Youtube Statistics Analysis

Jason Jackson & Andrew Watson

Introduction:

Our analysis delves into YouTube statistics, specifically focusing on different creator's overall viewership and subscriber numbers, to find factors which can improve a channel's performance on the platform. Our dataset was obtained using a web scraper and YouTube API's. We aimed to determine the connection between subscribers and average views per video, subscribers and total videos, outreach and video count, and total views and community score. We found positive correlations between subscribers and average views per video, outreach and video count, and total views and community score. We found no correlation between subscribers and total video count.
Presentation: Data Science Presentation 3
GitHub Repository: https://github.com/Drew-Watson-117/cs5830_project_3

Dataset:

The dataset used in our analysis proved more difficult to obtain than we had hoped. We began by acquiring an API key that is granted via YouTube and Google. We then attempted to make a scraper using that API key, that would give us data about different channels based on top charts. This proved to be more difficult, as we found that in order to use the API key, we needed the creator's channel ID, which is not easy to find. To remedy this, we created a python script that scraped YouTube to find channel IDs, then fed that information into another python script to obtain our data. The scraper used the Selenium library to scrape the dynamic YouTube webpage for YouTuber ID's. We finished with a variety of data, including over 200 channels with their username, view count, subscriber count, and video count. The set includes the top 100 most subscribed to channels on YouTube, as well as a variety of differently sized creators that fill out the rest of the data. The dataset consists of three primary variables: view count, subscriber count, and video count. From these three primary variables, three secondary variables can be derived:

- Average views/video: The average number of views a video on a particular channel gets. Calculated via (View Count)/(Video Count). This is what we should use for what "success" on YouTube looks like.
- Outreach: A measure of how much a channel's content extends beyond its subscriber base. Calculated via (View Count)/(Subscriber Count). This is what you had in your analysis for the size of the points.
- Community Score: A measure of how a channel's community size compares to the amount of content on the channel. Calculated via (Subscriber Count)/(Video Count).

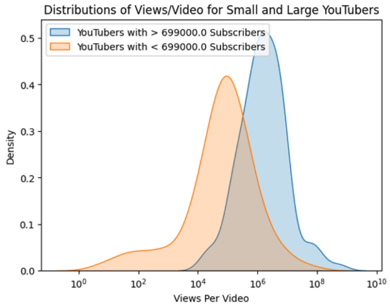
Analysis Technique:

For our analyses, we used a mixture of scatter plots and histograms. When appropriate, we included lines of best fit with scatterplots, and used kernel density estimates rather than histograms. When one variable was being compared across two populations, t-tests were used to determine statistical significance. When two variables were being compared within a single population, Pearson tests were used to determine statistical significance.

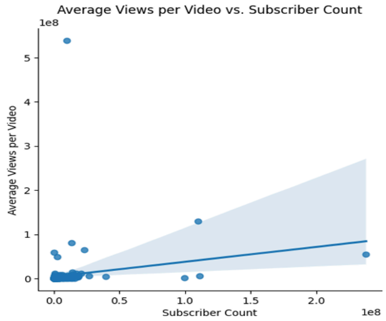
Results:

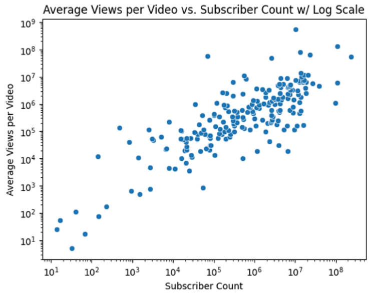
1.1 **The Relationship Between Average Views/Video and Subscriber Count**

The first thing we wanted to look at was the correlation between average views per video and subscriber count. By doing this, we can figure out if it is better for streamers to focus on getting more views to increase revenue, or focus on subscribers first, then views will come along with those subscribers. We first separated our dataset into small and large creators. Then we plotted the average views per video for both.



As you can see from the figure above, there is a difference in the distributions of views per video for small and large creators. This, coupled with the large difference in the median views per video supports the claim that there is a large difference in the average views per video of large creators and small creators. We calculated a p-value of 0.06 which close, but not statistically significant, still helps give more credibility to the claim that large creators get more views than smaller creators on average. We then plotted average views/video against total subscribers to find more evidence.

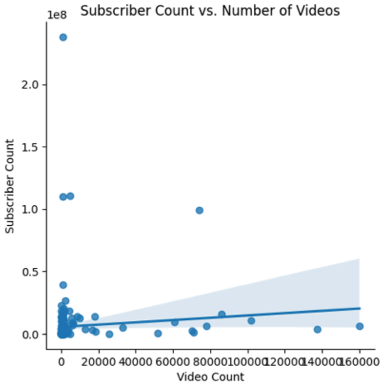




The figures above show that there is indeed a positive correlation between average views per video and total subscribers. This does make sense considering the more subscribers a channel has, the more viewers will regularly view their content. This will raise their average views per video. The low p-value of 0.0086549 supports this claim. There are some outliers being popular musicians, as they have a much smaller amount of videos, and each video tends to go viral. Overall, we can claim that channels with more subscribers will get more views on average than creators with less subscribers.

1.2 Subscribers vs Video Count

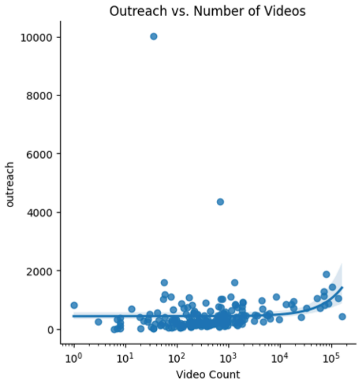
The second claim we wanted to dissect was if the number of subscribers to a channel is proportional to the number of videos that the channel has uploaded. This can show whether a good strategy to growing a fan base is to make more content, or focus on quality of videos.



The figure above suggests that there could be a weak correlation between the total number of videos and its subscriber count. The p-value we calculated was 0.19, which states that this is not significant. This tells us that quality of the videos is much more important than quantity of the videos.

1.3 Outreach vs Video Count

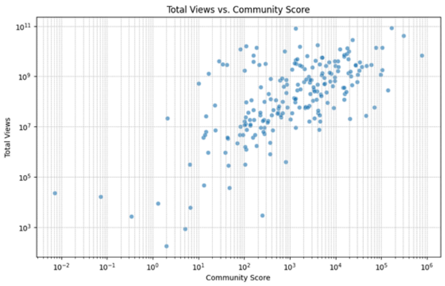
We then decided to look into if uploading more videos helps a content creator reach audiences outside their current subscriber base. We plotted the channels' outreach against their video count to see if there was a correlation.



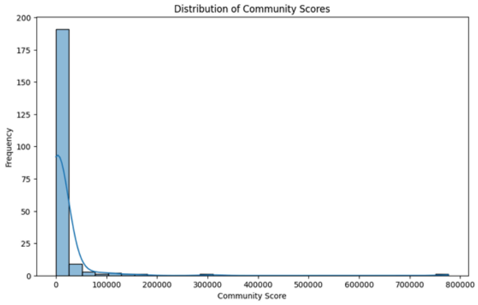
The plot showed an upward trend in outreach as the total videos increased, suggesting a positive correlation. Our p-value was 0.02, so the claim is positively correlated which makes sense, given that the more videos a creator puts out, the more likely they are to reach more audiences.

1.4 Total Views vs Community Score

The last relationship we looked into was the correlation between a healthy channel community and obtaining more views. To do this, we use a community score, and then compare that against total views.



This plot shows there is a significant positive correlation between the community score and the view count. This indicates that channels with a higher subscriber to video ratio, are associated with more views. Our p-value calculation for this plot was 0.0001002, hence confirming our claim.



This histogram shows a right skewed distribution, suggesting that most channels have a larger number of videos that is relative to their subscribers, meaning a lower community score. Few channels actually earn these high community scores. We calculated a p-value of 0.00402, which hints that channels with a healthier community have more views.

Technical:

The dataset required lots of preparatory steps before analysis. First, we had to obtain the channel ids from every channel we wanted to use. Then we had to use a YouTube api key with those channel ids to obtain the data we wanted. Then we did some calculations, such as average views per video, outreach, and community score. We calculated community score by dividing the subscriber count by the video count for each channel.

We used many different analytic techniques including descriptive statistics, correlation analysis, and t-tests to interpret the relationships and differences between our different pieces of data. The Pearson correlation analysis was suitable for quantifying the linear relationships between these continuous variables, such as subscriber count and view count.

Our analytical process included various hiccups, including contemplating what to analyze without overlapping too much. We ended up making multiple analyses, some of which didn’t tell a story, and some that were too similar to other ideas we had already created. We ended up deciding on the 4 analyses that told the best story overall given what we were trying to convey with our data.