YouTube Content Popularity Analysis

Project Update

Emma Crawford
emcr8954@colorado.edu

1 Problem Statement & Motivation

With this project, I hope to discover the key factors that determine the view count and overall popularity of YouTube videos in the United States. As someone aspiring to create YouTube content as an income stream, understanding the elements that contribute to video success is one of the first steps in deciding what to create. By analyzing a dataset of popular YouTube videos, I can learn about the patterns that attract users and optimize content creation strategies.

My motivation behind this project is a personal goal of establishing a successful YouTube channel and generating income from my content. I believe that data analysis, combined with knowledge of YouTube's trending video data, can provide valuable guidance for content creators. By identifying the features that influence popularity in the current market, I can make informed decisions to maximize my video's reach and income-generation potential.

My project will attempt answer three key questions. Firstly, what factors have the most significant impact on the monetization potential of YouTube videos in the US? Secondly, how do engagement metrics, such as likes, dislikes, comments, and subscriber count, affect audience retention and viewer loyalty? Finally, are there specific content genres or categories that consistently perform well in terms of views, audience, and revenue generation? By addressing these questions, I can make data-driven decisions to improve my content, engage my viewers effectively, and maximize my income on YouTube.

2 Literature Survey

In recent years, several studies have delved into the world of YouTube. One study by Gupta et al.[1] focused on using machine learning to develop models for YouTube's ranking mechanism and predicting views. They considered features like likes, dislikes, comments, and subscriber count to understand what drives video popularity. Another article by Yang et al.[2] investigated user engagement specifically with online science videos, uncovering patterns in video duration, social endorsement cues like likes, and their influence on viewer engagement. Additionally, Bärtl[3] conducted a statistical analysis examining YouTube channels, uploads, and views over the past decade. Their findings revealed the dominance of a small percentage of channels in obtaining the majority of views and emphasized the importance of genre selection for a lucrative channel.

While the existing literature gives great insights, my project aims to extend previous research by using the latest dataset of the top YouTube videos. By considering additional features and applying advanced machine learning algorithms, I will compile recommendations for content creators such as myself to enhance their YouTube prowess.

3 Proposed Work

The proposed work will involve the following steps:

3.1 Data Collection and Preprocessing

I will download a dataset instead of collecting my own data. I will preprocess the data by cleaning columns, removing inaccurate data, handling missing values, and converting relevant variables into appropriate formats.

3.2 Feature Engineering

I will engineer additional features from the available data. This may include creating new features such as video duration and tag categories (e.g., food, cars, etc.), as well as adding additional YouTube channel datasets to bolster information about subscribers and creators. By creating meaningful features, I can isolate the elements that have a significant impact on viewership.

3.3 Analysis and Modeling

I will begin by performing exploratory data analysis to understand the distribution and relationships of different variables. Statistical analysis will help me identify correlations between variables and potential predictors of views and other predictive indicators. In addition, I will be able to examine any interaction that may occur between variables that will affect my conclusions.

To predict view count and popularity, I hope to implement machine learning models tailored specifically to my dataset. Regression models will be trained using features such as likes, dislikes, comments, and other relevant attributes. Additionally, I will explore the use of advanced techniques like deep learning, natural language processing, and sentiment analysis to extract additional information from video descriptions and tags, if time allows.

4 Dataset

The dataset used for this project is a daily record of "top trending" YouTube. It includes data for many different countries, with up to 200 listed trending videos per day. However, I will be focusing on my market of interest: The US. The dataset encompasses video titles, channel titles, publish times, tags, views, likes, dislikes, comment counts, thumbnails, and additional metadata. This dataset, collected using the YouTube API, plays an important role in understanding the predictors of user behavior and tendencies on YouTube.

Access the dataset from the following URL:
Trending YouTube Videos Dataset

The dataset is downloaded on my machine and uploaded to the courses Jupyter environment. If I have additional time, I would like to add a dataset that includes additional information about the channel for each video.

5 Evaluation Methods

The evaluation of my models will involve the following:

Metric Evaluation: I will use mean squared error, mean absolute error, and R-squared to measure the performance of my predictive models. These metrics will quantify how well my models can predict the outcomes I have chosen in the analysis process.

Comparison to Existing Solutions: I will compare the performance of my models against existing solutions from the studies mentioned and from more casual data analysis on data-sharing sites. This comparison will help me understand if my techniques yield improvements in prediction accuracy. This project approximately replicates Gupta et al. [1] with a more recent dataset. Results will be compared to measure the effectiveness of my approach in the final report.

6 Tools

My project will be implemented using a combination of programming languages and tools. I plan to use Python as the primary programming language due to its ease of use for data analysis and machine learning, as well as my comfortability with the language. The following libraries and tools will be used:

1. Python: for data preprocessing, feature engineering, modeling, and analysis.
2. Pandas: for data manipulation and preprocessing.
3. Scikit-learn: for machine learning algorithms and evaluation metrics.
4. Matplotlib and Seaborn: for data visualization and exploratory analysis.

If I have additional time, I would like to explore the following advanced libraries:

1. TensorFlow: for building and training deep learning models.
2. Natural Language Toolkit: for natural language processing tasks including review of the more verbose description and tag fields.

As a result of the initial exploration, I found the following libraries to be additionally useful:

1. Isodate: to parse out the date formatting from YouTube

7 Milestones

To ensure progress within a 7-week timeline, I have created the following milestones:

- Milestone 1: Data collection and preprocessing completed (by Week 1).
- Milestone 2: Feature engineering and exploratory data analysis completed (by Week 2).
- Milestone 3: Development of predictive models (by Week 4).
- Milestone 4: Testing and evaluation of predictive models (by Week 5).
- Milestone 5: Exploration and implementation of advanced machine learning libraries and models (by Week 6).
- Milestone 6: Completion of the project report and presentation (by Week 7).

7.1 Milestones Completed

- Milestone 1: Data collection and preprocessing is complete. The YouTube dataset was loaded successfully. I cleaned the data by handling missing values and converting data types as necessary.
- Milestone 2: Feature engineering and exploratory data analysis are complete. New features were created, including the video duration in seconds, the day of the week, time of day, and tag/title keywords. Exploratory data analysis was completed.
- Milestone 3: Development of predictive models is started. Linear Regression and Random Forest models are included and currently being revised.

7.2 Milestones To Do

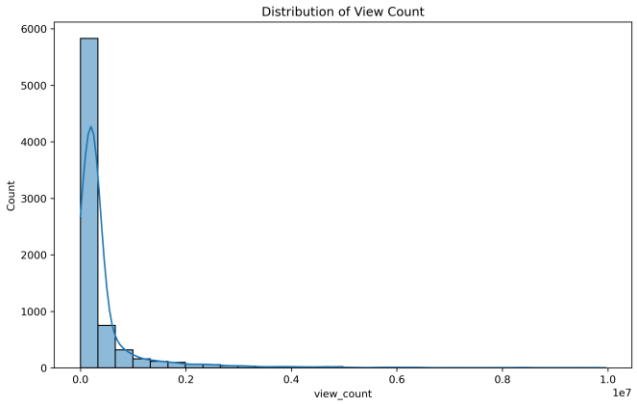
- Milestone 4: Testing and evaluation of predictive models.
- Milestone 5: Exploration and implementation of advanced machine learning libraries and models.
- Milestone 6: Completion of the project report and presentation.

8 Results So Far

Data Cleaning and Preprocessing: I am working with a subset of the data that has 7629 rows. I hope to be able to process around 50,000 rows for the final data analysis. The data was first cleaned, which involved handling missing values and converting data types as necessary. For instance, there were some missing values in the 'tags' and 'description' columns which were filled in with an empty string. The 'publish_time' was converted from a string to a datetime object for better handling. The 'duration' was converted from an ISO 8601 format to total seconds for easier analysis and modeling.

Feature Engineering: I created several new features from existing ones. The 'publish_time' was split into 'day_of_week' and 'time_of_day' to see if these factors might influence the view count. 'Duration' was converted to total seconds. I also parsed the 'tags', 'title', and 'description' to identify the most common words and added those words as columns to the data frame.

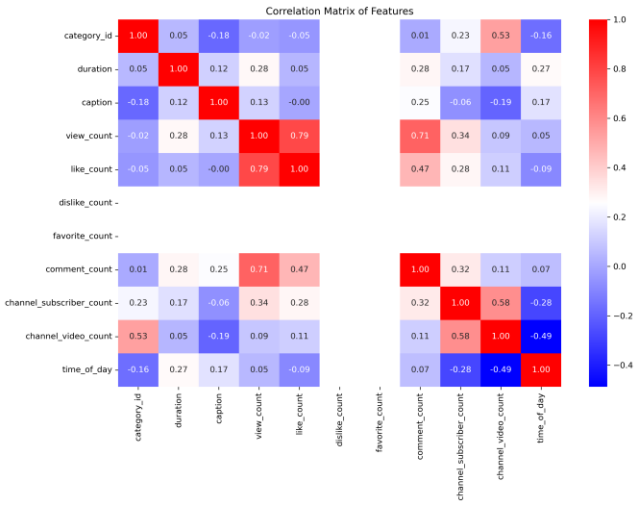
Exploratory Data Analysis: For the numerical features, I found that the distributions of view counts, likes, dislikes, and comment counts were heavily skewed to the right, suggesting a small number of videos receive a very high number of views, likes, dislikes, and comments. This matches with real-world intuition about the world of YouTube.



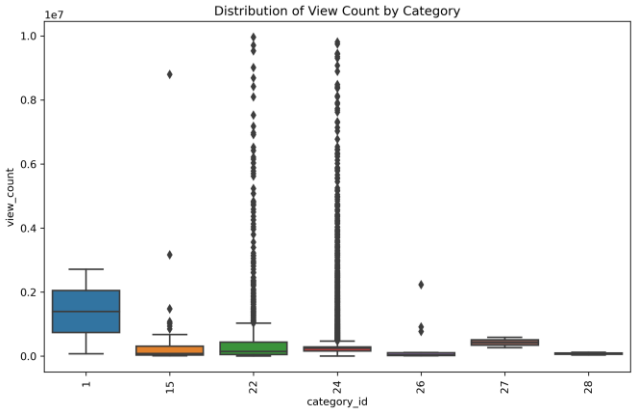
Here are the summary statistics for view count:

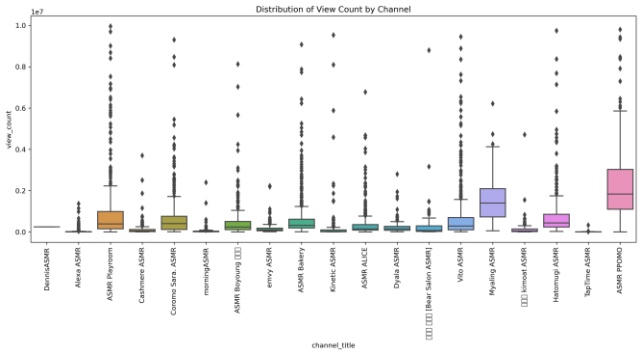
Count	7.629000e+03
Mean	4.622653e+05
Std	9.290508e+05
Min	0.000000e+00
25%	1.102170e+05
50%	2.461580e+05
75%	3.009900e+05
Max	9.955263e+06

I also found positive correlations between the view counts and likes, dislikes, and comment counts, which suggests that videos with more engagement (likes, dislikes, comments) tend to have more views. There is also some evidence of collinearity, which I will address in the testing and evaluation phase.

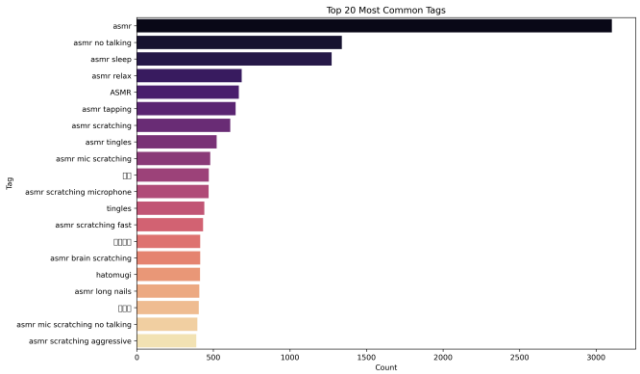


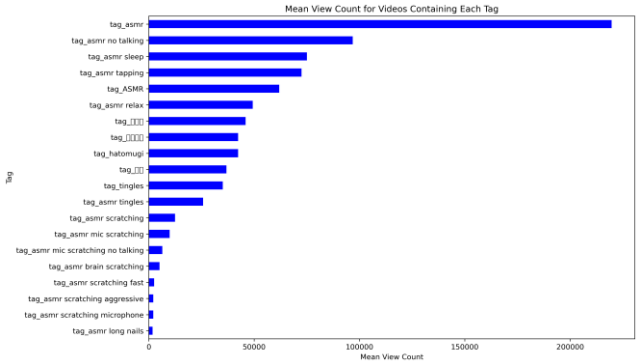
For the categorical features, I found that certain categories have significantly higher average view counts than others, and that certain channels also tend to have higher average view counts, with outliers in both features.

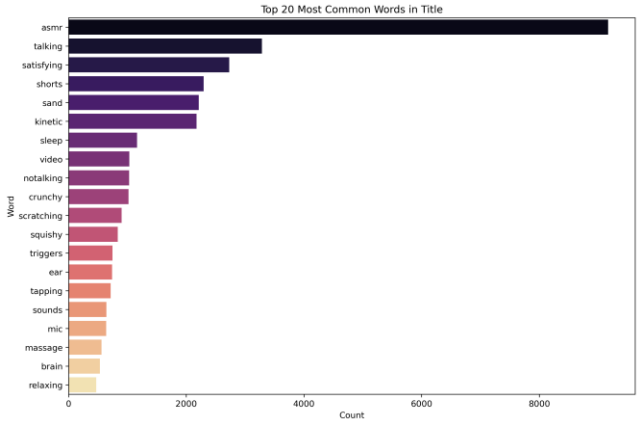


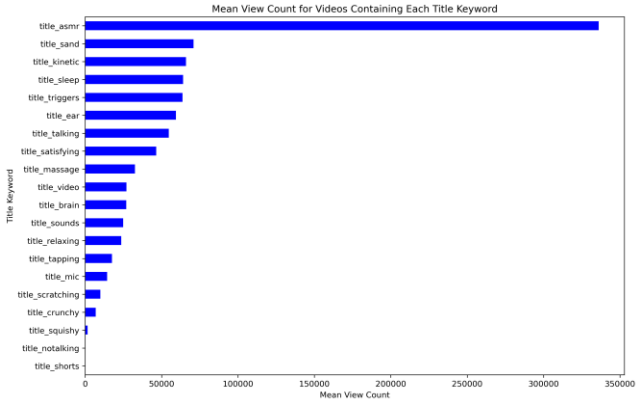


Text Analysis: I performed text analysis on the 'tags', 'title', and 'description' columns of the videos to identify the most common words and phrases. After combing through the output with a real-world eye, the top tags are ‘no talking’, ‘sleep’, ‘relax’, ‘tapping’, ‘scratching’, and ‘tingles’. The top title keywords are ‘satisfying’, ‘no talking’, ‘shorts’, ‘kinetic sand’, and ‘sleep’. The mean view count for videos containing each tag or keyword were calculated to determine which words had the most views.









Predictive Modeling: Linear Regression was performed and the initial output was:

Mean Squared Error (MSE): ~118232226129.28
R-squared: ~0.82

The MSE is very large which indicates that the model's predictions are quite far off from the actual values. The R^2 value is good, which indicates that the model is explaining the variation in the view count well. However, this is more of a baseline model for now, simply because I assumed the data would be more complex than a linear regression model was capable of handling. I will still do some tuning to this model to see if there is anything I can do to improve the MSE.

A tree-based Random Forest model was attempted next. Here are the summary outputs:

MSE: ~89462381650.91
R-squared: ~0.87

Because the MSE is lower than in the Linear Regression Model and because the R^2 value is higher, I feel comfortable saying that the tree-based model is better for this dataset and inquiry.

Upcoming Explorations: I will attempt the following techniques to attempt to improve performance of the models in the testing and evaluation phase:

- Feature Selection
- Hyperparameter Tuning
- Including/excluding Outliers
- Cross-Validation

Once I am satisfied with my model, I may look into doing sentiment analysis for the titles and descriptions to add an additional feature to the mix.

9 Conclusion

As someone aspiring to create YouTube content as an income stream, I recognize the importance of data analysis in making informed business decisions. Through this project, I hope to gain knowledge that

will provide inspiration, motivation, and data-driven excitement as I start my YouTube journey.

REFERENCES

[1] Vandit Gupta, Akshit Diwan, Chaitanya Chadha, Ashish Khanna, Deepak Gupta. Machine Learning enabled models for YouTube Ranking Mechanism and Views Prediction. arXiv:2211.11528 [cs.IR]. Available at: https://arxiv.org/abs/2211.11528.

[2] Shiyu Yang, Dominique Brossard, Dietram A. Scheufele, Michael A. Xenos. The science of YouTube: What factors influence user engagement with online science videos? PLOS ONE, 17(5), e0267697. DOI: https://doi.org/10.1371/journal.pone.0267697.

[3] Mathias Bärtl. YouTube channels, uploads and views: A statistical analysis of the past 10 years. Convergence: The International Journal of Research into New Media Technologies, 24(1), 1354856517736979. DOI: https://doi.org/10.1177/1354856517736979.