YouTube Analytics

Gavin Lampton
CECS
CSULB
Long Beach, USA
gavin.lampton@student.csulb.edu

Brian Cho
CECS
CSULB
Long Beach, CA
brian.cho02@student.csulb.edu

***Abstract*—We present a model to analyze whether the number of subscribers a YouTube channel has is correlated to how many views that channel will receive. Through collecting data from YouTube, cleaning and processing the data, and plotting the data, we observed that subscriber count does not heavily influence how many views a channel will receive.**

I. HTML Collection

To collect the data necessary, we first needed to decide where to focus our attention. We decided to use the top 100 most subscribed channels. This information was pulled from the website Social Blade. This is a website that tracks statistics for various media websites.

Once we had this list, we went to each channel to collect our data. By saving the page as an HTML file, we could parse through it to pull the information relevant to our project. The HTML file would only show information on the videos that had been loaded to that point, which meant we had to scroll to the bottom of the page to load all the videos on the channel. Some channels had hundreds of thousands of videos and attempting to load all of these onto the page would cause it to crash.

To solve this issue, we went into the browser settings and disabled images from loading. This helped with memory management and allowed us to load more videos, but the page would still crash at a certain threshold. This limited our data collection to channels that had under 9,000 videos.

Although an automated method may have been available, the YouTube API requires access to Google Cloud servers. This method was avoided to not incur any costs on our end.

II. HTML to CSV conversion

Two bash scripts were written to collect the information from the HTML file and then turn it into a CSV. The first was mistakenly written for Firefox's html formatting, but only some really early exploratory data was gathered in this format. The other was written for Google Chrome's mhtml formatting, which is the actual format of the gathered data. The script would narrow down the line count by using grep to get the html section which contained video information, then used tr to replace newlines, and sed to actually edit the information into a csv format. The completed CSV had ~ as the delimiter instead of the traditional comma. The initial information collected was the title of the video, the length of time since the video had been uploaded, and a link to the video.

The first issue we ran into with the CSV files was that the upload time was being read as a string in R. This meant that R would incorrectly sort it.

A bash script was created to translate the string containing time since upload into three separate integers to solve this issue. However, the Bash version, after two hours of runtime, only finished four files. The bash script was reworked to only feed filenames to a Python script, which would actually tackle the same issue. Writing it in Python dropped the runtime down to about 5-10 minutes.

III. EDA

A. Round 1

The first round of exploratory data analysis was performed after the first bash script was written, but before the Python script. For the first round, a line graph was created plotting the view of each video to the upload time. Upon inspecting the graphs, we noticed that they were sorted incorrectly. This was due to the time since upload column values. Some examples of these values are "1 month ago 14 minutes, 59 seconds", "10 years ago 1 minute, 22 seconds", and "2 years ago 14 minutes, 14 seconds". When the line graph was created, the values were being sorted in lexicographical order. This led to our graphs being sorted in an unexpected order and the graphs were not usable.

The Python script was then written to correct this issue and read the upload time values as integers.

B. Round II

After correcting the data with a Python script, the CSV was supposed to have the amount of time since upload in years, months, or days. However, the month and day columns ended up unusable because the data itself didn't contain month values for most videos over a year old or day values for videos over a month old. This was a surprise because the strings would still contain minute or second values. This is why the graphs don't make use of the month or day statistics.

With the data available we were able to start looking into two of our questions: which subscribers have the most viewed channels and do older videos still receive views.

IV. Most Subscribed Channels Vs Most Viewed Channels

Total Number of Views for Each Channel

To generalize whether subscriber count was correlated to the number of views the channel's videos would receive, a bar plot was created. We wrote a script in R that would read in the CSV for each file in the directory. The script would then add up the value in the "views" column. Then it would create a bar plot with the channel name and the number of views associated with that channel. Initially, the plot was not sorted in order of most to least subscribers, so it was updated. However, when creating the presentation, we realized that there was an error with a few of the channels and they were placed further to the right than was correct. This bar plot was not scaled to show how much a difference in views there were when comparing channels that had more subscribers to channels with fewer subscribers.

We made a few notable observations from this bar plot. The first was that although it looked like there was a correlation between subscriber count and view count, it appeared to be a weak correlation.

To show how weak or strong of a correlation there was, a linear regression model was added to the bar plot. While there was a line with a negative slope, indicating that channels with more subscribers received more views on their videos, the slope was not very steep. The slope of the line may be gentler due to the error in which some channels were plotted incorrectly and were further to the right than they should have been.

Another observation we made was that most channels had under 50 billion total views. The exceptions to this observation were the channels T-Series, Vlad and Niki, El Reino Infantil, Cocomelon, Like Nastya, and Kids Diana Show. Of these six channels, five of them are tailored for children. Due to the rise of tablet usage with children, it appears that channels with content created for children will do well. The sixth channel, T-Series, is a media channel for India's largest music record label. As the country with the second highest population, it is not surprising that a channel from India would do very well. T-Series is also currently the most subscribed YouTube channel.

Comparing Channels

Next, we created box plots to compare views per channel. The values that we wanted to compare were the minimum, maximum, and average views of each channel which would be best visualized with a box plot.

An R script was written to read in each CSV file from a directory. A box plot was then created using the values in the "views" column. A lot of this code could be recycled from the previous script. On the initial attempt, the range of values was so wide that the box in the box plots would not appear and looked more like a perpendicular line. To fix this, a logarithmic scale was applied to the plot. Using a logarithmic scaled fixed the issue and we now had working plots to sift through.

Something we noticed was the presence of outliers on many of the channels. We took a closer look and found the following observations.

Most of the channels had outliers on the top of the plot, indicating their videos received significantly more views than average. An example of this is the channel for BLACKPINK, a K-pop group. Upon examining the videos in this channel, we noticed that the views of the least watched videos ranged from 466 thousand to 732 thousand, but views from the most watched videos ranged from 496 million to 2 billion. When comparing the least to most watched videos, we found that the videos with the least number of views were lifestyle videos and were more of a vlog style whereas the most watched videos were of music and music videos from the group. Some of these outliers could be from cafes and restaurants adding these videos to a playlist and having it play multiple times a day which would inflate the view count, but the vlogs may be harder to watch for people who are interested only in music from the group.

An example of a channel with outliers on the lower end of the box plot is A4. He is a Russian content creator who makes videos that mimic popular trends. His least watched video had one million views and most watched video had 117 million views. The median video had 28 million views. Out of around 800 videos he has uploaded, most of them have fewer than 30 million views. This led to these outliers in the tail end of the graph. Similar to the BLACKPINK example, we took a look at the videos but it was difficult to see why there was such a large discrepancy between his views. One possible reason is that older videos seemed to be more in line with what he wanted to upload whereas newer videos would follow a trend.

We also compared videos from the most and least subscribed channels to those closer to the middle. Our first comparison was between the second most and 54th most subscribed channels, MrBeast and Bad Bunny. We saw that although MrBeast had more subscribers, Bad Bunny had more views. MrBeast is a content creator that creates videos made for entertainment with outlandish ideas such as giving away extravagant items for free while Bad Bunny is a music artist. Like the BLACKPINK channel, it is possible that Bad Bunny's videos are added to a playlist that is repeated throughout the day and not necessarily watched and paid attention to, but MrBeast videos would have a more accurate view as it would account for people who watched the video.

To account for the difference, we next compared two music channels, Shakira and One Direction. Surprisingly, One Direction had a higher median view count as well as a higher max view count. This could be due to the huge global reach One Direction had. Interestingly, Both comparisons contradicted our hypothesis that subscriber count correlated to view count.

Video Age vs View-count

This section is attempting to answer the question: which years had the most video views? The first attempt looked at median and average view-counts for videos. The median view-counts were a better indicator of overall site viewership, and this seemed to peak around 2018. The problem with average is that it was a better measure of the ratio of popular videos, usually baby YouTube video channel Cocomelon, to other top 100 channels in 2023. The problem is that most top 100 channels from 2023 were not uploading videos before 2010. This issue was the worst in 2006, where the video count was only 3 and they were all from Cocomelon. Cocomelon had a single viral video from 2006 that has almost 200 million views and this brought the average for the year up to 100 million views. This had to be removed because it blew out the scale of the graph and made the median view-count was too hard to see. The important interpretation of the first graph is that video virality is by far the most important aspect of video viewership and that lead to the question: which year had the biggest viral videos?

To approximate the biggest viral videos, we looked at the videos with the largest view-count by year. The most striking part of our entire dataset is contained in this graph: the viewership on the biggest YouTube video is about 1.75 times the approximate 8 billion population of Earth. The maximum view-count suggests that the largest viewership was on videos over three years old, which might be due to changes to the algorithm in 2019. One of the limitations of this dataset is that, again, we only got videos from the most subscribed channels on YouTube. This means that the most viewed videos by year are only the most viewed from the biggest channels. In hindsight it would have been great to collect video data from the largest videos playlist and combine it with our dataset to create a more accurate measure of the most viewed videos. So, to better answer the question from class: our dataset does not include data from Rick Astley. However, it turns out that the view-count of the highest view-count video in our data from 2009 was comparable to Never Gonna Give You Up so Rick-rolling might not be a significant factor compared to other trends. However, Never Gonna Give You Up is an outlier because the view-count is almost 350 times higher than the channel's subscriber count. Maximum channel views vs channel subscriber count would have made for an interesting graph but unfortunately, the final graphing process happened too late in the semester.

Average Video Views vs Channel Subscriber-count

We looked at the average view-count of all videos uploaded by a channel vs the channel's total subscriber count. This scatter plot demonstrates that there is not really a consistent relationship between subscriber count and average view count. The amounts seemed to fluctuate wildly; some of the most subscribed channels had low average view-counts and vice-versa. It would suggest that viral video count is much more important than subscriber count, which would be an interesting angle to cover if there were more time.

Conclusion

In this project we looked at the view-count on videos from most of the top 100 channels on YouTube. We were able to determine that there is a minimal correlation between view-count and subscriber count. We looked at the total number of views for each channel and the total count vs subscriber count and the trendline indicated as much, albeit the outliers had a massive effect on said trendline. When we looked at the box plots for each channel, it appeared that most channels had many outliers above the 25th percentile with a couple exceptions. When we looked at video views vs age it appeared that videos from around 2018 were the most popular in terms of average and median, if you discount the earlier videos where the average was propped up by a smaller number of high subscriber videos. The fact that most of the averages dwarfed the medians also indicates that viral videos are the most important factor in determining view-count. We also looked at the maximum values of videos in each year and found that the largest video on YouTube is monumental in terms of views: it is about 1.75 times the 8 billion population estimate of Earth's human population. Finally, we looked at the average view-count compared with subscriber count for channels and determined that there was no reliable correlation between views and subscribers. This also supports our conclusion that viral videos are more important than subscriber count when it comes to views.

REFERENCES

[1] "Top 100 YouTubers sorted by Subscribers - Socialblade YouTube Stats | YouTube Statistics," Social Blade, https://socialblade.com/youtube/top/100/mostsubscribed (accessed Oct. 13, 2023).