
**CSC 465**

**Ryan An - Joshua Crow - Daniel Kim - Sung-il Kim - Jacob Penrod**

# Contents

# Introduction

The dataset we have employed for this analysis is composed of data collected on a daily basis from videos featured in the Trending feed on youtube.com. Data was collected on videos in the United States, Great Britain, Germany, Canada, and France. The raw dataset can be found, and downloaded for free, here. The data has a number of variables associated with each trending video. A brief explanation of the variables is as follows:

*Date of Publication* - when the video was first uploaded to youtube.com

*Trending Date* - the date that the video was placed on the Trending feed

*Views* - a count of total video views

*Likes* - how many "Likes" a video received

*Dislikes* - how many "Dislikes" a video received

*Comment Count* - a count of the number of comments on a video

*Tags* - keywords that the videos have been "tagged" with

*Descriptions* - the text description of the video

*Category* - The category of the video has been placed in

The visualizations that follow will describe tenancies, trends, and nuances of YouTube's Trending feed.

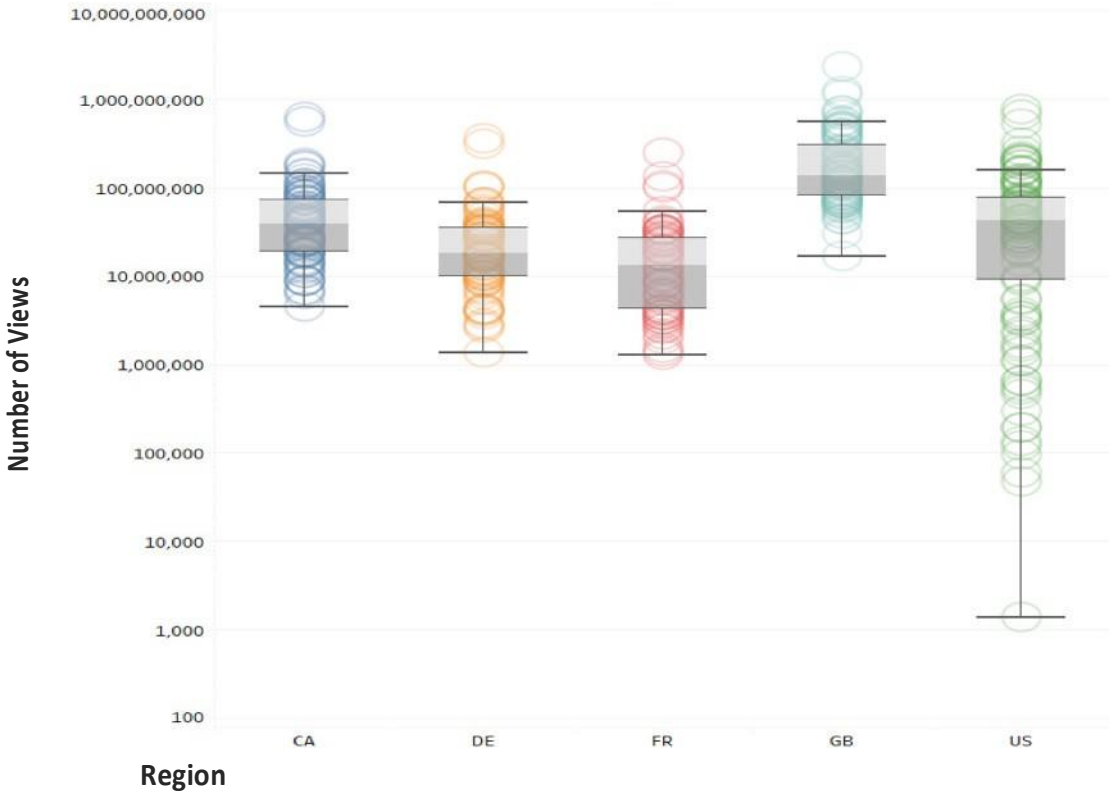
# Exploratory Analysis

Prior to analysis, some data cleaning was required and preformed. The data came in separate files by region and categories in a separate JSON file. First a column was added to represent the region (from the file name) [US,GB,CA,DE,-FR]. Then the files were concatenated into one large, master file. The JSON file was converted to csv and joined to the original data based on the Category ID field that already existed within data. This enabled analysis to begin.

A number of different avenues of data analysis were explored. This included views over time, likes/dislikes, variables by geographic location, and LDA of text data. A short overview of initial analysis can be found in the Appendix of this document.

# Visualizations

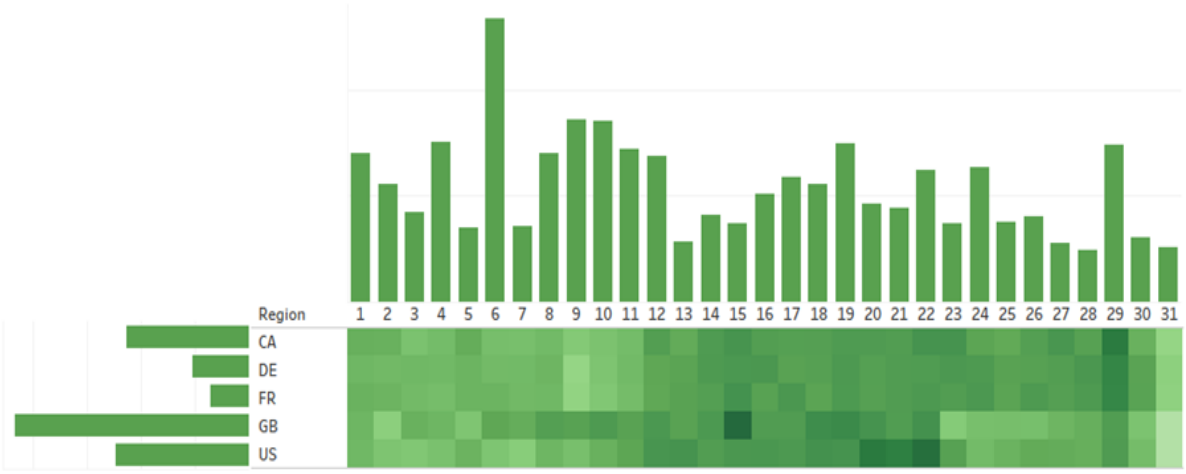
## Distribution of Video Views by Region




As shown in the graph above, Great Britain (GB) and the U.S. have videos that are viewed by large numbers of viewers. The overall distribution also indicates that there are more YouTube users in Great Britain and the U.S. compared to the rest of regions in this dataset. In addition, it can be suggested the market size of Great Britain and the U.S. can offer more opportunities for content creators and marketers who would want to earn money using YouTube. More viewers provide more opportunities for advertisement and exposure on web in Great Britain and the U.S. compared to all other regions.

# Visualizations

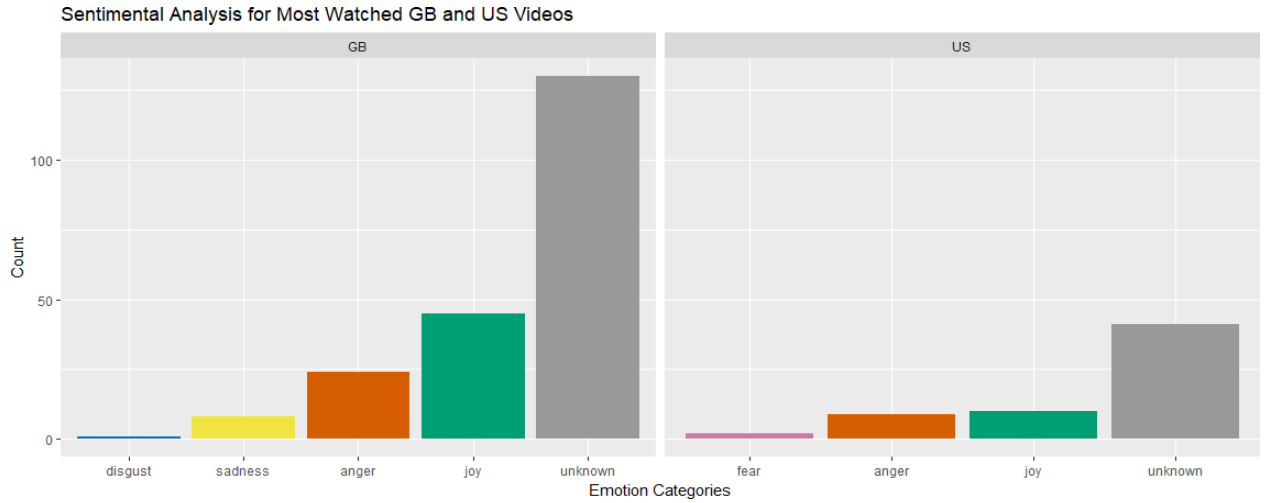
## Average Views per Region by Publication Day of Month

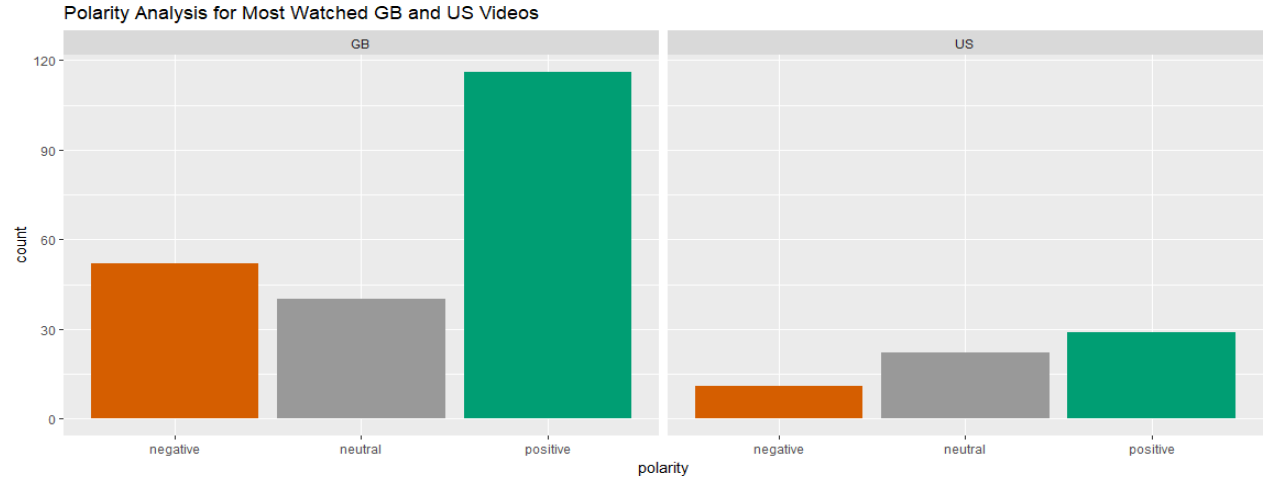



This visualization above represents the average video views on each day of the month on which the video published. It is of note that the average video views, which is located on the top side of the heatmap, are higher on the 6th, 9th, 10th, 19th, 22nd, 24th, and 29th, with the 6th being exceptional. The bar chart next to region represents the average view of all the videos in each region. Since the dataset contains more data from Great Britain (GB) compare to other regions, it is easier to observe the average viewing pattern of the GB. Based on visualization, it indicates that GB videos are likely to be trending between 15th and 19th since the bar chart above shows the increasing number of average views between these dates. An additional finding is that if there are videos uploaded around 29th day of month, it is more likely to be trending with help of views within the next 15 days, which is seen from the chart as average view changes between 1st and 12th date of a month. Overall, this chart suggests how to move more strategically if one tries to make a video as trending when deadline is given.

# Visualizations

## Sentiment and Polarity Analysis of US and GB Videos



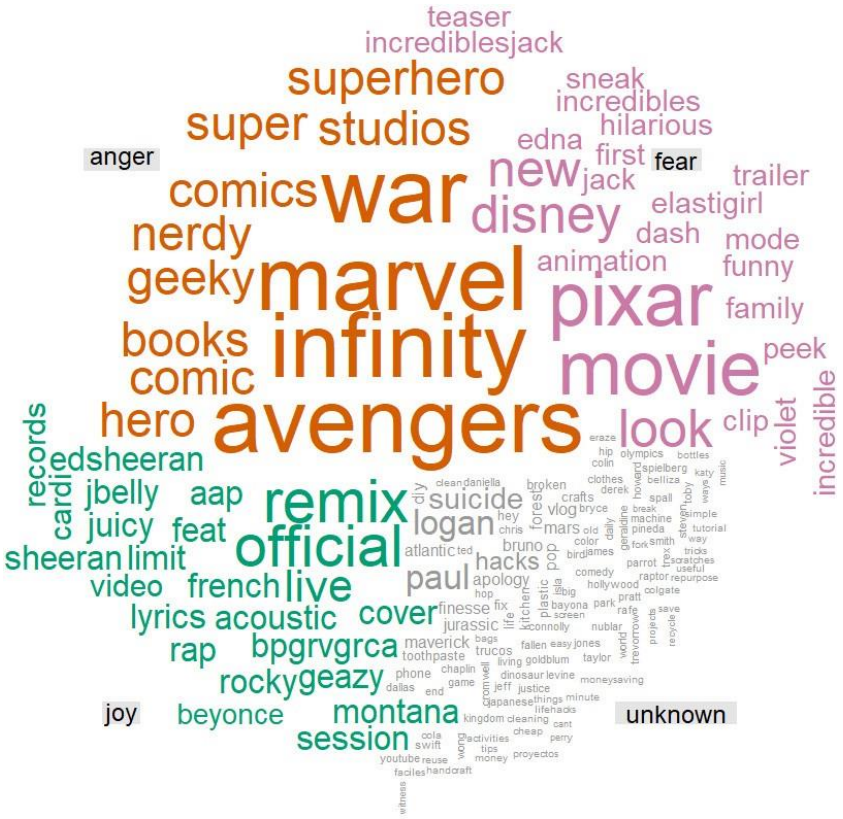




In order to focus on analyzing the tags of the most viewed videos for both US and GB, the data was filtered so that only videos with 30 million views or higher were included. The raw tags had too many non-essential characters such as foreign languages, punctuations, numbers, and so on. Therefore, the tags were filtered first to only include ascii characters. Only the non-ascii characters were removed, because removing the whole tag that contains non-ascii characters will leave out only a few observations due to punctuations. Sentiment/Polarity Analysis works by classifying the given word using a naïve Bayesian classifier (a machine learning algorithm) trained on Carlo Strapparava and Alessandro Valitutti's emotion lexicon. After classifying the tags into different emotion/polarity classes, small multiples barcharts of each emotion/polarity classes with region separations were made. Each class was subdivided by color in these graphs. There are a lot of “Unknowns” in the emotion graph. These unknowns do not necessarily mean neutral polarity, although a lot of the words in this class may belong in the neutral class. The “Unknown” simply means the algorithm was not able to classify the word, potentially due to pre-processing errors or it may be a foreign word in English alphabets. The sentiment/polarity small multiple barcharts demonstrate that the most viewed videos had a lot of joyful and positive tags in both GB and US. It also illustrates that there were more GB total tag count overall. This can mean that either more GB videos made it to 30 million views or above or most viewed GB videos had more tags in general.

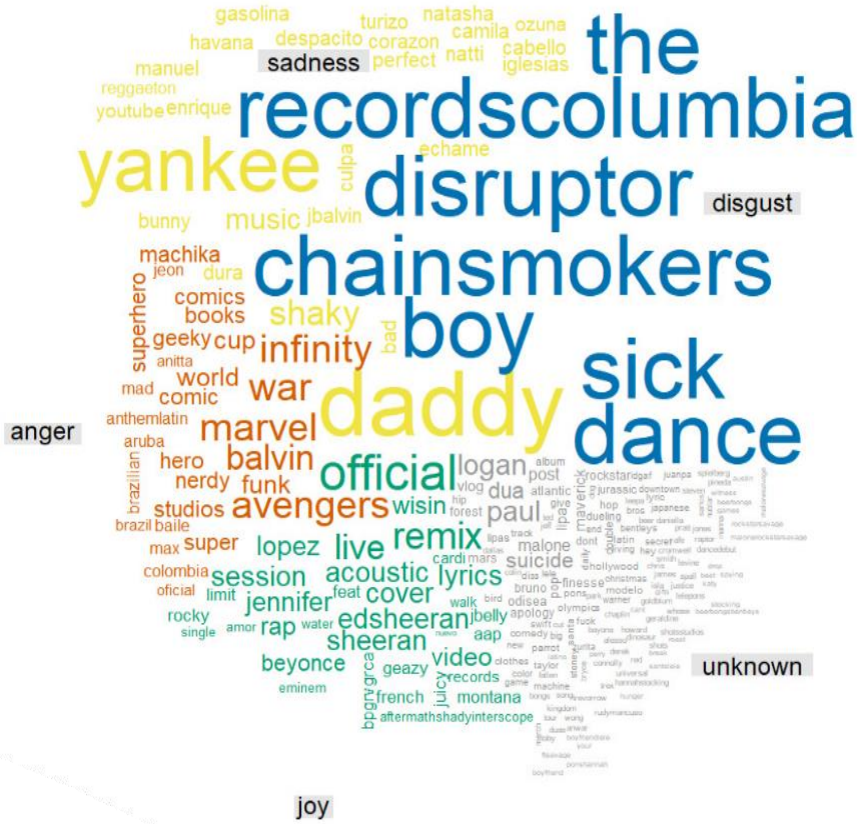
# Visualizations

## Word Cloud of Most Viewed Videos in Great Britain and US

US




GB



# Visualizations

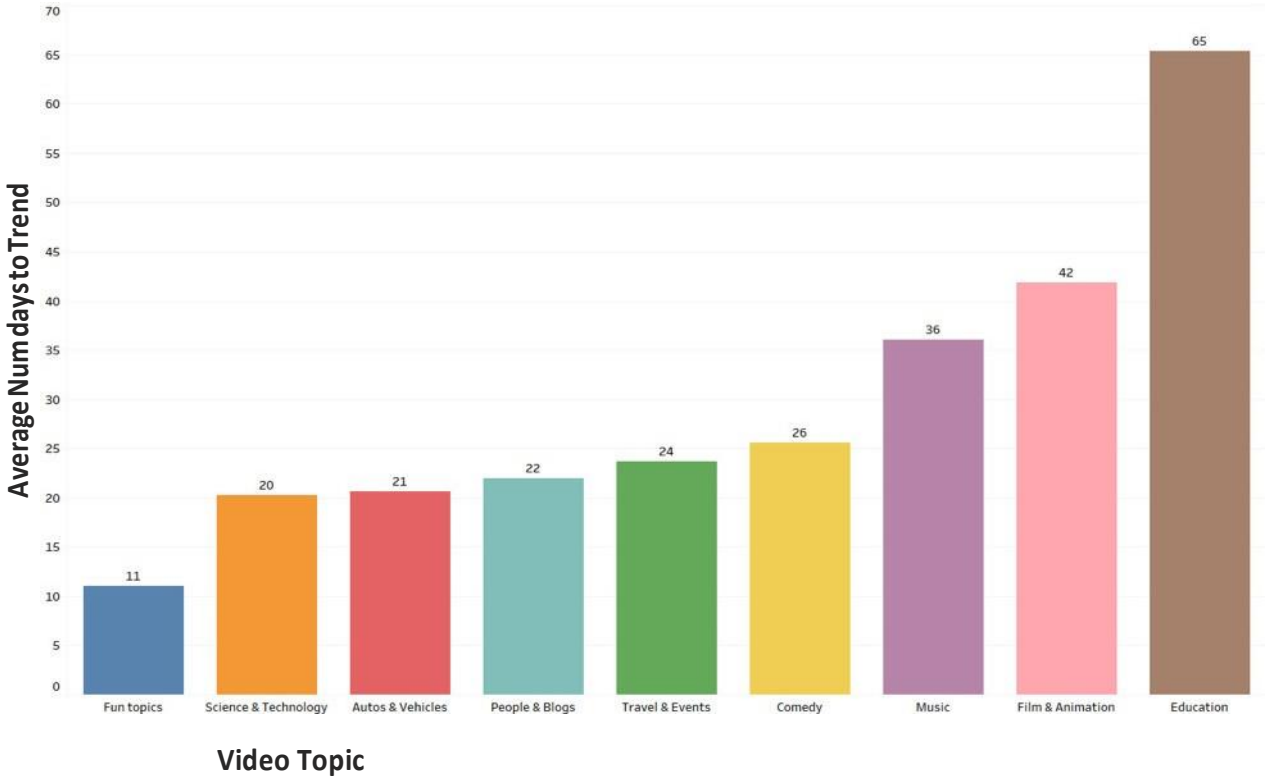
## Word Cloud of Most Viewed Videos in Great Britain and US

Note: This description refers to the word clouds on the previous page.

The tags of Most Viewed Videos in the US and GB were put into two different wordclouds. The same filter (above 30 million views) was used to only select the Most Viewed Videos for US and GB. These two wordclouds cannot be compared to each other, since the data was subset by region before each wordcloud was created. The tag size within the wordcloud represents the frequency of the tag used. The color illustrates the emotion the tags belong to (same emotion colors as the sentiment analysis graph and the emotion labels also show up by edge of the wordcloud). The size of the entire wordclouds was altered to capture all the tags in the Most Viewed Videos. There are some tags that are in a questionable emotion class. This is because the tags were compared to the emotion lexicon developed by Carlo Strapparava and Alessandro Valitutti a long time ago. Today's internet or slangs might not be properly classified due to this fact. This may also be true for the high “unknown” count in the sentiment analysis itself. The wordclouds suggest that movie related tags were used the most in the US and music/dance related tags were used the most in the GB.

# Visualizations

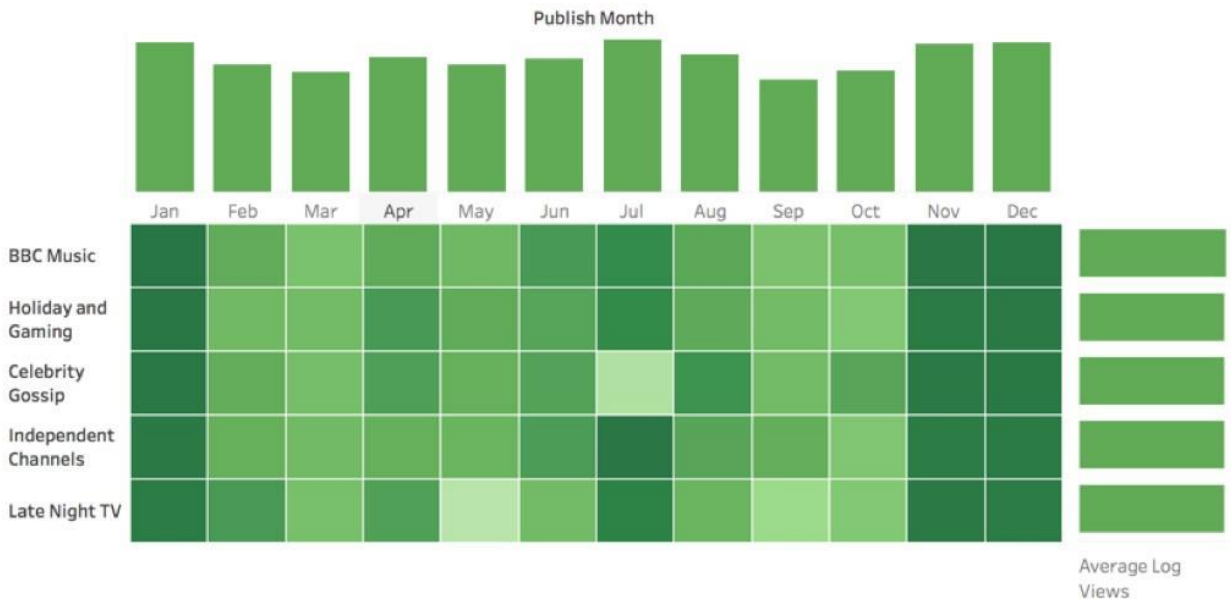
## Fun vs Other Topics, Average Length of Time (in days) from Publication to Trending Status




The chart above titled as so because the average trending days that fall under the “Fun topics” required the least number of days on average to become trending when compared to all other topics. Under Fun Topics, Entertainment, Gaming, Howto & Style, News & Politics, Nonprofits & Activism, Pets & Animals, Shows, and Sports are included. The chart includes data pertaining to all regions. It seems that of all categories, Education takes the longest to achieve trending status. This may indicate that people do not access YouTube trending videos to learn something, but rather use it for relaxing or fulfilling their enjoyment. The criteria to be included for Fun Topics is to have less than 20 days to become trending, and the categories included in Fun Topics all have less than 20 days to become trending, ends up having 11 days on average to become trending.

# Visualizations

## Average Log View By Topic and Published Month




This chart is a heat map of the average of log views by published months. Topics were discovered by a process known as Latent Dirichlet Allocation (LDA) on the video descriptions of all the English regions (US, GB, CA).

We can see some overarching trends, average views for all topics are up in general around January, November and December, post most of your videos in this timeframe. March is mediocre across all topics, so if you are a YouTube publisher, take your vacation in March. The remaining analysis shows how each topic differs from each other in terms of average viewership. For example, Celebrity Gossip has a low average view rating in July compared to the other topics, perhaps due to people being out of the office and school. The only topic to have a decent October is Celebrity Gossip. Late Night TV has off months in May and September, these months may represent changes in normal scheduling as May usually marks the beginning of Summer break and September is back to school. Holiday and Gaming has very good average numbers in November, December and January for obvious reasons. It has decent numbers in April (Easter), July and August (Summer Holidays). Independent channels seem to be consistent but have one noticeably down month in October.

# Analysis and Discussion

Our analysis clearly reveals some features of YouTube's trending videos data. The first visualization in this paper clearly shows that videos within the Trending feed with the highest number of views are from Great Britain and the U.S.. This is potentially due to the fact that this dataset was collected and created by a user in Great Britain. Since YouTube's trending feed page is customized per user, this explains the regional imbalance in terms of total trending video observation count within the dataset.

The second visualization illuminates days of the month that videos become Trending. According to our analysis, videos will receive the most views if they are published in Great Britain on the 6th of the month. As far as the content of tags of the most viewed videos is concerned, within Great Britain and the U.S., the predominate emotional category is joy within the analyzable range (not including the "unknowns"). The second strongest class of emotion for both regions is anger. Positive tags were the most frequently used in the most viewed Great Britain and U.S. videos, which could also be explained by the popular joyful emotion class. Negative tags of videos is second in Great Britain, but third in the U.S. This sentiment analysis can be seen with greater granularity in the wordclouds of the respective Regions. Although it is not possible to compare tags count between the two regions in these wordclouds, acknowledging which tags were used most often is possible within each region. In the U.S., movie related tags were used the most. In the Great Britain, music/dance related tags were used the most. This portrays that, according to the YouTube dataset, movie videos tend to get the most attention in the U.S. and music/dance videos get the most attention in the Great Britain.

The visualization found on page 10 demonstrates that videos that fall within the "Fun topics" category are most likely to go to the Trending feed faster than any other category.

The last visualization shows which type of videos most often are featured as we also looked at when is the best time to post your video by topic. We see that average views for all topics are up in general around January, November and December, definitely post most of your videos in this timeframe. March is mediocre across all topics, so if you are a YouTube publisher, take your vacation in March. Celebrity Gossip has a low average view rating in July. Independent Channels had a view drop in May and September. Although it is not clear why Celebrity Gossip had a decrease in July, the drop for Independent Channels possible can be explained by school/university break start and end time frame.

All the visualization techniques used portrays that tag's sentiment/polarity class affects the video's trending-ness and therefore views. The time of year and associated genre/type of video has a definite effect on views. However, this does not factualize that movie/music/dance videos will always be the most sought-after videos in the U.S. and the Great Britain. It also does not prove that Holiday/Gaming videos will not be popular in July. This is because the data has a lot of bias. It mainly focuses on only U.S. and the Great Britain from 10/31/2017 to 01/31/2018. This does not provide enough data and randomness within the dataset, especially with stochastic analytic methods. Although all the analysis done does not classify nor predict anything, all analysis done above illustrates and highlights some important features within the trending YouTube dataset. This will help building future analysis approaches for popular video trends in the YouTube community.

# Appendix - Individual Reports

Team Viral – Individual Report

**Ryan An**

**Data Cleaning**

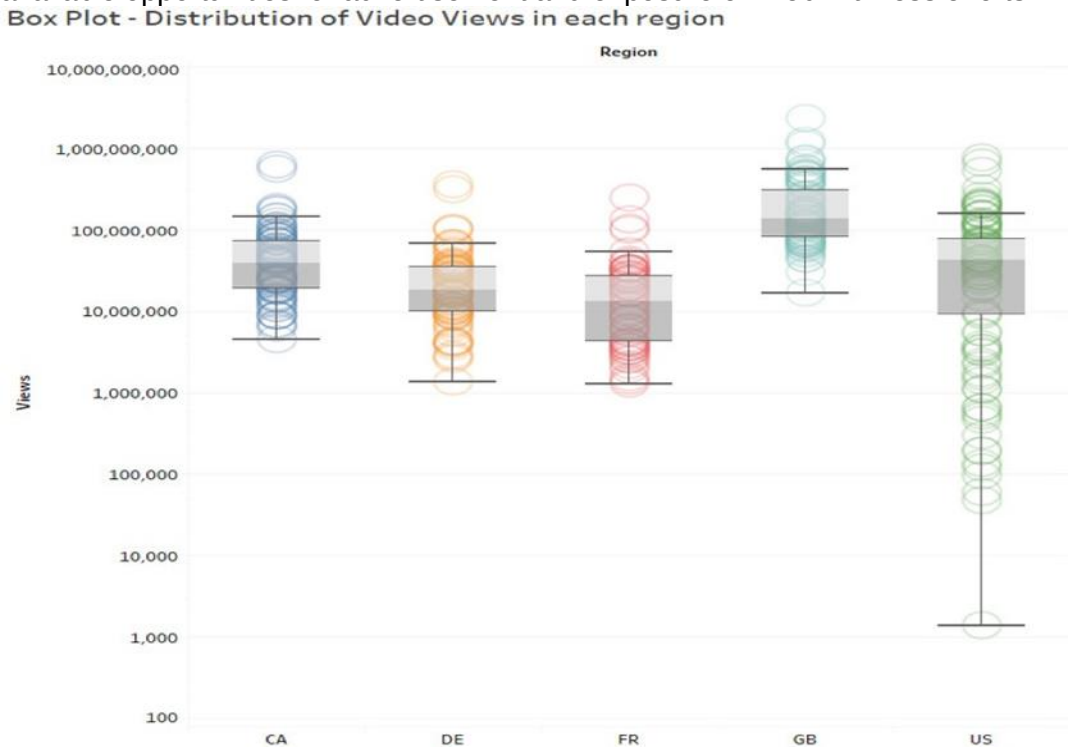
With the merged dataset, each column in the table was not appropriately set, so I fixed the data type of each column and did split into two groups. One was Numerical and another one was categorical table, then we could try further analysis.

**Exploratory Analysis**

I focused on the exploratory analysis, especially the relationship between key features and trending. Due to ASCII text issue, I extracted the numerical data only, and I created the number of views, like, dislikes, and comments of each video grouped by region. Among all the numerical features, the number of views indicates the trending of the videos the most. When we looked up the data, the number of videos uploaded from the UK was the highest, but the US videos were distributed the most.

**1. Box-whisker**

As referring to the result below, it seems that GB and US have videos that are viewed by lots of users. The overall distribution also indicate that there are more YouTube users in GB and US compare to the rest of regions in this dataset. In addition, it can be suggested the market size of GB and US can offer more opportunity for content creators and marketers who would like to earn money using YouTube. This is because having more active users indicate the more available opportunities for advertisement and exposure on web with less efforts.



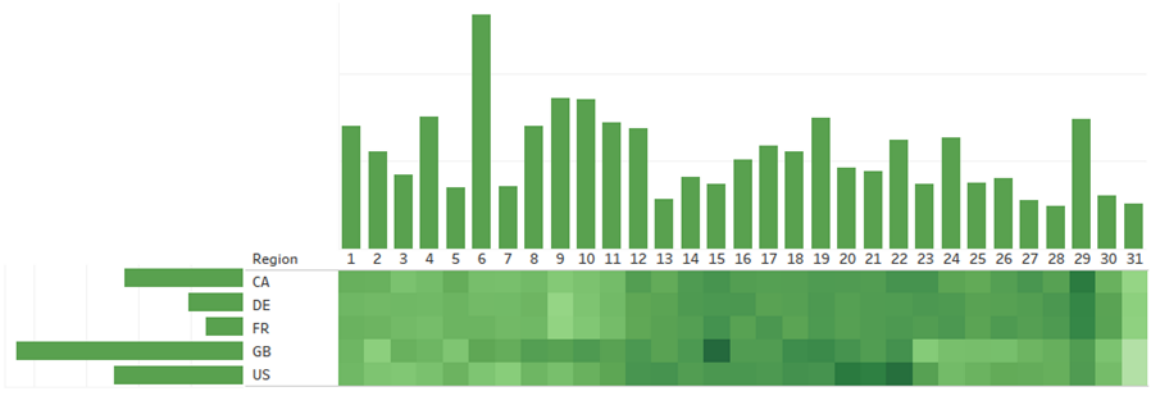

**2. Tree map**

According to the tree map below, among all the categories, music and the entertainment was the most popular. This tree map was created using Tableau, and the data was rolled up to category and grouped by region. What we can find from the chart is that the top view categories found from all the regions are Music and Entertainment, which clearly shows the role of YouTube – people mainly access to YouTube to browse fun videos or listen to music. There are other topics such as Comedy, Film & Animation, People & Blogs, etc. that are usually viewed by

users in YouTube. This clearly indicates that some of the categories are easier to be trending compare to others such as Education.


### 3. Heat Map

This visualization represents the average video views on each day of the month on which the video published. It is of note that the average video views, which is located on the top side of the heatmap, are higher on the 6th, 9th, 10th, 19th, 22nd, 24th, and 29th, with the 6th being exceptional. The bar chart next to region represents the average view of all the videos in each region. Since the dataset contains more data from Great Britain (GB) compare to other regions, it is easier to observe the average viewing pattern of the GB. Based on visualization, it indicates that GB videos are likely to be trending between 15th and 19th since the bar chart above shows the increasing number of average views between these dates. An additional finding is that if there are videos uploaded around 29th day of month, it is more likely to be trending with help of views within the next 15 days, which is seen from the chart as average view changes between 1st and 12th date of a month. Overall, this chart suggests how to move more strategically if one tries to make a video as trending when deadline is given.

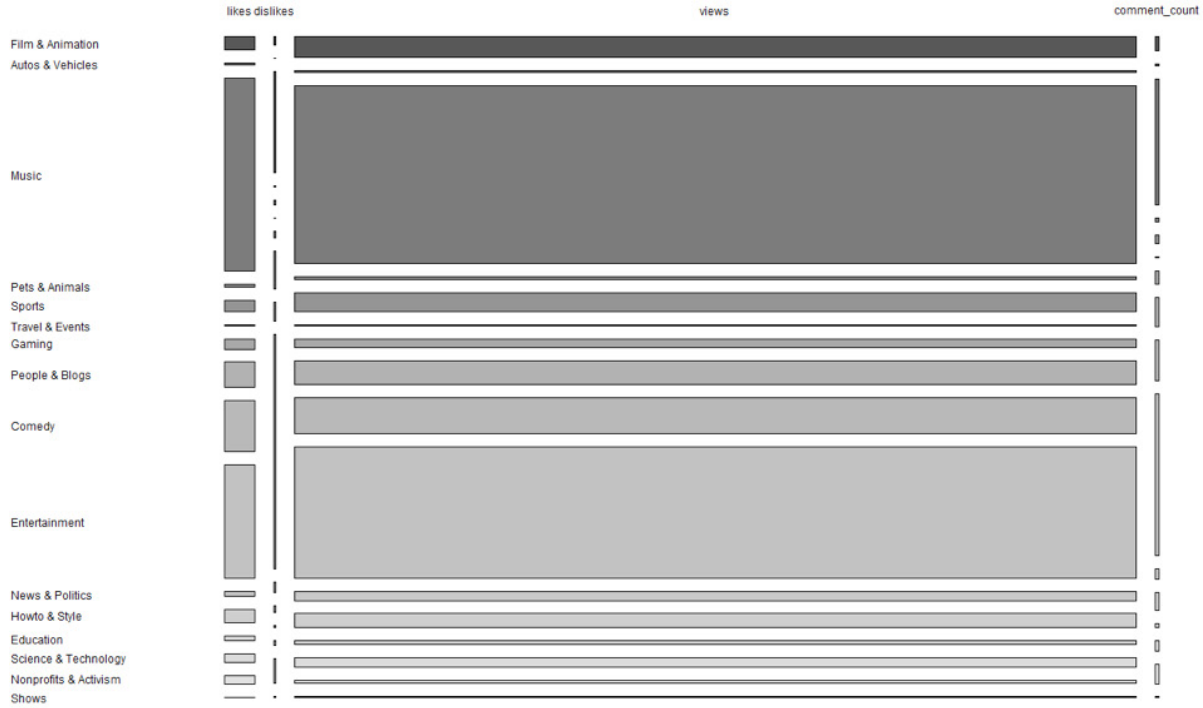


# Appendix - Individual Reports

**4. Mosaic Plot Mosaic Plot**

The mosaic plot below explains the portion of different reactions from YouTube users regarding videos in each of categories, including all the regions. Obviously, view has the largest portion because the number increases once a user clicks the video, then follow by likes, comments, and dislikes. The interesting finding from Music is that the length of likes is longer than views – this possibly indicates that some of the videos were more favored by YouTube users. Moreover, it is common to see that people like to express their feeling after they listen to music especially for their favorite musicians. Another interesting finding is that

Entertainment has the longest bar for dislikes, indicating that users might not like some of the videos and click dislike button without any hesitation – in this case, contents can be thought as inappropriate or controversial to users in common.




Code

```
if object_id('tempdb..#all') is not null
drop table #all

select *
into #all
from dbo.youtube_sub

if object_id('tempdb..#date') is not null
drop table #date

select
trending_duration = DATEDIFF(d, publish_time, trending_date),
    [trending_date] ,
    [publish_time] ,
    'views' = cast([views] as float) ,
    likes = cast([likes] as float),
    dislikes = cast([dislikes] as float),
    comment_count = cast([comment_count] as float),
    [region] ,
--  [CategoryID] ,
    [CategoryName]
into #date
from #all
```

# Appendix - Individual Reports

```
-- trending duration by category each region
select CategoryName, region, trending_duration = ceiling(avg(trending_duration))
from #date
group by region, CategoryName
order by region, avg(trending_duration), CategoryName
-- count the number of videos uploaded by region
select region, CategoryName, No_video = count(CategoryName)
from #date
group by region, CategoryName
order by region, count(CategoryName) desc
-- count the views of videos by region, category
select region, CategoryName, No_view = ceiling(avg([views]))
from #date
group by region, CategoryName
order by region, avg([views]) desc

select trending_year = year(trending_date), region, CategoryName,
tot_like = ceiling(avg(likes)), tot_view = ceiling(avg([views]))
from #date2
group by year(trending_date), region, CategoryName--, likes, [views]
order by region, year(trending_date) desc, avg(likes) desc -- , --cast(likes as float) desc
```

And the rest of cleaning job was done in excel

```
library(vcd)
setwd('C:/Users/RADK/Desktop/School/Winter 18/CSC 465/Project')

myd=data.frame(read.csv("youtube_sub2.csv",header=TRUE))
head(myd)

df = myd[,5:8]
cat = myd[14]
dt = cbind(df, cat)
head(dt)

#### Data converting ####
library(data.table)
DT <- data.table(dt)
newdt = DT[, list(sum(likes),sum(dislikes), sum(views), sum(comment_count)), by = CategoryName]

newdt = as.data.frame(newdt)
colnames(newdt) = c('Category','likes','dislikes','views', 'comment_count')

head(newdt)

### converting2: new format
rownames(newdt) <- newdt[,1]
dframe <- newdt[,-1]
head(dframe)
class(dframe)
mosaicplot(t(dframe),main=" ",las=1,cex=0.75,color=TRUE)
```

- Mosaic -

During this project, I could learn about how to approach to a specific dataset before explore and analyze further. In addition, I have also learned more data cleansing by utilizing different software and contribute the team by providing a clean dataset. Further, while visualizing categorical and numerical data altogether could not be done without putting much effort, I could learn about the better approach in terms of visualizing these two different types of data more effectively as I move toward the end of this project.

# Appendix - Individual Reports

**Joshua Crow**

I put together this document, the presentation, and provided editing and graphic design expertise. I helped initially with ideation and developing a narrative. I provided a lay-person's view point to help develop the general public perspective that this document represents.

With no background in data science, this project was an eye opener for me. I learned about dealing with data; the concept of cleaning data was previously foreign to me. Working with such a large dataset was a point of difficult to me. It took me a long time to start to comprehend how to approach the data and the story we wanted to tell. I helped with all document creation and writing involved in this project. I found that I may not be able to work easily with the data directly, but reading and interpreting information is something I can handle and contribute to. Working as the only out of state group member was a challenge in time management.

# Appendix - Individual Reports

**Daniel Kim**

Individual Summary

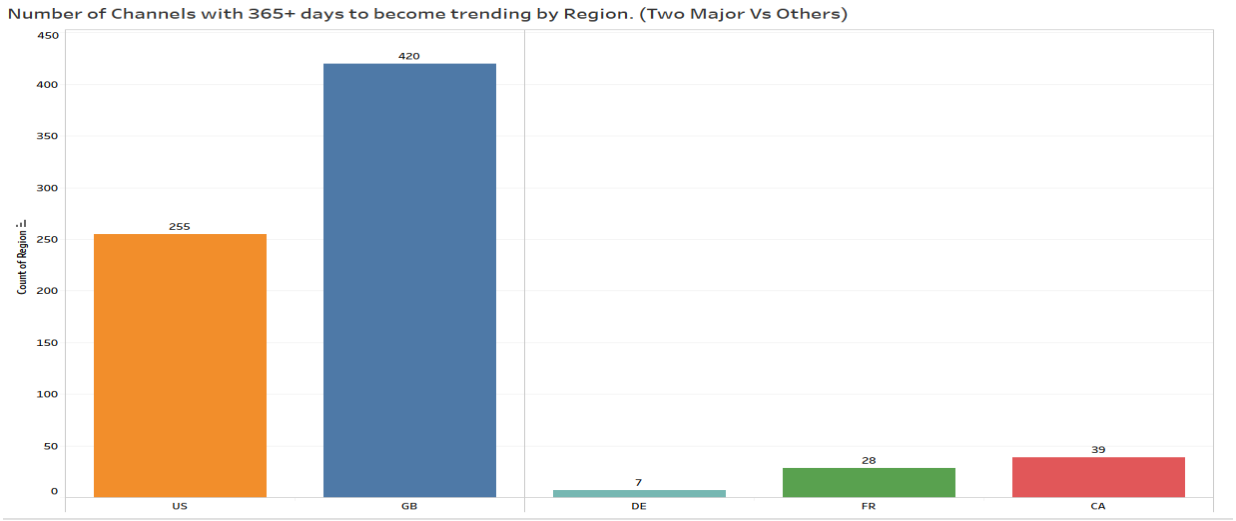
Our group decided to use Trending YouTube Videos dataset for our CSC 464 project. After we started our project, I have contributed the team by performing exploratory analysis, being responsible as class liaison and regularly checking in with Professor Brown regarding the direction of our project and collected in-depth feedback from the instructor after our team presentation.

In terms of visualization, I mainly used Tableau to create visualizations for the project by focusing on videos that have become trending after certain times after being published. There is a reason behind for doing this - the dataset contains all the videos that have become trending but solely highlighting videos with short duration to become trending does not seem to indicate the significance of using the dataset for our project.

Because of the above reason, I explored more into the dataset – by comparing regions and classifying outstanding videos, channels, and categories. After doing so, I could find several interesting facts by the following:

1. GB and US are the two major regions in the dataset. These two regions are dominant in terms of having cumulative views, comments, likes, and dislikes compare to other regions.

2. Along with their proportion in the dataset, the total number of channels that have spent longer than a year to have trending videos are found more in US and GB – aggregated number of channels under the condition for GB and US equal to 675, while the rest of regions have 74.

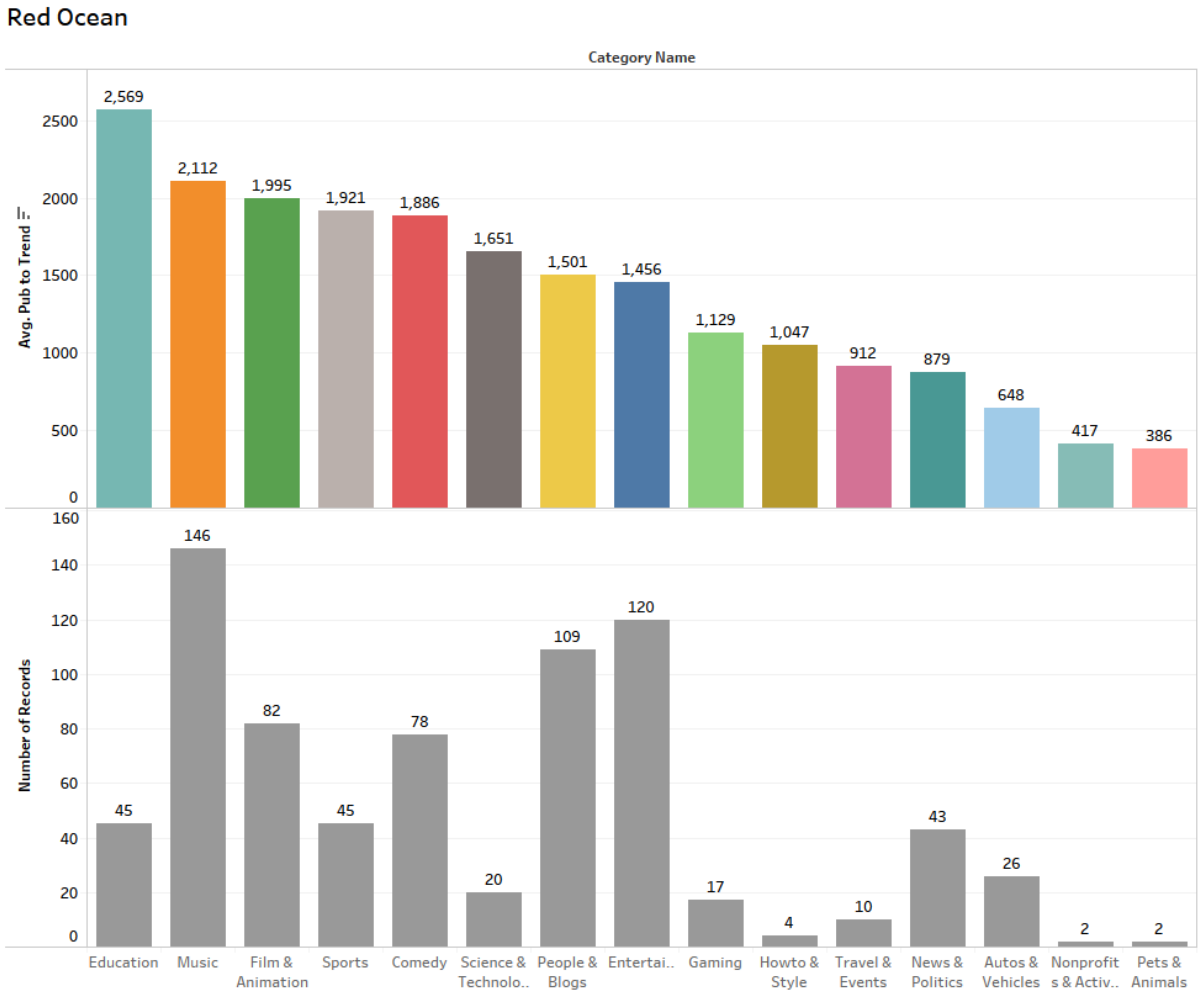



3. According to the visualizations made using Treemap, YouTube users mainly access to watch Entertainment, Music, and Comedy. Similar patterns were found in Likes, Dislikes, and Comments

# Appendix - Individual Reports

4. Contrast to its popularity, Music category needs 36 days on average to become trending. I could interpret this finding as 'Red Ocean' category in YouTube. Red Ocean defines a market with intensive competition and harder to capture existing demands against others.

5. Along with the above finding, I could find additional fact that Music, Entertainment, and People & Blogs are the top three categories that have outstanding channels (more than a year spent to become trending). 146 channels in Music, followed by Entertainment (120 channels), and People & Blogs (109 channels)




What I have learned from this project:

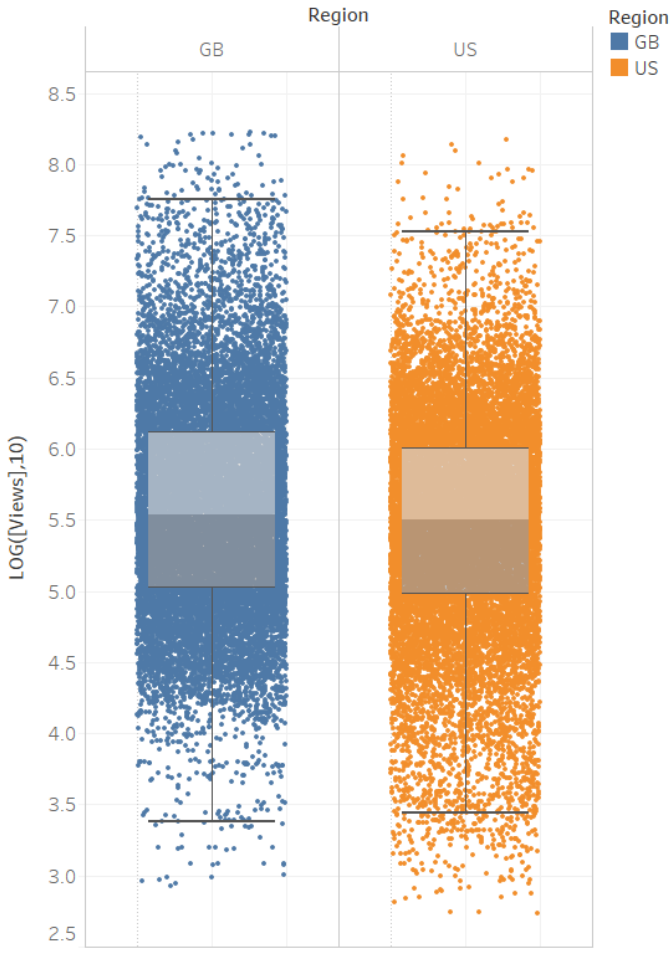
I could have learned a lot while participating for this project. To summarize, I could learn about how to visualize more effectively, exclude clutters to make visualizations more meaningful, the importance of sorting and grouping, have a moment thinking about how to drive the project toward altogether as a team, managing times for group meetings and provide appropriate feedbacks to team members regarding their work while listening to the feedbacks on my work as well. Additionally, using categorical and numerical variables altogether was not an easy step to take, but it seems that I have better understanding of visualizing these categories compare to the past.

# Appendix - Individual Reports

**Sungil Kim**

When I first explored the raw YouTube dataset, I noticed that there were a lot of categorical attributes. My first initial reaction was to avoid them, simply because I thought numerical variables were the most optimal for visualization purposes. However, I learned that the variable "tags" could be analyzed with text analysis. Since I only understand English out of all the regions, I specifically focused on comparing GB and US overall.

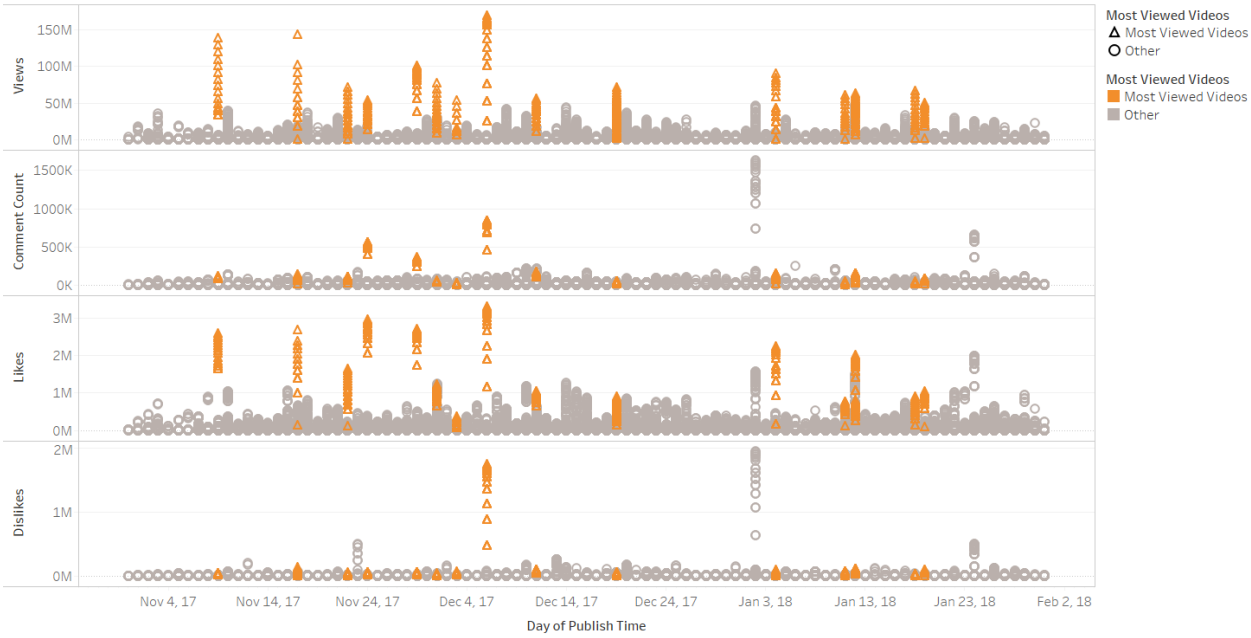
Box-and-whisker Plot of LOG10(Views) for GB and US




Random() vs. LOG([Views],10) broken down by Region. Color shows details about Region. The view is filtered on Region, which keeps GB and US.

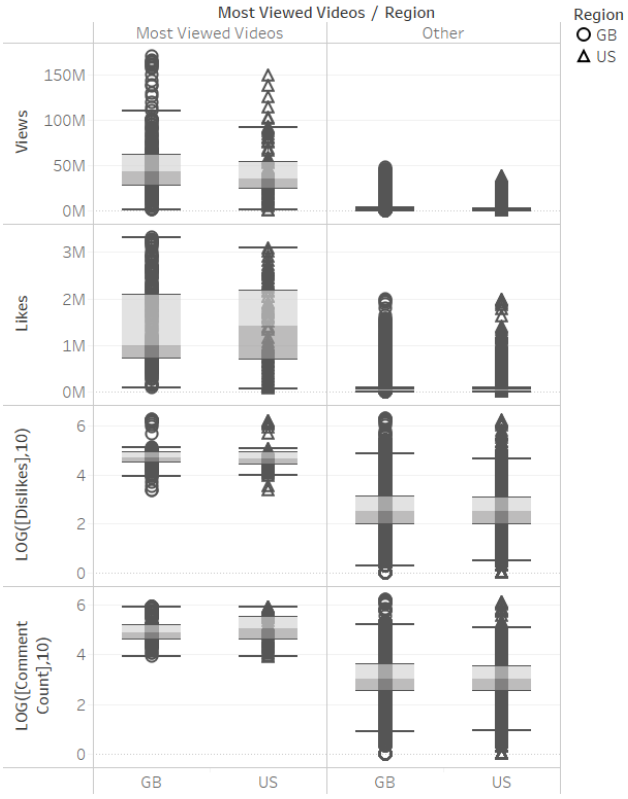
# Appendix - Individual Reports

Scatterplot of Views, Comment Count, Likes and Dislikes per Day from 10/31/2017 to 1/31/2018




The plots of Views, Comment Count, Likes and Dislikes for Day of Publish Time. Color shows details about Most Viewed Videos. Shape shows details about Most Viewed Videos. The view is filtered on Day of Publish Time, which ranges from October 31, 2017 to January 31, 2018.

Box-and-whisker Plot of Most Viewed Videos vs. Likes/Comment Count/Dislikes

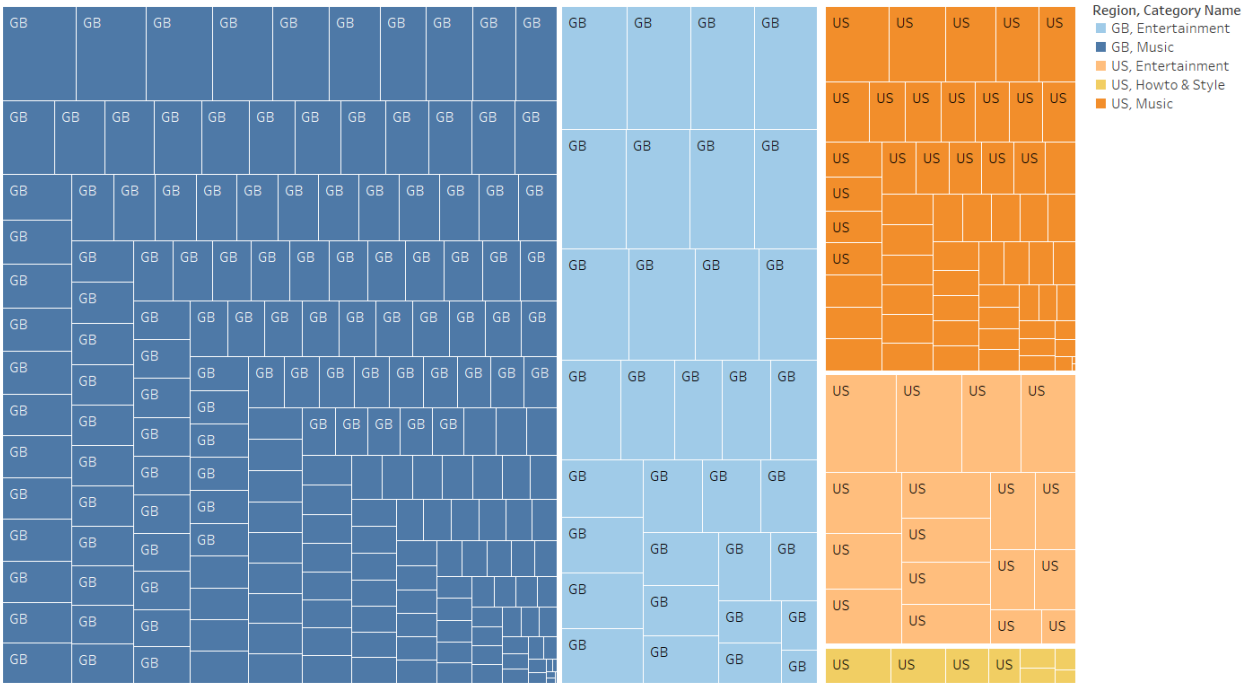



Views, Likes, LOG([Dislikes],10) and LOG([Comment Count],10) for each Region broken down by Most Viewed Videos. Shape shows details about Region. The view is filtered on Region, which keeps GB and US.

To start out, I explored the numerical variables first. I created a box-and-whisker plot of log10(views) for US and GB to look into the distribution of the views in US and GB. It turned out to be somewhat similar in terms of order of magnitude. After looking at the distribution of the views, I narrowed my focus

to most viewed videos (50 million views or higher) to see how the features of the top videos behave. Scatterplot of views, comment count, likes and dislikes was created to demonstrate the peaks of each variables after filtering to only include populated timeframes (10/31/2017 ~ 01/31/2018). The videos with 50 million views or higher were manually put into the Most Viewed Videos group. In order to analyze further with the numerical variables, I created a box-and-whisker plot of Most Viewed Videos' views/likes/comment count/dislikes per region. This graph also had a subdivision of Most Viewed Videos and Others for each region. The scatterplot and box-and-whisker plot illustrated that the videos with higher views tend to have higher likes, dislikes, and comment count. However, there were outliers to show that the linear relationship among views, likes, dislikes, and comment count is not always true.

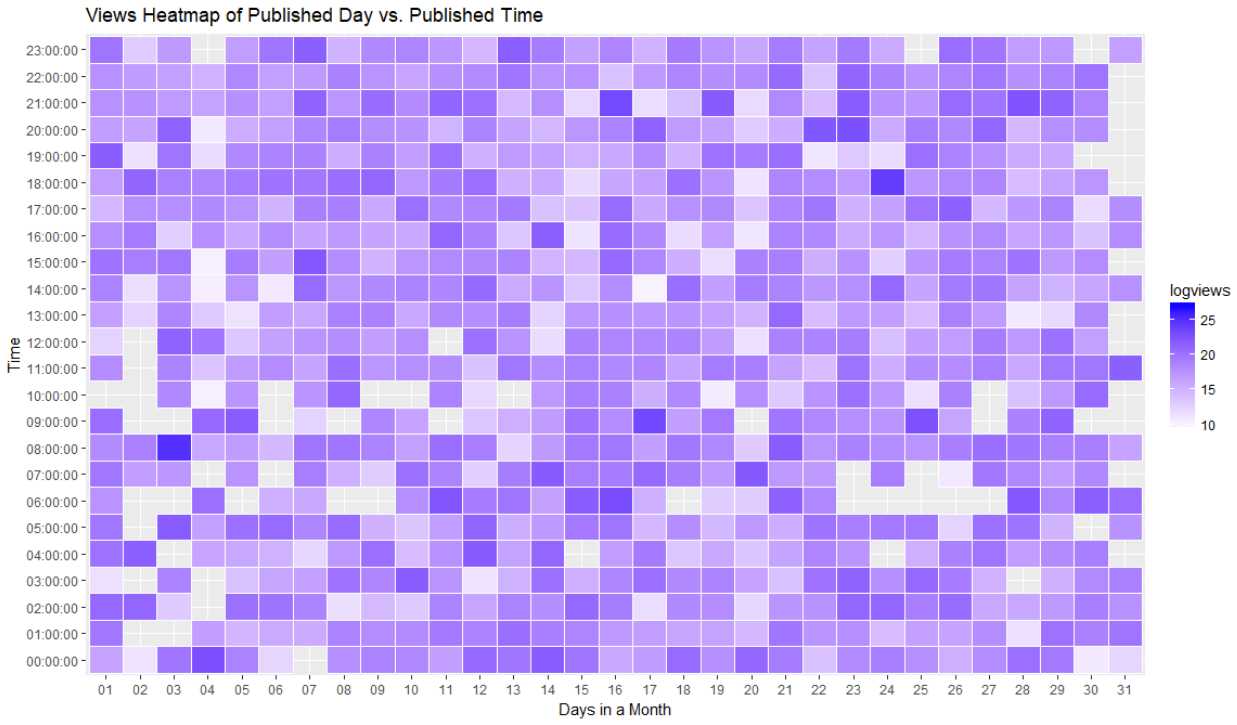
Treemap of Most Viewed Videos (above 50M views) per Region and Category

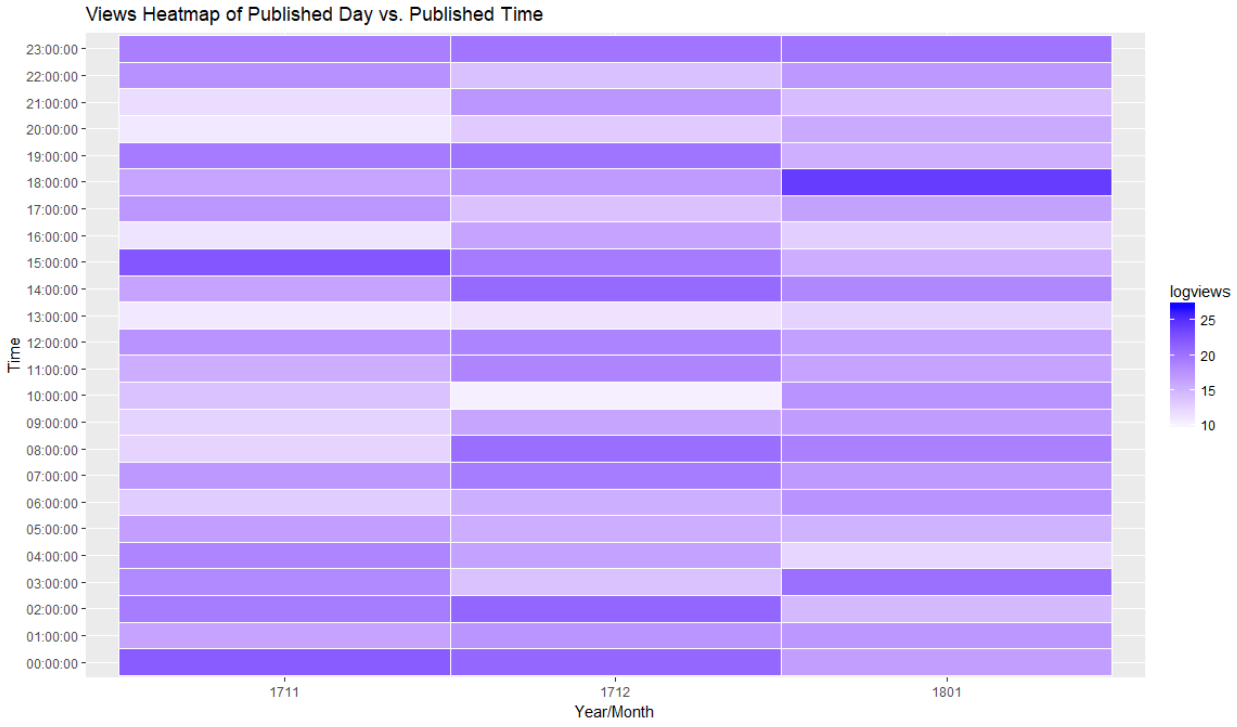



Region. Color shows details about Region and Category Name. Size shows Views. The marks are labeled by Region. The data is filtered on Most Viewed Videos, which keeps Most Viewed Videos. The view is filtered on Region, which keeps GB and US.

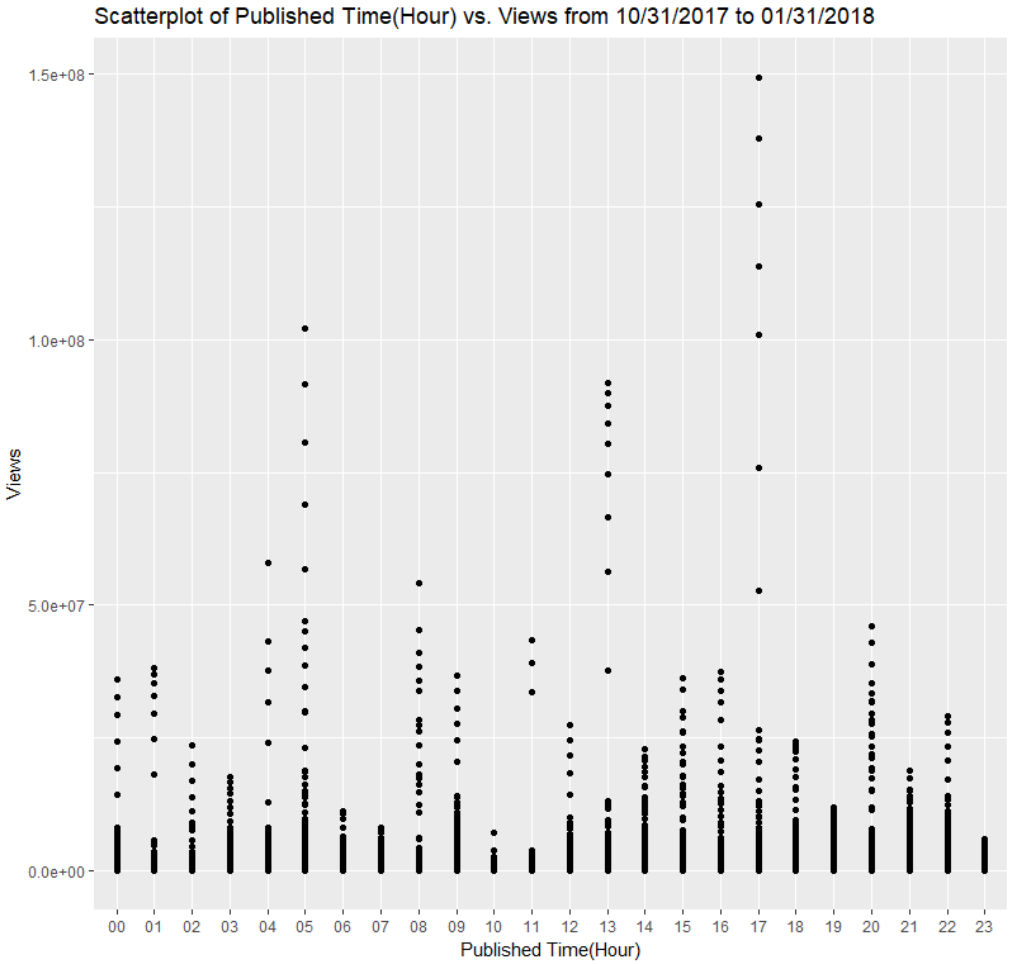
After playing around with views and other commonly related numerical variables, Treemap of Most Viewed Videos per region and category was developed for the purpose of portraying what categories of videos were popular per region. In US and GB, Entertainment and Music were the most popular categories. This graph however did not show a lot of information, so it was later replaced with a treemap of most popular category per region made by a group member.

# Appendix - Individual Reports



# Appendix - Individual Reports




The published day variable had a lot of preprocessing to do to be formatted into a timestamp. The published day was divided into different variables such as published day (in a month), year/month, hour, and so on to try different parameters. Logviews heatmap of Published Day in a month vs Published Time, Logviews heatmap of Published year/month vs Published Time, and scatterplot of Published Time (hour) vs. Views from 10/31/2017 – 01/31/2018 were developed to see the relationships between the published timestamp attribute and views. In the heatmaps and scatterplots, published time of 17, 5, 13 had the top three viewed videos. The views were the highest at day 3 around 8 - 9 am, day 24 around 6 - 7 pm, 3 - 4 pm in November 2017, and 6 - 7 pm in January 2018.

Then text analysis was utilized on the tags of US and GB Most Viewed Videos data (30 million and above to include more observations). I specifically chose sentiment and polarity analysis, to analyze what types of tags were involved in the top viewed videos. The tags had a lot of unnecessary characters, so the pre-processing part took a big chunk of the analysis. The tags involved in the sentiment analysis was later developed into a wordcloud per region, to deeply understand what kinds of tags were most frequently appeared. Detailed steps and results are included in the actual technical report under sentiment/polarity analysis and wordcloud visualization section.

In addition to writing my own analysis for the graphs in the final technical report and Appendix, I concluded the paper with proper analysis that supports our final story in the Analysis/Discussion section. This group project made me learn a lot about categorical data. Before this project, I believed that numerical data was the most useful in terms of performing analysis with or on. However, this YouTube project helped me understand the importance of considering categorical features. In the YouTube data alone, there were many categorical variables that could be analyzed differently, such as tags, description, categories, and so on. Categorical features had crucial roles in data visualization especially. They were used with color, shapes and even filtered to represent discrete axis. This project ultimately introduced me to unlimited data visualization possibilities by showing the importance of categorical variables.

# Appendix - Individual Reports

**Jacob Penrod**

# Data Cleaning

The initial data cleaning had a few steps. The data came in separate files by region and categories in a separate JSON file. I first added a column to represent the region (from the file name) [US,GB,CA,DE,FR] then I concatenated the files into one big file. I converted the JSON to csv and joined it to the original data based on the Category ID field that already existed in the data. After that I performed LDA (see section below).

# LDA

**Process**

Remove some common internet formatting – removing http:// and some common symbols used online. I created a list of words known as a Corpus, then I removed English and German stop words from that Corpus. Even though I restricted to English regions, German still showed up quite a bit. After that I created a and Document Terms Matrix and performed the LDA algorithm where k=5. I then assigned topics based on distance from the chosen topics to have a categorical variable to use.

**LDA Topics**

LDA Generated Topics

| late | christmas | tmz | images | jimmy |
|---|---|---|---|---|
| songs | nand | time | gey | movie |
| bbc | game | sports | patreon | song |
| top | out | voice | black | today |
| vlog | days | house | morning | tonighshow |
| **BBC Music** | **Holidays and Gaming** | **Celebrity Gossip** | **Late Night TV** | **Independent** |

From Left to right in the chart above

Topic 1 - Seems to be mostly music and specifically mentions BBC, so we call it BBC Music.
Topic 2 – Mentions Christmas, days out and games, so we call it Holidays and Gaming
Topic 3 - All about reality shows, sports and TMZ (a gossip publication), so we call it Celebrity Gossip
Topic 4 – Very little connection in terms but mentions Patreon which is very big in the independent channels (not corporate owned) so we call it Independent.
Topic 5 - mentions Jimmy [Fallon], tonight show and other interview topics, so we call it Late Night TV.
Once we have these assigned, we can use it in our visualizations.

# Hybrid Heat map and Bar chart

I made this chart in Tableau by combining 3 different charts in a dashboard. The first is a heat map of the Average of Log Views by Published months. We used the topics we discovered by a process known as Latent Dirichlet Allocation (LDA) on the Video descriptions of all the English regions (US,GB,CA).

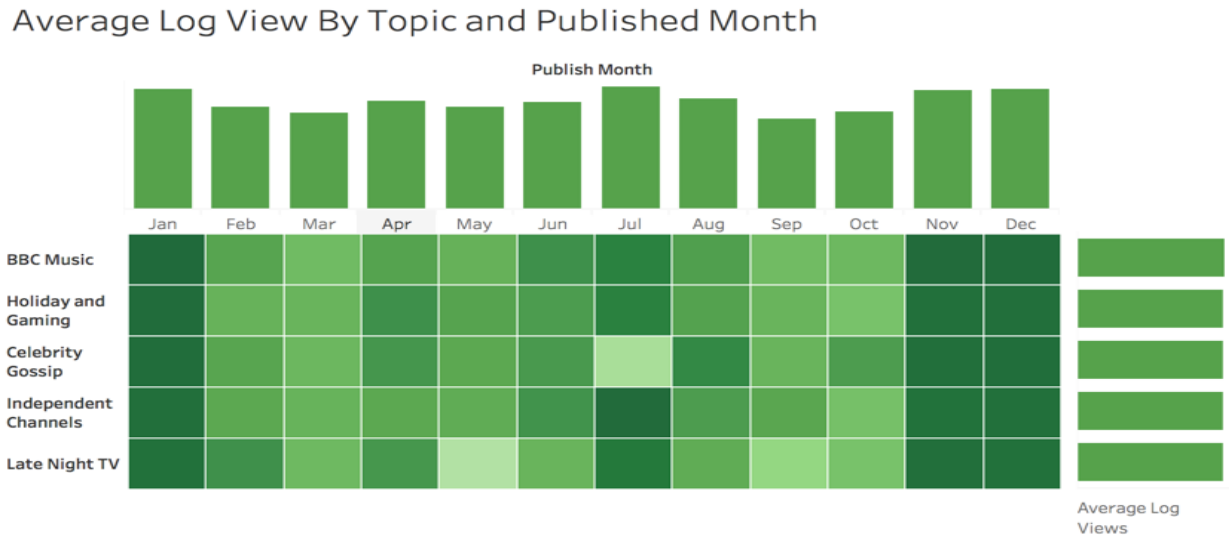
This chart shows that Holiday and Gaming has a low average view rating in July. Average views are up in general around January, November and December. Late Night TV had an off month in May and September, these months may represent changes in normal

scheduling as May usually marks the beginning of Summer break and September is back to school. Holiday and gaming has very good average numbers in November, December and January for obvious reasons. It has decent numbers in April (Easter) and July (Summer Holidays). Independent channels seem to be pretty consistent.

My process for this graph started with a simple heat map in R, and I started to look around for date ranges that showed consistent patterns. I started with trending months, but it turned out that the data only had three distinct months, so then I looked around at trending days and it looked really busy to me. I settled on published month as it had the full range of data and showed patterns that aligned with the LDA derived topics. I then picked green saturation for the heat map and green for the bar chart, it was originally blue, but green was chosen to help synchronize with the rest of our visualizations and it is easy to see different shades of green as we observed.

I also worked with different aggregates, and the sum total log view was very skewed in terms of published month toward November, December and January that I thought the visual would not be very interesting, so I decided on average over median even though they were very similar.

I then added the bar chart to add some overall context after observing the chart in other group presentations and I think it adds a good summary of the data for each row and column of the heat map. I then sorted the topics ascending by average log views so it would make the bar chart make sense, and left the months in their standard order.




# What I Did in the Project

I posted the original dataset, setup the Slack which was used for group coordination and discussions. I submitted the initial project survey. I helped to lead group discussions and delegate tasks. I did the initial data cleaning, performed LDA and LDA visualizations. I used R for my data pre-processing and for my initial designs, until I saw the hybrid heat map and bar chart that was used in one of the presentations. To make the visualization, I used Tableau to make a dashboard, as I thought it looked better than the R hybrid models I was finding online. I learned a lot about collaboration and coordination in this group project and how important communication portals such as slack and google drive can be for remote workers. This was my first mostly remote group and it had a very different feel to it, but I feel it was valuable experience for a world that is becoming more remote. I also learned it is important to start more difficult parts sooner, as I had some late nights trying to fully integrate the LDA data into the original set due to foreign languages and I should have had more time to perfect the process.

# R Code

# Appendix - Individual Reports

```
library(tidyverse)      # data manipulation & plotting
library(stringr)        # text cleaning and regular expressions
library(tidytext)
library(magrittr)
library(dplyr)

library(NLP)
library(topicmodels)
library(tm)

library(wordcloud)
library(slam)
library(topicmodels)

#For Data cleaning I mostly used Excel and MacOS terminal Commands
#I Added the region codes based on the file names
#cat GBvideos.csv USvideos.csv FRvideos.csv DEvideos.csv CAvideos.csv >
JoinedVideos_full.csv
#After that I used an online service to translate the JSON categories to a CSV (all the regions
have the same names)
library(readr)
#Import Categories - All regions have the same IDs and Descriptions, they only differ in meta
data we aren't using
catdf <- read_csv("~/Google Drive/DePaul/CSC465/youtube-new/US_category_id.csv")
#Import YouTube Dataset post Data Clean
ytraw <- read_csv("Google Drive/DePaul/CSC465/youtube-new/JoinedVideos_full.csv")
yt <- merge(ytraw,catdf,by=c("category_id","category_id"))

#Filter down by region
yt_eng <- yt[yt$region == 'US' | yt$region == 'GB' | yt$region == 'CA',]

#Filter out common problem strings, punctuation and whitespace
docs = gsub('(RT|via)((?:\\b\\W*@\\w+)+)','',yt_eng$description)
docs = gsub('http[^[:blank:]]+', '', docs)
docs = gsub('@\\w+', '', docs)
docs = gsub('[ t]{2,}', '', docs)
docs = gsub('^\\s+|\\s+$', '', docs)
docs <- gsub('\\d+', '', docs)
docs = gsub('[[:punct:]]', ' ', docs)
corpus = Corpus(VectorSource(docs))
corpus = tm_map(corpus,removePunctuation)
corpus = tm_map(corpus,stripWhitespace)
corpus = tm_map(corpus,tolower)
corpus = tm_map(corpus,removeWords,stopwords('english'))
corpus = tm_map(corpus,removeWords,stopwords('german'))
#these seem redundant with category
corpus = tm_map(corpus,removeWords,c('music','movie'))
# Creating a Term document Matrix
tdm = DocumentTermMatrix(corpus)
# create tf-idf matrix
term_tfidf <- tapply(tdm$v/row_sums(tdm)[tdm$i], tdm$j, mean) *
log2(nDocs(tdm)/col_sums(tdm > 0))
tdm <- tdm[,term_tfidf >= 0.1]
```

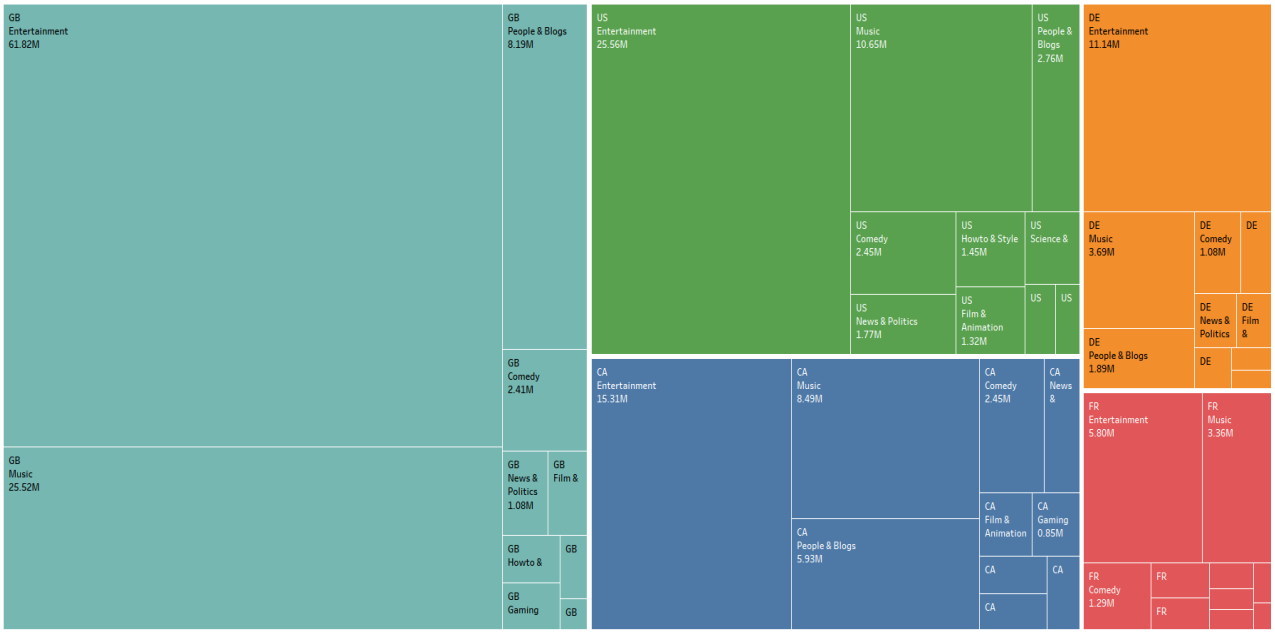
```
tdm <- tdm[row_sums(tdm) > 0,]
tdm <- tdm[,term_tfidf >= 0.1]
tdm <- tdm[row_sums(tdm) > 0,]
#Deciding best K value using Log-likelihood method
best.model <- lapply(seq(2, 5, by = 1), function(d){LDA(tdm, d)})
best.model.logLik <- as.data.frame(as.matrix(lapply(best.model, logLik)))
#calculating LDA
k = 5;#number of topics
SEED = 786; # Seed for reproducibility

l1 = LDA(tdm, k = 5, method = "Gibbs",control = list(seed = SEED, burnin = 1000,thin = 100, iter
= 1000))
l1.terms <- as.matrix(terms(l1,5))
#view the topic assignment for each document
tops <- topics(l1)
l1_topics <-as.matrix(topics(l1))
write.csv(l1_topics,file=paste("LDAGibbs",k,"DocsToTopics.csv"))
#Make an ID for joining
yt_eng$ID <- seq.int(nrow(yt_eng))
l1_topics$ID <- seq.int(nrow(l1_topics))
#Read in the IDs we just created
#LDA takes a while, so I added a check point
lda5 <- read_csv("Google Drive/DePaul/CSC465/youtube-new/LDAGibbs 5 DocsToTopics.csv")
#Join to main dataset
yt_eng$ID <-yt_eng$id
yt_eng <- merge(yt_eng,lda5,by=c("ID","ID"))

yt_eng$TopicID <- as.factor(yt_eng$TopicID)
yt_eng$loglikes <- log(yt_eng$likes)
yt_eng$logdislikes <- log(yt_eng$dislikes)
yt_eng$logviews <- log(yt_eng$views)

yt_eng$publish_month <- format(as.Date(yt_eng$publish_time,format="%m/%d/%Y"),"%m")
```

# Appendix - Exploratory Analysis

Code for visualizations can be found in this [Google Drive folder.]

The following visualizations represent some early exploratory analysis of the data discussed within this paper.

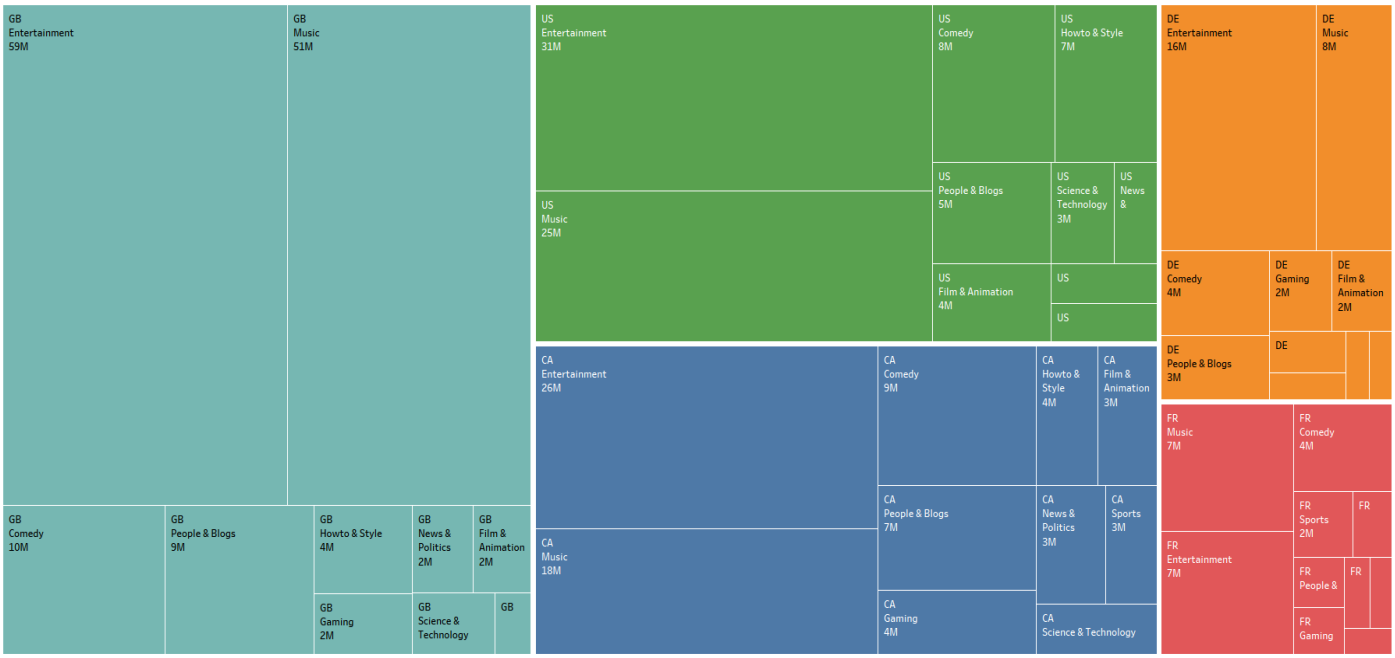
Top Dislike Categories from each region



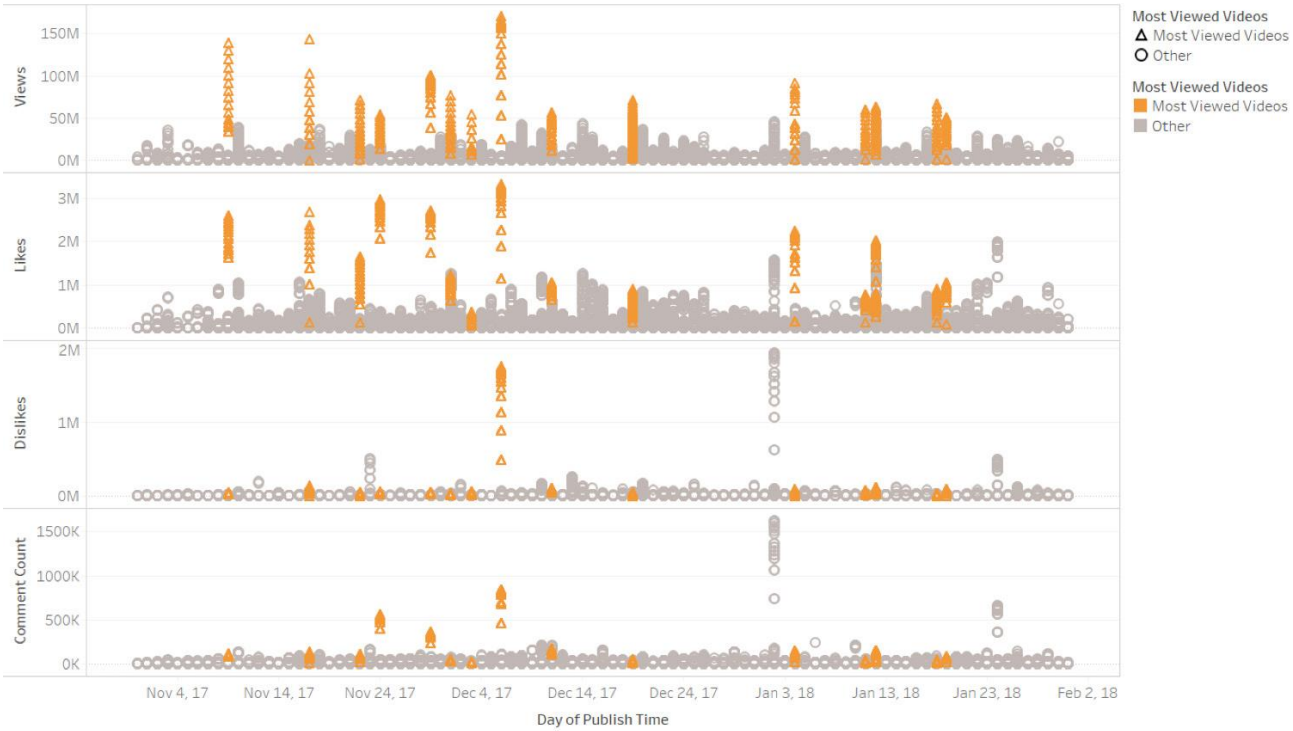

Top View Categories from each region

# Appendix

Top Commented Categories from each region



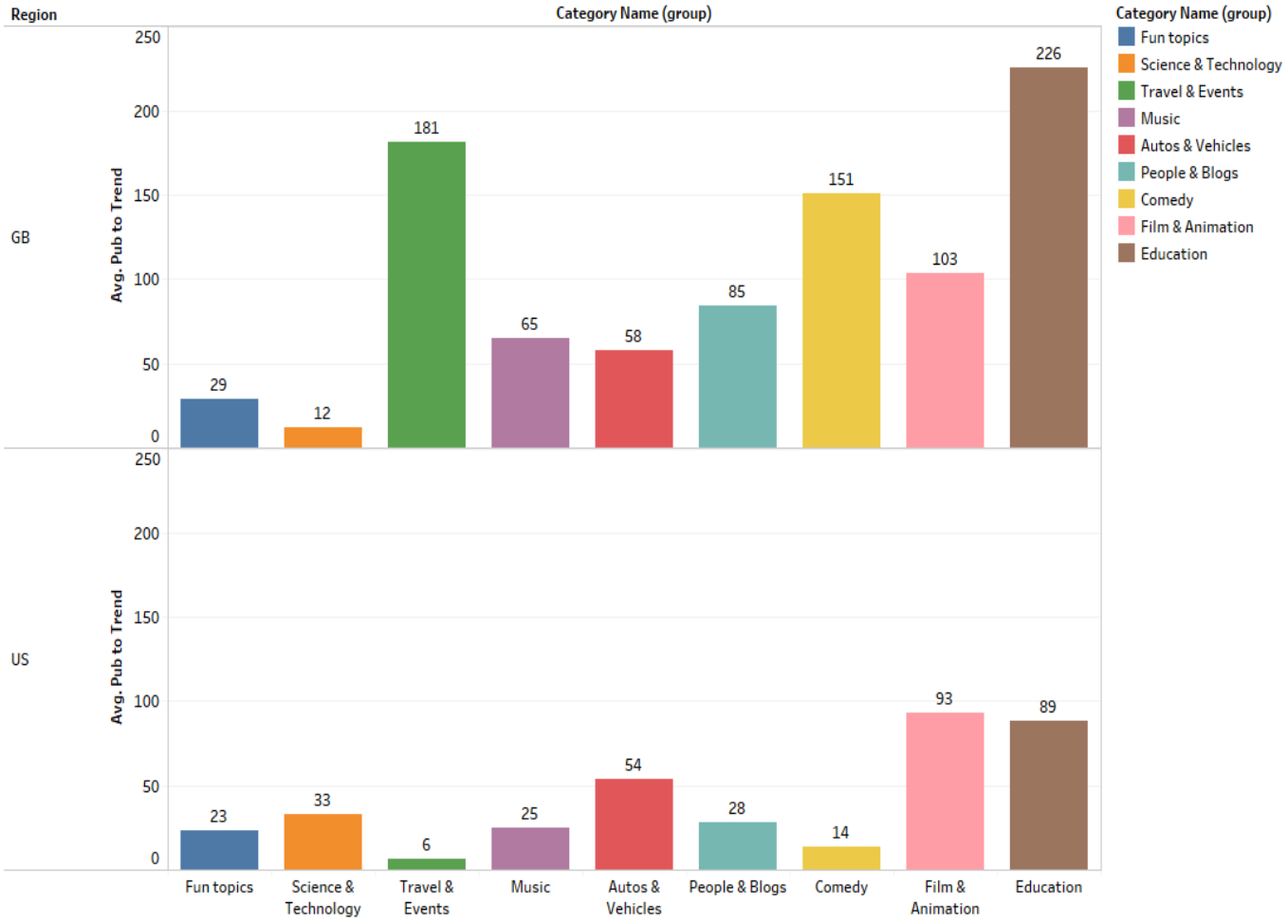

Scatterplot of Views, Comment Count, Likes and Dislikes per Day from 10/31/2017 to 1/31/2018



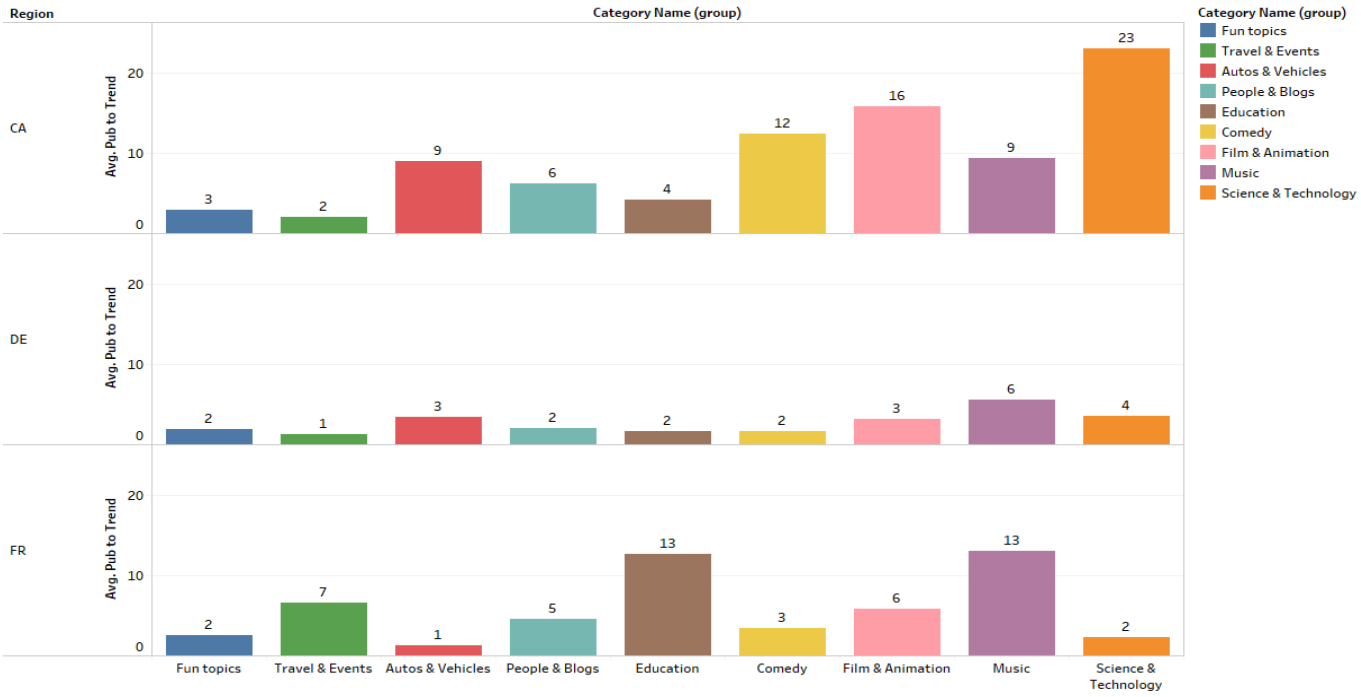

The plots of Views, Likes, Dislikes and Comment Count for Day of Publish Time. Color shows details about Most Viewed Videos. Shape shows details about Most Viewed Videos. The view is filtered on Day of Publish Time, which ranges from October 31, 2017 to January 31, 2018.

# Appendix

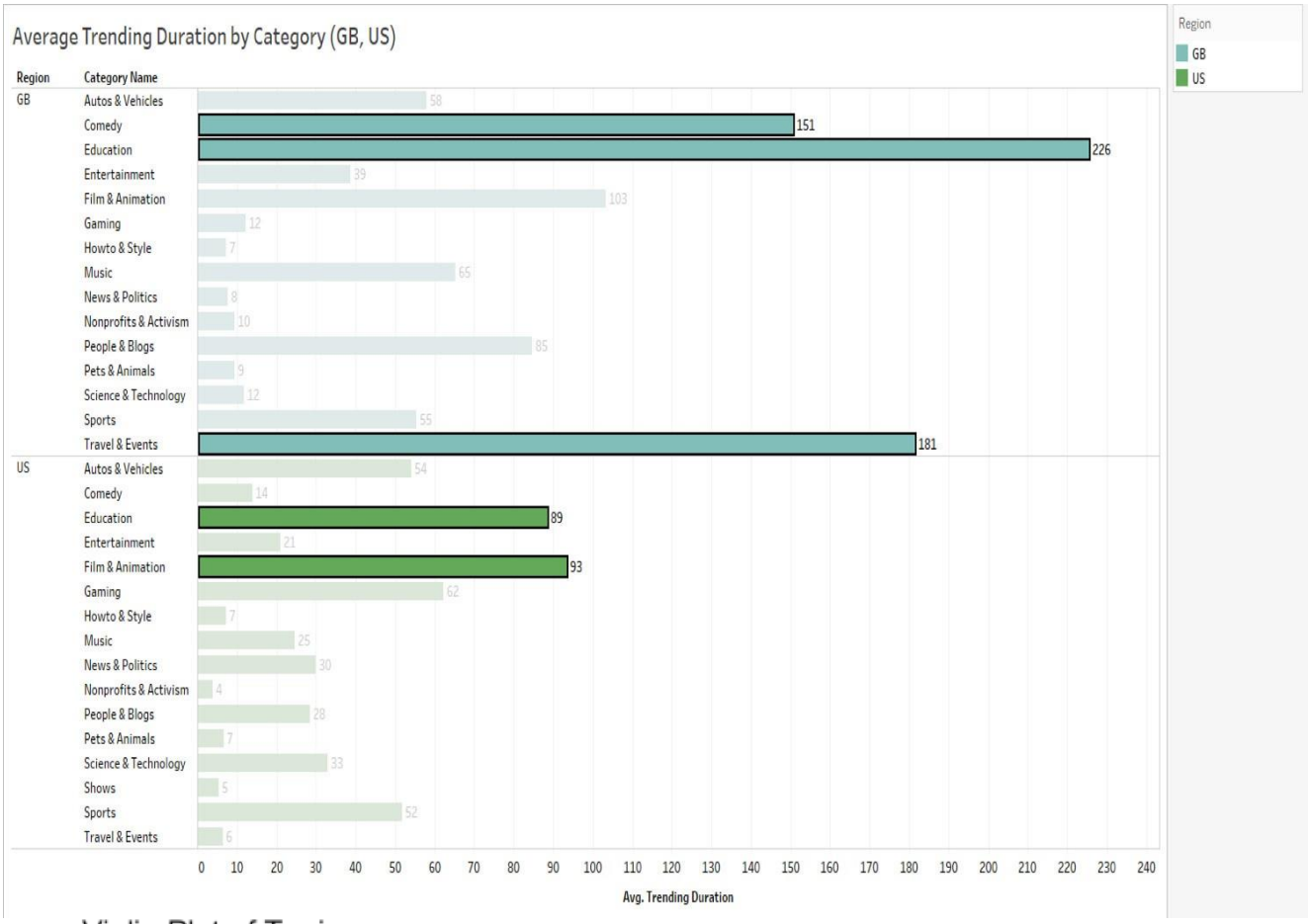
## Topic: Fun vs Others (US VS GB)

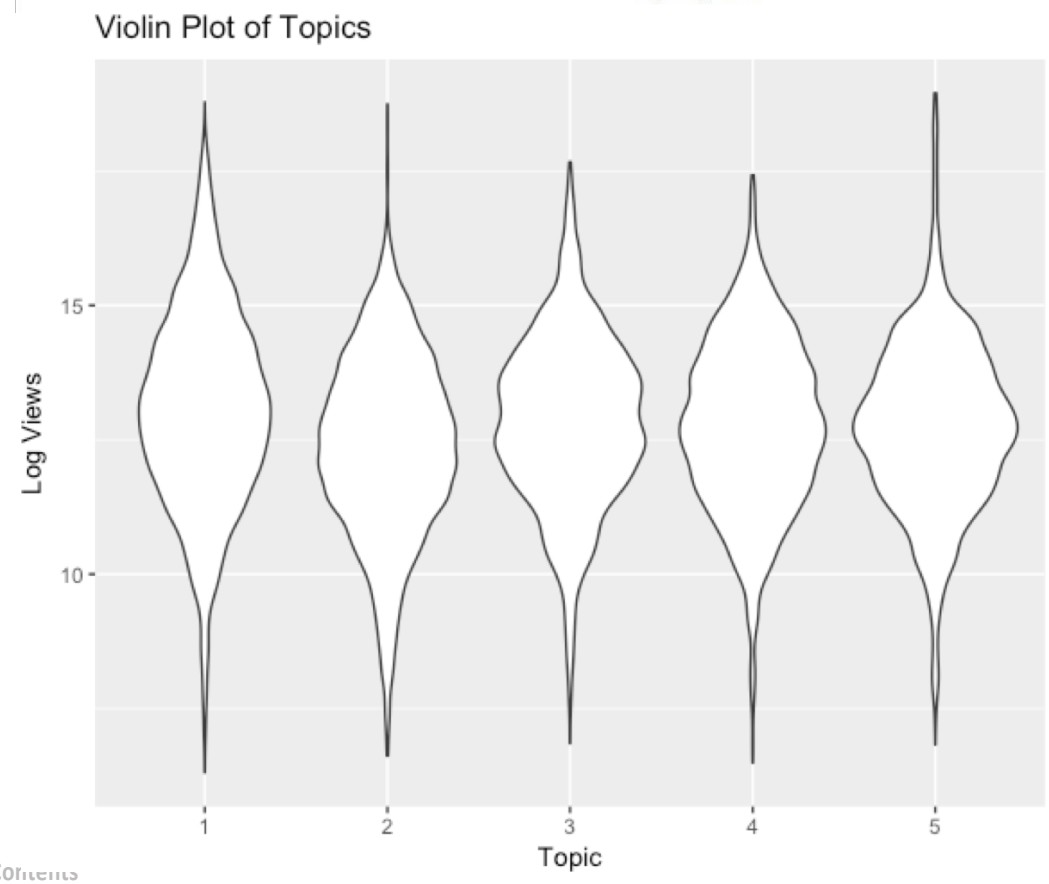



## Topic: Fun vs Others (CA vs DE vs FR)



# Appendix


Average Trending Duration by Category (GB, US)
Region
Category Name
GB
Autos & Vehicles 58
Comedy 151
Education 226
Entertainment 39
Film & Animation 103
Gaming 12
Howto & Style 7
Music 65
News & Politics 8
Nonprofits & Activism 10
People & Blogs 85
Pets & Animals 9
Science & Technology 12
Sports 55
Travel & Events 181
US
Autos & Vehicles 54
Comedy 14
Education 89
Entertainment 21
Film & Animation 93
Gaming 62
Howto & Style 7
Music 25
News & Politics 30
Nonprofits & Activism 4
People & Blogs 28
Pets & Animals 7
Science & Technology 33
Shows 5
Sports 52
Travel & Events 6
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200 210 220 230 240
Avg. Trending Duration
Region
GB
US


Violin Plot of Topics
Log Views
15
10
1
2
3
4
5
Topic