**Youtube trending video analysis**

|  |  |
| --- | --- |
| **Team member** | **SID** |
| LIN Chen | 1155147259 |
| LIN Ting-yuan | 1155148842 |
| Au-Yeung Tsing Anson | 1155148839 |
| LIU Zhengyu | 1155148850 |
| LIU Zengyu | 1155151541 |

Contents

1. Introduction about the project

1.1 Motivation

Our project is about Youtube Trending Videos Analysis ,using data collected by Youtube API. Exploratory data analysis and Regression method will be conducted to better understand the video patterns and get some insights, which may be helpful for those who want to gain popularity of their videos on Youtube.

1.2 Goals

Analyse what factors affect how popular a YouTube video will be using EDA. The information we will try to find is like:

* + How many views do our trending videos have? Do most of them have a large number of views? Is having a large number of views required for a video to become trending?
  + What are the most common words in trending video titles?
  + …
* Predicting likes using Regression model

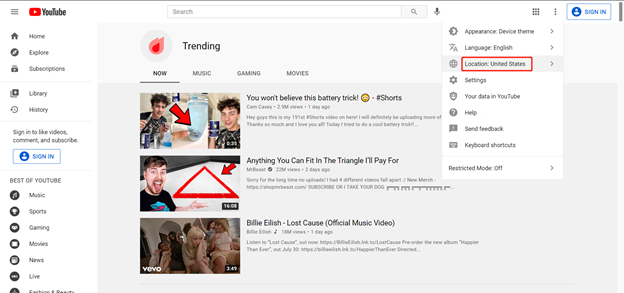
1.3 Dataset

This dataset focus solely on US region, including several months (and counting) of data on daily trending YouTube videos. We collect the data using API and combine these data with the data from kaggle, since there are only 200 pieces of data we can collect using API and we do not have enough time to get a large number of data solely depending on Youtube API.

Variables include the video title, channel title, category\_id, publish time, tags, views, likes and dislikes, description, and comment count etc.

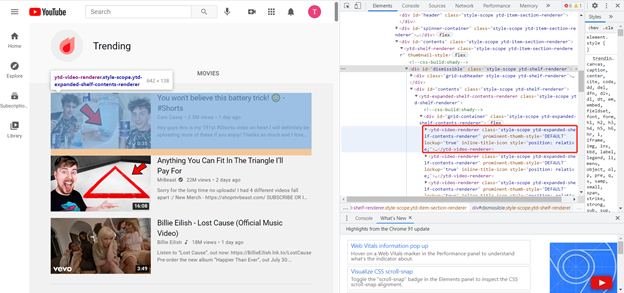
2. Webpage Analysis and Data Scraping

As stated above, we will be gathering and analyzing YouTube trending videos in the US region. The link of YouTube trending video is provided: <https://www.youtube.com/feed/trending>. This project requires users to set the YouTube location to the United States option beforehand.



2.1 Webpage Analysis

A quick glance at the trending video webpage is shown in the follow. After analyzing the webpage, we believe we could easily extract key data features of the trending videos from the html structure. We first looked into the div tag with attribute id equals “grid-container.” Then we ran several iterations in obtaining the trending video one by one. Ideally, the for-loop we ran will extract desirable data of each video till the very last trending video of the webpage.



This approach, however, has some obvious short-comings:

First, the webpage only shows the top 50 trending videos. In data analytics field, 50 record of data provide very limited information. Even though YouTube updates the trending list every 15 minutes, the fact that most of the trending videos could appear again and again for a long time because of the nature of visitors’ behavior. In other words, the trending video list from now could be almost identical to the one from an hour ago.

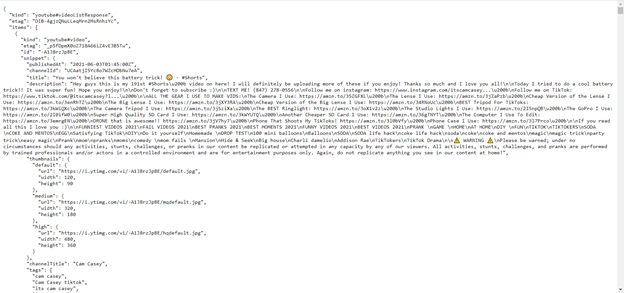
Secondly, since our goal was to merge extracted data with our original dataset from Kaggle, we found that several features is not directly included in the webpage. That means it could be less intuitive to extract data that completely match the features of the original dataset. Therefore, simply scraping data of the webpage will cause data inconsistency and is indeed a more time-consuming and inefficient way to gather data.

In summary, we believe the data from 50 trending videos are not enough for future analysis. Also, the approach is quite cumbersome in terms of gathering necessary features when implementing. Therefore, we decided to look for the API that supports trending video scraping on YouTube.

2.2 Data via YouTube API and Data Scraping

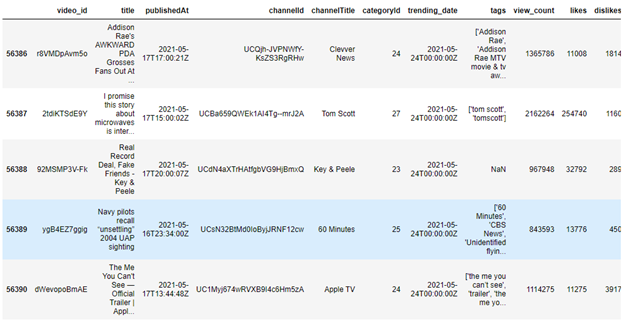
Given the short-comings of our first approach, we decided to use YouTube API for data collection. The Youtube Data API information can be obtained in the follow: <https://developers.google.com/youtube/v3/docs/videos/list>.

In this approach, we will download the trending video data via API instead of directly from a webpage. We could retrieve the data of current trending videos by HTTP request concatenate with our own api key which we previously applied from Youtube and with certain parameters added. The Youtube API request will then return a json format of information that includes all our desirable data. A snapshot of the json webpage is attached follow.



One thing that is worth mentioning in our project is that we believe API Key should always be kept confidential. That means people without permission of the API Key owner should not be able to directly observe API Key from the HTTP request url. To address this confidential issue, we create a separate text file that stores API Key. Other users may need to provide and type in their own API Key in the appointed text file before utilizing the code we provided.

With this API data gathering approach, we successfully retrieve 200 trending videos per day. It may be true that 200 record of data is considered trivial still, but we believe we could obtain the data of daily trending videos by executing our python code every day. Due to the time limitation in this project, we will only extract data of trending videos on May 24th and combine the 200 data record with original Kaggle dataset. The sample combined dataset is shown below.



3. Exploratory Data Analysis

3.1 Data preparation

* Data cleaning: replace null value in description column with ""
* Discription of data

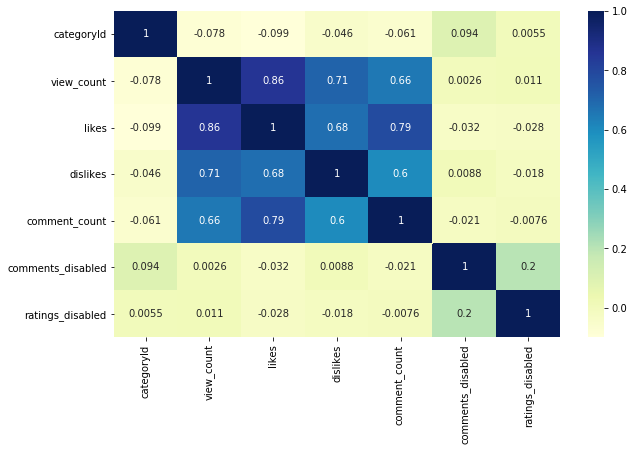
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | categoryId | view\_count | likes | dislikes | comment\_count |
| count | 56191 | 56191 | 56191 | 56191 | 56191 |
| mean | 18.62001 | 2.77E+06 | 1.52E+05 | 3419.742 | 1.57E+04 |
| std | 7.045633 | 6.40E+06 | 4.25E+05 | 15322.52 | 1.09E+05 |
| min | 1 | 0 | 0 | 0 | 0 |
| 25% | 10 | 5.51E+05 | 2.10E+04 | 394 | 1.91E+03 |
| 50% | 20 | 1.14E+06 | 5.32E+04 | 925 | 4.37E+03 |
| 75% | 24 | 2.55E+06 | 1.33E+05 | 2517 | 1.05E+04 |
| max | 29 | 2.33E+08 | 1.57E+07 | 879354 | 6.07E+06 |

We can see that the data of view\_count,likes and dislikes is very dispersed since the std is very large

Since the phenomenon of manufacturing statistics is very common and it is also one way to get more exposure for many video creators, we do not remove data that looks unusual(e.g. more likes than video views)

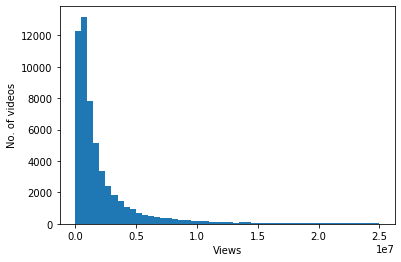
3.2 Analysis of numerical and categorical variables

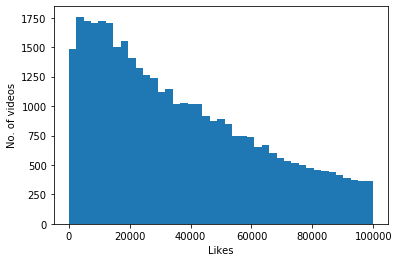
* The correlation between numerical variables

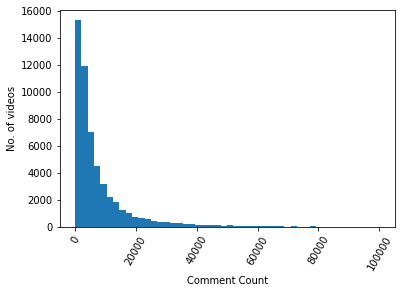


We can see that view\_count,like, dislikes and comment count have strong positive correlation, the coefficients are all above 0.6. While others simply does not correlate with each other.

* Distributions of view count, likes,and comment count

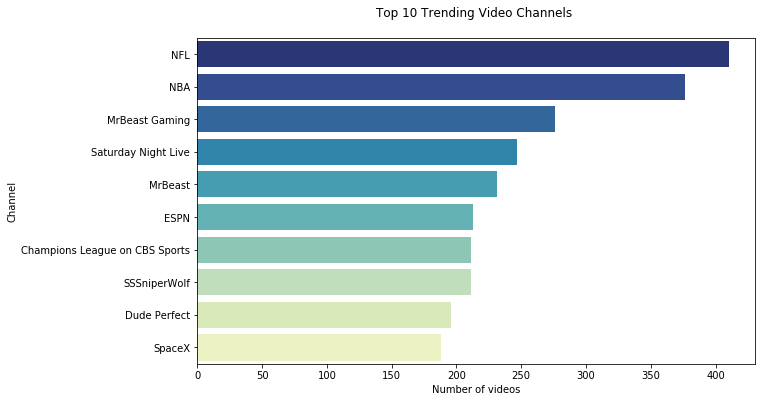




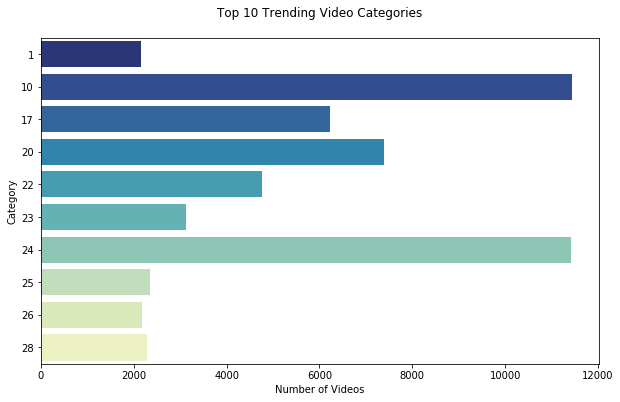


We use histograms and calculations to check the distributions of view count, likes and comment count and we find that around 60% of videos have views less than 1500000, 68% of videos have likes less than 100000 and as for comment count is, 73% of videos have comment no more than 10000.

* Video channel and category

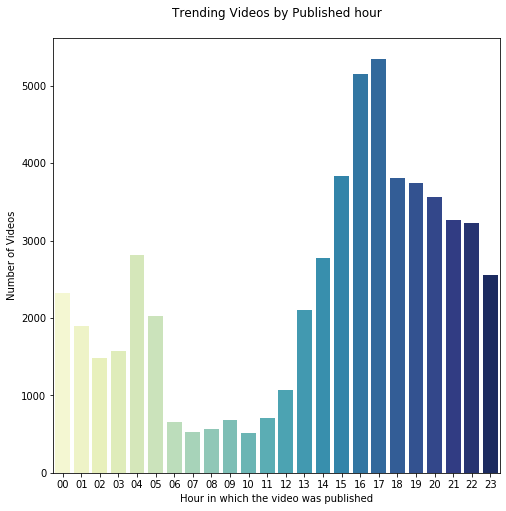


NFL rank top1, having 400 trending videos. And second follows NBA.

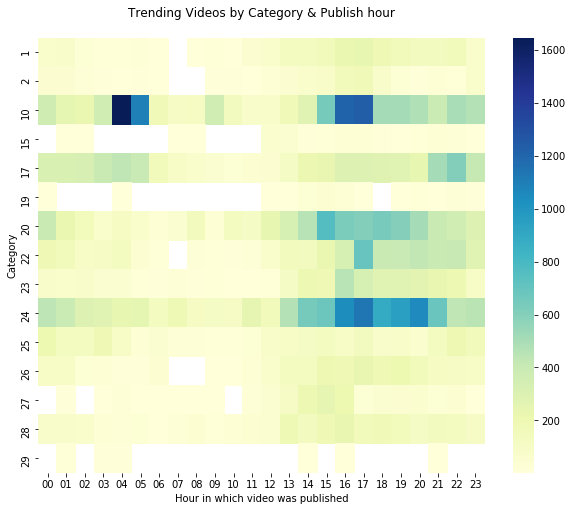


The top 3 popular video category's ids are 24,10 and 20 which are Entertainment, Music, and Gaming respectively.

* Pulished time



We can see that most videos were published between 15:00 and 22:00, and in contrast, very little number of videos were published between 6:00 and 11:00.

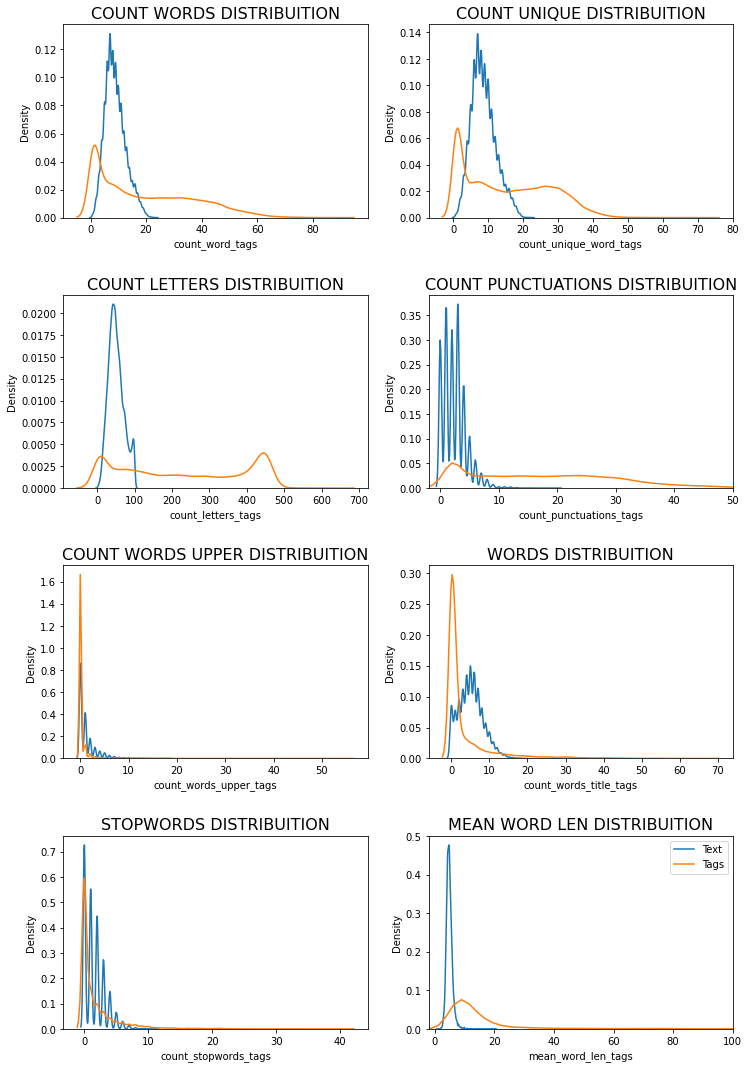


And combined with the video categoryID, we can see that the most popular video category are highly correlated with the most popular published time.

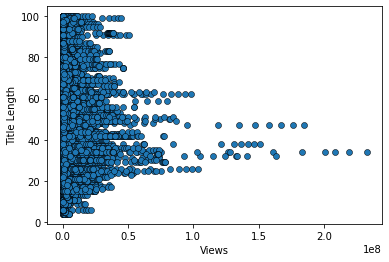
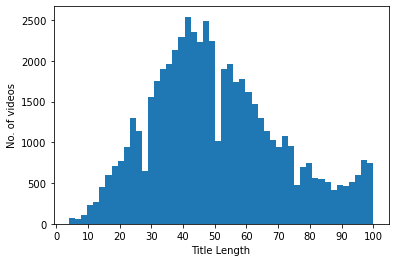
3.3 Analysis of text data

In order to study how do Youtube videos get into top trending videos, we also analyze the text data of each video, which is the title, description, and tags. We think that eye-catching titles and tags are the keys to attract viewers which are the crucial factors to help a video drive more views.

In this study, we made use of 'nltk' package, which is the Natural Language Toolkit that allows us to more text processing, to analyze the number of words, number of unique words, number of letters, letter cases, punctuations, stopwords, and the average length of words for both video titles and tags.



Base on the above outputs, we could conclude some key features. A trending video tends to have a title with around 10 words and the words in the title are tends to be unique. This means that those trending videos are probably on different topics or have some different and unique contents. Although we could not see any trends in the punctuations or letter cases, the word length tends to be short. For tags, it is usually a single word with around 10 letters, and which could also be meaningless.



After that, we try to see the relationship between the title length and views. Most trending videos have their title around 40 letters while videos with over 233 million kept their title with around 36 letters.







Furthermore, we try to highlight the words that highly used by word cloud. For titles, there are several with similar ranks, like 'GIRLFRIEND', 'ASKED', 'Apex' and 'Legends'. This reveals the contents of the trending videos is related to relationships and gaming which somehow echoing with our previous finding. Thus, video creators could always consider making their titles as direct as possible and they could also consider making content related to these themes.

For description, 'Subscribe' are the most frequently used word. This shows that those video creators usually ask their viewers to subscribe which is not surprising. According to Youtube policy, how quickly the video is generating views, as known as 'Temperature' is one of the signals that determines if that video could rank on trending. Once subscribers are notified that the channel they subscribed has updated a new video, they will tend to click and watch the video immediately. Thus this somehow proves that video creators are trying to make use of this factor to get on trending.

For tags, 'funny' ranked highest. This is a completely random word compare with those others in our study, revealing that tags could be unrelated to the content of the video except that it is a keyword that people search most.

4. Predicting Likes

Our project would like to explore the correlation of the channel and the like amount to find which channel is more helpful to obtain more like so that an entry-level youtuber could properly modify their topic selection.

Each video will be classified into a channel, and there are 15 channels available in the dataset. Here the project constructs a series of binary variables to demonstrate the channel that the video belongs to.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Channel1 | Ch2 | Ch3 | Ch4 | … | Ch15 |
| Video1 | 1 | 0 | 0 | 0 |  | 0 |
| Video2 | 0 | 1 | 0 | 0 |  | 0 |
| Video3 | 0 | 0 | 0 | 1 |  | 0 |

If the value equals to 1, the video belongs to that channel. For example, video 1 is belongs to channel 1.

In the previous part of the project, we conclude that there are several variables are highly correlated to the likes. So, we should select them as the control variables to ensure the validity of the model. The complete regression model is as below:

The result of the regression model indicates that Ch1, Ch2, Ch15, Ch17, Ch19, Ch25, Ch26, Ch28 play a negative effect on the like number. So, the remaining channels are recommended.

The R square of the model is 0.826 which is relatively acceptable. However, the current model of predicting likes still has many shortcomings. In the real world, the like number of a YouTube video is affected by countless factors both within the YouTube and outside the website. The number of variables in the current model is definitively not enough. So, the effectiveness of this model is not very satisfactory. The MSE is surprisingly large which is 29,672,739,891 and the RMSE is 172257. If more variables, such as the follower number of the youtuber; the google searching trend, were introduced from different information sources, the model would be more practicable.

5. Summary

We used 56191 data to explore Youtube trending videos and here are some interesting findings we can conclude from the above analysis:

* the most popular published time is between 3pm-10pm
* a unique title with short length which is no more than 10 words tends to attract more views
* Contents related to relationships and gaming are more likely to be popular

The project has some limitations when building predicting models. Maybe more variables should be included in further analysis.