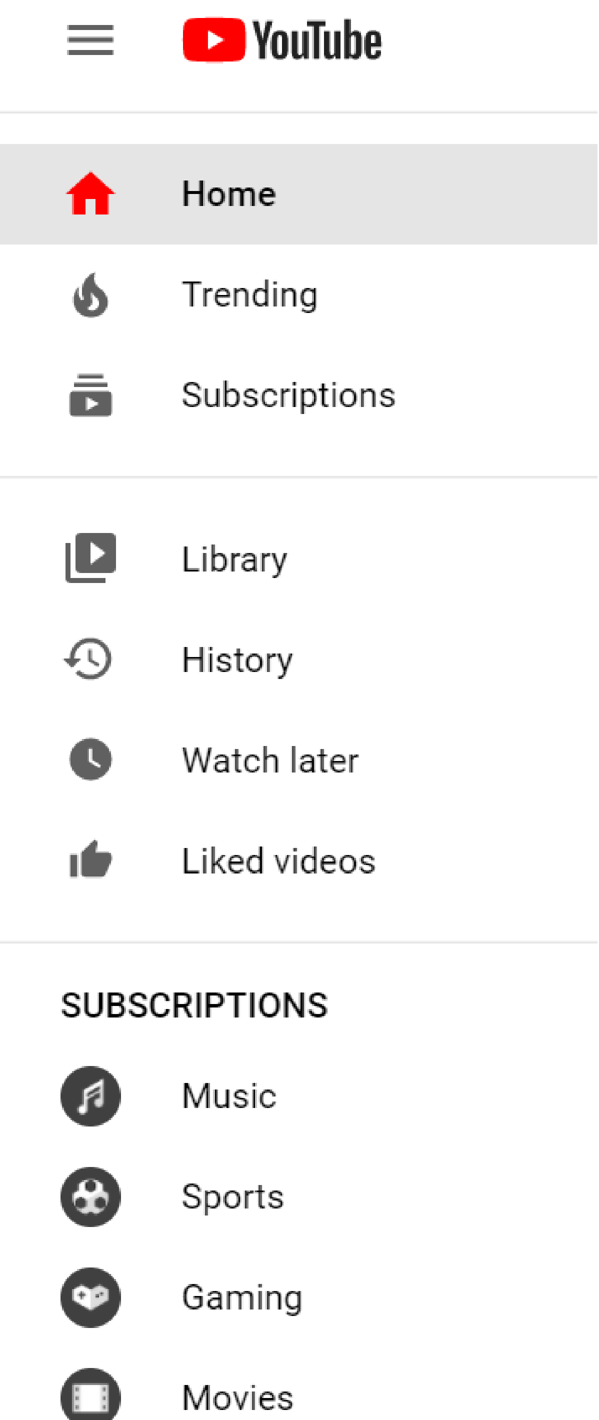
YouTube Trending Videos Analysis

AIDM7360 Big Data Management and Analytics

Group Name
Teletubbies
Leader
HUANG Zefei 20449496
Group Members
CAI Runlin 20426550
CHEN Xiaoqi 20465106
GUO Yuju 20465769

AGENDA

- **Introduction to the project**
- **The data**
- **Interface explanation**
- **News / template format**
- **Data Analysis and isualization**



Why we want to do this?
Where the dataset from?

Introduction

Why we want to do this?

The media landscape has changed radically since YouTube videos was launched. The amount and variety of content posted on the site was so great that it became the most popular video platform at one time. We analyzed YouTube videos, studied what people are looking at these days, and what are the hot spots.

Introduction

What questions we did for data management and visualization were as follows:

1. Which type of video is the most popular?

2. What is the popularity of each general category?(Template)

3. In today's flood of "entertainment traffic", is the publishing trend of educational videos declining?

4. The biggest hot spot in the United States recently is the "U.S. General Election."

Which channels have released relevant information?

Introduction

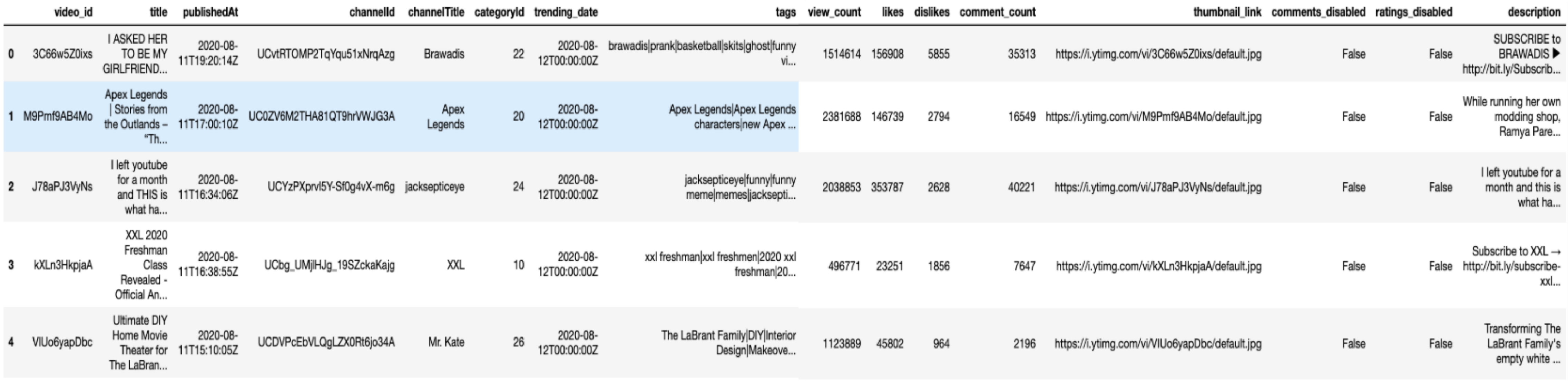
Where the data from?

https://www.kaggle.com/ammar111/youtube-trending-videos-analysis/data

We downloaded a data set of YouTube video trends from Kaggle that are relevant to the United States for data management and visualization, then we cleaned the data to get what we want.

Data Processing

Row data



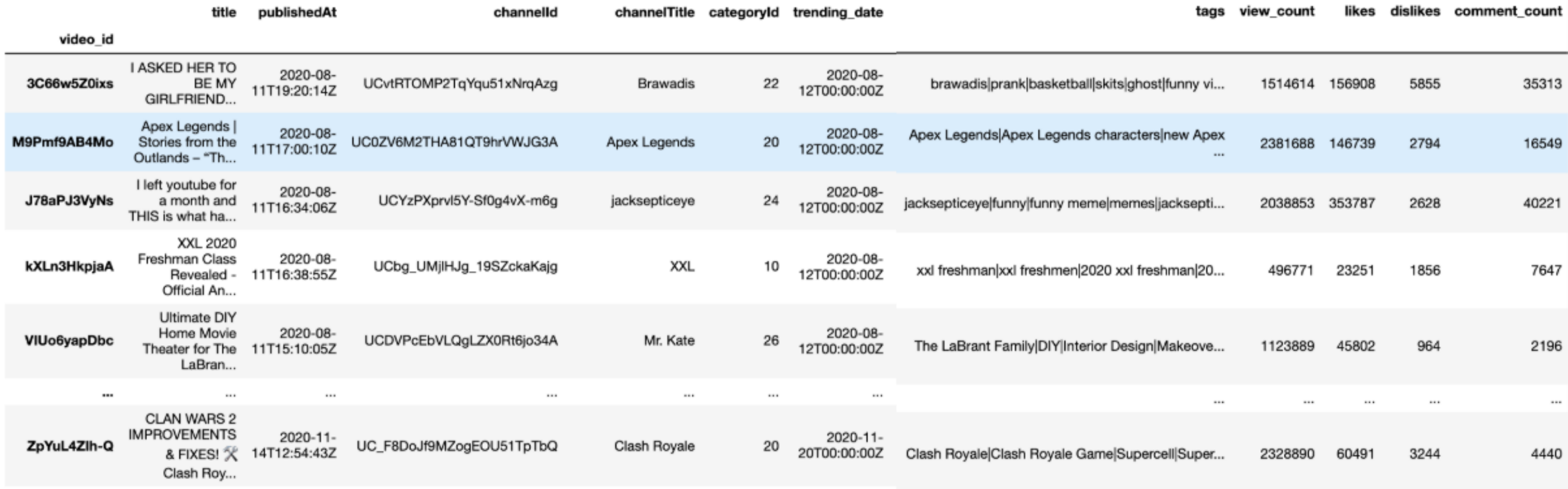
	video_id	title	publishedAt	channelId	channelTitle	categoryId	trending_date	tags	view_count	likes	dislikes	comment_count	thumbnail_link	comments_disabled	ratings_disabled	description
0	3C66w5Z0ixs	I ASKED HER TO BE MY GIRLFRIEND...	2020-08-11T19:20:14Z	UCvtRTOMP2TqYqu51xNrqAzg	Brawadis	22	2020-08-12T00:00:00Z	brawadis|prank|basketball|skits|ghost|funny vi...	1514614	156908	5855	35313	https://i.ytimg.com/vi/3C66w5Z0ixs/default.jpg	False	False	SUBSCRIBE to BRAWADIS ▶ http://bit.ly/Subscrib...
1	M9Pmf9AB4Mo	Apex Legends | Stories from the Outlands – “Th...	2020-08-11T17:00:10Z	UC0ZV6M2THA81QT9hrVWJG3A	Apex Legends	20	2020-08-12T00:00:00Z	Apex Legends|Apex Legends characters|new Apex ...	2381688	146739	2794	16549	https://i.ytimg.com/vi/M9Pmf9AB4Mo/default.jpg	False	False	While running her own modding shop, Ramya Pare...
2	J78aPJ3VyNs	I left youtube for a month and THIS is what ha...	2020-08-11T16:34:06Z	UCYzPXprvl5Y-Sf0g4vX-m6g	jacksepticeye	24	2020-08-12T00:00:00Z	jacksepticeye|funny|funny meme|memes|jacksepti...	2038853	353787	2628	40221	https://i.ytimg.com/vi/J78aPJ3VyNs/default.jpg	False	False	I left youtube for a month and this is what ha...
3	kXLn3HkpjaA	XXL 2020 Freshman Class Revealed - Official An...	2020-08-11T16:38:55Z	UCbg_UMjlHJg_19SZckaKajg	XXL	10	2020-08-12T00:00:00Z	xxl freshman|xxl freshmen|2020 xxl freshman|20...	496771	23251	1856	7647	https://i.ytimg.com/vi/kXLn3HkpjaA/default.jpg	False	False	Subscribe to XXL → http://bit.ly/subscribe-xxl...
4	VIUo6yapDbc	Ultimate DIY Home Movie Theater for The LaBran...	2020-08-11T15:10:05Z	UCDVPcEbVLQgLZX0Rt6jo34A	Mr. Kate	26	2020-08-12T00:00:00Z	The LaBrant Family|DIY|Interior Design|Makeove...	1123889	45802	964	2196	https://i.ytimg.com/vi/VIUo6yapDbc/default.jpg	False	False	Transforming The LaBrant Family's empty white ...

New data

Use the drop.() function to delete some data that is not important.

16 columns

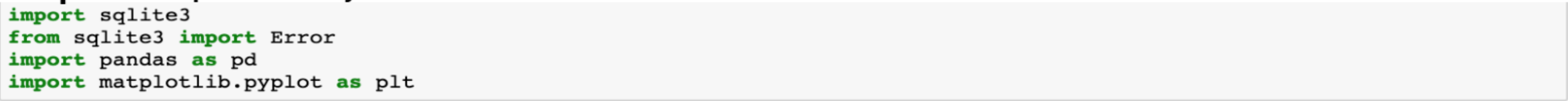
12 columns



video_id	title	publishedAt	channelId	channelTitle	categoryId	trending_date	tags	view_count	likes	dislikes	comment_count
3C66w5Z0ixs	I ASKED HER TO BE MY GIRLFRIEND...	2020-08-11T19:20:14Z	UCvtRTOMP2TqYqu51xNrqAzg	Brawadis	22	2020-08-12T00:00:00Z	brawadis|prank|basketball|skits|ghost|funny vi...	1514614	156908	5855	35313
M9Pmf9AB4Mo	Apex Legends | Stories from the Outlands – “Th...	2020-08-11T17:00:10Z	UC0ZV6M2THA81QT9hrVWJG3A	Apex Legends	20	2020-08-12T00:00:00Z	Apex Legends|Apex Legends characters|new Apex ...	2381688	146739	2794	16549
J78aPJ3VyNs	I left youtube for a month and THIS is what ha...	2020-08-11T16:34:06Z	UCYzPXprvl5Y-Sf0g4vX-m6g	jacksepticeye	24	2020-08-12T00:00:00Z	jacksepticeye|funny|funny meme|memes|jacksepti...	2038853	353787	2628	40221
kXLn3HkpjaA	XXL 2020 Freshman Class Revealed - Official An...	2020-08-11T16:38:55Z	UCbg_UMjlHJg_19SZckaKajg	XXL	10	2020-08-12T00:00:00Z	xxl freshman|xxl freshmen|2020 xxl freshman|20...	496771	23251	1856	7647
VIUo6yapDbc	Ultimate DIY Home Movie Theater for The LaBran...	2020-08-11T15:10:05Z	UCDVPcEbVLQgLZX0Rt6jo34A	Mr. Kate	26	2020-08-12T00:00:00Z	The LaBrant Family|DIY|Interior Design|Makeove...	1123889	45802	964	2196
...	...	...	...	...	...	...	...	...	...	...	...
ZpYuL4Zlh-Q	CLAN WARS 2 IMPROVEMENTS & FIXES! 🛠 Clash Roy...	2020-11-14T12:54:43Z	UC_F8DoJf9MZogEOU51TpTbQ	Clash Royale	20	2020-11-20T00:00:00Z	Clash Royale|Clash Royale Game|Supercell|Super...	2328890	60491	3244	4440

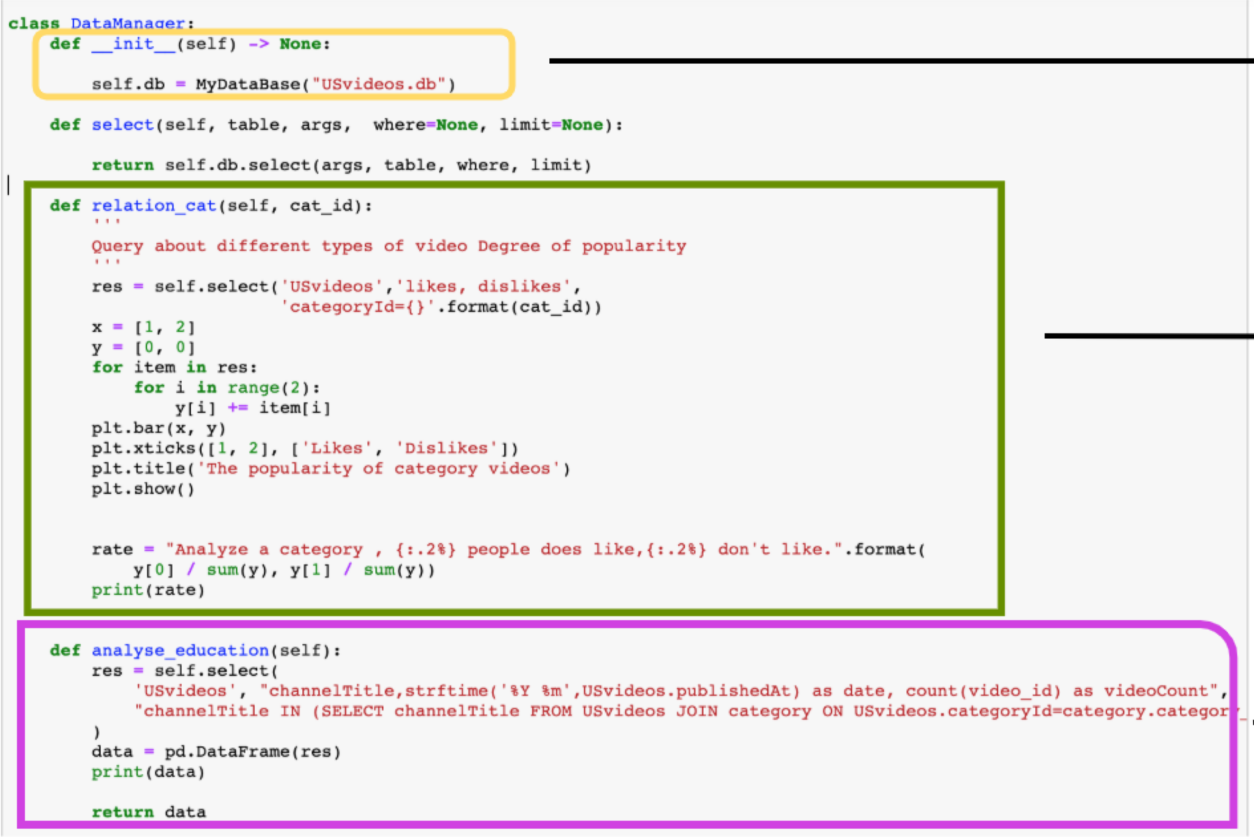
Interface Explanation

Step1 : Import library



```
import sqlite3
from sqlite3 import Error
import pandas as pd
import matplotlib.pyplot as plt
```

Step2 : Installs all the automated functions and databases in one class



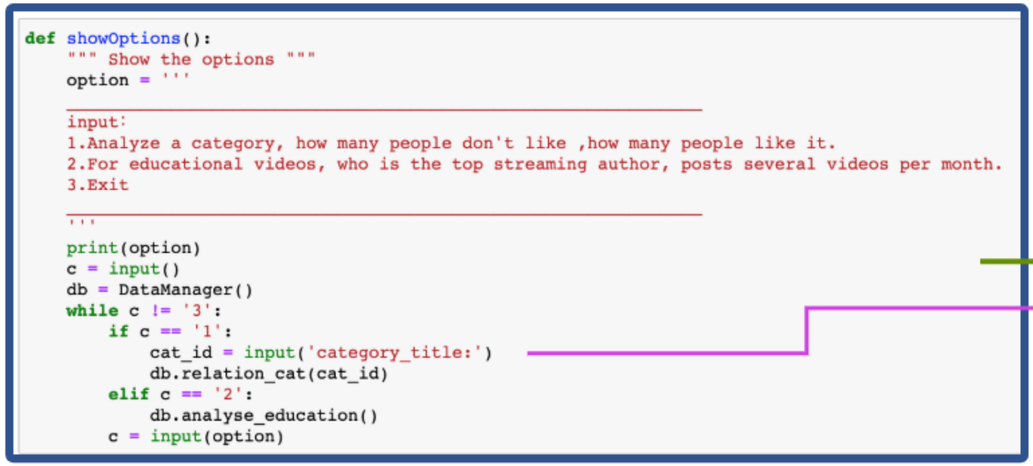
```
class DataManager:
    def __init__(self) -> None:

        self.db = MyDataBase("USvideos.db")

    def select(self, table, args,  where=None, limit=None):

        return self.db.select(args, table, where, limit)

    def relation_cat(self, cat_id):
        '''
        Query about different types of video Degree of popularity
        '''
        res = self.select('USvideos','likes, dislikes',
                          'categoryId={}'.format(cat_id))
        x = [1, 2]
        y = [0, 0]
        for item in res:
            for i in range(2):
                y[i] += item[i]
        plt.bar(x, y)
        plt.xticks([1, 2], ['Likes', 'Dislikes'])
        plt.title('The popularity of category videos')
        plt.show()


        rate = "Analyze a category , {:.2%} people does like,{:.2%} don't like.".format(
            y[0] / sum(y), y[1] / sum(y))
        print(rate)

    def analyse_education(self):
        res = self.select(
            'USvideos', "channelTitle,strftime('%Y %m',USvideos.publishedAt) as date, count(video_id) as videoCount",
            "channelTitle IN (SELECT channelTitle FROM USvideos JOIN category ON USvideos.categoryId=category.categor
        )
        data = pd.DataFrame(res)
        print(data)

        return data
```

Initializes and connects to the database

Template data replacement function and visualization

Subquery function, I use "Pandas ' to generate the DataFrame.

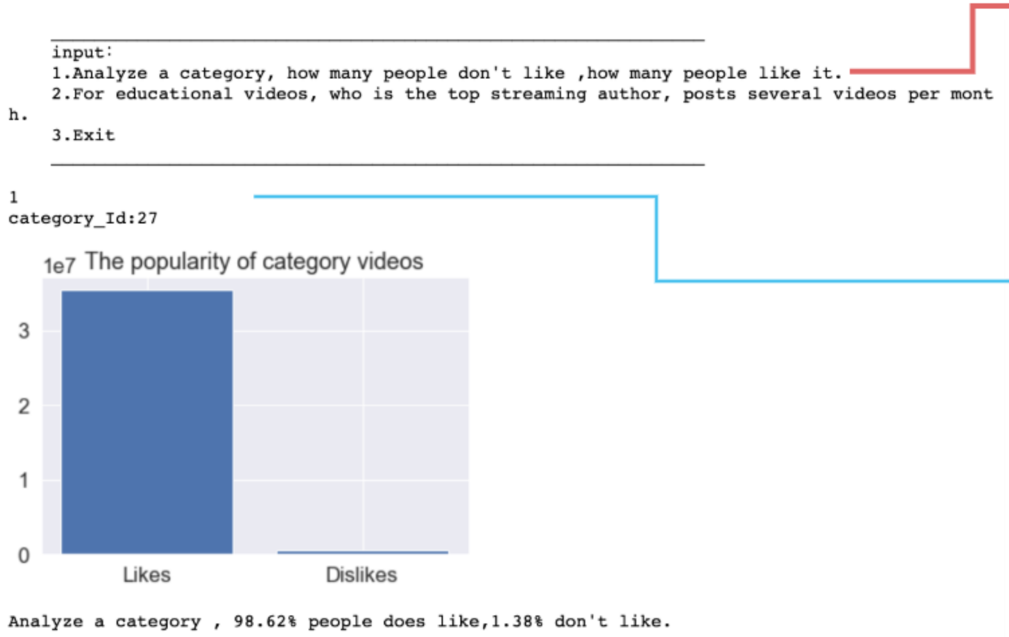
Interface Explanation



```
def showOptions():
    """ Show the options """
    option = '''
    ____________________________________________________________
    input:
    1.Analyze a category, how many people don't like ,how many people like it.
    2.For educational videos, who is the top streaming author, posts several videos per month.
    3.Exit
    ____________________________________________________________
    '''
    print(option)
    c = input()
    db = DataManager()
    while c != '3':
        if c == '1':
            cat_id = input('category_title:')
            db.relation_cat(cat_id)
        elif c == '2':
            db.analyse_education()
        c = input(option)
```

- This is an interactive system that faces object Settings.

- Users can extract the question they want from the input.



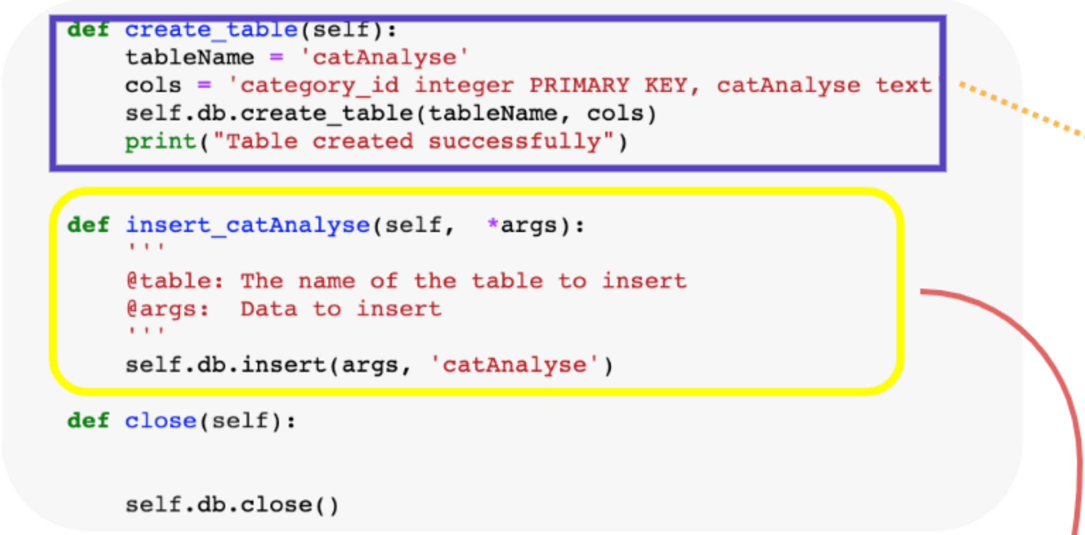
```
    ____________________________________________________________
    input:
    1.Analyze a category, how many people don't like ,how many people like it.
    2.For educational videos, who is the top streaming author, posts several videos per mont
h.
    3.Exit
    ____________________________________________________________

1
category_Id:27
```

```
Analyze a category , 98.62% people does like,1.38% don't like.
```

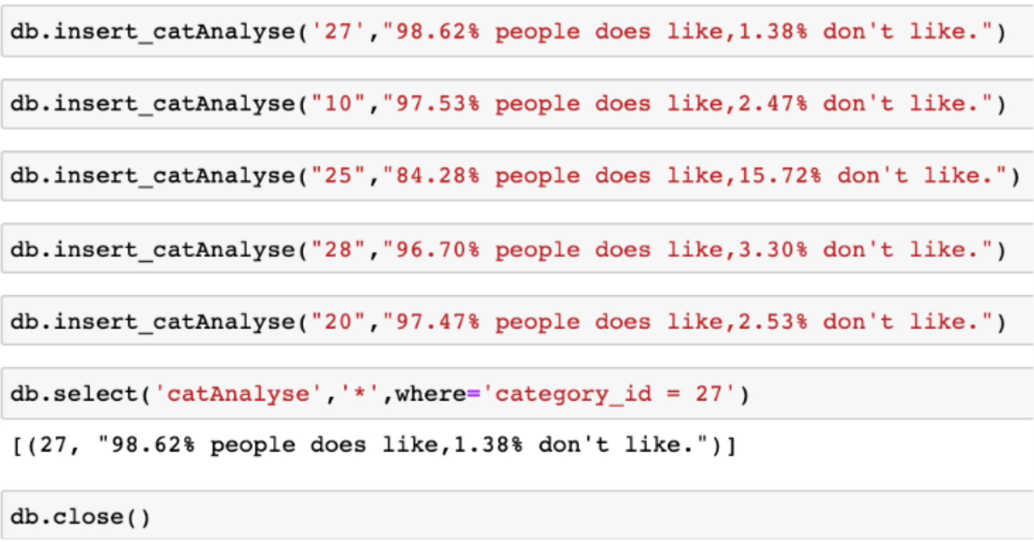
- For example, you want to know how many people like and dislike each category.

- You can select the first question, and then enter the id of the category.
You can present a visual data graph about popularity.

Interface Explanation



```
def create_table(self):
    tableName = 'catAnalyse'
    cols = 'category_id integer PRIMARY KEY, catAnalyse text'
    self.db.create_table(tableName, cols)
    print("Table created successfully")

def insert_catAnalyse(self,  *args):
    '''
    @table: The name of the table to insert
    @args:  Data to insert
    '''
    self.db.insert(args, 'catAnalyse')

def close(self):


    self.db.close()
```



```
db.insert_catAnalyse('27',"98.62% people does like,1.38% don't like.")
```

```
db.insert_catAnalyse("10","97.53% people does like,2.47% don't like.")
```

```
db.insert_catAnalyse("25","84.28% people does like,15.72% don't like.")
```

```
db.insert_catAnalyse("28","96.70% people does like,3.30% don't like.")
```

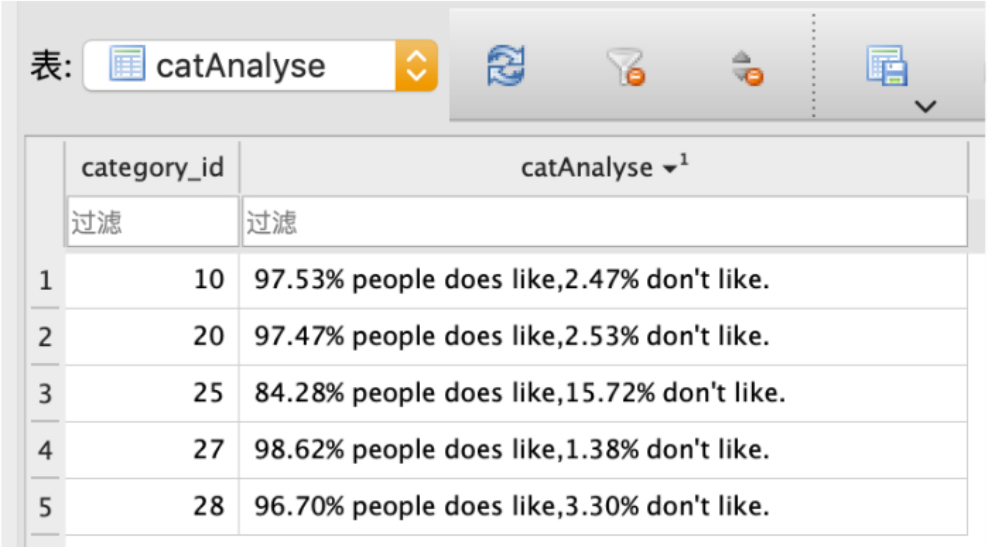
```
db.insert_catAnalyse("20","97.47% people does like,2.53% don't like.")
```

```
db.select('catAnalyse','*',where='category_id = 27')
```

```
[(27, "98.62% people does like,1.38% don't like.")]
```

```
db.close()
```

Beautiful



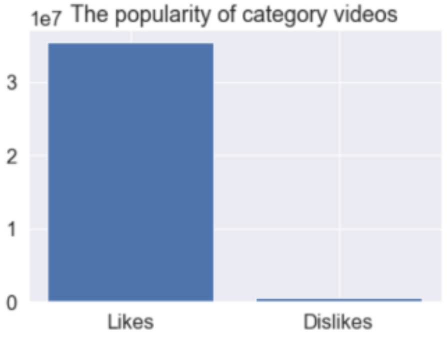
	category_id	catAnalyse
	过滤	过滤
1	10	97.53% people does like,2.47% don't like.
2	20	97.47% people does like,2.53% don't like.
3	25	84.28% people does like,15.72% don't like.
4	27	98.62% people does like,1.38% don't like.
5	28	96.70% people does like,3.30% don't like.

- Put the ability to create tables and insert data into aggregate data.
- Load the queried data into it.

News/Template Format

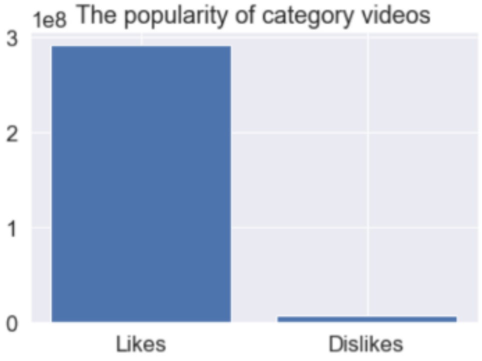
- The popularity of Education, {how many} like and {how many} don't like.
- The popularity of Gaming, {how many} like and {how many} don't like.
- The popularity of Music, {how many} like and {how many} don't like.
- The popularity of News&Politics, {how many} like and {how many} don't like.
- The number of views and comments of Music, {how many} view and {how many} comment.

```
1
category_Id:27
```



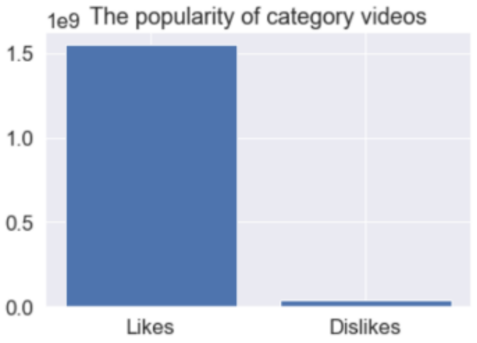
```
Analyze a category , 98.62% people does like,1.38% don't like
```

```
category_Id:20
```



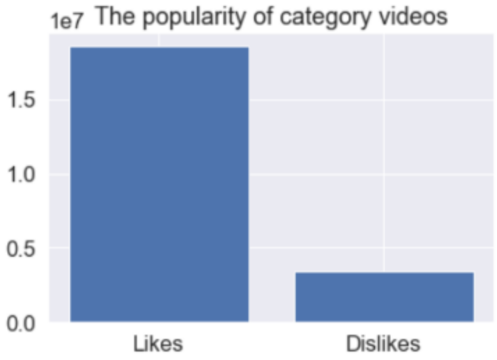
```
Analyze a category , 97.47% people does like,2.53% don't like.
```

```
category_Id:10
```



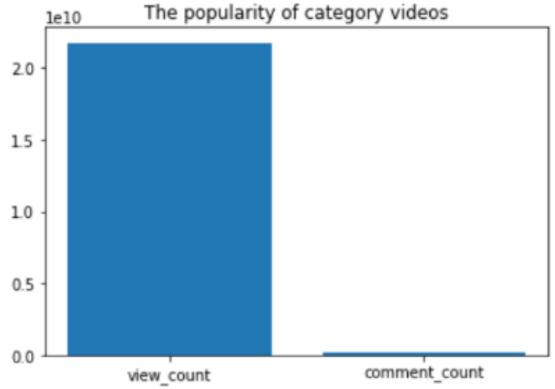
```
Analyze a category , 97.53% people does like,2.47% don't like.
```

```
category_Id:25
```



```
Analyze a category , 84.28% people does like,15.72% don't like.
```

```
category_Id:10
```

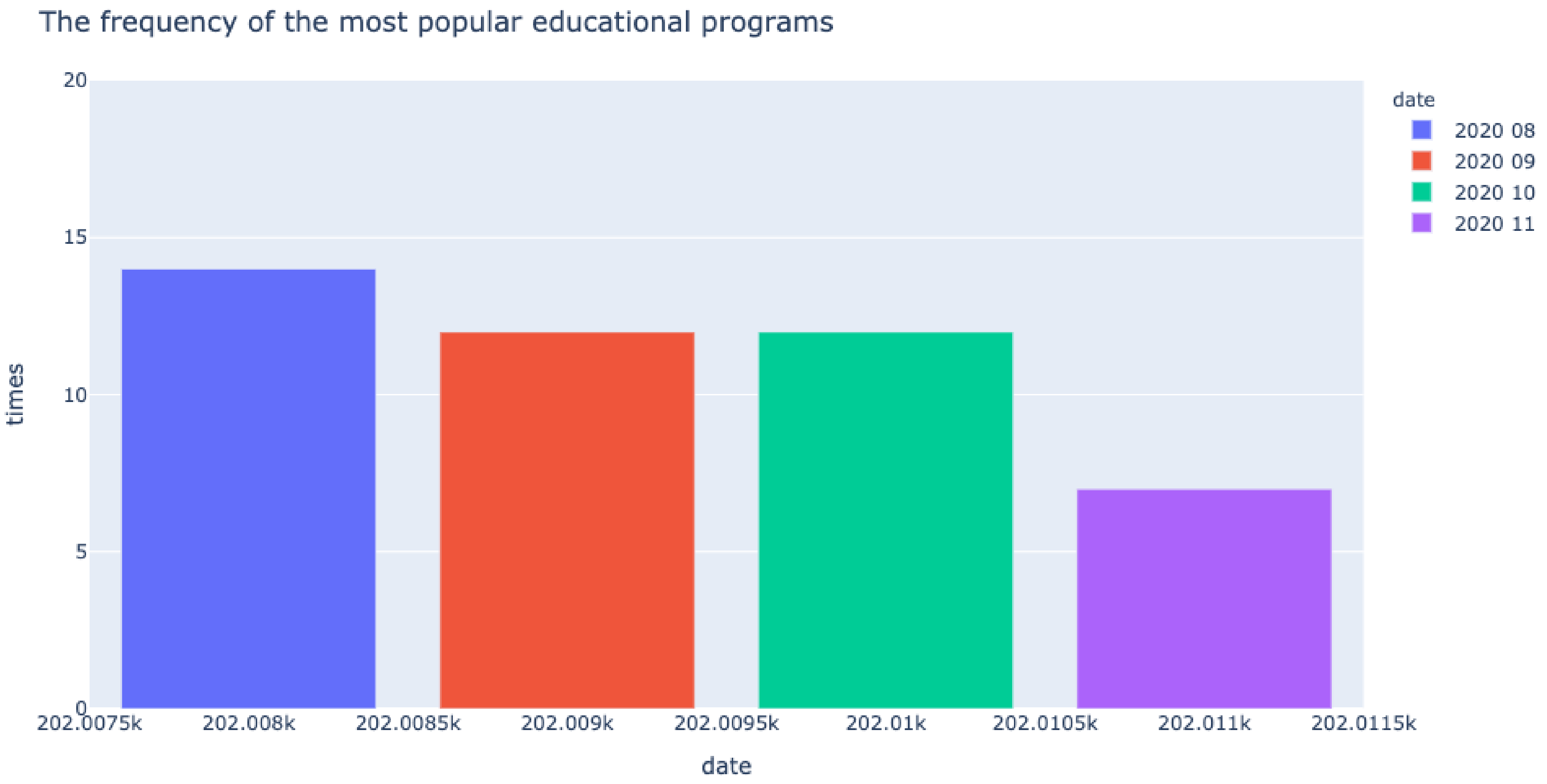


```
The category has 99.04% views ,0.96% comment
```

Data Analysis and Visualization

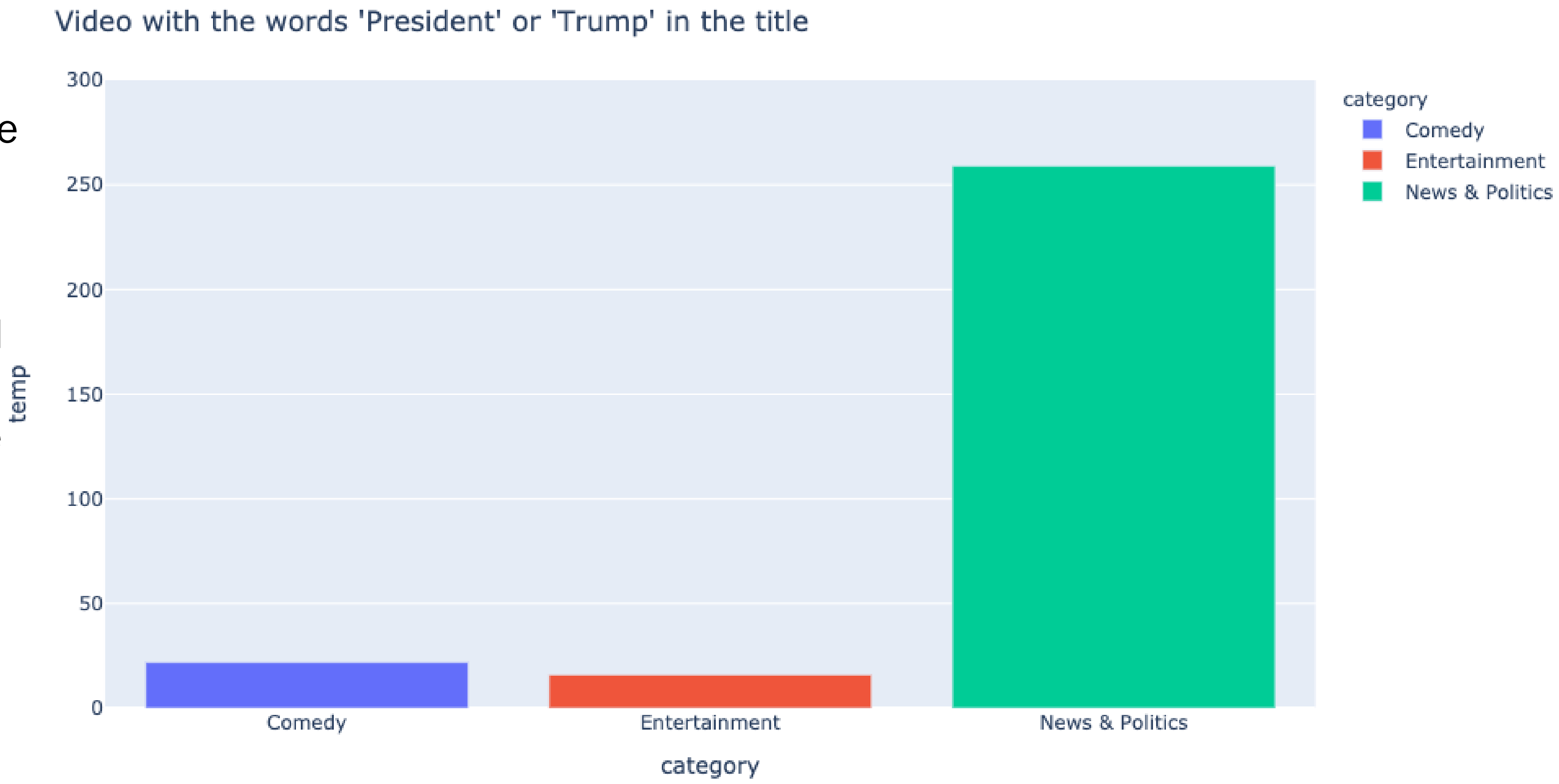
- First, we explored the frequency of the most popular educational programs. As shown in the chart, in August 2020, people visit the educational programs most.

- From August to September, the number of visits showed a downward trend. The reason may be that August are the time for school day, so this is a good time to create or watch some educational programs to do educational theme training.



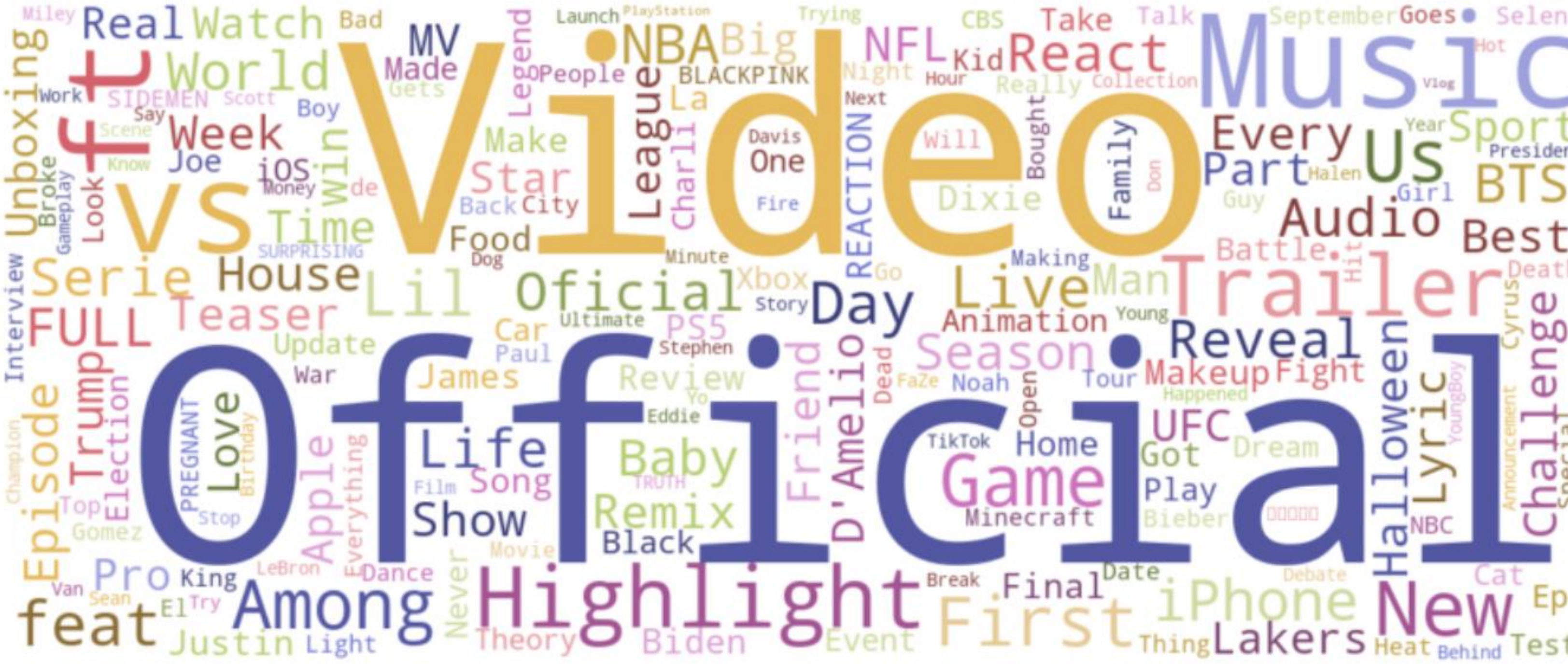
Data Analysis and Visualization

- Through the chart, we can find that videos with titles containing "President" or "Trump" appear the most in News & Politics, with more than 250 temps. Comedy and Entertainment have similar frequency, around 20temps. and 15temps respectively. On YouTube, users also like to make some comedies or funny videos for entertainment. And people don’t take politics so seriously and can use the president as a topic of entertainment.



Data Analysis and Visualization

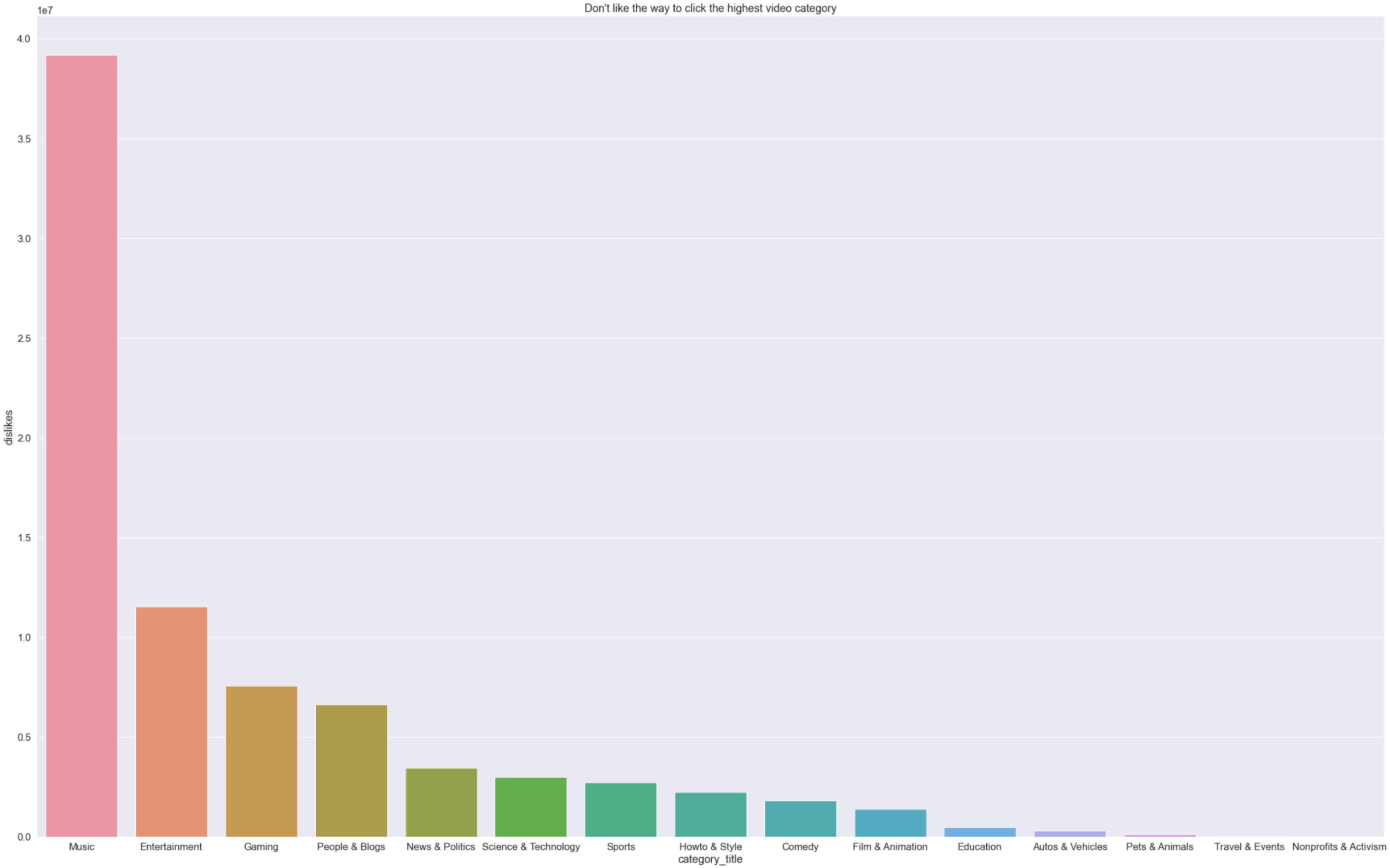
- We also use the titles of popular videos to create a word cloud. The results show that people are most interested in official, music and video. Because as the largest media platform, YouTube is a way to get the latest information from official media. On the other hand, due to the presentation of YouTube, people prefer to watch videos and music. People will also pay attention to some common interests, such as the NBA, BTS.



Data Analysis and Visualization

- We counted the types of videos with the highest number of clicks on "Dislike".

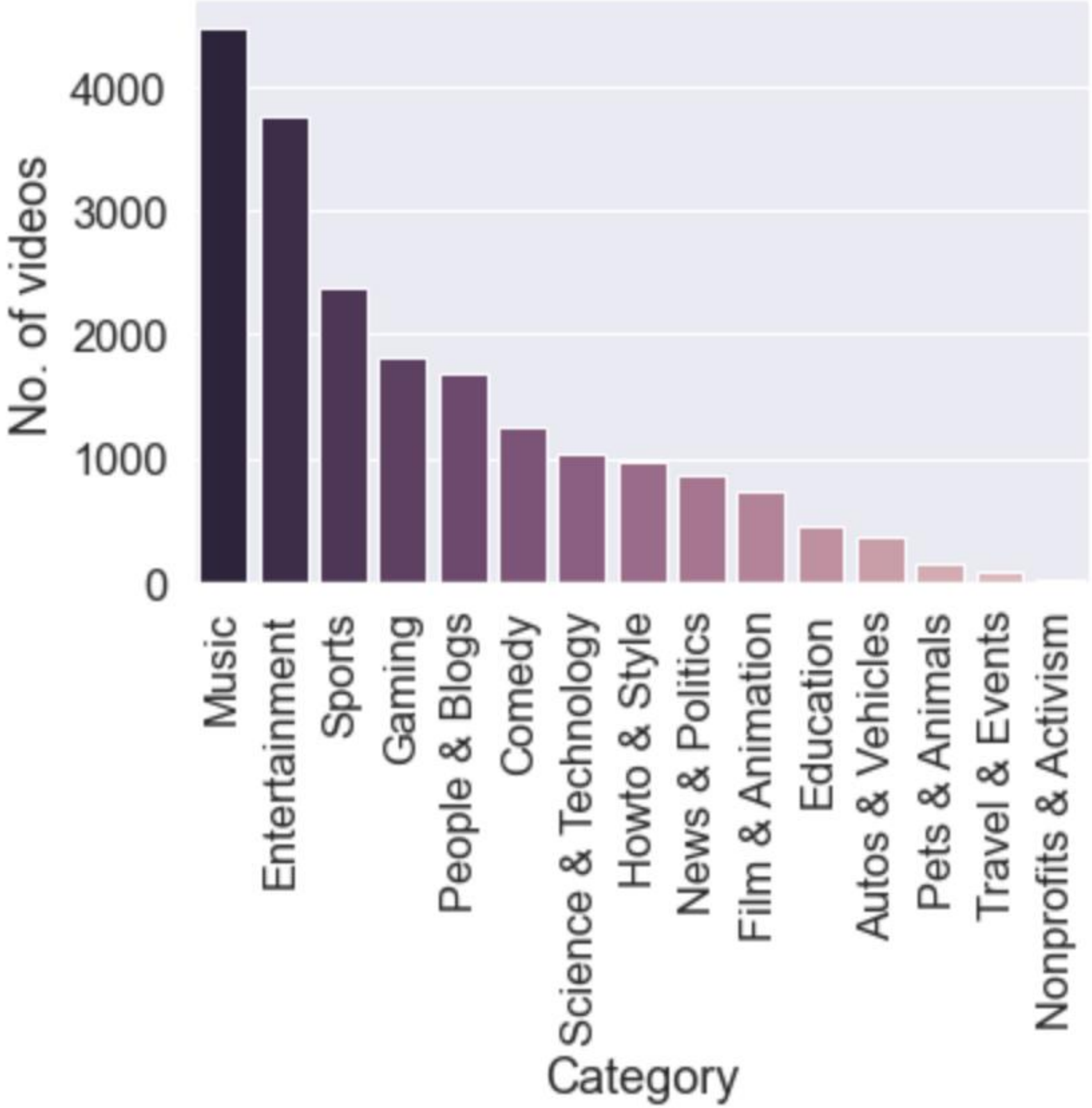
  Music category is the highest.



Data Analysis and Visualization

The most popular category of data visualization

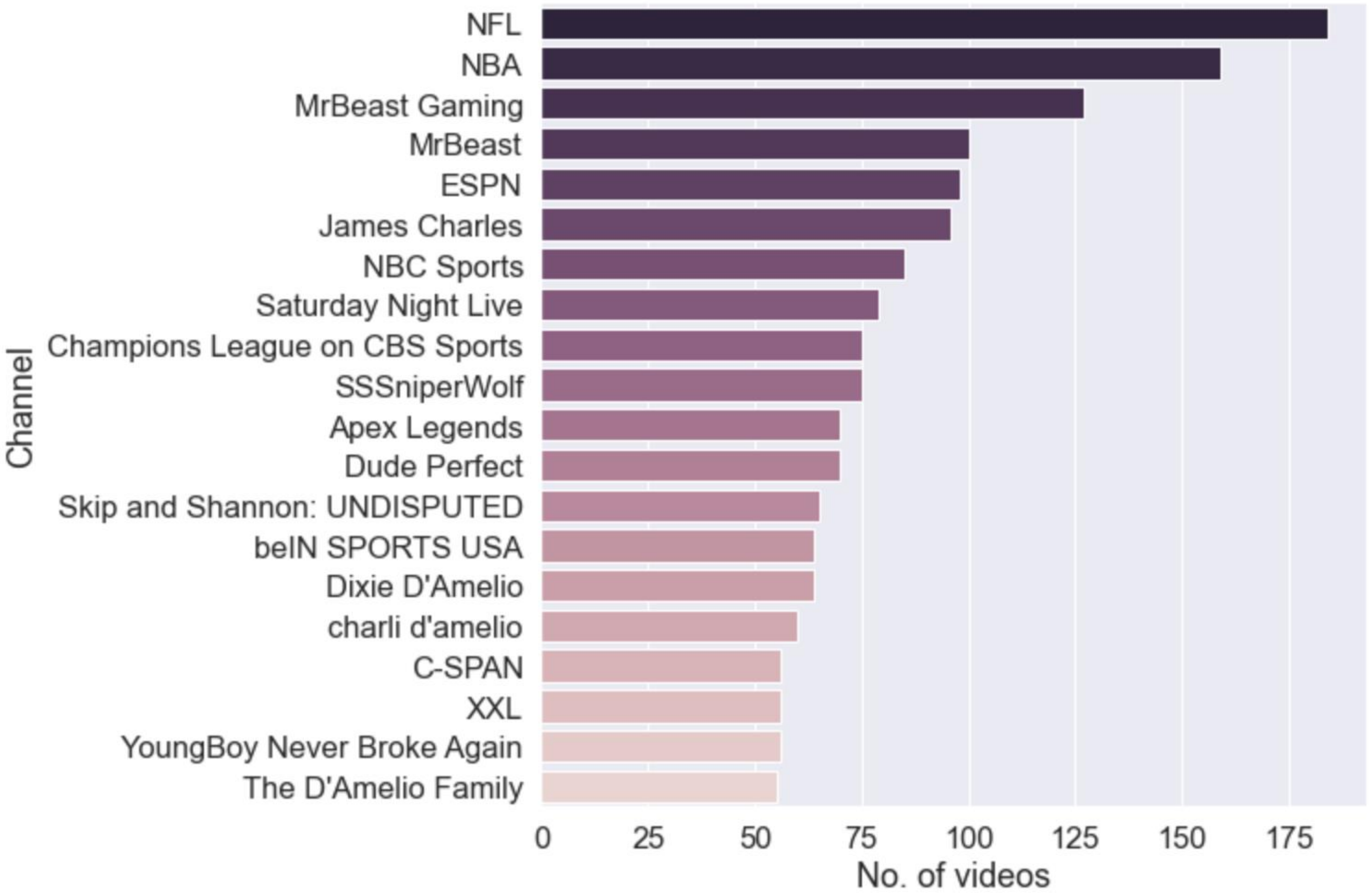
The picture shows that the most famous category is Music and Entertainment. It is easy to understand because Youtube is a way to post music videos and some funny videos. The third is sports, because of some official sports accounts(NFL(National Football League), and NBA). Gaming is also hot due to its addictive features.



Data Analysis and Visualization

Which is the most popular channel?

The most popular channel is NFL(National Football League), and next one is NBA and MrBeast Gaming(a game Youtuber)
In the picture we can see that the top of the YouTubers has two kinds, first is some official account and another kind is some famous YouTuber in different categories like games or music.

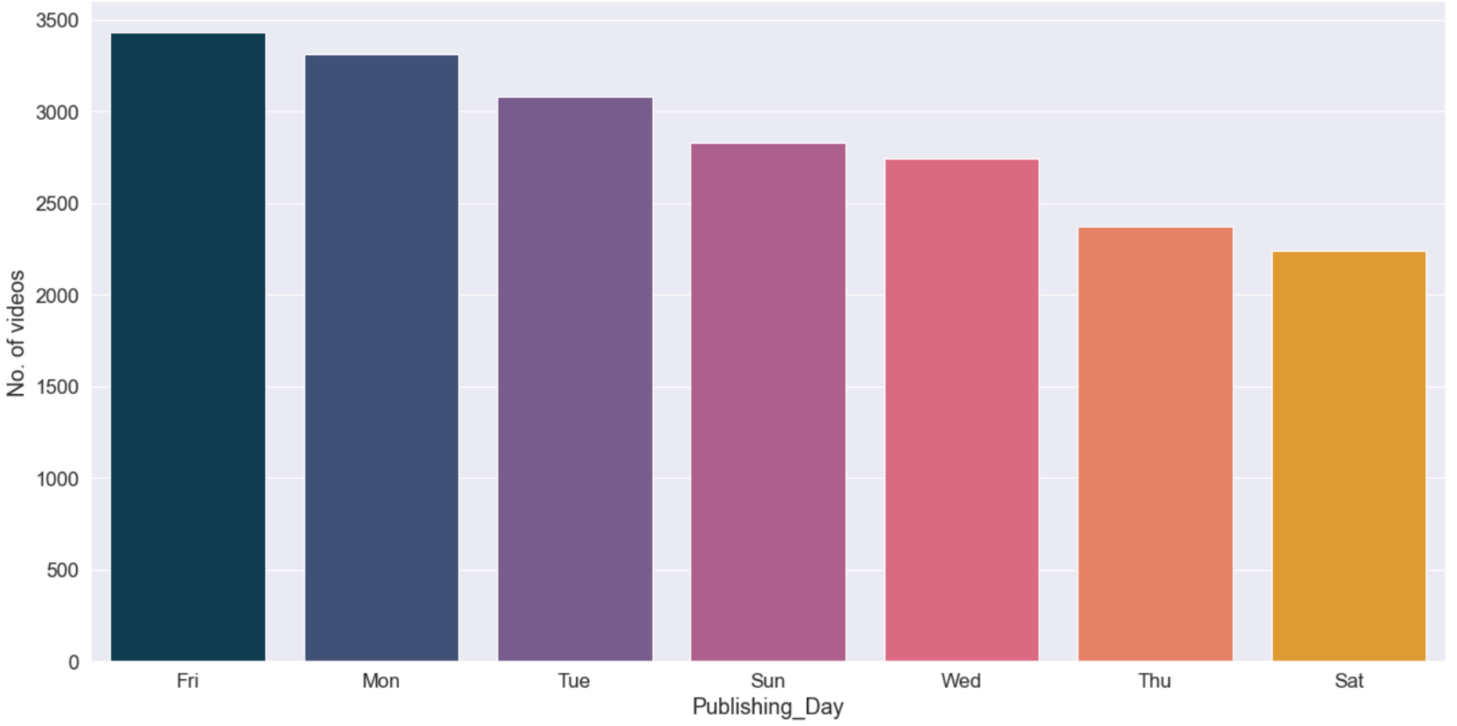


Data Analysis and Visualization

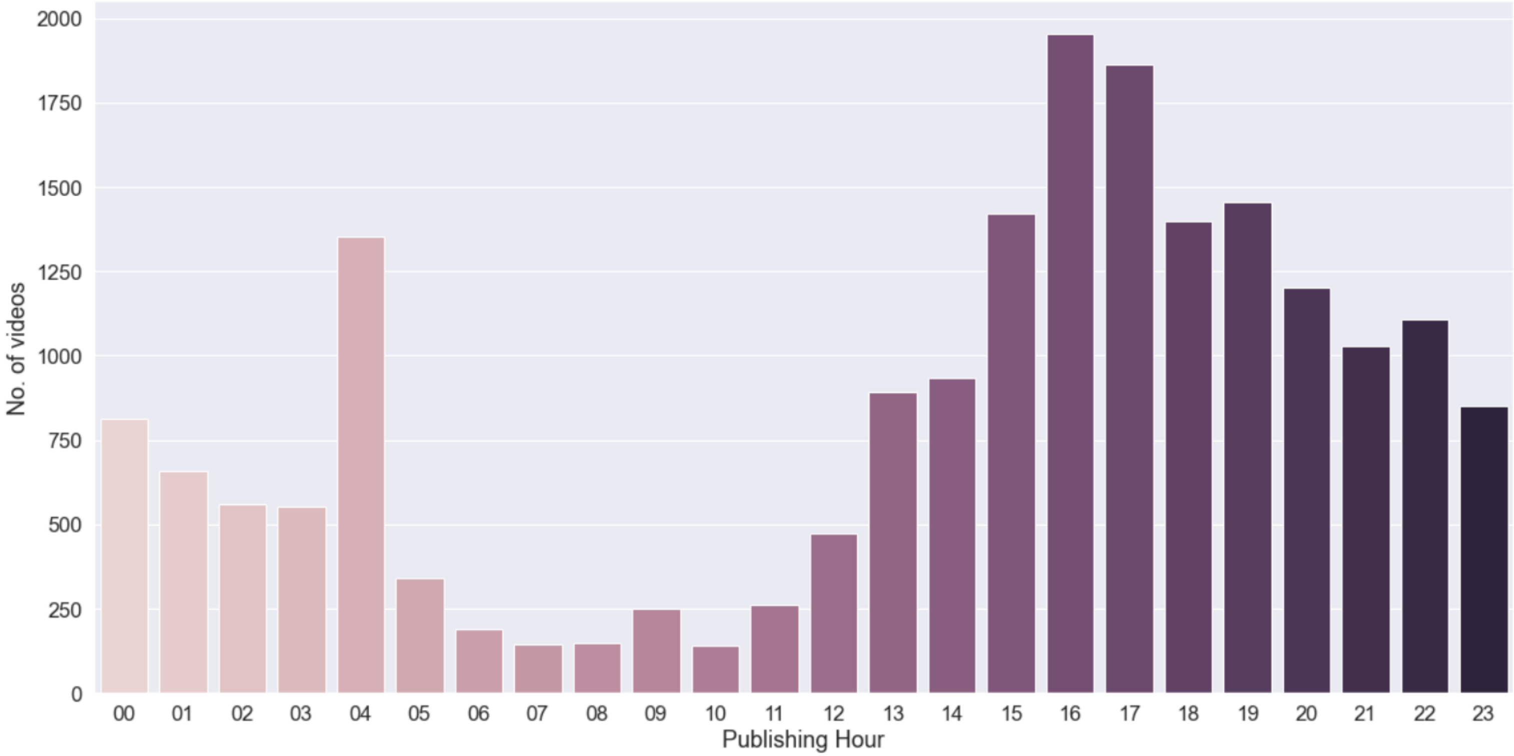
Audience popularity status of different types of videos.

Trending videos and their publishing time

As we can see, the number of top videos posted on Sunday and Saturday was significantly lower than the number posted on other days of the week.



Data Analysis and Visualization



Which channels have the largest number of trending videos?

Now let's use publishing hour column to see which publishing hours had the largest number of trending videos, as the picture’s, people watch YouTube videos post from 16:00-17:00 most, because at that time ,people finish work or class and they are free to watch videos. And the special phenomenon is videos post at 4.00,and We guess that the videos released during that time period are generally breaking news or information, like some new products are published.

Conclusion

People's attention on YouTube is very extensive, including many hot topics. More people show their interests on getting official information and the latest news, and they also watch some videos about their hobbies or listen to some music. Music is the hottest part of YouTube, but the 'dislike' of the music category is also the highest. Maybe it caused by people have multiple experiences and comments on different types of music or music video production styles.

On YouTube, users also like to make some comedies or funny videos for entertainment. And people don’t take politics so seriously and can use the president as a topic of entertainment.

THANKS

Group Project - Teletubbies

AIDM7360 Big Data Management and Analytics