YouTube Analytics

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

Brian Cho  
CECS  
CSULBLong 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.

Keywords—component, formatting, style, styling, insert (key words)

# 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.

# 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.

# EDA

## 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.

##### 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).