**Group 6**

**Emmanuel Ayoo, Katie Buckner,**

**Michelle Goodman, Polina Holland**

**YouTube Trending Statistics**

**Summary Introduction:**

As a group, we spent time searching available CSV’s and decided that one we found that focused on YouTube trending statistics would be of the most interest to all of us, as well as provide a good backdrop on which to practice the skills we have acquired thus far in the class. After much back and forth, we settled on four main categories within which we would develop our questions, and divided them up as follows:

* Emmanuel: Category ID’s
* Polina: Channels, Likes, and Dislikes
* Michelle: Publish Times/Days, Trending Dates
* Katie: Video Tags

We would each take our chosen category and try to find if these things had any impact on the total views of a video. **Could we find those things most likely to help a video trend and gain the most views?** Each of us did our own research, comparing, and narrowing down to create data and charts that we felt represented our data best. For that reason, we will each explain our process with a little more detail here.

**Emmanuel:**

A significant data point for the analysis of our CSV data set was the Category ID column. This column, though consistent, had YouTube videos categorized numerically based on an unspecified format. Each video had a Category ID classification. This was a vital data point in establishing how the Category ID affected the overall views and traffic of a video.

The Category ID was decoded through a query of the **YouTube Data API v3**. This Data API provides access to YouTube Data such as video statistics and channel data. An API call was made and a total of **32 content categories** were derived from our data set.

Once I established what each Category ID classification represented, I was able to compare each category ID with the number of views, likes, dislikes and comment counts. These are key indexes for video trends and channel traffic.

The **Top 5** Category IDs were:

**10. Music**

**24. Entertainment/Pop Culture**

**1. Film and Animation**

**34. Comedy**

**22. People & Blogs**

The indexes were visualized on a bar graph. These were the key conclusions that were derived:

* The Music category had significantly higher view counts and overall traffic count than any other category on YouTube.
* The Entertainment category was second. YouTube also classifies Entertainment as ‘Pop Culture’.
* Though all the other content categories had significant traffic and view counts, videos within Music and Entertainment/Pop Culture content categories have an extraordinarily strong probability of being the most  trending and generate significant channel traffic.

**Polina:**

Like everyone else, I started with getting to know my data by using a number of ways to explore our US YouTube dataset immediately followed by cleaning my data segment.

My individual part was to look closely into popular channels, likes, and views. Is there any relationship between them? How strong are they? What will the charts I created using my coding skills reveal?

My hypothesis was that the most popular channels get more views. If there is also a correlation between views and likes variables then we possibly know the right path for vloggers to make their videos to get more likes. Could I find answers to that with a little over forty thousand row dataset and a few columns that were of interest to me?

After a lot of thinking, planning and data analysis, I am ready to share some of my results:

* To discover the 15 most popular channels, I grouped the data by channel and the sum of total likes. The same was performed on the number of total dislikes. The seaborn bar plots included in my notebooks displayed the visuals, with IBighit, Dude Perfect, Logan Paul Vlogs… TayllorSwiftVEVO… EminemVEVO and the rest of the top 15. The Logan Paul Vlogs channel also got the highest number of dislikes out of the same top 15 channels.. hmm What kind of videos does that channel have??
* I then created a function for linear regression plots. While the Total Views and Total Dislikes indicated by Pearson(r)is equal to a moderate positive relationship of 0.47, Total Views and Total Likes displayed on the plot has r value of 0.85 which is a strong correlation. This tells us about a linear relationship between those variables.
* Having a set with many columns, I utilized a visualization tool (visualized the correlation matrix as a heatmap using the seaborn and matplotlib packages) to quickly check the correlations coefficients among columns. In this heatmap the lighter the color, the larger correlation magnitude. What are variables most correlated with likes? Views = 0.83 c. coef.

My conclusion is that my hypothesis was leading me in the right direction. The relationship between channels, views, and likes indeed exist.

I encountered a problem dealing with large numbers on my plots. I was thinking to simplify them for the visuals at first but decided to refresh my memory about reading scientific notations and be ready to clarify if questions arise. Here is a example: 1e8 in reference to the y-axis mean 1\*(10^8). It is s standard scientific notation, indicating scale factor for the y axis (2 actually indicates 2\*1e8 = 2e8 = 2 \* 10^8 = 200,000,000). Another obstacle was to extract a truly unique number of videos but with the help of great teammates it was done.

While I could share a lot more about my processes, analysis, outliers, limitations, more visuals and results, I have to stop right here and let Michelle talk about her US YouTube dataset adventure.

**Michelle:**

The very first thing I did when I got my topic was create a Jupyter Notebook that had comments throughout outlining what I **thought** would be a good plan of attack. I first had to read in the csv and pull/rename the data that I thought would be the most important to me. I quickly found that one notebook was going to have way too much information in it, and decided that doing separate notebooks for each of my tasks would be a better way to organize my thoughts. A few issues I ran into while doing my analyses:

* I realized that each video was listed for each date that it trended. So one video could be listed 15 times. And unfortunately, the total views were cumulative, and not individualized to that day. So I had to figure out how to sort the data and then only keep the final days’ total views.
* The times and dates were in an unusable-to-me format, so I spent quite a bit of time figuring out how to pull out only the information I needed.
* Since I was dealing mostly with timeframes, I quickly found out that scatterplots were not the way to go for me.

Finally, after trying lots of different ways to chart the information, to organize it, and to compare it to each other, I came up with answers to my questions:

* **Does the number of days a video is trending affect the total views it has?**
  + I found that the longer a video trended, the more total views it got in general.
* **Does the month of the year a video is published affect the number of days a video trends?**
  + I found that videos published from April to June trended for more days total in general.
* **Does the month of the year a video is published affect the number of views a video has?**
  + I found that videos published from April to June trended for more days total in general.
* **Does the day of the month a video is published affect the number of days a video trends?**
  + I found no correlation between the day of the month a video was published and the number of days a video trended.
* **Does the day of the month a video is published affect the number of views a video has?**
  + I found no correlation between the day of the month a video was published and the number of views a video received.
* **Does the hour of the day a video is published affect the number of days a video trends?**
  + I found no correlation between the hour of the day a video was published and the number of days a video trended.
* **Does the hour of the day a video is published affect the number of views a video has?**
  + I found no correlation between the hour of the day a video was published and the number of total views a video received.

**Katie:**

My part of this project involved analyzing what effect, if any, tags may have on driving viewership of a YouTube video. I was interested in looking into tags since they represented a qualitative value of our data set and my career in the past has had me working with a lot of qualitative data.

When I first started out working on this project, my goal was to see if there were any particular words that might drive YouTube traffic. I was hopeful that there were some words that appeared in videos more often that others. I was also curious if tags even mattered in driving views to a YouTube video.

Like my colleagues, I started out cleaning up my data so that it would be more usable for my analysis. I first decided I wanted as clean a DataFrame to work with as possible before I even began filtering my data. To do this I decided to split the publish time column into two separate columns, one representing publish time and one representing publish date. I then created a new DataFrame where I renamed columns, dropped ones I did not think would be useful for my analysis, and I removed the quotations from the “Tags” column so they would not get in my way later.

The “Tags” column to start out with posed a bit of a challenge, as each video had multiple tags separated by a “|” delimiter. Regardless, I first decided to do a value\_counts on this column to see if there were any multiple tag sets that matched. As it turned out, the most common tag set was [none], but some videos did have the same multiple tag sets. A tag set of “The Late Show|Stephen Colbert|Colbert|Late Show|celebrities|late night|talk show|skits|bit|monologue|The Late Late Show|Late Late Show|letterman|david letterman|comedian|impressions|CBS|joke|jokes|funny|funny video|funny videos|humor|celebrity|celeb|hollywood|famous|James Corden|Corden|Comedy” appeared in 25 of the 6351 videos.

Since [none] appeared so often in the “Tags” column I decided that it was worth some analysis of this value and focused my analysis on whether placing tags on a YouTube video really could account for more views to a video. I decided to create a sample of videos with Tags that was the same size as my No Tags sample in order to have a better statistical analysis. I discovered the following:

* The Tagged Video sample had 419,730,308 total views and the No Tag sample had 346,174,150 views. This is a difference of 73,556,158 views.
* Tags also play a role in the number of likes a video can expect to receive, the Tagged Video sample received 14,301,167 likes and the No Tag sample had 6,356,362 likes. This is a difference of 7,944,805 likes.
* Finally, using Tags can also generate comments on your video. The Tagged sample received 1,740,070 comments, while the No Tag sample received 685,975 comments. This is a difference of 1,054,095 comments.

It is fair to say that placing tags on your video can help you get more viewer engagement with your videos and more views. Based on Polina’s analysis, we know that comments and likes have a positive correlation with views. Since these things play a role in causing a YouTube video to gain more views and since Tags appear to have a positive correlation with comments and likes, it would be a good idea to attach Tags to your video if you wanted the maximum amount of views.

So in conclusion, we all agreed that the following inferences would be most helpful in posting a video that would trend and get the most views:

* Publishing in the Music category is definitely your best choice!
* Post in a well-visited channel if you can!
* For the most bang, publish your video in the late spring...April, May, or June
* Use tags! Any tags!

Working in a group presented challenges above and beyond just those working with the data, but overall this has been a good learning experience that will give us tools for our data analysis toolbelt.