**An Analysis on YouTube Data on Most Populated Categories**

By: Jack Carr, Dept. of Computer Science, Hood College, Frederick Maryland

**Abstract**

In this paper, I explore a select sample of YouTube videos and analysis trends that appear from it. The dataset includes data from YouTube videos as well as the data on the comments from said videos. The years that are included with this dataset range from 2009 to 2022. With this report, I hope to show some of the most popular trends, tags, dates, and comment sentiment on the selection of YouTube videos. Some predictions I had for the data where that there would be less activity in the beginning of the year than the middle of the year. Meaning that videos were getting less likes, views, and comments near the beginning of the year than later on in the year. It was determined that while yes videos were receiving less attention during the beginning of a given year, videos in the dataset where given more attention in the 2nd quarter and 3rd quarter of the year. This information will could likely help out advertisers know when people will see their ads and when to pay YouTubers to read out their ads.

**Introduction**

YouTube is a social media platform where users can view videos that other people have posted. It was created in 2005 by Steven Chen, Chad Hurley, and Jawed Karim, who were former employees of PayPal [2]. In 2008, YouTube begun to create advertising formats to help generate money for the site and its channels [3]. Advertisers saw the potential of this website and the audience that it reaches so they decided to put their ads on the platform. These ads can be generated after clicking a video and viewing a 5 second to 1 minute add that is possibly skippable. This data analysis should provide some more insight into what categories of videos, dates, and sentiment, advertisers and content creators should look toward before making decisions on the platform. For the sake of simplicity and size of the data, the dataset came with a little under 1,900 videos to look at. To put in perspective, YouTube currently has around 800 million videos [1] so this dataset is a very small look into the platform itself.

**Computing/Resources**

This project was completed on a single Windows 10 desktop computer. Kaggle was used to find the dataset along with understanding the basic idea of how the dataset worked. To visualize the data and combine the 2 datasets, the Microsoft program called Power BI and Power BI Dashboard were used. These 2 allowed me to visualize and see trends in the data. Dashboard was used to display the data given a valid custom input from the user.

Chart, sunburst chart

Description automatically generated

Figure 1. A representation of the data displayed in Power BI

**Methodology**

The data was obtained using the previously mentioned program in Power BI. The whole dataset was a combination of 2 datasets, one was video statistics, and one was on comment statistics. These were combined to simplify the project and to see if the videos statistics used the same video id as the comment statistics.

With the datasets successfully linked together, I started at looking at the video statistics first. This dataset is categorized by the number of comments on the video, the number of likes the video has, and the keyword that the video is categorized by. It has a separate way to look at the identification of certain videos. It does this by listing the title of the video, the video id, and the number of views attached to the video. I wanted to first look at what keyword was the most liked in the dataset which is represented in figure 1 above. Knowing what gets the likes is important for content creators and advertisers alike. The more likes a video has the more attention it gets. To continue down this trend, I wanted to know if that was the same for views for the keyword. Would Mr. Beast have the most views just like it did for likes?

A picture containing graphical user interface

Description automatically generated

Figure 2. A look at what is causing some staggers in the data

Chart, sunburst chart

Description automatically generated

Figure 3. A more visually appealing look at the amount of views on the keywords

This is where the dataset seems to confuse some keywords. This is more criticism for the original author of the dataset and how they categorized videos. In figure 2, it shows a music/meme video being a part of the google keyword which is not quite right. I feel like this should be in the music keyword rather than in google. Another questionable part is the animal keyword. This keyword has a double meaning in some cases. In figure 4, you can see that there are music videos in here with the name ‘animal’ somewhere in the title, whether it has to do with the song name or the artists. This has caused an unintentional stagger in the views when sorting by keyword. The animal category is supposed to be just actual animals like cat videos or nature videos, Music videos tend to have more views than likes because people will generally listen to the song more than once so the song will get a more staggering number of views than likes. Even after seeing these errors, it still is clear that the Mr. Beast keyword is still near or if not the top. This content creator has a massive following, so this automatically causes some disruption when looking at data. Just because you might have a styled video like Mr. Beast doesn't mean you will get the number of views he has.

Graphical user interface, application

Description automatically generated

Figure 4. Shows that 7 out of the 10 videos on this list are music videos just from the titles

When is comes to YouTube it is important to look at dates and when content creators are uploading videos. For this, I wanted to look at what time people were uploading at and how has this impacted the viewership they were getting. To start, I predicted that the beginning of any given year would have the least amount of views because there are less adds in January and February. Meaning that YouTubers will post less and get less views. When running the data, it turns out that I was almost right, the beginning quarter was the second lowest, only higher than the 4th quarter seen in figure 5. The highest quarter was the 2nd quarter. This could be due to a lot of things. Did someone big upload during that time, were a lot of creators posting during that quarter? Did Covid impact the amount of views during that quarter? While I can’t give a complete answer, the real answer is probably a mix of these hypotheses.

Chart, bar chart

Description automatically generated

Figure 5. A look at the amount of views in each quarter of the year in billions

Views on YouTube fluctuate all the time. Because of that, it is hard to tell if the platform is growing based on views alone. The dataset also seems to be missing some videos that were made in the years that are surprisingly low in figure 6. If I was pitching this to others, I would have to start showing results from 2018 onward because it properly shows some semi-correct data. The one guess to why 2018 is such an abnormality was because of that year’s YouTube rewind being as bad as it was which got it a lot of views. I don’t know what attributed to 2018 have such a high view count but besides that outlier, you can see that views steadily increase from 2019 onward. 2022 is cut off a bit because the dataset was probably cut off around the middle of the year, so it has less than what is predicted. The pandemic also played a role in this because more people were at home, off from work or school, watching videos.

Chart, bar chart

Description automatically generated

Figure 6. Data shows the amount of views per year with the given data

The next section is about comments and what are some of the highs and lows of them. Comments are a good way of viewers to express their opinions on the video and give feedback to the video creator. Usually, a good amount of comments on a video mean that the amount of views was also high. First, I wanted to look at what time of the year where users most likely to comment on a video. My predictions where that during the summer months where when videos would get the most comments because students would be off school, and that people would be off work to be at home looking at videos. When plugging in the data, it turned out that my prediction was correct and that there were more comments in August and the least amount in February, figure 7. August had nearly 1 million more comments than the 2nd most comment month which was April. The least commented months in the set were in October and February with a combined total of less than 1 million comments.

Chart, bar chart

Description automatically generated

Figure 7. The comments sorted by the months the videos where published in.

Another interesting feature is to see the amount of comments over the years. As mentioned before, comments are good way to see if the community on YouTube is alive. In figure 8 you can see a sharp increase in the amount of comments per year. This means that YouTube is still gaining popularity and more people are interacting with the videos.

Chart, histogram

Description automatically generated

Figure 8. This data shows the how the platform has had an increase in comments over the years.

Comments also show the impact a certain video had on the platform. Whether this video was just a meme video or that it made a big impression could determine if it has a lot of comments. My expectations were that in this list of top commented videos, I expected to see some of the most well-known videos on the platform there. Figure 9 shows that I was correct. From personal experience, these videos are either videos that everyone knows or are memorable music videos. While these are kind of hard to predict if they get viewed are, most of these non-music videos got famous by pure luck most of the time so if you are a content creator looking to see what made these videos good, its just luck and passion.

Chart, histogram

Description automatically generated

Figure 9. A bar graph showing what the most commented videos are on the set.

**Conclusion**

YouTube is a growing platform for anyone to publish videos, watch videos, and advertise products. While this dataset has its flaws with how it grouped certain videos and had a lack of videos for certain years, it provided good insight into the growth of the platform along with the “trendy” categories. Showing the most popular keywords like ‘Mr. Beast’ and showing what are the most active months are (see figures 5 and 7).

Some improvements to this study are to increase the amount of videos present in the dataset with some more variety in the dates that the videos were published. With YouTube being as large as it is, more data should be taken to accurately show the growth and popularity of the platform and its videos. Another fix is to make sure that each video is correctly categorized by its keyword. This is a reference to the animal category having music videos in them along with the google keyword having a music video as well.

With the information provided with the dataset, I expect this to be important information to people who are unfamiliar with the platform like advertisers that want their product be known. YouTube is a great way to get your product known either by having advertisements being played before the video starts or sponsoring YouTubers that have a lot of views or dominate a certain keyword. Whether your product is liked or not, if you play your advertisement enough or sponsor YouTubers frequently, people will know about it.

**References**

1. Aslam, Salman., March 14, 2022. *YouTube by the Numbers: Stats, Demographics & Fun Facts* <https://www.omnicoreagency.com/youtube-statistics/>
2. Hosch, W. L.. "YouTube." Encyclopedia Britannica, April 28, 2022. <https://www.britannica.com/topic/YouTube>
3. Google sites page for YouTube’s 5th year running <https://sites.google.com/a/pressatgoogle.com/youtube5year/home/history-of-monetization-at-youtube>
4. Kaggle dataset - <https://www.kaggle.com/datasets/advaypatil/youtube-statistics?select=comments.csv>