Data Analytics on YouTube Trending Video

1. Introduction

1.1 Social media could be part of our life

Social media has received much attention in recent years due to the development of information technologies. It is increasingly being used for communicating, learning, and entertaining.In January 2022, 58.4% of the world's population used social media. The average daily usage is 2 hours and 27 minutes and 4.62 billion people around the world now use social media(Global WebIndex). 424 million new users have come online within the last 12 months. According to data, the top three most used social media platforms in the world are, Facebook, YouTube and WhatsApp, respectively 2910, 2562, 2000 (million). This article is based on YouTube as it is the largest platform for UGC content.

YouTube is the most well-known video-hosting service in the social media domain. Unlike traditional media, YouTube allows users to interact, engage, view, collaborate, and primarily assess their system of communication (Gill, Arlitt, Li, & Mahanti, 2007). As famously known, YouTube was founded in February 2005 by three PayPal employees. Less than 2 years later, Google acquired YouTube for a fee of $1.65 billion, at a point when the major significance of a raft of new websites based on user-generated content, such as Wikipedia, Myspace and Facebook, was becoming increasingly apparent. YouTube is the most popular dedicated video-sharing applications, YouTube generated $19.7 billion revenue in 2020, 30.4 percent increase year-on-year. Over 2.3 billion people access YouTube once a month.

1.2 Participatory culture and user-generated content

Different from the one-way model of communication of most mass media, social media represent two-way communication between consumers and the materialization of the communication content(M Sigala et al.,2012). Participatory culture is one of the most prominent features of Youtube,in which users can develop, interact, and learn suck as express their emotions and attitudes in social media, as well as find attitudes and groups that share their interests.YouTube is an original content-media-sharing website that combine media production and distribution with social networking features, making them an ideal place to create, connect, collaborate, and circulate.It started out as a UGC-based platform, and as it continued to commercialize, PGC made its way onto youtube. User-generated content (UGC) means videos that other YouTube users upload to their channels. When their videos contain content you own, their video gets a claim and your match policy gets applied. The match policy claim tells YouTube what action to take with these videos (monetize, track, or block).

1.3 Implications

For this project, we used data from the US and India to compare whether there are differences in behavior on social media platforms (Youtube) between developing and developed countries. The significance of this study is to understand the preferences of people in countries with different development status and what kind of content would be popular; It is also possible to predict for the creator how many interaction metrics (retweets, comments, likes) the posted content will receive.

1.4 Aim of group project

The following are our research questions and hypotheses. Initially, describe data. How many videos each channel title has in the U.S and India? How much time do people spend on youtube each month in the above 2 countries? What about the change in the number of users in the above two countries in the two years 2020-2022. Then we searched the data for some more detailed information, such as which month of the year has the most popular videos in two countries?

In addition we make two hypotheses, firstly that the most popular channel categories differ between the two countries during the same period, and secondly that the number of likes on a video is related to the number of views, and we prove this hypothesis in the follow-up.

2. Related Work

YouTube is the second largest social media outlet in the world, it has risen to become one of the most relevant mass communication media in the last decade and is currently receiving a lot of attention in the academic world. YouTube also stores detailed information about user interaction and makes some data public, which greatly increases the opportunities for quantitative research. At present, a series of studies have realized the importance of youtube and used YouTube data to express some opinions on social phenomena. Most of these articles study the interactive factors of video and predict the popularity of video. This is very consistent with the research direction of this project. This section briefly introduces these works and their main findings.

One study introduces a random prefix sampling approach to address the issue of unknown number of videos hosted on YouTube and estimates roughly 500 million videos by May 2011 (Zhou et al.,2014).However,the popularity of YouTube content is not determined by the quantity of videos a channel uploads but by the views and engagement (YouTube, 2012).In the part of audience interaction, many youtubers will acquire a lot of followers within a few days after publishing a specific video, but most of these followers have not seen other videos of the Youtuber.(Cha et al., 2009).Further analysis of the change of video playback volume, some studies found some models. They assume that the first day after the release of most videos is the peak traffic of the video. Subsequent proof shows that this assumption seems to be feasible; it was found to produce reliable results by Szabo and Huberman (2010) and Pinto et al. (2013). They concluded that the initial number of views can not be regarded as a great indicator of whether a video is popular or not. Next year,according to the official youtube statement, the popularity of YouTube content is not determined by the quantity of videos a channel uploads but by the views and engagement (Youtube,2012). Cheng et al presents an in-depth and systematic measurement study on the characteristics of YouTube videos, added an exploration of social networks in video websites based on traditional research on YouTube videos.

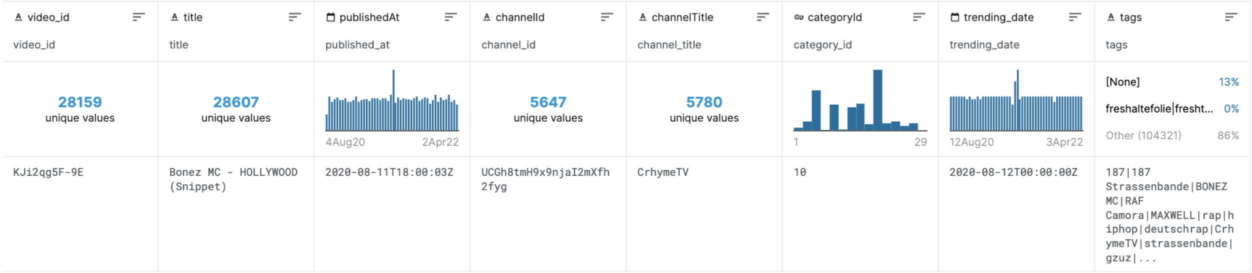
Other studies focused on the networking character of YouTube and investigated how videos are discovered and shared: Searching or clicking on YouTube suggestions was identified as the most common way to access a video (Figueiredo et al., 2011), while physical proximity of users, locality of interest (e.g. sports, news, politics) and language create barriers to geographic diffusion but, to some extent, can be overcome by social sharing (Brodersen et al., 2012). Some studies point out that youtube channels are now dominated by a few elite channels and that bloggers’ earnings vary greatly from country to country, then it shows that US channels dominate the platform, followed by India (Bernhard Rieder et al., 2020).

3. Data Description & Method

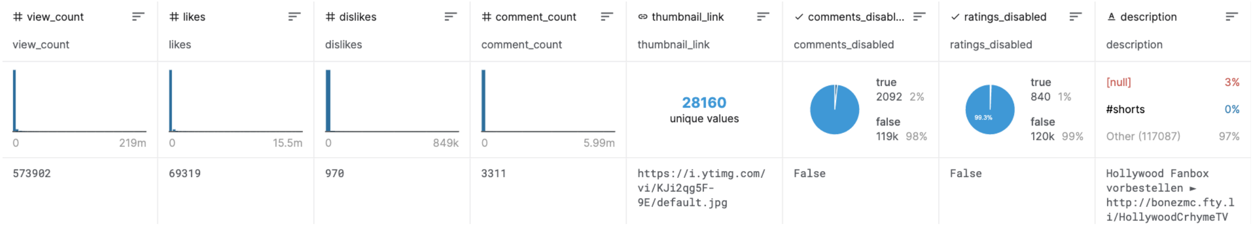
The dataset used in this project is from the top trending YouTube videos containing 14 countries (updated daily) in kaggle. , with up to 200 listed trending videos per day. Each region’ s data is in a separate file.

In our group project, we chose US and IN as the basis for our data from July 2020 to March 2022. The total number of data is 234,901, of which 120,391 are for the USA and 114,510 for India.

Data dictionary



video_id	title	publishedAt	channelId	channelTitle	categoryId	trending_date	tags
video_id	title	published_at	channel_id	channel_title	category_id	trending_date	tags
28159 unique values	28607 unique values	4Aug20 2Apr22	5647 unique values	5780 unique values	1 29	12Aug20 3Apr22	[None] 13% freshaltefolie|fresht... 0% Other (104321) 86%
KJi2qg5F-9E	Bonez MC - HOLLYWOOD (Snippet)	2020-08-11T18:00:03Z	UCGh8tmH9x9njaI2mXfh2fyg	CrhymeTV	10	2020-08-12T00:00:00Z	187|187 Strassenbande|BONEZ MC|RAF Camora|MAXWELL|rap|hiphop|deutschrap|CrhymeTV|strassenbande|gzuz|...



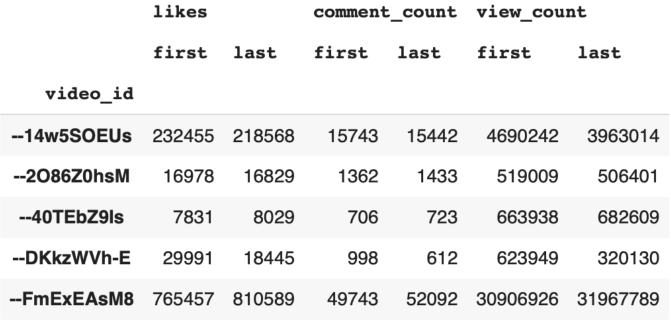
view_count	likes	dislikes	comment_count	thumbnail_link	comments_disabl...	ratings_disabled	description
view_count	likes	dislikes	comment_count	thumbnail_link	comments_disabled	ratings_disabled	description
0 219m	0 15.5m	0 849k	0 5.99m	28160 unique values	true 2092 2% false 119k 98%	99.3% true 840 1% false 120k 99%	[null] 3% #shorts 0% Other (117087) 97%
573902	69319	970	3311	https://i.ytimg.com/vi/KJi2qg5F-9E/default.jpg	False	False	Hollywood Fanbox vorbestellen ► http://bonezmc.fty.li/HollywoodCrhymeTV

Data pre-processing includes the removal of null values and the harmonization of time formats. Descriptive statistics and comparative analysis are used to show trends on youtube the U.S. and India, visualized as: bar charts, category charts, line charts. (package: pandas, seaborn, maltiplot) Using time series to show the heatmap that preferences in categories of the above two countries at different times. Then we did a correlation analysis of the video interaction metrics, where we used the (Kwak, et al., 2017) method to eliminate outliers from the data, which greatly improved the visualization, such as below figure. In the final section, Using NLP for content anaysis resulted in a comparison of topic convergence between the two countries; in addition, predictions were made using machine learning for trending-day versus numerical values, using multiple models for comparison experiments, setting a K value of 3 and finding that random forests had the highest classification accuracy.

4. Experimental Results

01. The number of likes per video in last time and first time

It can show the dynamic changes of likes , the number of comments and the number of views of each video.

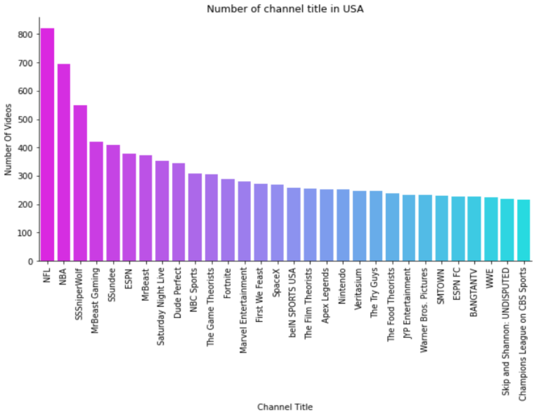


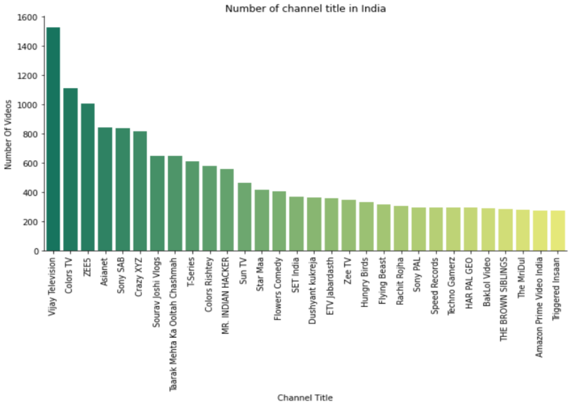
	likes		comment_count		view_count	
	first	last	first	last	first	last
video_id						
--14w5SOEUs	232455	218568	15743	15442	4690242	3963014
--2O86Z0hsM	16978	16829	1362	1433	519009	506401
--40TEbZ9ls	7831	8029	706	723	663938	682609
--DKkzWVh-E	29991	18445	998	612	623949	320130
--FmExEAsM8	765457	810589	49743	52092	30906926	31967789

02. **What is channel title distribution in the USA and India ?**

Use the bar chart to display the totals and distributions for each channel title in the two countries.

- In USA，NFL is the most popular channel title with over 800 videos, followed by NBA and SSSniperWolf
- In India,Vijay Television is the most popular channel title, with nearly 400 more videos than colors TV, while colors TV and ZEES are more popular in India.
- The channel titles that the two countries like are very different, to be specific, their top three and bottom three are completely different.



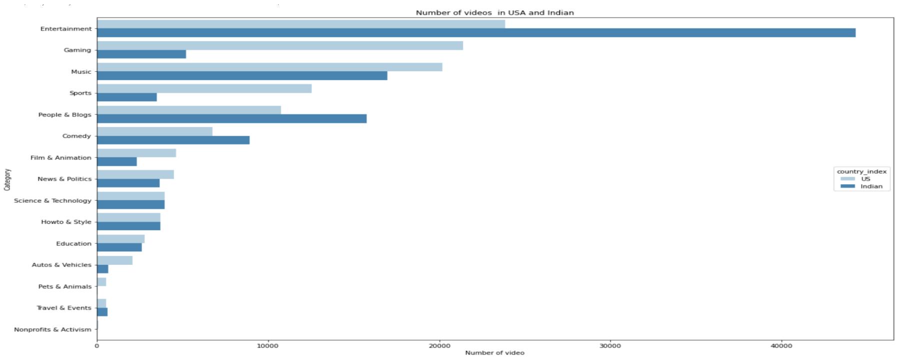


03.How many videos each categories has?

This plot uses a bar chart to show the comparison of the number of the video per category in the USA and India, the dark one symbol for USA and the light one means India.

- America's favorite genres are entertainment, games and music, and Indians' favorite entertainment, music, people & blog.
- Nonprofits & Activism are the least viewed in both countries.
- The most popular categories in the United States and India are entertainment videos, and the number of entertainment videos in India is nearly 2 times that of the United States.Music is also a category that people in both countries like.

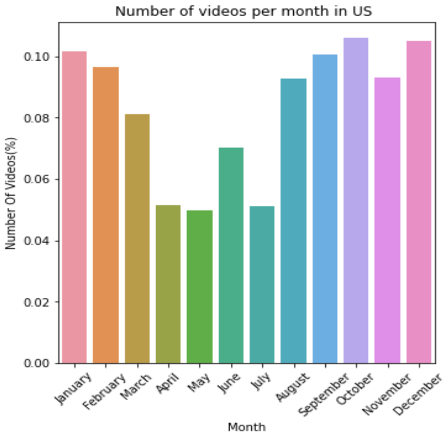
- It can be seen that Americans are very interested in game videos. The favorite is significantly higher than that of Indians in the top 3, which is nearly three times the number of viewings in India.

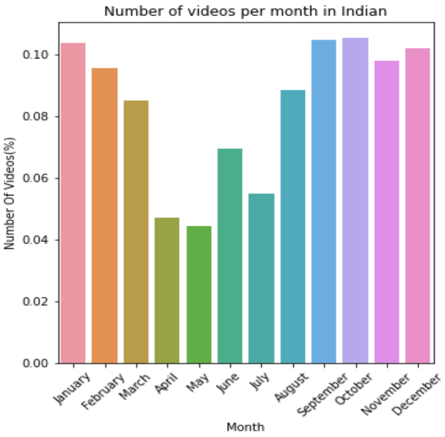


04. US and India YouTube trending monthly distribution

Use the bar chart to reflect the YouTube trending monthly distribution in the USA and India. The horizontal axis represents the month, and the vertical axis represents the proportion of the list.

- In the U.S. and India, both the beginning and end of the year are peak periods for YouTube trending, and the trends are the same in both countries.
- April, May, June, and July are all months with fewer videos in YouTube trending. January, September, October, and December are all months with more YouTube trending videos.

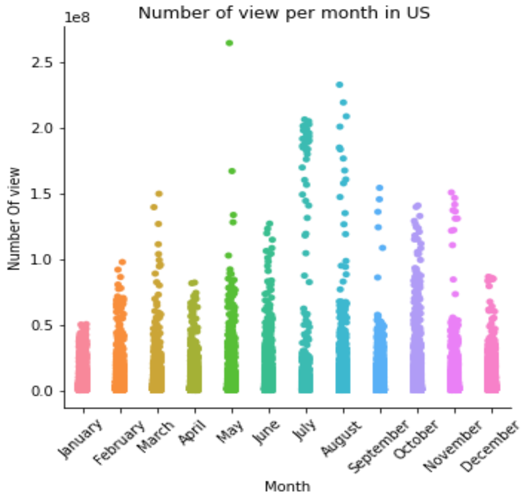


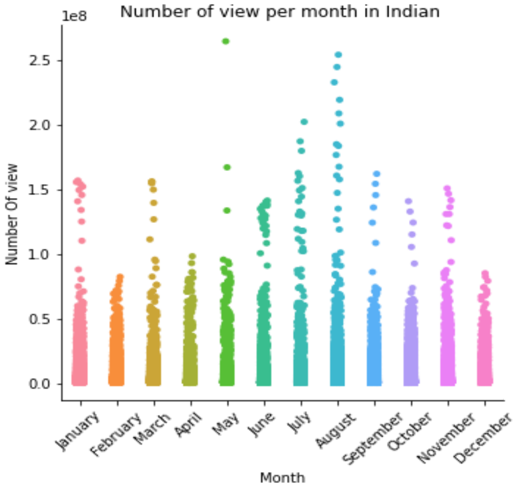


05. Number of view of videos per month

Use scatter plots to compare the maximum and minimum views of different videos and the distribution of views in each month

- Both the U.S. and India saw more views overall in August, and both saw the most viewed video of the year in May. The month with generally less streaming in the US is January; India is February
- In the United States, August has the highest dispersion of broadcasts and January the lowest; in India, August has the highest dispersion of broadcasts, and February and December are relatively concentrated



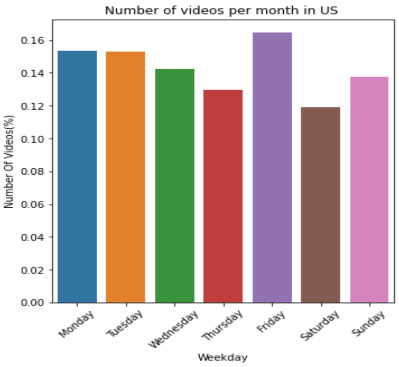


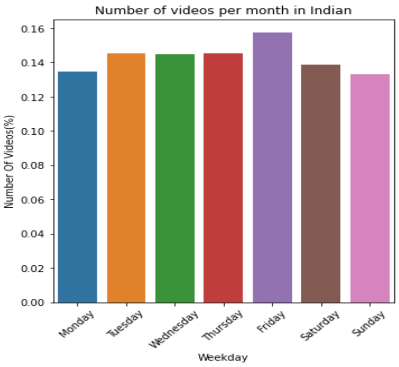
06. Number of YouTube trending videos per week

Count the number of videos published on YouTube Trending by two countries each week, and use the bar chart to display.

- The number of videos in the United States fluctuates greatly during the week, with the highest number of videos on the list on Friday and the least number of videos on the list on Saturday

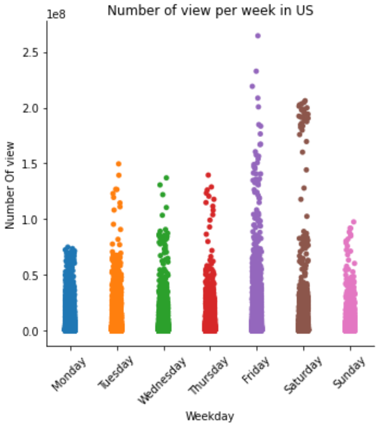
- India's intra-week distribution fluctuates less, and as in the United States, Friday is the day with the most videos on the list, but the day with the least is Monday.

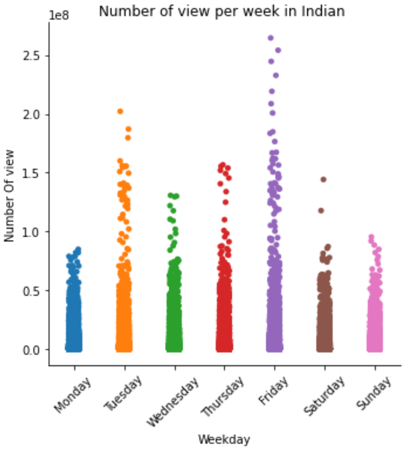




07.Number of view of videos per week

- The most weekly video views in the United States and India are on Fridays and the most discrete distribution; the most concentrated average views are on Mondays
- Weekly viewing fluctuations and dispersion are similar in the US and India

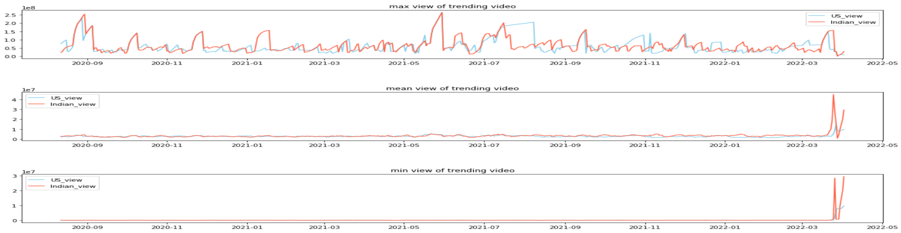




08.the max,mean,mini view of trending comparison

Using line chart compares the dynamic change between two countries from 2020.08 to 2022.05

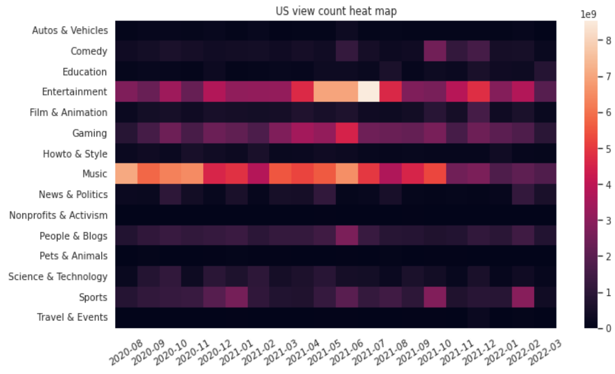
- Over time, the playback volume of the video is not much different between the maximum and the average.
- On the data, you can see that the average and maximum playback volume of late March and early April suddenly became very large
- Both the US and India have significant play peaks in September 2020 and late March 2022, and India has a peak in June 2021

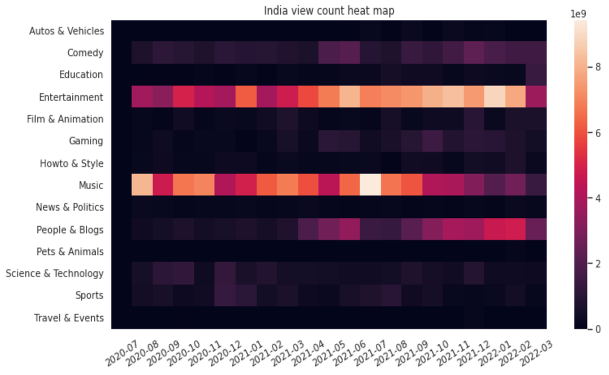


09. Changes in categories over time

Heatmap of different categories over time in the US and India

- In the U.S., entertainment, film & animation, animation and music are the most popular, and the popularity of music has weakened recently
- In India, entertainment and music are also very popular, and entertainment has become more and more popular in recent years, and the popularity of music has weakened

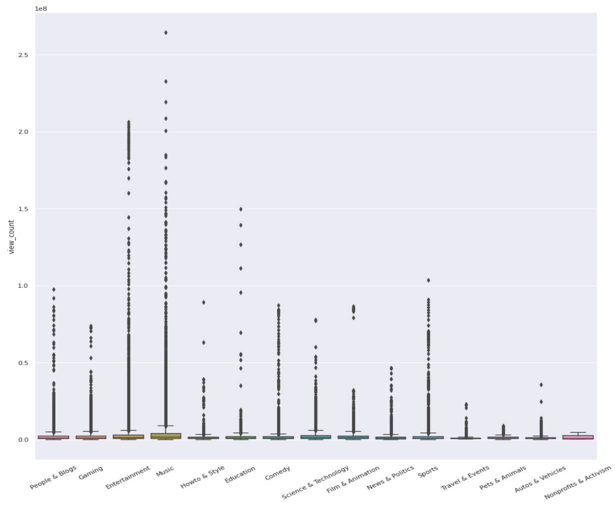


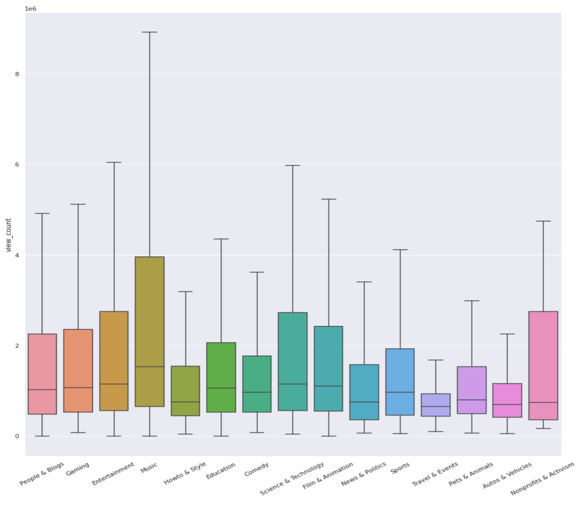


10. view_count with outliers and no outliers

When calculating the correlation, we need to delete outliers and eliminate interference

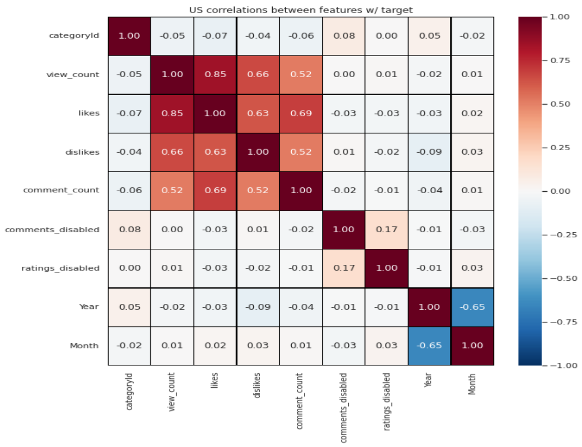
- reference:
  https://humansofdata.atlan.com/2018/03/when-delete-outliers-dataset/
  https://towardsdatascience.com/ways-to-detect-and-remove-the-outliers-404d16608dba



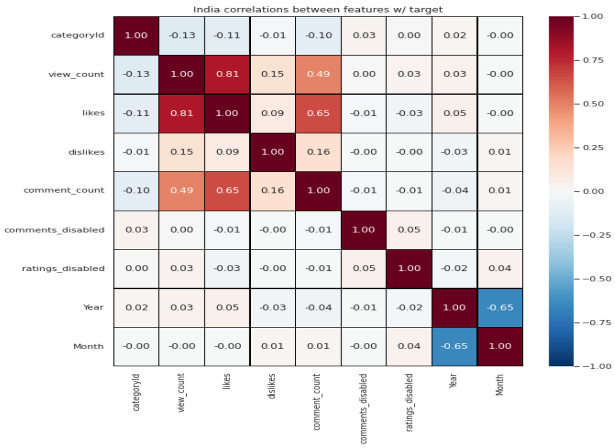


11. Correlation analysis between different columns

- In the United States, likes has the greatest impact on view_count, which is positively correlated; in addition, dislike and comment_count also have a certain impact, which is a positive correlation
- In India, view_count and like also have a strong positive correlation, and comment_count also has a relatively high correlation with view_count and like
- It can be found that dislike has little effect on other factors in India, but the opposite is true in the United States


US correlations between features w/ target

	categoryId	view_count	likes	dislikes	comment_count	comments_disabled	ratings_disabled	Year	Month
categoryId	1.00	-0.05	-0.07	-0.04	-0.06	0.08	0.00	0.05	-0.02
view_count	-0.05	1.00	0.85	0.66	0.52	0.00	0.01	-0.02	0.01
likes	-0.07	0.85	1.00	0.63	0.69	-0.03	-0.03	-0.03	0.02
dislikes	-0.04	0.66	0.63	1.00	0.52	0.01	-0.02	-0.09	0.03
comment_count	-0.06	0.52	0.69	0.52	1.00	-0.02	-0.01	-0.04	0.01
comments_disabled	0.08	0.00	-0.03	0.01	-0.02	1.00	0.17	-0.01	-0.03
ratings_disabled	0.00	0.01	-0.03	-0.02	-0.01	0.17	1.00	-0.01	0.03
Year	0.05	-0.02	-0.03	-0.09	-0.04	-0.01	-0.01	1.00	-0.65
Month	-0.02	0.01	0.02	0.03	0.01	-0.03	0.03	-0.65	1.00


India correlations between features w/ target

	categoryId	view_count	likes	dislikes	comment_count	comments_disabled	ratings_disabled	Year	Month
categoryId	1.00	-0.13	-0.11	-0.01	-0.10	0.03	0.00	0.02	-0.00
view_count	-0.13	1.00	0.81	0.15	0.49	0.00	0.03	0.03	-0.00
likes	-0.11	0.81	1.00	0.09	0.65	-0.01	-0.03	0.05	-0.00
dislikes	-0.01	0.15	0.09	1.00	0.16	-0.00	-0.00	-0.03	0.01
comment_count	-0.10	0.49	0.65	0.16	1.00	-0.01	-0.01	-0.04	0.01
comments_disabled	0.03	0.00	-0.01	-0.00	-0.01	1.00	0.05	-0.01	-0.00
ratings_disabled	0.00	0.03	-0.03	-0.00	-0.01	0.05	1.00	-0.02	0.04
Year	0.02	0.03	0.05	-0.03	-0.04	-0.01	-0.02	1.00	-0.65
Month	-0.00	-0.00	-0.00	0.01	0.01	-0.00	0.04	-0.65	1.00

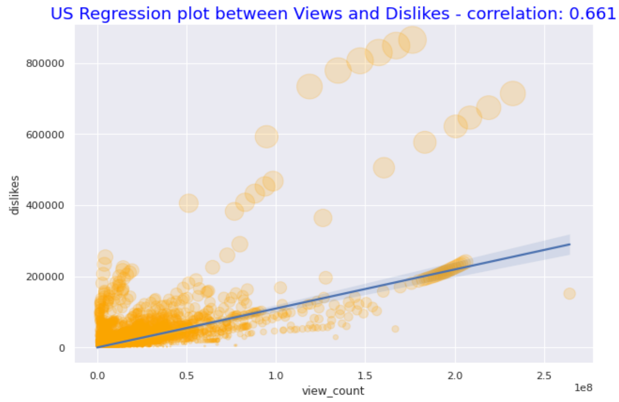
12. Comparison of dislikes and likes and the number of views in two countries

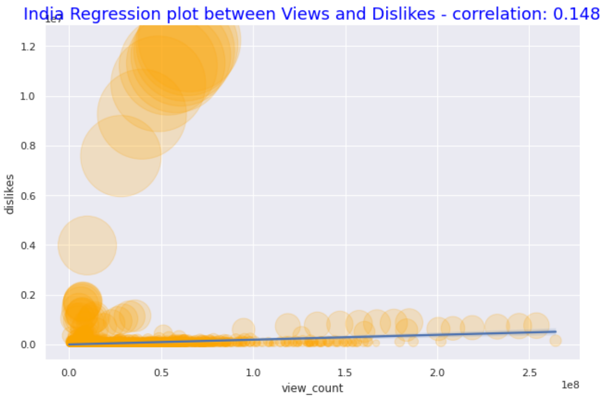
According to the above figure, it can be seen that the correlation between view_count and dislike in the two countries is quite different. Here, the linear regression model is used to separately explore the relationship between the number of views and likes and dislikes in the two countries

- In linear regression, the correlation between US video views and dislike is 0.661; but India is only 0.148, and India has a high dislike dispersion in the chart

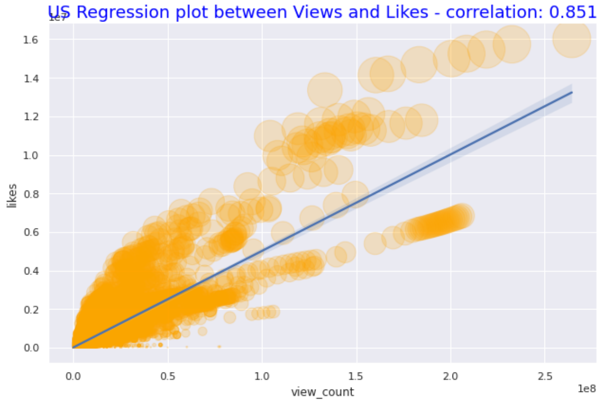
- The number of views in the United States and India has a strong correlation with likes, both exceeding 0.8, which is a strong correlation

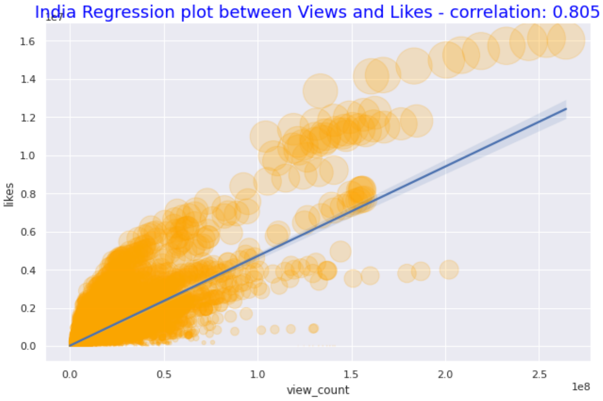
Dislike correlation in USA and India





Like correlation in USA and India

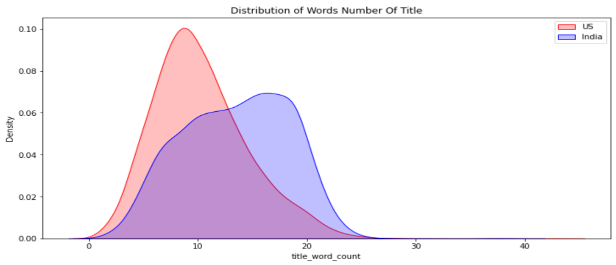


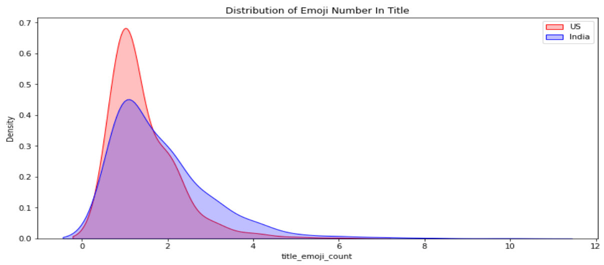


13. Comparison of the number of Title text and the number of expressions

Use stacked line chart to compare the number of text and emoji in US and Indian headers

- US and Indian video titles are written with extra emojis
- Indian titles are longer than US with text and emojis
- The U.S. uses a higher density of text and symbol titles than India

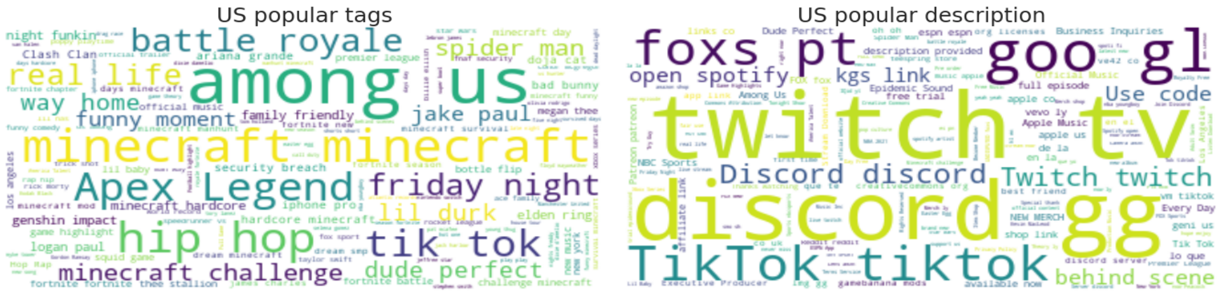




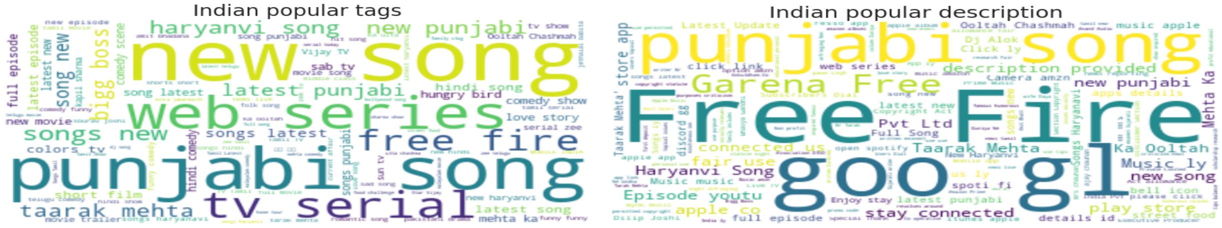
14. Comparison of word frequency between the two countries

- The description of the tag will be cleaner than the words of the description, and it will be more able to show hot topics
- The most famous tag in America is among us; the most popular description is discord, gg, twitch and TV
- The most popular video tags in India are music-related words such as new song, punjabi song, etc. At the same time, the most popular video description words in India are free, fire, etc., which also reflects the status quo of Indian society.

Wordcloud of USA

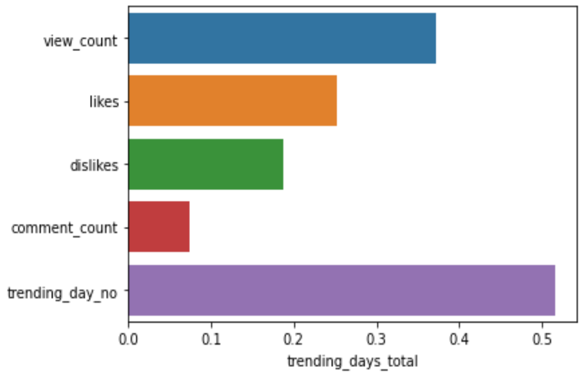


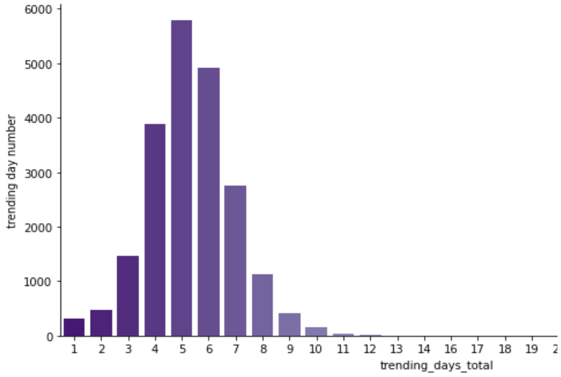
Wordcloud of India

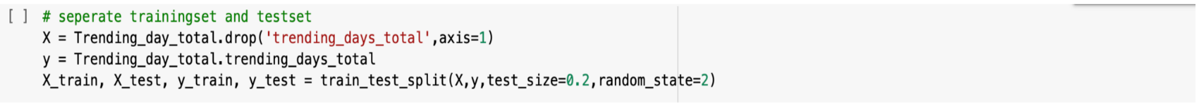


15. Predict the relationship between trending day and a numerical variable

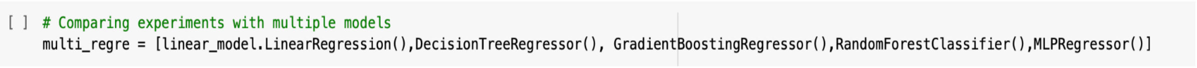
- You can see the impact of the relationship between view_count’, ’likes’, ’dislikes’, ’comment_count’ and date on trending day and predict trending day through these numerical variables
- Multiple models were used and the results showed that the random forest model had the highest accuracy.
- When R2_score = 1, the predicted value and the true value in the sample are exactly equal. When R2_score = 0, each predicted value of the sample is equal to the mean. And neg_mean_squared_error corresponds to the inverse of the mean squared error. The smaller its absolute value, the better.



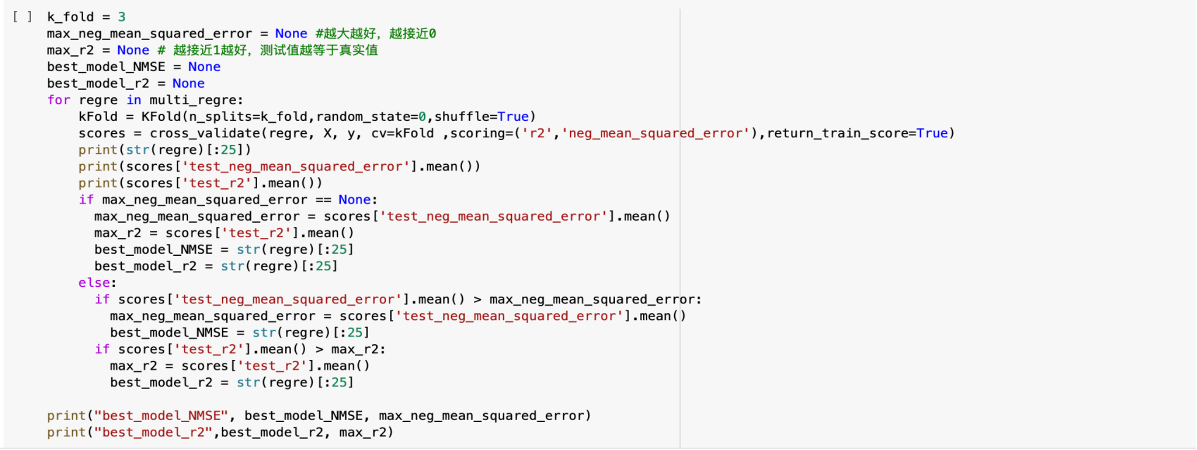




```
[ ] # seperate trainingset and testset
    X = Trending_day_total.drop('trending_days_total',axis=1)
    y = Trending_day_total.trending_days_total
    X_train, X_test, y_train, y_test = train_test_split(X,y,test_size=0.2,random_state=2)
```

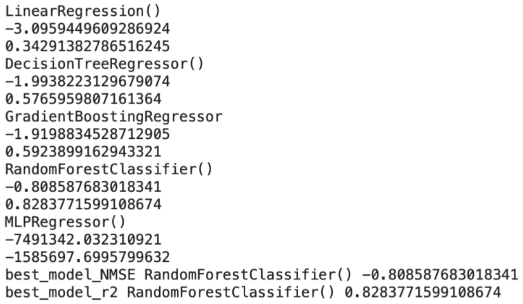


```
[ ] # Comparing experiments with multiple models
    multi_regre = [linear_model.LinearRegression(),DecisionTreeRegressor(), GradientBoostingRegressor(),RandomForestClassifier(),MLPRegressor()]
```



```
[ ] k_fold = 3
    max_neg_mean_squared_error = None #越大越好，越接近0
    max_r2 = None # 越接近1越好，测试值越等于真实值
    best_model_NMSE = None
    best_model_r2 = None
    for regre in multi_regre:
        kFold = KFold(n_splits=k_fold,random_state=0,shuffle=True)
        scores = cross_validate(regre, X, y, cv=kFold ,scoring=('r2','neg_mean_squared_error'),return_train_score=True)
        print(str(regre)[:25])
        print(scores['test_neg_mean_squared_error'].mean())
        print(scores['test_r2'].mean())
        if max_neg_mean_squared_error == None:
          max_neg_mean_squared_error = scores['test_neg_mean_squared_error'].mean()
          max_r2 = scores['test_r2'].mean()
          best_model_NMSE = str(regre)[:25]
          best_model_r2 = str(regre)[:25]
        else:
          if scores['test_neg_mean_squared_error'].mean() > max_neg_mean_squared_error:
            max_neg_mean_squared_error = scores['test_neg_mean_squared_error'].mean()
            best_model_NMSE = str(regre)[:25]
          if scores['test_r2'].mean() > max_r2:
            max_r2 = scores['test_r2'].mean()
            best_model_r2 = str(regre)[:25]

    print("best_model_NMSE", best_model_NMSE, max_neg_mean_squared_error)
    print("best_model_r2",best_model_r2, max_r2)
```



```
LinearRegression()
-3.0959449609286924
0.34291382786516245
DecisionTreeRegressor()
-1.9938223129679074
0.5765959807161364
GradientBoostingRegressor
-1.9198834528712905
0.5923899162943321
RandomForestClassifier()
-0.808587683018341
0.8283771599108674
MLPRegressor()
-7491342.032310921
-1585697.6995799632
best_model_NMSE RandomForestClassifier() -0.808587683018341
best_model_r2 RandomForestClassifier() 0.8283771599108674
```

5. Conclusion

In this project, we refer to the related research methods of YouTube video analysis in the past literature, and use the data set on kaggle to analyze the videos of the United States and India from 2020 to 2022. During the analysis, we used a variety of methods and graphs to compare the similarities and differences of YouTube videos in the US and India, and came up with the final results

6. Work Distribution Chart

Student Number	Student Name	Workload
21441243	DU He	Data Pre-processing; Correlation Analysis

21408041	JIANG Tianqi	Data Cleaning; Data Visualization
21434921	He Jiaqi	Data Visualization; Prediction
21448159	WU Jingyi	Introduction & Related work & Data description and Methods
21457328	Zhang Rui	Analysis of experimental results；text organization

7. References

Sigala, M., Christou, E., & Gretzel, U. (Eds.). (2012). Social media in travel, tourism and hospitality: Theory, practice and cases. Ashgate Publishing, Ltd..

Cha, M, Kwak, H, Rodriguez, P. (2009) Analyzing the video popularity characteristics of large-scale user generated content systems. Transaction on Networking 17(5): 1357–1370. New York: ACM

Szabo, G, Huberman, BA (2010) Predicting the Popularity of Online Content. Communications of the ACM 53(8): 80–88.

Borghol, Y, Mitra, S, Ardon, S. (2011) Characterizing and modeling popularity of user-generated videos. Performance Evaluation 68(11): 1037–1055.

Rieder, B., Coromina, Ò., & Matamoros-Fernández, A. (2020). Mapping YouTube: A quantitative exploration of a platformed media system. First Monday, 25(8), Article-number.

Kwak, S. K., & Kim, J. H. (2017). Statistical data preparation: management of missing values and outliers. Korean journal of anesthesiology, 70(4), 407–411.
https://doi.org/10.4097/kjae.2017.70.4.407