Emma Lynn May

Final Project Report

I worked with Ann Marie Humble on this project. Our slides are linked here and GitHub repo is linked here.

Introduction

In this project, we analyzed a dataset we created containing data on YouTube videos. We used many of the tools we discussed in class to answer questions about comment sections, descriptions, genres, and the popularity of YouTube videos. YouTube is a very popular (if not the most popular) video sharing platform on the internet today, and increasingly more individuals are making a living creating content and advertising on the platform. Understanding the trends and attributes of videos is very valuable for our stakeholders including aspiring YouTubers, established content creators, and advertisers.

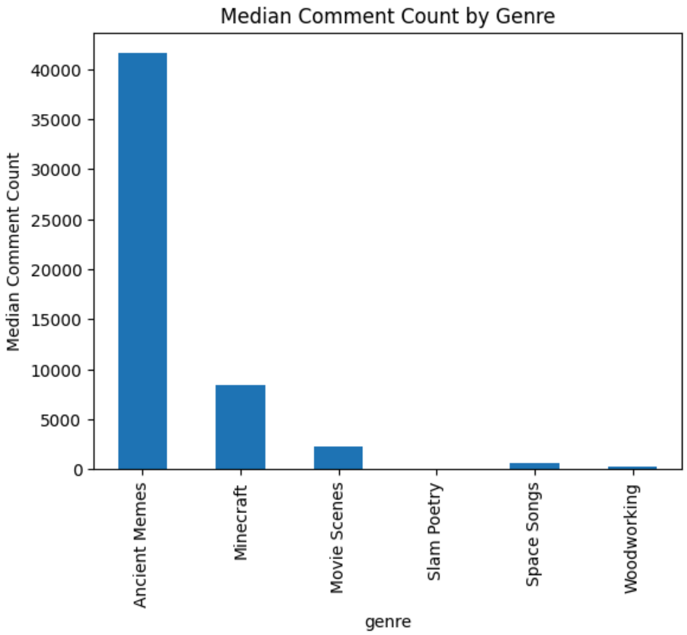
Dataset

We created our own dataset using the YouTube API. We selected 9 public playlists of YouTube videos to collect data on. 6 of these playlist topics were selected based on their varied topics and attributes: Slam Poetry, Minecraft, Movie Scenes, Woodworking, Space Songs, and Ancient Memes. The remaining 3 were selected because they are popularly advertised over, affecting the success of the video, which we'll discuss later. These playlists were Personal Finance, Travel, and Health. Then we wrote a Python script to collect data about each video in each of these playlists. We output the data to a .csv file, after removing attributes we weren't interested in and cleaning string values so they would not cause issues with the .csv format. This data is suitable for analysis because it is fairly large (with data for almost 1000 videos) and completely accurate since it comes directly from YouTube itself. It is also live data, which is exciting because it eliminates the problems that come from having out of date information. Up to date information can be retrieved and formatted into a dataset very quickly, which can then be run through our analyses and interpreted.

Analysis Technique

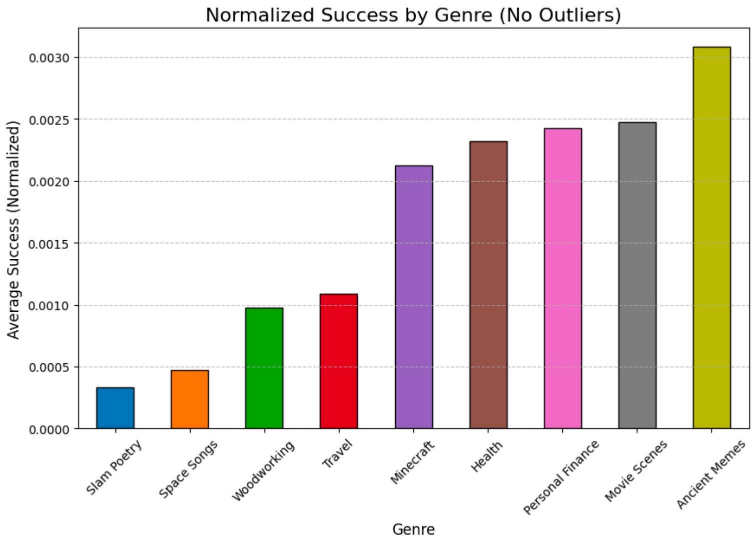
We used several different analysis techniques to interpret our data. To answer our questions about trends in comment sections, we started by visually comparing median view and comment counts by genre. We also calculated correlation coefficients between comment count & view count, and comment count & video duration. To explore what attributes make videos more successful, we created scatter plots and bar graphs for visual analysis, because we found that to be the easiest way to visualize differences in success between video genres. We also found Pearson correlation coefficients to determine which video attributes are most correlated to a video's success. To analyze trends in video description, we first looked at median and maximum video description lengths, and compared between genres. Then we attempted to use logistic regression to predict video genre based on description length. We tested using this method with both TF-IDF vectorization (which measures originality of words) and count vectorization (which measures the frequency of words). Finally, we attempted to predict video genres using a decision tree. These techniques are suitable for our data and domain because they helped us answer questions which provided us insights and knowledge about YouTube for our stakeholders.

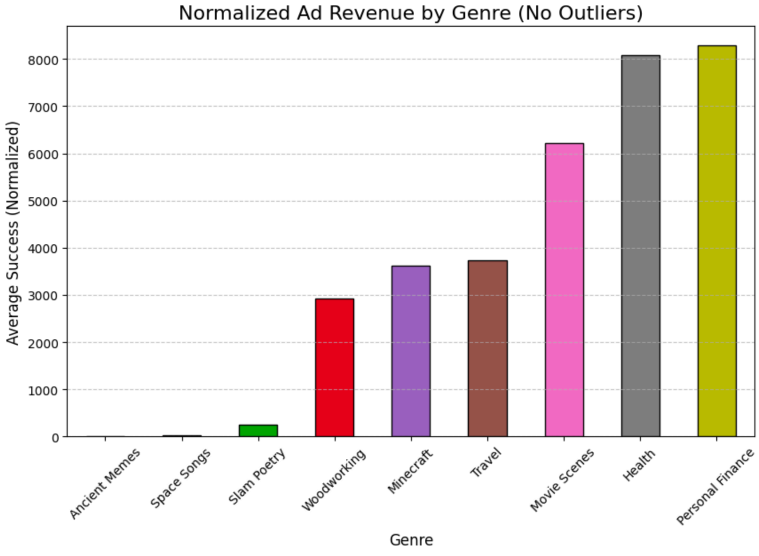
Results



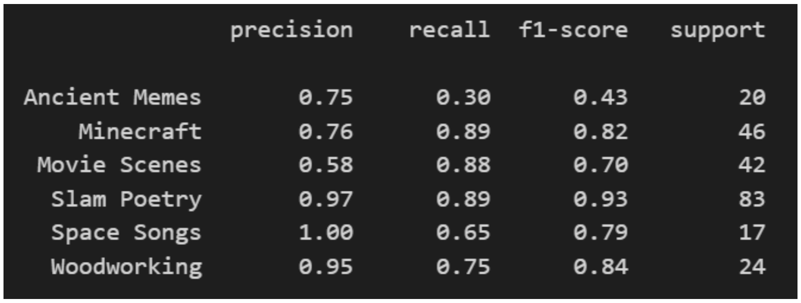
The first questions we wanted to answer were about trends in YouTube comment sections. We wondered what types of videos people tend to leave more comments on. We found that out of the playlists Ancient Memes, Minecraft, Movie Scenes, Slam Poetry, Space Songs, and Woodworking, videos in Ancient Memes by far had the highest median comment count. We also wondered if videos with longer runtimes had longer comment sections, but did not find evidence of a significant correlation, with the highest correlation coefficient we found being .2696 (Movie Scenes). This information is valuable to stakeholders wanting to increase the comments on their videos, who could potentially try making videos like those in Ancient Memes (using humor, nostalgia, virality, etc) to increase their comment counts.

Next we wanted to investigate what makes a video successful. We defined a video's success based on likes, views, and estimated ad revenue. We looked at these metrics of success individually and combined into a normalized success score. We saw that the Ancient Memes genre had the highest average overall success per video, despite ranking dead last in estimated ad revenue. We also noticed that Minecraft and Health ranked very similarly in estimated ad revenue, which was interesting because videos on Health make over 32 times the ad revenue of Minecraft videos per 1000 views. These conclusions are beneficial to our stakeholders because they demonstrate the power of a video going "viral", even if the video is on a topic that may not be the most highly valued by advertisers. It could also benefit advertisers as an insight into video genres that would be cheap to advertise over but still get a lot of visibility.





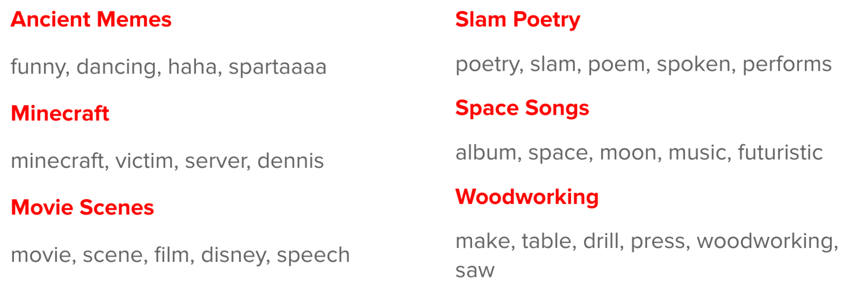
Last, we wanted to explore the potential applications of video description analysis. We were able to successfully predict video genre based on description content using Logistic Regression and a Count Vectorizer. We used F1-scores to evaluate this model, and got an overall F1-Score of 0.81. Our model performed the best at predicting Slam Poetry videos (which often have the category name in the description), and the worst at predicting Ancient Memes (which is, by nature, the most internally varied category).



```
               precision    recall  f1-score   support

Ancient Memes       0.75      0.30      0.43        20
    Minecraft       0.76      0.89      0.82        46
 Movie Scenes       0.58      0.88      0.70        42
  Slam Poetry       0.97      0.89      0.93        83
  Space Songs       1.00      0.65      0.79        17
  Woodworking       0.95      0.75      0.84        24
```

We also found that the best keyword predictors of each genre were as follows:



Ancient Memes

funny, dancing, haha, spartaaaa

Minecraft

minecraft, victim, server, dennis

Movie Scenes

movie, scene, film, disney, speech

Slam Poetry

poetry, slam, poem, spoken, performs

Space Songs

album, space, moon, music, futuristic

Woodworking

make, table, drill, press, woodworking, saw

This information is valuable to our stakeholders because it could help them categorize their own videos, which would assist in selecting video tags and identifying niche audiences to aim their content at. It could also help creators identify keywords to include in their video descriptions to increase their videos' visibility in a certain genre.

Technical

To prepare our dataset for analysis, we needed to drop rows that contained null values in columns relevant to each analysis (commentCount, duration, description, title, etc). To complete our analysis on video success, we needed to create some additional columns. We added a genre CPM (the dollar amount advertisers will pay per 1000 views to advertise over a specific genre of content) column, based on data from the 2023 fiscal year (Yahoo Finance, UberSuggest). Then we calculated estimated ad revenue using the video's CPM and view count. We also normalized estimated ad revenue, view count, and like count to create an overall “success” metric for each video. Finally, we removed outliers for some of our analyses (but left them in for some, as they gave us some interesting insights).

We used a variety of analysis techniques (calculating correlation coefficients, logistic regression, decision trees, manual inspection of visual representation of data, etc) in this project because of our varied dataset and goals. Our techniques were suitable for our dataset and goals because they gave us results that effectively addressed our research questions, offering actionable insights into YouTube trends for stakeholders.

To analyze trends in comment sections, we began by comparing view counts across genres. We also explored the relationship between comment counts and view counts by calculating correlation coefficients for each genre. We looked at the relationship between comment counts and video duration. While we found a fairly expected correlation between comment count and view count, we did not find any meaningful correlation between comment count and video duration.

To answer our questions about what attributes make videos more successful, we began by visually comparing video success by genre using a scatter plot. We saw that there were some extreme outliers in video success that made it hard to interpret results, as well as vastly different numbers of videos in each genre. To overcome these issues, we removed outliers from the scatter plot as well as looking at a bar graph of average video success per genre. We received somewhat disappointing results when looking at correlations between different video attributes and video success. While we got statistically significant results for correlations with caption status and description length, the correlations were not strong, at .168 and .237 respectively.

Video Attributes	**Pearson Correlation Coefficient**	**p-value**
Length	0.037	0.284
Caption Status	0.168	0.0
Description Length	0.237	0.0
Title Length	0.008	0.813

Paid Product Placement	0.01	0.773

An alternative approach / opportunity for future work would be to expand the video attributes being explored as explanatory variables. Metrics like the day of the week / time of day the video was published, amount of time the creator spent creating the video, and popularity of the video topic could also be promising attributes to investigate. We could also use comment count as an additional metric of engagement/success.

While exploring trends in video descriptions, we looked at description lengths over time, but did not find any clear patterns, except in the Slam Poetry genre which showed description lengths increasing over time. While predicting video genres using logistic regression, we were surprised that we got better results using a Count Vectorizer than using TF-IDF (by about 10%). While this was unexpected to us, since we thought more original keywords would be better predictors of our distinct genres, a possible explanation could be that video descriptions might contain many unique words specific to that video, but not specific to the genre. The Ancient Memes category was by far the hardest to predict, likely because the genre is much more eclectic than the others.

Conclusion

In conclusion, our analysis of YouTube video data provides valuable insights into the factors influencing comment section trends, video success, and genre-specific characteristics. By leveraging statistical and machine learning techniques, we identified correlations and patterns that can inform content creation, audience engagement, and advertising strategies. These findings are particularly useful for aspiring and established YouTubers aiming to optimize their content, as well as for advertisers looking to maximize impact on the platform. Overall, our results highlight the importance of using data-driven approaches to understand the evolving dynamics of the YouTube ecosystem, and other social media platforms.

References

Shaikh, Maliha. “20 Highest CPM YouTube Niches in 2023.” *Yahoo Finance*, 25 Oct. 2023, https://finance.yahoo.com/news/20-highest-cpm-youtube-niches-091413416.html.

Ubersuggest. https://app.neilpatel.com/en/ubersuggest/overview. Accessed 6 Dec. 2024.

Slam Poetry:
https://www.youtube.com/playlist?list=PL9rT6KUep7uqN8Xioo3R32cUyogH5wVMk

Minecraft:
https://www.youtube.com/playlist?list=PLKiPmtkaxvOYFJzP847fBEfpr4FZD3tpg

Movie Scenes:
https://www.youtube.com/playlist?list=PLYhREcq3PbZ8duNVstBQ_0jnHC1ffxJfu

Woodworking:
https://www.youtube.com/playlist?list=PLr6O8F0Sz5TrKaAzGdhmmkIoxE-vdRYDE

Space Songs:
https://www.youtube.com/playlist?list=PLG04vLHTI500i3IcAK5Xsgn9F2q66ZOgr

Ancient Memes:
https://www.youtube.com/playlist?list=PLHEBsC_NxJ_R3KnXJa4BfF7NEbnNvW_jg

Personal Finance:
https://www.youtube.com/playlist?list=PLyDivztPycqygz3KyF8LV49oEX-I8MtcI

Travel: https://www.youtube.com/playlist?list=PLf_KMOEjgle4-6QCdJ51QprYSSYPjltkf

Health: https://www.youtube.com/playlist?list=PL3q2mpVkaj0SzMx2niYaCzIf3hJlIZJae