Data Story

Background

Ranked as the second-most popular website in the world, the Google-owned video-hosting site YouTube claims that one billion hours of user-generated content is watched on the broadcast site each day. Yes, that's right. One billion hours per day. This global phenomenon is thanks in large part to the video site's machine learning recommendation capabilities. Although some trends are predictable, like when a new song is uploaded from a popular artist or when a new superhero movie trailer is released and gains millions of views in a single day, others are more surprising, like a viral cat video that transpires into an overnight sensation or a new dance craze that is quickly shared across the internet.

The record-breaking exposure channels can receive on the video platform has changed the game of marketing and afforded new career opportunities. "YouTubers," also known as YouTube personalities, celebrities, or content creators, are people who have gained popularity from their videos on their YouTube channels. Those who are able to amass a huge following can take advantage of corporate sponsorships and sign deals to have product lines of their own.

Introduction to the Problem

So how does one become a YouTuber? How does a YouTube video become popular in the first place? The answer partially lies in YouTube's Trending tab capabilities. The Trending tab features videos algorithmically, using a combination of factors, including view count, rate of growth in viewership, the age of the video and other user-engagement markers such as shares, comments, and likes. Videos that meet a certain threshold are then given a high-profile spot on YouTube's home page and can rack up millions or even billions of views in a short period of time. But the question still remains, how did the video get on the Trending tab to begin with?

Because there are a number of factors to consider when predicting what makes a video popular to one audience and not another, this project will attempt to solve the following problem:

What factors affect how popular a YouTube video will be and, using such factors, can we predict popularity for any video?

Potential Clients

Private individuals to large production companies have used YouTube to grow their audience base. Because of this large range of interests, there are several clients who would want to know the answer to this project's question. In particular, entertainment content creators such as movie studios or musical artists would be interested in knowing how to ensure that their videos reach a wide audience who would then purchase a movie ticket or buy a single to download. Likewise, advertisers who use music in their commercials to sell a product such as a luxury car would want to know which artists are trending to increase the likelihood of viewership for their sponsored video ads. Beyond entertainment products, clients may want to create a home improvement channel, document breathtaking travel destinations, or share information through video lessons designed for kids.

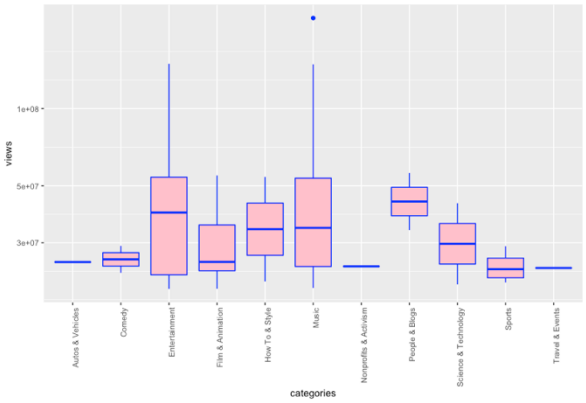
Deeper Dive into the Data Set

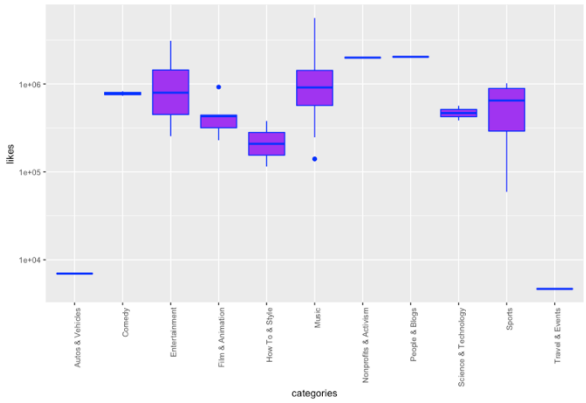
To solve the problem, **factors** and **popular** have to be classified. Based on the variables collected in the original data set, “tags” and “category_id” are **factors**. “Views”, “likes”, and “comment_count” will be summed to define what makes a video **popular**. Multiple time points are analyzed by comparing the “publish_time” to the “trending_date” to see how fast a video reached trending status. For this project, the “description” and “title” variables were not used due to their long character strings and because doing so could bias the results based on the particular style the content creators used to write them.

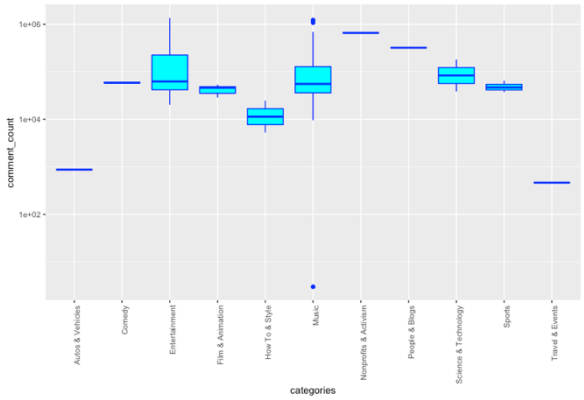
Ultimately, the purpose of the project will be to build a model that predicts the number of views, likes, and comments a YouTube video will receive based on the particular factors used (“tags”, “category_id”). For example, if the client uses _____ tags, they will gain _____ views in _____ amount of time from the publish date.

An important feature of the original data set is the observations collected across 10 different regions. Due to time constraints, the predictive model for this project will not account for the country in which a video trended. However, this information could be quite significant for clients who serve an international base.

After diving deeper into the data set, it was quickly discovered that the top viewed, liked, and commented on videos came from the *Music* and *Entertainment* categories. Because views, likes, and comments are measurements of user engagement that can be used to determine a video’s popularity, it’s important to find a representative sample that would be beneficial to multiple clients.







If the entire data set was used to generate a predictive model to answer this project's question, it would provide limited value to clients whose videos are in categories other than *Music* and *Entertainment*. Some of the factors associated with the videos, in particular the "tags" variable, are reflective of the category a particular video resides in. For example, the tag "acoustic" frequently appears for videos in the *Music* category. If a client whose videos were directed towards the *Science & Technology* category used this tag, it would most likely not effectively target the right audience for the video's content.

Although the entire data set contains 32 categories in all, only 12 are used in the final analysis because the other 20 did not have at least 100 observations, which was set as the limit for this project. The 12 categories included are: *Film & Animation, Music, Pets & Animals, Sports, Gaming, People & Blogs, Comedy, Entertainment, News & Politics, How To & Style, Education*, and *Science & Technology*.

Another limitation of the data set stems from the "tags" variable. It was quite messy to work with and required the data scientist who was cleaning and wrangling the data to make several judgment calls. In the original data set, the tags were compiled into a list separated by the pipe or vertical bar character. Because of this, some of the tags were stuck together, misspelled, or contained punctuation or spacing issues. For example, one tag is "ed sheeran" and another tag is "edsheeran". If these tags were not combined, this would result in two different tags each counted once instead of one tag being counted twice. It's too hard to tell if misspellings such as this were done intentionally to match the common search terms viewers use to look up a video. On the one hand, it makes sense to use the tags as they were originally entered; however, the frequencies would be quite low if similar tags were not compiled together, and thus not very useful for potential clients.

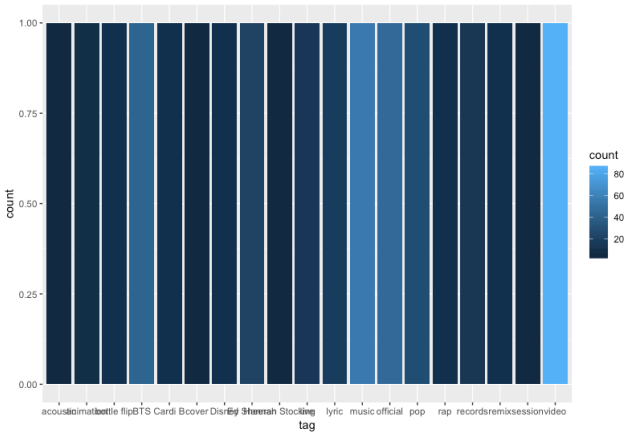
Tags
Childish Gambino|"Rap"|"This Is America"|"mcDJ Reco...
Rewind|"Rewind 2017"|"youtube rewind 2017"|"#You...
Ariana|"Grande"|"No"|"Tears"|"Left"|"To"|"Cry"|"Univer...
Becky G|"Natti Natasha"|"Natti Natasha Music"|"Natti ...
BIGHIT|"빅히트"|"방탄소년단"|"BTS"|"BANGTAN"|"방탄"|"FA...
The|"Weeknd"|"Call"|"Out"|"My"|"Name"
Luis|"Fonsi"|"Demi"|"Lovato"|"Échame"|"La"|"Culpa"|"U...
Cardi B|"I Like It"|"Invasion of Privacy"|"Bad Bunny"|"J ...
marvel|"comics"|"comic books"|"nerdy"|"geeky"|"supe...
Maluma Music|"Maluma Official Video"|"Maluma Vide...
BIGHIT|"빅히트"|"방탄소년단"|"BTS"|"BANGTAN"|"방탄"|"FA...
Taylor Swift|"Delicate"|"Big"|"Machine"|"Records"|"LLC...
calvin harris|"calvin harris one kiss"|"calvin harris dua...
TWICE What is Love|"TWICE What is Love?"|"TWICE 왓이...
Maroon|"Girls"|"Like"|"You"|"Interscope"|"Records*"|"P...

To make the tags more manageable, they were put into separate columns. Because clients need to know which tags to use to drive the most views, likes, and comments to their videos, the frequency with which each tag appeared needed to be calculated. It wouldn't be useful to count similar tags separately, so tags that contained the same root word or referred to the same song or person were consolidated into one tag name to increase the frequency count. For example, the tag *rap* would include "rap" and "hip hop" and "hip-hop". Likewise, the tag *animation* would include "Animation" and "animation" and "animator". Some of this process was limited based on the pop-culture knowledge of the data scientist who worked on this data set. For example, there are many tags that refer to the artist *Jennifer Lopez* but not everyone working on this data set would know that her stage name is *J. Lo* and that one of her nicknames

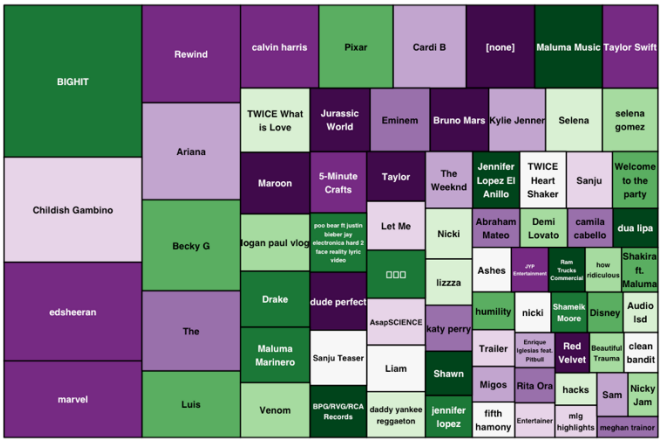
is *Jenny from the Block* based on her 2002 hit single of the same name. Therefore, if this information was not known, the frequency for the tag *jennifer lopez* may be higher or lower depending on the data scientist's fluency in pop-culture references.

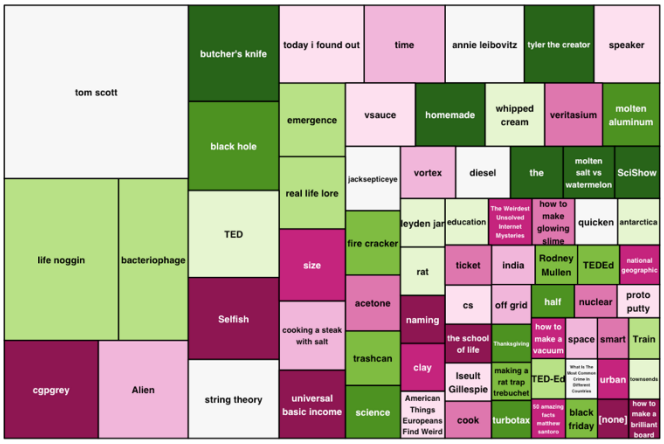
Preliminary Exploration and Initial Findings

Preliminary exploration reveals which tags occur most frequently. As shown in the bar graph to the right, tags such as *video*, *music*, and *BTS* appear often in the top 100 viewed videos for the *Music* category while *acoustic*, *session*, and *cover* appear less frequently.



Tree maps can also illustrate the frequency with which particular tags are used. In the left-hand side tree map, the tags for the top 100 viewed videos across all categories are shown. In right-hand side tree map, the tags for the top 100 viewed videos for the *Education* category are shown.





Approach

The approach has changed from the original question proposed. Comparing popularity across the 10 different regions will no longer be part of the final predictive model, only video results from the U.S. are analyzed. In addition, 12 video categories are assessed instead of the original 32. As previously mentioned, these changes were due to time constraints and limited data.

In regards to the findings from the preliminary exploration above, it's clear that the tags need to be consolidated; otherwise, there will be too many tags with each showing little variance between them. Tallying the frequency count in this way will provide more general utility for potential clients and produce a higher-quality analysis overall.