# final_report

Hantang Li

4/20/2022

## Introduction

YouTube, the world's third most popular online destination, has transformed from a video-sharing site into a job opportunity for content creators in both new and mainstream media. Holland (2016) Individuals who upload videos on Youtube, also known as YouTubers, could turn on monetization features. One of the major ways YouTubers earn money is through the number of ad views. For detailed information, you can visit the following youtube's official website: (link: How to earn money on YouTube).

**Since ad views directly depend on each video's views, we would like to analyze whether YouTube is still popular, what factors could result in a high view and how people's preferences have changed in recent years.**

We will use past Canadian area's youtube daily trending video datasets found online to answer this question. The largest past youtube trending video data set we found online is from Kaggle. The dataset contains detailed daily trending video information collected using YouTube Data API v3 ranging from 2017-12-01 to 2018-05-31. The download link is https://www.kaggle.com/rsrishav/youtube-trending-video-dataset.

For comparison, we found another Canadian area's youtube daily trending video dataset from Kaggle with a similar data format and was collected using YouTube Data API v3. The data ranges from 2020-08-12 to 2022-03-07. The download link is https://www.kaggle.com/rsrishav/youtube-trending-video-dataset.

Since both datasets do not contain detailed information about the video, such as the duration, I registered YouTube Data API v3 and used this API to retrieve that information. Since the second dataset contains a dataset that ranges around two years and could easily exceed the API's quota if I call API on all the videos, I decided to clip the second data to only include data ranges from 2021-08-01 to 2022-03-01. The reason is the clipped data's time interval in a year overlaps the first dataset the most, and now both datasets contain six months of data, so the annual activity's affection on people's preference could be reduced, and the affection of those annual activities is hard to analysis.

For reference, the YouTube Data API v3 documentation can be found here video data annotation:

For naming convention, since the first dataset started in 2017, I named it CA_2017, and for the second dataset ends in 2022, I named it CA_2022. As creators for each dataset used slightly different naming conventions, I listed a detailed explanation for all the column names for each dataset.

| **CA_2017** | |
|---|---|
| video_id | The ID that YouTube uses to uniquely identify the video. |
| trending_date | Date that the video is on trending |
| title | The video's title. |
| channel_title | the channel name that the video was uploaded to |
| category_id | " Varies between regions. To retrieve the categories for a specific video |

| CA_2017 | |
|---|---|
| publish_time | Time that the video is published |
| tags | The keyword tag suggested for the video. |
| views | The number of times the video has been viewed. |
| likes | The number of users who have indicated that they liked the video. |
| dislikes | The number of users who have indicated that they disliked the video. |
| comment_count | The number of comments for the video. |
| thumbnail_link | A link to the thumbnail images associated with the video. |
| comments_disabled | Whether the author allowed people to leave a comment. |
| ratings_disabled | Whether the author publicly displays a number of likes or dislikes. |
| video_error_or_removed | Whether the video can be viewed on a browser at the time data is recorded. |
| description | The video's description. The property value has a maximum length of 5000 bytes and may contain all valid UTF-8 characters except?. |

| CA_2022 | |
|---|---|
| video_id | The ID that YouTube uses to uniquely identify the video. |
| trending_date | Date that the video is on trending |
| title | The video's title. |
| channelTitle | The channel name that the video was uploaded to |
| categoryId | " Varies between regions. To retrieve the categories for a specific video, find it in the downloaded JSON file CA_category_id.json. |
| publishedAt | Time that the video is published |
| tags | The keyword tag suggested for the video. |
| view_count | The number of times the video has been viewed. |
| likes | The number of users who have indicated that they liked the video. |
| dislikes | The number of users who have indicated that they disliked the video. (note this feature has been removed from Youtube, so it will be zero in this dataset.) |
| comment_count | The number of comments for the video. |
| thumbnail_link | A link to the thumbnail images associated with the video. |
| comments_disabled | Whether the author allowed people to leave a comment. |
| ratings_disabled | Whether the author publicly display number of likes or dislikes. |
| channelId | The channel ID that the video was uploaded to |
| description | The video's description. The property value has a maximum length of 5000 bytes and may contain all valid UTF-8 characters except?. |

# Methods

## Download the data

- For CA_2017: Open (link 2017 YT trending) inside the browser, and download CAvideos.csv.
- For CA_2022: Open (link 2020 YT trending) inside the browser, and download CA_youtube_trending_data.csv.
- We observed that each dataset only contains category id s without showing the category id's corresponding name. We downloaded an additional JSON dictionary to match each category ID to their name. Since both datasets use the same dictionary file, I downloaded it from the website for the '2017 YT trending data. Open (link 2017 YT trending) inside the browser, and download CA_category_id.json.(I have included it in the repository)

## Clip the 2022 trending video data

As mentioned in the Introduction, I decided to clip the second data to only include data ranging from 2021-08-01 to 2022-03-01, which includes six months of data ranging from the end of a year to the beginning of a year. The reason is the clipped data's time interval in a year overlaps the first dataset the most, and now both datasets contain six months of data, so the annual activity's affection on people's preference could be reduced, and the affection of those annual activities is hard to analysis. We will use the "filter" method to remove all the CA_2022 data with trending_date larger than 2021-08-01 and smaller than 2022-03-01

## Obtain additional data using API

For each video, to obtain video duration, video dimension, and whether the video has a caption, we need to use YouTube Data API v3.

The API key is obtained through the Google cloud. The method we will use is video:list(link: video list method). This method is able to return a maximum of 50 videos in one call, so we first obtain a unique vector of all video ids for both data sets and then split the vector into a list of chunks. Each chunk contains 50 video ids. Then, we loop over the list of chunks and call API on each chunk of video ids. Note that the chunk of ids needs to be formulated as a comma-separated string. At last, we convert the API's result to a data frame and save it as a file called All_ca_trend_vid_content.csv, so we can load the data frame directly for future usage.

One problem that occurred is that We were unable to get the video detail for some videos due to the video becoming unavailable on youtube. One example is the video with the id CYl1YwAO-ew, video_link. By clicking the link and trying to watch the video on the browser, it will display that it is a "Private video." And for those kinds of videos, we cannot get detailed information through API. Since 5364 out of 33557 videos are currently unavailable, and most of those are from the CA_2017 dataset, we need to consider those missing values in the following analysis.

## Data Cleaning and Wrangling

- **Reformatting date and time using python script**: We observe CA_2017's trending_date is stored as characters. For example, 2017-11-14 is stored as character 17.14.11 we will need to convert it to POSIXct DateTime format. Since R is hard to perform string operations, it takes quite a while applying .POSIXct or as.Date on a data frame with a huge amount of rows. We use python to solve this problem. With the help of pandas, we can process it in a few seconds. The script is "insert_words.py." After processing the time format using the python script, we save the file as "CAvideos2.csv" and load it directly while running the R script.
- **Renaming column names**: We Renamed the CA_2022 dataset to have the same variable names as CA_2017, here is the modification:

| CA_2017 | CA_2022 |
|---|---|
| publish_time | publishedAt |
| channel_title | channelTitle |
| category_id | categoryId |
| views | view_count |

- **Remove unnecessarry variables**: In addition, CA_2022 contains an additional channel_id for each video and does not include video_error_or_removed information. Since we will be focusing on analyzing information for each video instead of the channel that the video belongs to, we will remove the channel_id column from CA_2022.

**Remove outliers:**

We use data.table to remove the data we want using logical functions.

- **Remove videos with the error or deleted**: For CA_2017, by observing videos with video_error_or_removed is TRUE, we can see that some videos' title is labelled as Deleted, or it has zero likes and dislikes. Since the amount of error or removed video in CA_2017 is 27, which is small compared to the whole dataset size of 40881, we can safely remove those videos from the CA_2017 data set and then remove the video_error_or_removed column.
- **Remove videos with zero views, zero likes and zero comments**: Since those are important attributes for a video and those variables will be used in the analysis, we remove those videos. A total of 5 videos are being removed.

**Observation of other variables with extreme values**

Further, we observed NAs and extreme values occurred in the following variables:

- likes: 604 videos have zero likes. Since YouTubers can set their videos to disable counting likes, we will keep those variables, and in the future analysis, if needed, we can remove those variables.
- vid_duration: 9268 videos have no video durations, and two videos have zero duration since we explained that those 9268 videos could be missing while we are using API to retrieve the duration, so we can keep them while the analysis does not involve using the duration variable, those two videos with zero duration could be caused by error, and after we tried to open the video online the video is unable to open but recommended to many people, considering that, there might be a glitch on youtube's algorithm, so we removed it.Another problem is the video duration is stored as characters in ISO 8601 duration, so we will use lubridate to parse it to seconds.
- 14 Videos have over 100000000 views. After opening those videos online, we found those are normal videos, just very popular.
- Trending date: for each data set CA_2017 and CA_2022, respectively, we created a vector of all the date ranges from the start to the end trending date and listed all the trending dates that are not recorded in the data. Since only eight days of data are missing from the CA_2017 dataset and three days of data are missing from the CA_2022 data set, we do not need to worry a lot about it as it does not affect a lot while analyzing for video views.
- Since one video could be trending on two different days, we need to check whether one same video occurred twice on the same day. The method we use is to construct two data frames, one lists all the video trending date and video id pairs, and another is the unique video trending date and video id pairs. We compare whether those two data frames contain the same data.

## Join all the datasets

We use the merge function to merge the CA_2017 and CA_2022 datasets with the data frame that was downloaded using API, which contains additional video information, including video duration, video dimension and whether the video has a caption.

Since category_id only contains an id that points to a specific category, we will load the category id dictionary from "CA_category_id.json" and merge the dataset with the id dictionary for the Canada region and obtain category names that each video belongs to. The id dictionary are downloaded from the Kaggle link provided in the introduction. After merging, we observe that the category with id = 29 does not match any category name in the dictionary after research (cite: https://gist.github.com/dgp/1b24bf2961521bd75d6c) category ID 29 corresponds to the category name Nonprofits & Activism, so we replace all the NA category names as Nonprofits & Activism.

For the last step, we add an indicator variable to CA_2017 and CA_2022 to indicate which dataset the video is from and then concatenate the two datasets into one data set called df_CA_trending.

## Questions and their analysis methods

### 1. Do more people are using youtube?

Due to the pandemic and the development of society, there could be more and more people starting to watch youtube. We will use a box plot to compare the number of views for each video between the years 2017 and 2022 for all trending videos to observe whether there is an increasing trend.

### 2. Does short videos still popular today?

As the short video has arisen since 2016, we will use a histogram to compare the video length between the years 2017 and 2022 for all trending videos to observe whether there is a decreasing trend.

To examine whether there is a relationship between youtube videos' view and duration, we will use advanced linear regression with a cubic regression spline on the video's length and view.

### 3. What words do those popular videos use?

Since all the videos are Canada's trending videos, we can analyze using text mining with stop words removal.

### 4. Which category is most popular?

We will use mostly barplot and histograms to show the distribution of each category, like view and comment numbers. We will also need to apply normalization to those variables.

### 5. What time to publish those videos could obtain the most views?

We will use a box plot for each hour to show the view's distribution.

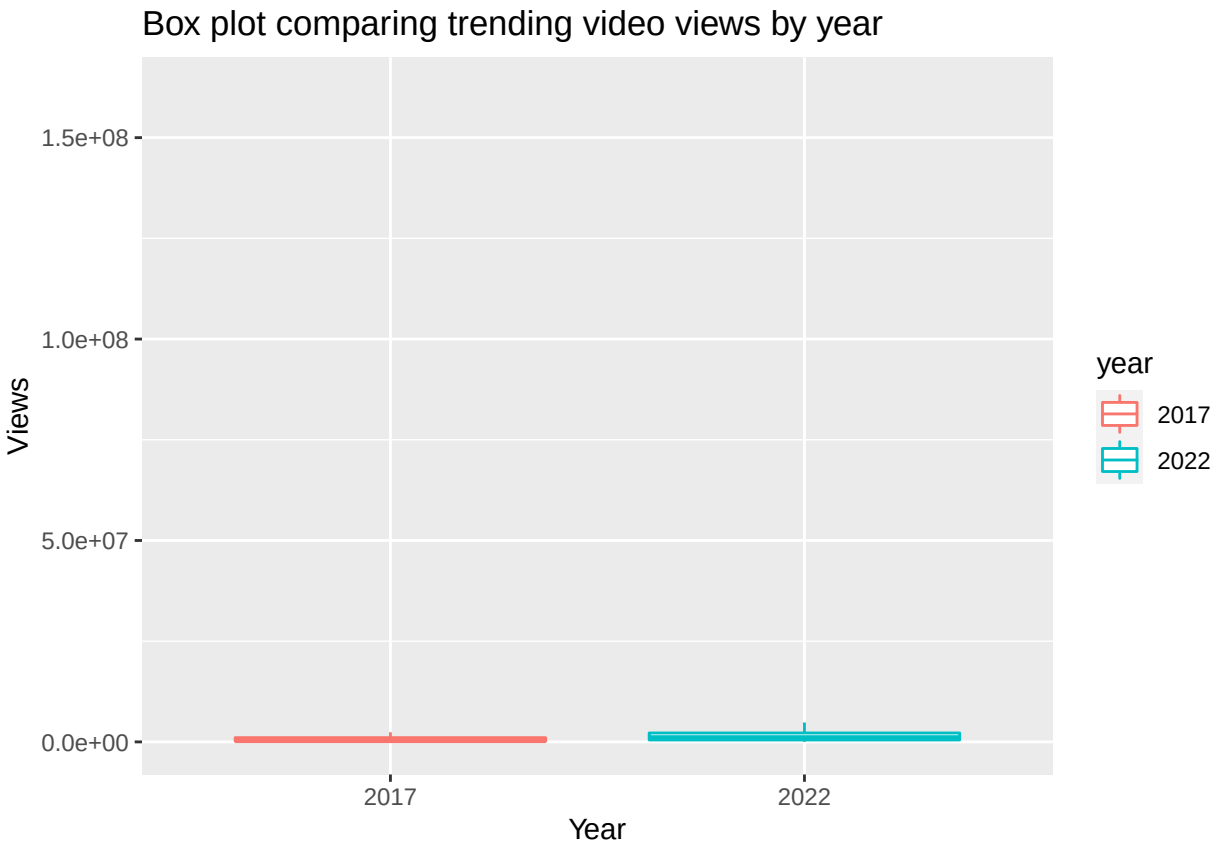
### 6. Can we fit a classification model?

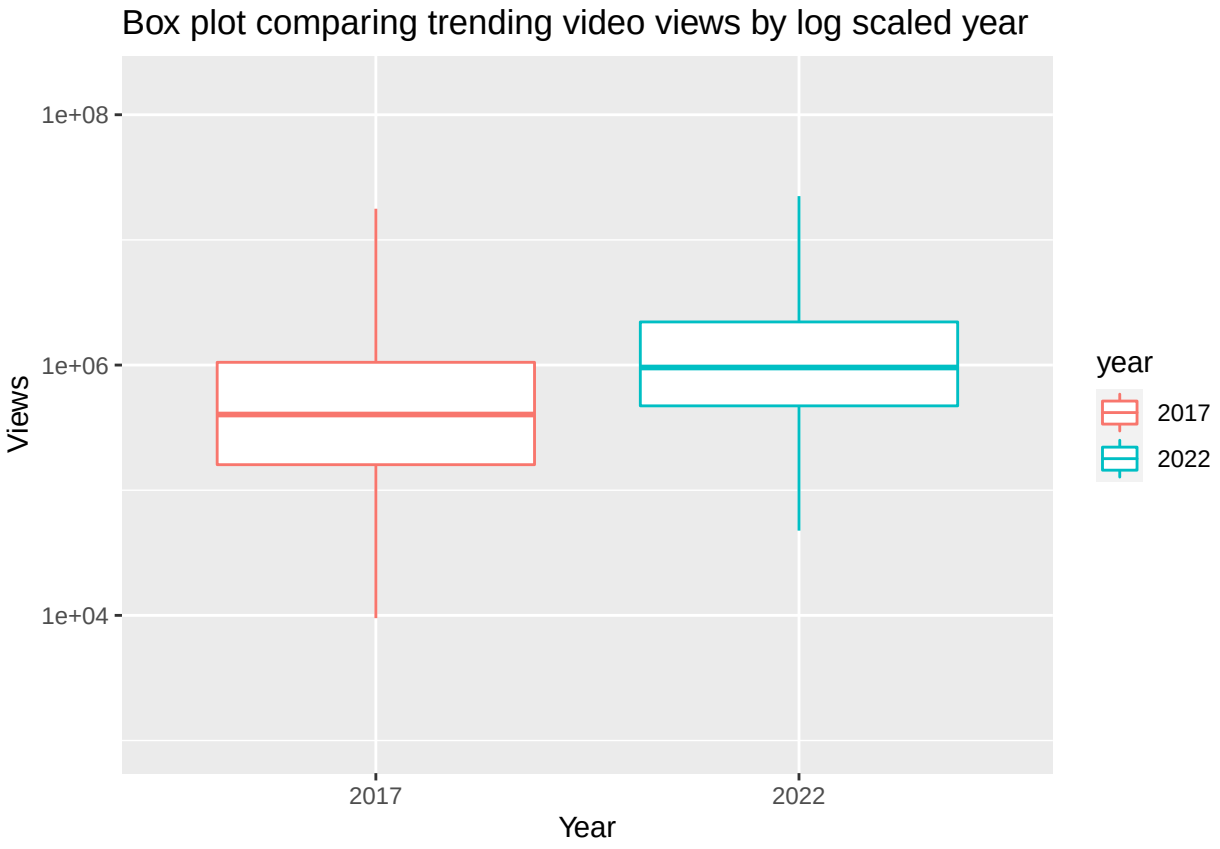
We would like to, but we need more variables to fit a model with views. In the following analysis, I will show that both like and comment count are linearly correlated with views. For the category variable and video duration, we can easily show its relationship with views using diagrams. We could design an algorithm that encodes video title, description and even profile picture, then use that information to fit a machine learning model, but that exceeded my current ability, and I will try to do that in the future.

# Results

```
CA_2022<-df_CA_trending[year==2022,]
CA_2017<-df_CA_trending[year==2017,]
```

1. Does more people are using youtube?

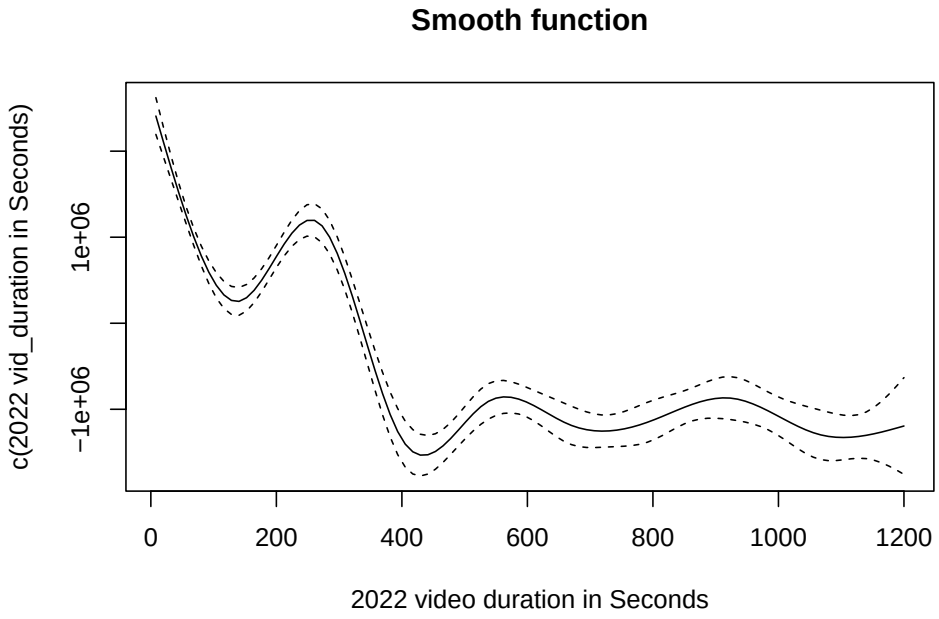


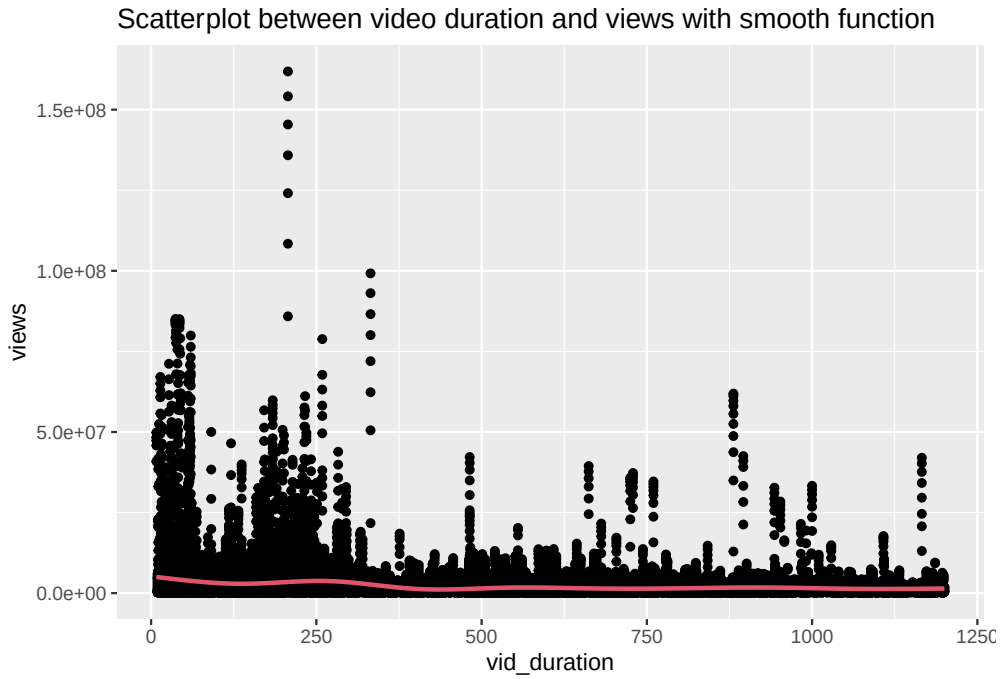



From the first plot, we can see that if we plot the views directly, we are unable to see the trend. Since each year has, a small portion of trending videos obtain many views, and those videos can be seen as a part of extreme outliers. To solve that, I applied log10 scaled the y axis and plotted plot 2. Here we can see that clearly trending videos are getting more views, and videos from 2022 have much higher median views and higher lower fence.

Further, since the F test shows two view samples are likely to have different variances (p-value close to 0), we performed a t-test with unequal variance. The p-value $<$ 2.2e-16 shows that the average video view is significantly different between the two years. Thus, we are confident that videos will get more views in 2022, and YouTube is getting more popular.

**2. Does short videos popular today?**






To examine whether short videos are popular today, here I use advanced linear regression with a cubic regression spline on the video's length and view.

The adjusted r-square of the spline model is 0.0372, which shows the model performs badly on estimating views using duration.

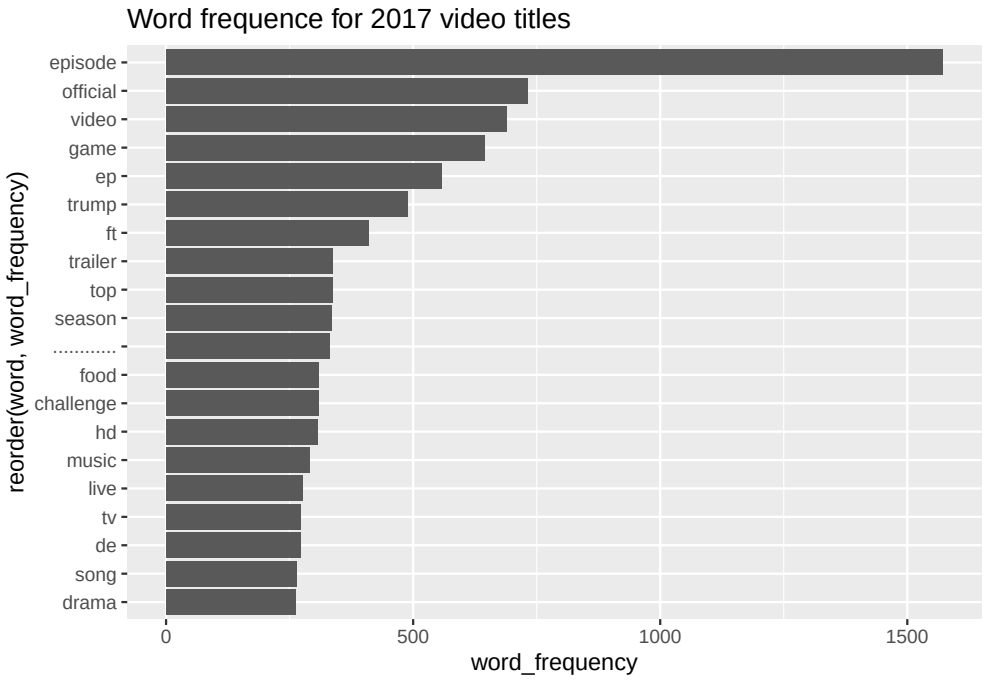
But, as the p_value for both intercept and significance of smoothing parameters is less than 0.05, we are confident that the trend shown in the smooth function graph could show a trend of the relationship between video duration and the number of views. We observe that the trend is decreasing and there is a mode of the number of views when the video duration is around 0, 210, 550, 900. The model interprets that the view number is highest when video duration is low (when video duration is around 0, 210, 550, 900 seconds), and the view is lowest when video duration is around 400 seconds.

Although video views are not interpretable using video duration, we can estimate from the fitted smooth function that shorter videos are more popular than long-length videos nowadays.

### 3. What words do those popular videos use on their title and description?

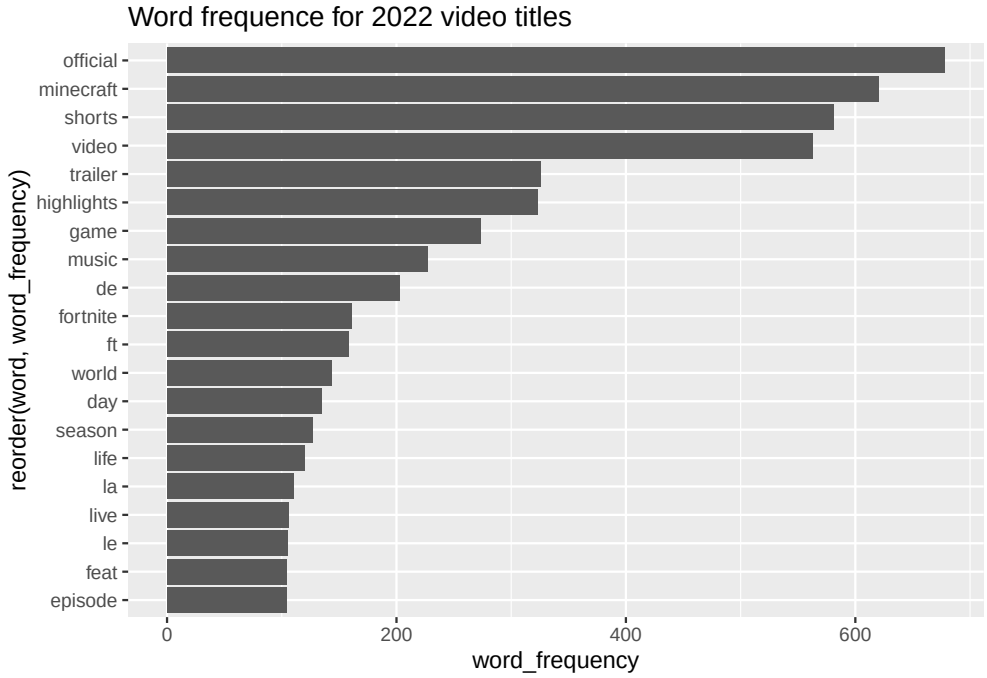
We will use Text Mining to obtain word frequence for both the video title and description. Full word frequence plot will shown in the website.

**Title**





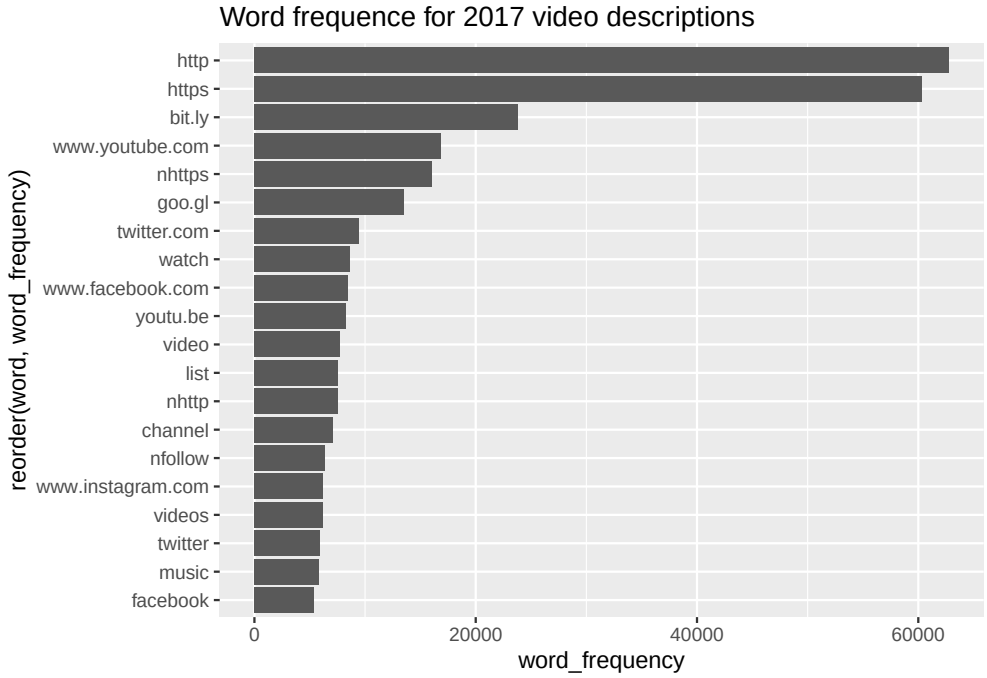
After removing stopwords, the word frequency gives us an excellent understanding of what those popular videos are talking about. For 2017 videos, we can see that most videos seem to be episodic and official. Game videos were popular at that time, and there are a lot of videos about Trump as well. From the words: "trailer, food, challenge, music," we can interpret that movie trailers, food videos and challenge videos were popular as well. Some non-English words show Canada has many non-English speaking people.

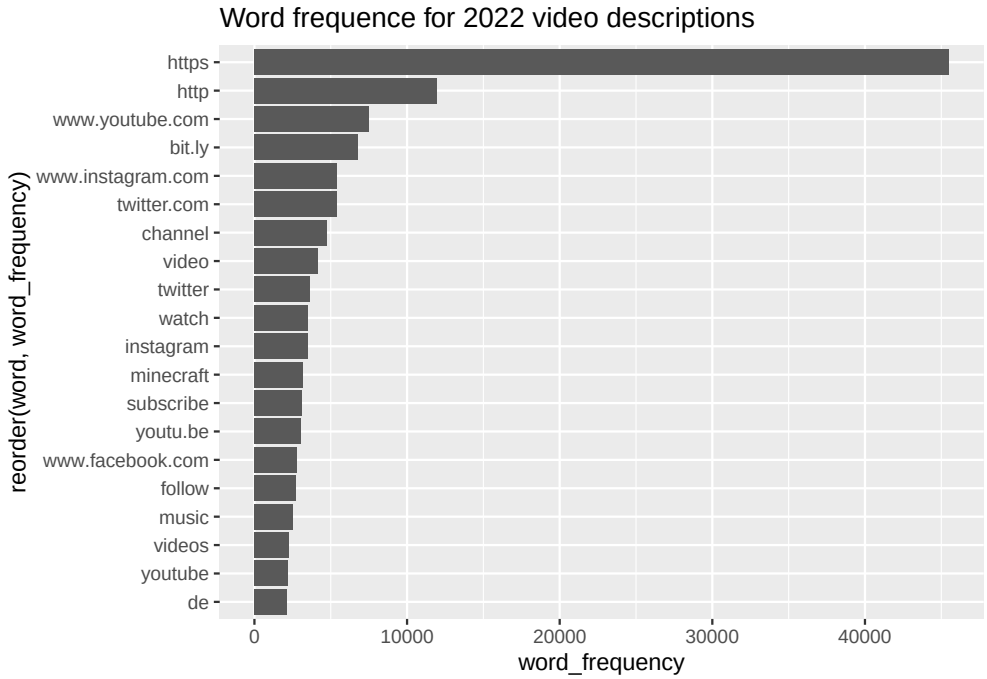




For 2022 videos, the most popular words are almost all about games. We can spot some famous game names such as Minecraft and Fortnite. Highlights and shorts are popular words showing most videos tend to be short, which verifies my conclusion in the previous part. There are some French words, such as la and le, which match Canada's official language.

**Description**

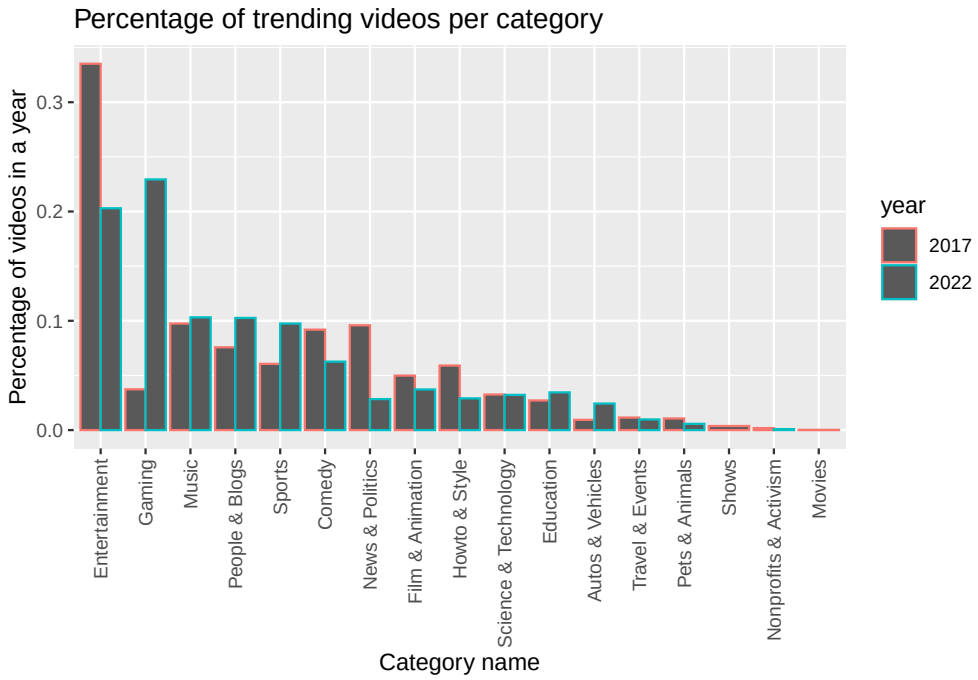


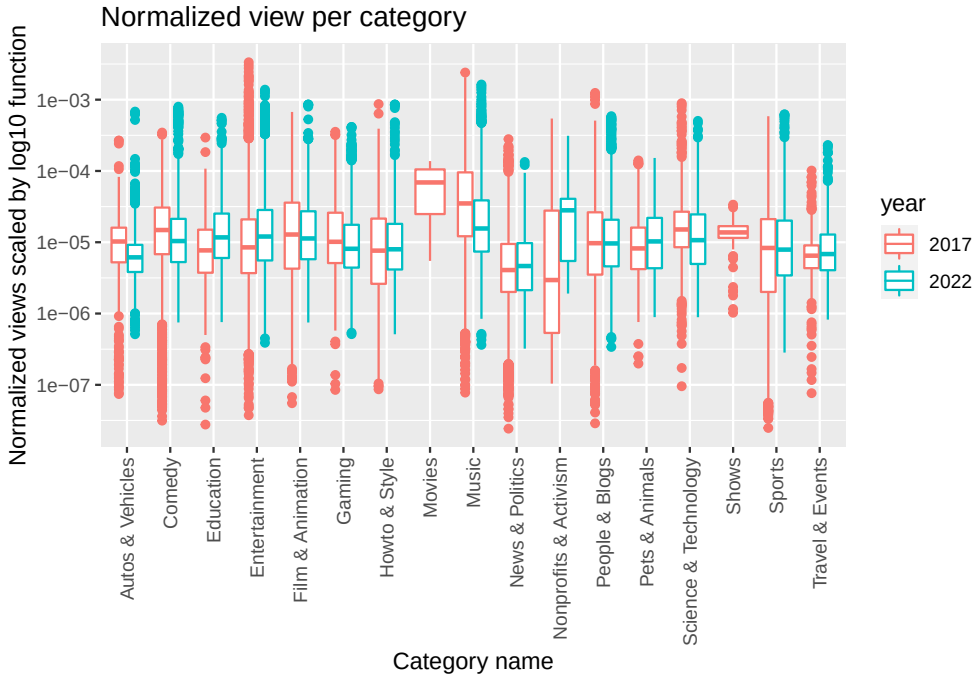

The most popular word in the description does not change much. Most of them are social network websites. And the purpose is probably to remind viewers to subscribe to their social networks. Instagram's frequency increased from 2017 to 2022, while facebook's frequency decreased might hint that Instagram is getting more popular than Facebook.

**4. Which category is most popular?**

We will analyze based on the number of videos in each category as well as video views and likes.



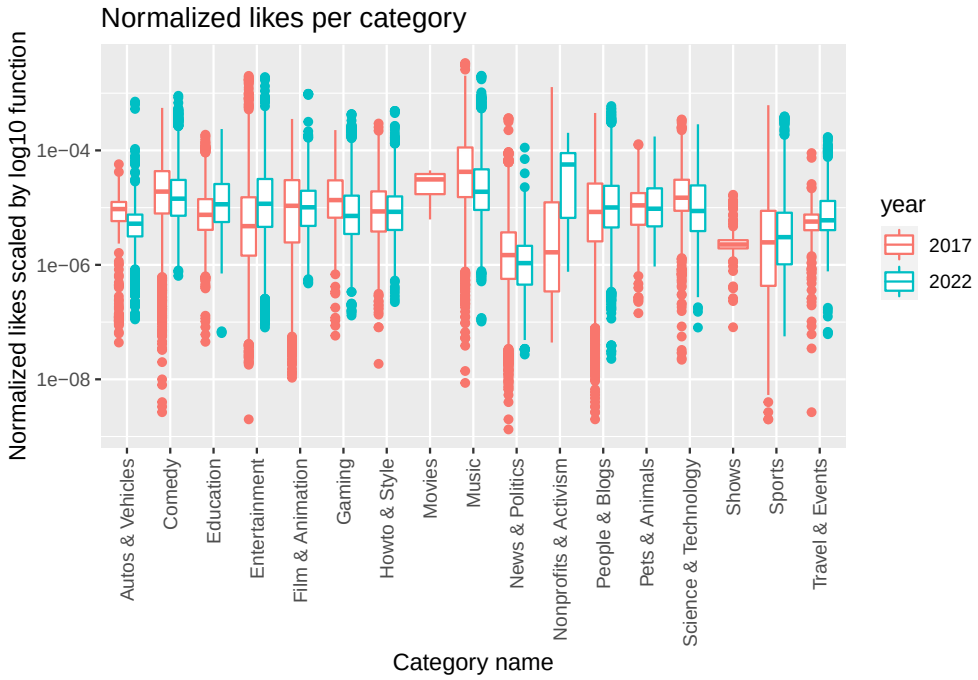

From the first plot, we can observe that in 2017, most videos are categorized in the Entertainment category, while in 2022, more videos are categorized in the Gaming category. This shows people's entertainment methods tend to skew toward video games instead of general entertainment videos.




For the second plot, I normalized each video's view by dividing each year's total number of views. The normalized views for autos and vehicles and music decreased significantly since 2017, while education and nonprofit video view increased significantly. It could be due to the pandemic, students started to get used to online education, and people are more aware of the global issues.

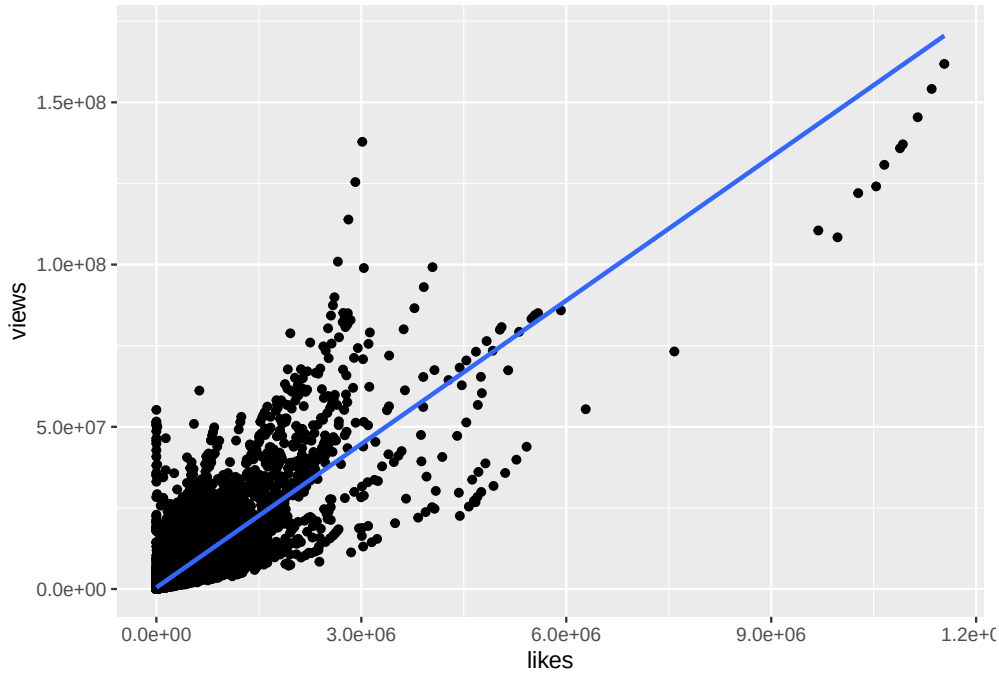
One thing to notice is that although the number of the video that belongs to entertainment has decreased

since 2017, the normalized view per video increased in 2022, which shows



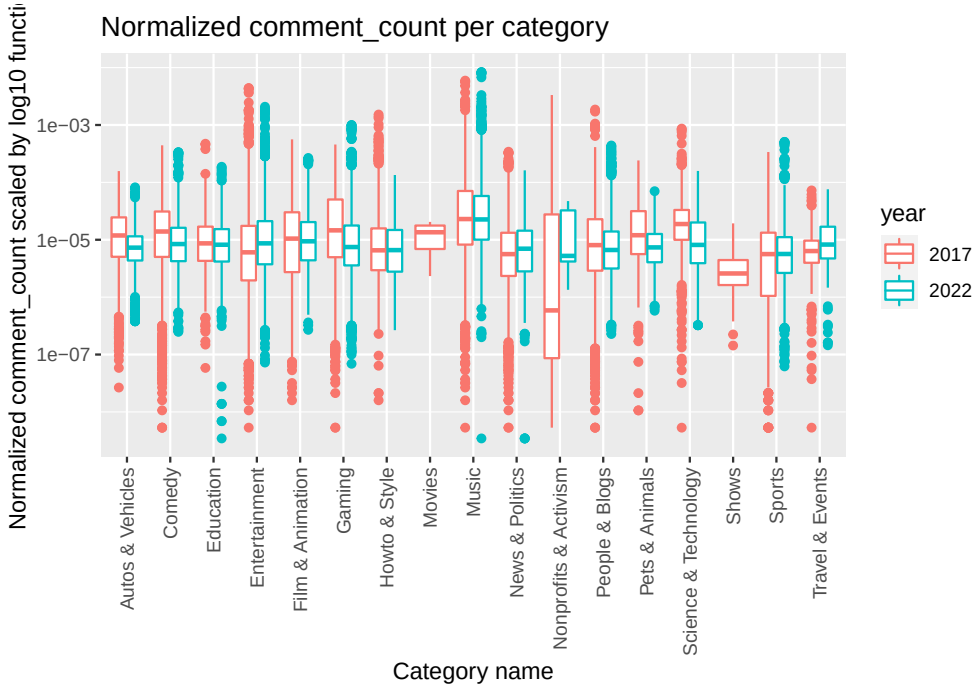

For the third plot, I normalized like for each video per category. One thing that interested me is that the second plot and the third plot are almost the same! So I will analyze whether there is some linear relationship between likes and views.

**Linear regression between like and number of views**



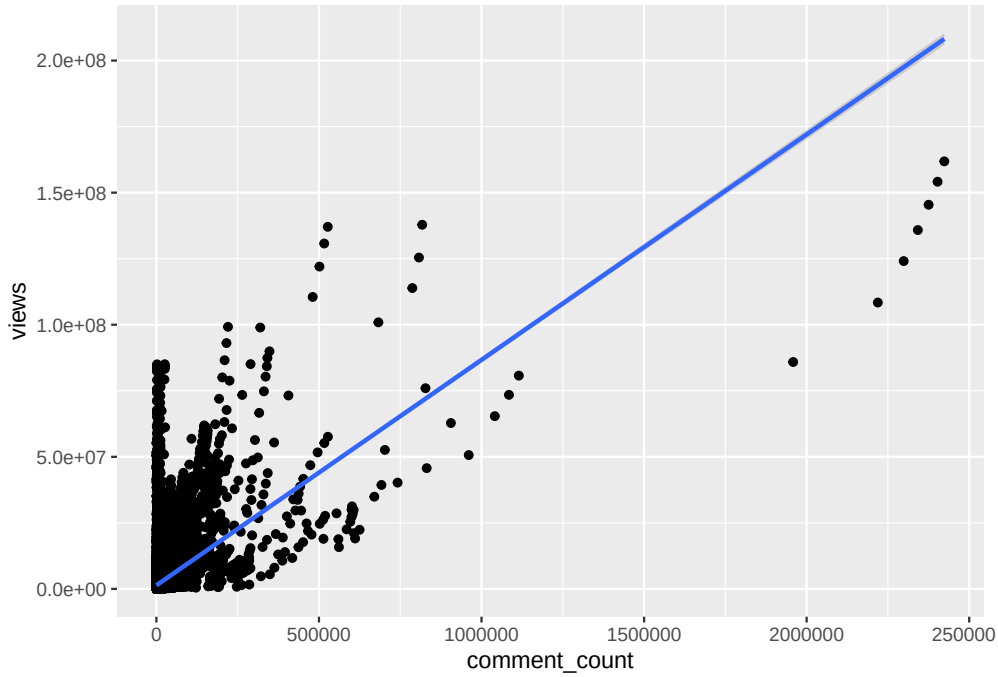

Here we observed a significant positive linear relationship between likes and view, as the p_value for coefficients is less than 0.05. Further, the r-squared is 0.7, which is close to one, and it shows we can even estimate the number of views based on the number of likes using the linear regression model. And we are confident that

trending videos tend to get more likes when more people view them.




From the fourth plot, we observed that nonprofit videos tend to get fewer comments than other videos. For music videos, although the normalized like and normalized view have both increased since 2017, the median number of normalized comments remains the same, which shows people tend to comment less on nowadays videos.

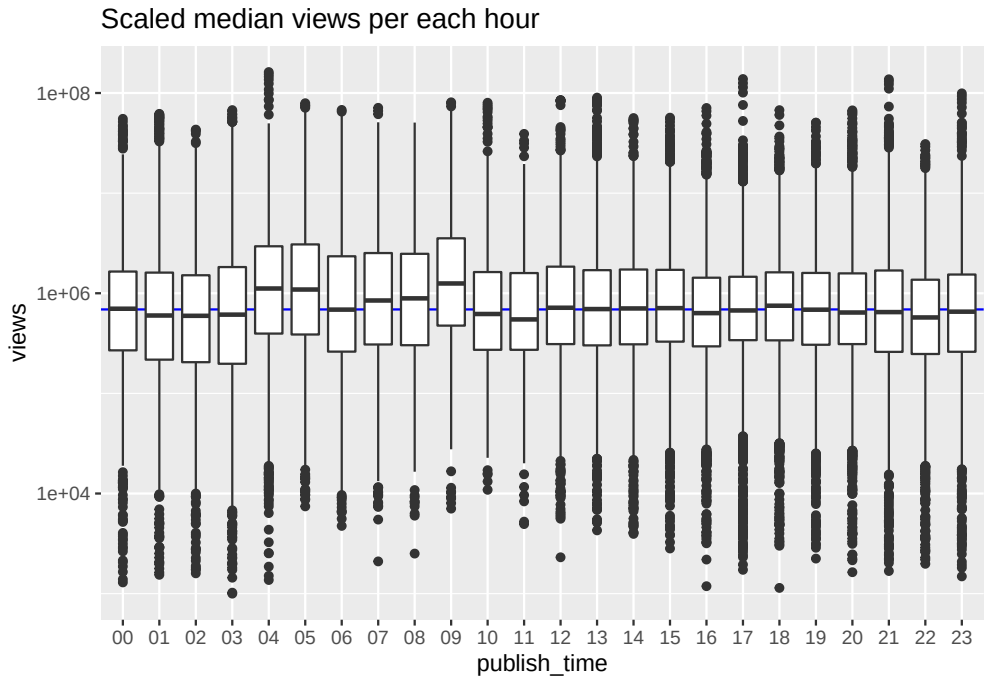
**Linear regression between comment count and number of views**




Since many videos disable viewers from leaving a comment, we remove videos with a disable comment setting and create a linear regression model between comment_count and views. Here we observed a significant

positive linear relationship between comment count and view, as the p_value for coefficients is less than 0.05.

**5. What time to publish those videos could obtain most views?**




Out of curiosity, we want to know whether there are view differences if we publish videos at different times of the day. From the plot, we found that video published at 4, 5 and 8 in the morning tends to have higher views.

# Summary

In summary,

Holland, Margaret. 2016. "How YouTube Developed into a Successful ... - Elon University." *Elon Journal of Undergraduate Research in Communications*. https://www.elon.edu/u/academics/communications/journal/wp-content/uploads/sites/153/2017/06/06_Margaret_Holland.pdf.