ORIE 3120 Final Project: Data Visualization

Group 29: Selena Kang, Andrew Xiao, Maria Silaban, Toshi Tokuyama

Data

For our project, we decided to use the YouTube dataset which consists of various data on 200 trending videos across 205 days for 10 different countries. In this project, we focused on the data from the US videos (USvideos.csv) to create visualizations.

The dataset consists of the following columns:

- video_id - String, Unique value
- trending_date - Date in the format of *year.date.month*, date the video was trending
- title - String, Title of the video
- channel_title - String, Channel of the video posted
- category_id - Integer, Category of the video
- publish_time - Date time format in an ISO-8601 date representation, Date and time the video was published
- tags - String, tags separated by "|", Tags attached to the video
- views - Integer, Number of times the video was viewed
- likes - Integer, Number of likes on the video
- dislikes - Integer, Number of dislikes on the video
- comment_count - Integer, Number of comments on the video
- thumbnail_link - URL, Link of the video
- comments_disabled - Binary data, Indicates whether comment for the video was disabled, Possible values: [True, False]
- ratings_disabled - Binary data, Indicates whether rating for the video was disabled, Possible values: [True, False]
- video_error_or_removed - Binary data, Indicates whether the video was removed or had errors, Possible values: [True, False]
- description - String, Description of the video

Questions

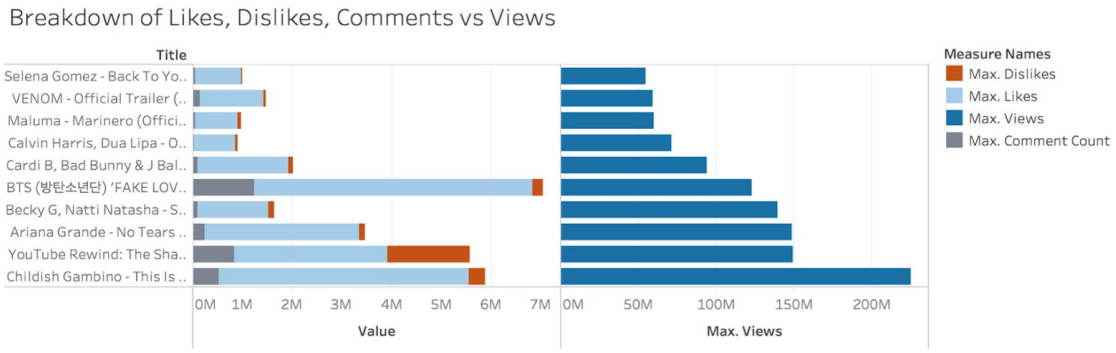
To get more insight into the data, we came up with the following questions to help us analyze the data:

1. How do popular videos compare in terms of the number of likes, dislikes, and comments?
2. Are views affected by the time of year?
3. Do videos with more tags get more comments?
4. Are the videos trending on a particular day a dependent variable of what videos trended on previous days? If so, in what way?

Visualization

We created the following visualizations to get a deep understanding of the questions which can help conduct analysis.

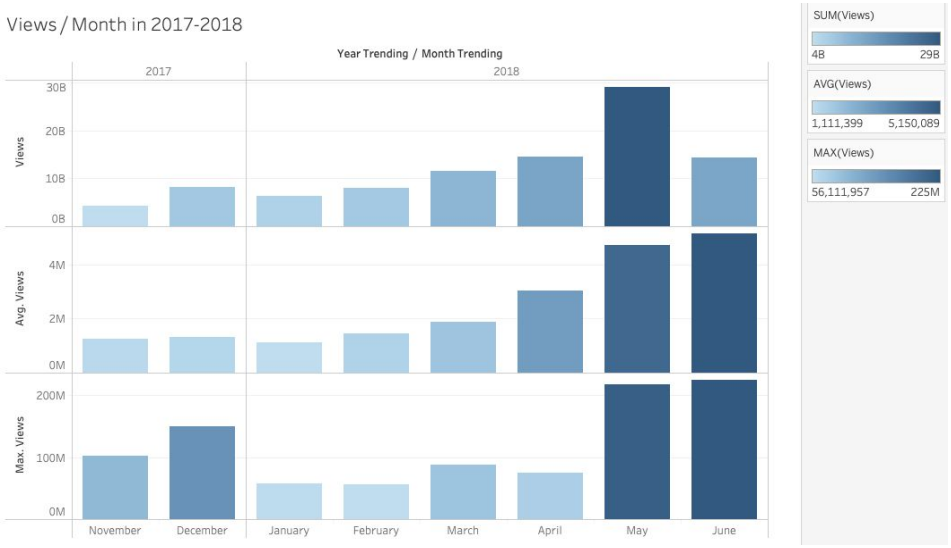
Visualization for Question 1



Max. Dislikes, Max. Likes, Max. Comment Count, Max. Views and Max. Views for each Title. Color shows details about Max. Dislikes, Max. Likes, Max. Comment Count and Max. Views. Details are shown for Max. Dislikes, Max. Likes, Max. Comment Count and Max. Views. The view is filtered on Title, which keeps 10 of 6,455 members.

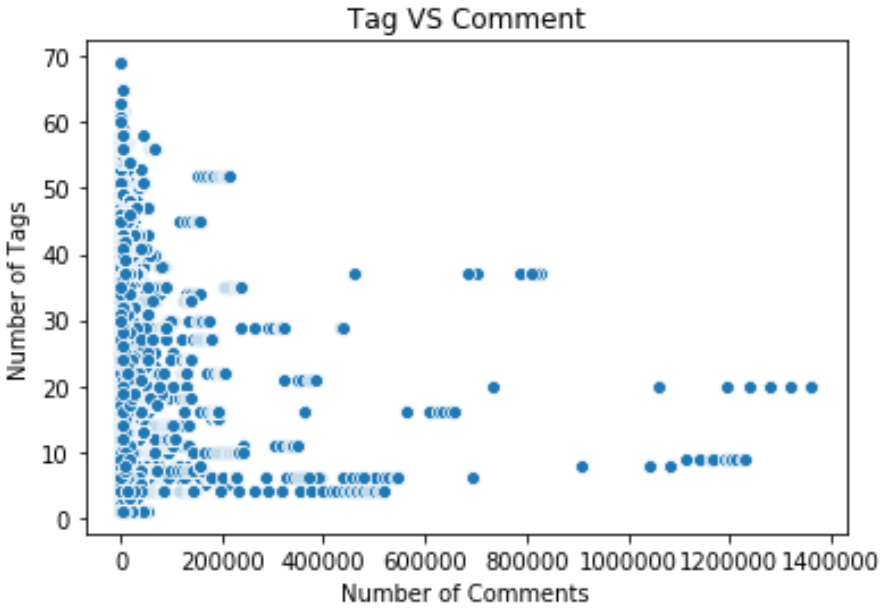
This visualization aimed to break down the number of views for the top 10 most popular videos and see how their likes, dislikes, and comments compare to each other. We decided to choose the top 10 most popular videos as it would give us the best chance at seeing relationships between views and the number of dislikes, likes, and comments. From this visualization, we noticed that all videos have significantly fewer "reactions" (reactions being dislikes, likes, and comments) than they do the number of views. This discrepancy can be clearly seen in the Childish Gambino video as it has over 200 million views but less than 6 million reactions. We also saw that most viewers would prefer to express that they "like" a video rather than "dislike" except for the Youtube Rewind where there was a significant amount of dislikes in comparison to its number of likes. We also noticed that the number of comments on a video tends to be greater if there are a lot of dislikes. The 3 videos that garnered substantial amounts of comments were the 3 videos that also had a visible portion of dislikes. We had hypothesized that videos with the most views would have the most reactions but in this visualization, we can see that the BTS video had fewer views but more reactions than the most viewed video.

Visualization for Question 2



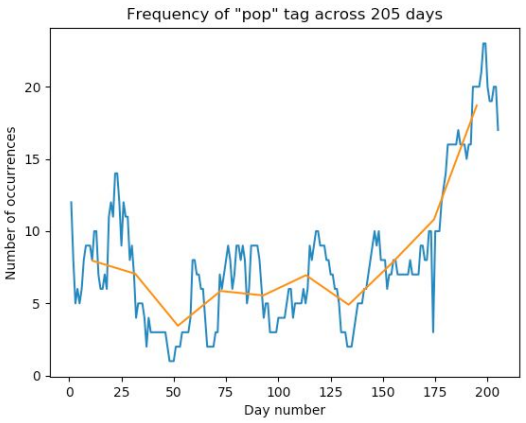
The bar graphs in “Views / Month in 2017-2018” show visualization of the total views, average views, and maximum views per month of top-trending videos in the year 2017-2018. Each gives slightly different interpretations of how views are affected by the time of the year, and the darkness of the bars corresponds to the larger number of views. From the given data and visualization, in all three cases of examining and defining “views”, warm months of May and June are seen to have the highest views (as shown in all three bar graphs). For instance, the sum of the views from top-trending videos that occurred in May 2018 exceeded by more than two times the sum of views for other months. Generally, one can interpret that the views are indeed affected by the time of the year as the graphs show a trend in the number of views - a general increase in the number of views from January to May. However, it would be wise to gain more data from other years to be able to safely make the relationship between views and the time of the year. This will also help in analyzing and forecasting future trends and may be conducive for many content creators or publishers on Youtube.

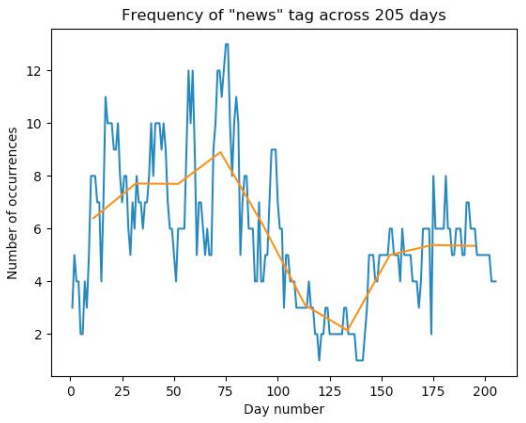
Visualization for Question 3



The scatter plot (*Tag VS Comment)* shows the relationship between the number of comments and the number of tags attached to each video that did not have comments disabled. The scatter plot shows that most points are concentrated on the left side indicating that the number of comments on the majority of videos is less than 200,000. As the number of comments increases, the number of tags attached to the video starts to decrease. This is an interesting trend as normally we would think that if more tags are attached then the video would more likely to appear on the user’s recommended videos. From the graph, we can fit a linear regression in which we transform the regression from linear to exponential. Conducting the regression analysis would let us have a better understanding of the relationship between the number of tags and the number of comments.

Visualization for Question 4





funny	4123
comedy	3573
how to	1652
pop	1627
music	1562
[none]	1535
trailer	1278
food	1255
2018	1251
review	1236
humor	1196
science	1190
news	1184
makeup	1164
celebrity	1081

The fourth question seeks to determine whether the videos that trend on a given day influence which videos will trend on the subsequent day. If such a relationship exists, it would be within expectation to observe the perpetuation of existing trends throughout time. For example, if Ariana Grande's new music video begins to garner views on a given day, it would be expected that the music video's views would increase for some time. This is an example where a *positive* trend was perpetuated across multiple days for the *same video*. It could also be possible for a trend to be perpetuated across multiple days for a *type* of video (e.g. category), a certain *topic* of video (e.g. tags), or even videos made by a particular content creator (e.g. channel). Furthermore, the type of trend that is perpetuated can be quite varied as well, the range of which includes (but is not limited to) positive trends, negative trends, alternating trends etc.

For this visualization, the focus was placed on positive and negative trends for videos of the same topic (i.e. column 'tags'). Displayed on the left is a table of the 15 most commonly observed tags, along with their number of occurrences throughout the 205 day period. Of these 15 tags, 'pop' and 'news' most evidently demonstrated perpetuation of existing trends over time. As shown by the two line plots above, the number of 'pop' related videos trending per day increased, and the number of 'news' related videos trending per day decreased.

The rationale behind selecting such a question is that, especially in the online environment, today's data can depend on yesterday's data and decisions. The environment is changing in unpredictable ways and the data may not always be coming from a distribution.