INFO 6105: Data Science Engineering Methods SEC 02 Spring 2020

Trending YouTube Video Statistics Report

Team 11:

001443861 Hao Wu

001305642 Fangqing Wu

001082876 Yangyang Liu

001405396 Mengzhe Zhang

001302072 Lihang Zhou

Date: 4/15/202

Content

1 Introduction

1.1 Background

YouTube is one of the world's largest platforms for creating, sharing and discovering video content. As the world-famous video website, YouTube maintains an enormous large-scale dataset, and the recommendation system is responsible for helping more than a billion users discover personalized content from an ever-growing corpus of videos. To determine the year's top-trending videos, YouTube uses a combination of factors, including measuring users' interactions (number of views, shares, comments and likes).

Therefore, this project is based on dataset for recording daily record of the top-trending YouTube videos, which is aiming to perform analysis, visualization, prediction and recommendation.

Based on the results, we could plan to be extracted about personal preferences of users from all over the world in this dataset, which will contribute to the training of several prediction models.

1.2 Objectives

According to the conclusions of the analysis, several features are planned to be extracted about personal preferences of users from all over the world in this dataset, which will contribute to the training of several prediction models. The functionalities are as follows:

1. For a user, three possibilities of favorite kinds of video would be predicted according to the behaviors of other users who share the similar attributes among the dataset (e.g. residence in the same region, similar watching history, similar subscribing video channels, etc.).

2. For a video to be published or just published, the number of comments and likes for it would be predicted, as well as the attitudes of most comments according to its details, tags and the quality of content.

3. Even taking the stand of YouTube the company, a model would be established for recommendation of specific kinds of video towards the specific kinds of users (e.g.

region, age group, kind of business, etc.).

1.3 Methodology

In order to achieve this project goals, obtain reliable conclusions and offering scientific recommendation, data science applications, programming tools and algorithms are significant necessary to our methodology. The following is the detailed method list to be shown in this project:

- Programming language: Python
- Basic tools: NumPy, pandas, seaborn, matplotlib, json, datetime, IPython, glob
- Machine Learning tools: train_test_split, LinearRegression, PolynomialFeatures, r2_scores， Polynomial Regression, Random Forest, K-means Cluster
- Advanced Tools: nltk, WordCloud, spacy, TextBlob
- Algorithm:
  Machine Learning (Linear Regression, Polynomial Regression, Random Forest, K-fold training, K-means Cluster,r2_score)
  Natural Language Processing:(Word frequency statistic, Sentiment recognition)

1.4 Method Solving

Based on the data science methods, here are the specific ways to implement methodology and get analytical results:

- Using pandas, NumPy and matplotlib processed a general visualization of data: e.g. The most influential video publisher; diverse topics; the videos with the most dislikes; the distribution of LIKES & DISLIKES and so on.
- Using IPython and glob to display a Html-source dataset (photo grids).
- After getting a view of dislikes, likes, comment-counts, and views, getting correlations and heatmap for the modified dataset. This relationship would be used in latter prediction model, by using Polynomial Regression and R^2 score for evaluation.
- Using wordcloud to count the frequency of words in description, tags, and title.

- Using nltk to analyze the sentiments of words and getting a view of it.

1.5 Task Description

Main tasks need to be identified to support the completion of this project and accuracy of the results. Several basic steps are listed below to guarantee our progress:

- Distribution of attributes
- Correlation between Views & Likes
- Prediction about number of Likes
- Recommendation system
- Comment sentiment recognition

2 Data Preprocessing and Feature Engineering

2.1 Data Source

The data is from Kaggle and the reference is:

https://www.kaggle.com/datasnaek/youtube-new

2.2 Data description

This dataset includes several months (and counting) of data on daily trending YouTube videos. Data is included daily top-trending videos' information from 10 regions (up to 200 per day): US, GB, DE, CA, FR, RU, MX, KR, JP, IN (USA, Great Britain, Germany, Canada, France,Russia, Mexico, South Korea, Japan and India). Each region's data is in a separate file.

Every dataset for different countries has 16 columns in total and they are: video_id, trending_date, title, channel_title, category_id, publish_time, tags, views, likes, dislikes, comment_count, thumbnail_link, comments_disabled, ratings_disabled, video_error_or_removed, description

These csv files are diverse and huge enough to conduct a comprehensive and systematic study. What’s more, the data quality is superb and complete with seldom missing values to deal with. According to these conditions, a trustworthy and precise report is predictable.

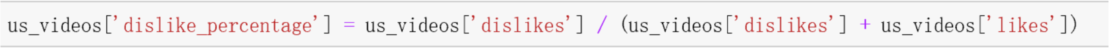
Data selection is important for the purpose of highly efficient project procedure. Among the large data quantity, obviously the key parameters are video title, channel title, publish time, tags, views, likes and dislikes, description, and comment count, which are also the main columns chosen to study on in this project. Besides, foreign languages from several areas are hard to be recognized by programming language and may cause interference in the system, so our project will focus on English-based countries, and the entire process is mainly based on the data from the United States

2.3 Data Summarization

First of all, we need to figure out the scope of the predictors and be familiar with the general appearance of the parameters. Summarization and feature engineering are essential parts. After some data preprocessing steps show in the python file, several results are presented:

2.3.1 Dislike Percentage

For the Dislike category, we can use the formula below to learn about a few cases:



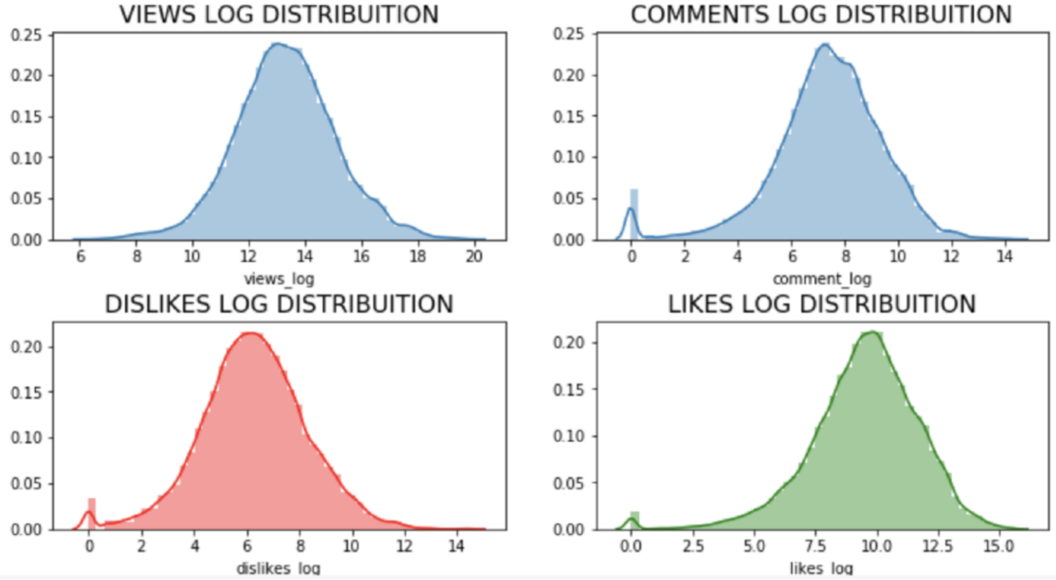
```
us_videos['dislike_percentage'] = us_videos['dislikes'] / (us_videos['dislikes'] + us_videos['likes'])
```

And the percentage results are shown:

```
0       0.014112
1       0.053712
2       0.028718
3       0.044185
4       0.047064
          ...
529     0.035998
530     0.097533
531     0.136068
532     0.036491
533     0.054054
Name: dislike_percentage, Length: 534, dtype: float64
```

2.3.2 Distribution of basic parameters

For the parameters of views, comment count, likes and dislikes, we can use log form to display their distribution separately with curves:

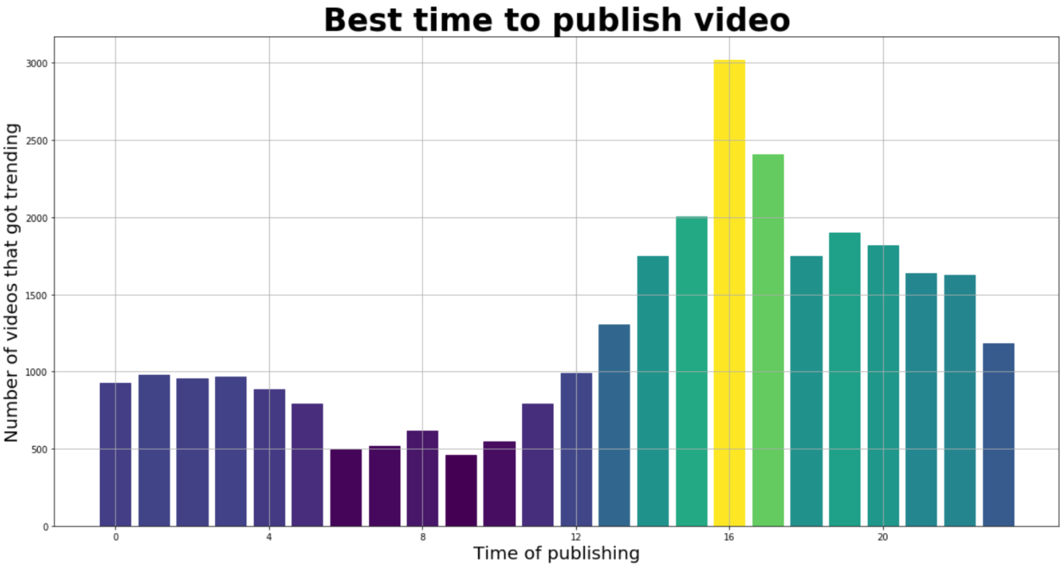


After log function, the variables: views, comment, dislikes and likes follow normal distribution. Then we can do analysis and model based on this condition.

3 Visualization and Analysis

In order to further study and analyze the relationship among parameters, we carry on more visualization missions to exhibit the characteristics of the video, so that the preferences of the audience are clearer to explore.

3.1 Best Time to Publish Video



From the char above, we can conclude the publishing hour should between 14 and 19, and then it will be easier to get trending.

3.2 Most Trending Videos

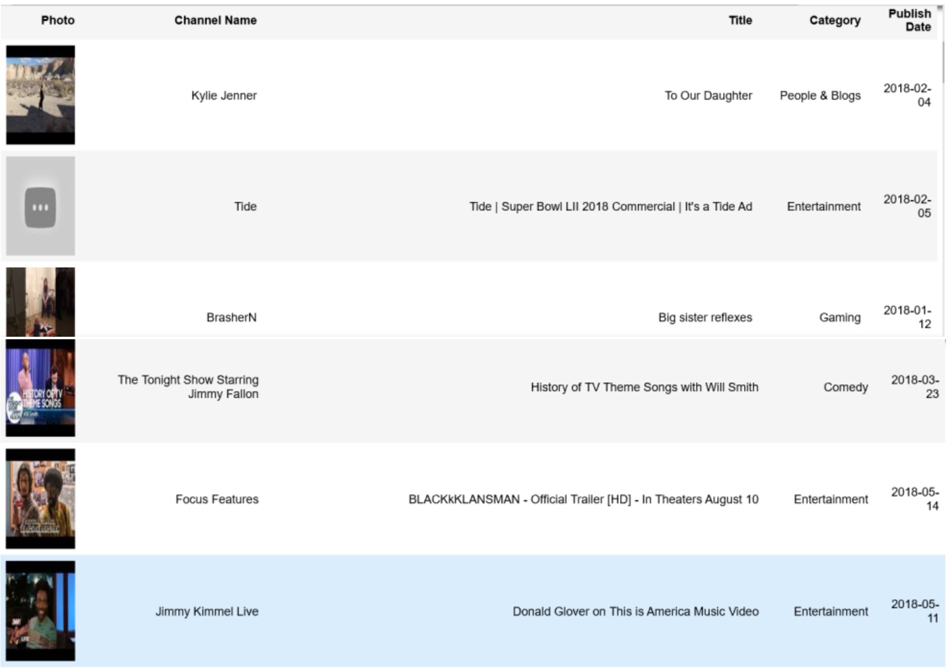
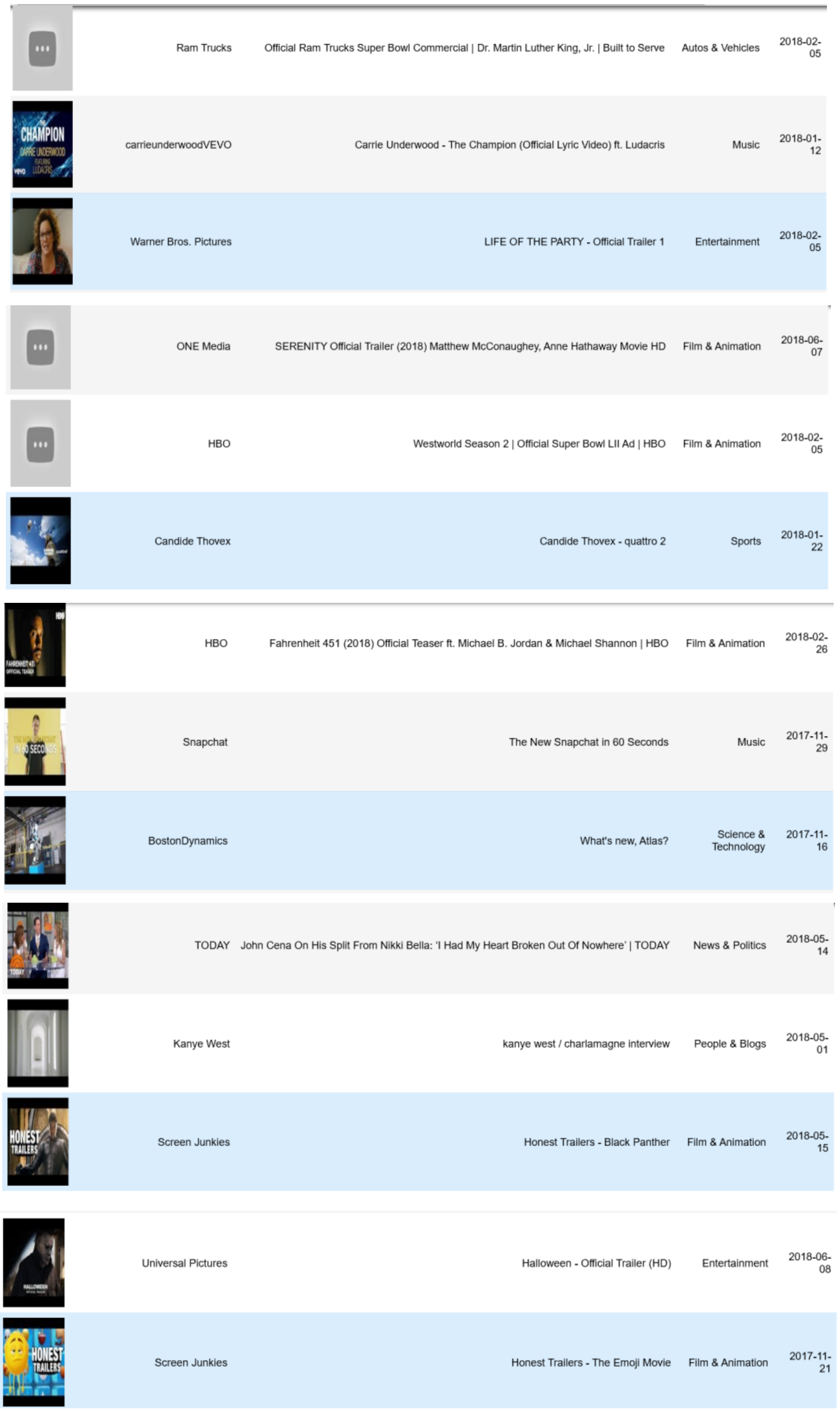


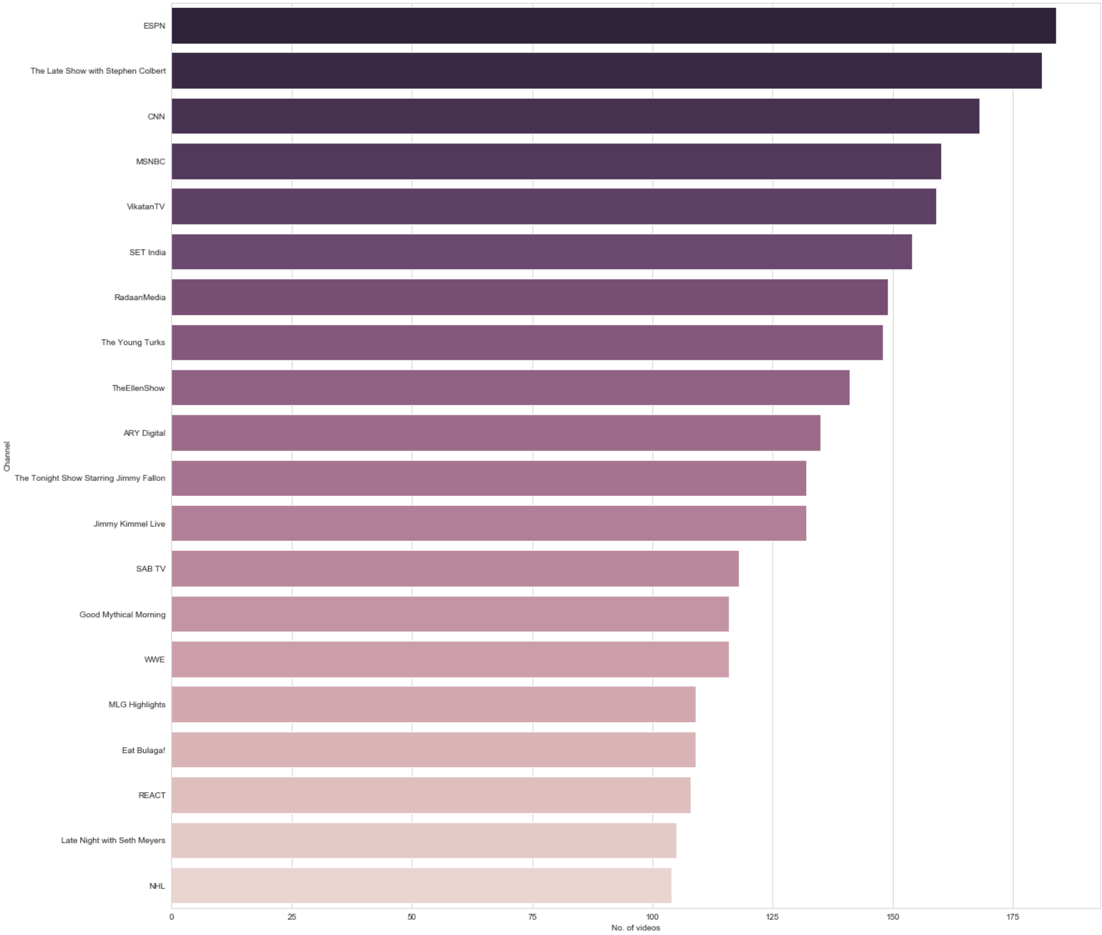
Photo	Channel Name	Title	Category	Publish Date
	Kylie Jenner	To Our Daughter	People & Blogs	2018-02-04
	Tide	Tide | Super Bowl LII 2018 Commercial | It's a Tide Ad	Entertainment	2018-02-05
	BrasherN	Big sister reflexes	Gaming	2018-01-12
	The Tonight Show Starring Jimmy Fallon	History of TV Theme Songs with Will Smith	Comedy	2018-03-23
	Focus Features	BLACKkKLANSMAN - Official Trailer [HD] - In Theaters August 10	Entertainment	2018-05-14
	Jimmy Kimmel Live	Donald Glover on This is America Music Video	Entertainment	2018-05-11



	Ram Trucks	Official Ram Trucks Super Bowl Commercial | Dr. Martin Luther King, Jr. | Built to Serve	Autos & Vehicles	2018-02-05
	carrieunderwoodVEVO	Carrie Underwood - The Champion (Official Lyric Video) ft. Ludacris	Music	2018-01-12
	Warner Bros. Pictures	LIFE OF THE PARTY - Official Trailer 1	Entertainment	2018-02-05
	ONE Media	SERENITY Official Trailer (2018) Matthew McConaughey, Anne Hathaway Movie HD	Film & Animation	2018-06-07
	HBO	Westworld Season 2 | Official Super Bowl LII Ad | HBO	Film & Animation	2018-02-05
	Candide Thovex	Candide Thovex - quattro 2	Sports	2018-01-22
	HBO	Fahrenheit 451 (2018) Official Teaser ft. Michael B. Jordan & Michael Shannon | HBO	Film & Animation	2018-02-26
	Snapchat	The New Snapchat in 60 Seconds	Music	2017-11-29
	BostonDynamics	What's new, Atlas?	Science & Technology	2017-11-16
	TODAY	John Cena On His Split From Nikki Bella: 'I Had My Heart Broken Out Of Nowhere' | TODAY	News & Politics	2018-05-14
	Kanye West	kanye west / charlamagne interview	People & Blogs	2018-05-01
	Screen Junkies	Honest Trailers - Black Panther	Film & Animation	2018-05-15
	Universal Pictures	Halloween - Official Trailer (HD)	Entertainment	2018-06-08
	Screen Junkies	Honest Trailers - The Emoji Movie	Film & Animation	2017-11-21

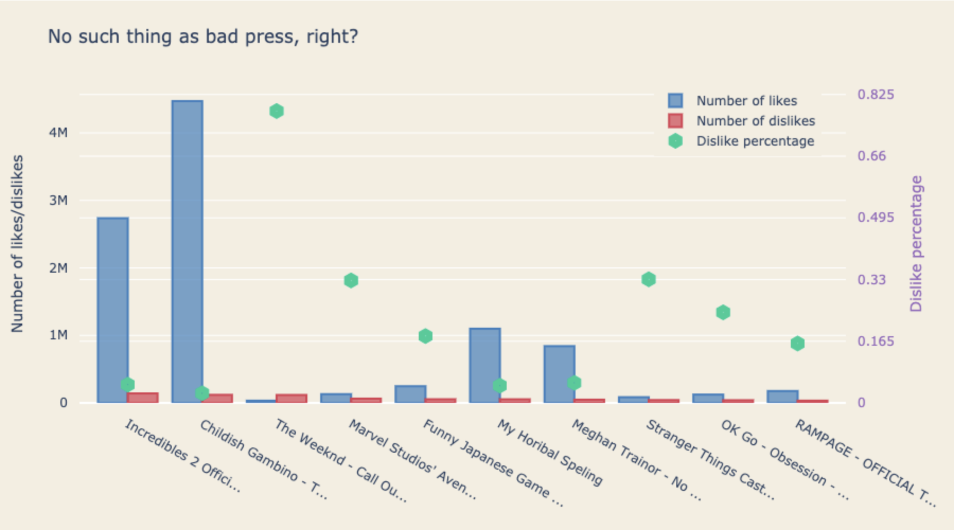
We get the information of the top popular videos, which have good reference meaning for preference study, and the first one is To Our Daughter from Kylie Jenner Channel.

3.3 The Most Influential Video Publisher

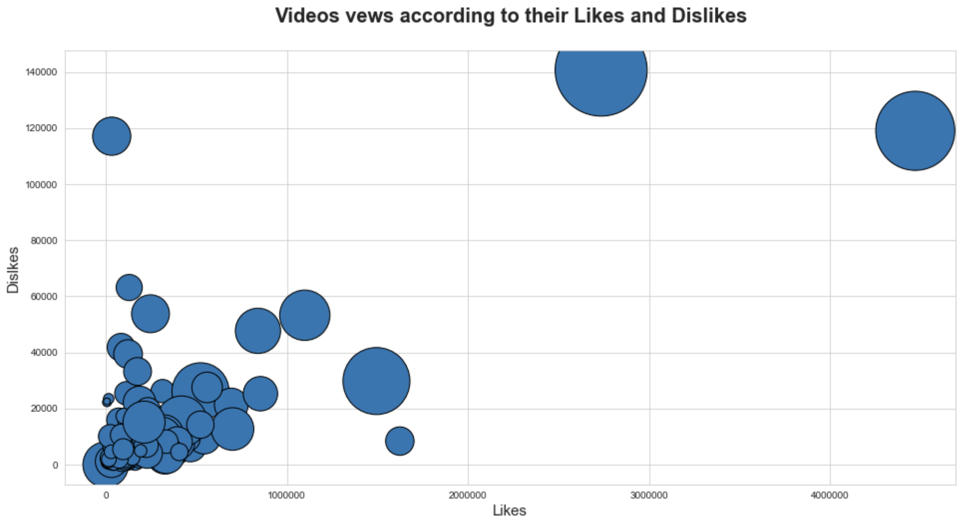


We got the most influential creators and channels, which are listed in the diagram. The tops are ESPN, The Late Show with Stephen Colbert, CNN…

3.4 The Videos with The Most Dislikes

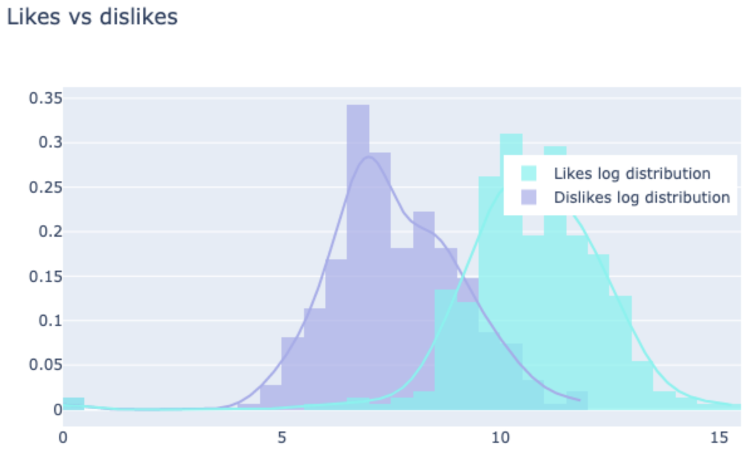


The videos get most dislikes number are shown, but most of them still get much more likes number.



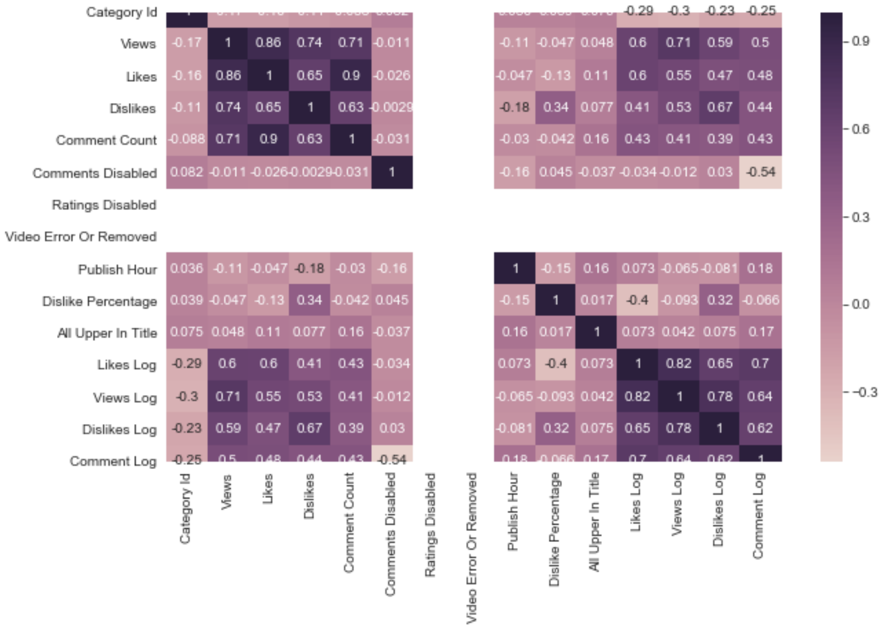
It shows videos relation between its own likes and dislikes number, which presents it is roughly linear, but for many videos, the likes number is much higher than its dislikes number.

3.5 The Distribution of LIKES & DISLIKES

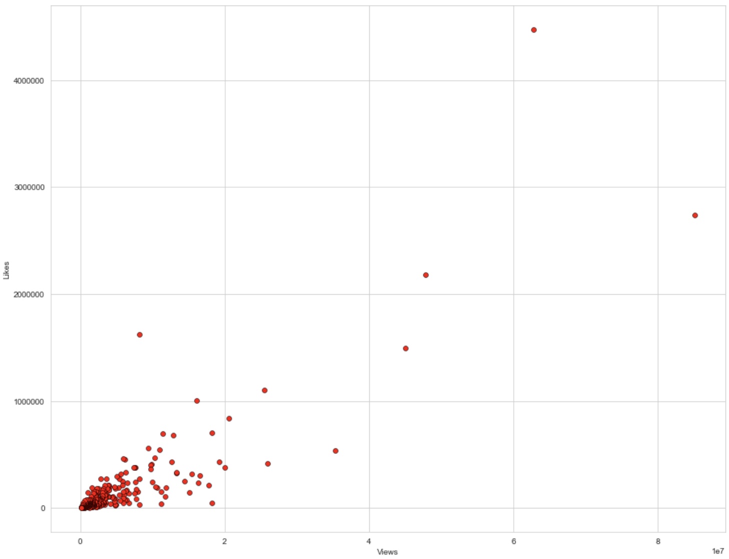


From the distribution contrast, it shows the vide groups that easily get concentrated likes or dislikes.

3.6 Correlation between Dataset Variables

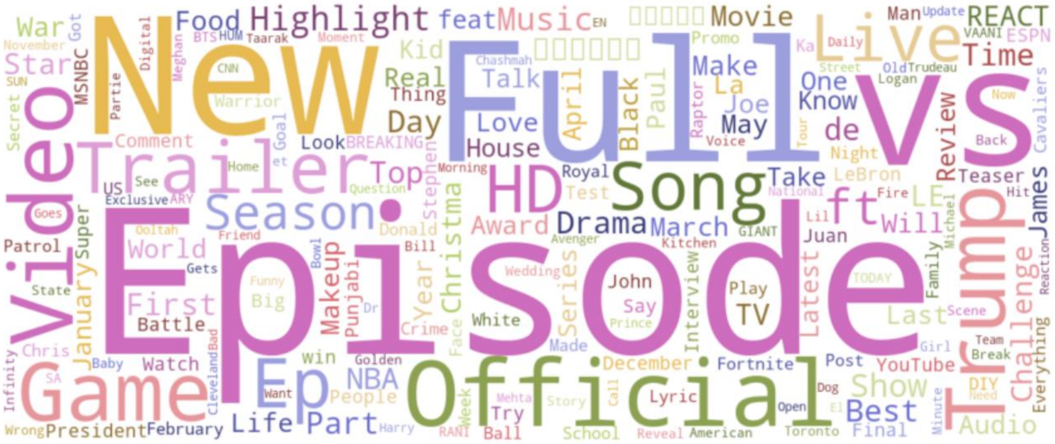


From the heat map, the variable correlations are shown clearly. The deeper color, the stronger relationship. We get that the variables likes and views, likes and comment count, dislikes and views, comment count and views are highly correlated.



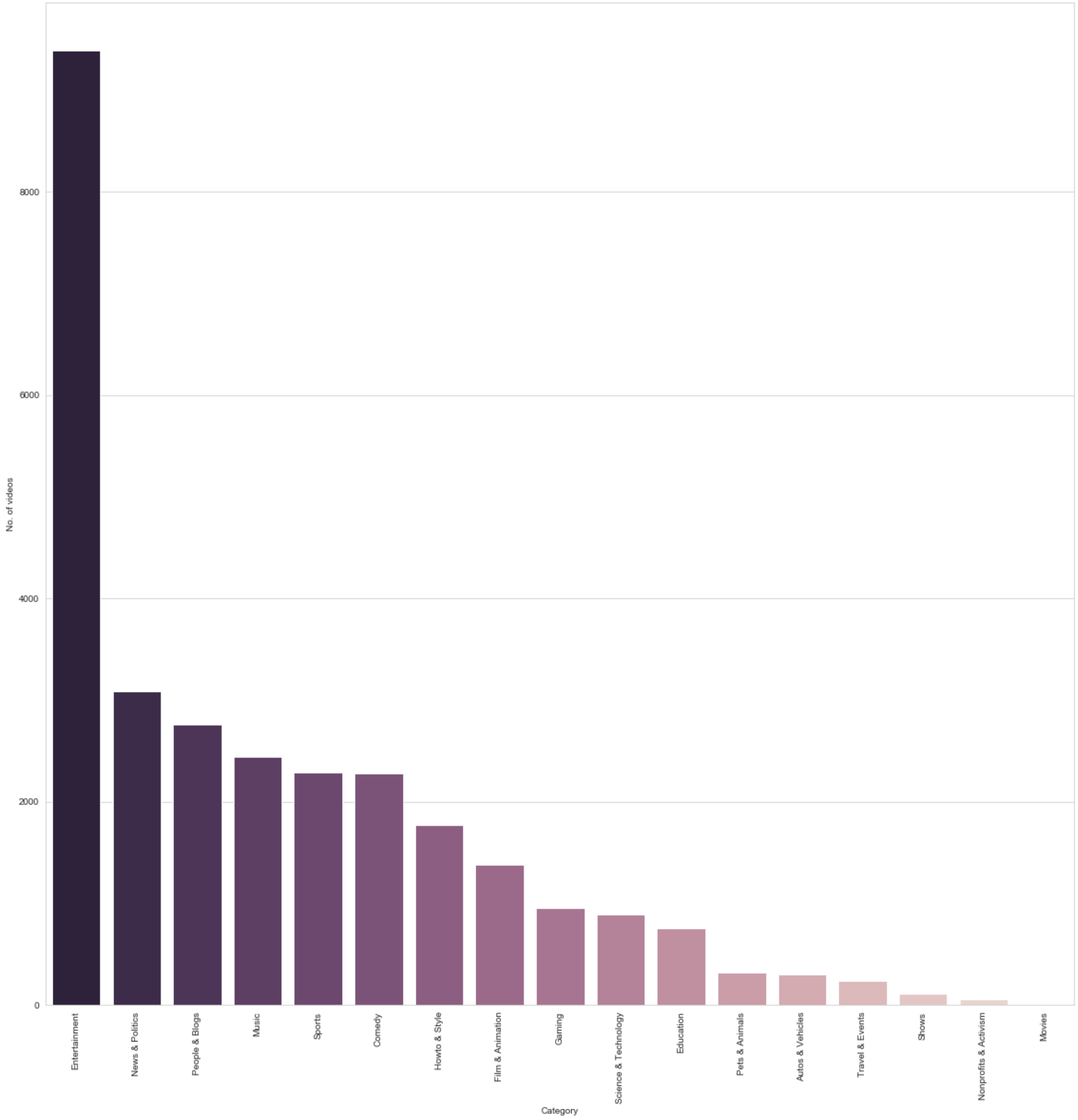
By the scatter plot between views and likes, they generally have a positive linear correlation.

3.7 Most Common Words in Video Titles

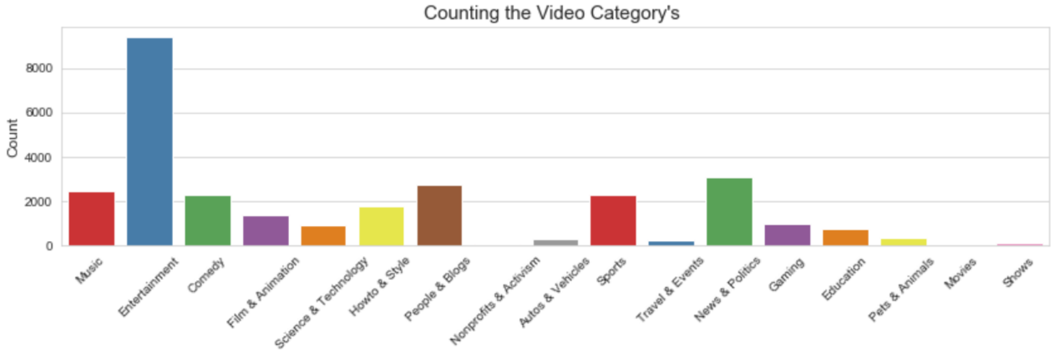


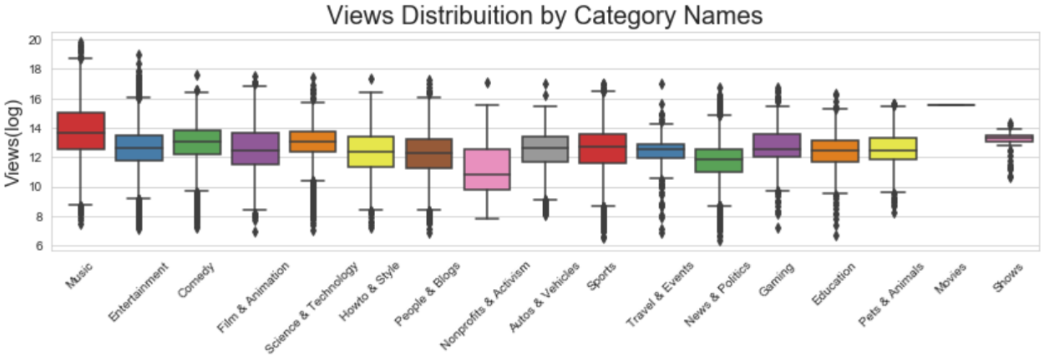
The popular words in videos are obviously to get, such as episode, official, full, game, song, Trump and so on, which are good information to keep an eye on.

3.8 The Distribution of Videos from Categories



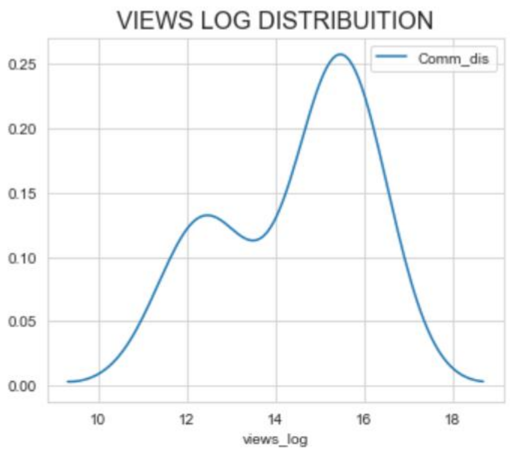
Most popular categories IS entertainment, and Its trending value is absolutely high.

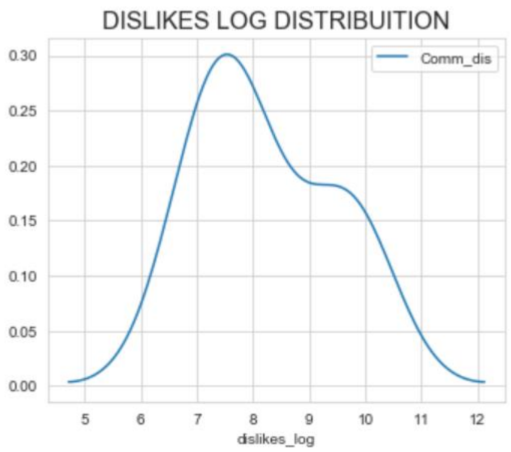


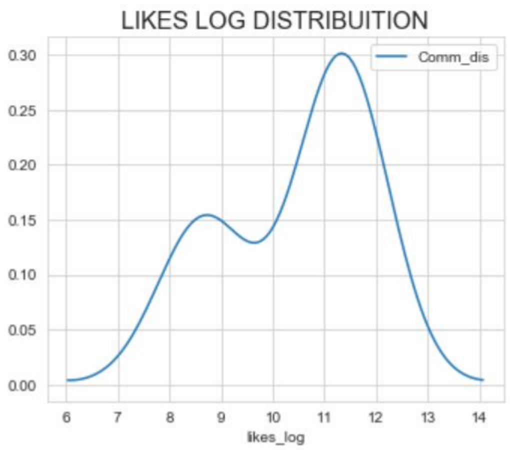


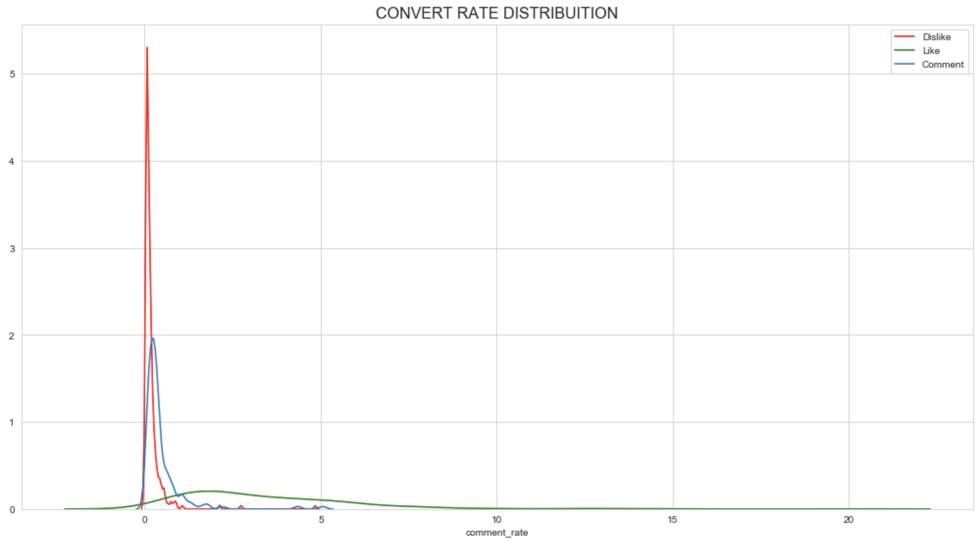
From the information from distribution and box charts, we can get their median, 5th, 25th, 75th, 95th percentile and minimum and maximum values. In general, categories of music, film & animation, science & technology, sports are easier to get more views.

3.9 Simple Feature Extraction from The Number of Likes, Dislikes and Comments

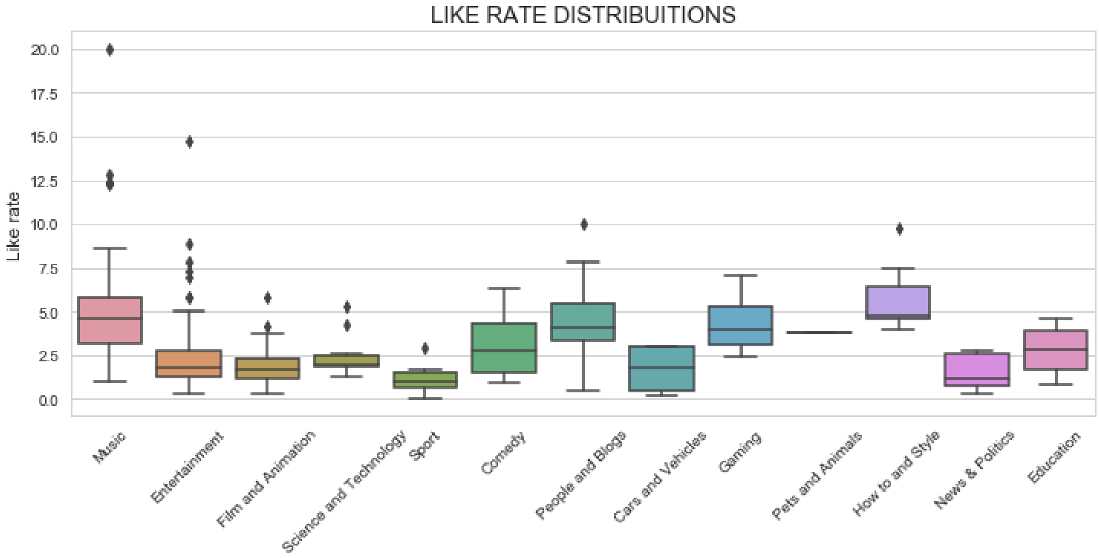


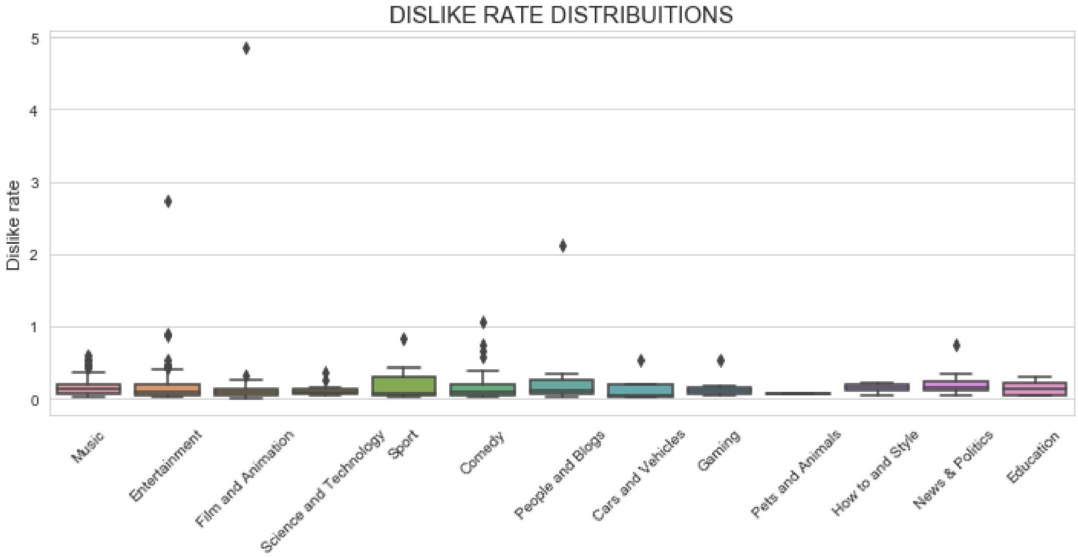


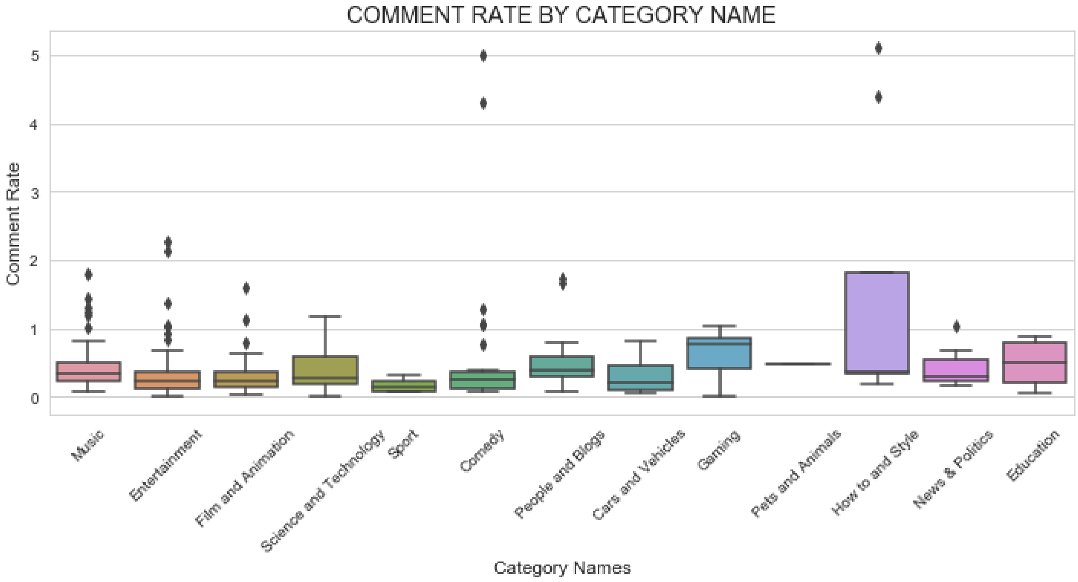




There is some important information in our dataset. The Comment & Dislike are similarly distributed. And then we need to have deeper research.


LIKE RATE DISTRIBUITIONS
Like rate
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
Music
Entertainment
Film and Animation
Science and Technology
Sport
Comedy
People and Blogs
Cars and Vehicles
Gaming
Pets and Animals
How to and Style
News & Politics
Education


DISLIKE RATE DISTRIBUITIONS
Dislike rate
5
4
3
2
1
0
Music
Entertainment
Film and Animation
Science and Technology
Sport
Comedy
People and Blogs
Cars and Vehicles
Gaming
Pets and Animals
How to and Style
News & Politics
Education

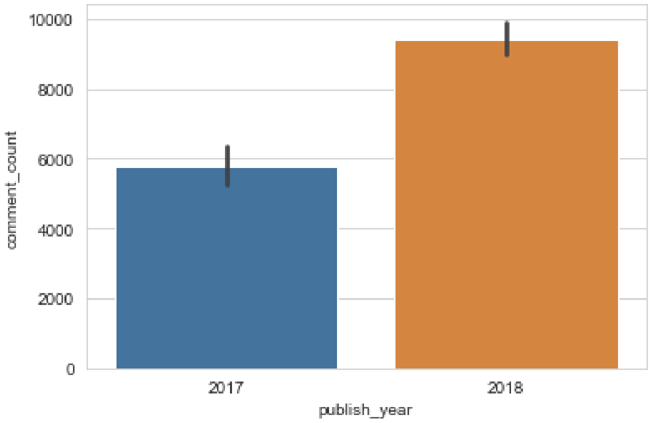

COMMENT RATE BY CATEGORY NAME
Comment Rate
5
4
3
2
1
0
Music
Entertainment
Film and Animation
Science and Technology
Sport
Comedy
People and Blogs
Cars and Vehicles
Gaming
Pets and Animals
How to and Style
News & Politics
Education
Category Names

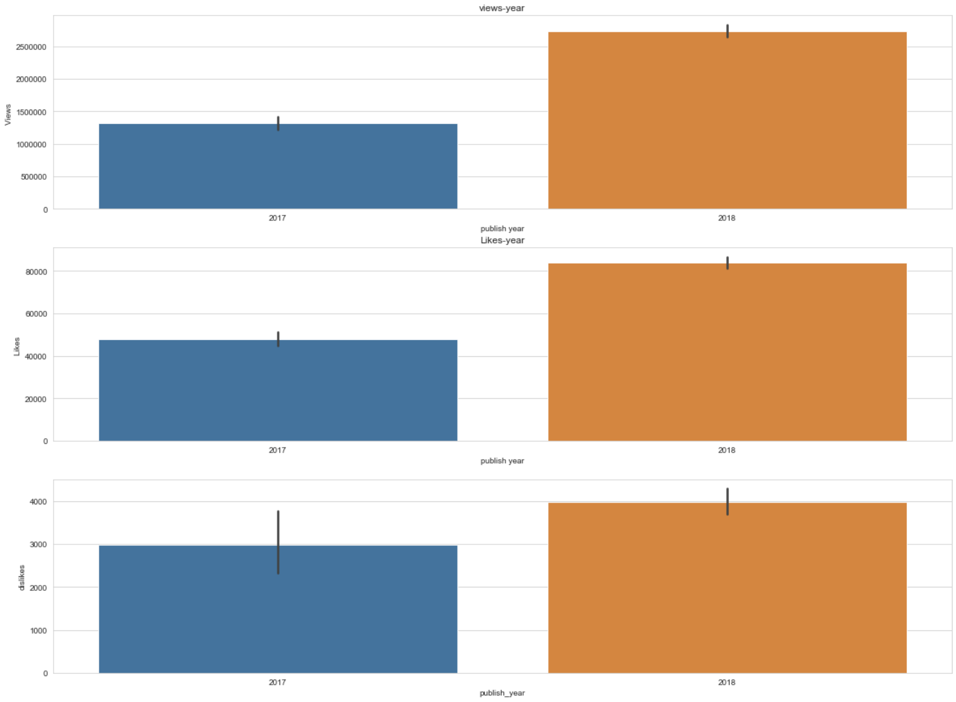
At the meantime, we can see that the Dislike rate is very low in almost all categories but some outliers in 'People and Blogs' and News & Politics that we can might can consider "Normal".

The mean of like distribution is less than 5% but in music we have a very interesting pattern of like rate. The music category has the highest engagement rate.

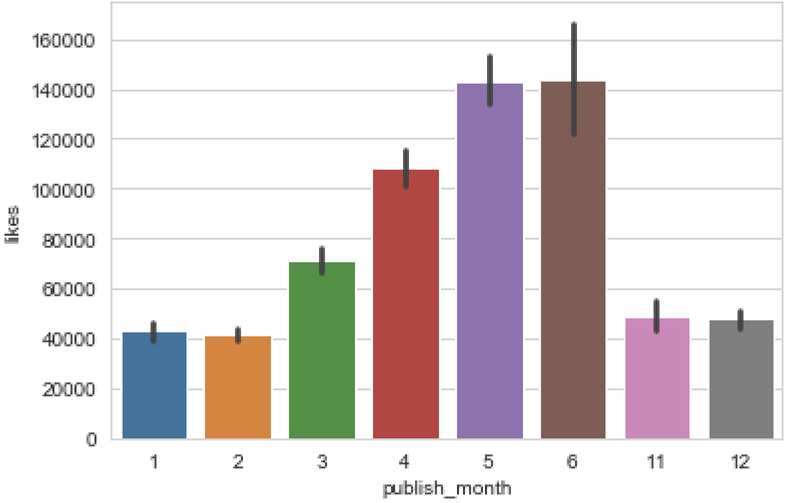
And at the Comment rate we can see the how-to category with the highest rates of comments.

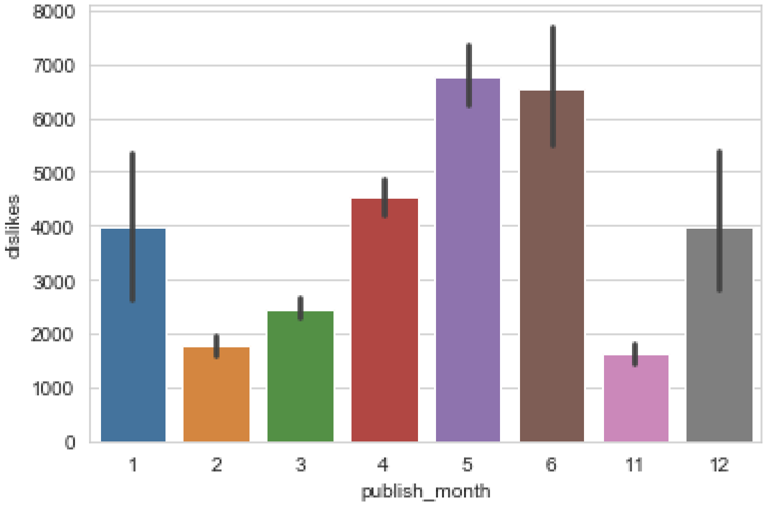
3.10 Visualization on publishing time

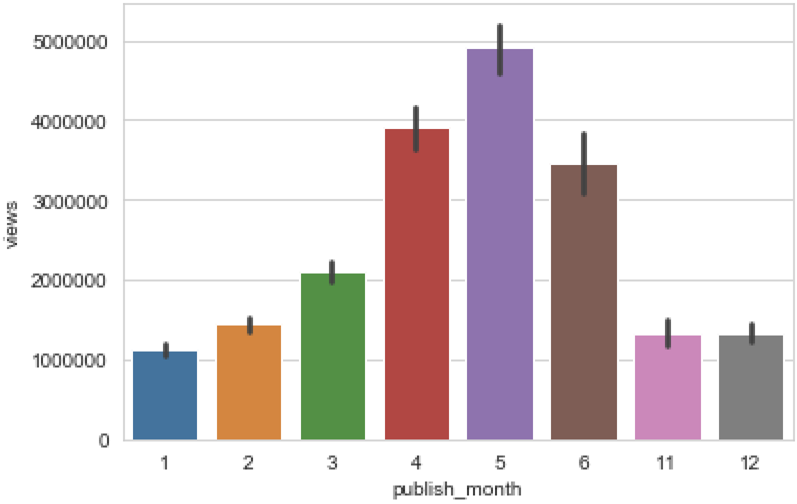


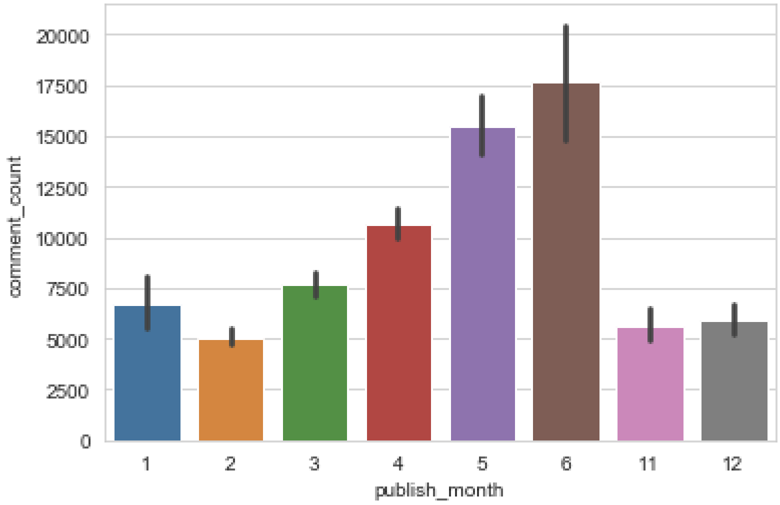


Comparing these two year: 2017 and 2018, all the four variables: comment_count, views, likes, and dislikes get increased. We can preliminarily infer that it is because the user base has increased.





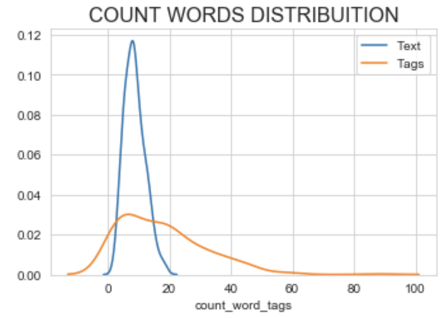


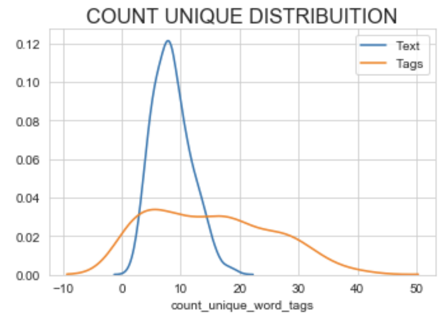


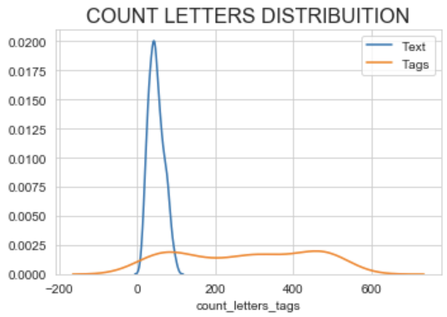
From the four parameters' situation, because all the values become higher in the summer, we can get that YouTube users are more active from April to June, and videos are more likely to get attentions during these months.

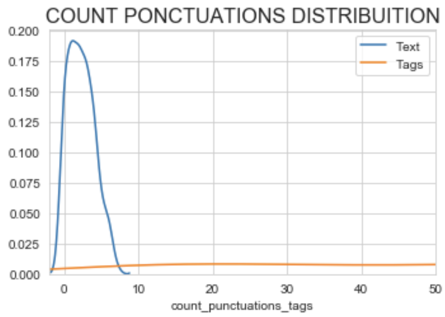
4 Extract information from Tags & Description

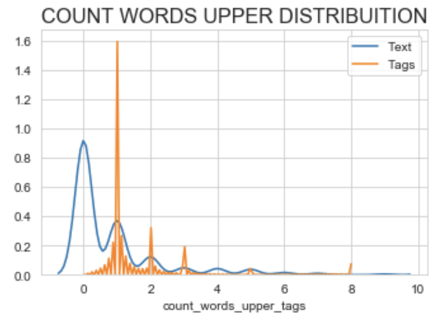
4.1 Statistic & Preprocessing

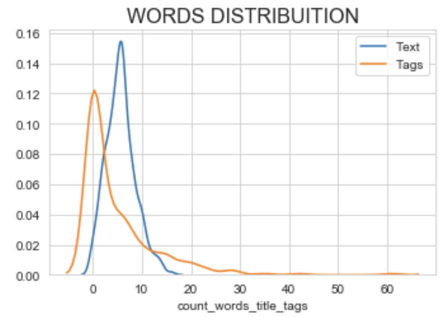

COUNT WORDS DISTRIBUITION

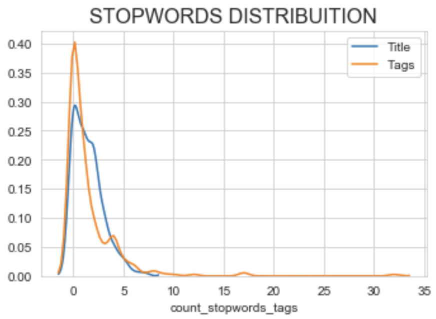

COUNT UNIQUE DISTRIBUITION

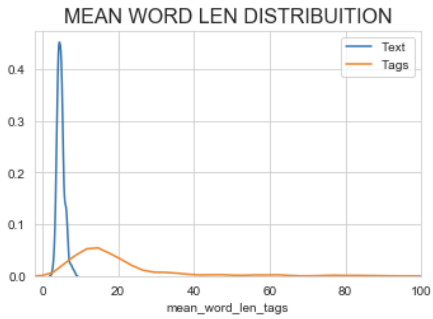

COUNT LETTERS DISTRIBUITION


COUNT PONCTUATIONS DISTRIBUITION


COUNT WORDS UPPER DISTRIBUITION

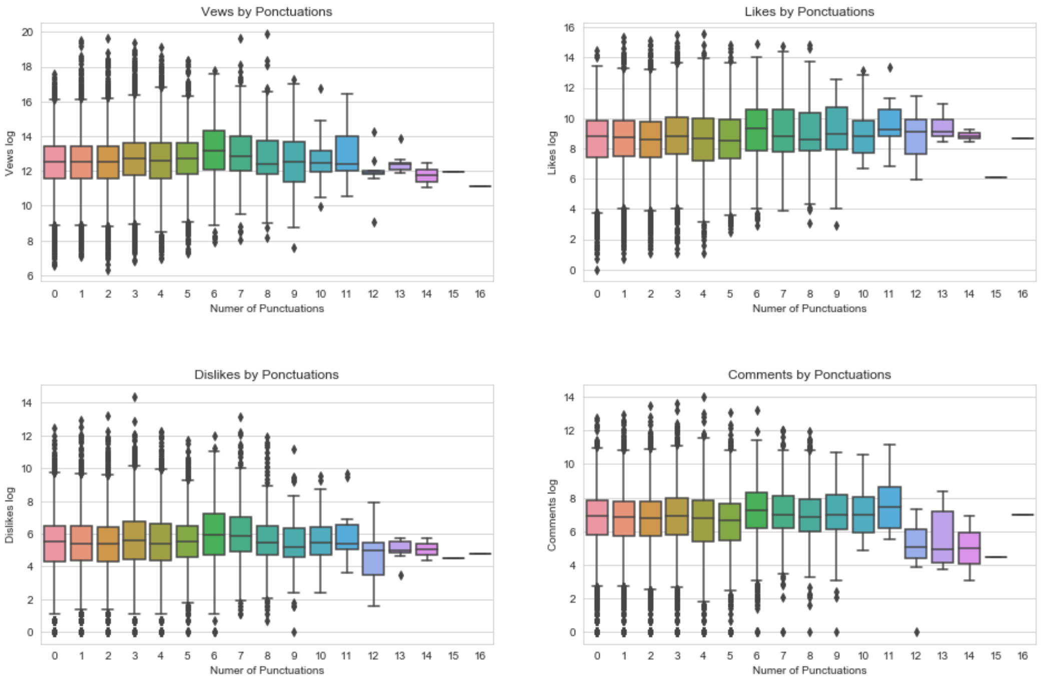

WORDS DISTRIBUITION


STOPWORDS DISTRIBUITION

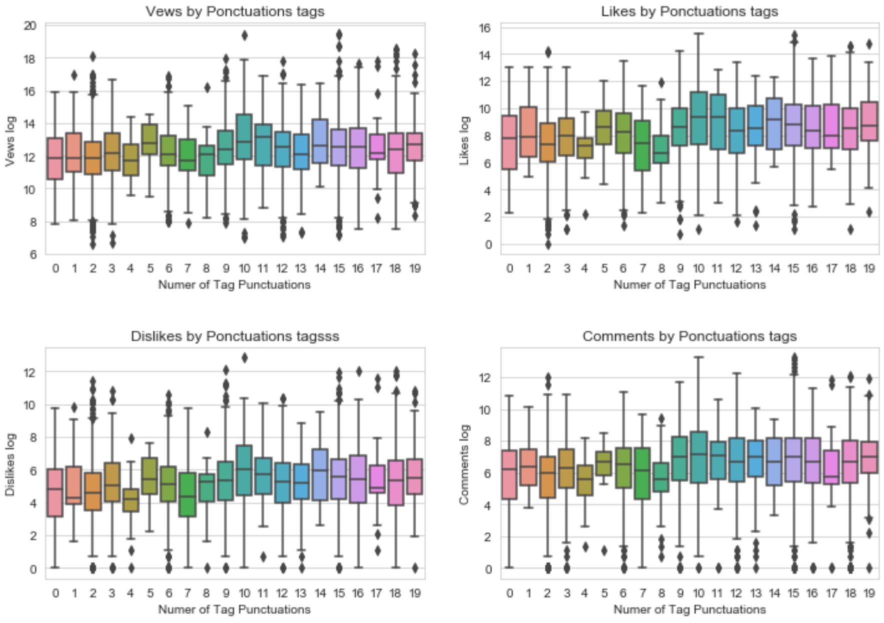

MEAN WORD LEN DISTRIBUITION

To start the analysis on the tag and other word information, we get several distribution graphs.

4.2 The Relationship between Information from Tag & Description and The Number of Likes, Dislikes, Views & Comments

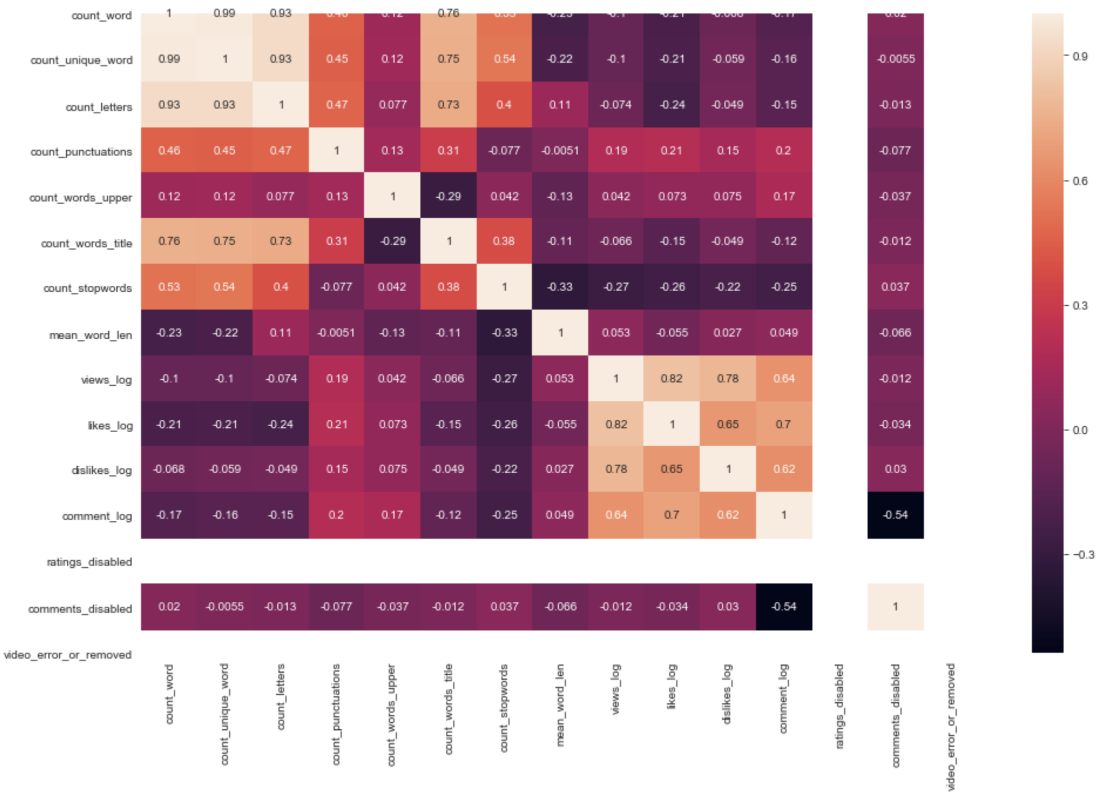


4.3 The Statstics of Puncuation Value



From box charts, parameters don’t show obvious relationship with number of tag punctuations, and the distribution are slightly changed with number of tag punctuations change.

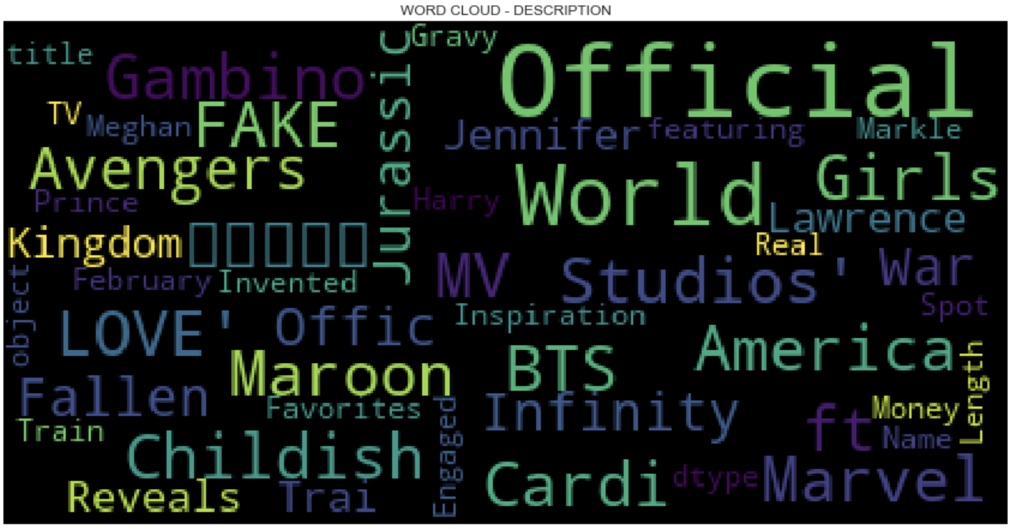
4.4 The Correlation between the Text and Number of Views



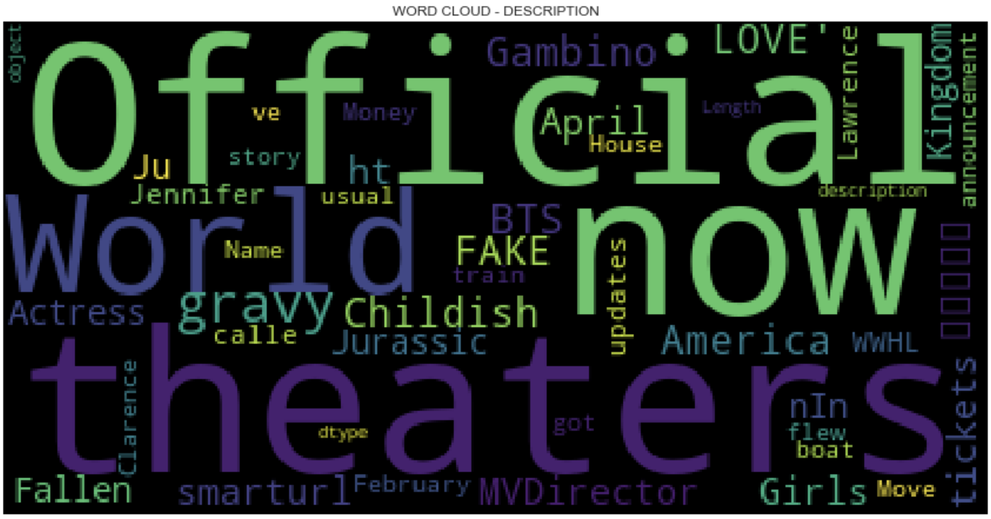
Generally speaking, most of the correlation between text and numbers of views shows dark color, which means it is weak

.

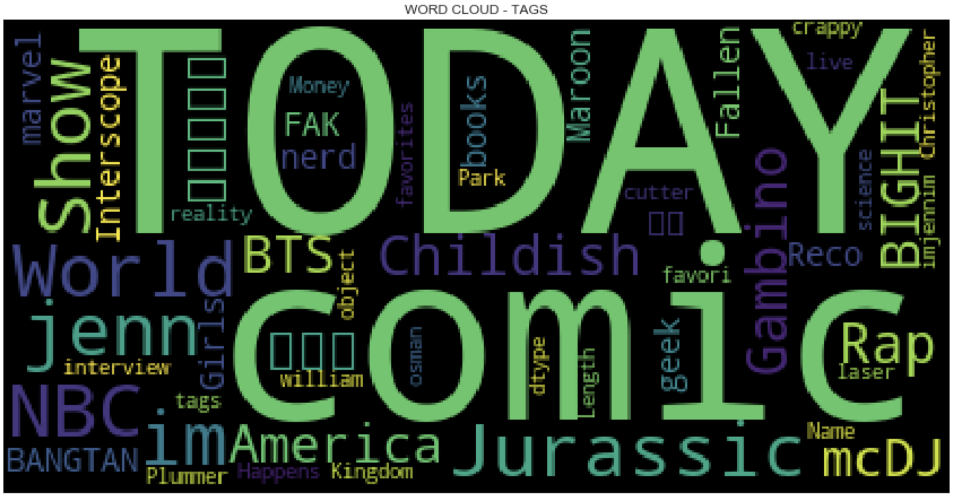
4.5 The Visualization of Titles, Tags and Descriptions by Wordcloud



For titles, the popular words are official, world, girls, America, childish, maroon, fake, avengers, Cardi, studios, love and so on.



For description, the popular words are official, now, world, theaters, gravy and so on.

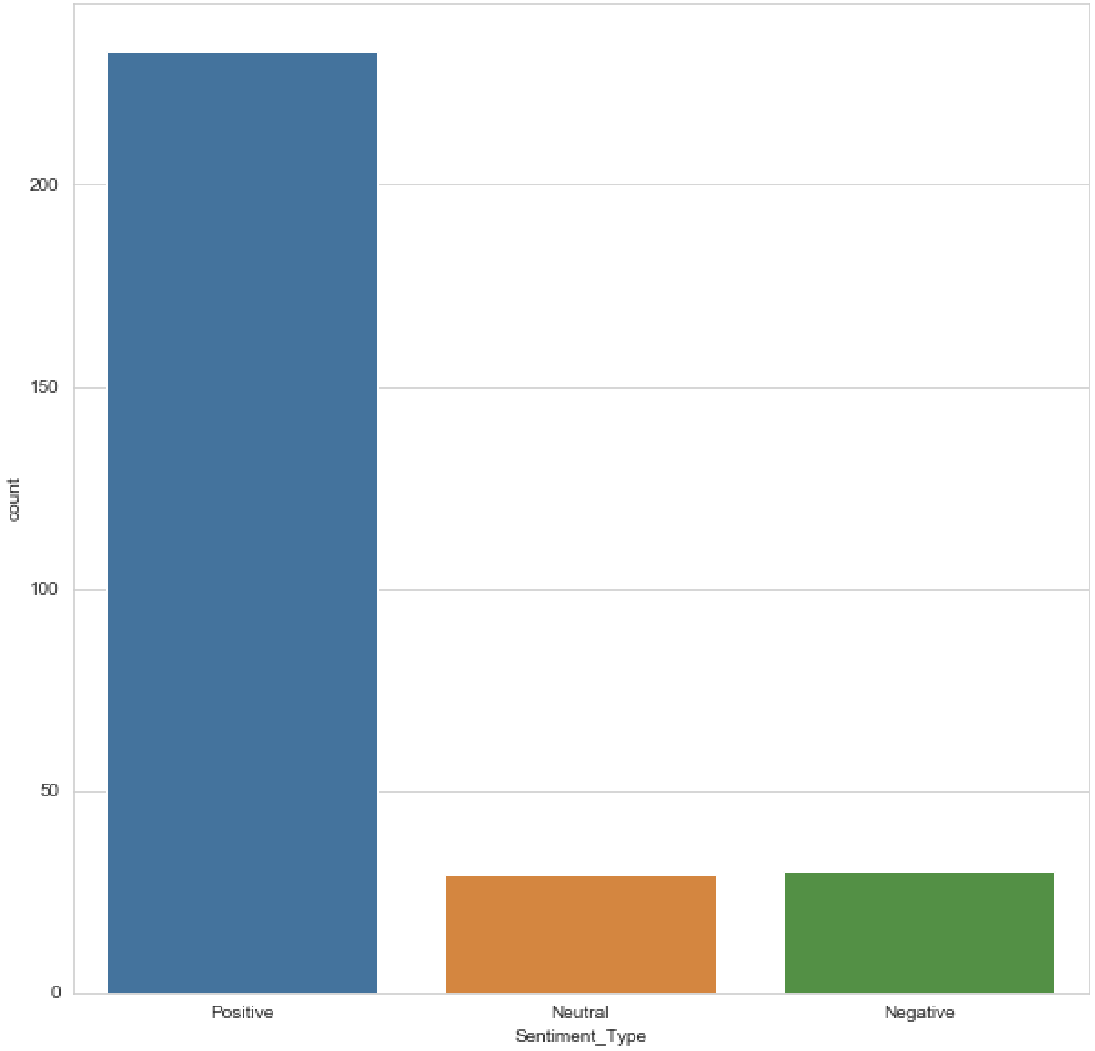


For tags, the popular words are today, comic, Jurassic, world, NBC and so on.

By comparing, we can find some words are high repeated, like: official, world, America and so on…

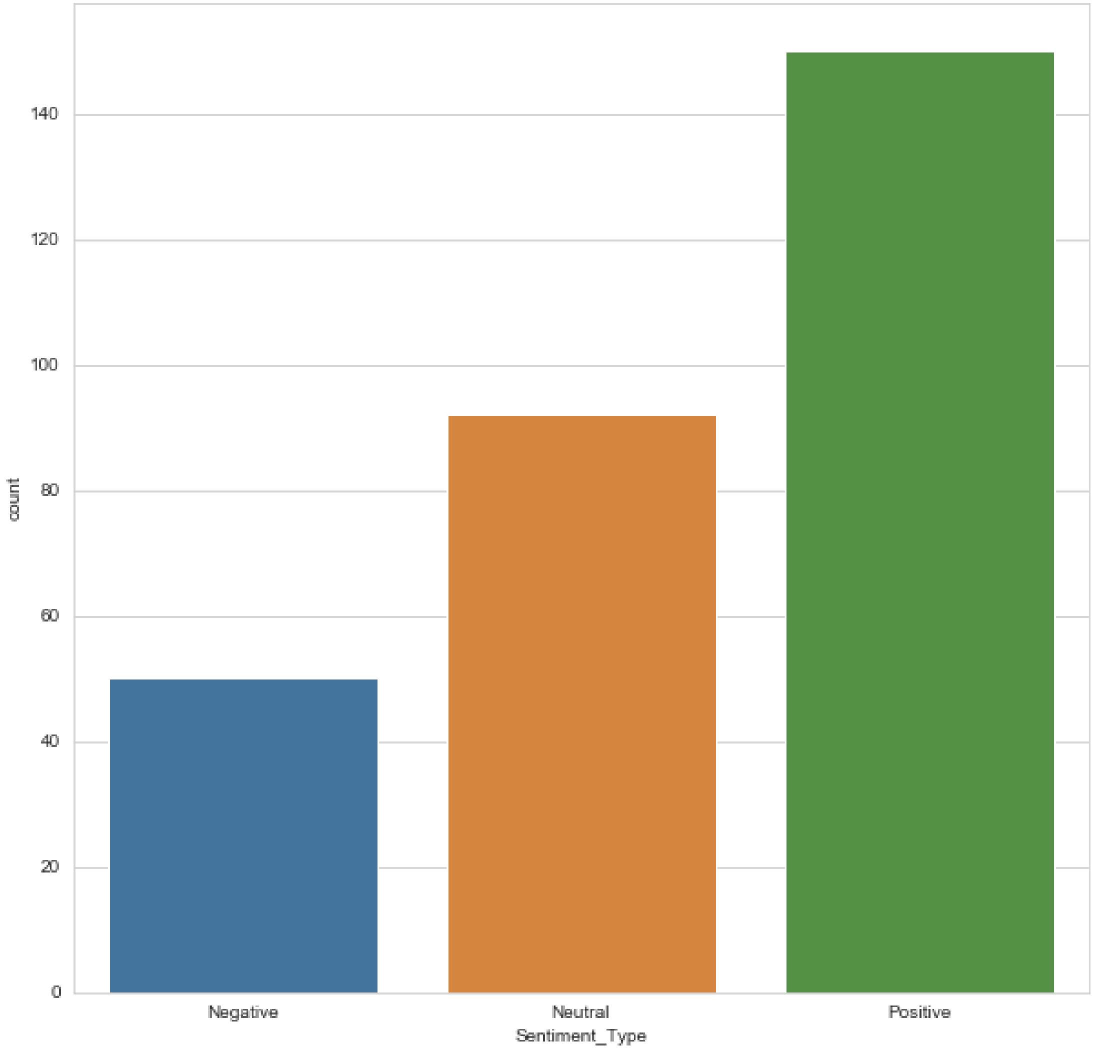
4.6 Sentiment Analysis of Titles, Tags and Descriptions by TextBlob

Sentiment Analysis of Descriptions



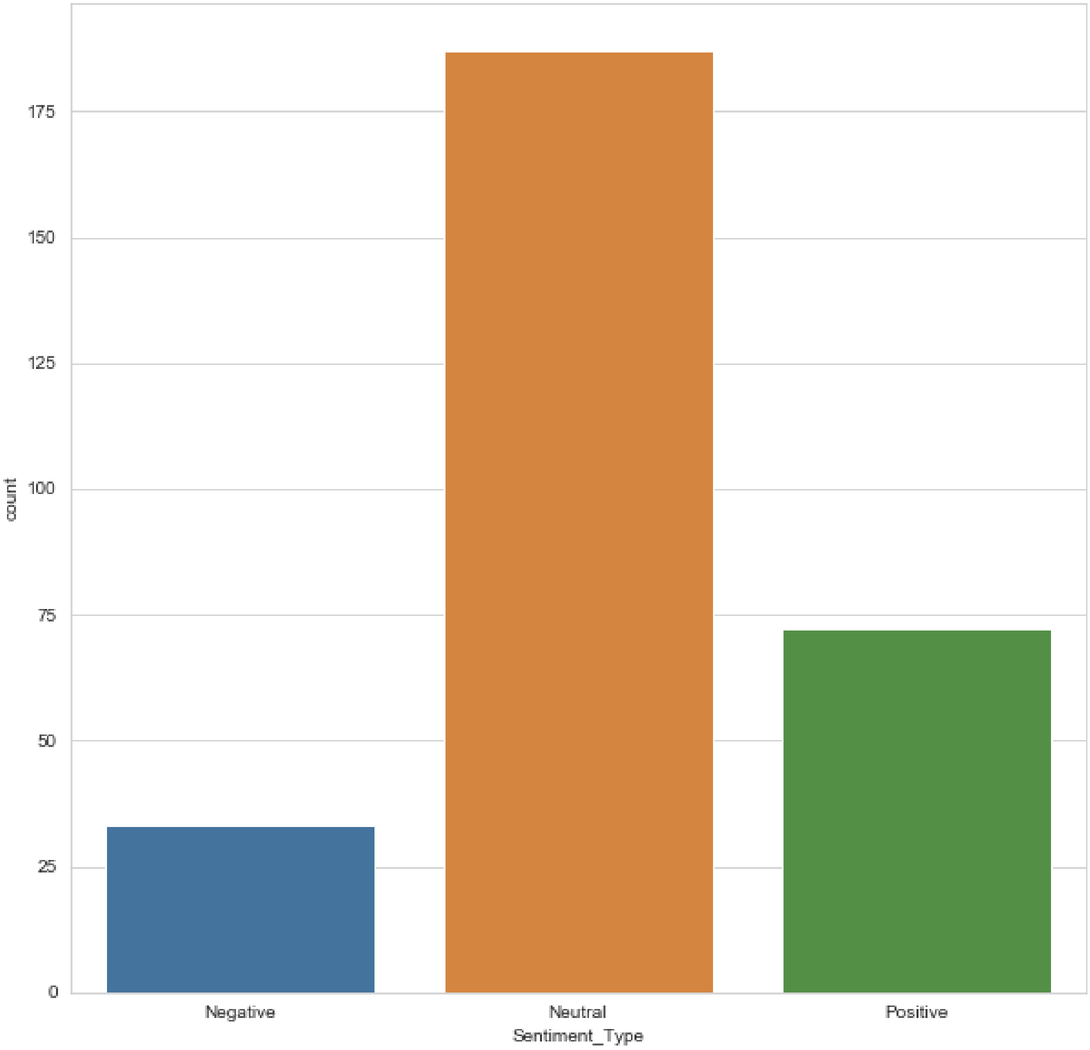
The positive sentiment counts the most in the descriptions.

Sentiment Analysis of Tags



The positive sentiment takes the highest parts of tags.

Sentiment Analysis of Titles

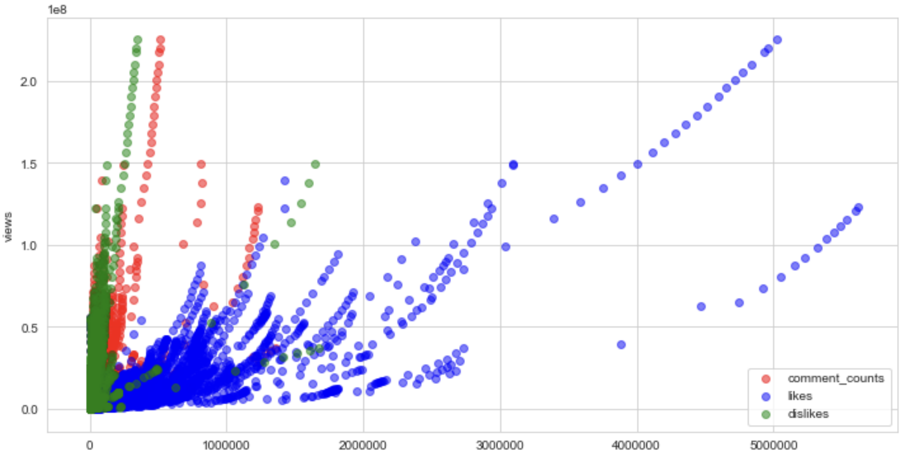


The neutral sentiment type is highest for titles.

5 Algorithm Models and Evaluation

5.1 Data Preprocessing and Plot

Before modeling, we need to do data preparation including reconstruction of dataset, cleaning duplicated data, transfer data type and so on. Then a plot for overviewing is given:



From the scatter plot, we can conclude the comment counts parameter is partly consistent with likes and dislikes

5.2 Predicting views

Owing to get highly performance prediction and outstanding accuracy and based on the observation of quantities of visualization and plots, we believe polynomial regression, linear regression, random forest could fit this data well. Firstly, we use Views as response.

5.2.1Train the Model by Polynomial Regression

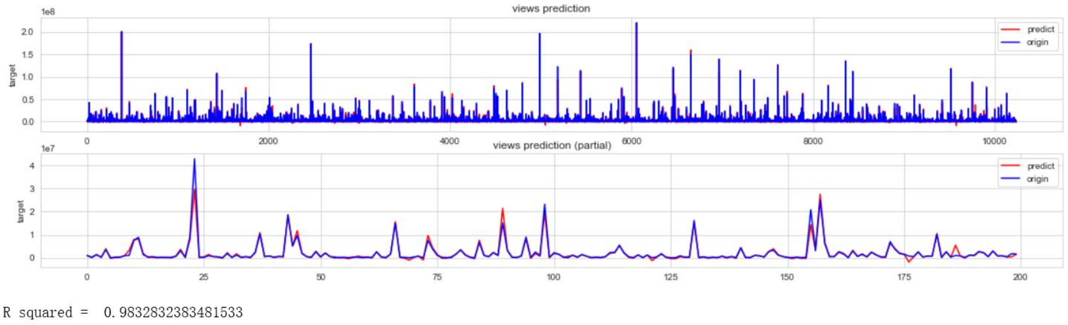
Polynomial Regression, as a type of linear regression, has the independent variable x and dependent variable y with nth degree polynomial relation. It is expressed as E (y |x), meaning a kind of nonlinear relationship between x and the corresponding conditional mean of y.

Vedio_id, trending_date, pulish_date, publish_time and publish_hour are selected as X and Y is views. Then we separate training data set and test data set and applied polynomial regression model to the training data.

5.2.2 Prediction with R^2 Score and Plot

We make prediction by polynomial regression model and show the prediction situation by the plots below, from the curve comparation of prediction and original data, we find most parts of these curves fit well, which means the prediction accuracy is high.

Due to verify our opinion by the specific value, we use R^2 score to check the accuracy rate. In statistics, the coefficient of determination, denoted R^2 or r^2 and pronounced "R squared", is the proportion of the variance in the dependent variable that is predictable from the independent variable (s). It is a statistic used in the context of statistical models whose main purpose is either the prediction of future outcomes or the testing of hypotheses, on the basis of other related information. It provides a measure of how well observed outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model.



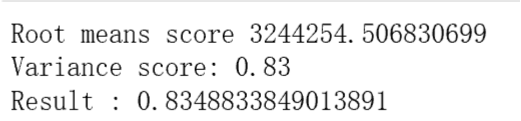
```
R squared =  0.9832832383481533
```

R^2 value is 0.9833, which proves that the prediction result is excellent and the model is very successful

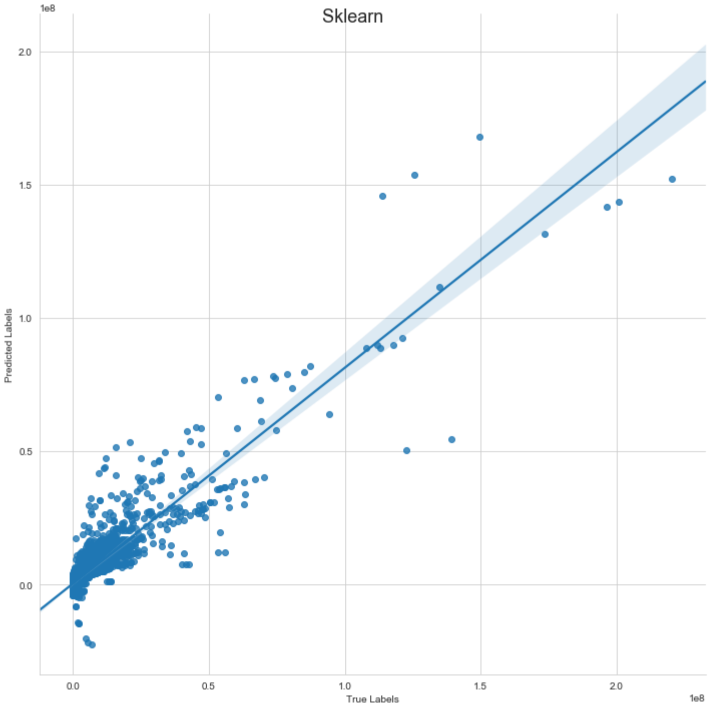
5.2.3 Train the Model by Linear Regression

In statistics, linear regression is a linear approach to modeling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regression.

We use linear regression to apply the train data and predict our test set. Then we get its accuracy score and prediction figure.



```
Root means score 3244254.506830699
Variance score: 0.83
Result : 0.8348833849013891
```

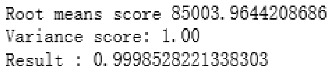


The prediction result is quite good with 0.83 accuracy.

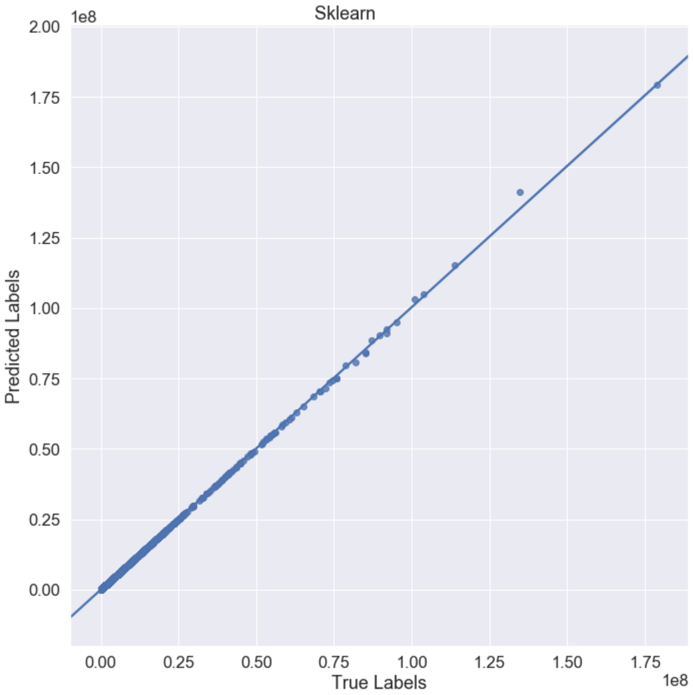
5.2.4 Train the Model by Random Forest

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.[Random decision forests correct for decision trees' habit of overfitting to their training set.

We use random forest to apply the train data and predict our test set. Then we get its accuracy score and prediction figure.



```
Root means score 85003.9644208686
Variance score: 1.00
Result : 0.9998528221338303
```



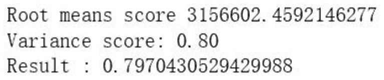
The prediction result is outstanding with very high accuracy close to 1.

5.3 Predicting Likes

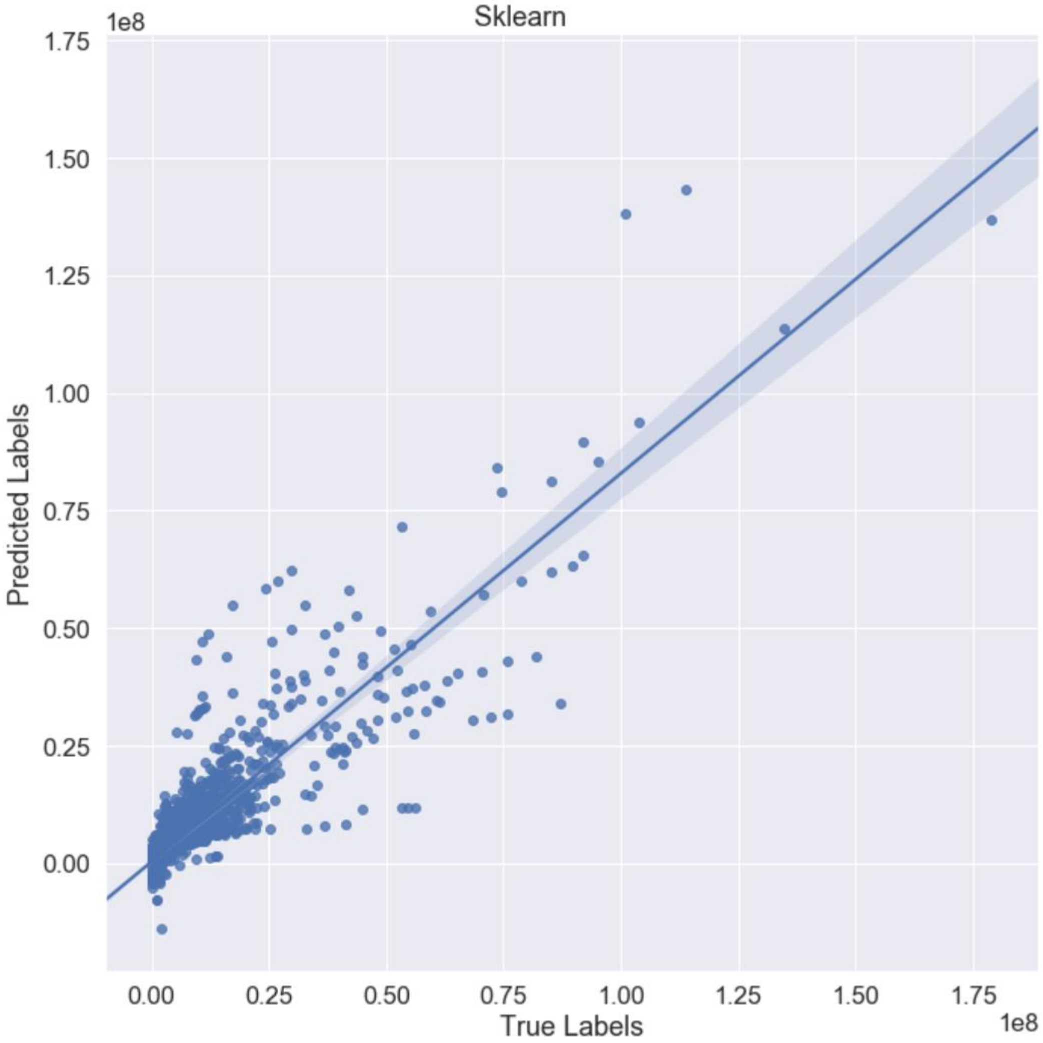
We take parameter Likes as target to build model by linear regression and random forest.

5.3.1 Linear Regression

We take Views as Y and separate data by test size= 0.2. Then we train the linear regression model and get the following prediction results:



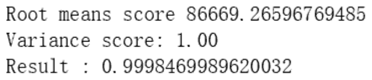
```
Root means score 3156602.4592146277
Variance score: 0.80
Result : 0.7970430529429988
```



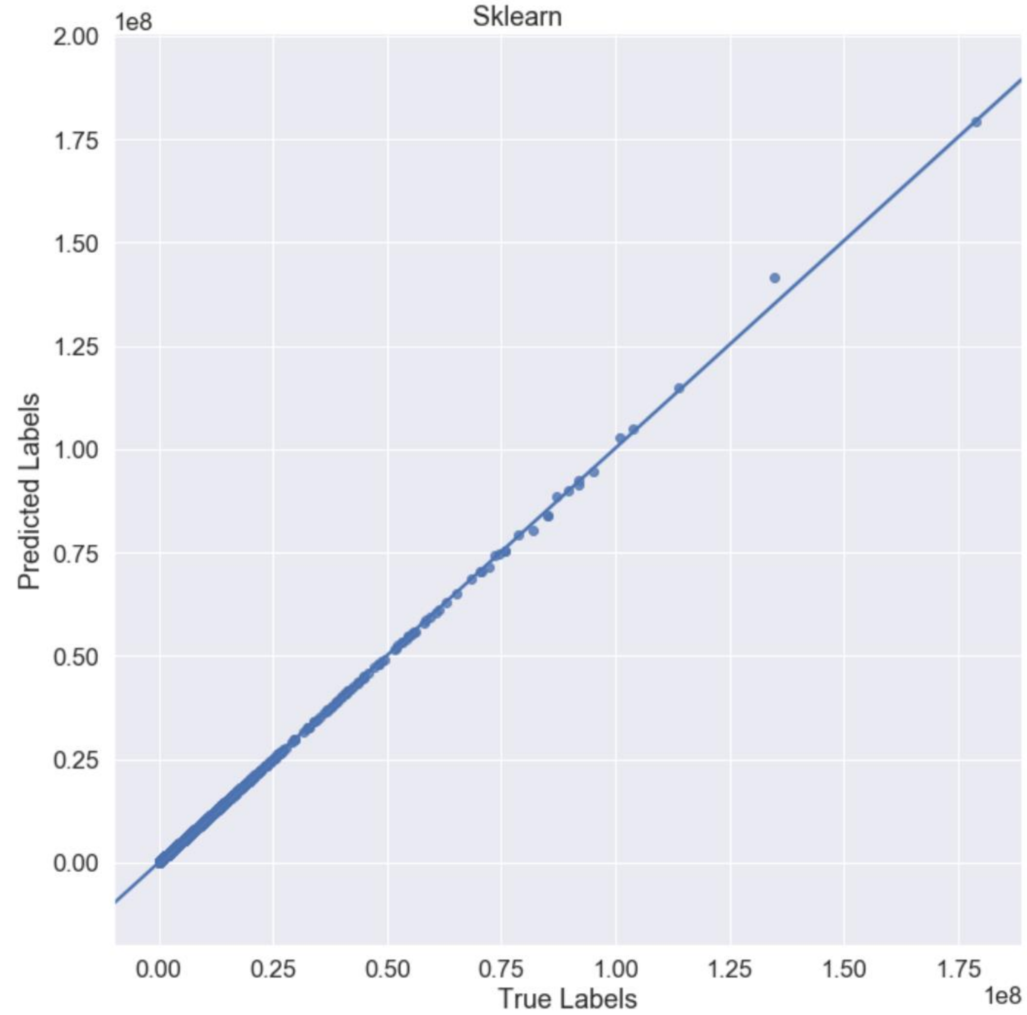
The variance score is 0.80 and this model fit well.

5.3.2 Random Forest

We apply random forest and take Likes as target. The prediction outcome is here:



```
Root means score 86669.26596769485
Variance score: 1.00
Result : 0.9998469989620032
```



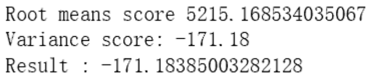
The prediction shows random forest works very excellently on this model with very high variance score.

5.4 Predicting Number of Comments

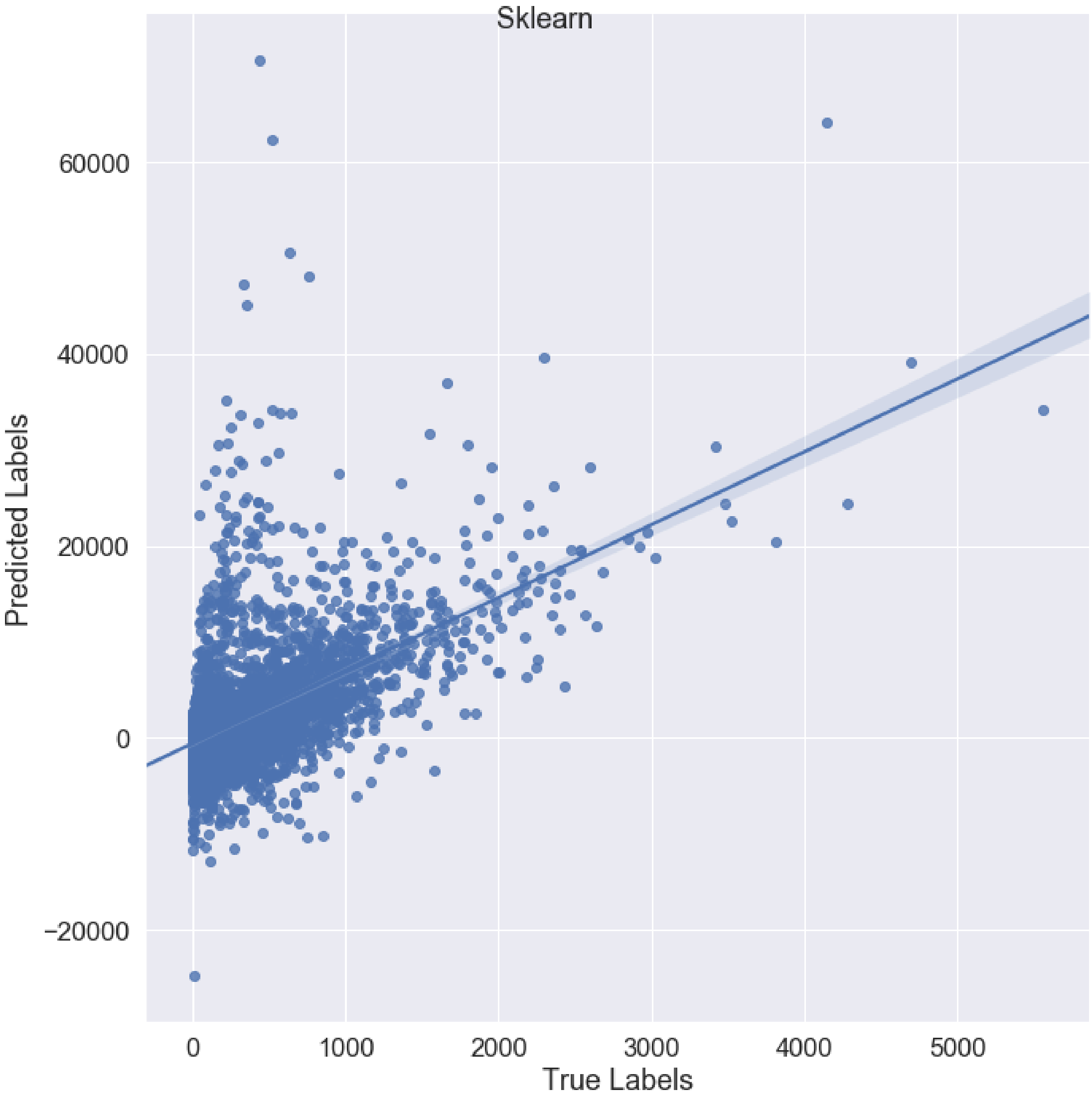
We take parameter Comments as target to build model by linear regression and random forest.

5.3.1 Linear Regression

We take comment_count as Y and separate data by test size= 0.2. Then we train the linear regression model and get the following prediction results:



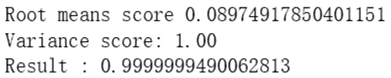
```
Root means score 5215.168534035067
Variance score: -171.18
Result : -171.18385003282128
```



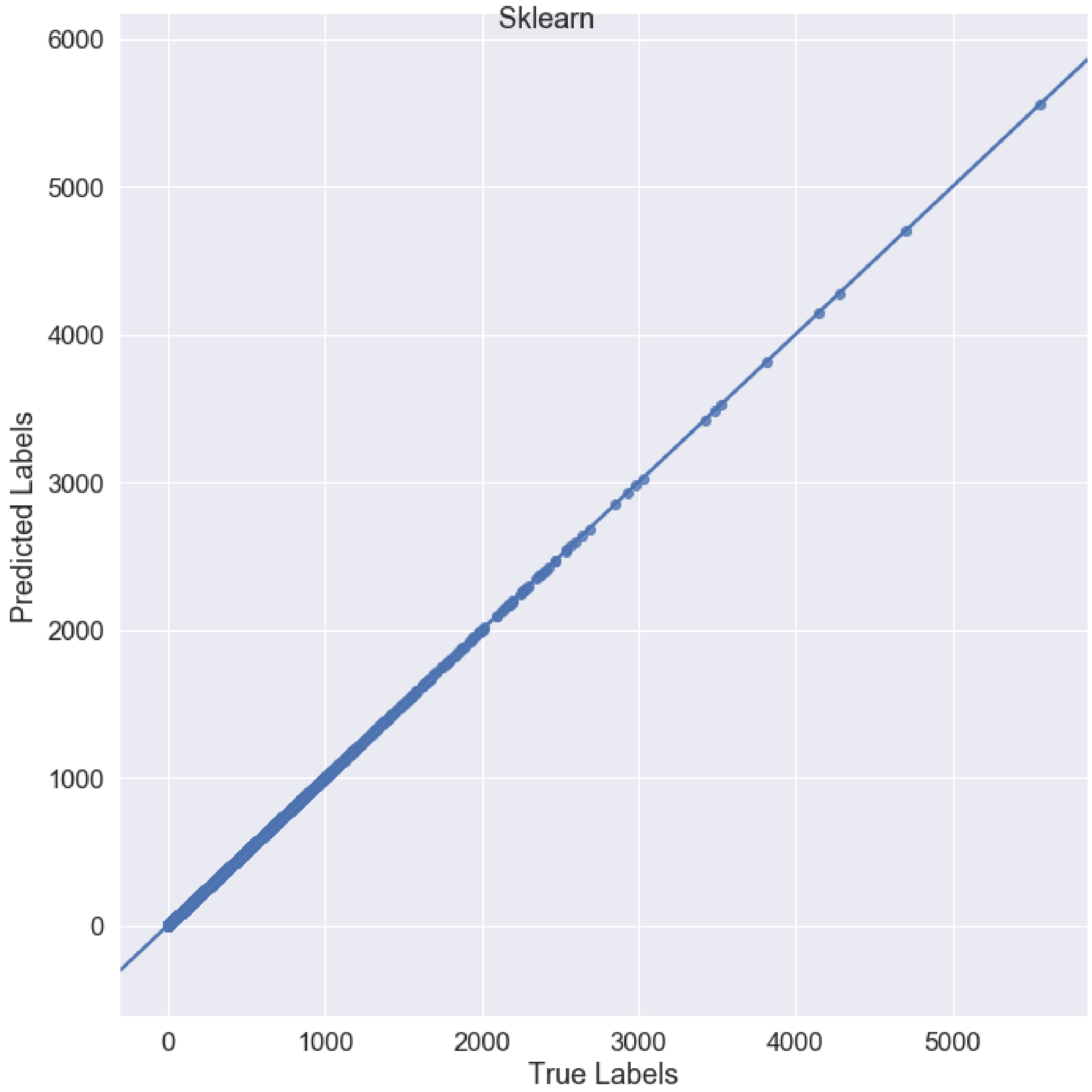
This model doesn't work well, and prediction result is not ideal.

5.3.2 Random Forest

We apply random forest and take comment_count as target. The prediction outcome is here:



```
Root means score 0.08974917850401151
Variance score: 1.00
Result : 0.9999999490062813
```

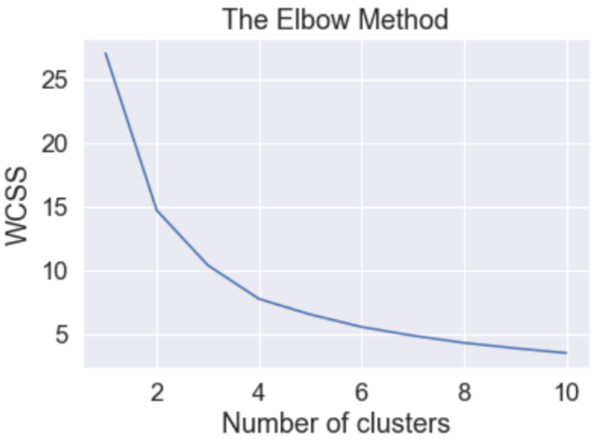


The prediction shows random forest works very excellently on this model with very high variance score.

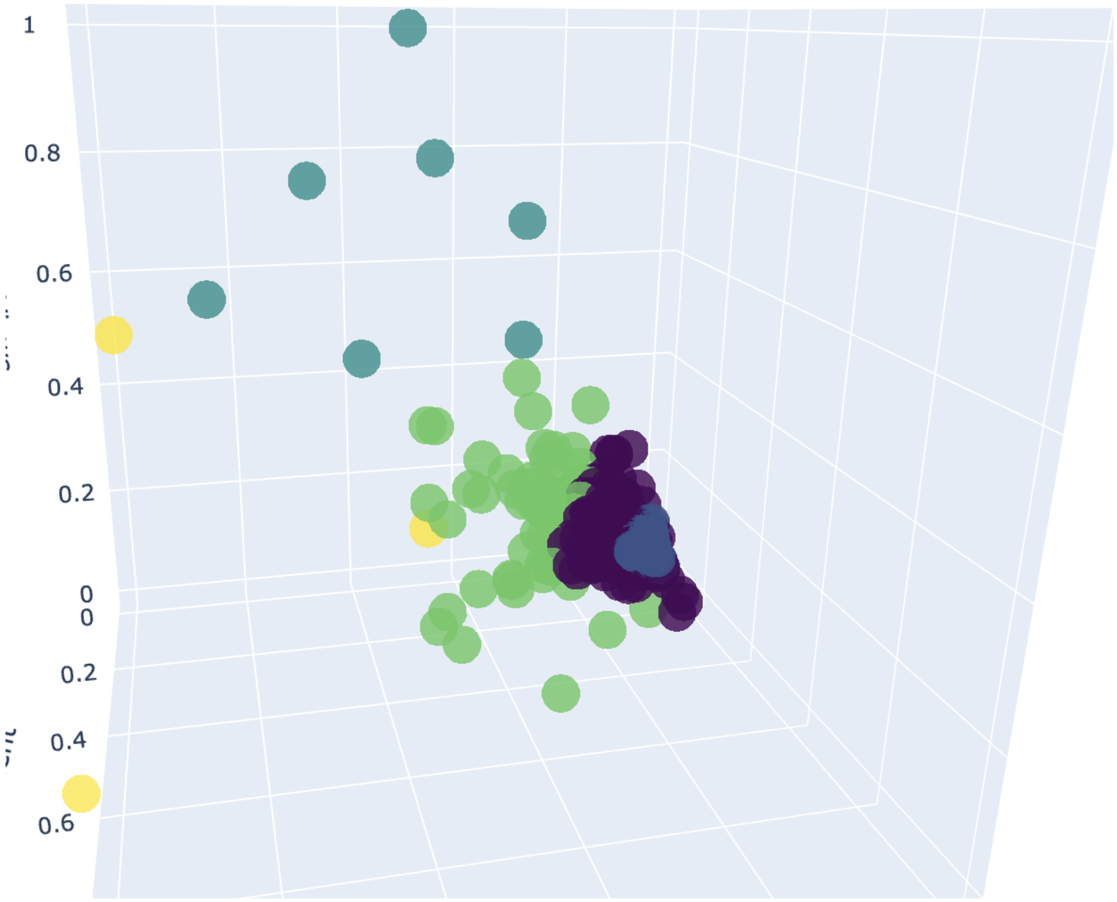
5.5 Classification by K-Means Clustering

k-means clustering is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells. It is popular for cluster analysis in data mining. We decide use K-means clustering as classification model.

The rating rate after normalization:



We set 5 cluster to do the classification and get the final 3d result (more effects shown in the ipynb file).



6 Recommendation and Conclusions

6.1 Recommendation

Through a large amount of data visualization and analysis comparison, we get large quantities of effective information and advice on how to catch the people's preferences to watch and like videos.

We get some conclusion that the publish time around 16 is easier to get audiences' attention, the top channel is ESPN, The Late Show with Stephen Colbert and so on, the popular words in titles are Episode, New, Full, Official, Game, Trump and so on, the popular categories are entertainment, new & politics, people & blogs, music and so on…

In addition, the content of this report is not limited to what are mentioned above. Our project also contains many other useful information and suggestions, hoping to help, who have interest in this topic.

6.2 Conclusion

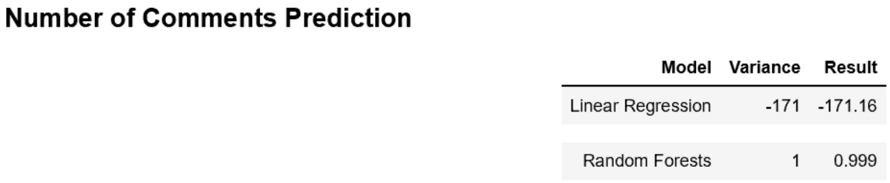
We take Views, Likes, Number of Comments as target separately and build model by Linear Regression and Random Forests. The following is the accuracy report for these models.

Views Predicition

Model	Variance	Result
Linear Regression	0.83	0.834
Random Forests	1	0.994

Likes Predicition

Model	Variance	Result
Linear Regression	0.8	0.799
Random Forests	1	0.999



Number of Comments Prediction

Model	Variance	Result
Linear Regression	-171	-171.16
Random Forests	1	0.999

By comparing the prediction variance score, we can conclude Random Forest is the best model to fit this YouTube dataset for prediction, which has accuracy around 0.99.

7 Reference and Appendix

7.1 Reference

The dataset link from kaggle:
https://www.kaggle.com/datasnaek/youtube-new

Concept introduction form Wikipedia:
https://en.wikipedia.org/wiki/Coefficient_of_determination
https://en.wikipedia.org/wiki/Polynomial_regression
https://en.wikipedia.org/wiki/Linear_regression
https://en.wikipedia.org/wiki/Random_forest
https://en.wikipedia.org/wiki/Random_forest

7.2 Appendix

More analysis content and detailed python code is shown in the attached ipynb file.