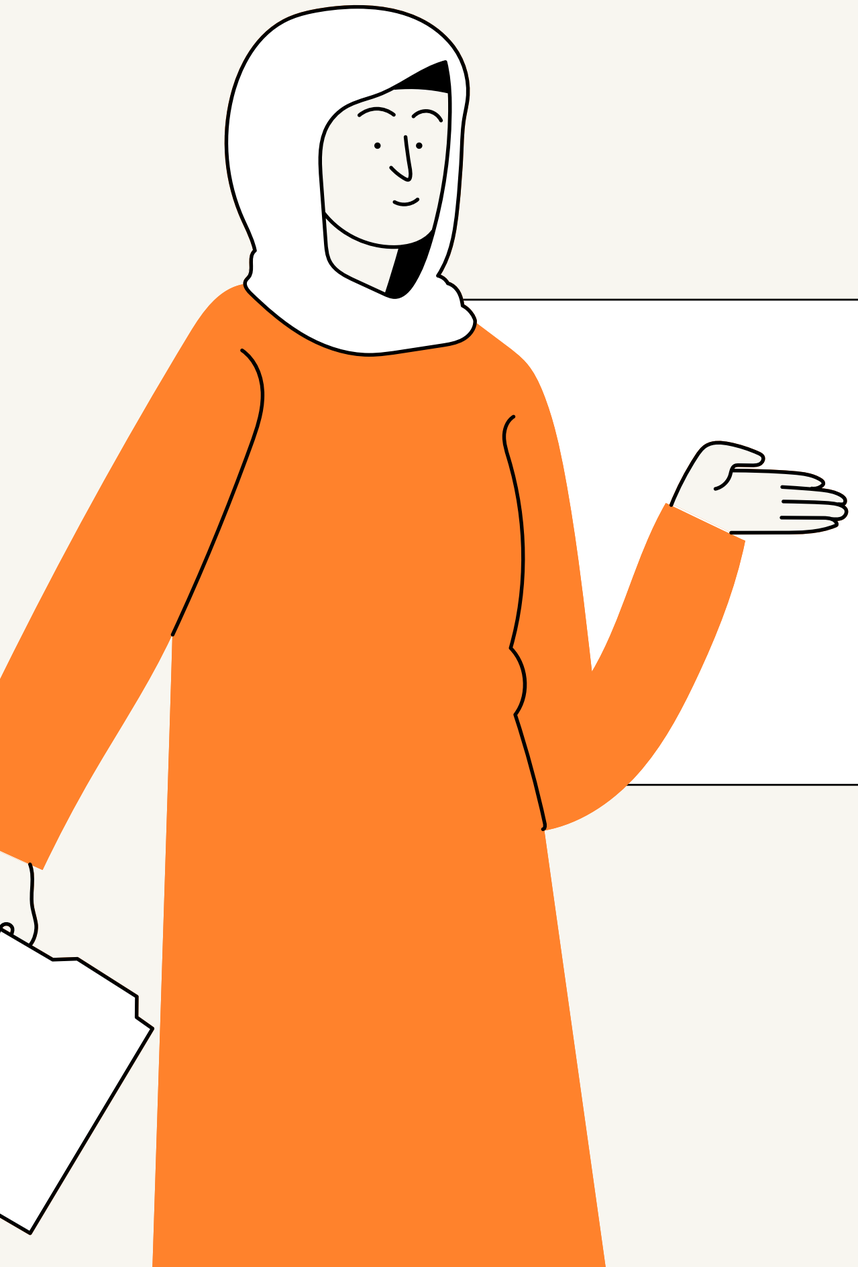
Data Analysis on YouTube Trending Video between India and the USA

Group Name: Small Ants

Overview

- Introduction
- Data Description & Method
- Data Visualization
- Experimental Results
- Conclusion

Introduction

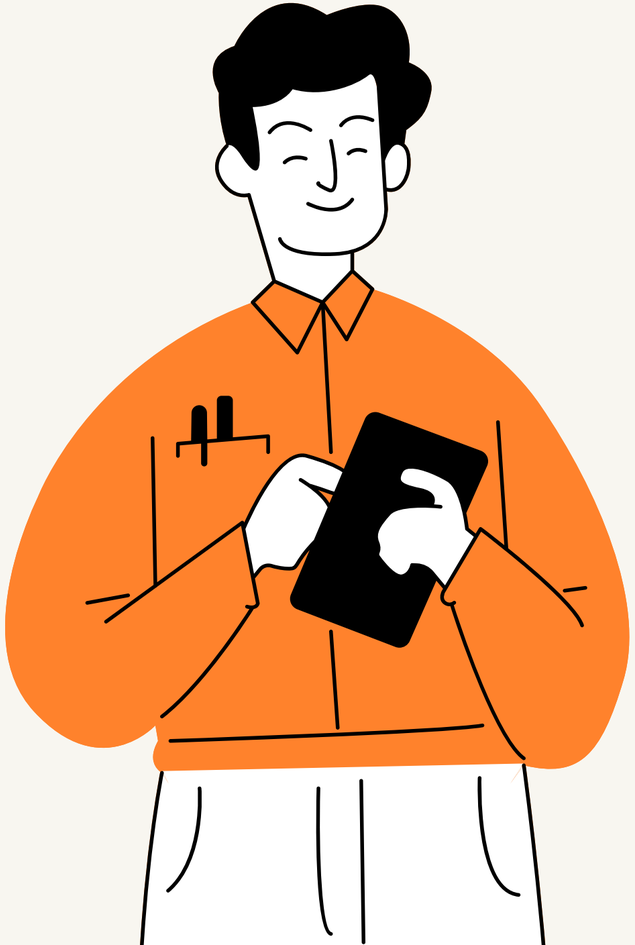


Interest of our project

- We would to know the different usage preference between developed and developing countries in social media.

- USA vs India ; YouTube

Hypothesis



Duirng the same period, two counties has different preference distribution

The categories of the trending videos differ between the two countries during the same period

The interactive metrics, likes and views wil be positive correlation

Data Description & Method

2 OUT 5	
95%	

kaggle

Create

Home

Competitions

Datasets

Code

Discussions

Courses

Dataset

- This dataset includes several months of data on daily trending YouTube videos
- Includes data several regions over the same time period
- We choose USA and IN from July 2020 to March 2022

Data Description

- There are 118478 rows and 22 columns in the USA dataset and 112360 rows and 22 columns in the India dataset.

- Qualitative Variables are title,channelId,channelTitle,categoryId are most important attributes for our analysis

- Quantitative Variables are PublishedAt,trending_date,view_count,likes, dislikes ,comment_count

Method

Practical Method In Python

- Pre-processing such as filling in missing values, processing time format
- Data visualization, correlation (by eliminate outliers)
- Prediction, multiple models for numerical value of trending day (by LinearRegression; DecisionTree; GradientBoosting; RandomForestClassifier etc)

Data Preprocessing

目录

目录

Cooperation

- Google Colab
- Build Content
- Team Work

Function： my shape print

Data truncation

```
def my_shape_print(source, df):
  print('-----------------', source, '-----------------')
  print ("Rows     : " , df.shape[0]) #Displays numbers of rows .
  print ("Columns  : " , df.shape[1]) # and column our dataset contains.
  print ("\nFeatures : \n", df.columns.tolist())#displays column names
  print ("\nMissing values :  ", df.isnull().sum().values.sum())   #find missing values
  print ("\nUnique values :  \n", df.nunique())  # Count distinct observations

#since the dataset may change i will be creating a subset till 20 March 2022).
us_data=us_data.loc[us_data['publishedAt'] < '2022-03-20']
in_data=in_data.loc[in_data['publishedAt'] < '2022-03-20']
my_shape_print('us_data', us_data)
my_shape_print('in_data', in_data)
```

```
us_data.description.fillna('No description provided',inplace=True)
display('-----us_data------', us_data.isna().sum())

in_data.description.fillna('No description provided',inplace=True)
in_data.channelTitle.fillna('No description provided',inplace=True)
display('-----in_data------', in_data.isna().sum())
```

Function： add category

```
def add_category(jsonPath, df):
  print('------------', jsonPath, '------------')
  with open(jsonPath) as file:
    categories = json.load(file)
  file.close()

  category = []
  for cate in categories["items"]:
      category.append([cate["id"],cate["snippet"]["title"]])

  df_category = pd.DataFrame(category, columns=['categoryId','category'])
  display(df_category)

  df_category.categoryId=df_category.categoryId.astype('int64')

  return df.merge(df_category, on = 'categoryId', how = 'inner')

us_data = add_category('/content/US_category_id.json', us_data)
display('----- us_data -----', us_data.head(5))

in_data = add_category('/content/IN_category_id.json', in_data)
display('----- in_data -----', in_data.head(5))
```

Function： add year month

```
def add_year_month(df):
  #changing published date , trending_date to datetime datatype.
  df.publishedAt= pd.to_datetime(df.publishedAt)
  df.trending_date= pd.to_datetime(df.trending_date)
  df['Year']=df['publishedAt'].dt.year
  df['Month']=df['publishedAt'].dt.month
  df['Weekday']=df['publishedAt'].dt.weekday

add_year_month(us_data)
display('----us_data-----', us_data.head(5))
add_year_month(in_data)
display('----in_data-----', in_data.head(5))
```

Function： add words and emoji count

```
def add_words_and_emoji_count(df):
  title_word_count_list = []
  title_emoji_count_list = []
  for index, row in tqdm(df.iterrows(), total=df.shape[0]):
    if row['title']:
      words = nltk.word_tokenize(row['title'])
      title_word_count_list.append(len(words))
      emoji_summary = adv.extract_emoji([row['title']])
      title_emoji_count_list.append(emoji_summary['emoji_counts'][0])
    else:
      title_word_count_list.append(0)
      title_emoji_count_list.append(0)
  df['title_word_count'] = title_word_count_list
  df['title_emoji_count'] = title_emoji_count_list

add_words_and_emoji_count(us_data)
add_words_and_emoji_count(in_data)
```

pandas methods -- general description

nlargest()
nsmallest()

Top and bottom view_count

```
#Select and order top 10 entries
us_data.nlargest(10,'view_count')
```

```
[ ] #Select and order bottom 10 entries.
    us_data.nsmallest(10, 'view_count')
```

	video_id	title	publishedAt	channelId	channelTitle	categoryId	trending_date	tags	vie
4122	mCY4b6GGkb4	The Funeral of The Duke of Edinburgh	2021-04-17 14:56:38+00:00	UCTkC3Jt91QkqNAE4FGWkEIQ	The Royal Family	22	2021-04-18 00:00:00+00:00	[None]	
35008	r7nYQXsxJdU	HBCU Homecoming 2020: Meet Me On The Yard	2020-10-25 01:40:31+00:00	UCqVDpXKLmKeBU_yyt_QkItQ	YouTube Originals	24	2020-10-27 00:00:00+00:00	2 CHAINZDESI BANKSILIONEL RICHIE ILANCE GROSS...	
35065	r7nYQXsxJdU	HBCU Homecoming 2020: Meet Me On The Yard	2020-10-25 01:40:31+00:00	UCqVDpXKLmKeBU_yyt_QkItQ	YouTube Originals	24	2020-10-29 00:00:00+00:00	2 CHAINZDESI BANKSILIONEL RICHIE ILANCE GROSS...	
45441	ifJYb2An7wE	Gay And Not Proud - Daniel Howell I YouTube Pr...	2021-06-25 21:04:38+00:00	UCGjylN-4QCpn8XJ1uY-UOgA	Daniel Howell	24	2021-06-27 00:00:00+00:00	PridelPride 2021IYouTube PridelYouTube Pride 2...	
45442	kmk5vciFbek	Demi Lovato performs	2021-06-26	UCZkURf9tDolEQeuw_4RD7XQ	Demi Lovato	24	2021-06-27	PridelPride 2021IYouTube	

pandas methods -- general description

first()
last()

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

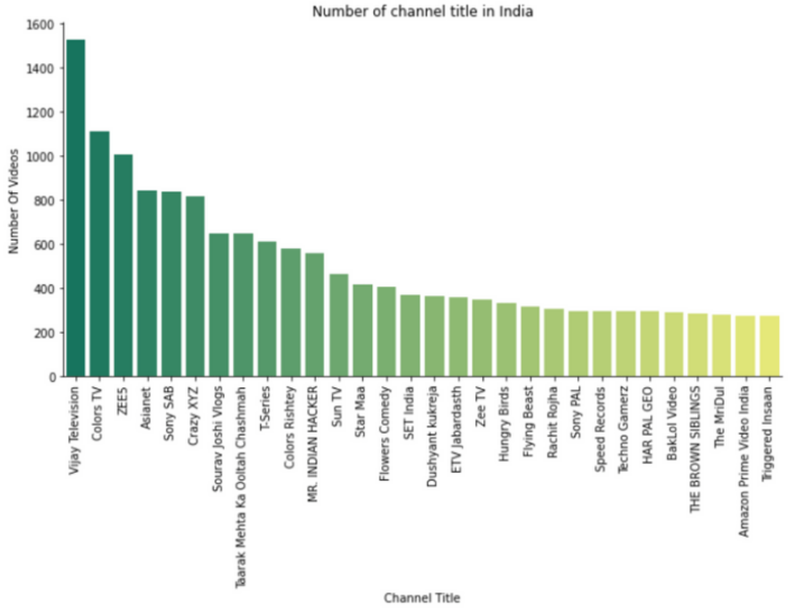
	likes		comment_count		view_count	
	first	last	first	last	first	last
video_id						
--14w5SOEUs	232455	218568	15743	15442	4690242	3963014
--2O86Z0hsM	16978	16829	1362	1433	519009	506401
--40TEbZ9Is	7831	8029	706	723	663938	682609
--DKkzWVh-E	29991	18445	998	612	623949	320130
--FmExEAsM8	765457	810589	49743	52092	30906926	31967789

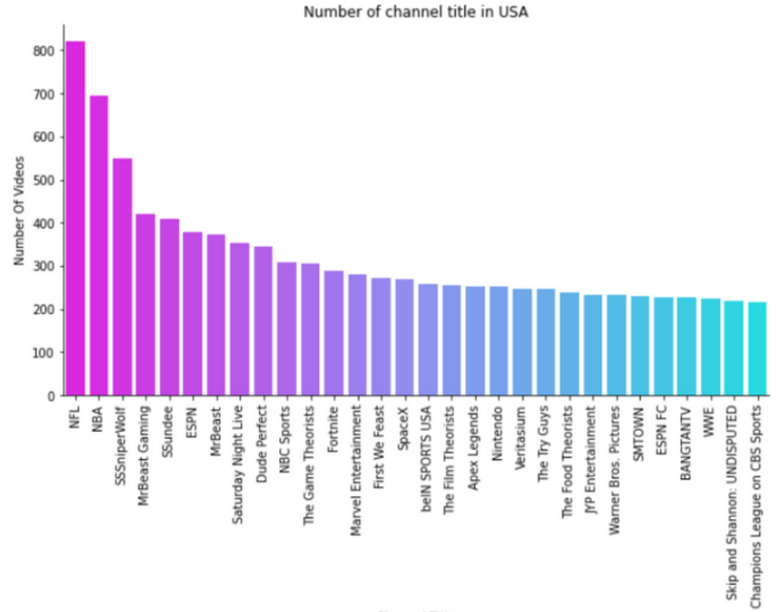
Data Visualization

What is channel title distribution in USA and India?

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

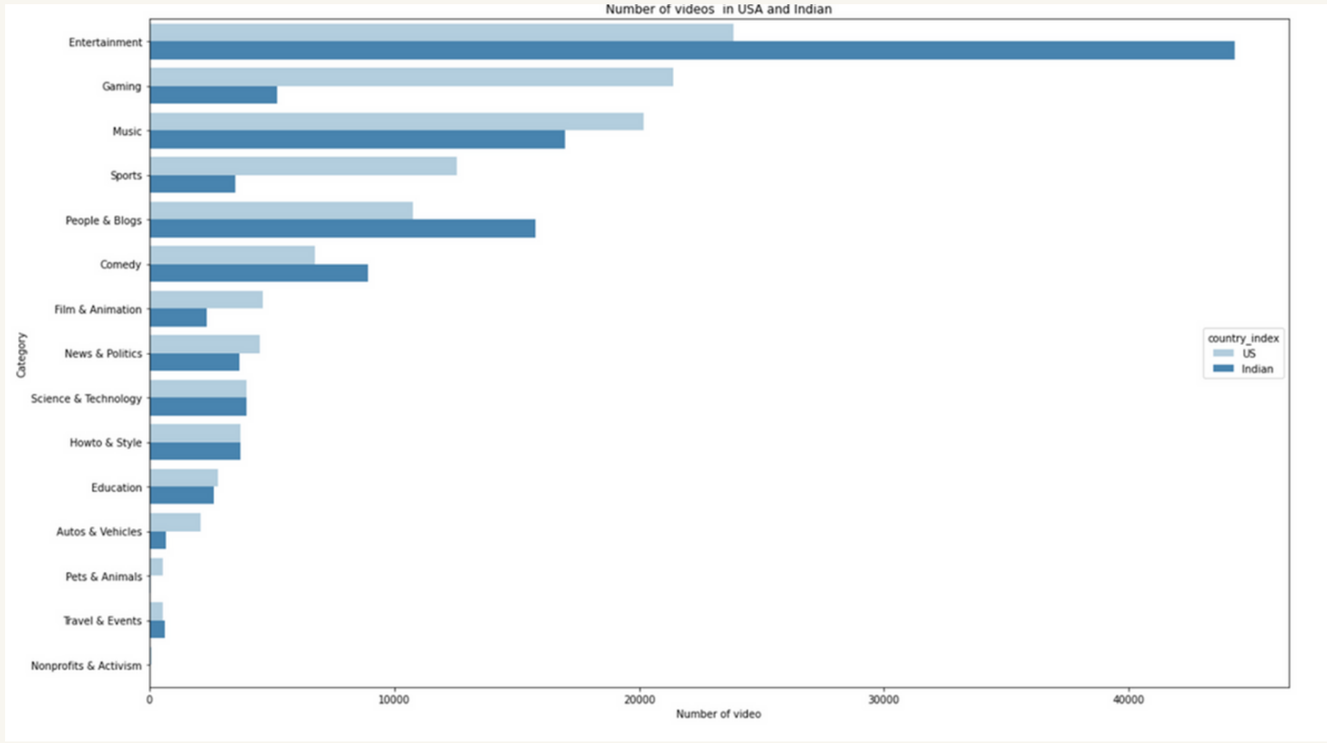
- In USA，NFL is the most popular channel title with over 800 videos, followed by NBA and SSSniperWolf
- In India, Vijay Television is the most popular channel title, followed by colors TV and ZEES.
- The channel titles are very different in two countries





How many videos each categories has?

- bar chart
- America's favorite genres are entertainment, games and music. Indians' favorite are entertainment, music, people & blog.
- Nonprofits & Activisim are the least viewed in both countries.
- Entertainment video, as the largest category, is almost twice as common in India as in the United States.
- Americans are very interested in video games, more than the music category, which is the second largest in India.

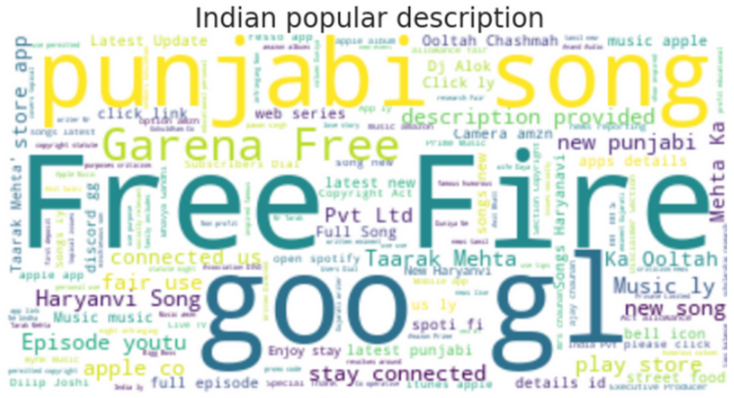


Comparison of word frequency between the two countries

- The tag will be more able to show hot topics than description, because it is cleaner.
- The most famous tag in America is among us about a game; while India are music-related words such as new song, punjabi song.
- The most popular descriptions are discord, gg, twitch and TV in US, while in India they are free fire, punjabi song.
- They almost all focus on entertainment, but the focus is different

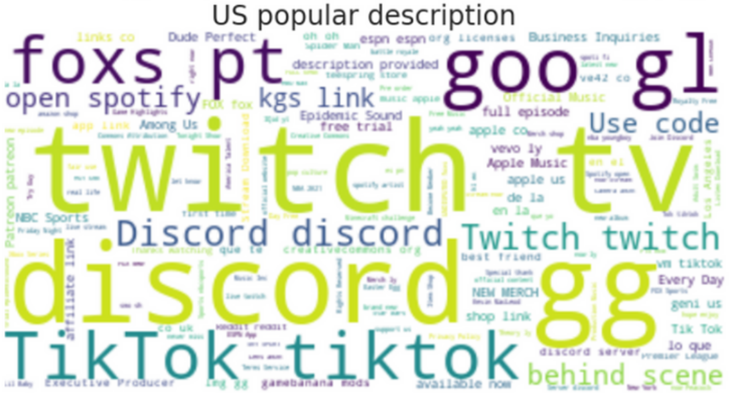
Indian popular tags

Indian popular description



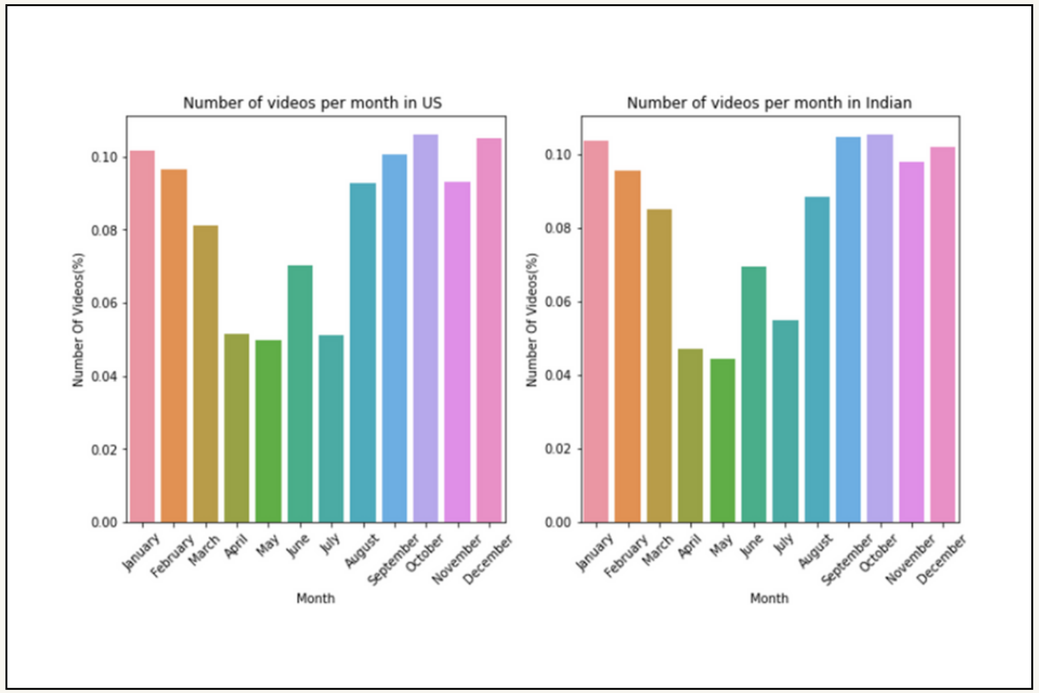
US popular tags

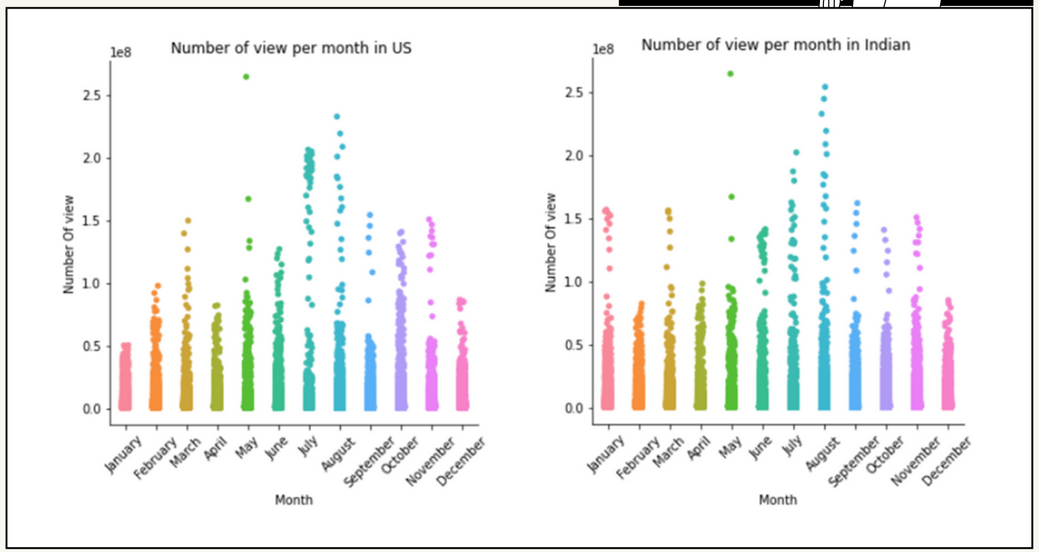
US popular description



US and India YouTube trending monthly distribution

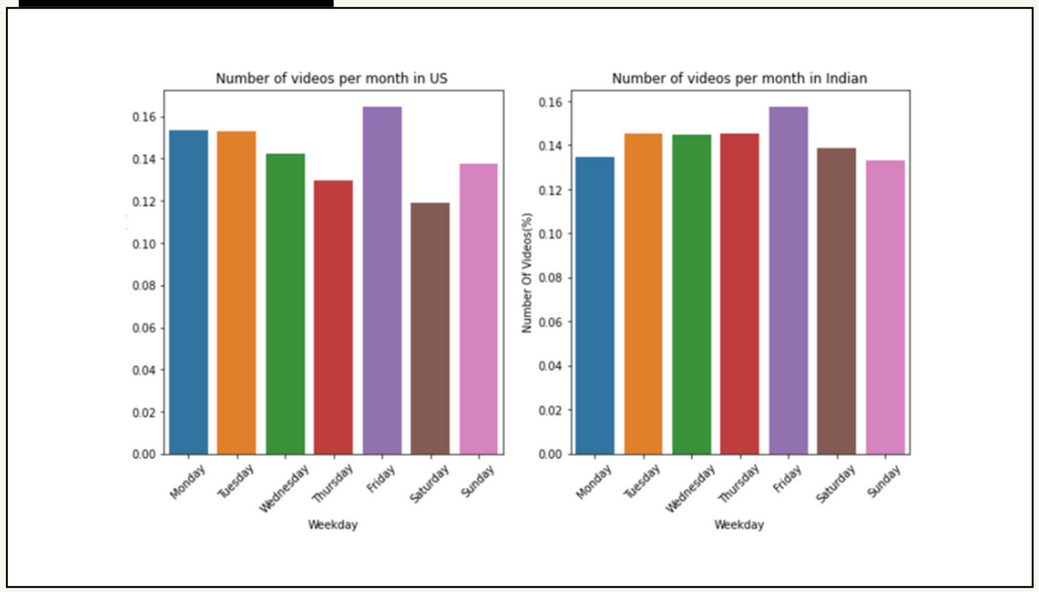
- Bar chart
- both the beginning and end of the year are peak
- The monthly trends are the same in both countries.
- April to July are all months with fewer videos
- September to January are with more YouTube trending videos.





Number of view of videos per month

- catplots
- Both saw more views overall in August
- The most viewed videos of the year are both in May. The least viewed video of the US is in January, India is in February.
- In the US, August has the highest dispersion and January the lowest; in India, August has the highest dispersion and February and December are relatively concentrated.

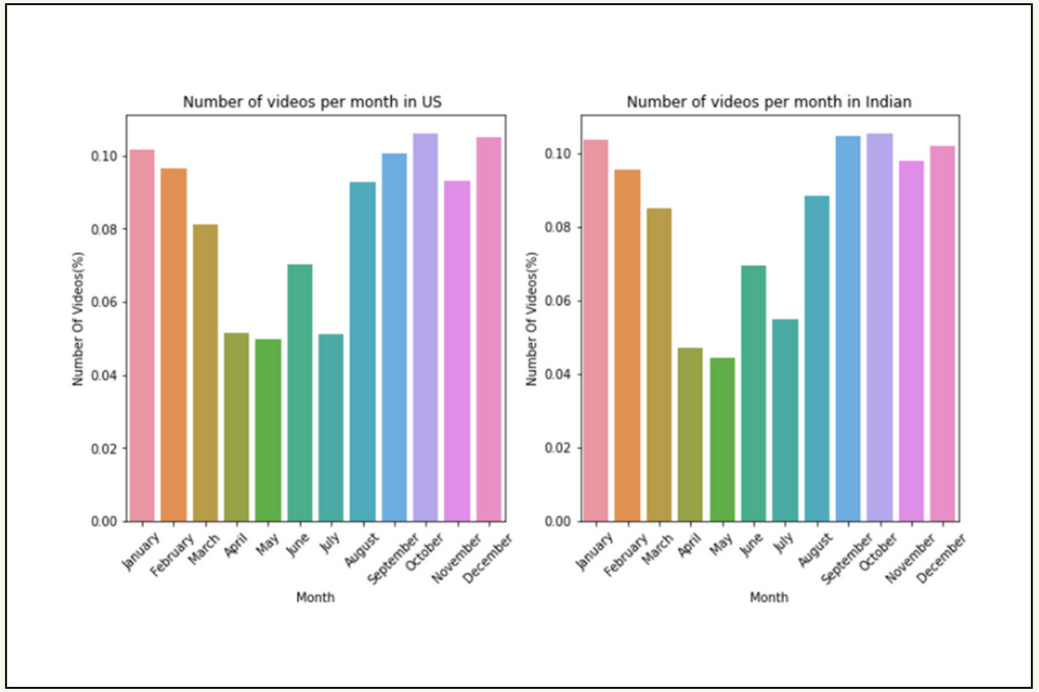


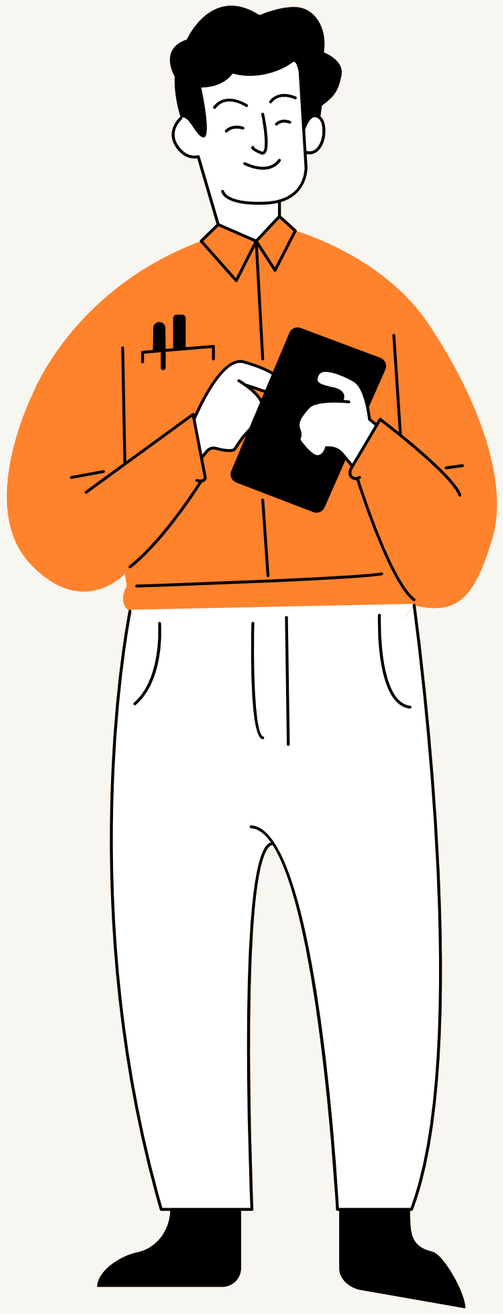
Number of YouTube trending videos weekly distribution

- bar chat
- The number of videos in the US fluctuates greatly during the week, India's distribution fluctuates less.
- The highest number of videos are both on Friday.
- The least number of videos on Saturday in the US, while Indian is Monday.

Number of view of videos per week

- catplots
- Weekly viewing fluctuations and dispersion are similar in both countries
- The most weekly video views and most discrete distribution are both on Fridays.
- The most concentrated views are on Monday



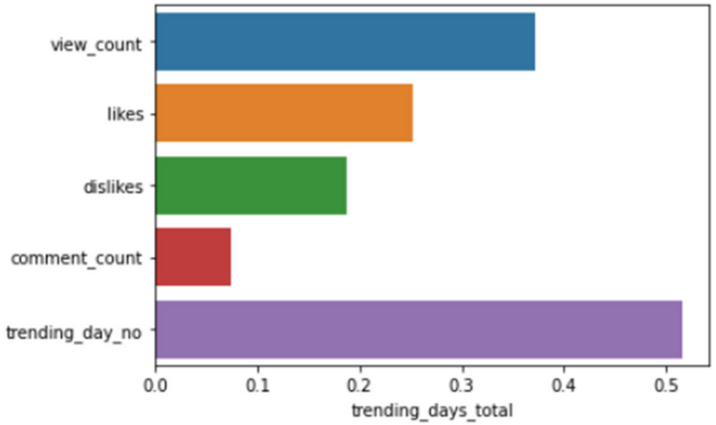


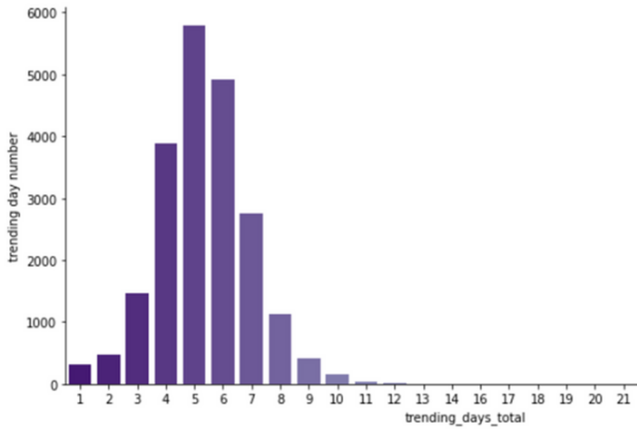
The max,mean,mini view of trending comparison

- Using line chart compares the change from 2020.08 to 2022.05
- The view counts of the video are not much different between the maximum and the average.
- The average and maximum view counts of late March and early April suddenly became very large.
- Both countries have significant play peaks in September 2020 and late March 2022, and India has a peak in June 2021.

Predict thetrending day using a numerical variable

- View Count has the largest positive correlation with Trending Day
- Videos typically last about five days





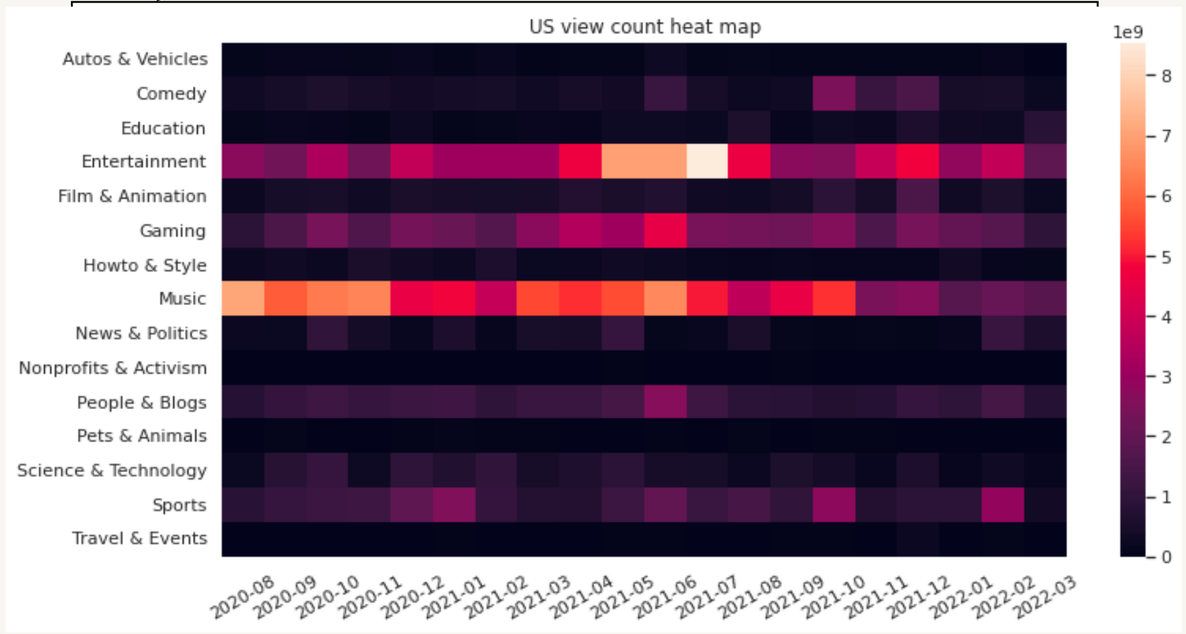

```
LinearRegression()
-3.0959449609286924
0.34291382786516245
DecisionTreeRegressor()
-1.9938223129679074
0.5765959807161364
GradientBoostingRegressor
-1.9198834528712905
0.5923899162943321
RandomForestClassifier()
-0.808587683018341
0.8283771599108674
MLPRegressor()
-7491342.032310921
-1585697.6995799632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

$$R^2 = 1 - \frac{\sum_i \left(\hat{y}^{(i)} - y^{(i)}\right)^2}{\sum_i \left(\overline{y} - y^{(i)}\right)^2}$$

$$\frac{1}{m}\sum_{i=1}^{m}\left(y_{\text{test}}^{(i)} - \hat{y}_{\text{test}}^{(i)}\right)^2 = MSE$$

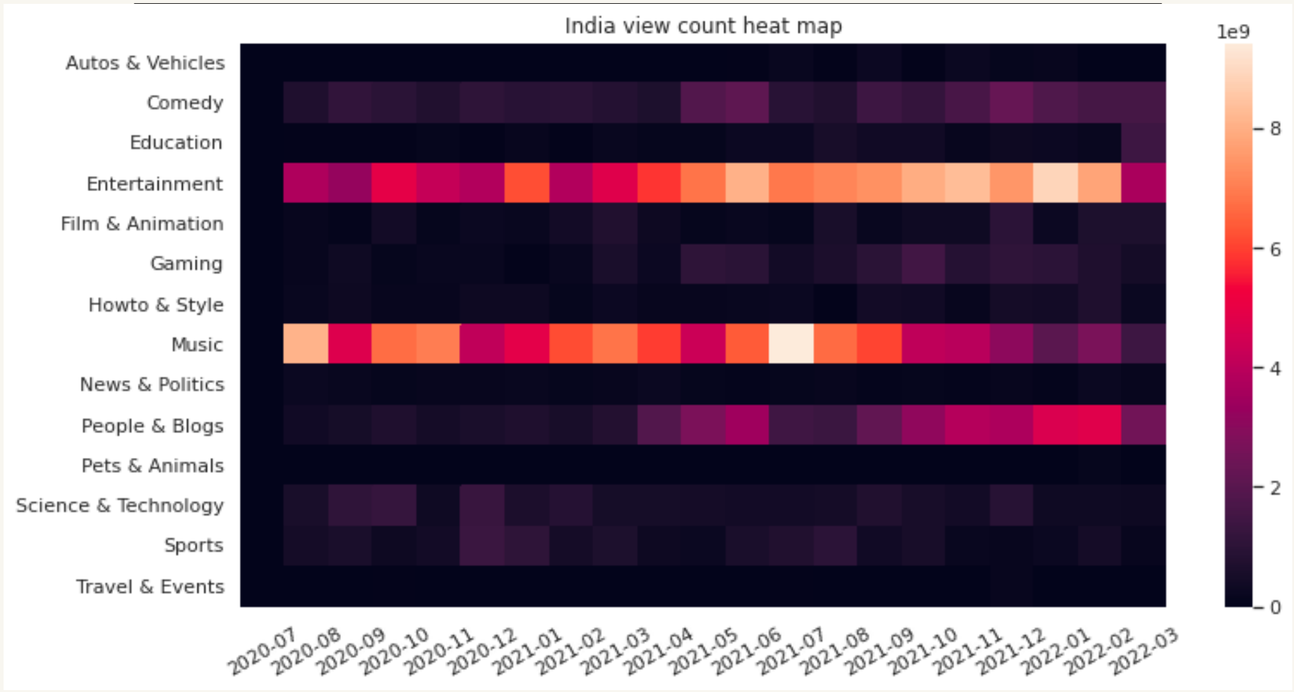
- R2_score ranges from 0 to 1. The closer you get to 1, the closer you get to the true value
- neg_mean_squared_error corresponds to the inverse of the mean squared error. The smaller its absolute value, the better.
- Multiple models were used and the results showed that the random forest model had the highest accuracy.

Changes In Categories Over Time

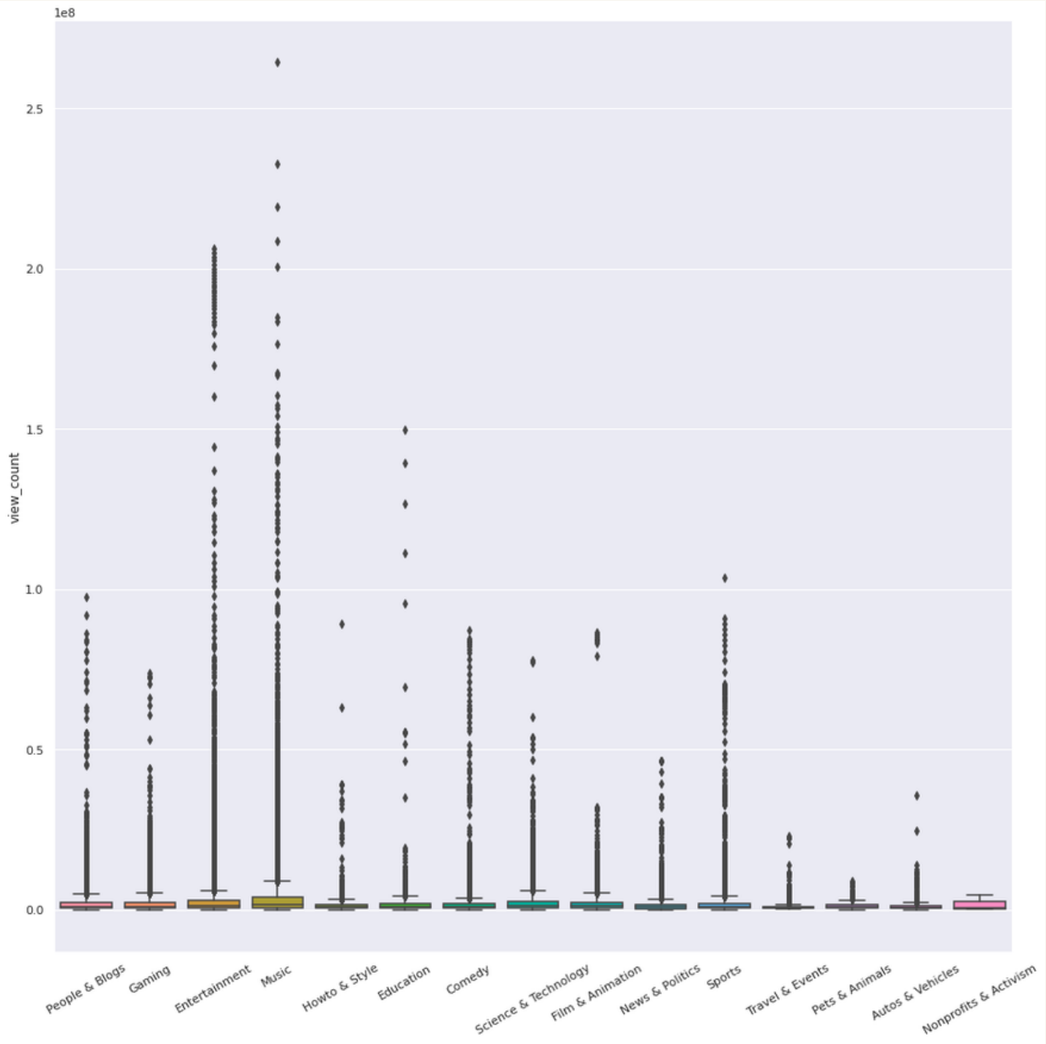


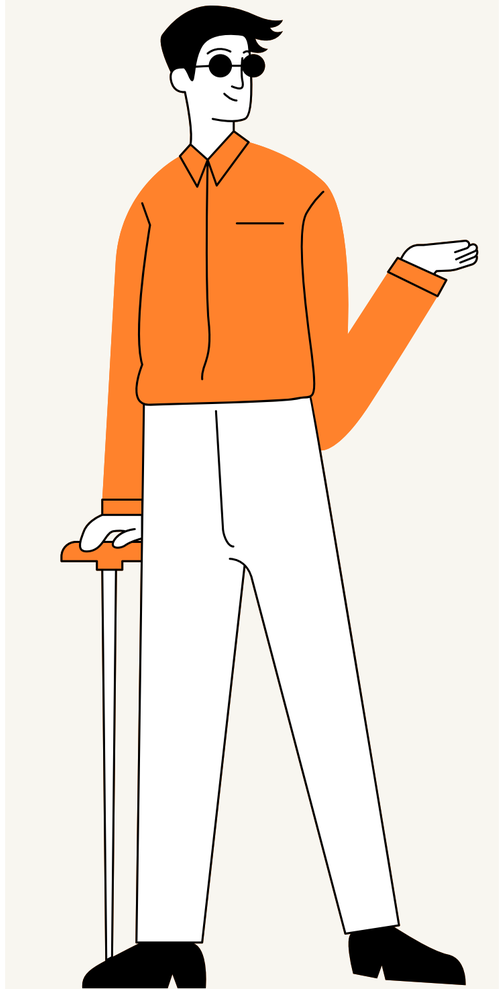
In the U.S., entertainment, film & animation, animation and music are the most popular, and the popularity of music has weakened recently

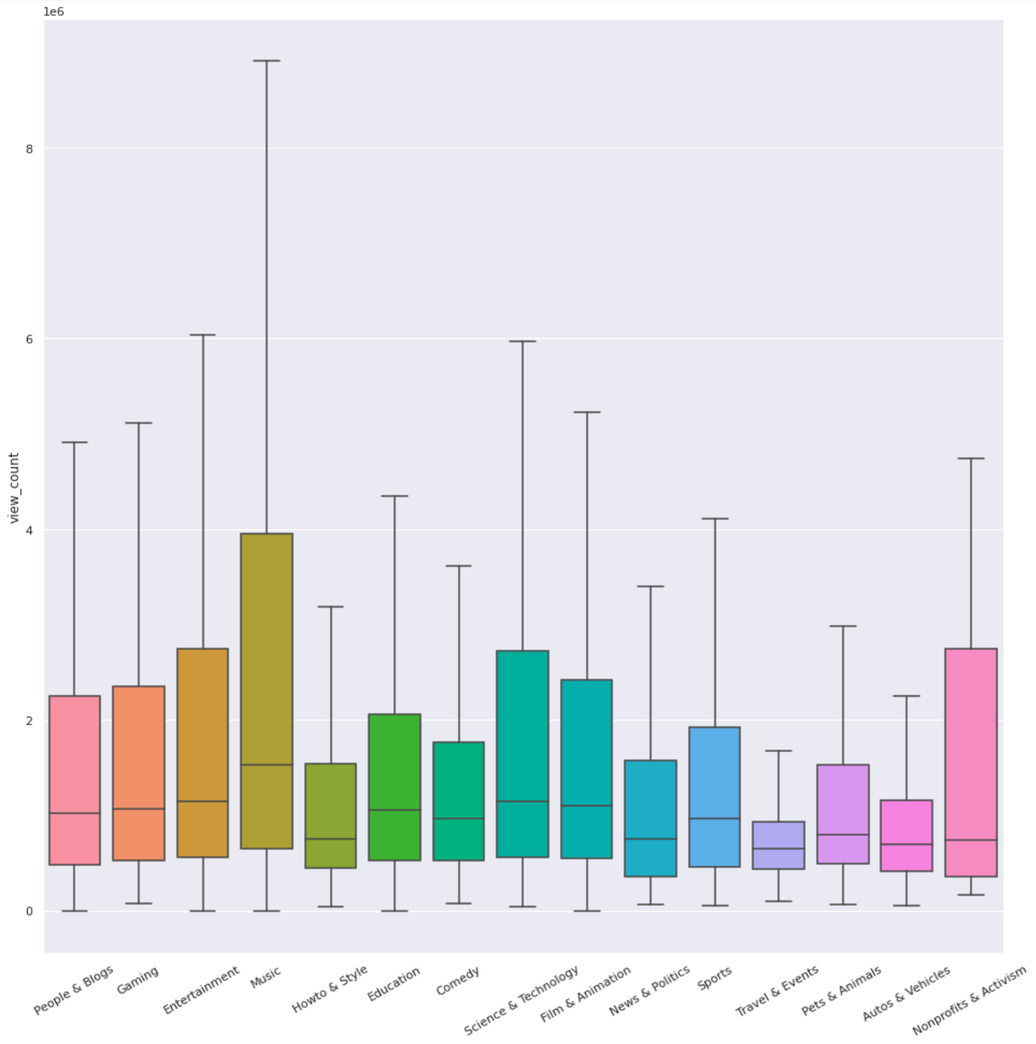
In India, entertainment and music are also very popular, and entertainment has become more and more popular in recent years, and the popularity of music has weakened



view_count with outliers and no outliers

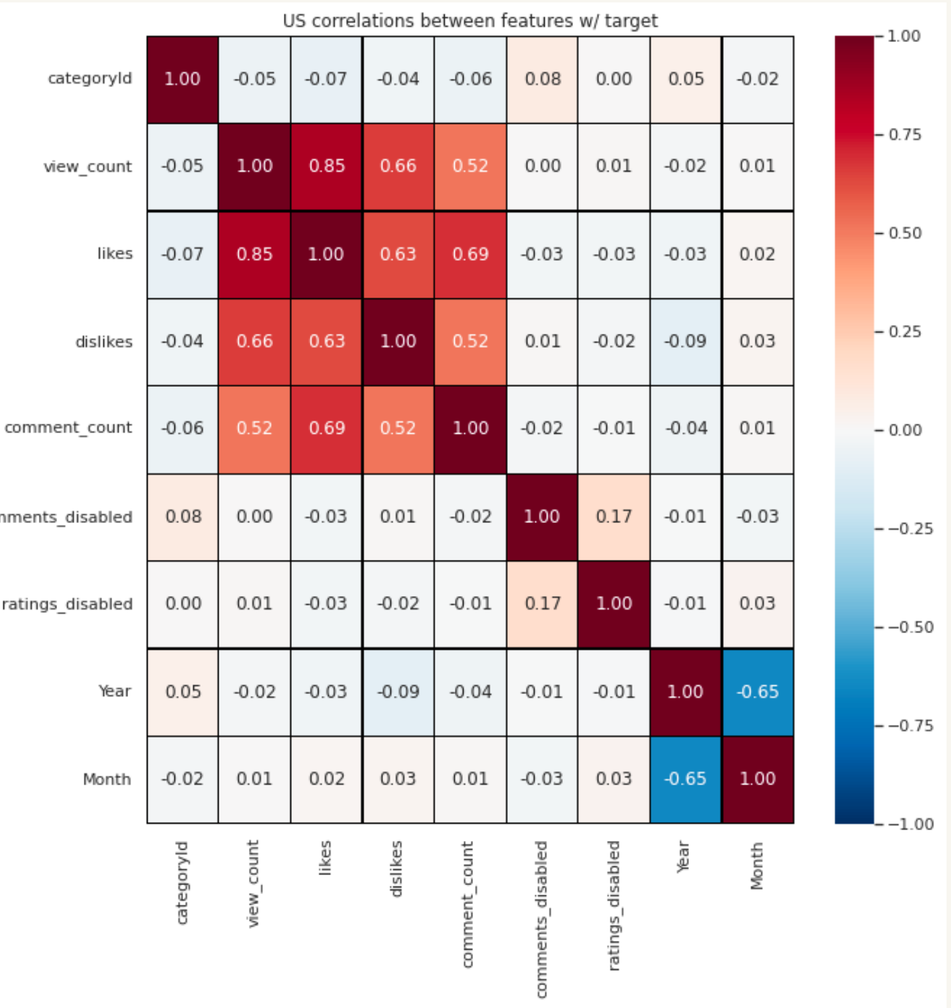


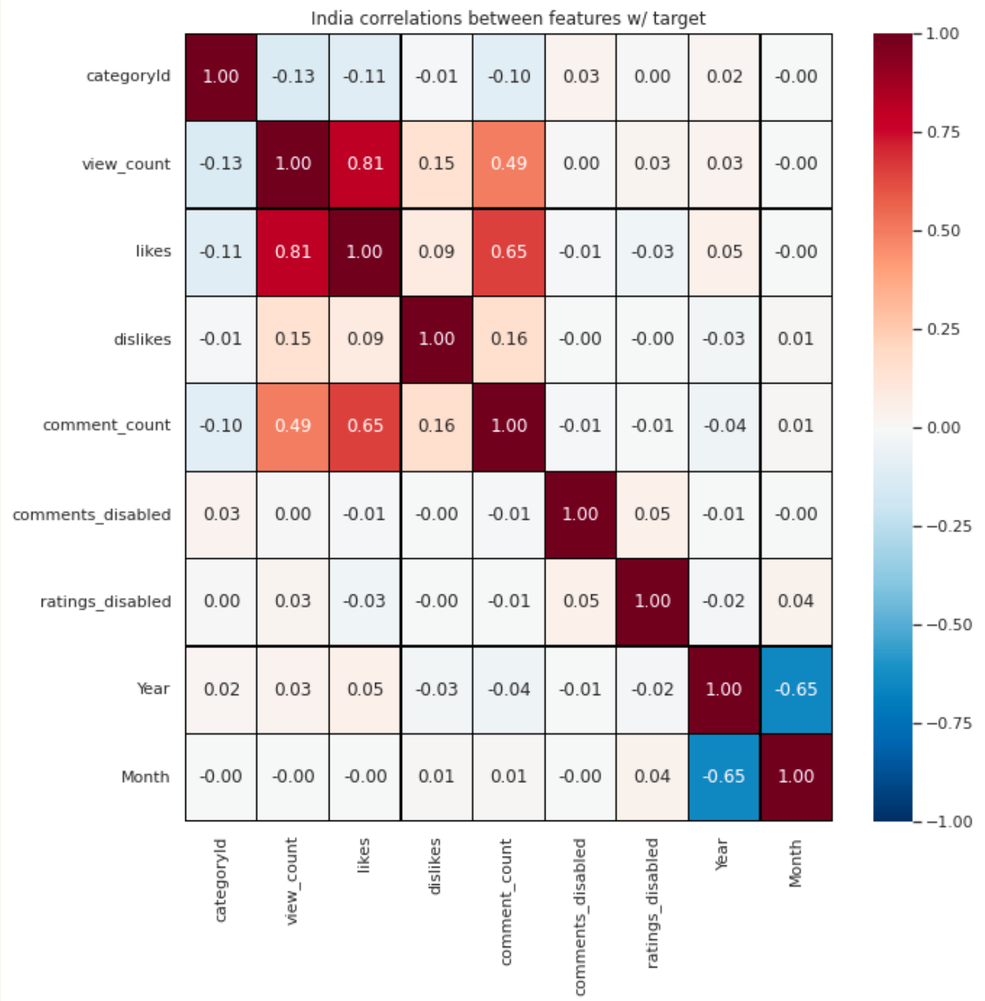




Correlation analysis between different columns

- 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

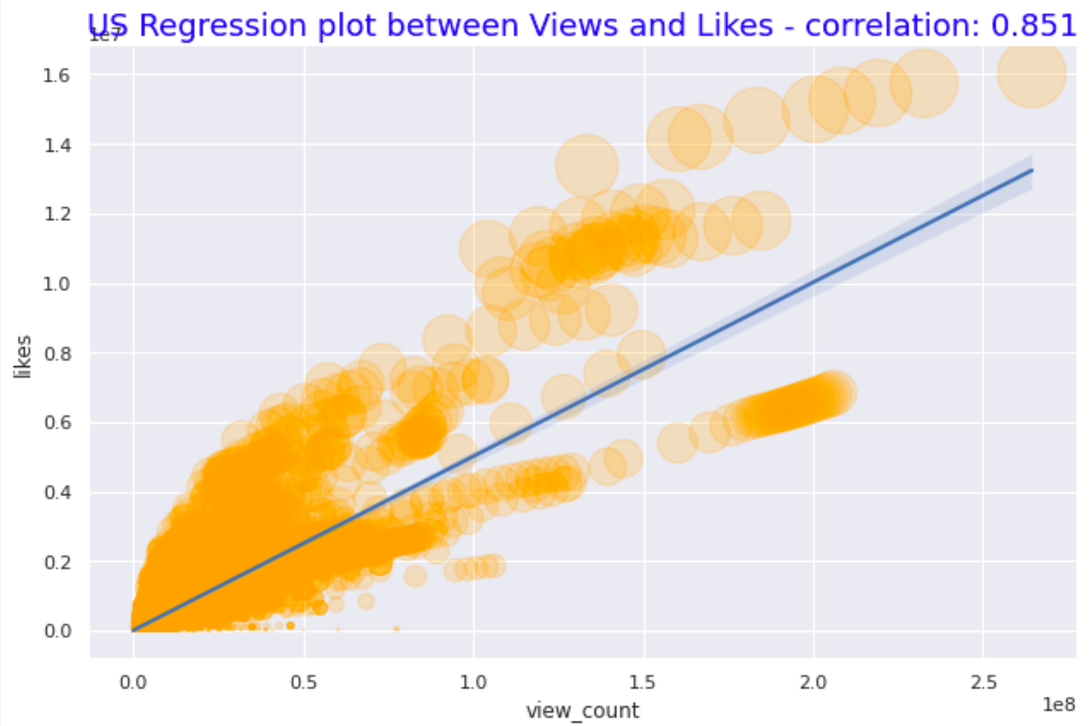


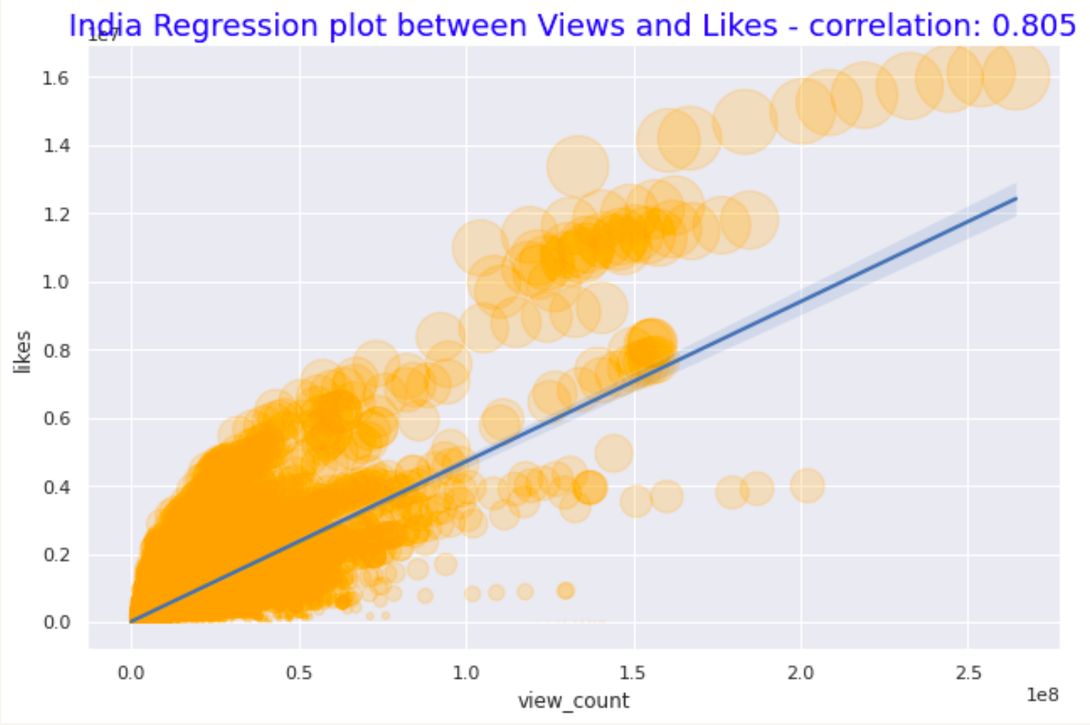


- In India, view_count and like also have a strong positive correlation, and comment_count also has a relatively high correlation with view_count and like
- It can be found that dislike has little effect on other factors in India, but the opposite is true in the United States

Like correlation in USA and India

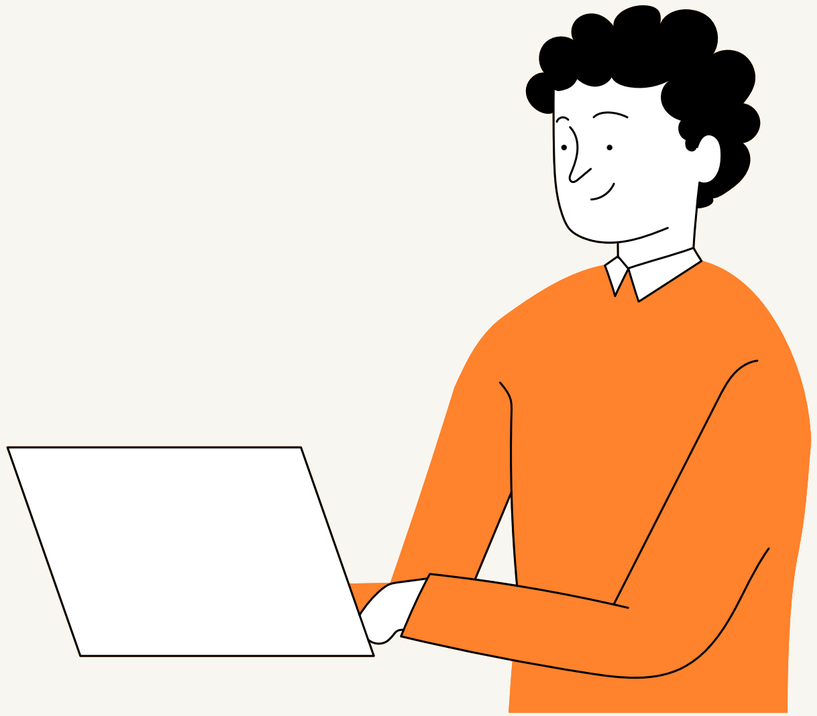
The number of views in the United States and India has a strong correlation with likes, both exceeding 0.8, which is a strong correlation

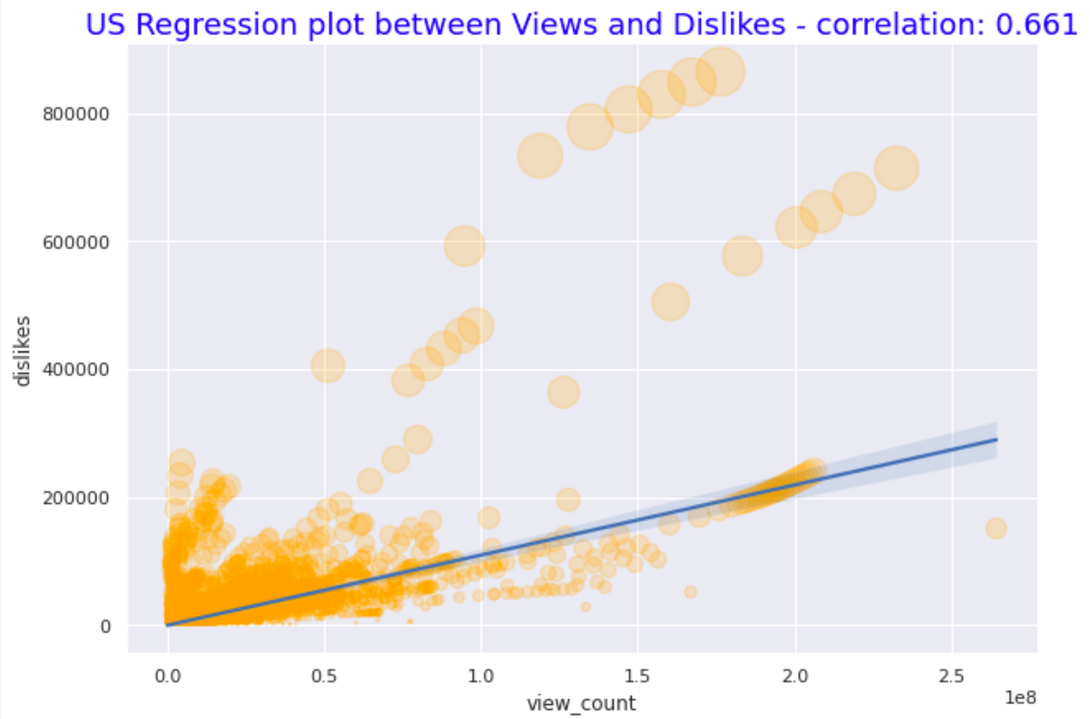


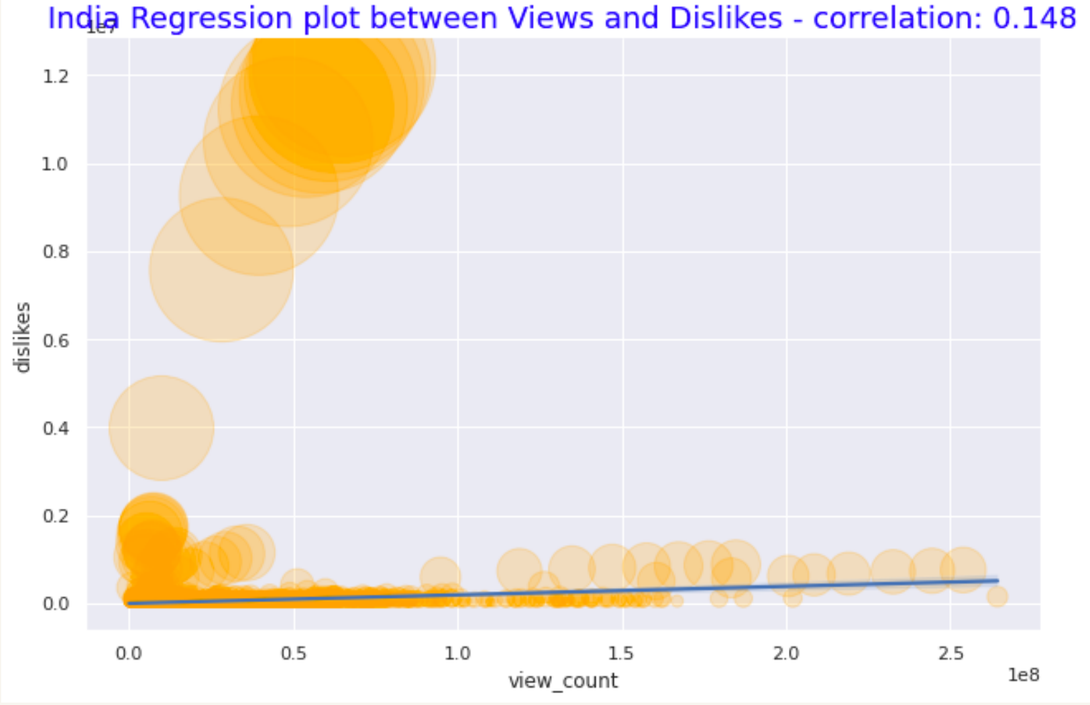


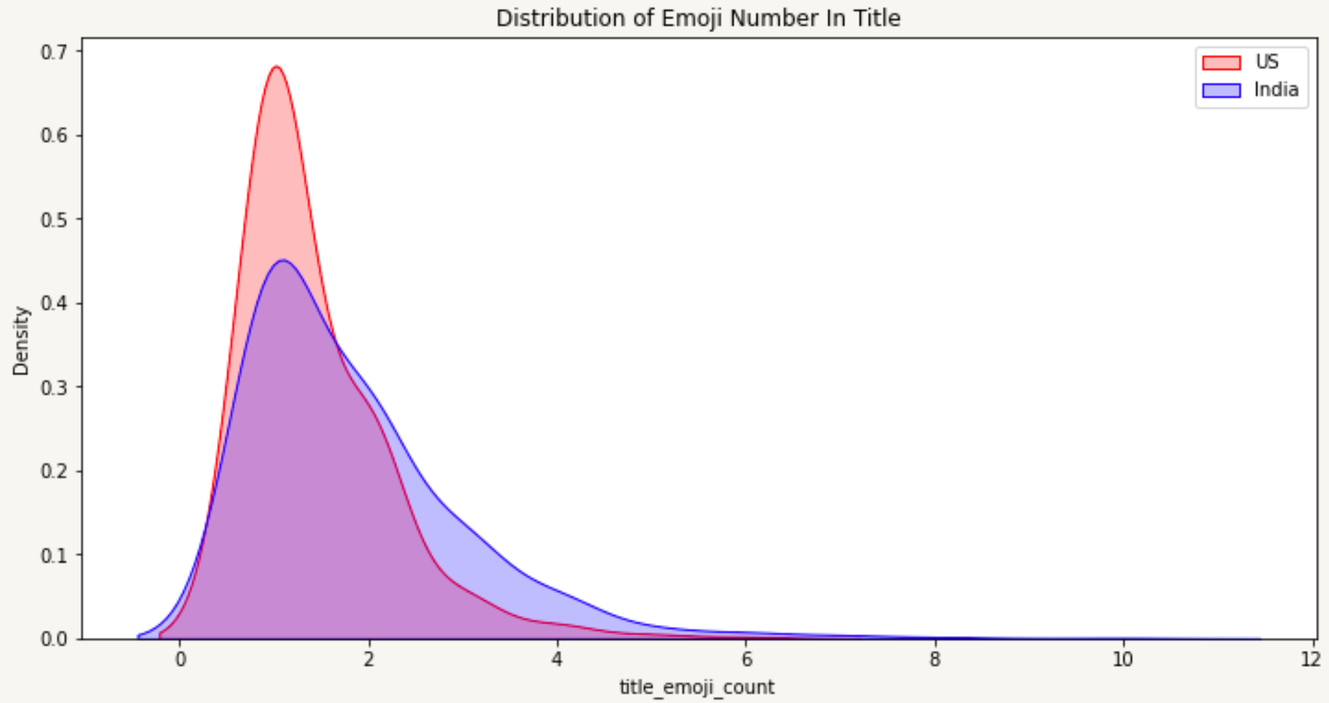
Dislike correlation in USA and India

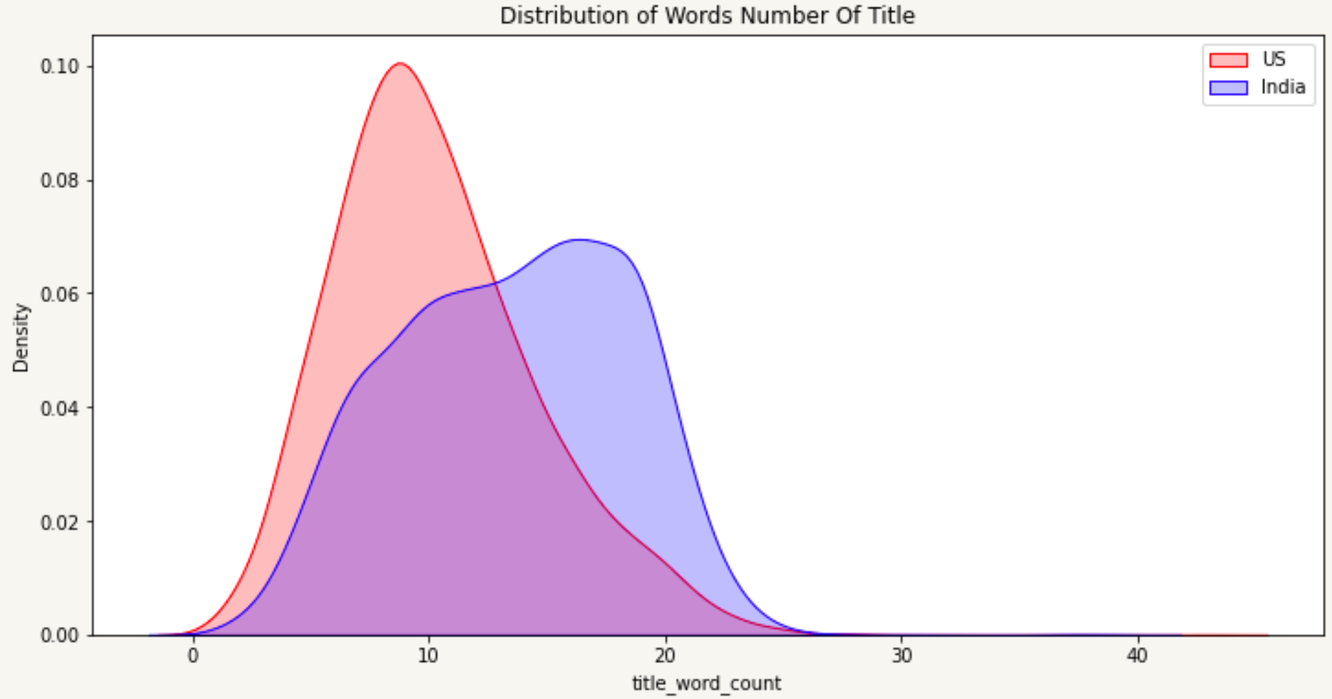
In linear regression, the correlation between US video views and dislike is 0.661; but India is only 0.148, and India has a high dislike dispersion in the chart











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

- 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

Conclusion

Analysis

The First Analysis of Obtained Results

- the two countries have a big difference in channel preferences; the difference in favorite categories is small.

The Second Analysis of Obtained Results

- In correlation research, likes are the most influential factor on the number of views, and the number of comments also has a certain impact on the number of views

The Third Analysis of Obtained Results

- There is a difference in the correlation between dislike and the number of views in the two countries

Thank you for listening!

Student Number	*Student Name*
21448159	*Wu Jingyi*
21408041	*JIANG Tianqi*
21434921	*He Jiaqi*
21441243	*DU He*
21457328	*Zhang Rui*