Chandler Ebrahimi

Final Report

YouTube Like Predictions for Trending Videos

Problem Statement:

For mainstream youtubers it is important for their video to hit the homepage for trending videos. The goal is to find out what criteria and conditions need to be met in order for a YouTube video to be trending on the homepage of YouTube. The purpose is to give data to youtuber's a goal of what numbers to hit on their video to hit trending.

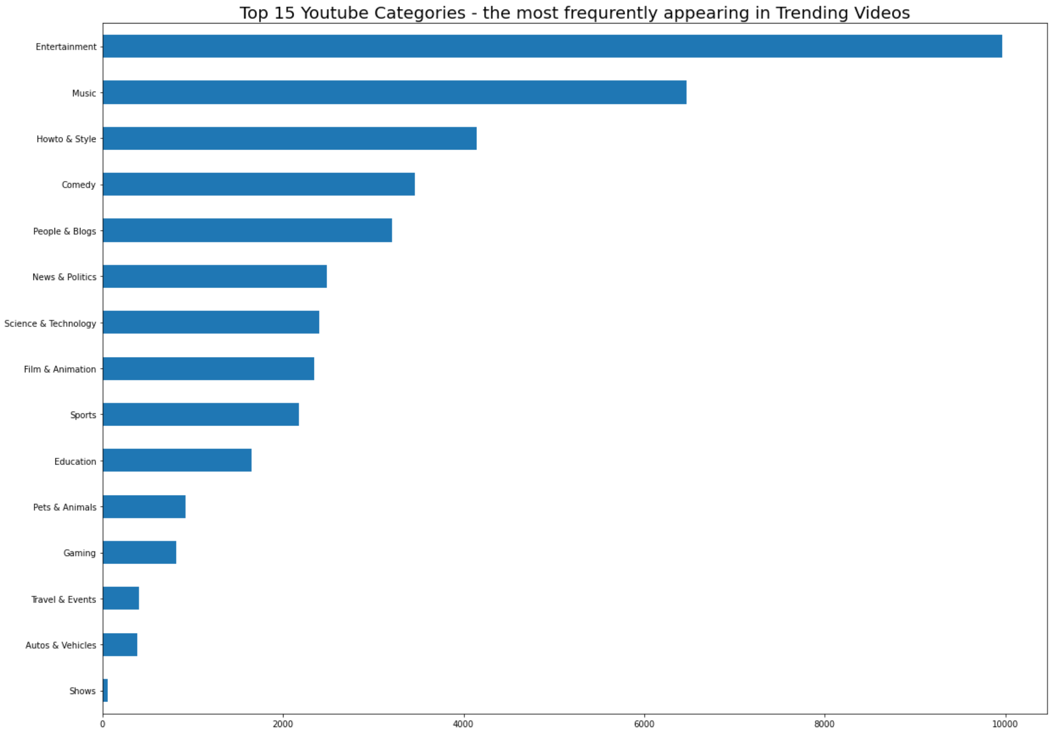
I will be using a dataset that contains trending video that occurred in the year 2017 from several different countries, that was provided on kaggle as CCO: public domain, free of use. In this dataset I will be using a combination of the following attributes such as trending date, category ID, video publish date, tags on video, views, likes, dislikes, comment count to determine how a video becomes trending.

Data Wrangling:

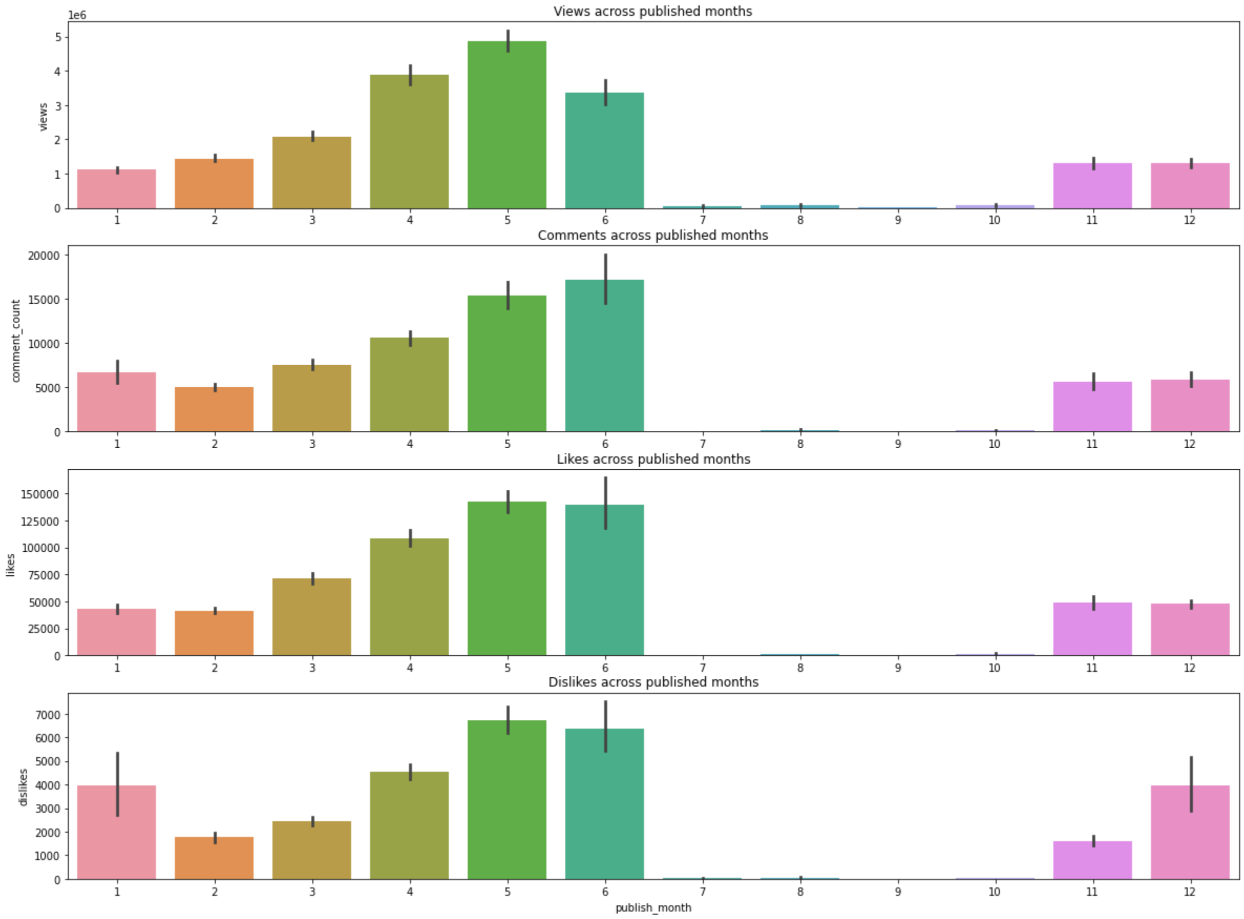
The original dataset has 40949 rows with 16 columns with most of the data coming from the year 2017. Some of the prominent columns or features of the dataset were views, likes, dislikes, comment count, and category id. Category Id had a supporting json file that I had to convert and merge content into the data frame. Some of the features that were dropped quickly were thumbnail_link, ratings_disabled, video_error_or_removed, comments_disabled, description. The final features that were going to be used for EDA are: trending_date, publish_time, views, likes, dislikes, comment_count, category

Exploratory Data Analysis:

To start off I wanted to plot which category was most often trending on the home page. As seen below Entertainment, Music, and Howto & Style are the 3 most frequent appearing in Trending Videos



Breaking up the numerical features likes, dislikes, comment count, and views into each month we can see which month(s) increase your chance of getting your video trending.


Views across published months

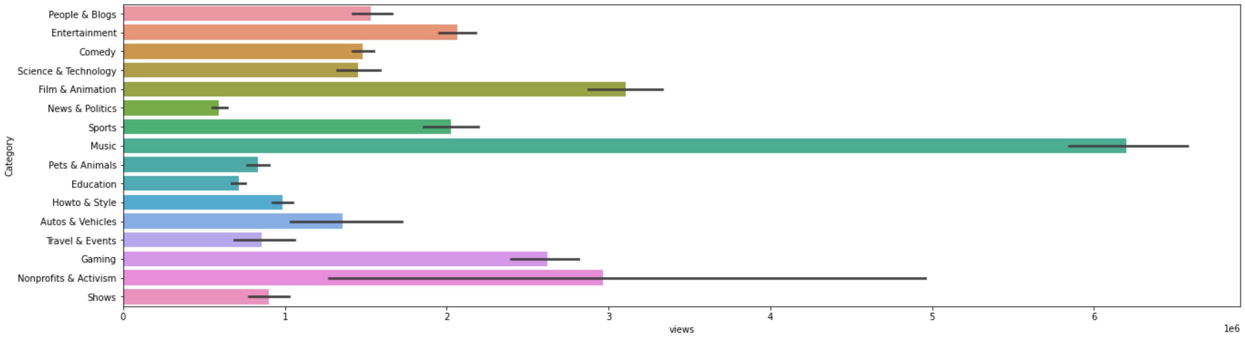
Comments across published months

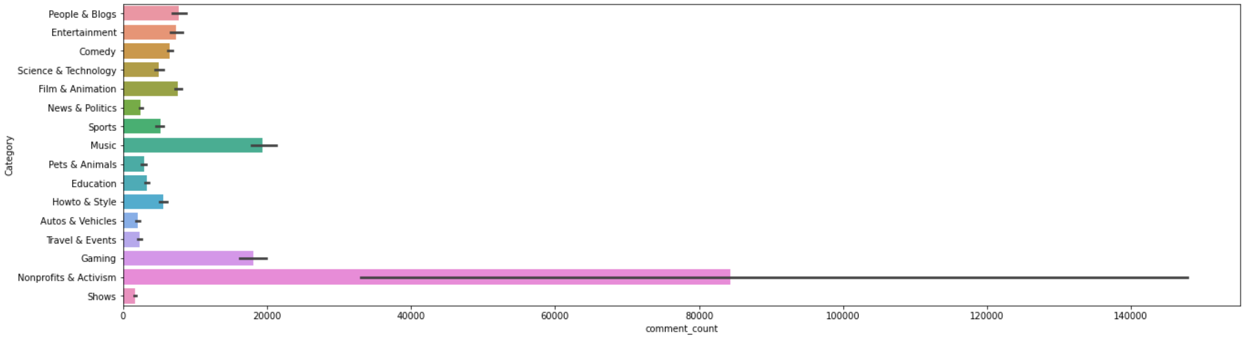
Likes across published months

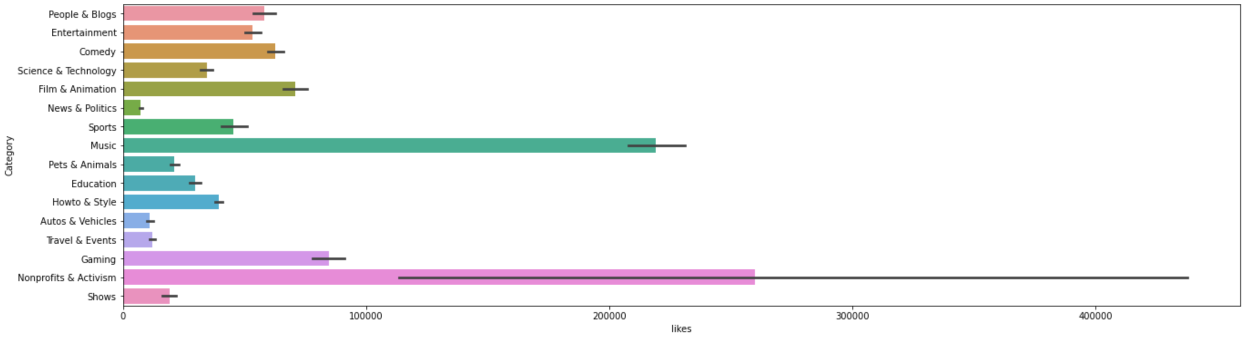
Dislikes across published months

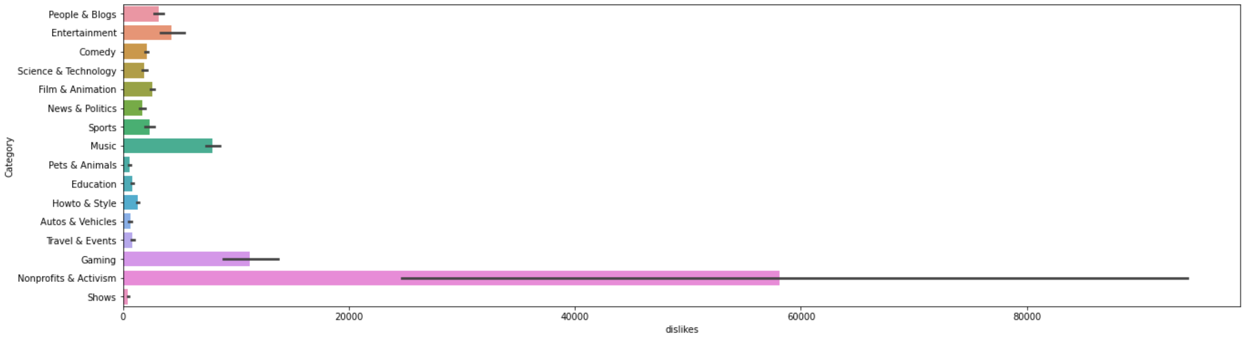
The summer months (5,6) appear to have the most amount of general attraction, as well as the beginning and end of the year to a certain degree.

Next I broke up the same numerical features likes, dislikes, comment count, and views, grouping them by the categories to see if their were any outliers or interesting trends to notice.



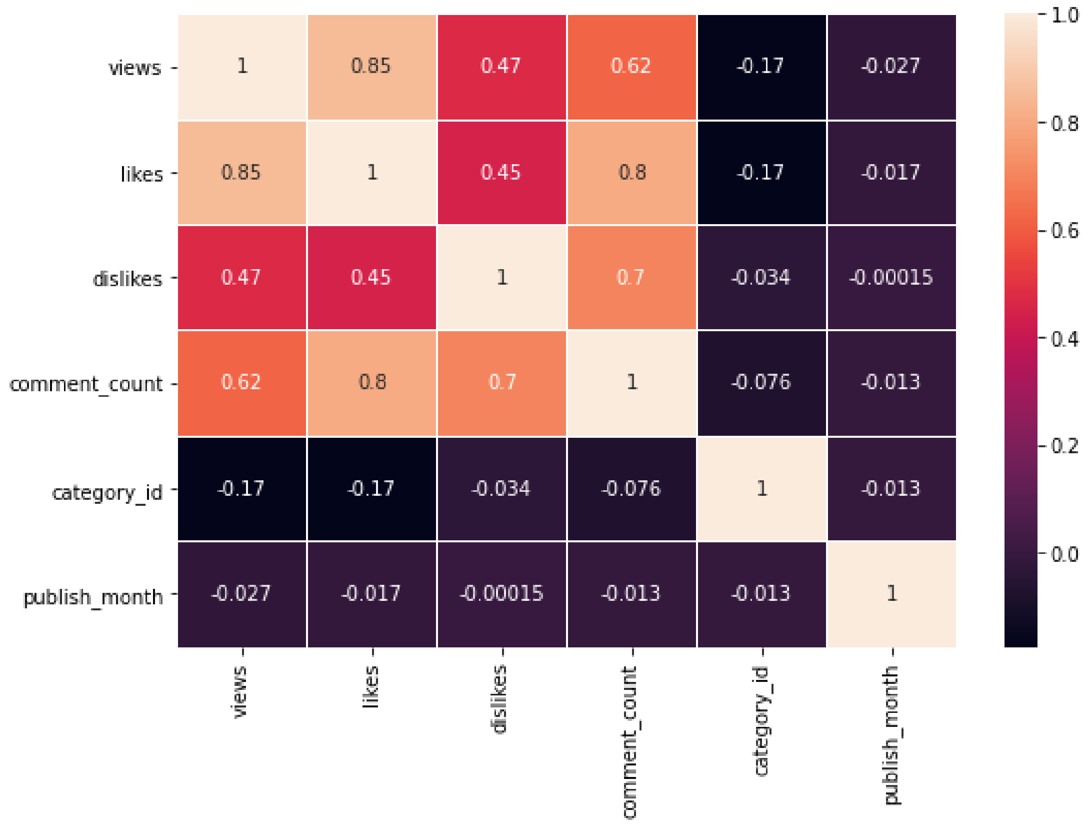






Comparing the numerical features to category Nonprofits & Activism has the highest average, however they also have the biggest standard deviation as well.

I focused more on the numerical values because after doing the correlation between them I saw that naturally likes and views had a strong correlation with each other.



Modeling:

I tested modeling with 5 different modeling techniques: Linear Regression, Random Forest Classifier, Decision Tree Classifier, K-Neighbors Classifier, and SVC. I took the numerical values of likes, dislikes, comment count, and views and split those into a test_train_split, with likes being the target value.

After getting the results back, one model came out more successful in testing than the others. The linear Regression model came back with an accuracy of 89.22%, while the other models were less significant. The Random Forest Classifier had 46.4%, the Decision Tree Classifier had 47.2%, the K-Neighbors Classifier had a puny 1. 5%, and SVC had even worse, .4%.

Future Research:

After doing this project I realized I could implement categories into a better way, perhaps create dummy variables for them and separate the modeling by category to see, in each category what kind of numbers does a video need to achieve for it to be on the trending YouTube homepage.

Furthermore, the dataset included data from different country which could provide meaningful conclusion on which categories of YouTube videos where could provide better results.