# Team Project Exploratory Visualization

**by Group 8**

*YouTube Trending Video Analysis*

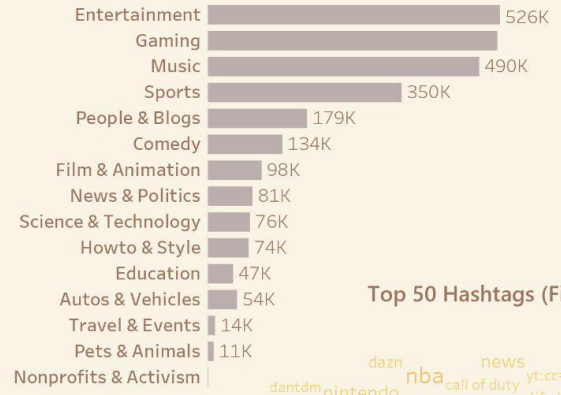
## 1. Hypothesis 1: The more hashtags a video uses, the more views it receives.

- **Motivation:** This analysis can help people understand the correlation between hashtags and other relevant attributes. It will benefit YouTubers or analysts who wonder how many hashtags a video should have and how effective they are. In this exercise, we assume more hashtags can help the videos to reach more customers and potentially increase the “views” and “likes”.
- **Main attributes from the dataset for this hypothesis**
  - **tags -** String
  - **count of tags (for each video) -** Integer
  - **view_count -** Integer
  - **likes -** Integer
  - **Channel Category -** String
- **Exploration process with raw data**
  - Get familiar with all the data attributes and their content.
  - Learn the categories for each attribute, for instance, “Channel Category”.
  - Generate a few “quick numbers” to have a big picture of the dataset, learning from a high-level view.
- **First thoughts and actions**
  - In order to have a better understanding of the data, we generated three quick numbers. From there, we know there are 161,449 trending videos in the dataset with 15 unique channel categories. And these videos were published on YouTube from the year 2020 to 2022.
  - From the second horizontal bar chart named "Video Count for Each Category", We can conclude that "Entertainment", "Gaming" and "Music" are the top three categories by video counts.
  - Some keywords stand out from the word cloud chart in the middle of the dashboard page. More YouTubers tagged their videos with "minecraft" and "fortnite", which also aligned with the bar chart above. It shows that "Entertainment" has more trending videos than the others, which answers why gaming hashtags are more in use.
- **Refine process**
  - Analyzing "Video Count for Each Category" shows that the relationship between hashtags, views, and likes do not match our assumption. The "Music" category has fewer hashtags but generates more "views" and "likes" compared to the "Gaming" category. Similar to "Education", it doesn't seem the number of hashtags matters much in this case.
  - Refine process 1, according to "Correlation Between Hashtags vs Views", we can clearly tell that many videos with zero hashtags in the

Entertainment, Gaming, and Music categories generated many views. Furthermore, the correlation overall between the count of hashtags and the count of views is minimal.
    - Refine process 2, according to "Video Count for Each Category (filtered by top 30 views)", the visual was sorted by the number of hashtags. Videos with more hashtags are listed on the top, and "views" and "likes" are stated on the right-hand side. As the number of hashtags decreases, the "views" and "like" aren't following the same pattern but are distributed randomly. Also, videos marked with "#Shorts" perform very well with zero hashtags.
- **Final conclusions:**
    - Using hashtags properly can be helpful, but only in some cases. Study shows as the number of hashtags increase, the fewer views it receives.
    - YouTube "#Shorts" trending videos with zero hashtags labeled resulted in sizeable "Views" and "Likes".

| Total Video Counts | Total Channel Categories | Video Published Years |
|---|---|---|
| **161,449** | **15** | **2020 2021 2022** |

### Video Count for Each Category




# YouTube

## *Trending Video Dashboard*

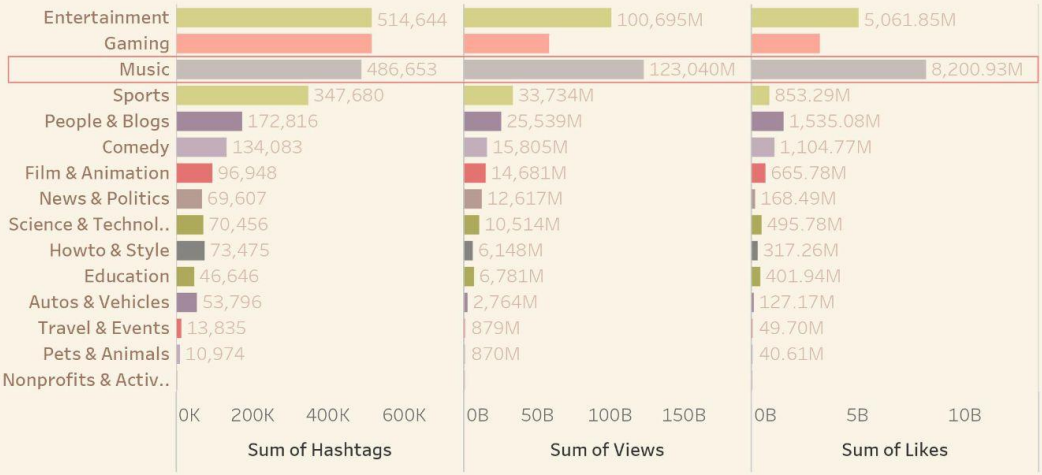
*- Hashtag Analysis*


Group 8

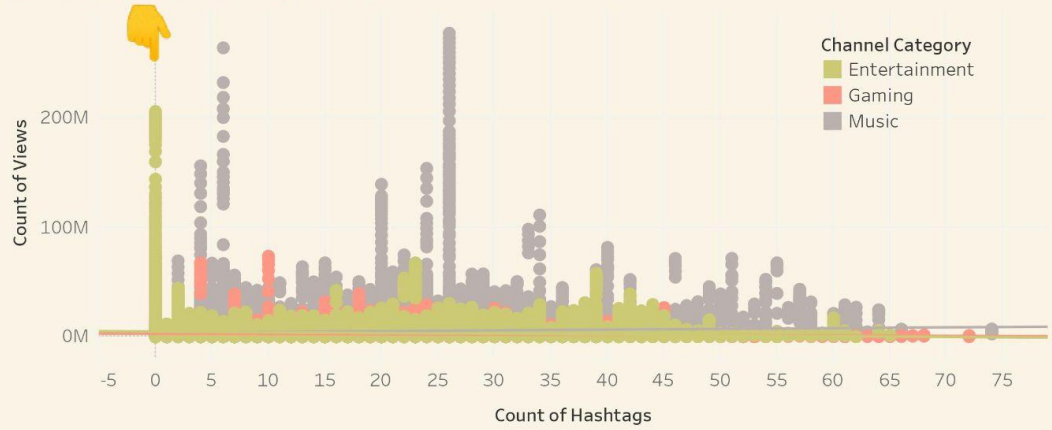
Tamalika Basu
Yilei Ge
Ching-Wen(Jenny) Huang
Sungho Lee (He/Him/His)
Shulang (Simon) Ning


### Top 50 Hashtags (First Choice)

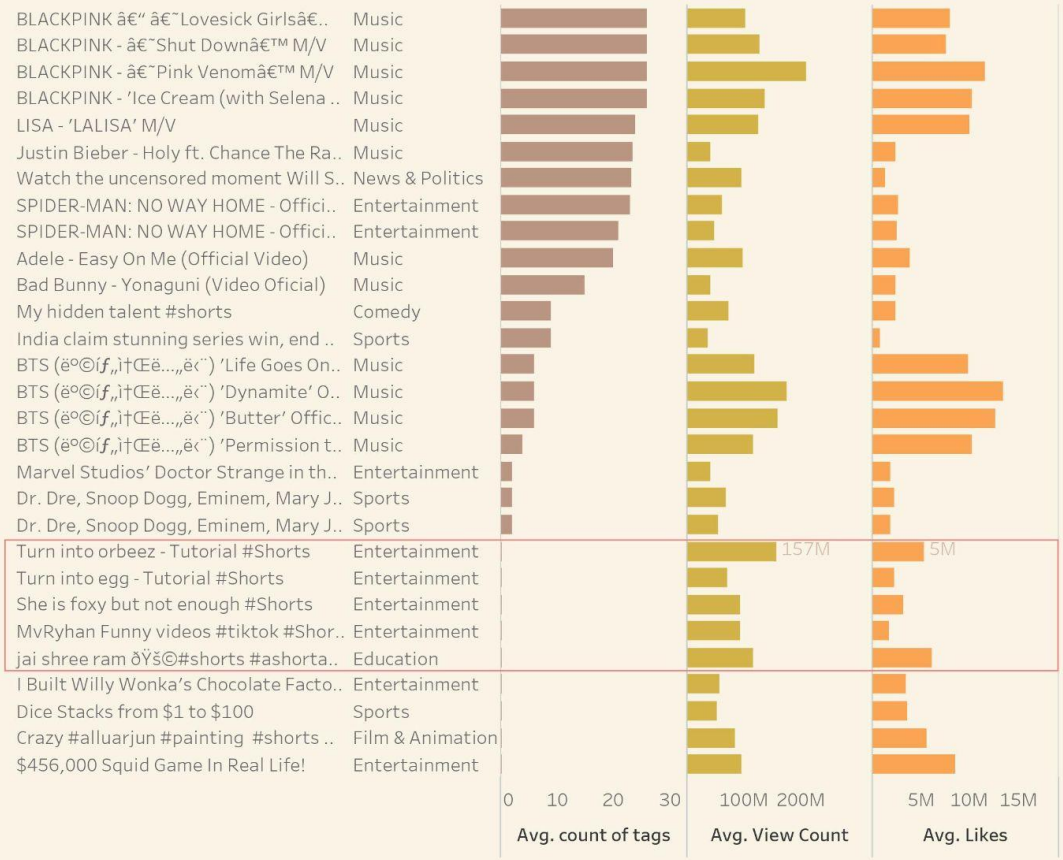

### Video Count for Each Category



## Correlation Between Hashtags vs Views

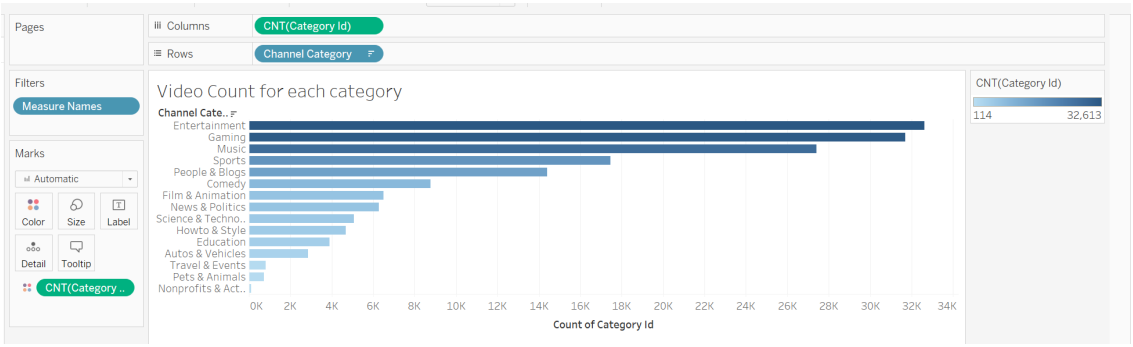



## Video Count for Each Category



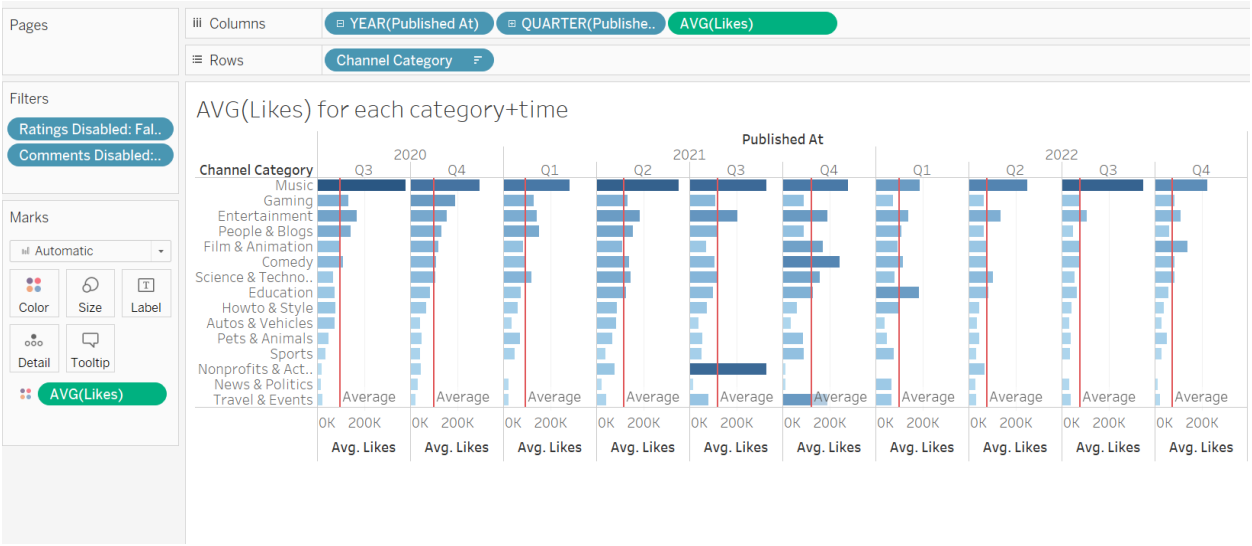
## 2. Hypothesis 2: ‘Entertainment’ category is the most popular among viewers.

- **Motivation:** The motivation behind selecting this hypothesis comes from a layman's understanding of the video viewing pattern. The entertainment category is something that people across demographics relate to and watch. This category does not require any prior knowledge, skill, or understanding and mostly complements the leisure mood for any viewer, thus naturally indicating to be the most popular
- **Attributes from the dataset for this hypothesis:**
  - **channel category -** string
  - **view_count -** integer
  - **likes -** integer
  - **comments -** integer
- **Exploration process & Visualization:**
  - To assess the hypothesis, we first checked the count of the number of entries under each channel type. This will help us understand whether the data is biased toward a particular channel. As we see in the observation below, entries for the entertainment channel are the highest, thus we determine to normalize the data in some way so that this bias-ness does not impact our analysis

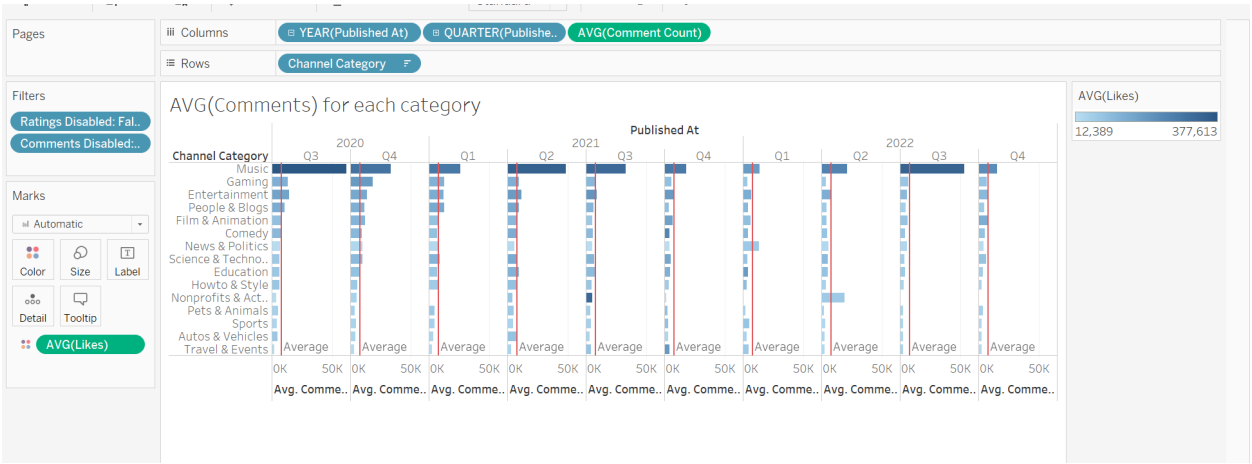



  - Next, to assess the popularity of the channels, we examine 3 KPIs: likes, comments, and view counts. Because the dataset has a large count of certain channels, we consider taking the ‘average’ across metrics to normalize the effect of counts across channel types.
  - Thus, we look at the effect of average likes per category across the years 2020, 2021, and 2022. For 2020, we observe data present for only 2 quarters therefore analyzing our assumption for 2020 will not be fair, and similarly, for 2022, the data might not be complete because 2022 is the running year. Therefore, analyzing 2022 will not be fair. Considering only

2021, we see that ‘entertainment’ did not have the highest average number of likes. The ‘Music’ category had the largest average likes.



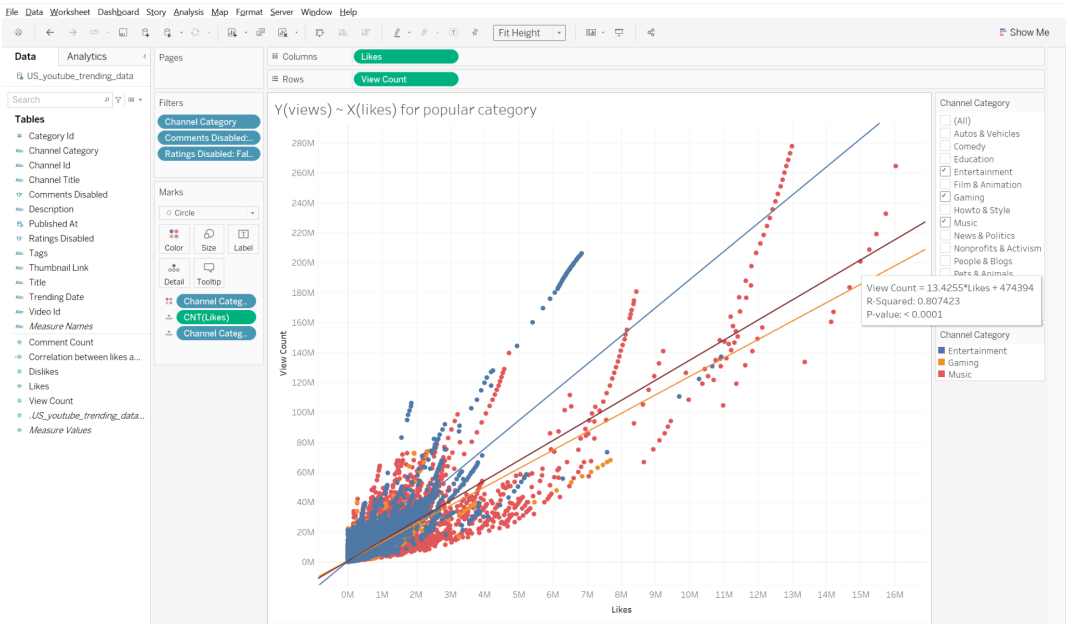

- Similarly, we want to assess the ‘comments’ metric across 2020, 2021, and 2022 to understand the pattern across channels. Adhering to the same understanding of the data pattern over the 3 years, we see an inaccurate representation of 2020 and 2022 in our dataset thus we analyze only 2021. The ‘Music’ category seems to have the highest average number of comments in 2021



- Similarly, we want to assess the ‘comments’ metric across 2020, 2021, and 2022 to understand the pattern across channels. Adhering to the same understanding of the data pattern over the 3 years, we see an inaccurate representation of 2020 and 2022 in our dataset thus we analyze only 2021. The ‘Music’ category seems to have the highest average number of comments in 2021
- Based on the above, we identified 3 popular categories that typically had higher popularity in terms of likes and comments - Entertainment, Gaming, and Music. Further, we want to assess the view count across the top 3 channels and filter out the 2020 and 2022 datasets because of unreliability. For 2021, we observe that ‘gaming’ is the channel with the most average views.


- Further, we observe 2 patterns - 1. The ‘Music’ category has the highest average likes and comments and 2. The ‘Gaming’ category has the highest average views. At this point, we can conclude that our hypothesis about the ‘entertainment’ category being the most popular has been rejected.

- **Final Conclusions:**
    - Therefore, we can conclude that our hypothesis was incorrect to assume that the ‘entertainment’ category is the most popular.
    - Finally, to conclude our analysis, we calculate the $R^2$ (coefficient of determination) to understand which channel type has a better correlation between likes and views to determine the most popular channel in our dataset
    - We observe that with an $R^2$ of **~80%** or R-value of **~89%** (R is the correlation coefficient)**,** the ‘Music’ category has the best positive correlation between its likes and views, thereby indicating it to be the most popular amongst viewers.



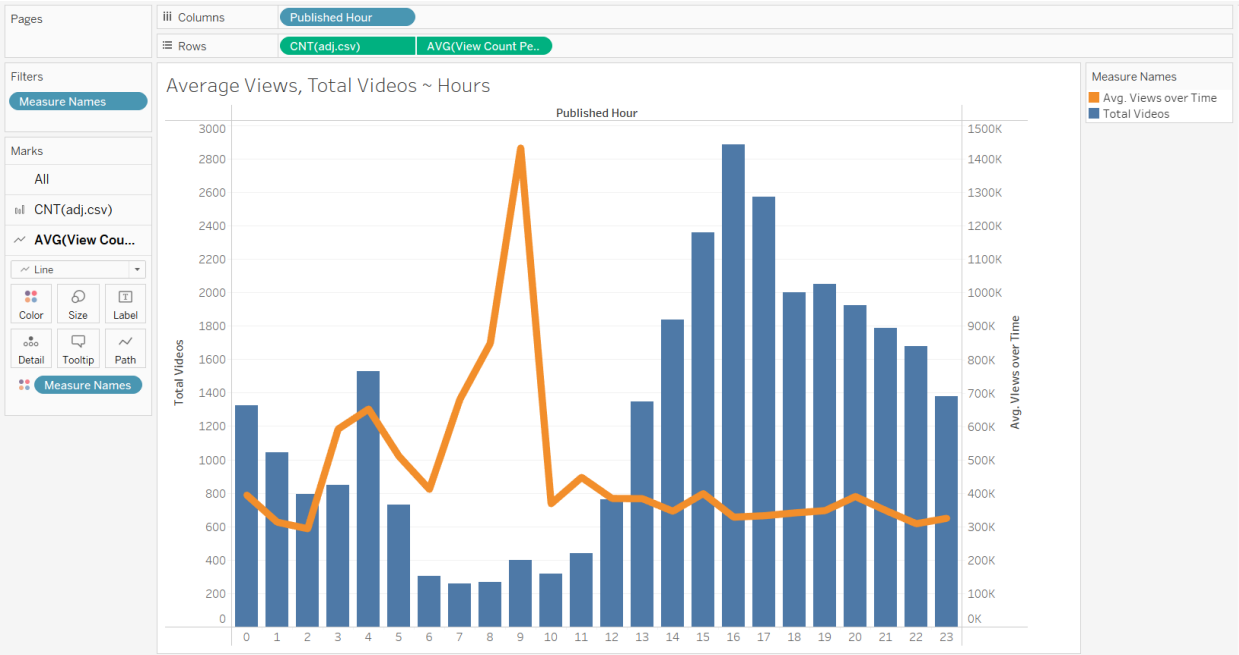
## 3. Hypothesis 3: Youtubers are considering posting time to get the best popularity of their videos.

- **Motivation:** As time goes by, competition for content produced by YouTubers is intensifying day by day. They are doing their best to make the videos they post popular. We usually think that YouTubers are also considering publishing timing for the best popularity. We are going to test whether this is true. If not true, we can present them with the best video upload timing. Also, it comes to another question which is - Does long video exposure time lead to a high number of views, likes, and comments?
- **Attributes from the dataset for this hypothesis:**
  - **publishedAt -** date and time
  - **trending_date -** date and time
  - **comments_disabled -** logical
  - **ratings_disabled –** logical
  - **(derived) views_over_time -** integer
  - **(derived) likes_over_time -** integer
  - **(derived) dislikes_over_time -** integer
  - **(derived) comments_over_time –** integer
  - **(derived) exposure_time –** double
  - **(derived) day -** char
- **Exploration process & Visualization:**
  - To assess the hypothesis, we preprocessed the original data to find publishing time for each published video and examined KPIs such as views, likes, dislikes, and comments at midnight every day for three to five days.
  - The problem is that the number of days counted for each video is different, making direct comparison difficult. Therefore, we induced new variables to calculate the popularity per unit day as follows:

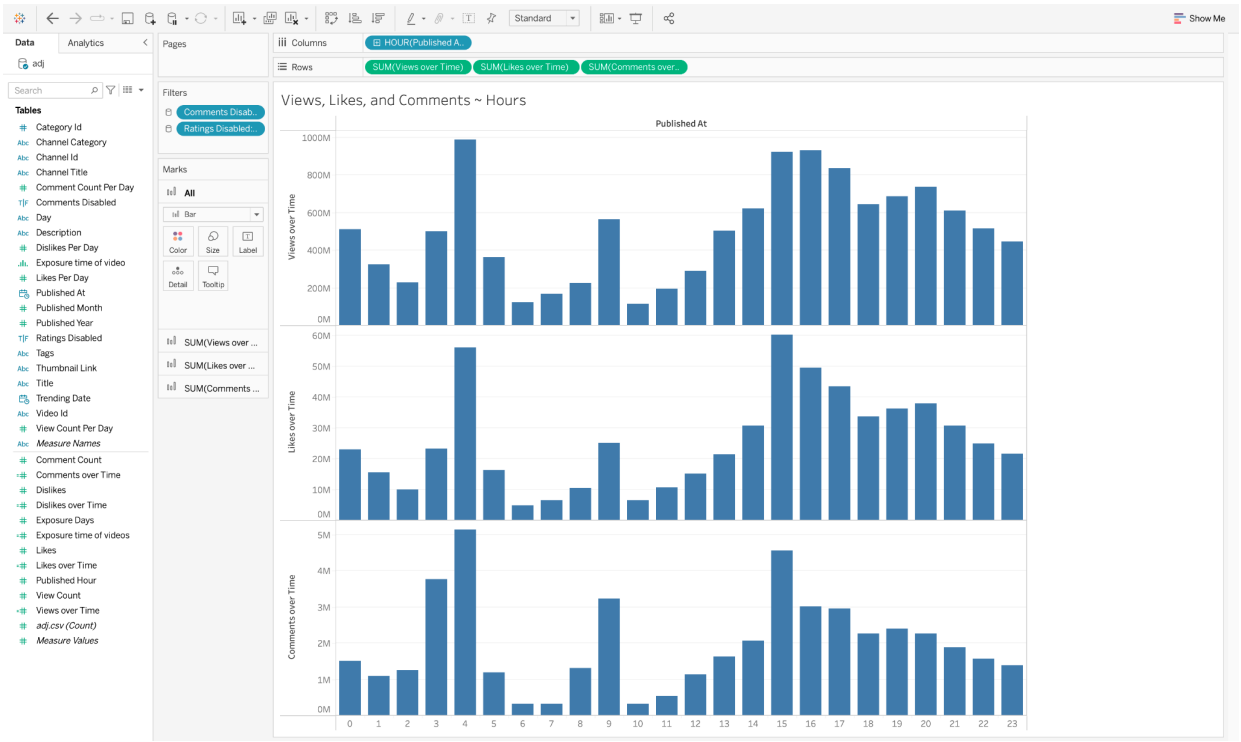
$$\text{popularity per day} = \frac{Lastest\ number\ of\ \text{popularity}}{Exposed\ days}$$

$$= \frac{Lastest\ number\ of\ \text{popilarity}}{Latest\ value\ of\ treanding\ date - Published\ date}$$

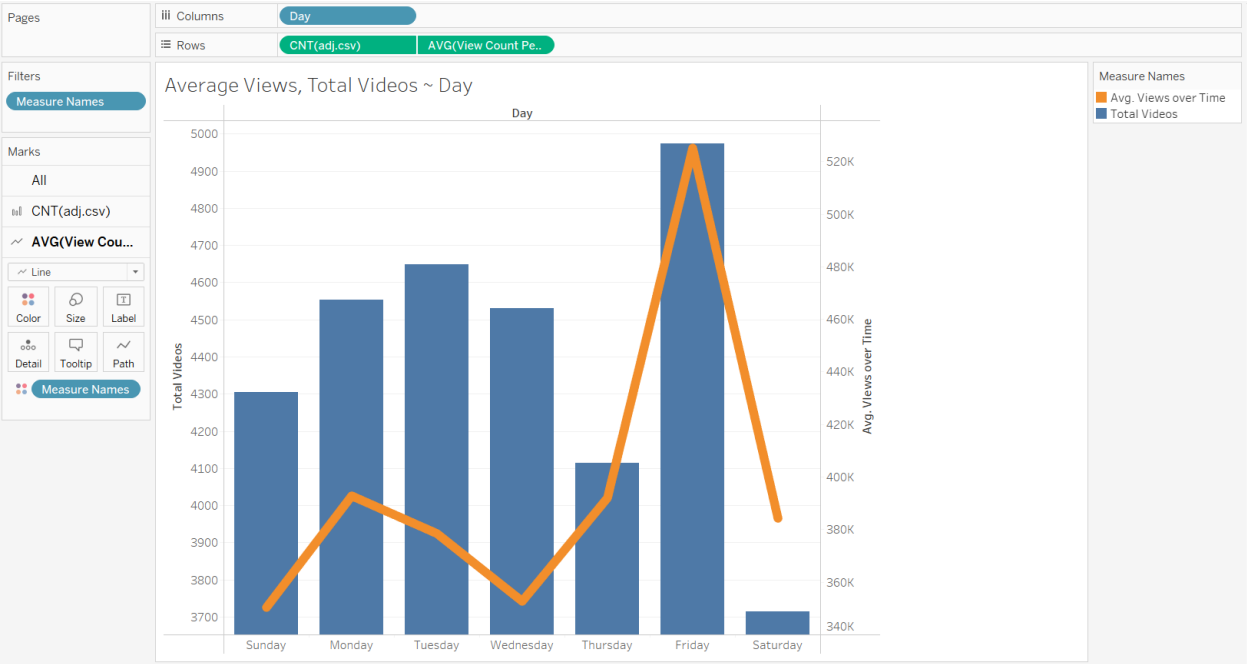
  - In addition, the data were grouped using video ID and channel ID, and data cleaning such as data noise removal was performed. Video publishing time and views of the publishing time were compared as follows, which shows that YouTubers tended to publish their videos mainly in the afternoon (usually at 4 p.m). Surprisingly, however, viewers tended to watch a lot of content published in the morning, especially between 8 and 9 a.m.



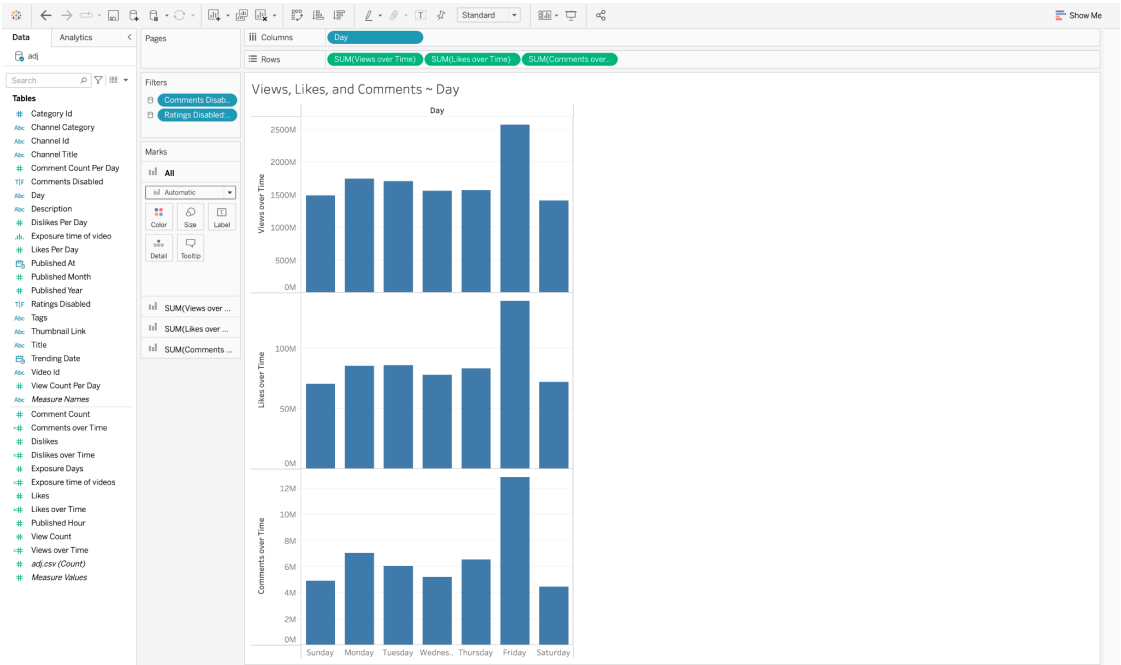

- Also from the graph, we can see that the likes and comments are following the trend of the views. The only difference is that the likes and comments decrease a little after the video was posted for 16 hours.



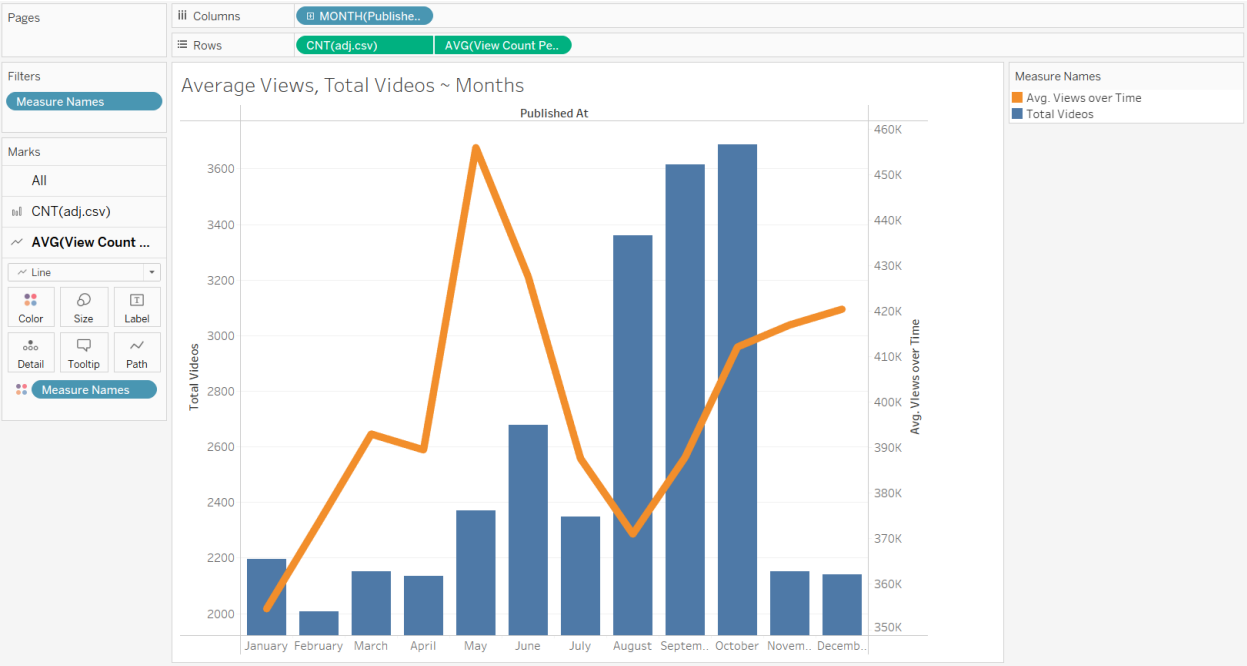
- Next, we visualized to compare the characteristics of the day of the week. YouTubers published their videos the most on Friday and showed the lowest number on Saturday. Viewers also watched the video, which was published on Friday, the most. Interestingly, viewers tended not to watch videos published on Wednesday.



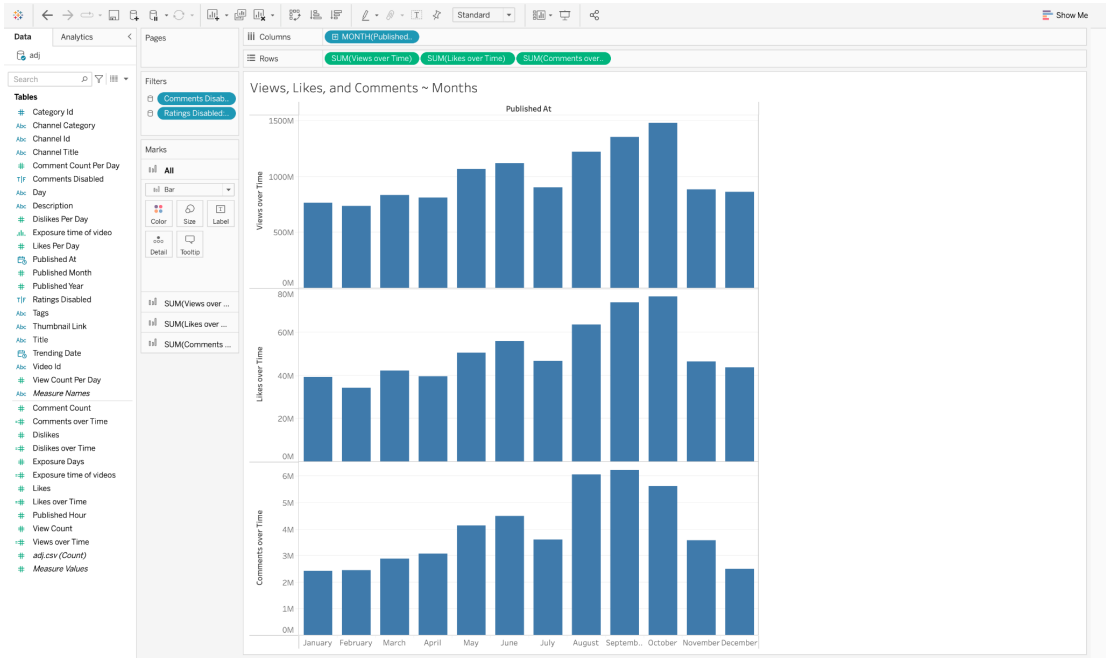

- Also, the viewers give likes and comments on Friday mostly. Besides Friday, we can see that other days have uniform distribution. The reason that people like to watch videos on Friday maybe is that it is the first night of the weekend, and people want to relax from watching the videos.



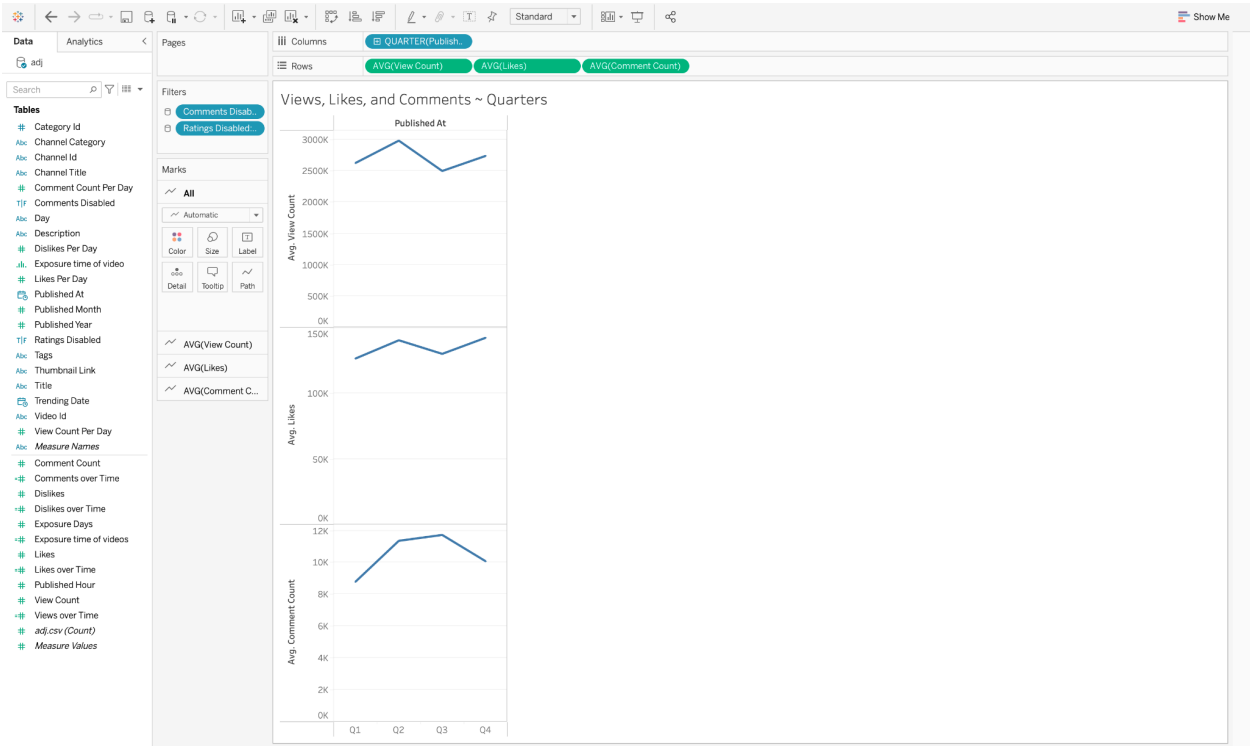
- We checked whether the difference between these YouTubers and viewers appears on a monthly basis. As shown in the chart below, the difference was also shown by month. While YouTubers usually publish videos from August to October, viewers mainly watch videos published in May, November, and December.



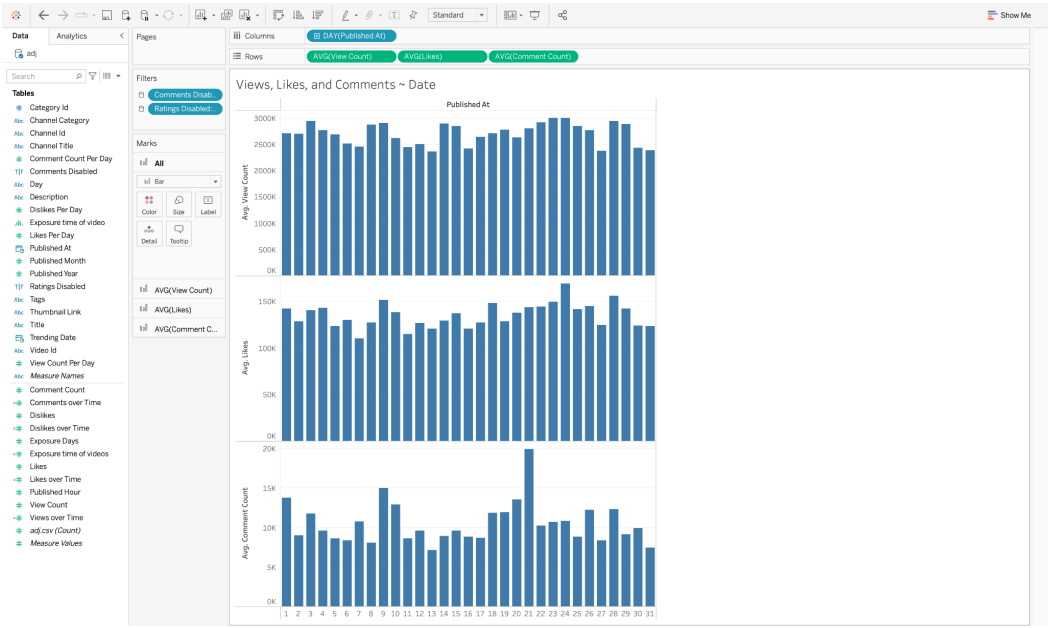

- The number of likes and comments follows the trends of the views, they like to give likes and comments in August, September, and October.



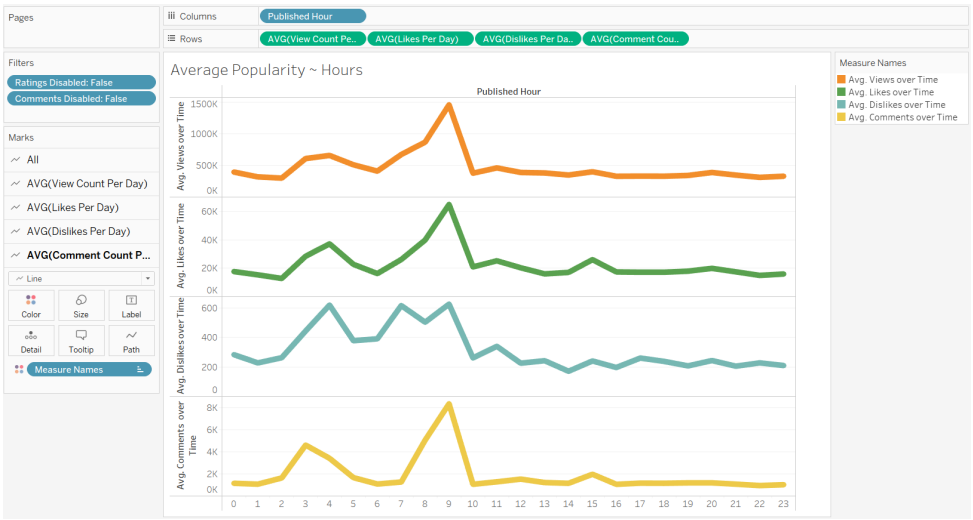
- We checked the relationship between the number of views, likes, and comments with the quarter. From this visualization, we saw that the second quarter of the year has the most views and likes, but in the third quarter of the year, it has the least views and likes. However, there is a different situation regarding the number of comments since the highest number of comments is given in the third quarter, and the least number of comments is given in the first quarter of the year.



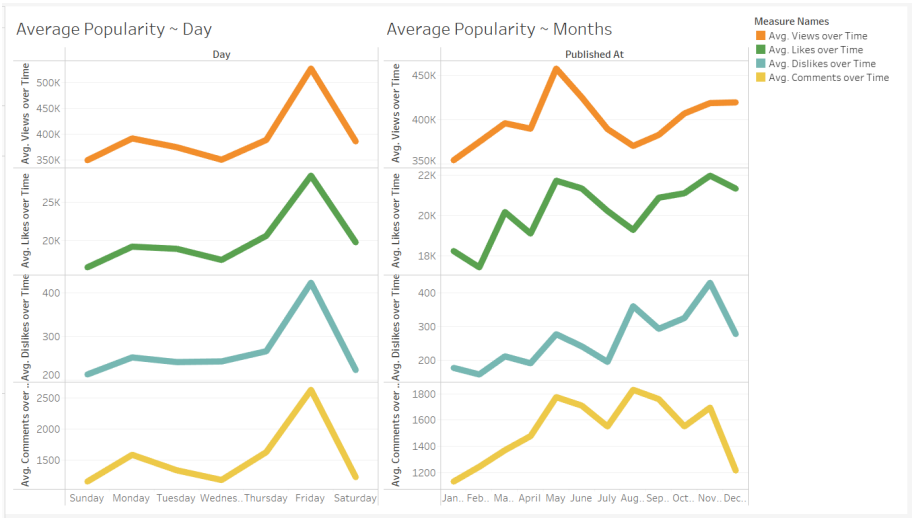
- We also checked if the number of views, likes, and comments have any relationship with the date of the month from the 1st to the 31st. From the visualization, they are a kind of uniform distribution. The highest view and likes given are on the 24th, and the most comments given are on the 21st. However, the different date of the video post does not give much difference in views and likes.



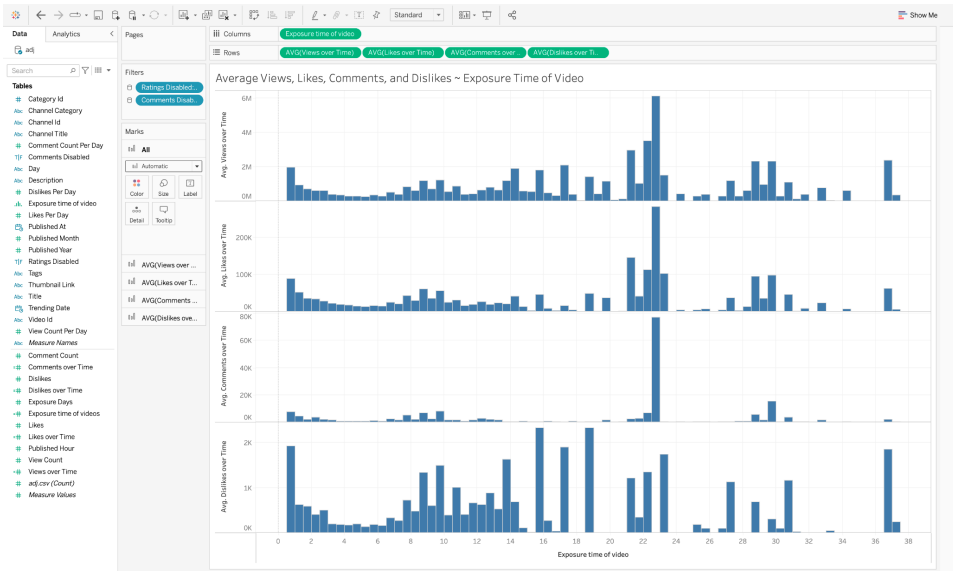

- We looked at the relationship between Views, likes, dislikes, and comments over time. Our first prediction was that the trends in Views and the rest of the variables would be quite similar. However, as a result, ‘likes’ and ‘comments’ followed the trend of ‘Views’, but ‘dislikes’ were noticeably high in videos published at 4 a.m. and 7 a.m.



- The above comparison was conducted by day of the week and month. In the trend by day of the week, all four variables showed similar trends, but in the monthly trend, viewers tended to press a lot of ‘dislikes’ on content published from October to December.




- Moreover, we checked if the video popularity will be affected by the video exposure time or not. In the visualization below, we can see that some of the videos are really popular which means that the video got views, likes, and comments immediately after the video was posted. Then the popularity decreases somehow and then increases a little bit until it hits 22.5 hours after the video post. At 22.5 hours, the video's popularity went to the highest and then went back to peace. The dislike of the video does not have this trend, it has a high dislike for the video just posted and somewhere between 13.5 and 18.5 hours, and at 36.5 hours.



- **Final Conclusions:**
  - Viewers usually watch a lot of videos published in the morning, while YouTubers usually publish videos in the afternoon.
  - Viewers watch the least videos published Wednesday, while YouTubers publish the least on Saturday
  - Viewers usually watch videos published between May and June, November and December, while YouTubers usually publish videos from August to October.
  - In general, 'likes' and 'comments' were not much different compared to 'Views', but 'dislikes' were mostly pressed on videos published at 4 a.m., and 7 a.m., and from October to December.
  - As a result, we can conclude that our hypothesis was incorrect. Mainly, Youtubers are not considering posting time to get the best popularity of their videos.
  - The popularity of the video will reach its highest in 22.5 hours after the video is posted. Also, some videos get views immediately after the video get posted.