Capstone Project 2
Milestone Report
Youtube Recommender System

Problem Statement

Building a recommender system that recommends a video similar to the videos a user has watched.

Client

Youtube. As someone who admittedly falls into the “Youtube rabbit hole” quite often, I see how effective these recommender systems can be. My goal is to create a simple and efficient recommender system based on the latest data I was able to find.

Data Wrangling and Cleaning

All of the data used in this project is from Kaggle. The following files are going to be used in this project:

- `GB_category_id.csv`
- `GBvideos.csv`
- `US_category_id.csv`
- `USvideos.csv`

Through data wrangling, we see that this data contains information on the top trending videos in Great Britain and the United States from 11/14/17 to 06/14/18. Each day, there are around 150-200 videos included on that list for each country.

`GBvideos.csv & USvideos.csv`

video_id
trending_date
title
channel_title
category_id
publish_time
tags
views
likes
dislikes
comment_count
thumbnail_link
comments_disabled
ratings_disabled
video_error_or_removed
description

Both csv files contain the columns shown on the left. The variables provide the videos’ statistics such as the number of likes and dislikes, number of views, number of comments, and etc.

The `tags` column provides all the tags included in the videos. Video tags/ Youtube tags “are words or phrases used to give YouTube context about a video.” Since we would want to see the relationship of the number of tags included in the video and the number of views that video gets, we create a new column called `tag_count`.

The `trending_date` column, which contains the date when that video was on the trending list, currently has an object type. To convert this to a datetime type, we need to change its current format from year/day/month to year/month/day.

Our data for Great Britain contains some null values. Since this is just .23% of the data, we simply drop it. Lastly, we merge both data frames and make it our main data frame. To account for the countries these videos went viral in, we add a column called `country` to indicate whether the data is from GB or US.

`GB_category_id.json & US_category_id.json`

The dataframe of the JSON files look like this:

	kind	etag	items
0	youtube#videoCategoryListResponse	"m2yskBQFythfE4irbTIeOgYYfBU/1v2mrzYSYG6onNLt2...	{'kind': 'youtube#videoCategory', 'etag': '"m2...
1	youtube#videoCategoryListResponse	"m2yskBQFythfE4irbTIeOgYYfBU/1v2mrzYSYG6onNLt2...	{'kind': 'youtube#videoCategory', 'etag': '"m2...

Since the GBvideos and USvideos dataframes only contain the category_id of each video, we use these JSON files to get the category title that is associated with that id. With that said, we are only interested in the `items` column. Each row in this column contains the following dictionary:

```
{'kind': 'youtube#videoCategory',
 'etag': '"m2yskBQFythfE4irbTIeOgYYfBU/Xy1mB4_yLrHy_BmKmPBggty2mZQ"',
 'id': '1',
 'snippet': {'channelId': 'UCBR8-60-B28hp2BmDPdntcQ',
  'title': 'Film & Animation',
  'assignable': True}}
```

We extract the id and category information from this using a forloop.

With this information, we add another column to our main dataframe--`category`. This will be one of our key variables in building a recommender system.

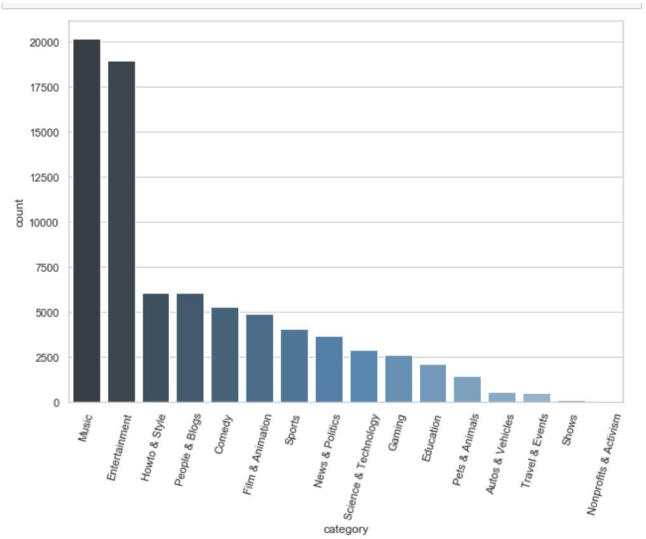
There are some null values in the category column but since this is just .23% of the data, we drop it.

Exploratory Data Analysis

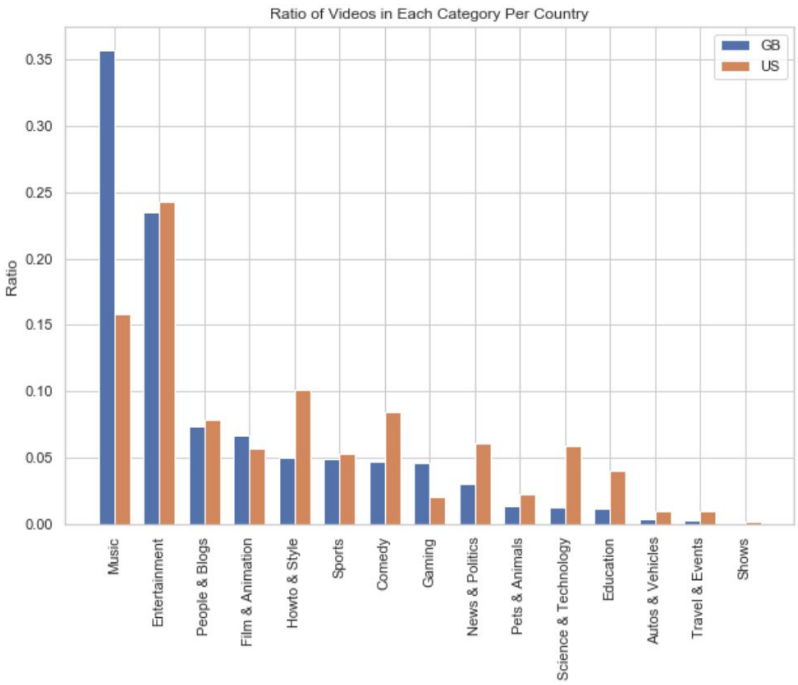
Next, we study trends across the different categories.

First, we see which categories are often on the trending videos list.

As seen in the count plot, videos that fall in the category of music and entertainment make up for the majority of the trending videos.

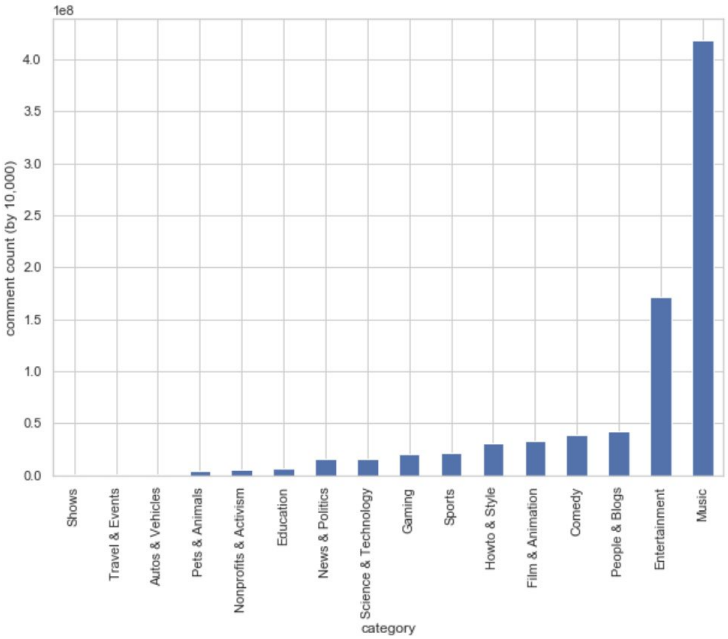


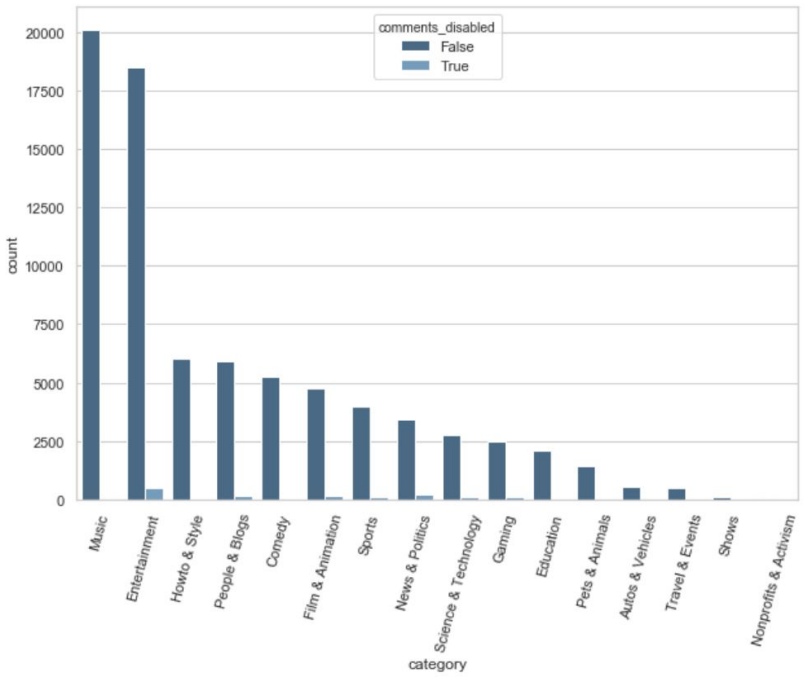
Next, we see if the popularity of these categories is the same in both countries. Since we have more data for US, we cannot simply make a countplot. Instead, we calculate the percentage of trending videos that fall in each category for each country.



As shown in the barplot, the most popular category in Great Britain is Music, while the most popular one in the US is Entertainment.

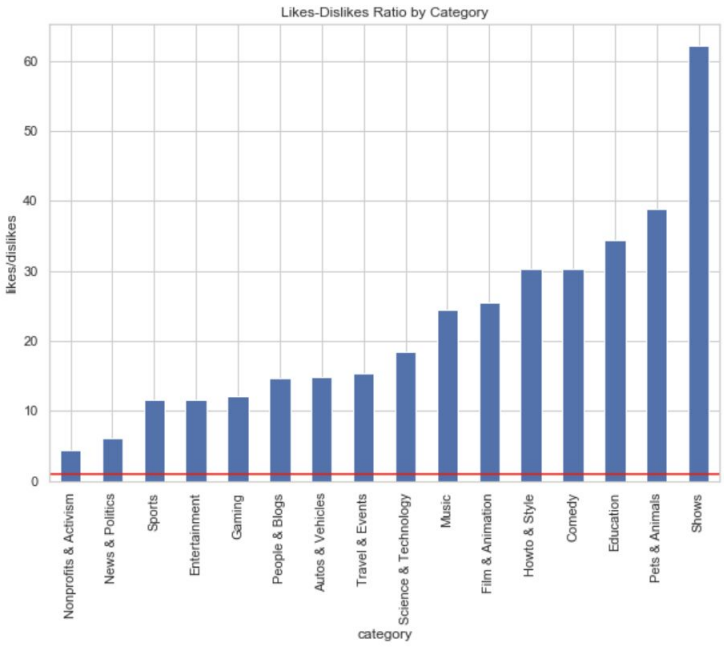
Next, we try to see if comment counts vary across categories. As expected, the most popular categories also happen to have most of the comment counts on Youtube.

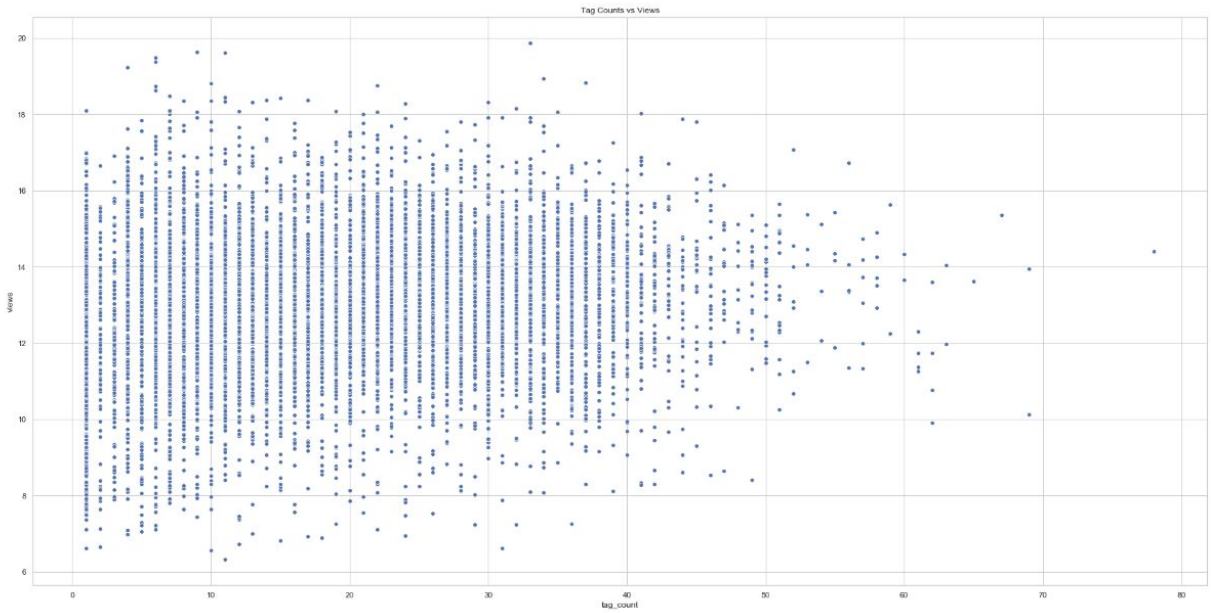




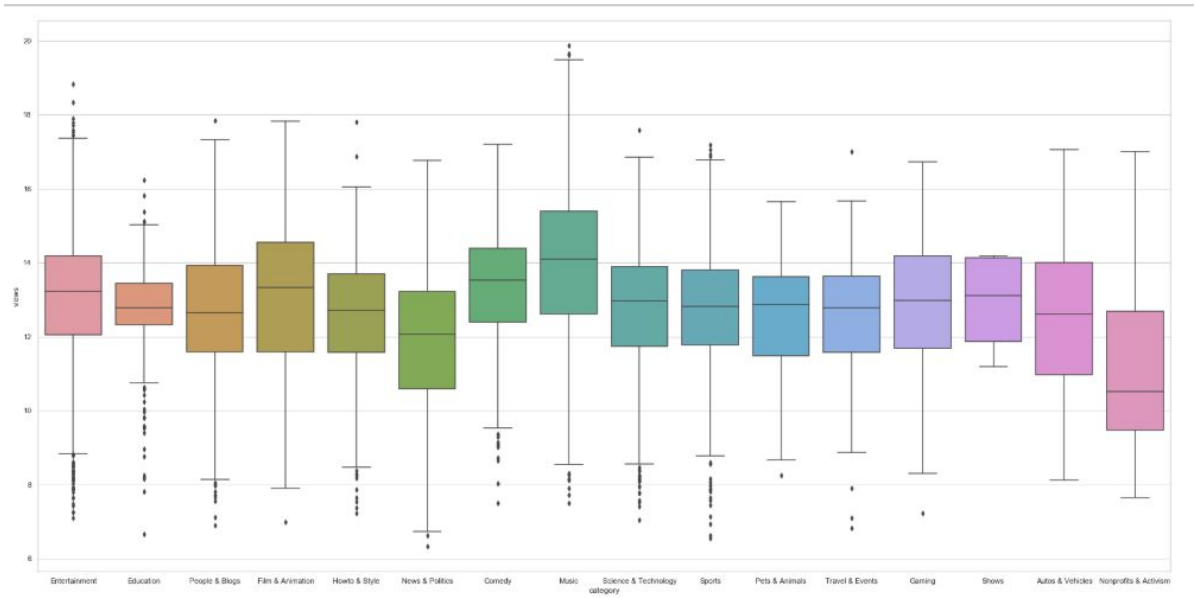
With this plot, we try too see which categories often usually disable comments on their videos. As we can see, most of the categories have very few videos in which comments are disabled. This is helpful information because this shows us that this variable would not explain much about the category.

By getting the likes-dislikes ratio of each category, we see whether or not most of these videos are trending due to positive or negative reviews. A horizontal line is drawn when y=1 (i.e. there are the same amount of likes as there are dislikes. Since no categories fall under the threshold, this means all videos are often well-liked. Although, we can see how close categories such as “Nonprofits & Activism” and “News & Politics” fall closely to the threshold. This makes sense since these videos are often controversial and divisive.





The main purpose of video tags is to help viewers find the video's content. Does this mean videos with more tags often get more views? The scatterplot above tells us it doesn't. The relationship seems random.



In the previous count plots, we see that most of the videos on the trending list belong to the music category. The boxplot above shows that this category also often have the most amount of views. It is important to note that since videos can show up in the trending list more than once, for this plot, we extracted the latest view count for each video.

Next, we try and see which videos showed up on the daily trending list the most between November 2017 to June 2018 in each country. In the US, the most a video has been part of the daily trending list during this time period was 38 times. There were six videos that were on the trending list for 38 days, and three of them belong in the music category.

	video_id	count	title	category
0	NooW_RbfdWI	38	Jurassic World: Fallen Kingdom - Official Trai...	Entertainment
1	Il-an3K9pjg	38	Anne-Marie - 2002 [Official Video]	Music
2	2z3EUY1aXdY	38	Justin Timberlake's FULL Pepsi Super Bowl LII ...	Sports
3	BhIEIO0vaBE	38	To Our Daughter	People & Blogs
4	u_C4onVrr8U	38	Miguel - Come Through and Chill ft. J. Cole, S...	Music
5	u_C4onVrr8U	38	Miguel - Come Through and Chill (Official Vide...	Music

In Great Britain, the most times a video showed up on the trending list is 30. This was only true for one video, which fell in the entertainment category.

	video_id	count	title	category
0	j4KvrAUjn6c	30	WE MADE OUR MOM CRY...HER DREAM CAME TRUE!	Entertainment

Inferential Statistics

Now that we've seen the difference between different categories, we want to see if the differences are statistically significant. Since there are 16 categories that we want to compare, we can use a statistical technique called Analysis of variance (ANOVA). ANOVA uses F-tests to statistically test the equality of means in each group. If the p-value of these F-tests are lower than 5%, we reject the null that the mean across the categories are equal. This will help us determine whether we should include the variable in our predictive model or not.

Views By Category

> Null Hypothesis: The number of views each category gets are not statistically different from each other.
>
> Alternative Hypothesis: The number of views each category gets are statistically different from each other.

```
f-stat: 29.91229
p-value:0.0000000000
The number of views each category gets are statistically different from each other.
```

Likes By Category

Null Hypothesis: The number of likes each category gets are not statistically different from each other.

Alternative Hypothesis: The number of likes each category gets are statistically different from each other.

```
f-stat: 544.81234
p-value:0.0000000000
The number of likes each category gets are statistically different from each other.
```

Dislikes By Category

Null Hypothesis: The number of dislikes each category gets are not statistically different from each other.

Alternative Hypothesis: The number of dislikes each category gets are statistically different from each other.

```
f-stat: 60.05653
p-value:0.0000000000
The number of dislikes each category gets are statistically different from each other.
```

Comment Count By Category

Null Hypothesis: The number of comments each category gets are not statistically different from each other.

Alternative Hypothesis: The number of comments each category gets are statistically different from each other.

```
f-stat: 148.42400
p-value:0.0000000000
The number of comments each category gets are statistically different from each other.
```

Tag Count By Category

Null Hypothesis: The number of tags the videos in each category are not statistically different from each other.

Alternative Hypothesis: The number of tags the videos in each category are statistically different from each other.

```
f-stat: 51.10537
p-value:0.0000000000
The number of tags the videos in each category has are statistically different from each other.
```

In conclusion, all of these features are statistically different across categories. This tells us that all of these features should be included in our recommendation system. When recommending videos, we can find similar videos that fall in the same category.