Youtube Trends

Nicholas Crnic, Durgesh Vaishnav, Billy Cheeley, Justin Covert

October 20, 2018

Introduction

The question to be answered with data was, ***What is the user behavior based on the most viewed, liked, disliked and commented YouTube videos during the time frame of November 2017 to June 2018?***

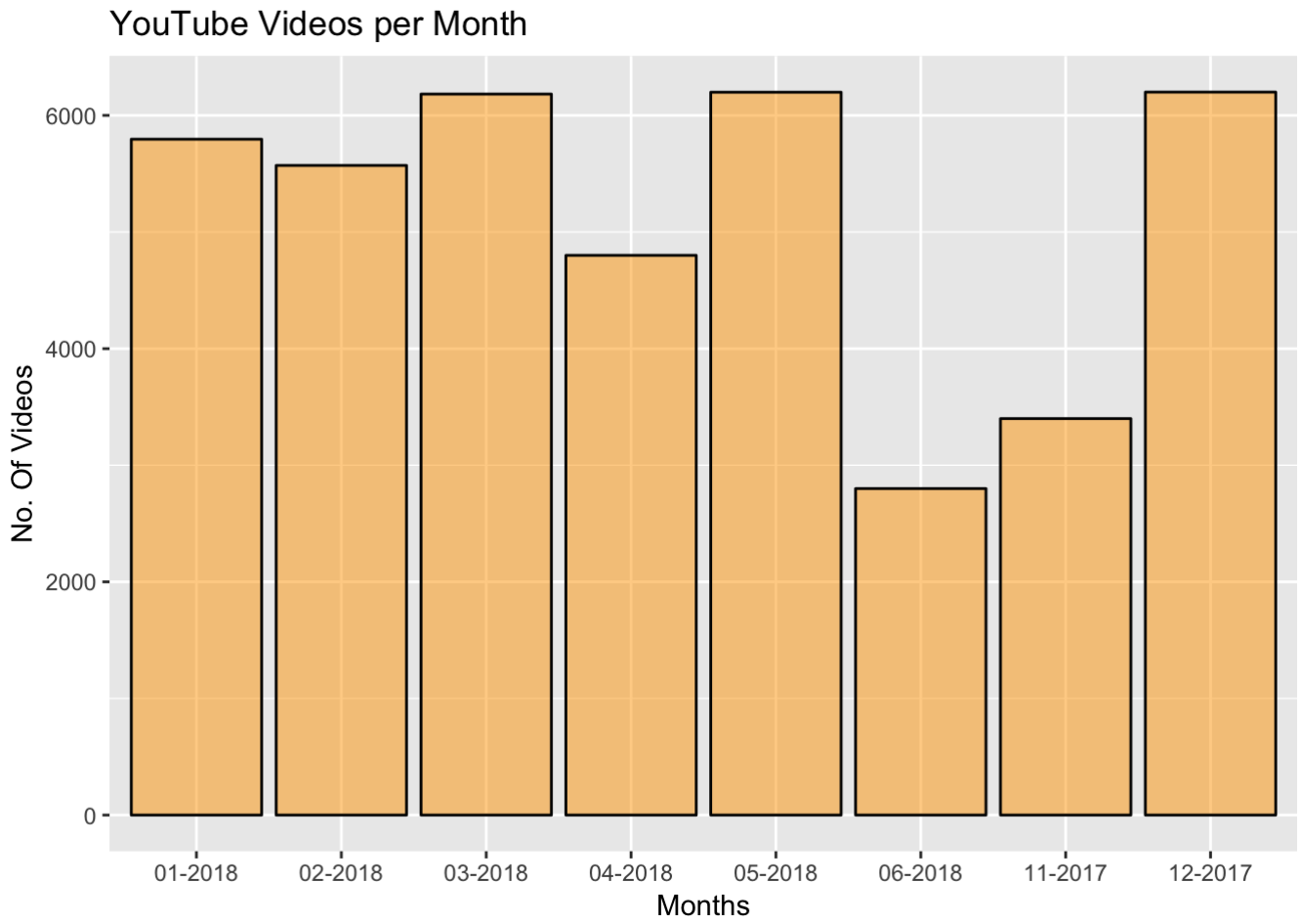
We used 50,000 lines of data from kaggle which can be found at https://www.kaggle.com/datasnaek/youtube-new#USvideos.csv (https://www.kaggle.com/datasnaek/youtube-new#USvideos.csv) to determine the answer to this question. The data provided us with many variables including but not limited to number of views, likes, dislikes, video Ids and categories.

As we set out to determine the answer to the question a variety of questions arrose.
1. What variable or variables should be used to determine the most popular video?
2. How do you determine that?
3. What variables make up a great video etc.?

Ultimately, the answer to the previous three questions is “in the eye of the beholder.” In the beginning, we thought the number of likes could quickly and easily determine the best video but we were incorrect.

Analysis # 1 (YouTube Videos uploaded per month)

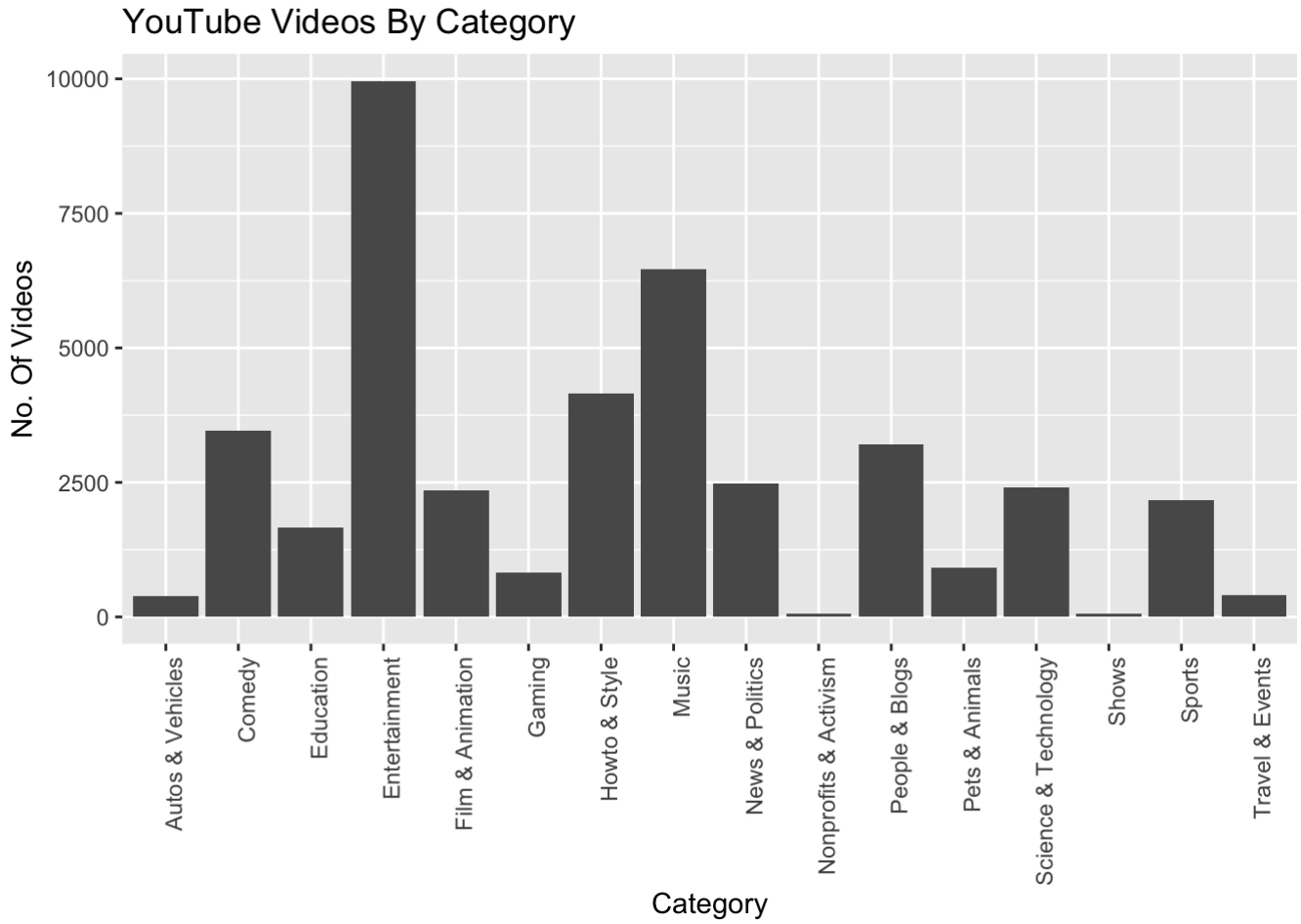
Take a look at point graph for YouTube Videos uploaded per month



This shows that in more than 6000 videos uploaded on the month of December, March and May and less than 3000 videos uploaded in months of June.

Analysis # 2 (YouTube videos by category)

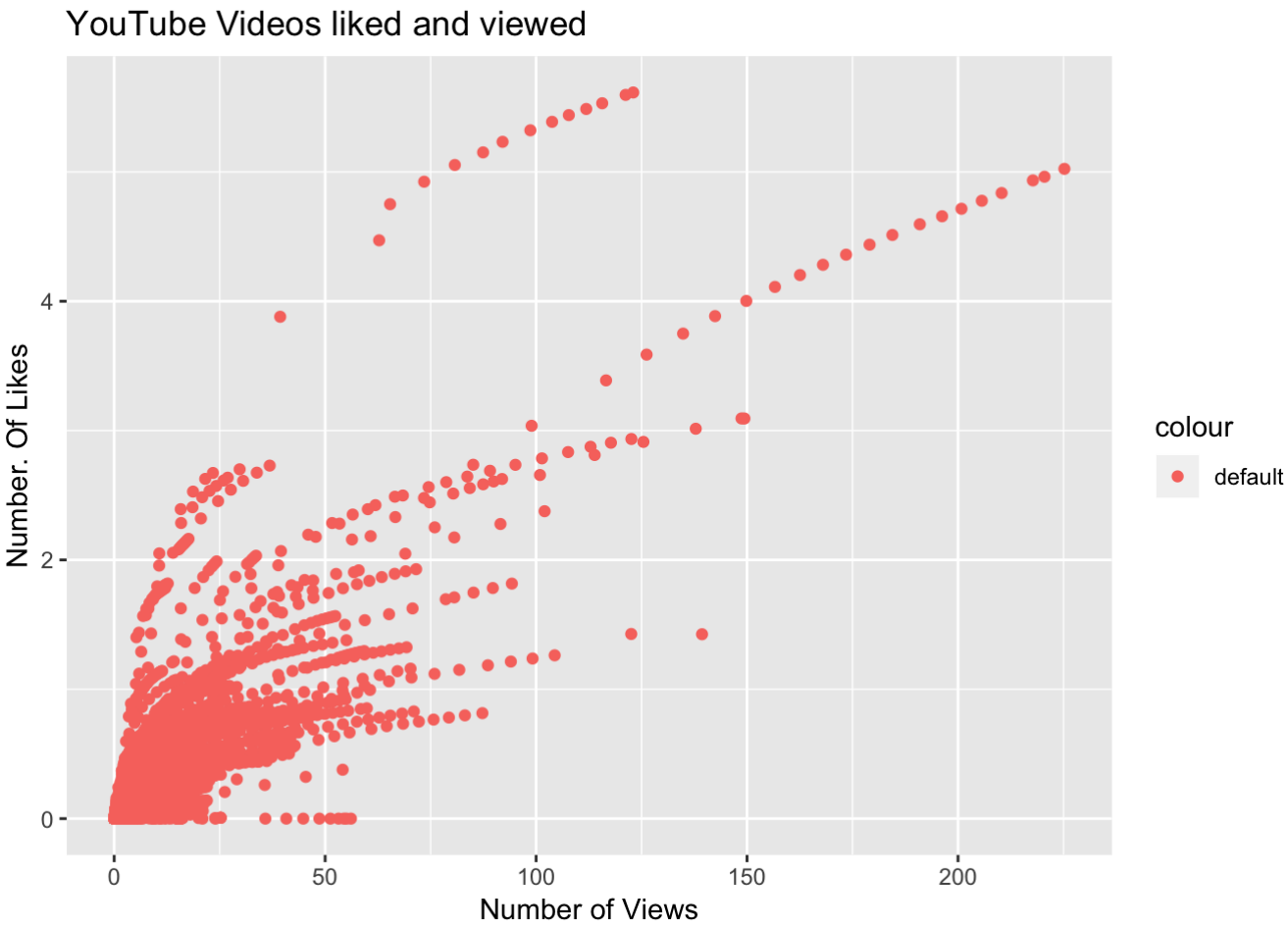
Our raw data came with categoryIds and we found another json data file *US_category.id.json* to map categoryId to category titles. below graph is generated from these two different data sources.

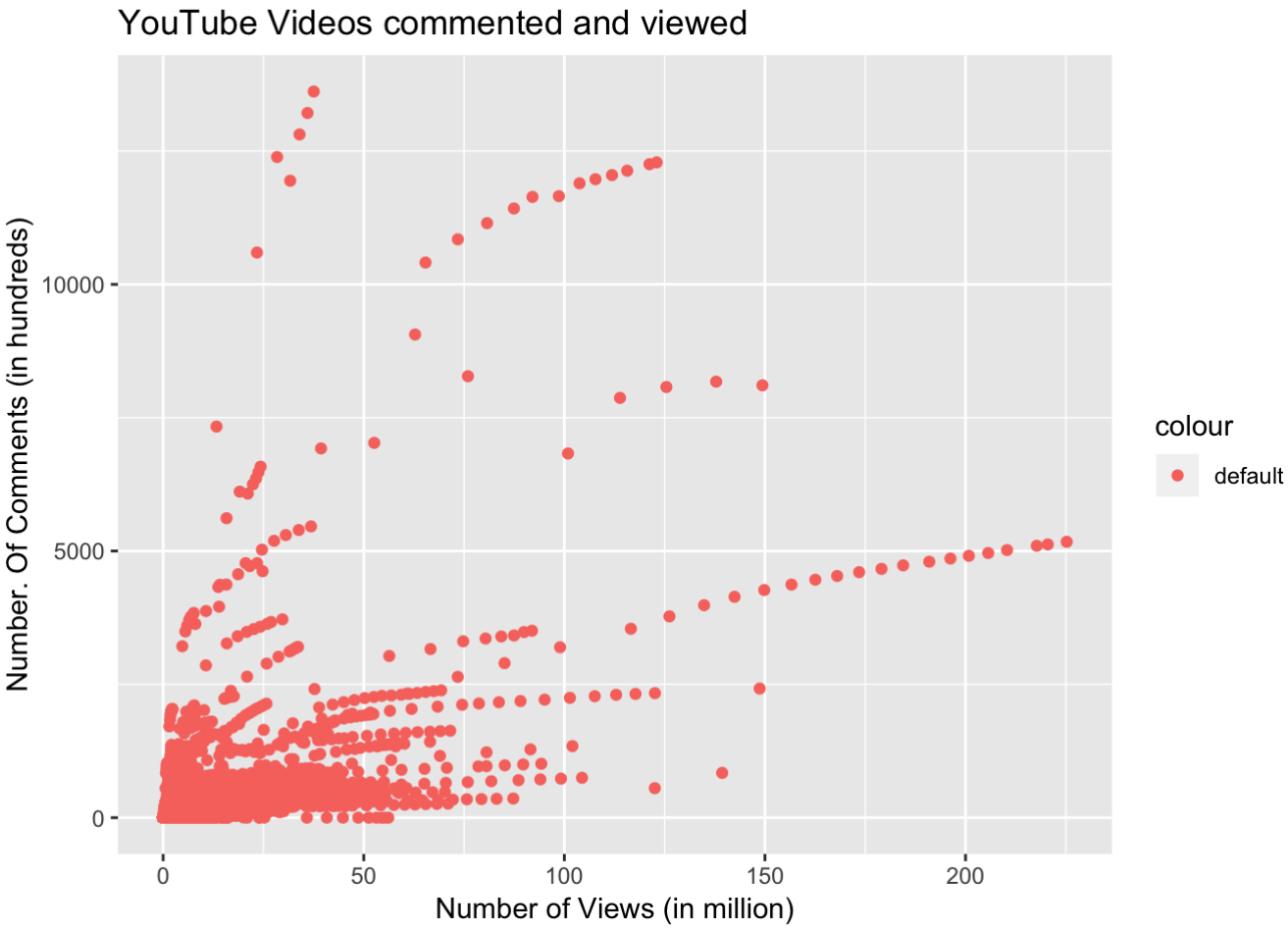


This analysis shows that “Entertainment” is the popular category where about 10,000 videos uploaded and a few videos uploaded to “Shows” and “Nonprofit and Activism” category.

Analysis # 3 (YouTube Videos viewed as well as liked and commented)

Take a look at point graph for YouTube Videos viewed as well as liked (in millions).


YouTube Videos liked and viewed
Number. Of Likes
0
2
4
0
50
100
150
200
Number of Views
colour
default

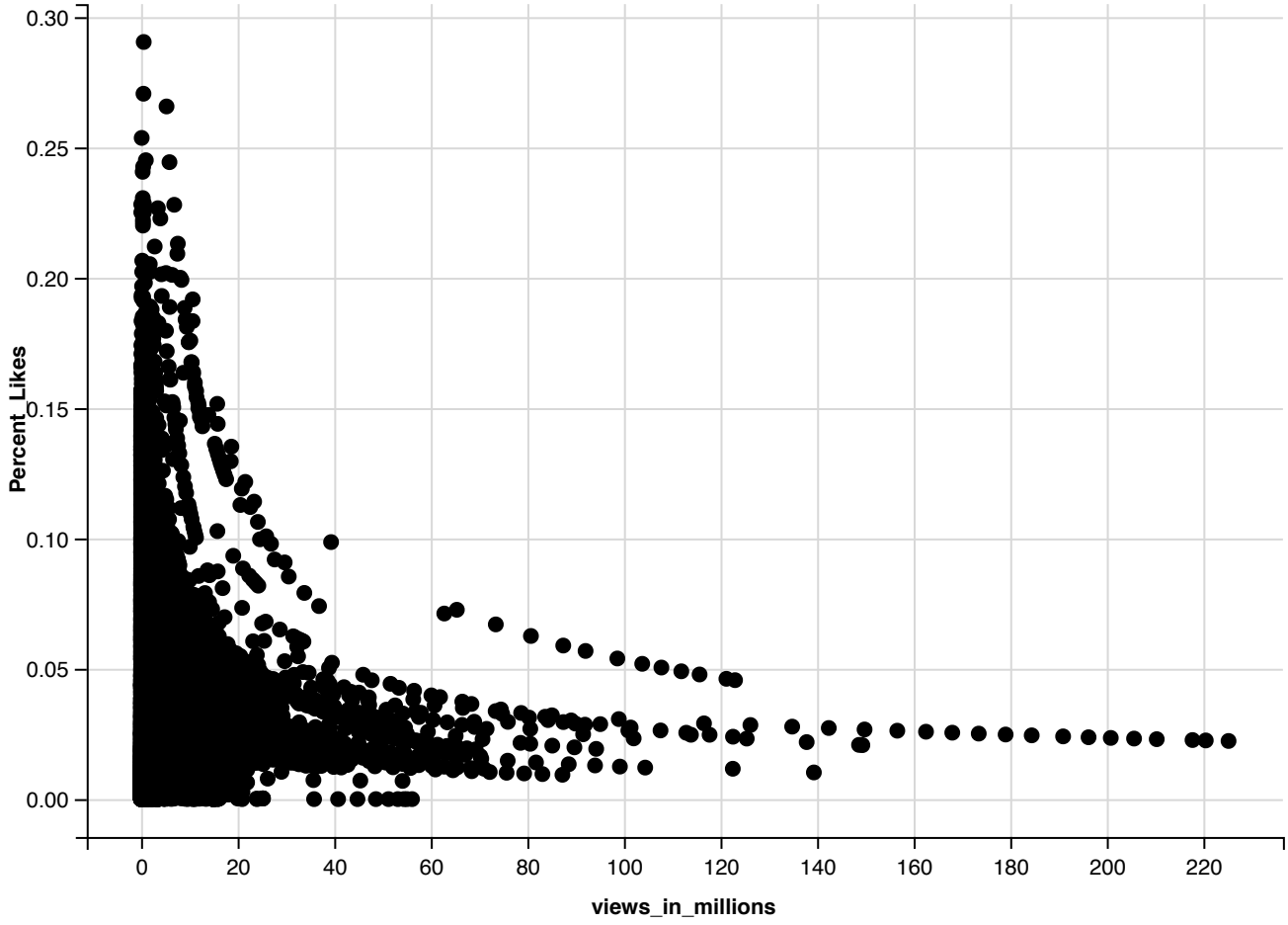

YouTube Videos commented and viewed
Number. Of Comments (in hundreds)
0
5000
10000
0
50
100
150
200
Number of Views (in million)
colour
default

The Question continued…

The graph above exhibit an upward trend in likes that are at a direct result of number of views. What this means is that based on this data the number of likes is dependent upon the number of views a video received. We didn't believe that this was an accurate depiction of the best video during the time frame because the number of views is so dependent upon marketing efforts. Whether the efforts be guerilla marketing tactics or paid for marketing tactics, so we set out to discover a different method for determining the popular video.

Analysis # 4 (YouTube Videos % likes)

The first method we came up with taking the percentage of likes in comparison to the number views. Initially, we thought that this would give us a more accurate and clearer depiction of the popular video during the sample pool time frame. So we used the dplyr package to add another variable to our data set and we called this variable Percent_Likes. As you can see in the below graph the results were virtually the same.

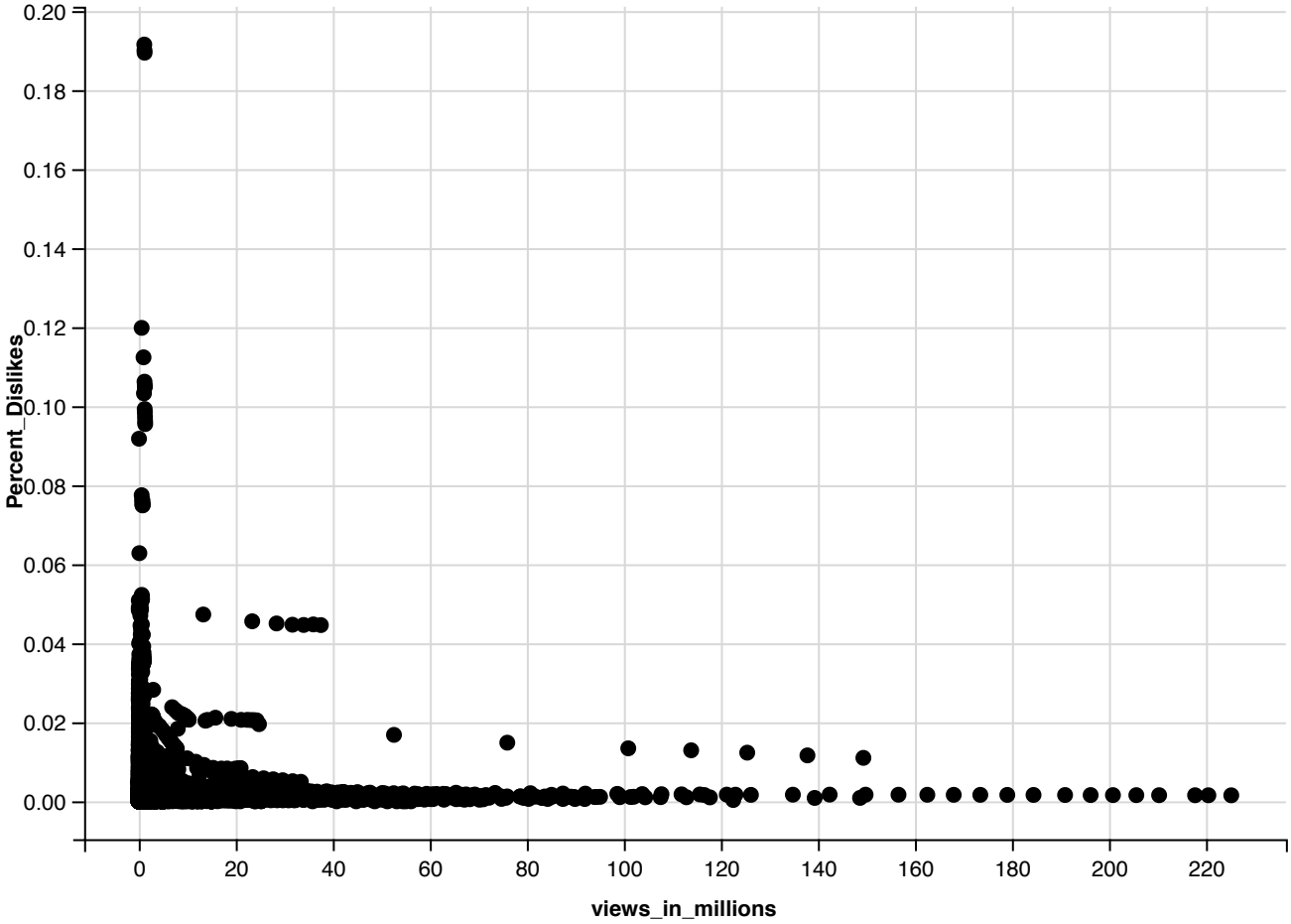


analysis #4 continued…

As you can see from the above graph the trend was downward and told us virtually the same thing as the first to graphs. The percentage of likes per view when down as the number of views went up. What this means again is that the number of likes of a youtube video during this time frame is directly dependent on the number of views a video obtains and the views are directly dependent upon marketing dollars and marketing effort. As a result we determined that this metric alone could not clearly define the popular youtube video. We needed more metrics and increase in-depth analysis.

Analysis # 5 (YouTube Videos % dislikes)

Almost instantly we thought what about taking the percentage of dislikes in comparison to the number of views a video received? So we created another variable and we called this variable Percent_Dislikes:



As you can see from the graph above the percentage of dislikes in comparison to views displayed the exact same trend as the percentage of likes in comparison to views. The percentage of dislikes was completely dependent upon the number of views once again!

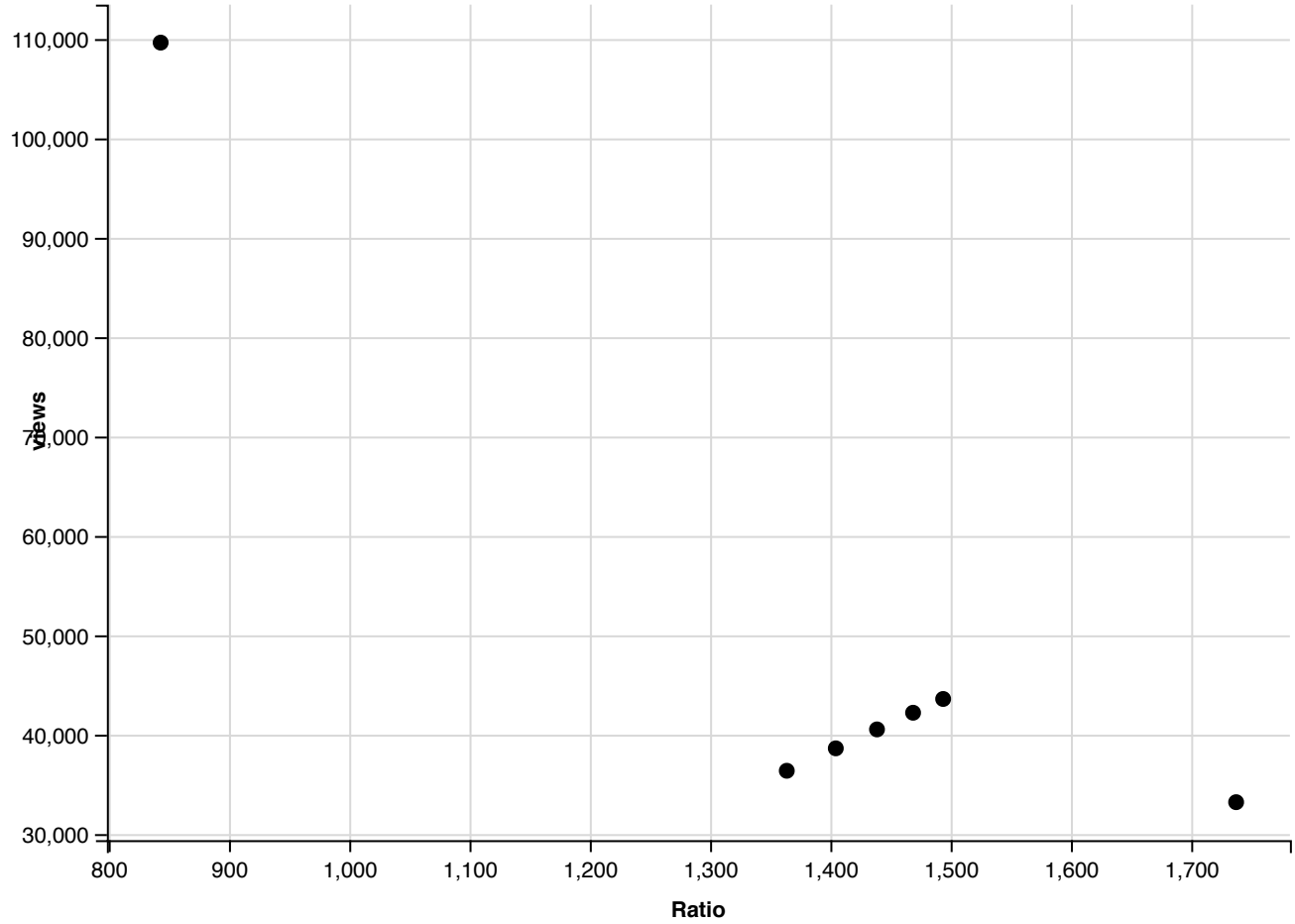
Analysis # 5 continued…

Since the results of the previous analysis have determined the same result further analysis was needed to determine the answer to our main question. After brainstorming we came up with another metric. The metric was the percentage of likes to views divided by the percentage of dislikes to views. As a result, we created another variable. We called this variable Ratio. When trying graph the Variable Ratio many problems arrose being that the ratios were small in comparison to our sample size. So we had to narrow our focus in order to achive diagnosable results.

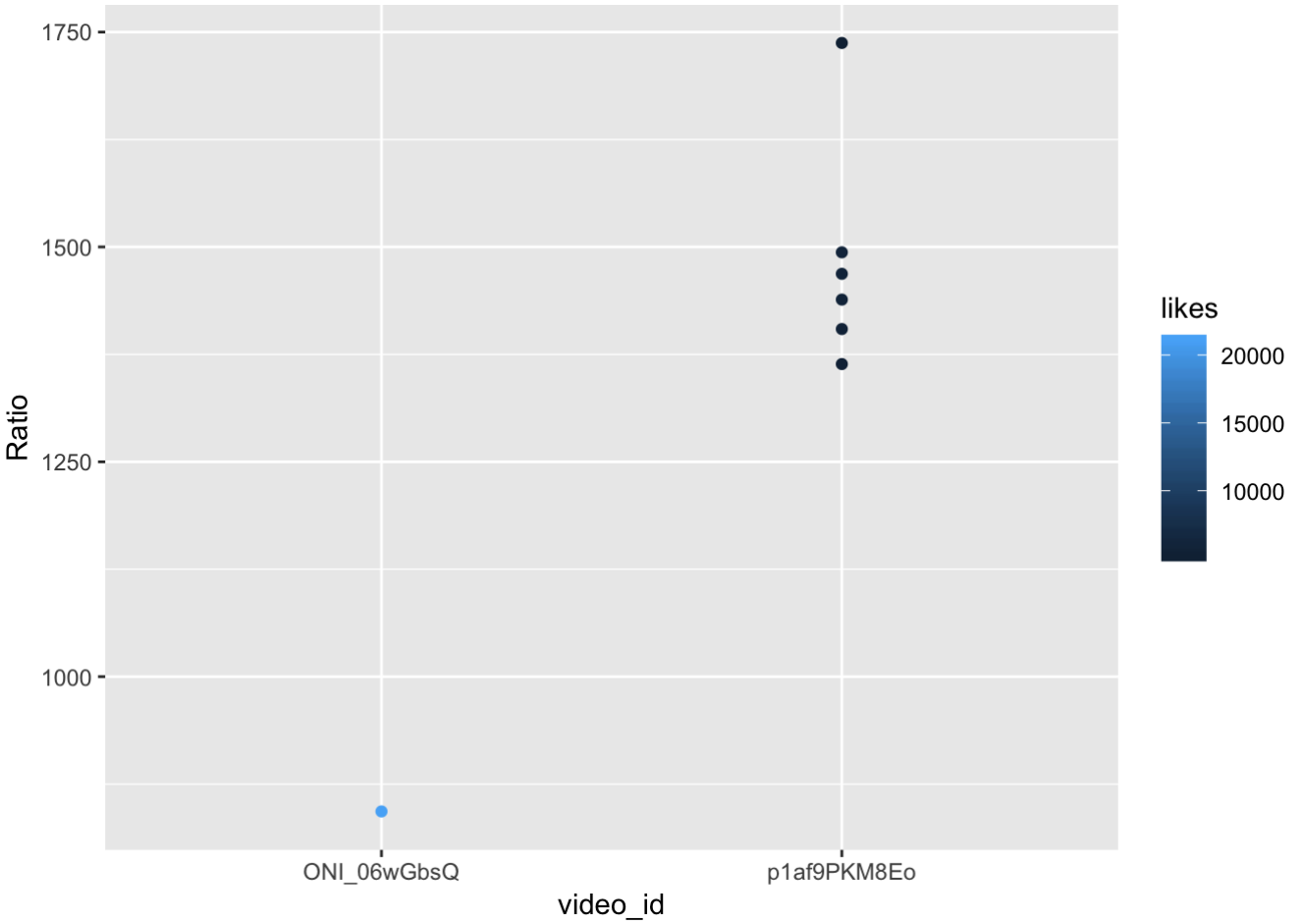
Analysis # 6 (Narrowing the Focus/Minimizing the Sample Size)

In order to minimize the dataset we thought it was important to disregard the number of views. The previous graphed trends showed us that the number of likes and dislikes on any particular video was dictated and directly impacted by the number of views. Therefore the number of views was proven irrelevant in determining the popular youtube video and the ratio of dislikes compared to likes would be a more accurate depiction of the best video. So once again we had to do some more data cleaning! So we needed to extract the top 50 videos with the highest Ratio.

Our data analysis shows that Maximum Ratio is **1737.3333333** and mean Ratio **43.968451** and generate below graph above mean ratio.



After analyzing the graph it was easy to see that there were 6 videos with a ratio greater than 1000. These six videos all had the percentage of likes per view divided by the dislikes per views with a ratio of 1000 or greater. Which means the percentage like per views divided the percentage of dislikes per views was greater than 1000 times. At last, we were able to find the top videos!



Based on the graph above the video with the highest ratio is p1af9PKM8Eo but ONI_06wGbsQ has a high ratio as well and quite a few more views. Based on ratio alone, which was the metric we chose to determine the best video p1af9PKM8Eo was the winner. Let's take a look at the second place video first.https://www.youtube.com/watch?v=ONI_06wGbsQ (https://www.youtube.com/watch?v=ONI_06wGbsQ). As you can see this video had 36,000 likes and 66 dislikes would make the Ratio very high and prove that our analysis was accurate. The first place video can be found here https://www.youtube.com/watch?v=p1af9PKM8Eo (https://www.youtube.com/watch?v=p1af9PKM8Eo). This video had 9,500 likes and 14 dislikes which once again would create a high Ratio and prove are analytics accurate.

The Business Behind the Data

You are probably wondering why it matters what video on Youtube had the greatest Ratio of likes to dislikes in comparison to views. One reason is that if you didn't notice neither of these videos had adds placed on them. With little marketing and at a low cost these videos could increase their views dramatically and quickly being that the majority of the people that watch them like them. Which would make youtube more revenuw at a lower cost. The Ration could also be used as a market research tool for music/music videos, movies/movie trailers, web shows and TV shows. A company could release a video prior to releasing the single, movie or actual show to test it's Ratio in-order to give the company an idea of how well the content will perform. This could be offered as another service by Youtube. Finally, it was extremely apparent that the number of dislikes and likes was extremely dependent on the number of views a video received and the percentage of people choosing one or the other was very small. Which tells us that there is definitely room for improvement when it comes to getting people to respond to a video at all. Many things could be done to increase the percentage of likes or dislikes per view so a more accurate picture can be displayed on how good the video is or not.