Project 3 Report

By Venkata Shiva Sai Dheeraj Kotte and
Bradley Leavitt

Introduction

For our data analysis project we chose to collect data about Youtube. Specifically, we wanted to find data on the most subscribed and viewed youtube channels. We were able to collect attribute data such as the country each channel originates from, the view, subscribers and video counts, and more.

From this data, we looked for meaningful information on what influences the success of youtube channels. How do different countries and cultures come together through youtube? How does the age of a channel affect its sub count? Does sub count really give you more video views? After all, YouTubers are paid for the number of views they get. Hopefully having more subscribers increases the odds of consistent viewership.

Github link:- https://github.com/Dheeraj0650/cs5830-project3
Presentation link -
https://docs.google.com/presentation/d/1834EsKXIXuWc--gua5uBUbOtLJu2gUtWdyVe80YGY64/edit?usp=sharing

The dataset

Our data was collected from youtubers.me and Wikipedia pages containing lists of the top YouTubers along with the stats we wanted to analyze. The sub-count was by far the most focused feature of the dataset other important ones were: Content type, views, age (by years), number of videos, and country population. Some entries had symbols and data types that needed cleaning/formatting to get raw number data. Taking all the tables obtained we were able to merge them into a few data frames for easy graphing/analyzing.

Here are the links to the sites we scraped:

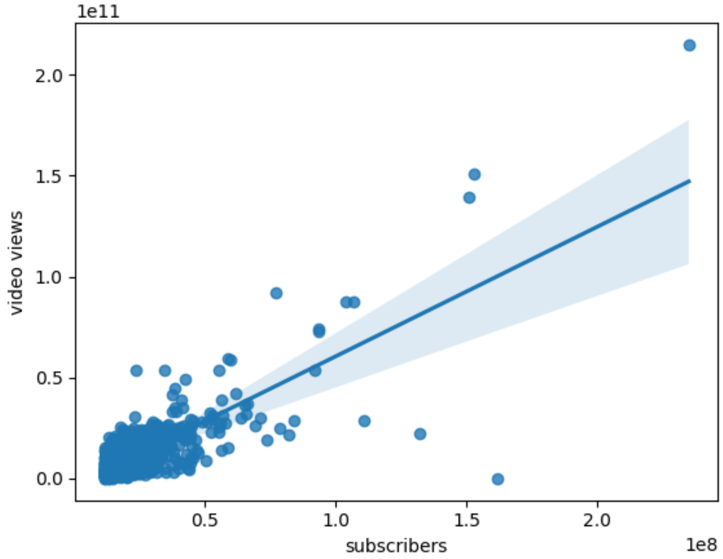
- https://us.youtubers.me/global/all/top-1000-most-subscribed-youtube-channels
- https://www.worldometers.info/world-population/population-by-country
- https://en.wikipedia.org/w/index.php?title=List_of_most-subscribed_YouTube_channels&oldid=1137714711

Analysis Technique

Many of our analysis used grouping by different categories and plotting them. Often numerical data were summed in those categories to show the sheer volume of subscriptions and views each channel is getting. We used Correlations to show the possible influence of attributes on subscriber and view counts. We were also able to use figures such as scatter plots, regression lines, and bar charts to visualize the data. We found scatter plots useful, to sum up, and compare numerical features with many data points. We had the top 1000 channels to visualize and discover meaningful patterns from. Barcharts were useful to visually compare categorical data points such as the language targeted or country of origin.

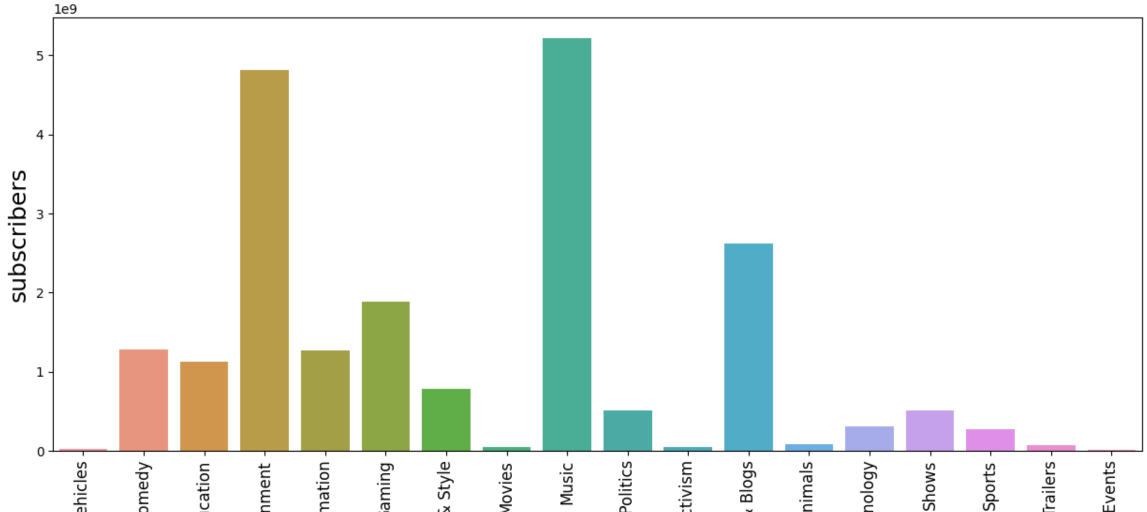
Results

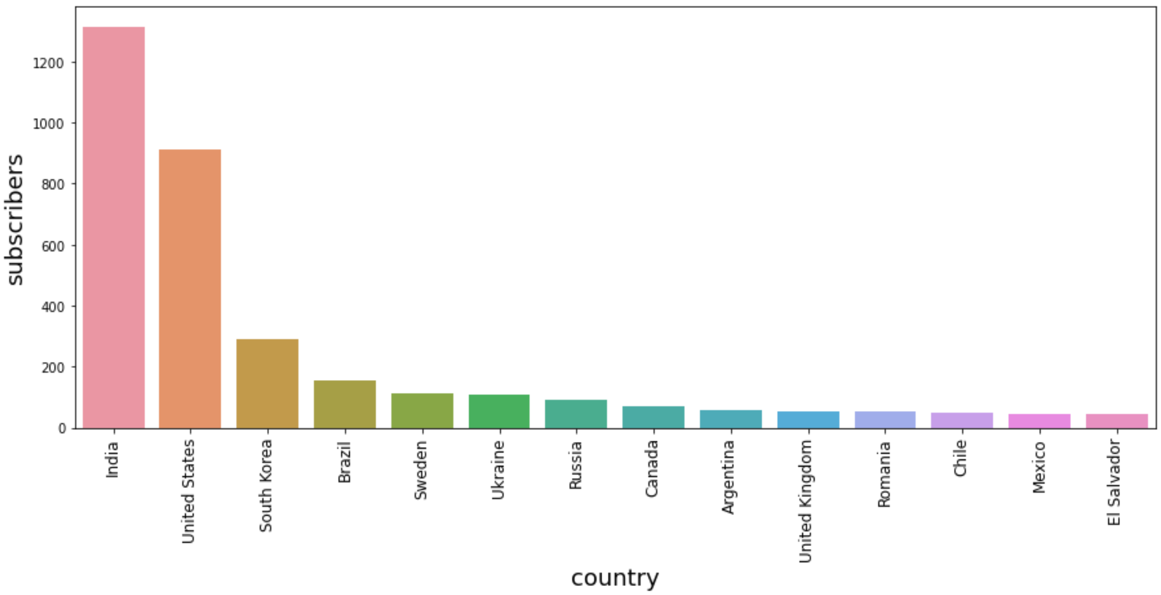
Our **first analysis**. Does having a high subscriber count really increase the views a channel gets? Observing the data displayed, there is in fact a significant positive correlation between total video views and channel subscriptions. With a minuscule p-value to support it! This supports the assumption that subscribers will watch more videos on the same channel because they are updated about uploads. There are some outliers towards the bottom of the plot showing a high subscriber count but low views. There are a couple of ideas that may explain this. First, there are plenty of youtube channels that have fake bot subscribers. They use these to boost the youtube algorithm to be put closer to the youtube trending/front page. Another is that the channel is rapidly growing but has posted a small number of videos compared to other channels hence fewer videos to view.



Correlation coefficient: r= .792, p = 2.149e -209

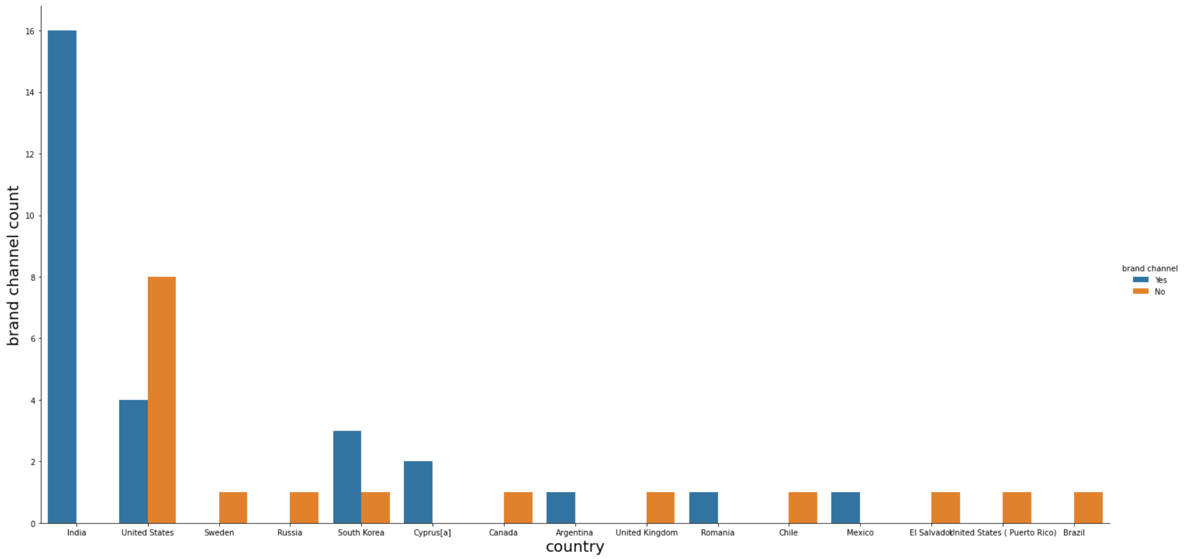
Second analysis. We explored which categorical features a channel with many subscribers would have. Most of the YouTube channels whose category is music have the highest number of subscribers. This is because the majority of YouTube subscribers are from India and Population in India is really high.





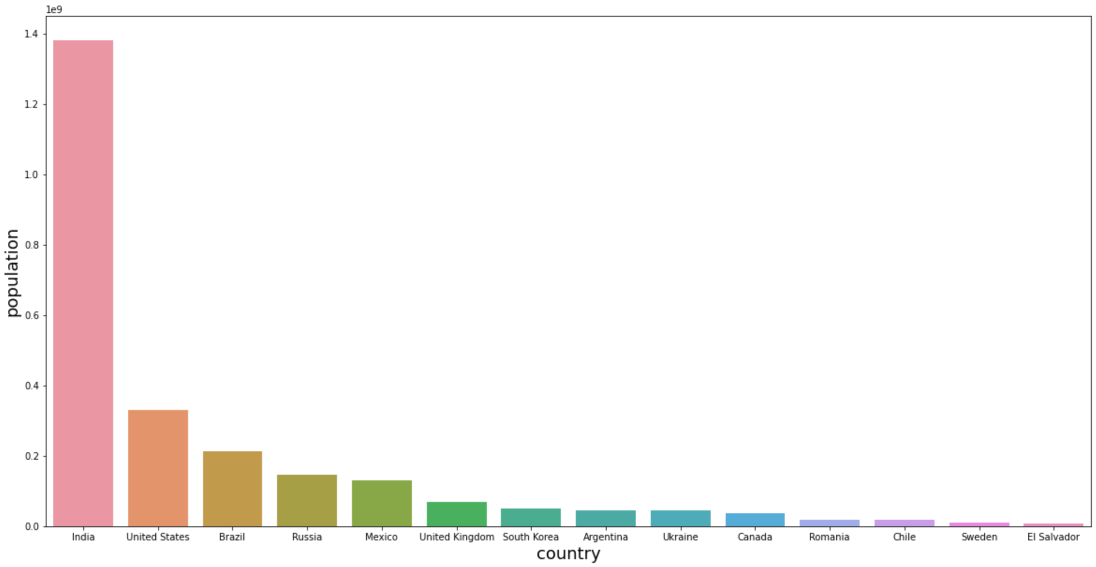
People in India are very much interested in music because of the trends of music in movies, and tradition and it's been one of the stress busters for the people doing jobs in India. Most of the subscribers subscribed to the channels that make content in English and Hindi because they are commonly used languages in most populated countries like the U.S.A. India where they produce most of the YouTubers who give entertainment. Chile has the least number of subscribers because it's very less populated. Subscription count depends on the content we are making, in which language we are making, and from which place we are from.

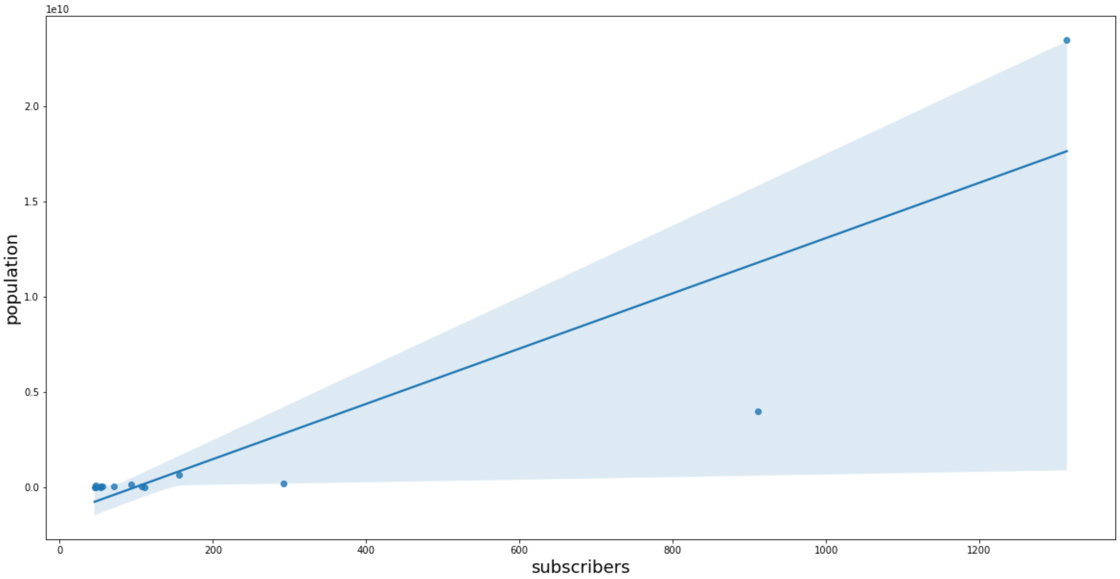
Another interesting find (**analysis 3**) was if the highest channels were branded or not and if they had a company associated with them. (See graph below)



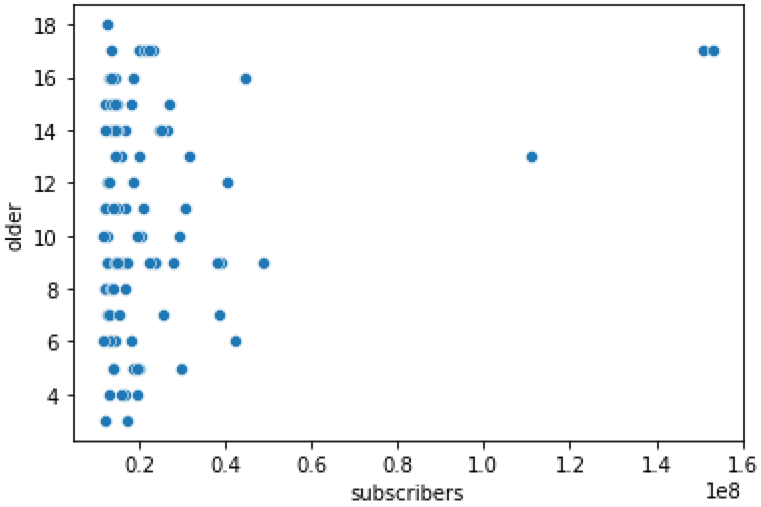
India has a higher number of brand channels than other countries because most of the people in India are middle-class and below-middle-class people. Where in a country like the USA you will find few brand channels and few are not brand channels because here are well settled and you can find many jobs to earn money. Few counties have very few YouTubers and those are the people who are passionate. Most of those people might don't about YouTube and how to use it. The very successful channels in India have businesses backing them and maintaining them.

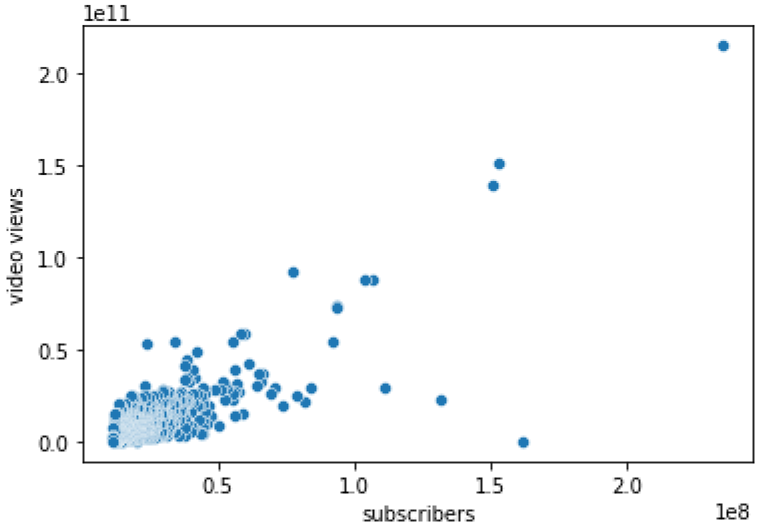
A **fourth analysis** was looking at how a country's population affects the number of subscribers to a channel within the country. India strikingly towers over almost every country population-wise. Second only to China. Why is China not on the list? Well, China has banned most media including youtube, therefore, there are no channels to become big. when a country has more population, it's likely to get more targeted traffic, exposure, and views because most people can get to know about the channel giving more support to the channel. As per the Pearson correlation coefficient of r = .88 when the population increases subscriber count increases.





The **last finding** was that the age in years of a channel is not strongly correlated to the subscriber count. We were given a correlation test stat of .19 and a p-value of .06 which may not be considered significant anyways.





Technical

Things that went well were that even though some preprocessing was necessary where information wasn't available (NaN's). At the same time all the tables were nearly complete so dropping rows with empty spots didn't cause any problems. Dealing with numbers in the trillions gets hectic so Making graphs to visualize it all was extremely helpful. The data frames seemed to merge easily as well. An important part of making the data useful was formatting the numerical data correctly. Numerical data fields were converted to floats or int 64 data types to work with the graphs and for correlation tests. For our analysis with the bar charts, we summed the data on subscribers per category. We did this to reveal the obvious outliers and hoped to gather what made them special compared to the channels just below them. Using Correlation tests for numerical data was useful, especially when comparing numerical features to subscribers it became apparent whether certain variables had an influence on each other or not.

The beautiful soup python library was the best way we found to pull data off our chosen sites. The library itself was efficient at obtaining rows and columns for our dataset. At first, we were hoping to get the top youtube channels through the years to analyze how youtube channels grew. However, the tables made for each update to the Wikipedia page were inconsistent making data of that size difficult to manage. Things we could have done more of is played around with population averages and other numerical features to see how much bias the analysis results have. Also, t-testing distributions would have also been an interesting take. We could explore whether or not channels have a similar distribution of subscribers. Based on content type or language.