar projects or problems in the future. After I did some research, I found in 2005-2012, the view is considered as the primary indicator of trending. For now, even the more engaging of users is considered, the view still plays the dominant role in trending video. But what role other interaction factors play? Does channel or tags affect views? Does the relationship differ from categories? Therefore, the question I am going to figure out is that what are the factors, and what extent that affect how popular YouTube videos would be, in other word, the "views", in different categories. To see that, I will use a linear mixed effect (multilevel) model to explore that internal relationship.

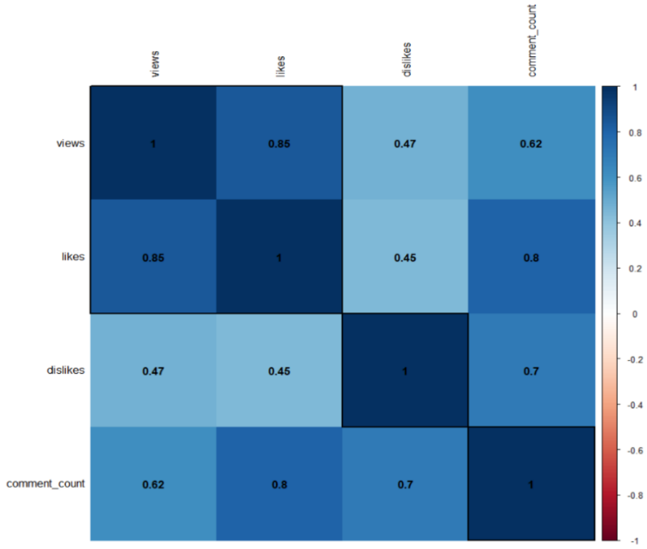
Method

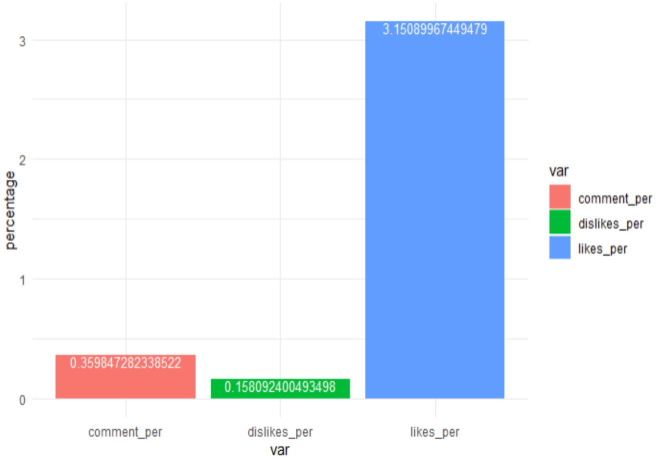
Data

The dataset I used is YouTube's daily list of trending videos from 2017 to 2018, provided by Mitchell J in Kaggle. His data provided data covers different countries, and I selected data from the United States for the following analysis. There are 16 meaningful variables in this dataset, such as publish time, tags, number of views, like or dislike, etc., total of 40726 observations, which provide me with enough predictors for using a multilevel model.

I cleaned data by eliminating NA values, and some missing values and error. Also, I transformed the format of date, and re-organized “tag” variable since it was messed up with punctuations. When I was doing EDA, I subset original data because it is too large, and I only need partial of it for visualization.

I did couples of exploratory data analysis to figure out what variables I should put in the regression model. I selected top channels and tags from the data, and I think they are not good predictors since they are in text from. I also made histograms and distributions of user interactions (views, likes, dislikes and comments), and I found they are in pretty similar patterns. Further, I generated the correlation matrix. It shows high correlation between likes and views, also between likes and comments. To support that, I got the mean of user interactions, and divided them by mean of views to see the percentage. The histogram figure shows that, on average, the number of likes accounts for 3% of the number of views.





Therefore, I extracted 5 variables from dataset for modeling. They are views, likes, dislikes, comments, and category id. I used “lme4” and “sjPlot” package to generate and check the multilevel model.

Models

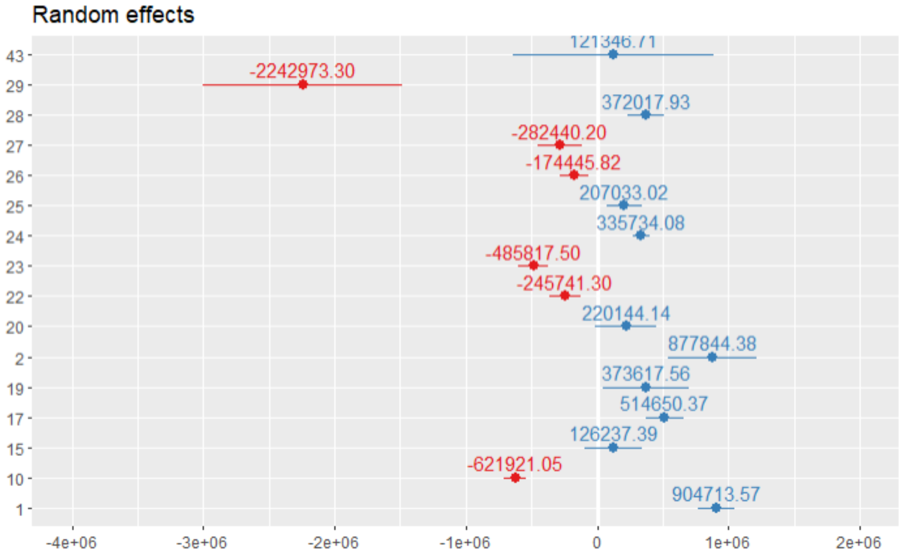
I tried multiple linear model, multilevel model, and Poisson model to fit this dataset.

My outcome is views, the number of views of a trending video. And I set likes, dislikes, comments as predictors. For multilevel model, I use video categories (category id) to be my random effect to account the difference among categories of YouTube.

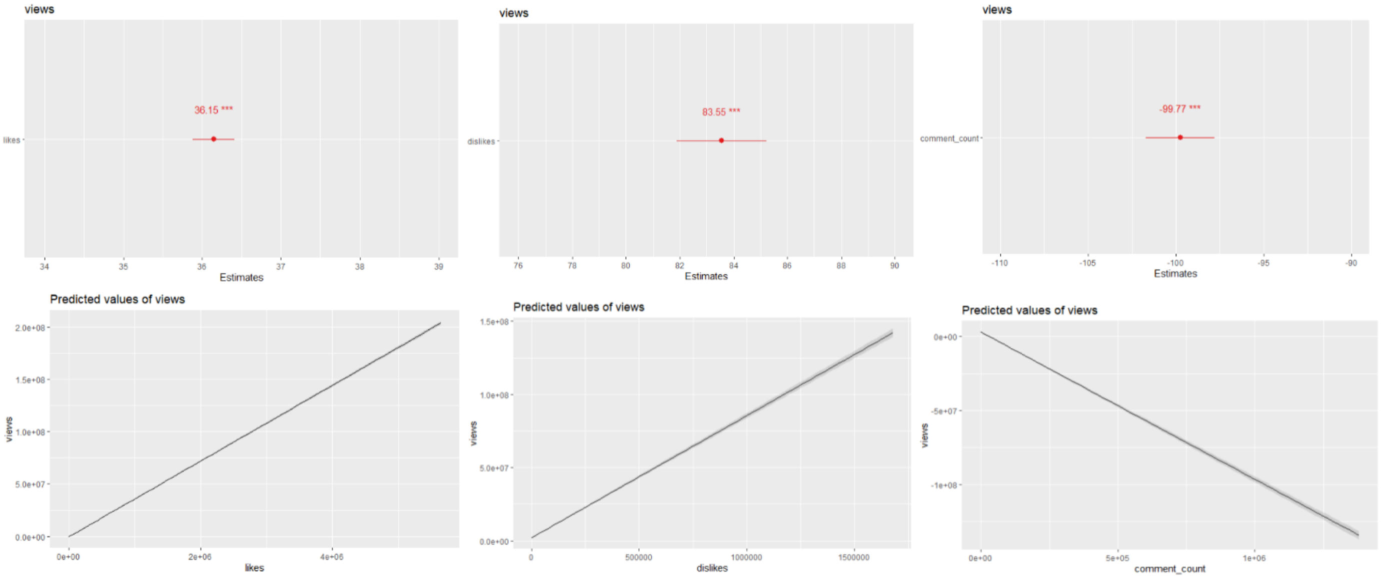
Since some of variables (views, likes etc.) are classified as character in original data, I transformed them into numeric form. I also converted category id into factors.

Result

```
model <- lmer(views ~ likes + dislikes+ comment_count+ (1|category_id),data=data)
```

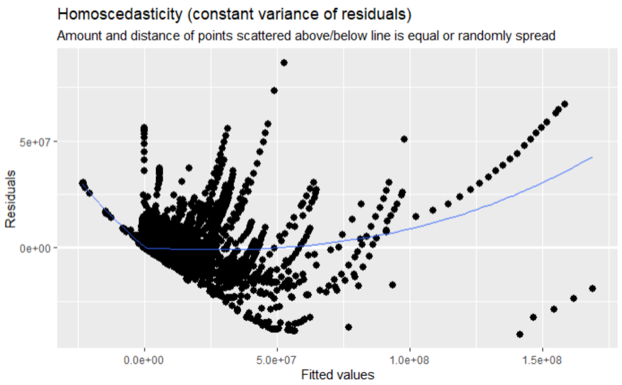


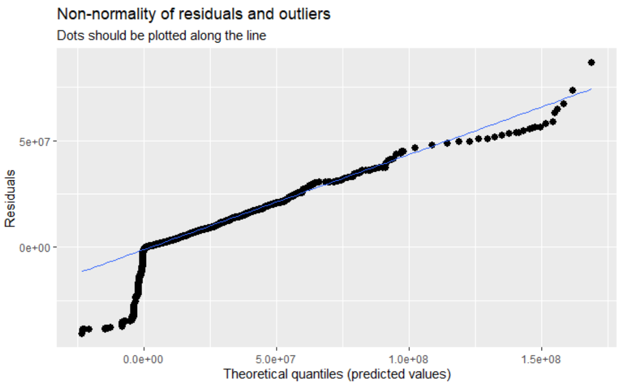
The figure above shows the random effects from categories. (The corresponding category id list is attached in Appendix) Different categories has different effect on views, for example, the Nonprofits & Activism (id=29) has especially wide error bars and very negative effect on views. That maybe due to the particularity of the views who are interesting in this kind of video.

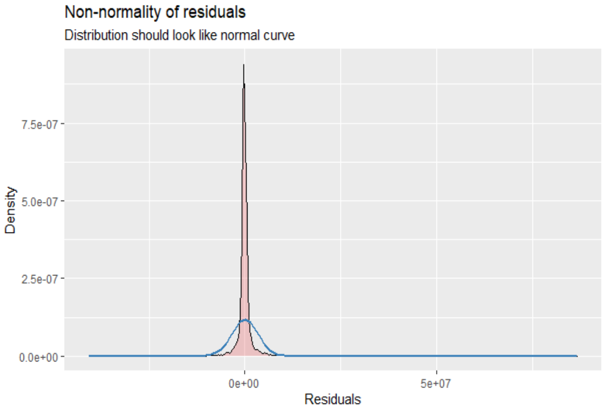


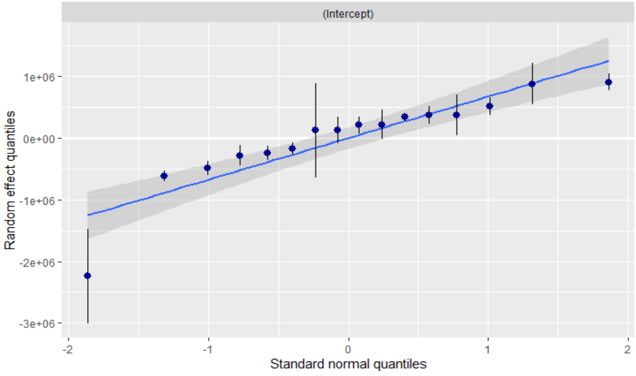
The figure above shows the estimates of three predictors and their corresponding prediction plot. To my surprise, dislikes has larger estimate than likes. With one increase of dislike, on average, the number of views would increase around 83.55. And as like increases in one unit, the number of views will increase around 36.15 on average. The comments have very negative effect on views that with one increase of comment, the views would averagely decrease 99.77. While watching the video, people are pleased to leave a like or dislike, however, most people are reluctant to comment since it costs time.

For the prediction, all of three predictors have good prediction. The likes was predicted most accurate, and the dislikes and comments have relatively bigger uncertainty. To illustrate that, I think because of difference of posting time, especially for some new trending videos, the number of comments and dislikes collected are kind of low compares to likes.









The figures above are the diagnostic plot of this multilevel model. As we can see, the residual plot shows a clustered shape and some outliers, which indicates that it is not a good fit for this regression model. I also tried multiple linear and Poisson model for that data, but they even show worse residual plot and normality. The Q-Q plot (figure2, 4) show a straight looking dashed line which indicates that our residuals are mostly normal distributed, and figure 3 shows the same thing.

Discussion

The results of this model are different from what I expected. Before that, I guessed that the number of likes is most related to the views, but the result is that the dislikes has a greater impact on views. In addition, the negative impact of comments was unexpected. However, this can be explained. This is actually a very common phenomenon. We often see that with the increase of the number of video views, the growth of comments is relatively slow, and the gap is getting larger and larger.

To be honest, I think this model produced a good prediction, since they all have thin certainty intervals. However, the residual plot shows my model did not fit very well, it is clustered and even shows Heteroscedasticity. I have tried to fix that. I rescaled predictors and added log, but there is no obvious change in the residual plot. In addition, changing the multilevel model did not solve the problem. Therefore, I think it may be due to the data itself. After doing the relevant research, I found that if there are extremely large or small values in the data, which may lead to the outlier. As we know, there are some videos extremely popular with pretty high number of user interactions.

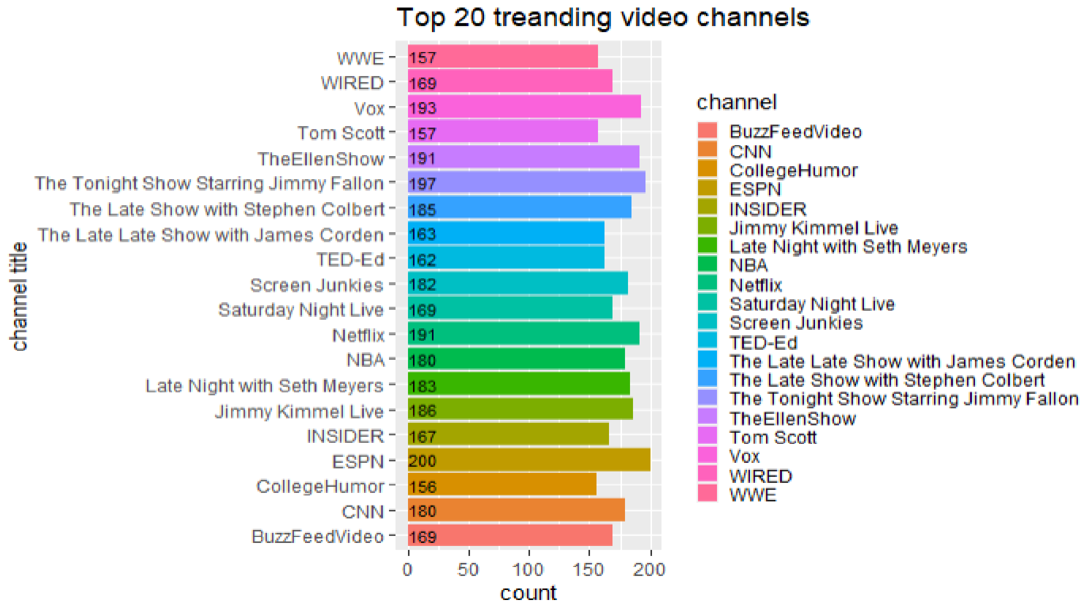
As I mentioned earlier, when the author of this data collecting, each video takes a different time after it is posted, so the data of each video will be quite different, even though they are popular videos. In order to get a more accurate prediction, it may be necessary to select videos that have experienced the same time. In addition, because I only selected the data from the United States, it cannot guarantee that other countries will have the same conclusion. In some other countries, the likes may affect views more than dislikes.

Appendix

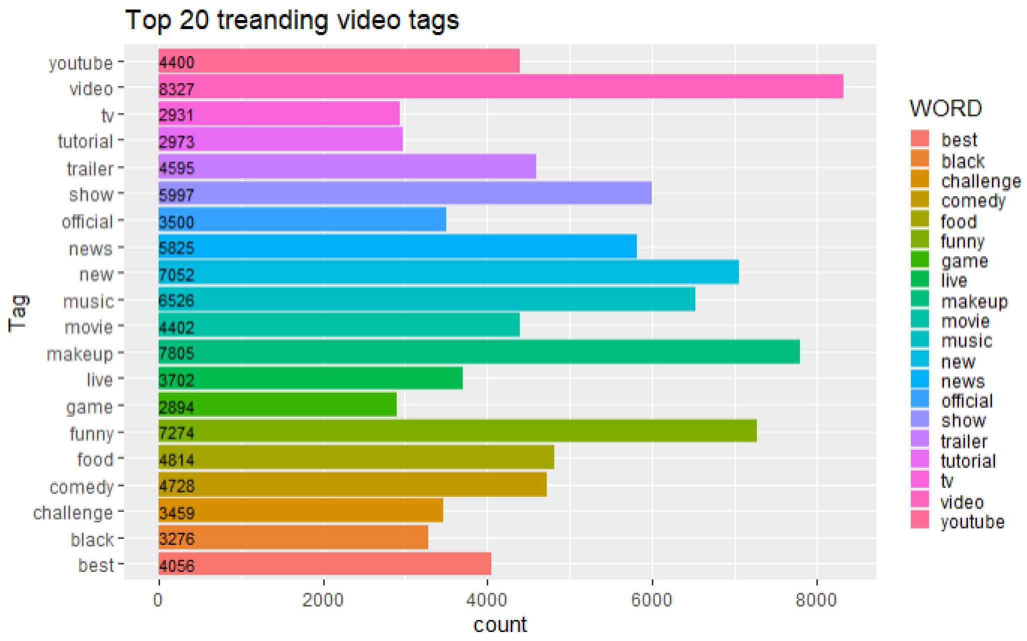
#the list of category id of YouTube

```
2 - Autos & Vehicles
1 -  Film & Animation
10 - Music
15 - Pets & Animals
17 - Sports
18 - Short Movies
19 - Travel & Events
20 - Gaming
21 - Videoblogging
22 - People & Blogs
23 - Comedy
24 - Entertainment
25 - News & Politics
26 - Howto & Style
27 - Education
28 - Science & Technology
29 - Nonprofits & Activism
30 - Movies
31 - Anime/Animation
32 - Action/Adventure
33 - Classics
34 - Comedy
35 - Documentary
36 - Drama
37 - Family
38 - Foreign
39 - Horror
40 - Sci-Fi/Fantasy
41 - Thriller
42 - Shorts
43 - Shows
44 - Trailers
```

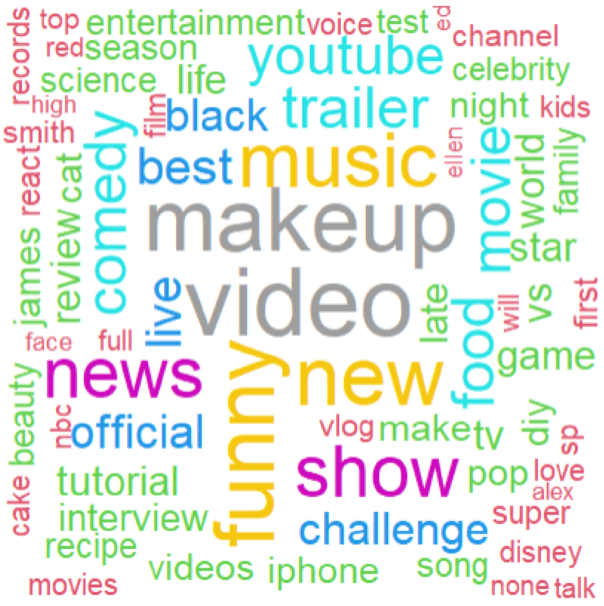
#the bar plot of top 20 channels



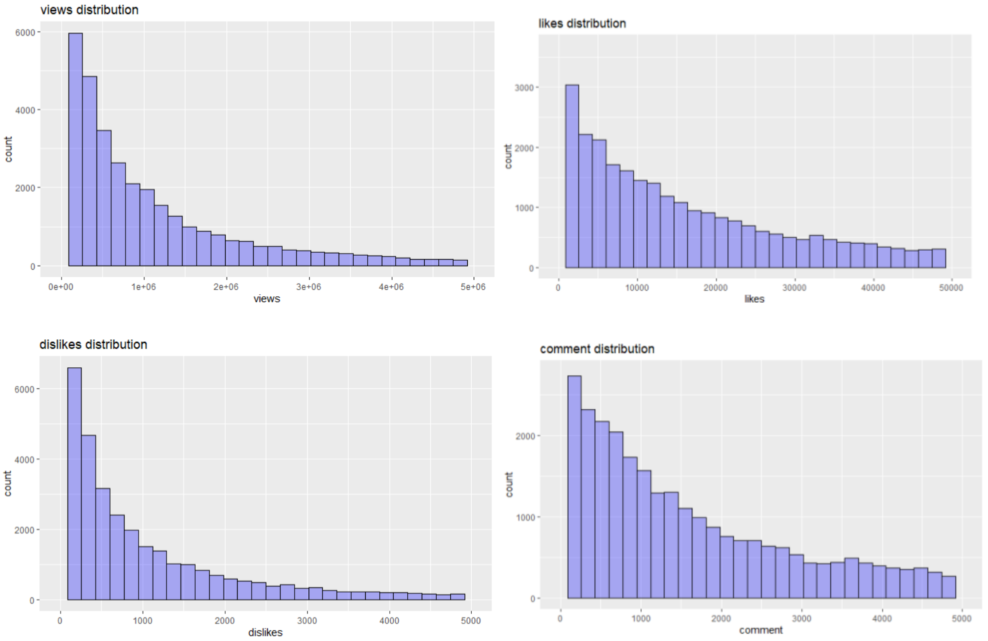
#the bar plot of top 20 tags



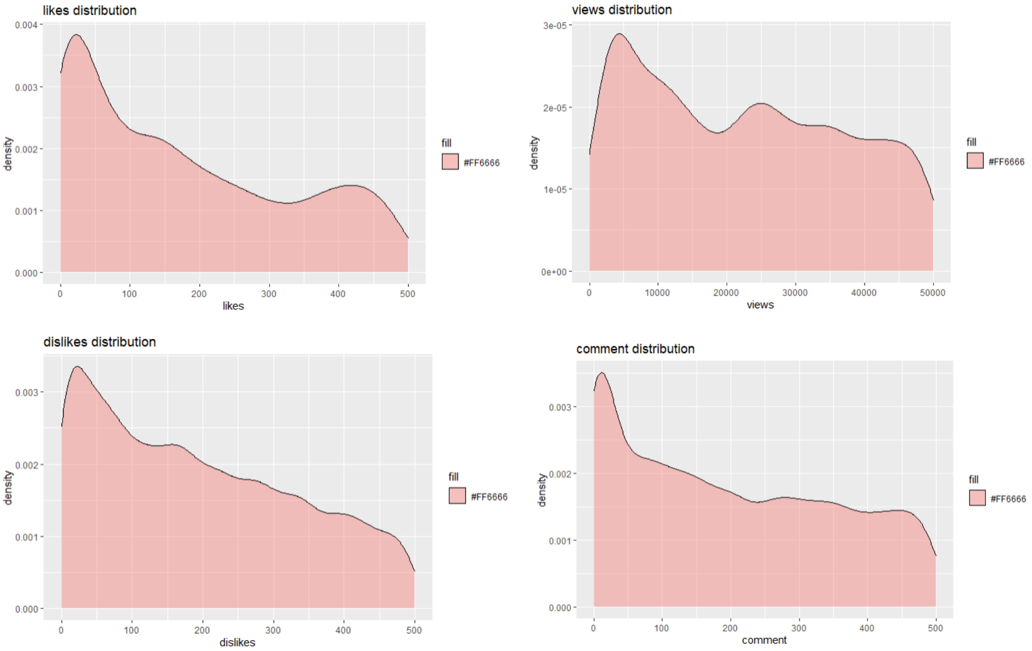
#the wordcloud of more top tags



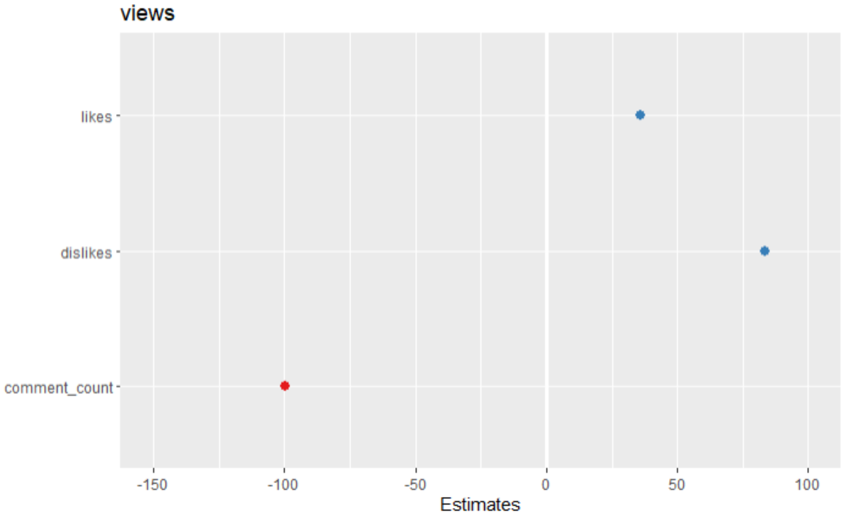
#the histogram plot of user interactions



#the distribution of user interactions



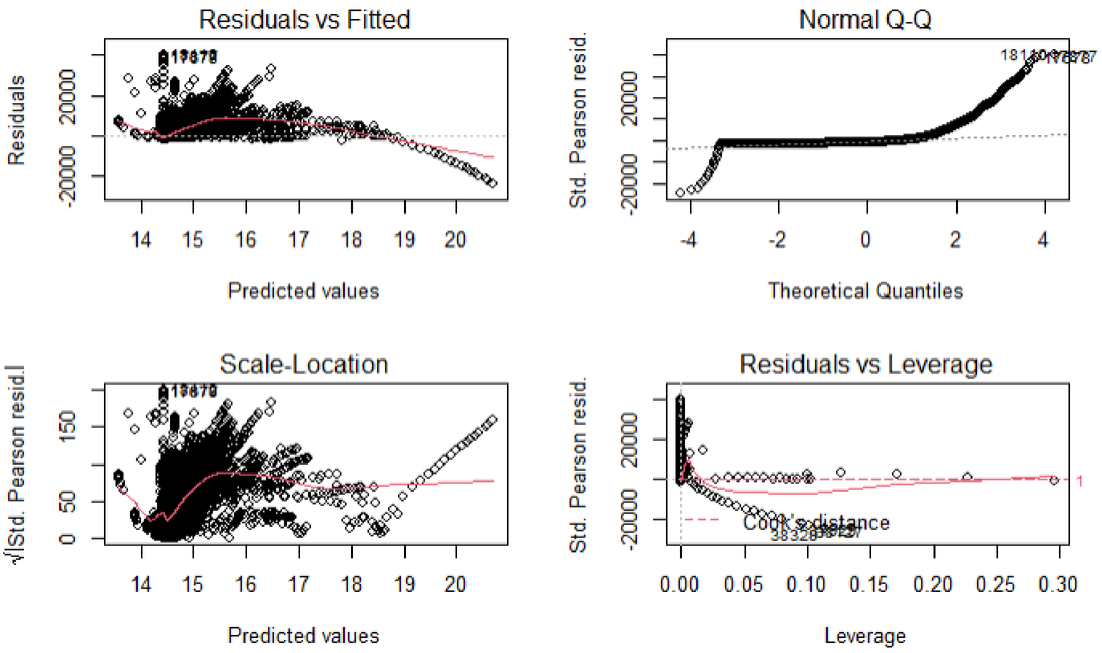
#estmate plot of user interactions



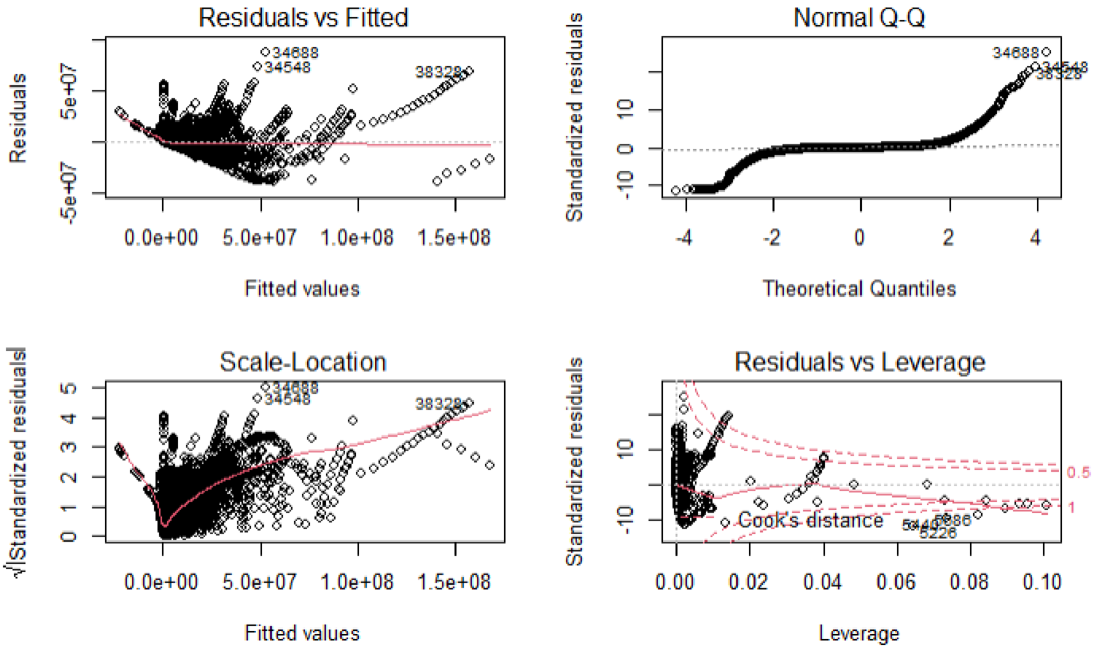
#output of lmer

```
Formula: views ~ likes + dislikes + comment_count + (1 + 1 | category_id)
    Data: data5

REML criterion at convergence: 1335270

Scaled residuals:
     Min         1Q     Median        3Q       Max
-11.8470   -0.1557   -0.0476     0.0896   25.2755

Random effects:
 Groups        Name          Variance   Std.Dev.
 category_id (Intercept) 5.731e+11   757030
 Residual                  1.174e+13 3426757
Number of obs: 40545, groups:   category_id, 16

Fixed effects:
                  Estimate Std. Error          df t value Pr(>|t|)
(Intercept)      1.826e+05   1.939e+05   1.074e+01     0.942      0.367
likes              3.615e+01   1.354e-01   4.052e+04 267.054     <2e-16 ***
dislikes          8.355e+01   8.503e-01   4.052e+04   98.251     <2e-16 ***
comment_count -9.977e+01   1.003e+00   4.051e+04 -99.504     <2e-16 ***
---
Signif. codes:   0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
              (Intr) likes   dislks
likes         -0.009
dislikes     -0.002   0.271
comment_cnt -0.005 -0.770 -0.637
```

#diagnostic plot of Poisson model



#diagnostic plot of multiple linear model


Residuals vs Fitted
Residuals
5e+07
-5e+07
34688
34548
38328
0.0e+00
5.0e+07
1.0e+08
1.5e+08
Fitted values
Normal Q-Q
Standardized residuals
10
0
-10
34688
34548
38328
-4
-2
0
2
4
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
5
4
3
2
1
0
34688
34548
38328
0.0e+00
5.0e+07
1.0e+08
1.5e+08
Fitted values
Residuals vs Leverage
Standardized residuals
10
-10
Cook's distance
5440
5686
5226
0.5
1
0.00
0.02
0.04
0.06
0.08
0.10
Leverage