YouTube Trends of Countries Around the World

STAT 3355.001 Project

Group 10 Faaiz Nadeem Gabe Marchant

11/16/2020

Introduction

The purpose of this project was to utilize R and practice using the skills we learned in class to analyze a large set of data. Our group used a data from Kaggle by Mitchell J called "Trending Youtube Video Statistics"[1]. This dataset contained several months of data on daily YouTube videos in 10 different countries. The countries included USA, Great Britain, Germany, Canada, France, Russia, New Mexico, South Korea, Japan, and India from 2017-2018. The dataset contains over 375,000 observations on variables describing the videos including:

- 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

Seeing that YouTube is one of the biggest social media platforms in the world with close to 2 billion users since 2019, our group thought it would be interesting to see the trends on YouTube around the world. Questions we had immediately upon looking at the dataset that we pushed to answer in this report include:

- Which categories are most trendy around the world?
- Are some categories trending more in certain countries than in others?
- Have category interests shifted over time in all countries from 2017 - 2018?
- Does controversy (like/dislike ratio/ comments, etc.) have a strong effect over trends?
- Are certain categories more popular than others around the world?
- Do specific channels/creators trend more in other countries?

Data Cleaning

Our group began our exploratory data analysis by taking the csv files of the individual data sets of the countries and deciding which countries would be best to look at for this project. We decided that combining countries multiple countries from at least one continent would give us more than enough data to discover what we were looking for. Because of this, we chose the datasets for the United States, Mexico, Canada, Germany, France, Great Britain, Russia, India and Japan. This way, we had three countries per region by continent, North America, Europe and Asia, respectfully.

Afterwards, in order to find universal trends across all of the countries, we merged all the countries we selected into a new data set. This way we could directly compare trends universally across all the countries. Next thing we worked on was change the category ids from numerical values to their actual names according to the respective json file of the country. Names of the new categories included:

- Autos and Vehicles
- Comedy
- Education
- Entertainment
- Film and Animation
- Gaming
- How to Style
- Movies
- Music
- News and Politics
- Nonprofits and Activism
- People and Blogs
- Pets and Animals
- Science and Technology
- Shows
- Sports
- Trailers
- Travel and Events

To summarize, the process of data cleaning in this project was not difficult. After we were able to rename the categories and create another data set that combined all the other countries together, it was simple to continue our analysis of the data. The only tedious part of the process was renaming the categories for all the individual data sets from their ids to their proper names.

```
# Read in the Data, and Create country variable
US <- read.csv("USvideos.csv")
US[1:40949, 'Country'] = 'US'

RU <- read.csv("RUvideos.csv")
RU[1:40739, 'Country'] = 'RU'

IN <- read.csv("INvideos.csv")
IN[1:37352, 'Country'] = 'IN'

MX <- read.csv("MXvideos.csv")
MX[1:40451, 'Country'] = 'MX'

GB <- read.csv("GBvideos.csv")
GB[1:38916, 'Country'] = 'GB'

DE <- read.csv("DEvideos.csv")
DE[1:40840, 'Country'] = 'DE'

CA <- read.csv("CAvideos.csv")
CA[1:40881, 'Country'] = 'CA'

FR <- read.csv("FRvideos.csv")
FR[1:40724, 'Country'] = 'FR'

JP <- read.csv("JPvideos.csv")
JP[1:20523, 'Country'] = 'JP'
```

```
# Create one DF with all countries data
youtube = rbind(US, CA, GB, DE, FR, IN, JP, RU, MX)
```

```
# Converting numerical category identifier to character description
youtube$category_id[youtube$category_id == 1] = 'Film & Animation'
youtube$category_id[youtube$category_id == 2] = 'Autos & Vehicles'
youtube$category_id[youtube$category_id == 10] = 'Music'
youtube$category_id[youtube$category_id == 15] = 'Pets & Animals'
youtube$category_id[youtube$category_id == 17] = 'Sports'
youtube$category_id[youtube$category_id == 18] = 'Short Movies'
youtube$category_id[youtube$category_id == 19] = 'Travel & Events'
youtube$category_id[youtube$category_id == 20] = 'Gaming'
youtube$category_id[youtube$category_id == 21] = 'Videoblogging'
youtube$category_id[youtube$category_id == 22] = 'People & Blogs'
youtube$category_id[youtube$category_id == 23] = 'Comedy'
youtube$category_id[youtube$category_id == 24] = 'Entertainment'
youtube$category_id[youtube$category_id == 25] = 'News & Politics'
youtube$category_id[youtube$category_id == 26] = 'Howto & Style'
youtube$category_id[youtube$category_id == 27] = 'Education'
youtube$category_id[youtube$category_id == 28] = 'Science & Technology'
youtube$category_id[youtube$category_id == 29] = 'Nonprofits & Activism'
youtube$category_id[youtube$category_id == 30] = 'Movies'
youtube$category_id[youtube$category_id == 43] = 'Shows'
youtube$category_id[youtube$category_id == 44] = 'Trailers'
```

```r
#Some useful variables for data analysis
youtube$like_to_dislike_ratio = youtube$likes / youtube$dislikes
youtube$trend_month = substr(youtube$trending_date, 7, 8)
youtube$trend_day = substr(youtube$trending_date, 4, 5)
youtube$trend_year = substr(youtube$trending_date, 1, 2)
youtube$post_month = substr(youtube$publish_time, 6, 7)
youtube$post_year = substr(youtube$publish_time, 3, 4)
youtube$post_day = substr(youtube$publish_time, 9, 10)
youtube$likes_per_view = youtube$likes / youtube$views
youtube$dislikes_per_view = youtube$dislikes / youtube$views
youtube$ratings_per_view = (youtube$likes + youtube$dislikes) / youtube$views
youtube$ratings = (youtube$likes + youtube$dislikes)
youtube$comments_per_view = youtube$comment_count / youtube$views
```

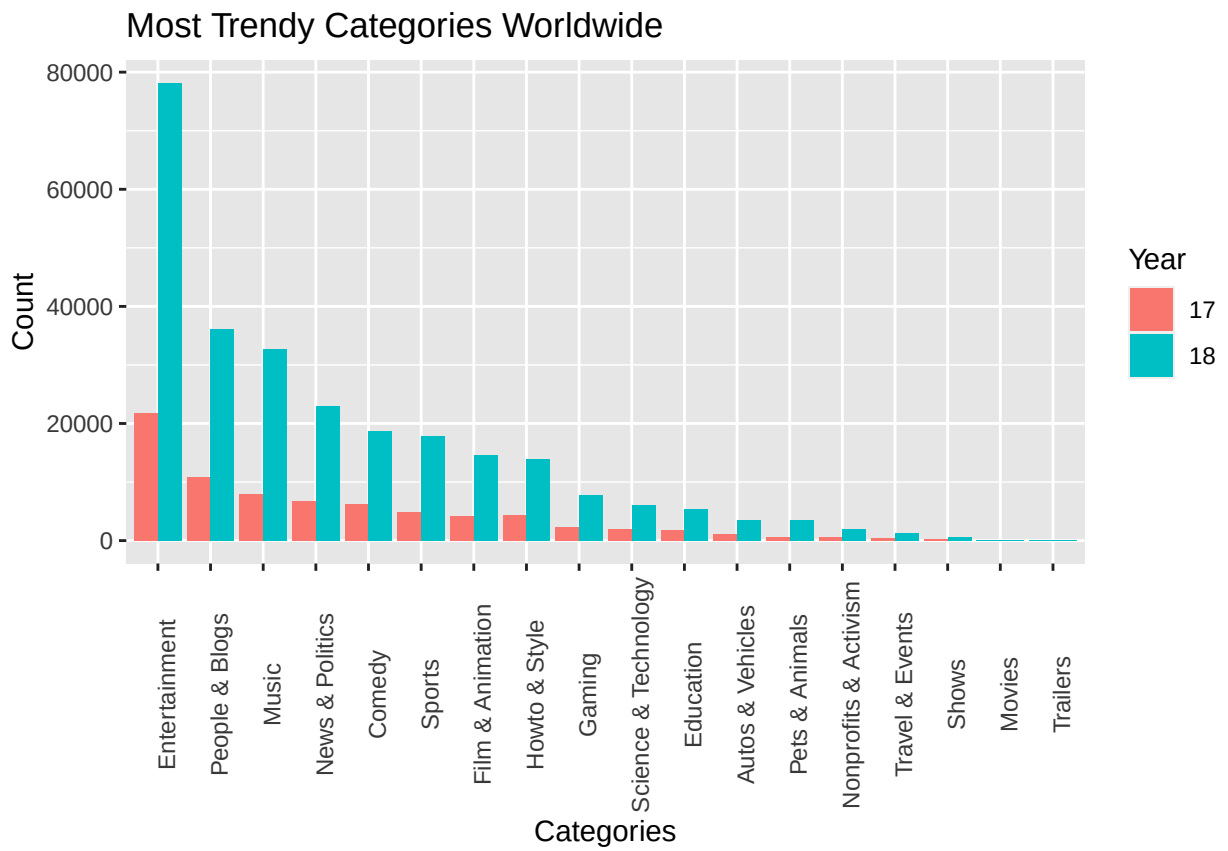
Analysis and Findings

Throughout this section of the report, we will discuss our findings of the data based on the questions we had in the beginning. We also included other findings that we thought would be interesting when it came to the trends. Questions throughout the report that drew more interest from us will have more detail.

Top Trending Categories

There are a few things we can tell by the graphic below. Firstly to answer one of our initial questions, Entertainment by far is the most occuring category among trending videos worldwide. The categories 'People & Blogs' and 'Music' follow. Also, we see that overall we have much less data from 2017 than 2018, but as we can see in respect to popularity of category, the two years are following similar trends.

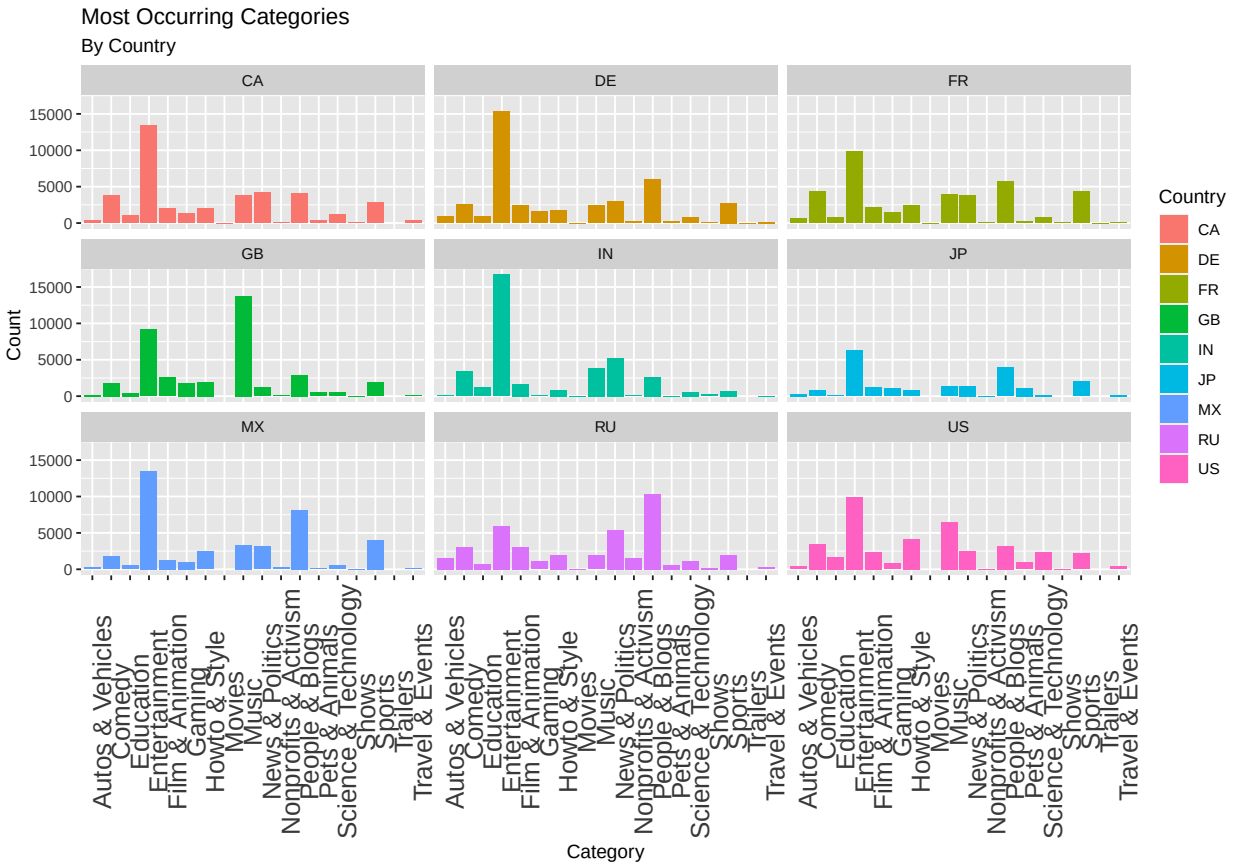
```r
library(ggplot2)
ggplot(data = youtube) +
  geom_bar(mapping = aes(x = as.factor(category_id), fill = trend_year), position = 'dodge') +
  scale_x_discrete(limits = c('Entertainment', 'People & Blogs', 'Music', 'News & Politics',
                              'Comedy', 'Sports',
                              'Film & Animation',
                              'Howto & Style', 'Gaming', 'Science & Technology', 'Education',
                              'Autos & Vehicles',
                              'Pets & Animals','Nonprofits & Activism','Travel & Events',
                              'Shows','Movies', 'Trailers')) +
  theme(axis.text.x = element_text(angle = 90)) +
  labs(x = 'Categories', y = 'Count', fill = 'Year', title = 'Most Trendy Categories Worldwide')
```



However, on the opposite side of the spectrum we were also able to see what categories trended the least in all the countries. These included shows, movies, and trailers, with movies and trailers having close to 0 total videos on the trending page. Reasons for this can include users having separate streaming services for shows and movies such as Netflix or Hulu.

Trending Categories across Countries

```
ggplot(data = youtube) +
  geom_bar(mapping = aes(x = as.factor(category_id), fill = Country)) +
  theme(axis.text.x = element_text(angle = 90, size = 15)) +
  facet_wrap(.~Country) +
  labs(x = 'Category', y = 'Count', title = 'Most Occurring Categories',
       subtitle = 'By Country')
```

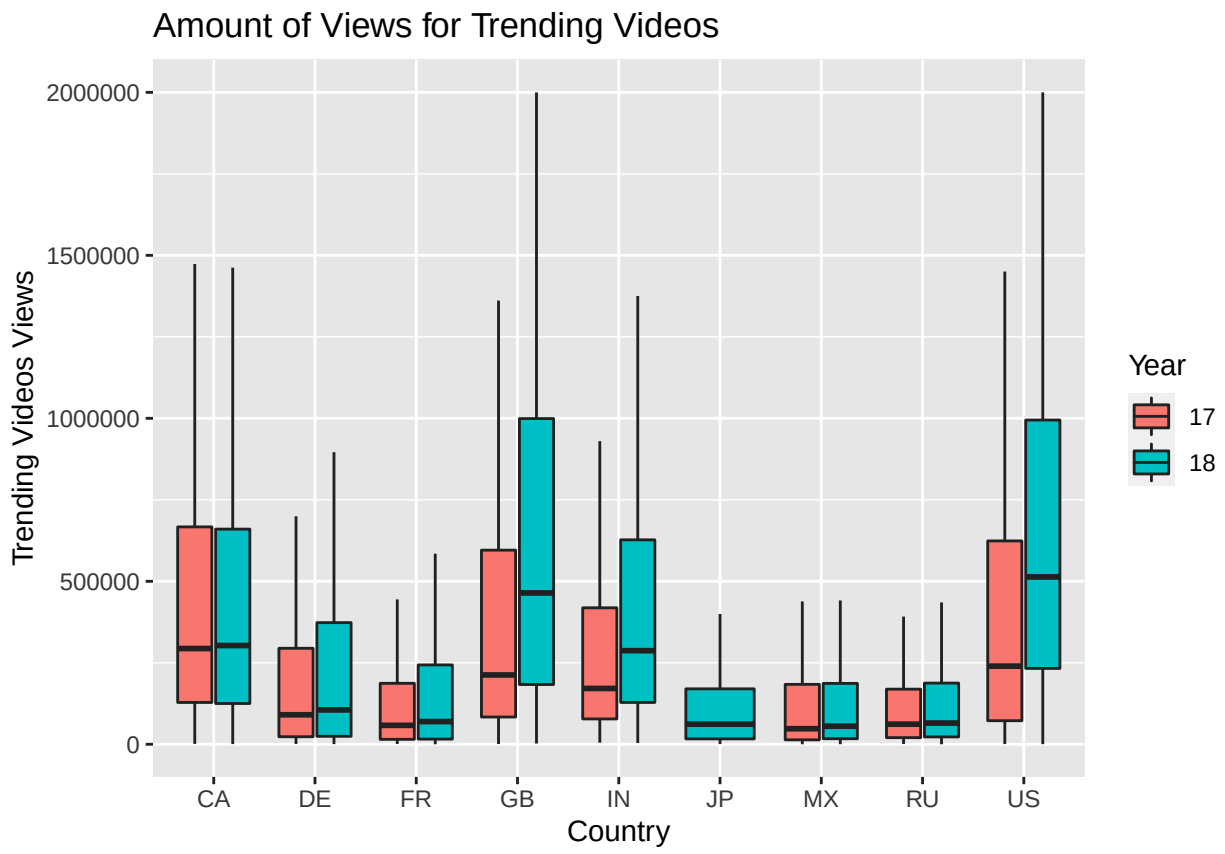


After looking at the categories of YouTube in a universal lens, we wanted to see how the categories were looking based on the country itself. Several commonalities were found as soon as we did this. 77% of the countries had entertainment as their highest video count for the trending page. The remaining 23% (Great Britain and Russia) had Music and People and Blogs as their highest trending category respectively. 66% of the countries had People and Blogs as the second highest trending category. Movies and Shows stayed consistently the lowest trending across all the countries while other categories such as sports and politics varied per country.

Viewership Across Countries

By looking at the graphic below we can see the average amount of views for a trending video in each respective country. From 2017 to 2018 the average amount of views for trending videos in Great Britain, India, and the United States increased by about 250,000 views each, while the other countries stayed the same. Does this mean that YouTube usage grew from 2017 to 2018? Not necessarily there is not enough here to make that conclusion. Maybe YouTube changed the criteria for trending videos in these countries, making a higher threshold for view count. There are many other factors that may attribute to this, but we do know the trend is there.

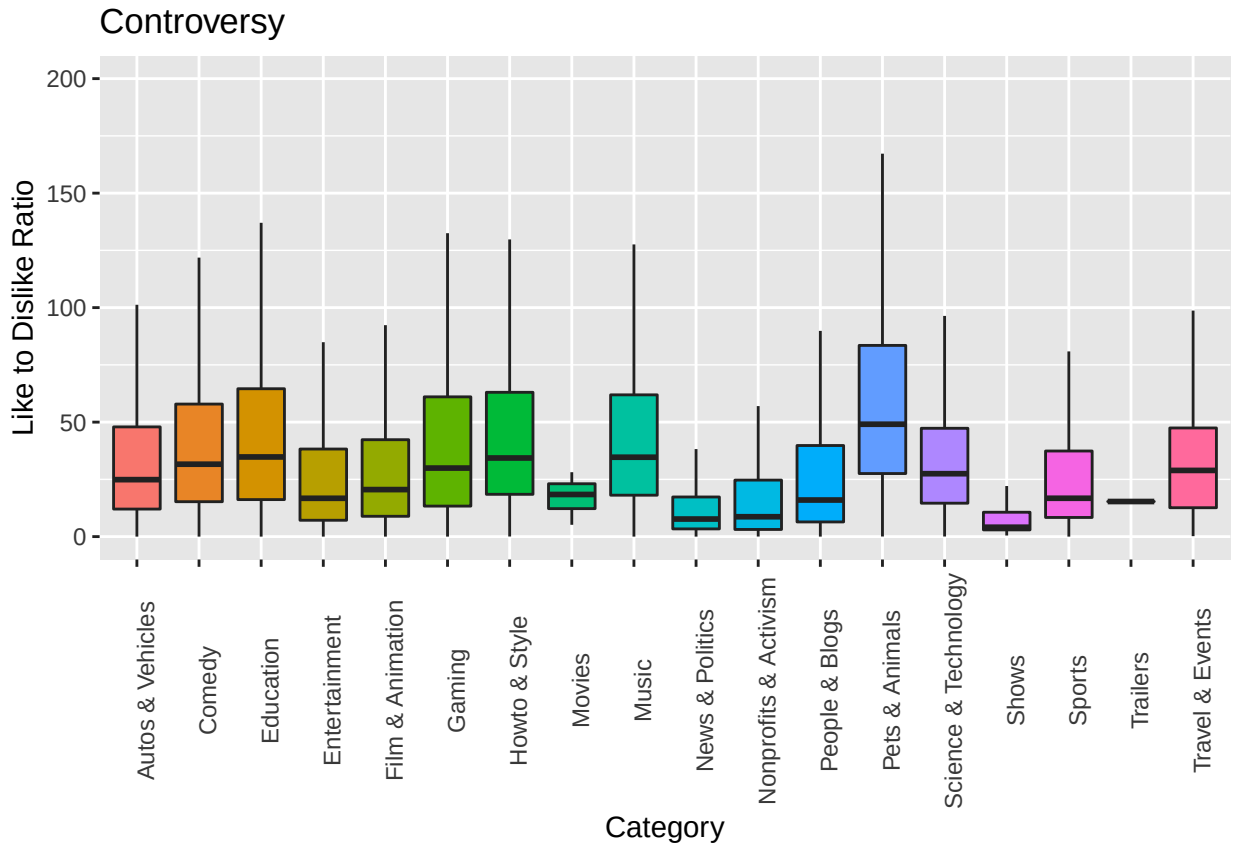
```
ggplot(data = youtube) +
  geom_boxplot(mapping = aes(x = Country, y = views, fill = trend_year), outlier.alpha = 0) +
  ylim(0, 2000000) +
  labs(x = 'Country', y = 'Trending Videos Views', title = 'Amount of Views for Trending Videos',
       fill = 'Year')
```



Controversy Basics

The graphic below visualizes a new variable that we must introduce. This is the Like to Dislike ratio, which will tell us how many likes a video has for each dislike. Looking at the 'Pets & Animals' category the average video has 50 Likes for every Dislike. Now lets define controversy. Controversy is a disagreement among the public, that is usually prolonged. With this being said, we are assuming that the closer the Like to Dislike ratio is to 1 the more controversial it is, because a ratio of 1 would mean for every dislike there is one like, therefore, there is no clear consensus by majority opinion among the public as to whether they Like or Dislike the video. With that being said 'Pets & Animals' is the least controversial category. And we can see that the 'News & Politics' video's Likes to Dislikes ratio is close to one, as we would expect, meaning this category is 'more controversial'.

```
ggplot(data = youtube) +
  geom_boxplot(mapping = aes(x = as.factor(category_id),
                             y = like_to_dislike_ratio, fill = as.factor(category_id)),
               outlier.alpha = 0, show.legend = FALSE) +
  ylim(0, 200) +
  labs(x = 'Category', y = 'Like to Dislike Ratio', title = 'Controversy')+
  theme(axis.text.x = element_text(angle = 90))
```



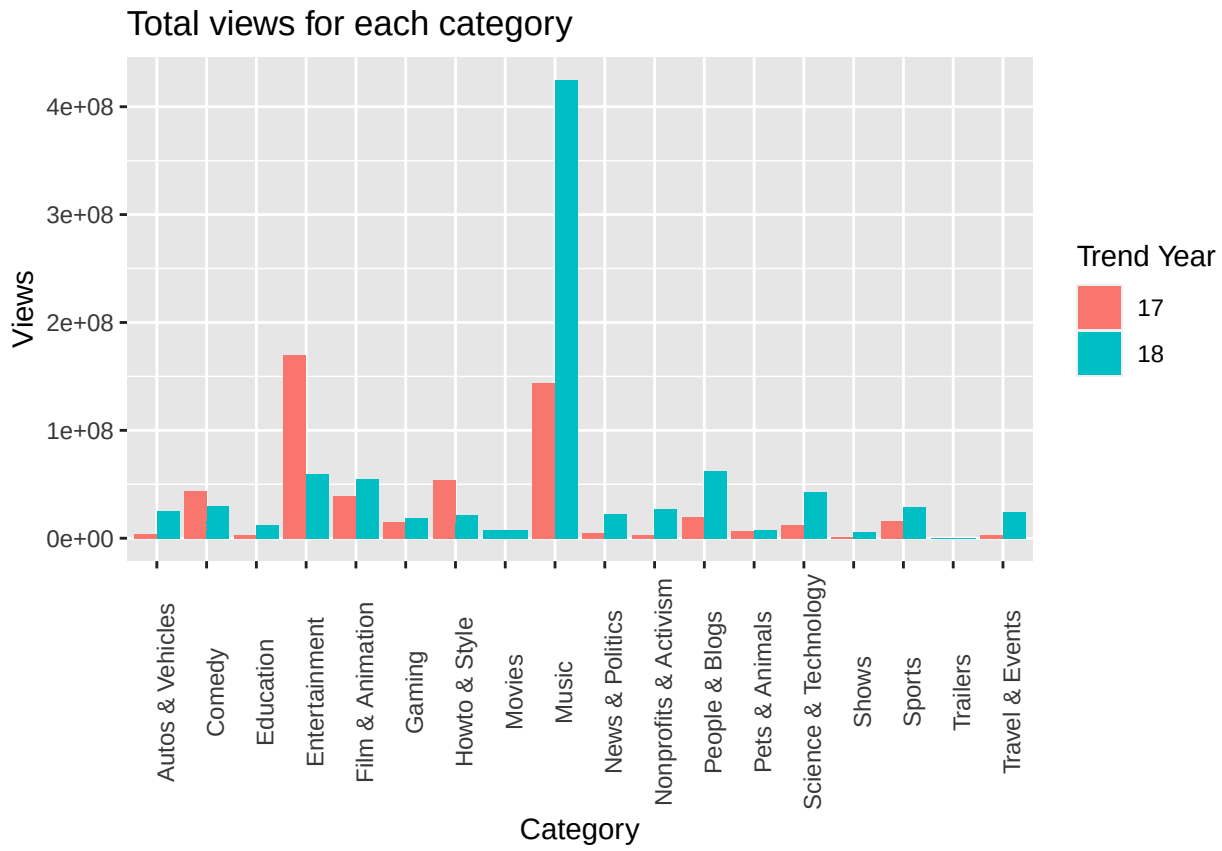
Viewership Across Category

As we can see by looking at this graphic, although the category entertainment occurred the most over 2017 and 2018's trending videos, the music category has many more total views in 2018. What does this mean? Even though there are more 'trending' entertainment videos, music videos have grossed more total views, by a longshot. This means that the average trending music video has more views than the average trending entertainment video. Why is there such a big change from 2017 to 2018 in the amount of trending music video's views? Was there a breakout artist that caused this increase? The channel with the most trending videos in 2018 in the music category was SMTOWN, a South Korean music collective. Maybe SMTOWN's dominance in the trending video realm with 1747 trending videos in 2018 was a cause of the large spike in trending Music videos total views.

```
artist = which(youtube$trend_year == 18)
artist_1 = which(youtube$category_id == 'Music')
artist_2 = youtube[artist, ]
artist_3 = youtube[artist_1, ]
which.max(table(artist_3$channel_title))
```

```
## SMTOWN
##   1747
```

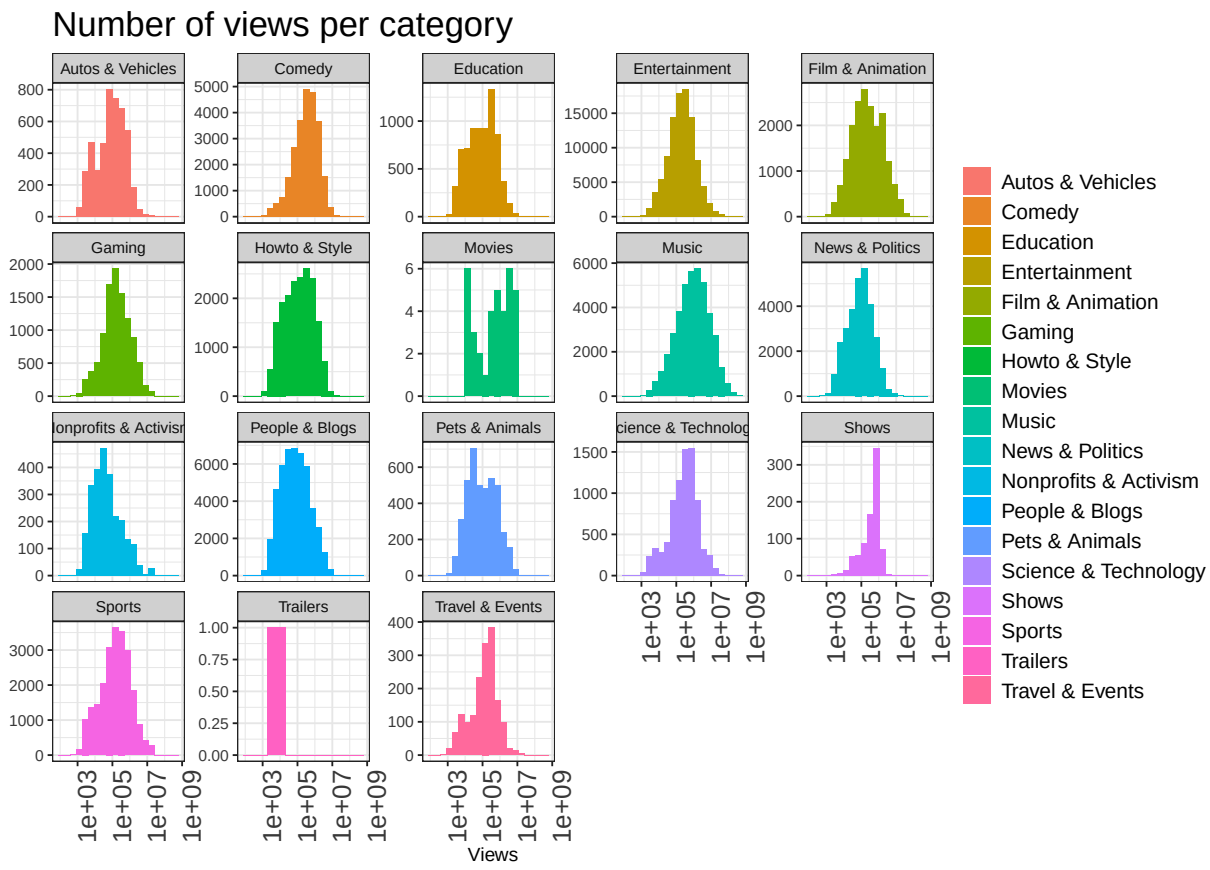
```
ggplot(data = youtube) +
  geom_bar(mapping = aes(x = category_id, y = views, fill = as.factor(trend_year)),
           position = 'dodge', stat = 'identity') +
  theme(axis.text.x = element_text(angle = 90))+
  labs(x = 'Category', y = 'Views', title = 'Total views for each category', fill = 'Trend Year')
```



After looking at the category interest on YouTube universally and individually by country, we questioned which category had the most views universally. To solve this we made the figure below, which shows the number of views per category. We chose views because we believed that it was a good indicator of how many people interacted with a specific category since liking and disliking is a voluntary response to the video viewed. Whereas views on a video happen as soon as a user begins watching the video. After plotting the figure below, we noticed that entertainment, gaming, how to style, and film & animation all had at least over 1500 videos with a million views. Entertainment had the highest with close to 20,000 videos, while gaming was the lowest with close to 2000. Categories such as movies, shows, and trailers had few videos to pass one million views with movies having around 24 videos, and shows having around 500 videos.

```
ggplot(youtube, aes(views, fill = category_id)) +
  geom_histogram(bins = 20) +
  scale_x_log10() +
  theme_bw() +
  labs(x = 'Views', y = '', fill = '') +
  theme(plot.title = element_text(size = 20),
        legend.text = element_text(size = 12),
        axis.text.x = element_text(angle = 90, size = 15)) +
```

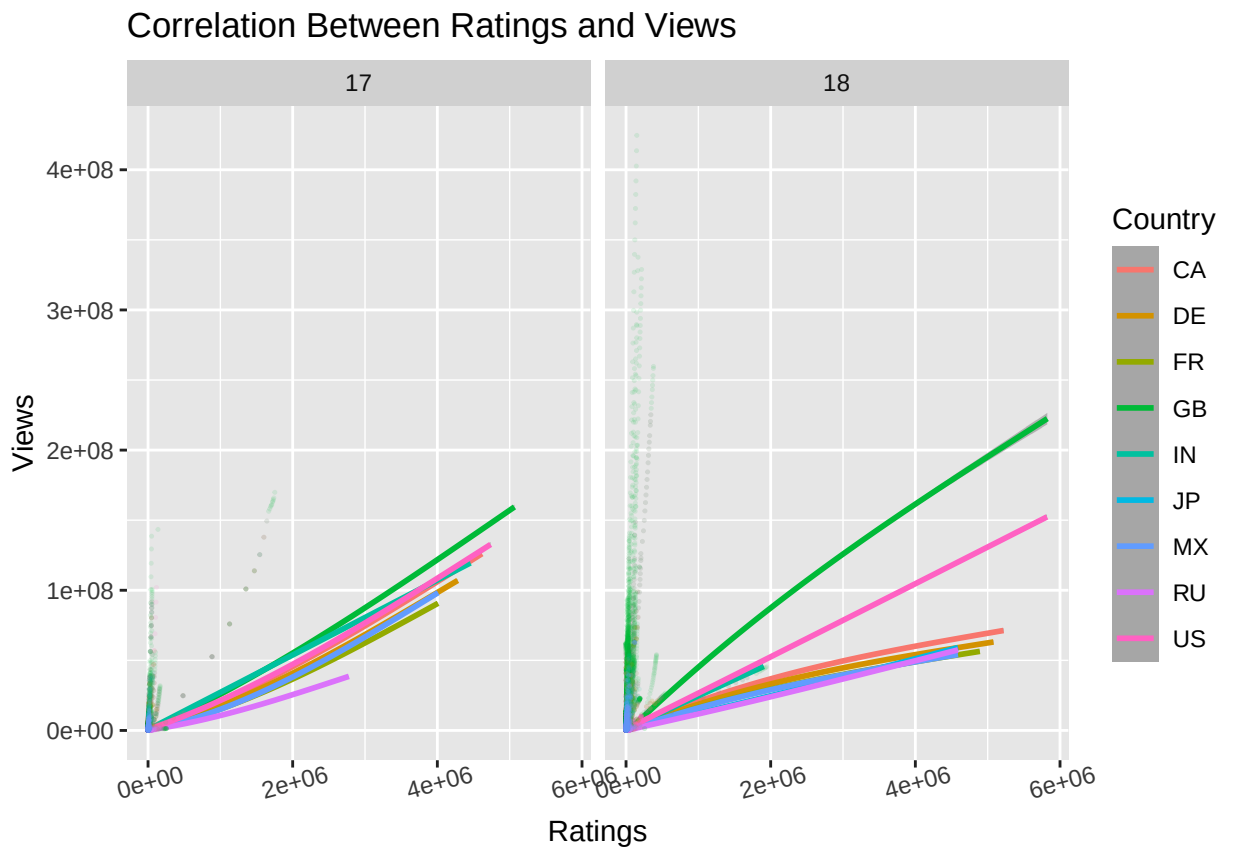
```
  facet_wrap(~ category_id, scale = 'free_y') +
  ggtitle("Number of views per category")
```



Analyzing Relationships in our data between ratings, views, comments, and more.

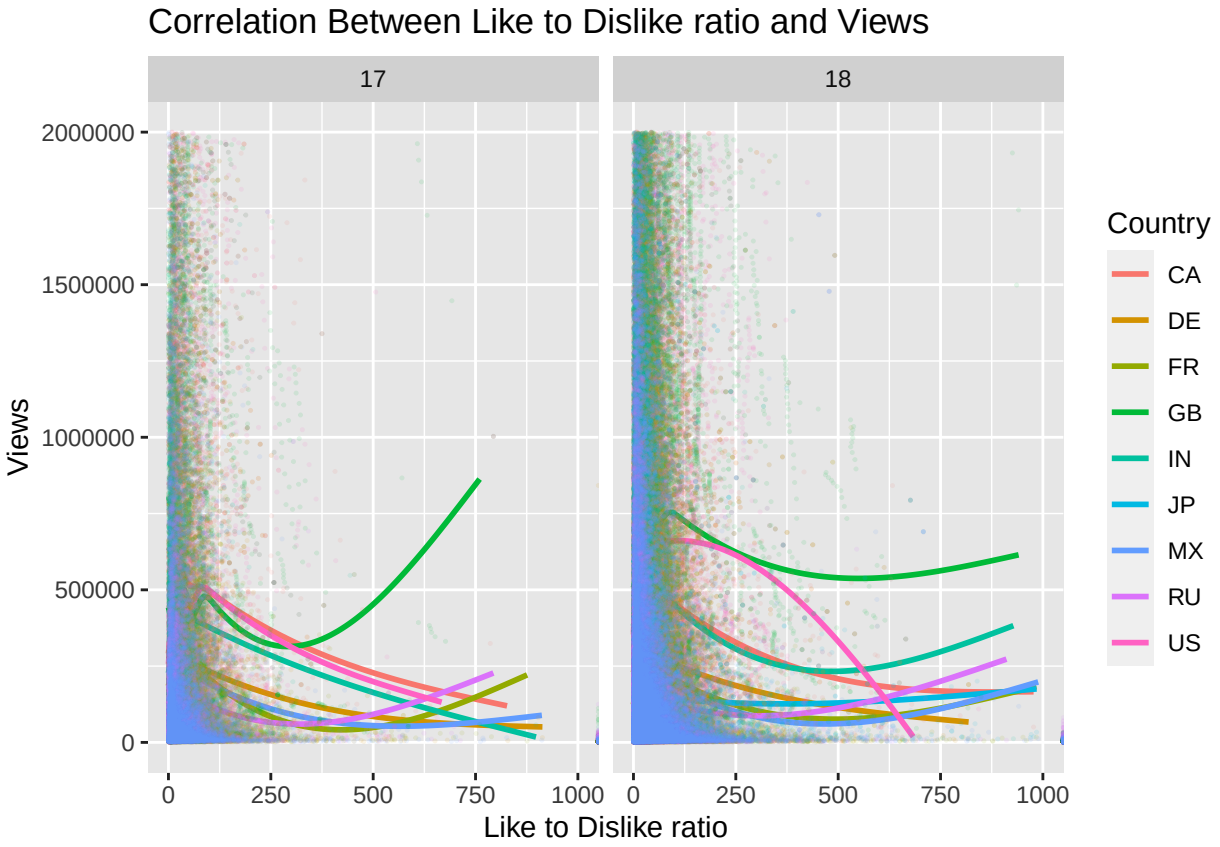
In this graph we visualized the relationship between a trending video's total amount of ratings (likes and dislikes) and it's total amount of views. What we can infer from this graph is that this relationship is positively correlated across each country. If we had a trending video then we would know from this relationship that the more ratings we get the more views we get. Across all countries the average amount of ratings is 4.2773237×10^4. Across all countries the average amount of views is 1.4178647×10^6.

```
ggplot(data = youtube) +
  geom_smooth(mapping = aes(x = ratings, y = views, color = as.factor(Country)),
              alpha = .7)+
  geom_point(mapping = aes(x = dislikes, y = views, color = as.factor(Country)),
             size = .2, alpha = .1)+
  facet_grid(.~trend_year)+
  labs(x = 'Ratings', y = 'Views',
       title = 'Correlation Between Ratings and Views', color = 'Country')+
  theme(axis.text.x = element_text(angle = 15))
```



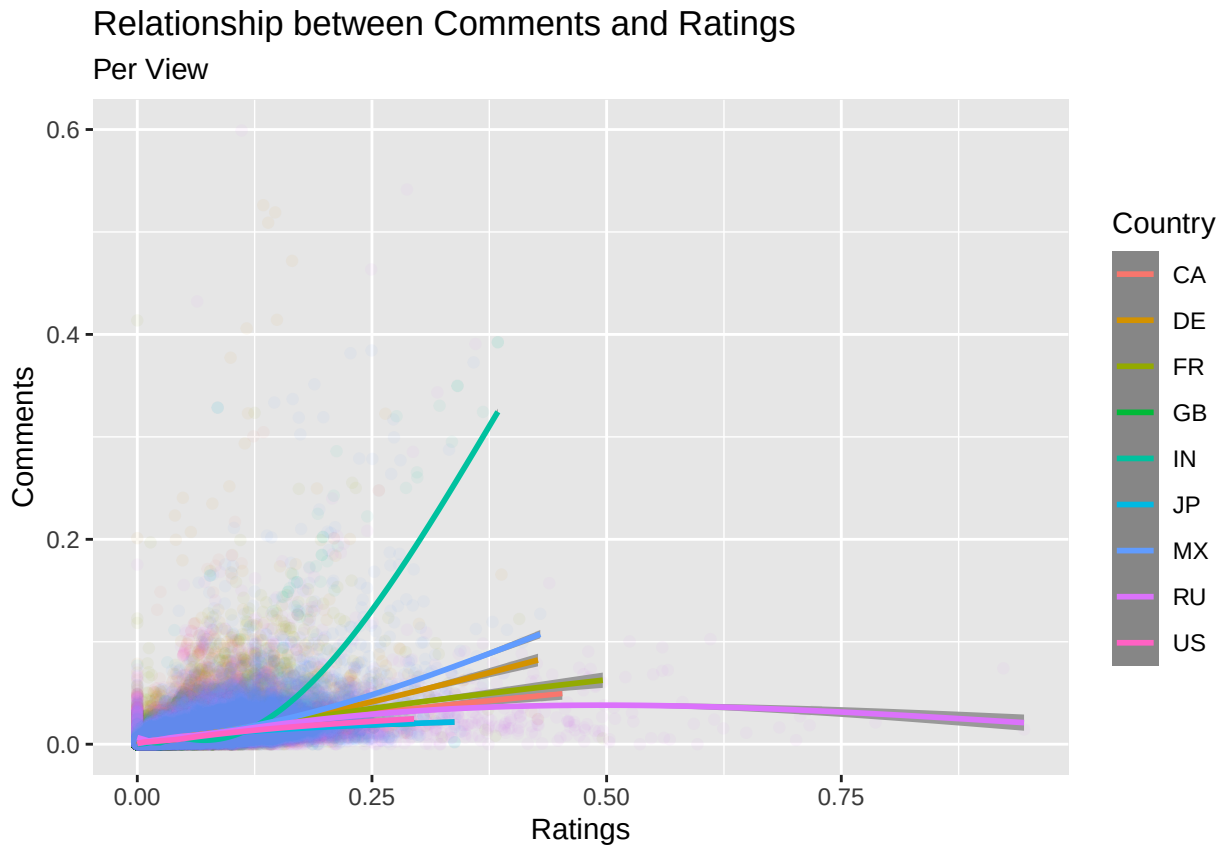
This graph looks at the relationship between the like to dislike ratio and the amount of views. In 2017 we can actually see in countries like the US, CA, and IN controversial videos actually did much better views wise than their less controversial counterparts. We can make the conclusion by looking at this graph that controversy sells in terms of gaining more views.

```
ggplot(data = youtube) +
  geom_smooth(mapping = aes(x = like_to_dislike_ratio, y = views, color = as.factor(Country))
              , se = FALSE, alpha = 1)+
  geom_point(mapping = aes(x = like_to_dislike_ratio, y = views, color = as.factor(Country))
             , size = .2, alpha = .1)+
  facet_grid(.~trend_year)+
  labs(x = 'Like to Dislike ratio', y = 'Views',
       title = 'Correlation Between Like to Dislike ratio and Views',
       color = 'Country')+
  ylim(0, 2000000)+
  xlim(0, 1000)
```



By looking at this graphic below we can tell that watchers in India are more likely to comment on a video than anywhere else in the world. Whereas a Russain watcher is much more likely to leave a rating than a comment on a trending video. 0.5579917 percent of viewers of trending videos leave a comment, and 3.9407059 percent of viewers of trending videos leave a rating.

```
ggplot(data = youtube) +
  geom_point(mapping = aes(y = comments_per_view, x = ratings_per_view, color = Country),
             alpha = .05) +
  geom_smooth(mapping = aes(y = comments_per_view, x = ratings_per_view, color = Country),
              alpha = 1) +
  labs(x = 'Ratings', y = 'Comments', title = 'Relationship between Comments and Ratings',
       subtitle = 'Per View')
```

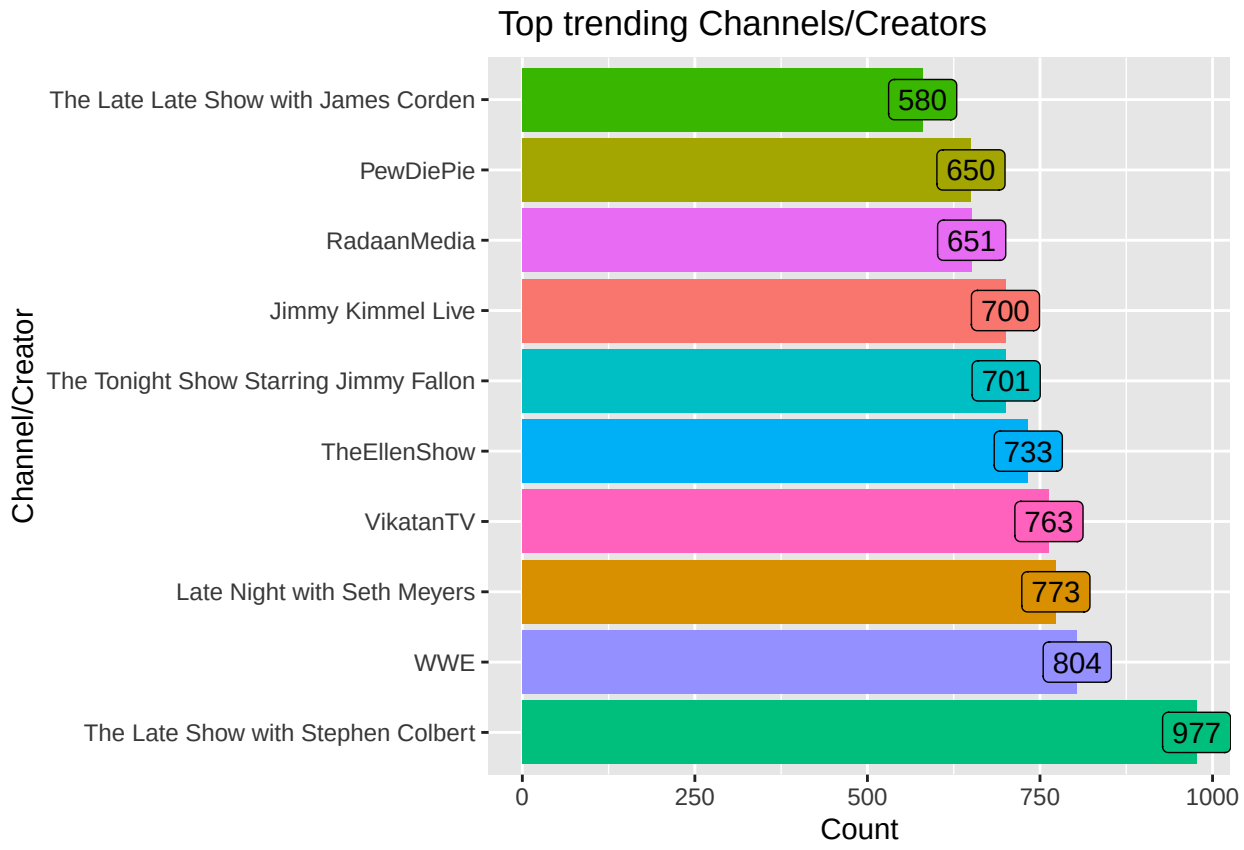


Channel Trends Across Countries

Another thing we wanted to check was if there were certain channels that appeared on the trending page more than others across all the countries. In the figure below we took the 10 channel names that had the most amount of videos on the trending page. Across all the countries, sixty percent of the channels were late night shows or talk shows such as The Late Show with Stephen Colbert which was 977 individual appearances on the trending page, and Ellen with 733 individual appearances on the trending page. Twenty percent of the channels included Entertainment channels such as VikitanTV with 763 individual appearances and RadaanMedia with 651 individual appearances. Ten percent of the channels included Sports channel for the WWE which had 804 individual appearances, with the remaining ten percent being the only individual creator that appeared on the list was PewDiePie with a total of 650 individual appearances on the trending page.

```
library(data.table)
library(dplyr)
library(lubridate)
library(RColorBrewer)
youtube_v1 = data.table(youtube)
ggplot(youtube_v1[,.N,by=channel_title][order(-N)][1:10],aes(reorder(channel_title,-N),
                                                             N,fill=channel_title))+
  geom_bar(stat="identity")+
  geom_label(aes(label=N))+
  guides(fill="none")+
```

```
labs(title=" Top trending Channels/Creators")+
xlab("Channel/Creator") + ylab("Count") +
coord_flip()
```



Conclusion

After cleaning and analyzing our data thoroughly we got the chance to answer some of our initial questions and gain a multitude of insight regarding YouTube's trending videos. Across the world the most trending videos are of the type we all love and know: Entertainment, People & Blogs, as well as Music. This should be no surprise as now more than ever people are tuning into the art and content that people are posting on YouTube. I believe we will see this trend continue forward as we move into the future and people turn more and more to YouTube for entertainment. Our worldly neighbors residing in the countries close and far to us around the globe are not too much different from us. As we saw from analyzing the most viewed categories across the different countries, the majority of the time we are viewing the same categories. Whether it be a different language or style unique to one's culture, it's refreshing to see that there are still some things that may unite us all such as our love for music and entertainment, in that sense we are all the same! People all around the world are tuning into the same channels, such as Ellen, PewDiePie, and The Late Show and adding to their viewership. Some categories are gaining popularity as the years go by, and this trend should continue as YouTube's viewership is projected to continue to grow, and as YouTube turns into one of the most popular Websites in the world! Then looking more at the technical aspects of things, we got an inside look into how to categorize a video as trending. YouTube creators around the world are working to figure

out the algorithms that classify their video as trending or make it recommendable to others. Analyzing things such as the ratings and views and comments is very insightful for this purpose. We can determine what the average trending video looks like, and then tell creators what they should aim for to get a trending video. We could tell them for their respective category what works best, and even the Like to Dislike ratio they should aim for. These things are all very interesting, especially as we move into the world of machine learning. If given stats for non-trending YouTube videos we could create a supervised learning classifier that would classify very accurately whether a video is trending or not. This is where this project is leading Faaiz and Myself, and what we plan on working on next.