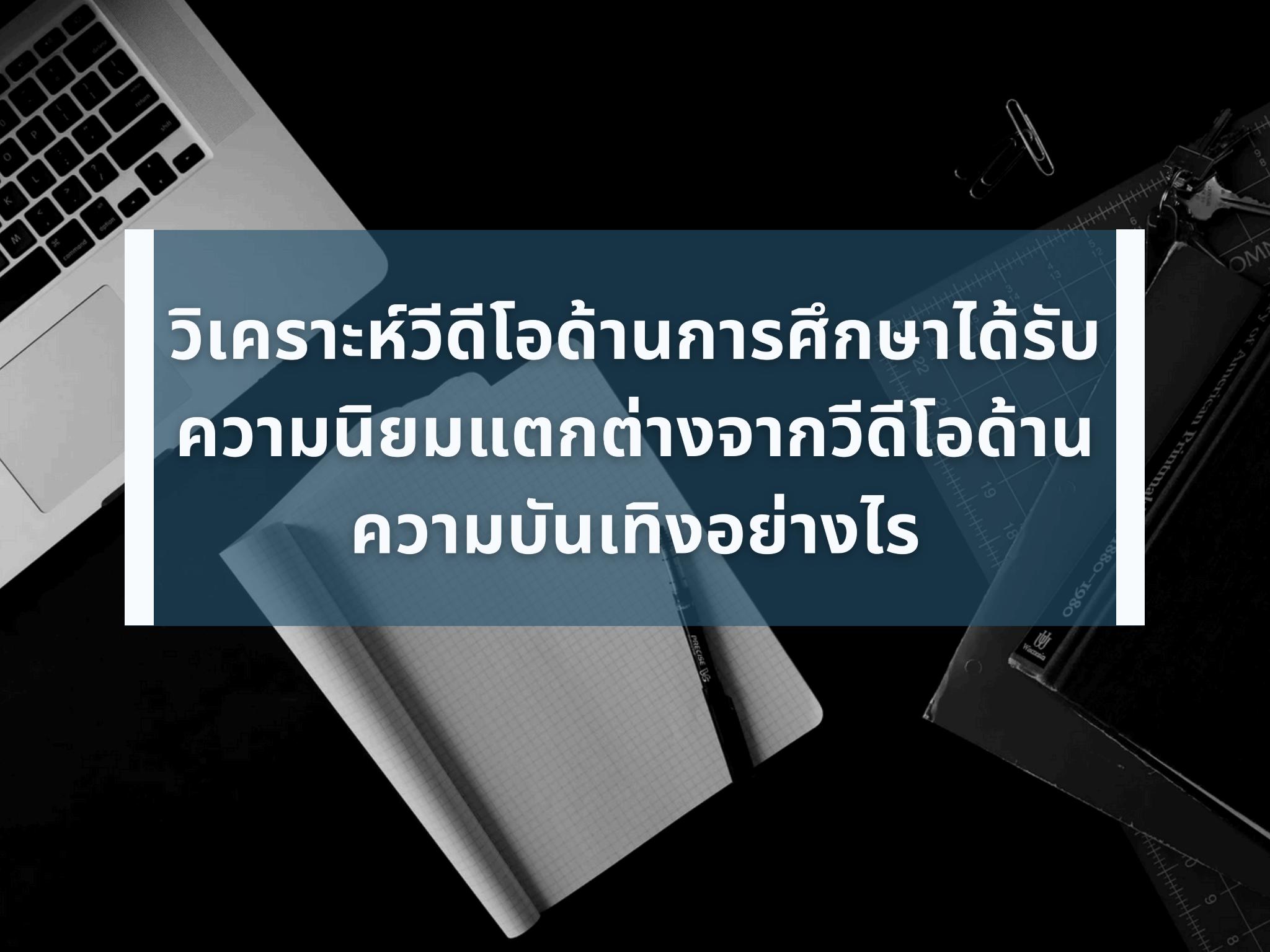


วิเคราะห์ดีໂອด้านการศึกษาได้รับ
ความนิยมແຕກຕ່າງຈາກວິດໂອດ້ານ
ความบันເທິງຍ່າງໄປ



ที่มาและความสำคัญ

ปัจจุบันมีคลิปวิดีโอบนโซเชียลมีเดียเป็นจำนวนมาก ซึ่งหนึ่งในแพลตฟอร์มที่มีผู้ใช้เป็นจำนวนมาก ในการสร้างและแชร์วิดีโอนั่นคือ YouTube

การคนละผู้จัดทำต้องการนำข้อมูลราย เอียงของแต่ละคลิปวิดีโอบน YouTube เช่น จำนวนผู้เข้าชม จำนวนยอดในการกดไลค์ ความยาวของวิดีโอ วันที่ลงคลิปวิดีโอ จำนวน การแสดงความคิดเห็น เป็นต้น เพื่อวิเคราะห์ พร้อมกับทำความเข้าใจเนื้อหาและลักษณะของ วิดีโอดังที่ได้รับความนิยม และสามารถนำเสนอ พฤติกรรมของผู้ชมว่าสนใจด้านการศึกษา แตกต่างจากด้านความบันเทิงอย่างไร



วัตถุประสงค์ ของโครง งาน



1. วิเคราะห์วิดีโอบน YouTube ว่าในแต่ละเดือนหรือแต่ละช่วงเวลา ผู้คนให้ความสนใจวิดีโอประเภทการศึกษาแตกต่างจากวิดีโอประเภทความบันเทิงอย่างไร



2. เพื่อศึกษาการใช้ API ใน การนำเข้าข้อมูล จากผู้ให้บริการ YouTube API



3. เพื่อศึกษาและอภิแบบฐานข้อมูลในการเก็บข้อมูลที่ได้จากแหล่งข้อมูลที่ต่างกัน

ขอบเขตข้อมูลที่สนใจ

1. ชื่อคลิปวิดีโอ
2. วันที่ลงคลิป
3. ID ของที่เป็นเจ้าของคลิป
4. หมวดหมู่ของวิดีโอ
5. จำนวนการรับชม
6. จำนวนการแสดงความคิดเห็น
7. จำนวนการกดถูกใจ
8. ความยาวของวิดีโอ



ตัวอย่างข้อมูลจาก YOUTUBE API

```
{ □  
  "id": "qNtx40b17Qs",  
  "snippet": { □  
    "publishedAt": "2023-12-24T04:00:16Z",  
    "channelId": "UCPPaFQnKHu73s6MZDxmNtUg",  
    "title": "สวัสดีค่ะ การเดินทางไปเมืองไทย | ภาระน้ำหนักต้องลดลง | หนังสือเรียนภาษาไทย",  
    "channelTitle": "Mr. FanTube",  
    "categoryId": "24"  
  },  
  "contentDetails": { □  
    "duration": "PT1H37M46S"  
  },  
  "statistics": { □  
    "viewCount": "882898",  
    "likeCount": "7069",  
    "commentCount": "108"  
  }  
}
```

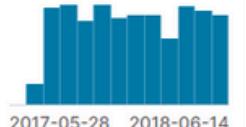
categoryId คือหมวดหมู่วิดีโอใน YouTube

channelId คือ Id ของช่องที่อัปโหลดวิดีโอ

duration คือ ความยาวของวิดีโอด้วยรูปแบบ ISO 8601

อีกหนึ่งส่วน **statistics** เก็บข้อมูลสถิติเกี่ยวกับวิดีโอ

ຕົວຢ່າງຂ້ອມູລ ຈາກ KAGGLE

▲ video_id	▲ title	▲ channel_title	category_id	publish_time	# views	# likes	# comment_count
#NAME?	1%	16721 unique values	1426 unique values			# views	# likes
rRr1qiJRsXk	0%						
Other (36825)	99%				4024	125m	0
kzwfHumJyYc	Sharry Mann: Cute Munda (Song Teaser) Parmish Verma Releasing on 17 November	Lokdhun Punjabi	1	2017-11-12T12:20:39.000Z	1096327	33966	882
zUZ1z7FwLc8	पीरियद्वारा के समय, पेट पर पति करता ऐसा, देखकर दंग रह जायेंगे	HJ NEWS	25	2017-11-13T05:43:56.000Z	590101	735	0
10L1hZ9qa58	Stylish Star Allu Arjun @ ChaySam Wedding Reception TFPC	TFPC	24	2017-11-12T15:48:08.000Z	473988	2011	149
N1vE8iiEg64	Eruma Saani Tamil vs English	Eruma Saani	23	2017-11-12T07:08:48.000Z	1242680	70353	2684
kJzGH0PVQHQ	why Samantha became EMOTIONAL @ Samantha naga chaithanya marriage Reception Filmylooks	Filmylooks	24	2017-11-13T01:14:16.000Z	464015	492	66
il_pSa5198w	MCA (Middle Class Abbayi) TEASER - Nani,Sai Pallavi, Dil Raju, Devi Sri Prasad, Sriram Venu	Dil Raju	24	2017-11-10T04:29:50.000Z	6106669	98612	4763

ຂ້ອມູລທີ່ໄດ້ຈະອູນໃນຮູບແບບໄຟຣ csv ທີ່ມີຂໍ້ອຄອລັນນີ້ ຈະສັນພັນນີ້ແລະມີຄວາມໝາຍເໜືອນກັນ
ກັບຊື່ຝລດ໌ຂອງຂ້ອມູລທີ່ໄດ້ຈາກ YouTube API

YOUTUBE API

1. สร้างฟังก์ชันดึงข้อมูลจาก YouTube API

```
[ ] # ฟังก์ชันในการค้นหาวีดีโอจาก Youtube API
def search_videos(Rcode, categoryId, number_results, qWord, nextPageToken=None):
    youtube = build('youtube', 'v3', developerKey=api_key)
    request = youtube.search().list(
        part='snippet',
        type='video',
        regionCode=Rcode,
        maxResults=number_results,
        videoCategoryId=categoryId,
        q = qWord,
        pageToken=nextPageToken if nextPageToken else None
    )
    response = request.execute()
    videos = response['items']
    nextPageToken = response.get('nextPageToken')

    if nextPageToken:
        more_videos = search_videos(Rcode, categoryId, number_results, qWord, nextPageToken)
        videos.extend(more_videos)

    return videos
```

ใช้ **method .search()** เพื่อค้นหาวีดีโอจาก YouTube โดยเลือกเอาข้อมูลในส่วน **snippet** คือ ข้อมูลพื้นฐาน เช่น ชื่อ คำอธิบาย **id** เป็นต้น **type** เป็นวีดีโอ **regionCode** ใช้เพื่อแสดงผลการค้นหาของวีดีโอดูได้ในประเทศที่ระบุ **maxResults** คือจำนวนจำกัดที่สามารถค้นหาได้ใน สูงสุด 50 **q** คือคำค้นหา **pageToken** ใช้เพื่อดึงข้อมูลหน้าถัดไป ถ้ายังมีโควต้า **nextPageToken** จะเรียกใช้ฟังก์ชัน **search_videos** ซ้ำ ในรูปแบบเวียนเกิด

YOUTUBE API (CONT)

2. สร้างฟังก์ชันดึงข้อมูลรายละเอียดวีดีโอใช้ YouTube API

```
# ฟังก์ชันในการดึงรายละเอียดต่างๆของวีดีโอด้วย id
def get_video_details(v_id):
    youtube = build('youtube', 'v3', developerKey=api_key)
    request = youtube.videos().list(
        part = "snippet,statistics,contentDetails",
        id=''.join(v_id[:50]),
        fields = "items(id,snippet(channelId, title,channelTitle,publishedAt,categoryId),statistics(viewCount,likeCount,dislikeCount,commentCount),contentDetails(duration))"
    )
    response = request.execute()

    return response
```

ใช้ **method .videos()** เพื่อ filter รายละเอียดภายในวีดีโອ่ที่เราสนใจ โดยเลือกเอาข้อมูลในส่วน **snippet, statistics, contentDetails**

id คือ รหัสวีดีโອ่ที่ต้องการดึงข้อมูล สูงสุด 50 รหัสคั่นด้วย คอมม่า

fields ใช้เพื่อรับข้อมูลเฉพาะที่ต้องการดึง

โดยที่ ID ของแต่ละวีดีโອ่ที่ได้จาก **method .search()**

YOUTUBE API (CONT)

3. สร้างฟังก์ชันดึงชื่อประเทศ จากรหัสช่อง

```
# ฟังก์ชันในค้นหาประเทศจาก channel id
def get_video_country(v_id):
    youtube = build('youtube', 'v3', developerKey=api_key)
    request = youtube.channels().list(
        part = "snippet",
        id=''.join(v_id[:50]),
    )
    response = request.execute()

    return response
```

ใช้ **method .channels()** เพื่อดึงเอาทรัพยากรณ์ของวิดีโอ YouTube โดยเลือกเอาข้อมูลในส่วน **snippet** เนื่องจากข้อมูลในส่วนนี้จะมีการเก็บชื่อประเทศที่เชื่อมโยงกับช่องนั้นๆอยู่ **id** คือ รหัสช่องที่ต้องการดึงข้อมูล สูงสุด 50 รหัสคันด้วย คอมมา โดยที่ **channel ID** ของแต่ละช่องได้จาก **method .video()**

YOUTUBE API (CONT)

4. ดึงข้อมูล Categories

```
[ ] # ดึงข้อมูล videoCategories จาก YouTube API  
yt = build('youtube' , 'v3', developerKey = api_key)  
lis = []  
for i in range(0,50):  
    rq = yt.videoCategories().list(  
        part="snippet",  
        id = f'{i}')  
    x = rq.execute()  
    lis.append(x)  
lis
```

ใช้ **method .videoCategory()** เพื่อค้นหาข้อมูลหมวดหมู่วิดีโอที่มีรหัส ตั้งแต่ 0 - 49 โดยเลือกเอาข้อมูลในส่วน **snippet** เนื่องจากข้อมูลในส่วนนี้จะมีการเก็บรหัสและชื่อหมวดหมู่ **id** คือ รหัสหมวดหมู่ที่ต้องการดึง เก็บข้อมูลที่ได้ไว้ในตัวแปร **lis** เพื่อไปเลือกข้อมูลและทำความสะอาด

YOUTUBE API (CONT)

5. เลือกข้อมูล และแสดงข้อมูล category

```
▶ # เลือกข้อมูลโดยจะเอาข้อมูลที่ฟิลด์ items ไม่เป็นลิสต์ว่างไปเก็บใน videoCategoryList  
videoCategoryList = []  
for i in lis:  
    if i[ 'items' ] != []:  
        videoCategoryList.append(i[ 'items' ])  
  
# จัดเก็บข้อมูลที่ได้และนำเอาเฉพาะข้อมูลในฟิลด์ id และ title ของ videoCategoryList ไปสร้าง DataFrame  
categoryId , categoryName = list() , list()  
  
for i in videoCategoryList:  
    categoryId.append(i[0][ 'id' ])  
    categoryName.append(i[0][ 'snippet' ][ 'title' ])  
  
videoCategoryDataFrame = pd.DataFrame({ 'categoryId':categoryId , 'categoryName':categoryName })  
videoCategoryDataFrame
```

24	Entertainment
25	News & Politics
26	Howto & Style
27	Education

จะพบว่า รหัสหมวดหมู่ของวีดิโอที่ตรงกับ
ความต้องการของผู้จัดทำคือ 24 และ 27
ที่มีชื่อหมวดหมู่เป็น Entertainment
และ Education ตามลำดับ

KAGGLE DATASET

1.จัดการข้อมูลที่ได้จาก dataset

```
▶ # สร้าง dataframe ของอินเดีย  
kaggle_IN = pd.read_csv("INvideos.csv")  
  
# id ของวีดีโอมหาดมังเหงิง  
IN = kaggle_IN.query("category_id == 24")  
IN = IN.drop(columns=['trending_date']).drop_duplicates().sort_values("views", ascending=False)  
df_IN_entertainment = IN.drop_duplicates().reset_index(drop=True)  
# id ของวีดีโอมหาการศึกษา  
IN = kaggle_IN.query("category_id == 27")  
IN = IN.drop(columns=['trending_date']).sort_values("views", ascending=False).reset_index(drop=True)  
df_IN_education = IN.drop_duplicates().reset_index(drop=True)  
  
# List ของ id ในวีดีโอดังต่อไปนี้  
ID_form_kaggle_education = df_IN_education['video_id'].unique().tolist()  
ID_form_kaggle_entertainment = df_IN_entertainment['video_id'].unique().tolist()  
print('วีดีโອดีบดิชของ education ที่ได้จาก kaggle มีทั้งหมด {} วีดีโอ')  
print('วีดีโอดีบดิชของ entertainment ที่ได้จาก kaggle มีทั้งหมด {} วีดีโอ')
```

▶ วีดีโอดีบดิชของ education ที่ได้จาก kaggle มีทั้งหมด 799 วีดีโอ
วีดีโอดีบดิชของ entertainment ที่ได้จาก kaggle มีทั้งหมด 7564 วีดีโอ

คัดเลือกเอาข้อมูลจาก dataset คือ เอาข้อมูล Id ของวีดีโอกลุ่มนี้ที่มีรหัสหมวดหมู่เป็น 24 และ 27 มาเก็บไว้ในตัวแปร ID_form_kaggle_education และ ID_form_kaggle_entertainment ตามลำดับ และนำ ID ที่ได้มาไปค้นหาข้อมูลรายละเอียดวีดีโองาน ฟังก์ชัน video_details ที่ได้สร้างไว้ในพาร์ท YouTube API

```
▶ # เรียกฟังก์ชันดึงรายละเอียดต่างๆในวีดีโอกลุ่มนี้มาไว้ใน variable คือ (id ของ video จาก kaggle)  
details = video_details(ID_form_kaggle_education)  
details = video_details(ID_form_kaggle_entertainment)
```

MONGO DB

1. เชื่อมต่อฐานข้อมูล MongoDB และ Insert ข้อมูลลงฐานข้อมูล

```
[ ] # เชื่อมต่อ database จาก MongoDB
uri = "mongodb+srv://projectYoutube:FvGPuWp4Dv6PNFnO@project.r bqekbu.mongodb.net"
client = pymongo.MongoClient(uri)
# client.list_database_names()
db = client.ProjectDB

education = db.Education
entertainment = db.Entertainment
country = db.Country
videocategoty = db.VideoCategory
```

collection ที่ได้ทำการสร้างไว้ มี 4 collection

```
[ ] # พิ่งก์ชัน insert data จาก Youtube api ไปยัง MongoDB
def insert_data(col, data):
    for d in data:
        if len(d['items']) > 0:
            yt_data = d['items']
            col.insert_many(yt_data)
```

ฟังก์ชันสำหรับ Insert ข้อมูลที่เก็บมา ลงฐานข้อมูล MongoDB

MONGO DB

ตัวอย่างข้อมูลใน collection Country

```
_id: ObjectId('65cc53774afaf40a985defa')
kind: "youtube#channel"
etag: "5ppR0w3yJEM3pqL9RmEhb1PCTxE"
id: "UCnIG60Ir971Wpa0nkMlViDQ"
  snippet: Object
    title: "พีแม็ค The NEW GEN"
    description: ""
    customUrl: "@macthenewgen"
    publishedAt: "2020-10-20T13:43:14.112793Z"
  thumbnails: Object
  localized: Object
  country: "TH"
```

id คือรหัสช่องที่ลงไว้ดีโอ

country คือชื่อตัวย่อประเทศที่ช่องนั้นๆ เชื่อมโยง

ตัวอย่างข้อมูลใน collection Education

```
_id: ObjectId('65cc51f24afaf40a985dc9d')
id: "mbmckE6eJkQ"
  snippet: Object
    publishedAt: "2023-05-27T12:00:09Z"
    channelId: "UCXDJB-XhmWAf6B_vDfJHDNg"
    title: "#ดิวสอนครุภู่ช่วย [EP.1] ภาค ก คณิตศาสตร์&เหตุผล By...แก้วดิวเดอร์"
    channelTitle: "แก้วดิวเดอร์ คณิตศาสตร์ขั้นมหาวน"
    categoryId: "27"
  contentDetails: Object
    duration: "PT1H57M36S"
  statistics: Object
    viewCount: "62305"
    likeCount: "1550"
    commentCount: "52"
```

channelId คือรหัสช่อง เป็นข้อมูลเดียวกันกับ **id** ใน collection country

publishedAt คือวันเวลาที่ลงคลิปไว้ดีโอ

duration คือ ความยาวคลิปในรูปแบบ ISO 8601

MONGO DB

ตัวอย่างข้อมูลใน collection Entertainment

```
_id: ObjectId('65cc71bb4afa1f40a985f161')
id: "qNtx40b17Qs"
snippet: Object
  publishedAt: "2023-12-24T04:00:16Z"
  channelId: "UCPPaFQnKHu73s6MZDxmNtUg"
  title: "สวรรค์แห่งการพากไทย | ภาคยนตร์ศิลปะการต่อสู้ | หนังจีนพากย์ไทย"
  channelTitle: "Mr. FanTube"
  categoryId: "24"
contentDetails: Object
  duration: "PT1H37M46S"
statistics: Object
  viewCount: "882898"
  likeCount: "7069"
  commentCount: "108"
```

ตัวอย่างข้อมูลใน collection VideoCategory

```
1: "Film & Animation"
2: "Autos & Vehicles"
10: "Music"
15: "Pets & Animals"
17: "Sports"
18: "Short Movies"
19: "Travel & Events"
20: "Gaming"
21: "Videoblogging"
22: "People & Blogs"
23: "Comedy"
24: "Entertainment"
25: "News & Politics"
26: "Howto & Style"
27: "Education"
28: "Science & Technology"
29: "Nonprofits & Activism"
```

channelId คือรหัสช่อง เป็นข้อมูลเดียวกันกับ **id** ใน collection country

publishedAt คือวันเวลาที่ลงคลิปวีดีโอ

duration คือ ความยาวคลิปในรูปแบบ ISO 8601

ประกอบด้วยข้อมูลคือ รหัสหมวดหมู่ และชื่อหมวดหมู่

จัดการและทำความรู้จักข้อมูล

จัดการข้อมูลจาก collection Country

```
[ ] # สร้าง dataframe จาก document ใน country ที่ query มาจาก MongoDB
j = json.dumps([list(country.find({},{'_id': 0, 'id':1, 'snippet.country':1, 'snippet.title':1, 'snippet.description':1}))])
df = pd.json_normalize(json.loads(j))
df.rename(columns = {'id':'channel_id', 'snippet.country':'country', 'snippet.title':'channel_title', 'snippet.description':'description'}, inplace = True)

def lang_detect(x):
    try:
        result = detect(x)
    except LangDetectException as e:
        result = str(e)
    return result

df['language'] = np.where(np.logical_or(df['channel_title'].apply(lang_detect) == 'th', df['description'].apply(lang_detect) == 'th'), 'th', 'other')
index = df[df['country'].isna()].query("language == 'th'").index
df.loc[index, "country"] = "TH"

df_country = df.query('country == "TH"').drop(columns=['language','description'])
df_country
```

บาง document จะไม่มี field **country** ทำให้ข้อมูลคอลัมน์ **country** ใน DataFrame มีค่าว่าง จึงใช้การตรวจสอบภาษา ถ้ามีชื่อช่อง หรือ คำอธิบาย มีภาษาไทยอยู่ จะกำหนดให้ คอลัมน์ **country** ใน DataFrame เป็น TH

ตัวอย่างข้อมูลใน DataFrame

	channel_id	channel_title	country
1	UC_Ddwm4jjuCqEE7hW3ExScA	คณิตเพลสกี้	TH
2	UCDaffzFTDEy319jqxJks9EA	CoursewareMaster	TH
3	UCdYLNljIsIokdEnKbPPqIBA	ครูฟันคนสอนเลข	TH
4	UCEDTSRW-0iUxL_45MR3ESPg	Proj14 ม.3	TH
5	UCq1yn0fvaU4KfReXIV4GgCQ	NockAcademy -ไฟฟ์สอนสด อันดับ1-	TH

จัดการและทำความรู้จักข้อมูล

จัดการข้อมูลจาก collection Education

```
# สร้าง dataframe จาก document ใน education ที่ query มาจาก MongoDB
j = json.dumps(list(education.find({}, {"_id": 0})))
df = pd.json_normalize(json.loads(j))
df.rename(columns = {"id":'video_id', 'snippet.title':'video_title', 'snippet.channelId':'channel_id',
df_education = df

# ทำการ merge dataframe เพื่อ filter video จากประเทศไทย
df = df_education.merge(df_country)
df.dropna(inplace=True)
df[['views','likes','comment_count']] = df[['views','likes','comment_count']].astype('int')
df_TH = df

# dataframe education จากประเทศไทยอันเดียว
j = json.dumps(list(education.find({"id": {"$in": ID_form_kaggle_education}}, {"_id": 0})))
df = pd.json_normalize(json.loads(j))
df.rename(columns = {"id":'video_id', 'snippet.title':'video_title', 'snippet.channelId':'channel_id',
df.dropna(inplace=True)
df['country'] = 'IN'
df[['views','likes','comment_count']] = df[['views','likes','comment_count']].astype('int')
df_IN = df

# ทำการ concat datacaion จากทั้ง 2 ประเทศ และเลือกมา 5000 แล้ว
df_education = pd.concat([df_TH, df_IN])
df_education['category_id'] = df_education['category_id'].astype("str")
df_education = df_education.sort_values("views", ascending=False)
df_education = df_education.query("category_id == '27'")[:5000].reset_index(drop=True)
df_education
```

1. ทำการเปลี่ยนชื่อคอลัมน์ และรวม DataFrame df_education กับ df_country จะได้ข้อมูลที่มีรหัสวีดีโอตรงกับรหัสวีดีโอใน dataframe df_country เก็บในตัวแปร df_TH

2. คัดเลือกข้อมูลและเอาเฉพาะข้อมูลที่มี id ตรงกับ id_from_kaggle_education ซึ่งข้อมูลจาก kaggle จะเป็นข้อมูลประเทศไทยเดียว ดังนั้น จึงกำหนดให้คอลัมน์ country ใน DataFrame เป็น IN และเก็บไว้ในตัวแปร df_IN

3. ทำการ concat df_TH และ df_IN พร้อมเลือกข้อมูลมา 5000 แล้ว

ตัวอย่างข้อมูลใน DataFrame

	video_id	published_time	channel_id	video_title	channel_title	category_id	duration	views	likes	comment_count	country
0	JEm6Ftxk2y8	2015-05-14T06:52:43Z	UCB8p9XDpbCoX9slaPPP7sgw	เพลง ABC Song บทเพลง คำนาร์ เพลงเด็ก @KidsOn...	KidsOnCloud	27	PT4M30S	332329485	733662	0	TH
1	HjWMxuC4X6c	2015-05-22T13:26:12Z	UCB8p9XDpbCoX9slaPPP7sgw	เพลงเพื่อเด็กวันนี้ อักษร A-Z แบบ เรียนเพลง บทเพลงคร...	KidsOnCloud	27	PT5M42S	39242959	129058	0	TH
2	gxtLzMqFzEk	2018-02-07T03:52:57Z	UCKjjXmZxk1SITUQUB28Khpw	วินา รุก, วินา อะคู อะเมสตี้ กีซ่า แบล็ค How...	TsMadaan	27	PT7M17S	21966977	568560	27626	IN
3	QjI6o3dZBwc	2016-10-13T14:01:56Z	UC103Yv2QS7C2Vf17XIMk0AQ	ก ใจ ฝึกเรียน ฝึกอ่าน ก-ศ สำหรับเด็กอนุบาล 🍀 ...	Indysong Kids เพลงเด็ก น้อย มีท่านป่องเป็ลลิ่นต์	27	PT11M43S	18031502	57371	0	TH
4	nDDfFws3BfA	2017-12-17T18:00:03Z	UCYenDLnIHsoqQ6smwKXQ7Hg	10,000 Years Into the Future in 10 Minutes	#Mind Warehouse	27	PT13M15S	16284675	122603	27987	IN

จัดการและทำความรู้จักข้อมูล

จัดการข้อมูลจาก collection Entertainment

```
# สร้าง dataframe จาก document ใน entertainment ที่ query มาจาก MongoDB
j = json.dumps(list(entertainment.find({},{'_id': 0})))
df = pd.json_normalize(json.loads(j))
df.rename(columns = {'id':'video_id', 'snippet.title':'video_title', 'snippet.channelId':'channel_id', 'snip
df_entertainment = df

# ทำการ merge dataframe เพื่อ filter video จากประเทศไทย
df = df_entertainment.merge(df_country)
df.dropna(inplace=True)
df[['views','likes','comment_count']] = df[['views','likes','comment_count']].astype('int')
df.drop_duplicates(inplace=True)
df_TH = df

# dataframe entertainment จากประเทศไทยอีกเดียว
j = json.dumps(list(entertainment.find({"id":{"$in":ID_form_kaggle_entertainment}},{'_id': 0})))
df = pd.json_normalize(json.loads(j))
df.rename(columns = {'id':'video_id', 'snippet.title':'video_title', 'snippet.channelId':'channel_id', 'snip
df.dropna(inplace=True)
df['country'] = 'IN'
df[['views','likes','comment_count']] = df[['views','likes','comment_count']].astype('int')
df_IN = df

# ทำการ concat dataframe entertainment จากทั้ง 2 ประเทศ และเลือกมา 5000 แถว
df_entertainment = pd.concat([df_IN, df_TH])
df_entertainment['category_id'] = df_entertainment['category_id'].astype("str")
df_entertainment = df_entertainment.sort_values("views",ascending=False)
df_entertainment = df_entertainment.query("category_id == '24'")[:5000].reset_index(drop=True)
df_entertainment
```

- ทำการเปลี่ยนชื่อคอลัมน์ และรวม DataFrame df_entertainment กับ df_country จะได้ข้อมูลที่มีรหัสวีดีโอตรงกับรหัสวีดีโอใน dataframe df_country เก็บในตัวแปร df_TH
- คัดเลือกข้อมูลและเอาเฉพาะข้อมูลที่มี id ตรงกับ id_from_kaggle_entertainment ซึ่งข้อมูลจาก kaggle จะเป็นข้อมูลประเทศไทยอันเดียดังนั้น จึงกำหนดให้คอลัมน์ country ใน DataFrame เป็น IN และเก็บไว้ในตัวแปร df_IN
- ทำการ concat df_TH และ df_IN พร้อมเลือกข้อมูลมา 5000 แถว

ตัวอย่างข้อมูลใน DataFrame

	video_id	published_time	channel_id	video_title	channel_title	category_id	duration	views	likes	comment_count	country
0	6ZfuNTqbHE8	2017-11-29T13:26:24Z	UCvC4D8onUfxzvJTOM-dBFfEA	Marvel Studios' Avengers: Infinity War Official...	Marvel Entertainment	24	PT2M25S	261874712	4129985	453710	IN
1	1L6Nrdjyp_4	2018-03-08T11:30:01Z	UC6-F5tO8uklgE9Zy8IvbdFw	Tina's First Day With The Gada Family Tapu S...	Sony SAB	24	PT13M29S	251786703	624159	15899	IN
2	FlsCjmMhFmw	2017-12-06T17:58:51Z	UCBR8-60-B28hp2BmDPdntcQ	YouTube Rewind: The Shape of 2017 #YouTubeRe...	YouTube	24	PT7M15S	241896188	4692502	909353	IN
3	n4tFuxxKhgo	2018-09-12T13:31:21Z	UC92shgTDA-nXNcl9sWORblQ	สายแนนหัวใจ - ก้อง หวยไร่ (เพลง ประกอบภาพยนตร์...	Search Group (Official)	24	PT3M57S	190996263	609928	18477	TH
4	mK9C6wViwn4	2018-03-21T12:07:49Z	UCLtCejNI8eAg4PO_9lf2Tig	Raambo 2 Chuttu Chuttu 4K Video Song Sha...	Anand Audio	24	PT3M30S	181112323	446651	15239	IN

TRANSFORM DATA

รวมข้อมูลและเลือกคอลัมน์ที่ต้องการ

```
# ทำการรวม Dataframe ของวิดีโอทั้ง 2 ประเภท  
yt = pd.concat([df_education, df_entertainment], ignore_index=True)
```

```
# เลือกเฉพาะ column ที่ต้องการใช้
```

```
youtube = yt[['channel_title', 'video_title', 'category_id', 'duration', 'views', 'likes', 'comment_count', 'country', 'published_time']].copy()  
youtube
```

จะได้ข้อมูลใน dataframe เป็น 10000 row พร้อมกับเลือกเอาเฉพาะคอลัมน์ที่ต้องใช้

แปลงหน่วยแบบเวลา

```
# ฟังก์ชันแปลงเวลาจาก ISO-8601 เป็นนาที
```

```
def convert_time(time):  
    time_to_seconds = isodate.parse_duration(time)  
    return (time_to_seconds.total_seconds() / 60)
```

```
youtube['duration'] = youtube['duration'].apply(convert_time)  
youtube = youtube.rename(columns={"duration": "duration(minutes)"})  
youtube
```

แปลงเวลาจาก ISO-8601 เป็นหน่วยนาที คอลัมน์ที่ทำการแปลงคือ duration เก็บความยาวคลิปวิดีโอ และเปลี่ยนชื่อคอลัมน์เป็นduration(minutes)

ตัวอย่างข้อมูลใน dataframe

	channel_title	video_title	category_id	duration(minutes)	views	likes	comment_count	country	published_time
0	KidsOnCloud	เพลง ABC Song บทเพลงความรู้ เพลงเด็ก @KidsOn...	27	4.500000	332329485	733662	0	TH	2015-05-14T06:52:43Z
1	KidsOnCloud	เพลงเพื่อนสัตว์น่ารัก A-Z แนวเดิมเพลง บทเพลงค...	27	5.700000	39242959	129058	0	TH	2015-05-22T13:26:12Z
2	TsMadaan	บินา รุก, บินา อะ�เค ॲंगรีzi กैसे बोलें How...	27	7.283333	21966977	568560	27626	IN	2018-02-07T03:52:57Z
3	Indysong Kids	เพลงเด็กน้อย มีหานน่องเป๊กอินดี้ ก ໄກ ฝึกเขียน ฝึกอ่าน ก-ช สำหรับเด็กอนุบาล 🌻 ...	27	11.716667	18031502	57371	0	TH	2016-10-13T14:01:56Z
4	#Mind Warehouse	10,000 Years Into the Future in 10 Minutes	27	13.250000	16284675	122603	27987	IN	2017-12-17T18:00:03Z

TRANSFORM DATA

แปลงรหัสหมวดหมู่วิดีโอเป็นชื่อหมวดหมู่

```
# สร้าง dataframe ของประเภทของวีดีโอ
```

```
df = pd.json_normalize(([list(videoCategory.find({}))]).transpose()
```

```
df = df.iloc[1:, :].reset_index().rename(columns={0:"category_name", "index":"category_id"})
```

```
# ทำการรวม dataframe ที่มี category ID ตรงกัน
```

```
youtube = youtube.merge(df).drop(columns='category_id')
```

```
youtube
```

สร้าง dataframe ประเภทวีดีโอ โดยเอาข้อมูลจาก collection videoCategory และ merge dataframe youtube กับมีรหัสหมวดหมู่ตรงกัน พร้อมกับต่อไปคอลัมน์ category_id

ตัวอย่างข้อมูลใน dataframe

	channel_title	video_title	duration(minutes)	views	likes	comment_count	country	published_time	category_name
0	KidsOnCloud	เพลง ABC Song บทเพลงความรู้ เพลงเด็ก @KidsOn...	4.500000	332329485	733662	0	TH	2015-05-14T06:52:43Z	Education
1	KidsOnCloud	เพลงเพื่อนสัตว์น่ารัก A-Z แบนเนิมเพลง บทเพลงคุ...	5.700000	39242959	129058	0	TH	2015-05-22T13:26:12Z	Education
2	TsMadaan	बिना रुके, बिना अटके अंग्रेजी कैसे बोलें How...	7.283333	21966977	568560	27626	IN	2018-02-07T03:52:57Z	Education
3	Indysong Kids เพลงเด็กน้อย นิทานน้อดปั๊ด อินดี้	ก ໄກ ฝึกเขียน ฝึกอ่าน ก-ษ สำหรับเด็กอนุบาล 🍀 ...	11.716667	18031502	57371	0	TH	2016-10-13T14:01:56Z	Education
4	#Mind Warehouse	10,000 Years Into the Future in 10 Minutes	13.250000	16284675	122603	27987	IN	2017-12-17T18:00:03Z	Education
...
9995	SPY Channel	8 อันดับ การคุณคนไทย ที่ไม่มีอะไรเทียบ (อันนี้...	7.766667	394380	15033	2903	TH	2021-06-19T10:00:03Z	Entertainment
9996	คิมจี สปอร์ต	{สปอร์ตออนไลน์} เมื่อยากเข้าสูตรให้ต้องมาเลี้ยง...	123.866667	394259	13054	205	TH	2022-09-23T01:00:01Z	Entertainment
9997	GMM25Thailand	ละครคน EP.24 [1/5]	17.250000	394187	828	91	TH	2017-07-11T15:00:06Z	Entertainment

TRANSFORM DATA

ใช้ regular expressions แยกคอลัมน์ published_time

```
# Dataframe ของ datetime
datetime = youtube.copy()
datetime[['published_time', 'date', 'time']] = datetime['published_time'].str.extract(r'((\d{4}-\d{2}-\d{2})\w(\d{2}:\d{2}:\d{2})\w)')
datetime[['year', 'month', 'day']] = datetime.date.str.split("-", expand=True)
datetime[['hours', 'minuts', 'seconds']] = datetime.time.str.split(":", expand=True)
datetime = datetime[['day', 'month', 'year', 'hours', 'minuts', 'seconds']]
datetime
```

ตัวอย่างข้อมูลใน dataframe

	day	month	year	hours	minuts	seconds
0	14	05	2015	06	52	43
1	22	05	2015	13	26	12
2	07	02	2018	03	52	57
3	13	10	2016	14	01	56

Merge dataframe

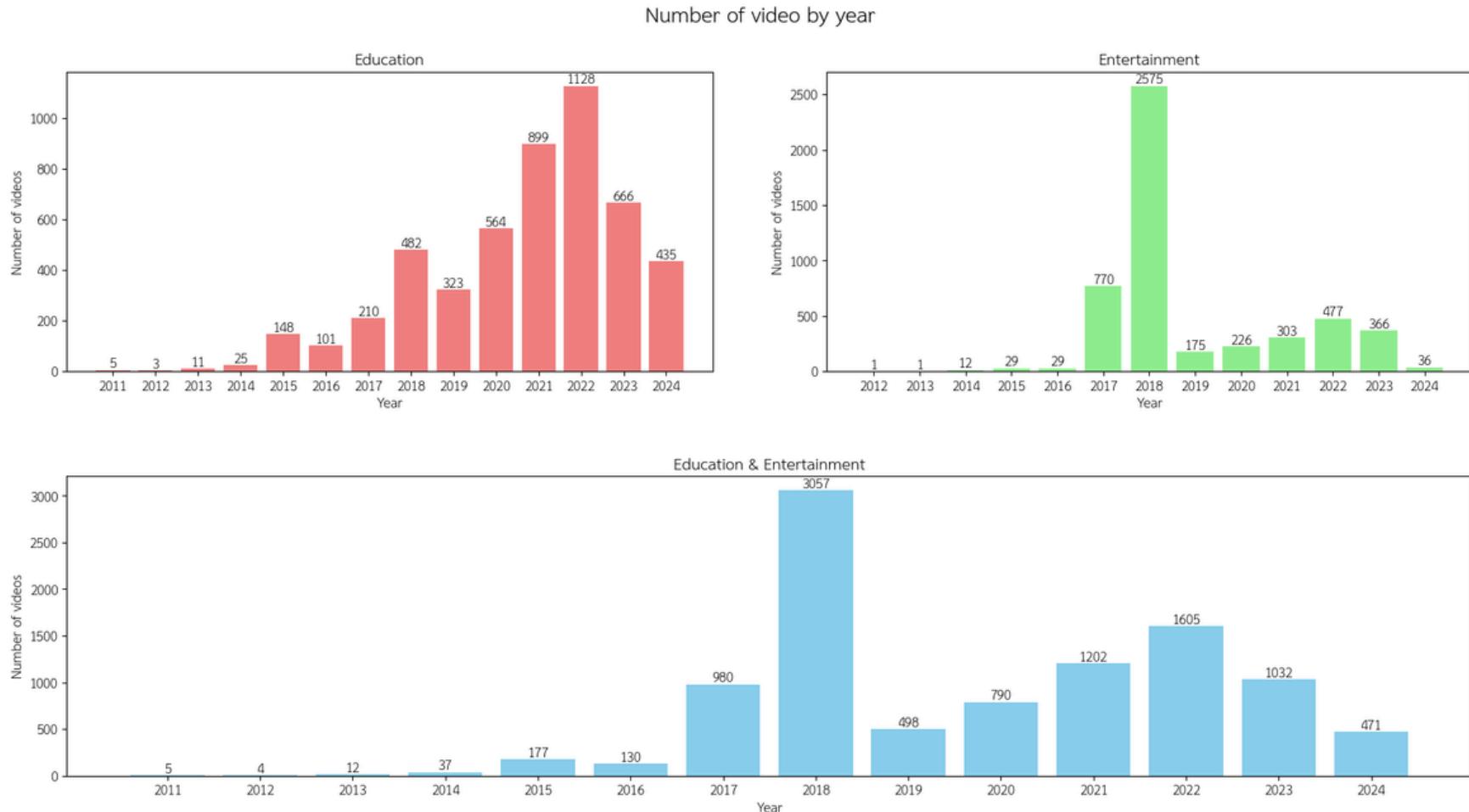
```
# ทำการ merge dataframe
youtube = youtube.merge(datetime, left_index=True, right_index=True).drop(columns='published_time')
youtube
```

ตัวอย่างข้อมูลใน dataframe ที่ทำการ merge เสร็จเรียบร้อย

	channel_title	video_title	duration(minutes)	views	likes	comment_count	country	category_name	day	month	year	hours	minuts	seconds
0	KidsOnCloud	เพลง ABC Song บทเพลงความรู้ เพลงเด็ก @KidsOn...	4.500000	332329485	733662	0	TH	Education	14	05	2015	06	52	43
1	KidsOnCloud	เพลงเพื่อนสัตว์น่ารัก A-Z แบบเต็ม เพลง บทเพลงคุ...	5.700000	39242959	129058	0	TH	Education	22	05	2015	13	26	12
2	TsMadaan	बिना रुके, बिना अटके अंग्रेजी कैसे बोले How...	7.283333	21966977	568560	27626	IN	Education	07	02	2018	03	52	57
3	Indysong Kids	เพลงเด็กน้อย นิทานน่องเป็ดอินดี้ ก ไก่ ฝึกเขียน ฝึกอ่าน ก-ช ส่าหรับเด็กอนุบาล 🍀 ...	11.716667	18031502	57371	0	TH	Education	13	10	2016	14	01	56
4	#Mind Warehouse	10,000 Years Into the Future in 10 Minutes	13.250000	16284675	122603	27987	IN	Education	17	12	2017	18	00	03

DATA ANALYSIS

กราฟแท่งแสดงจำนวนวิดีโอที่ปล่อยในแต่ละปี

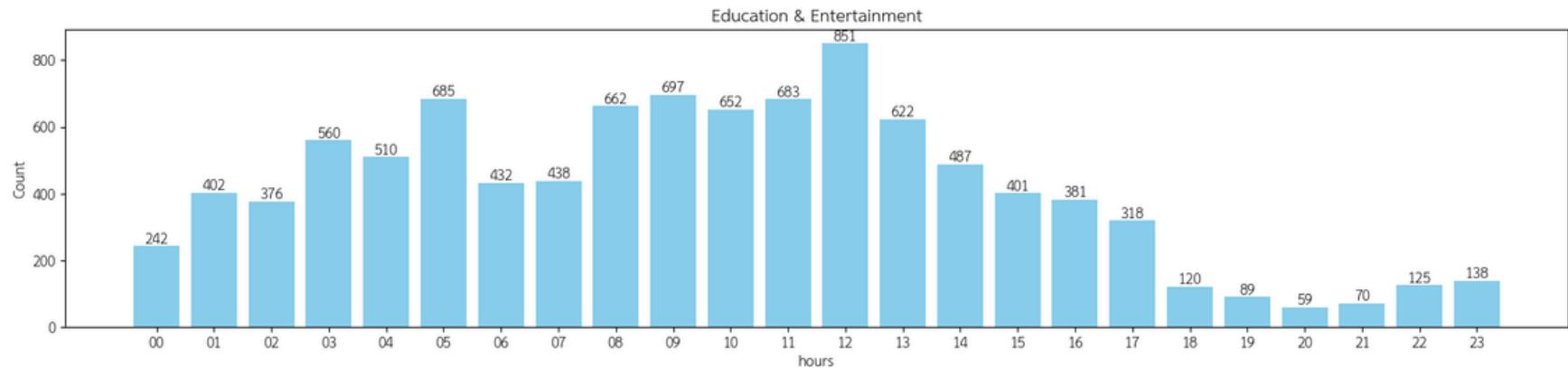
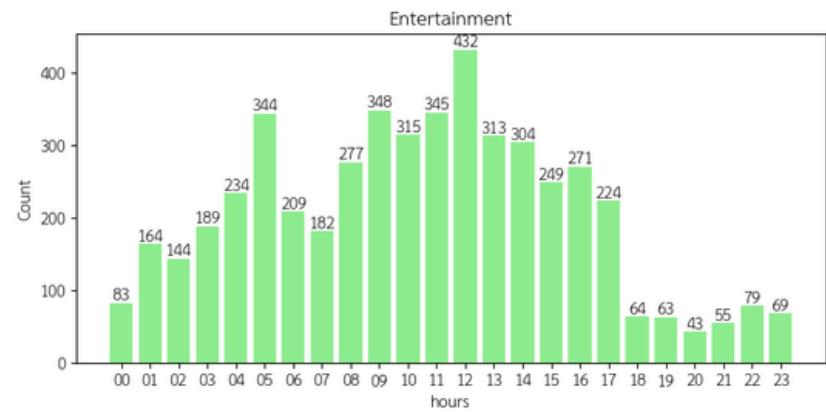
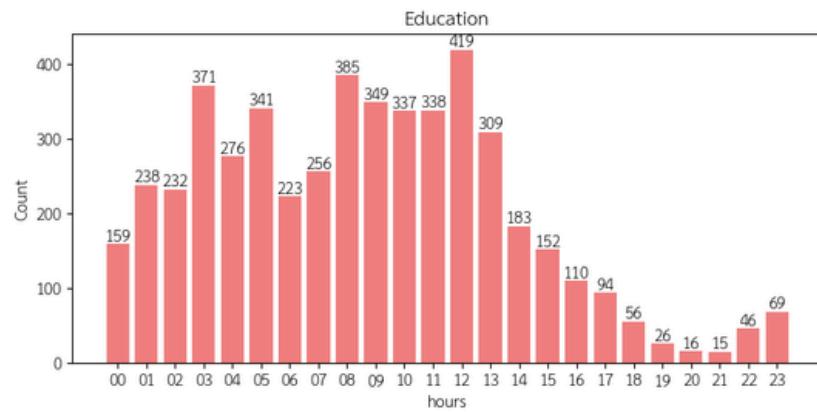


จากการพบร่วมกัน ปี 2022 จะเป็นปีที่มีการอัปโหลดวิดีโอหมวดหมู่ **education** มากที่สุด
ปี 2018 จะเป็นปีที่มีการอัปโหลดวิดีโอหมวดหมู่ **entertainment** มากที่สุด
เมื่อร่วมกันจำนวนของวิดีโอทั้งสองหมวดหมู่จะได้ว่า ปี 2018 คือปีที่มีการอัปโหลดวิดีโอมากที่สุด

DATA ANALYSIS

กราฟแท่งแสดงจำนวนวิดีโอในแต่ละช่วงเวลา

Number of video by hours

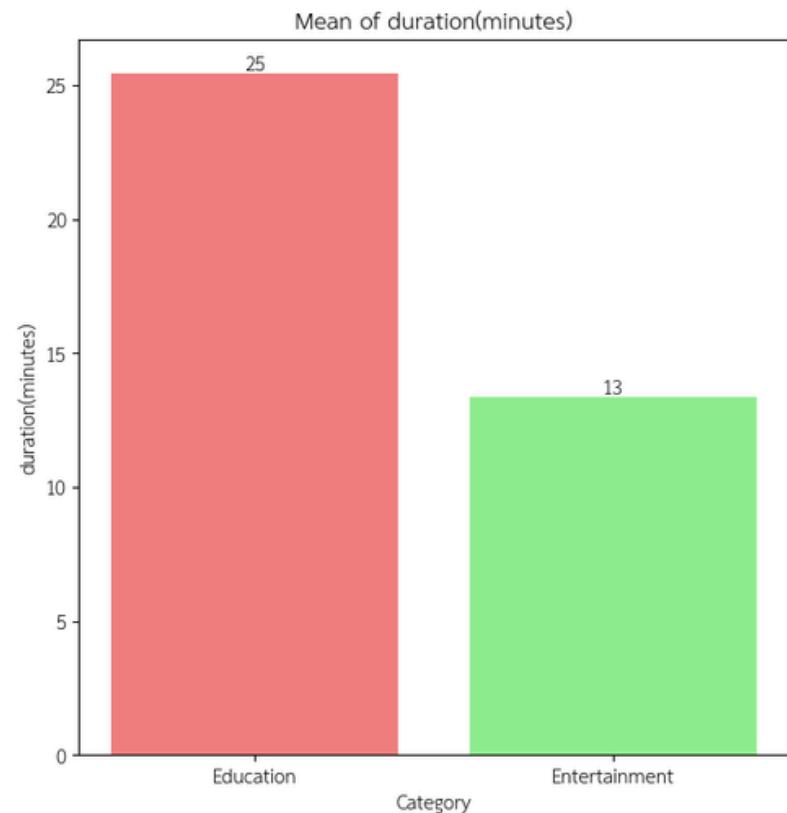
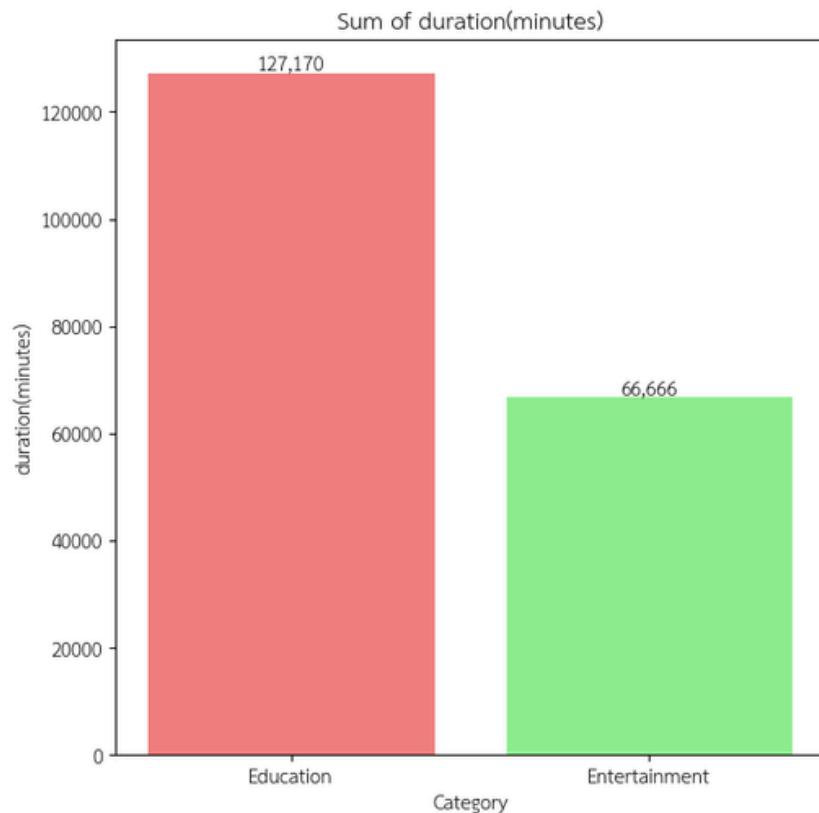


พบว่าวิดีโอกึ้งสองประเภทถูกอัปโหลดลง YouTube มากที่สุดคือช่วงเวลาเที่ยงวัน

DATA ANALYSIS

กราฟแท่งแสดงผลรวมและค่าเฉลี่ยของความยาววิดีโอ

duration(minutes) of video

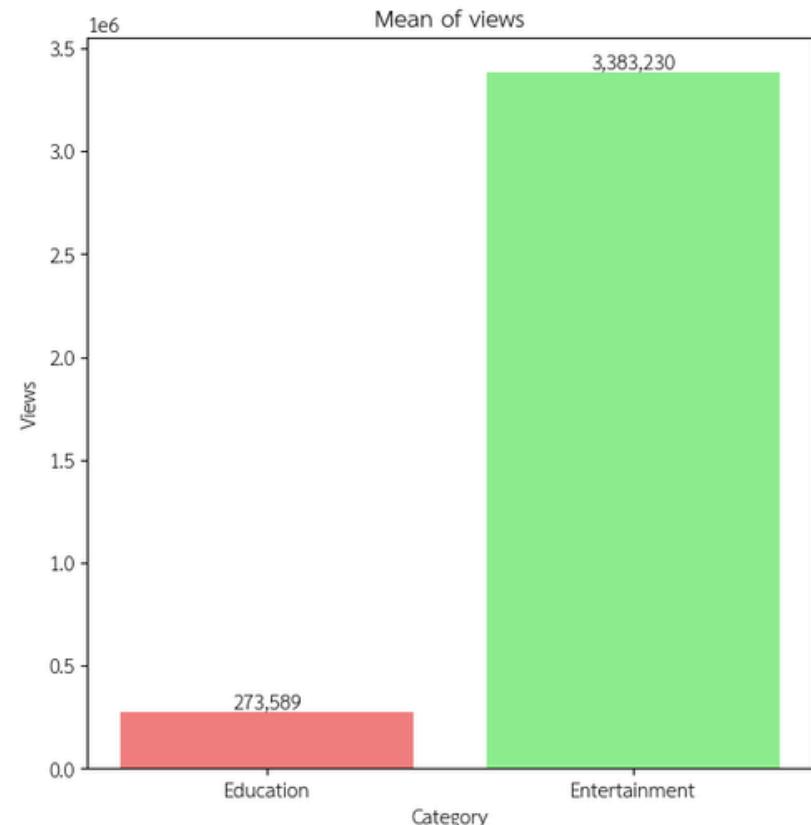
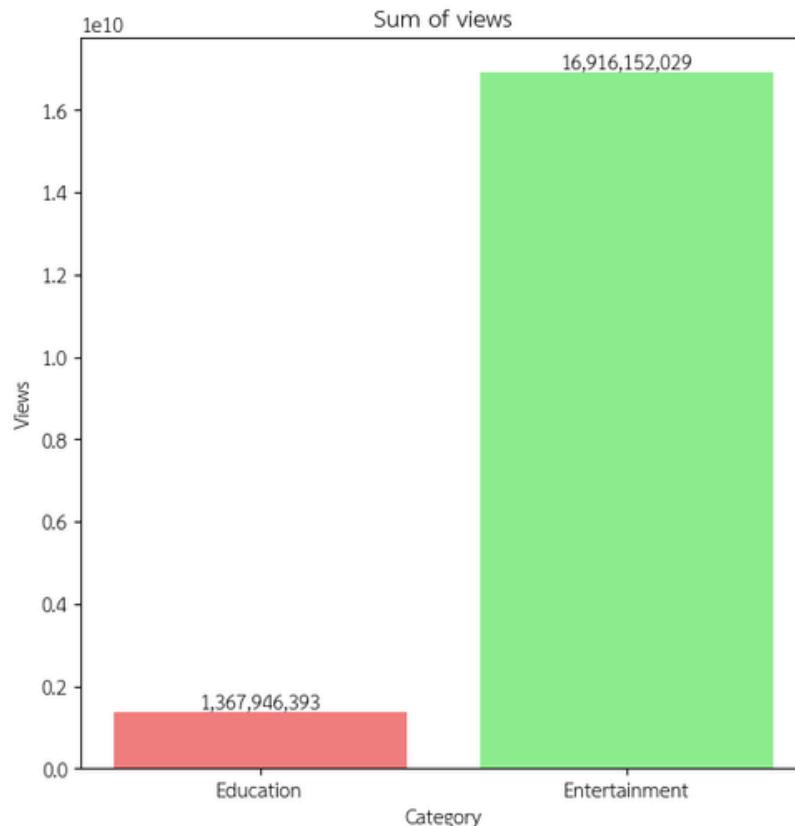


พบว่าวิดีโอมีหมวดหมู่ การศึกษา จะมีความยาวคลิปมากกว่าวิดีโอมีหมวดหมู่บันเทิง
ค่าเฉลี่ยของความยาววิดีโอด้วยหมวดหมู่การศึกษา และ หมวดหมู่บันเทิง อยู่ที่ 25นาที และ 13นาที ตามลำดับ

DATA ANALYSIS

กราฟแท่งแสดงผลรวมและค่าเฉลี่ยของยอดการรับชม

Views of video

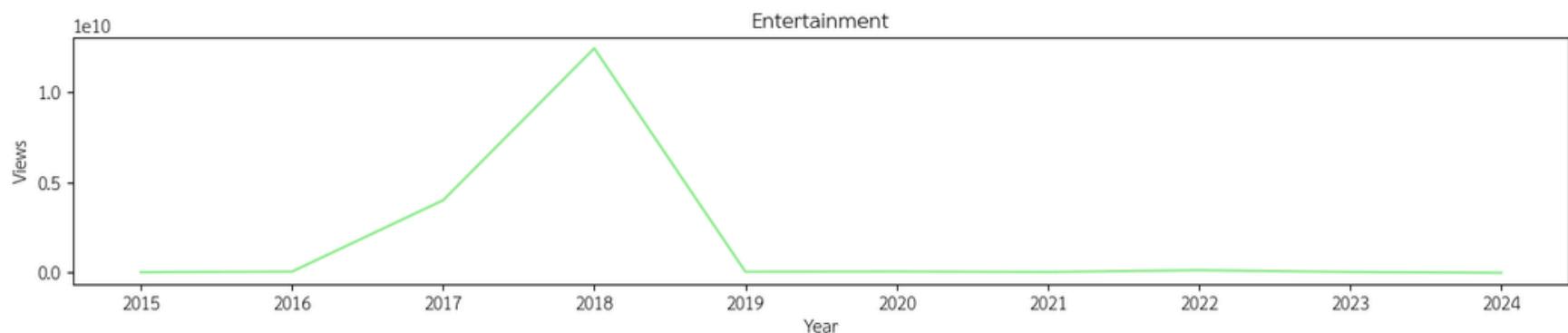
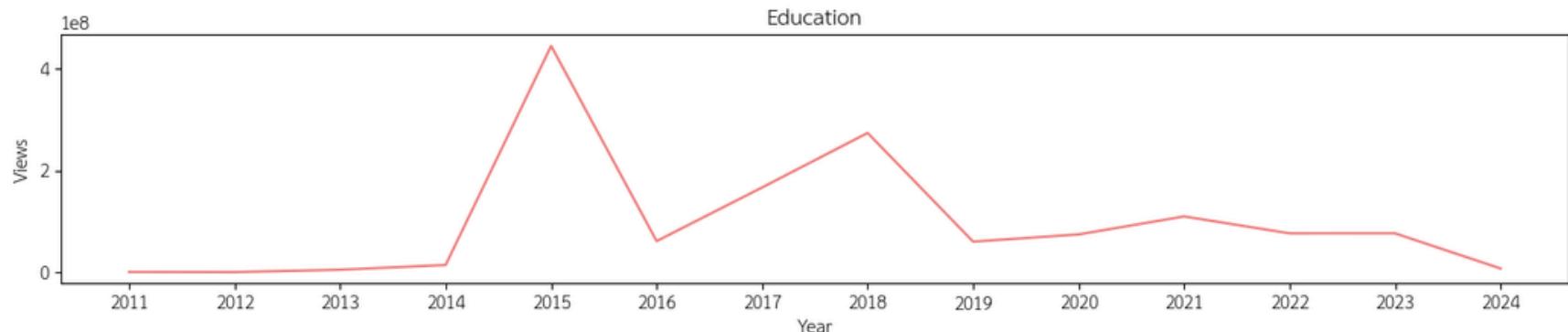


พบว่ายอดการรับชมวีดีโอหมวดหมู่ บันเทิง สูงกว่าหมวดหมู่การศึกษา

DATA ANALYSIS

กราฟเส้นเปรียบเทียบจำนวนวันยอดการรับชมของวีดีโອที่ปล่อยในแต่ละปี

comparing views of videos published per year



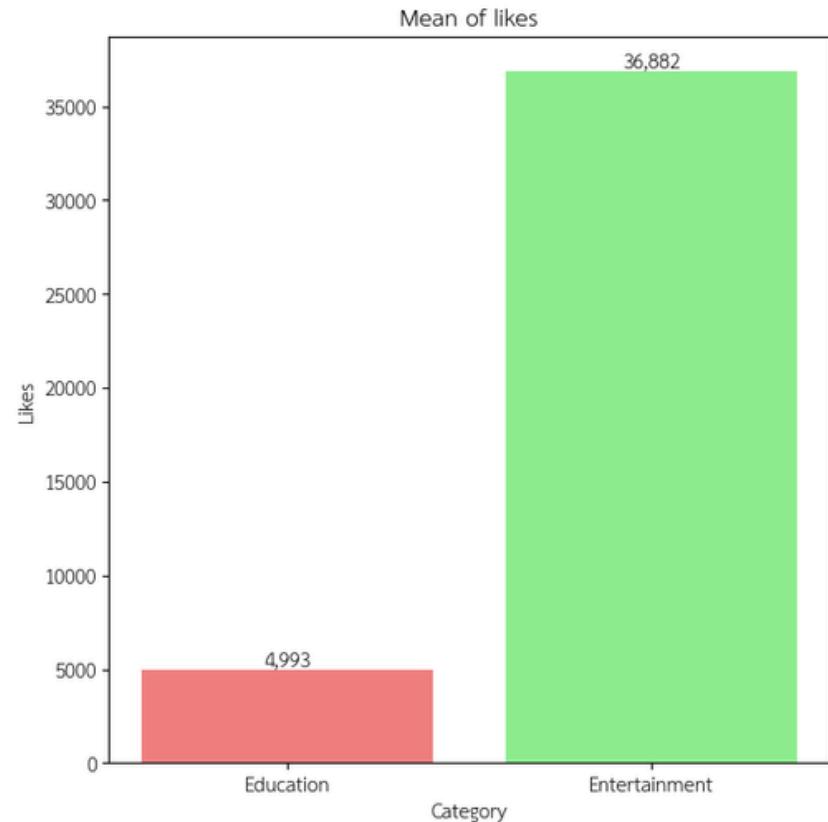
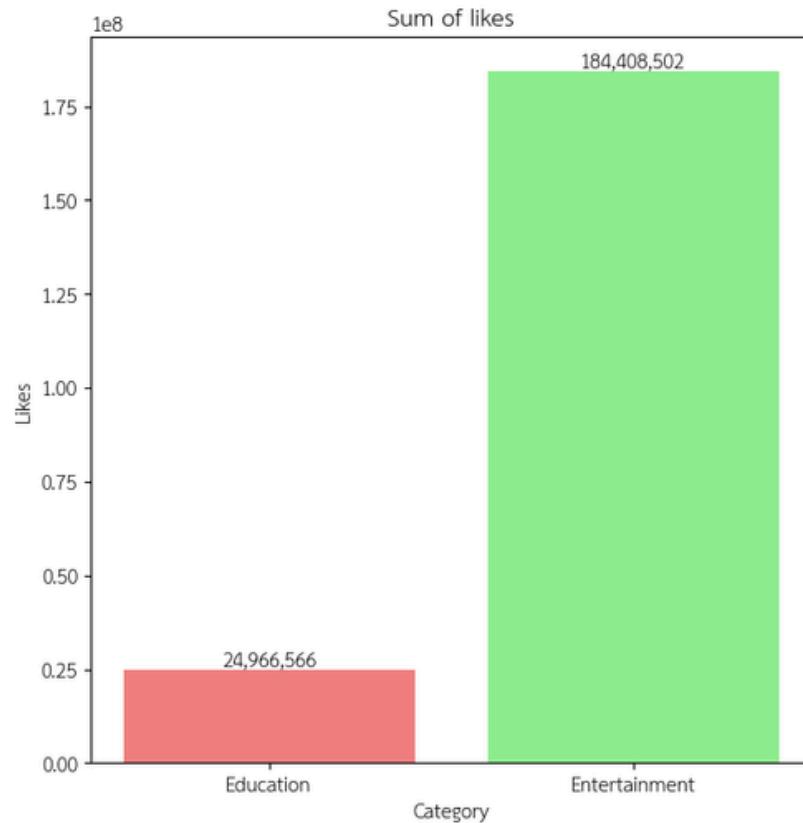
วีดีโอดหมวดหมู่การศึกษามียอดการรับชมมากที่สุดในปี 2015 รองลงมาคือ 2017-2018

วีดีโอดหมวดหมู่บันเทิงมียอดการรับชมมากที่สุดในปี 2016 - 2018

DATA ANALYSIS

กราฟแท่งแสดงผลรวมและเฉลี่ยของการกดถูกใจในวีดีโอทั้ง 2 ประเภท

likes of video

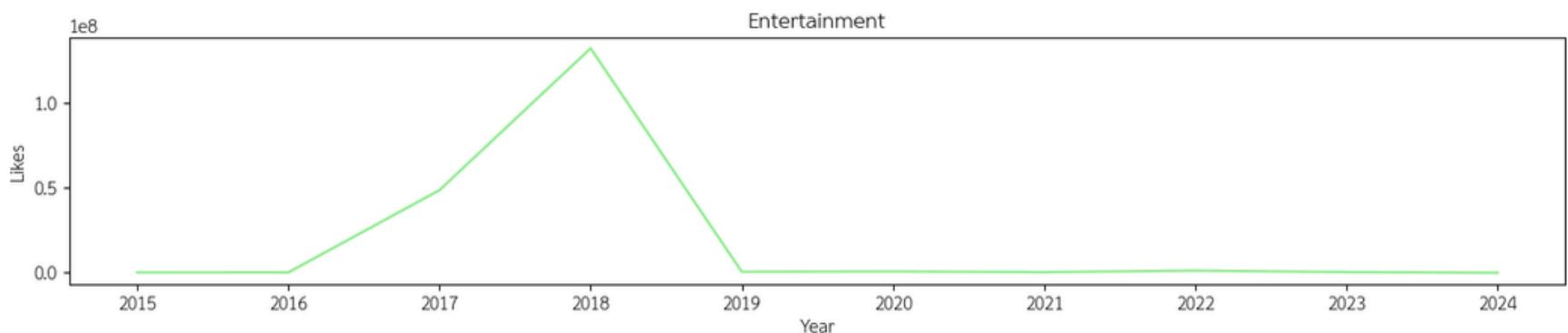
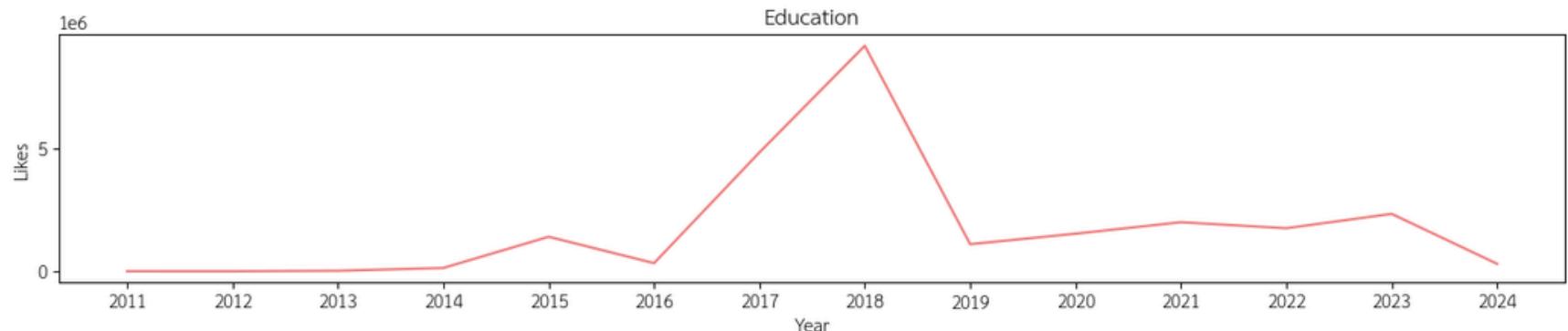


พบว่าวีดีโอมีจำนวนการกดถูกใจน้อยกว่าวีดีโอมีบันเทิง
ค่าเฉลี่ยอยู่ที่ 4,993 และ 36,882 ตามลำดับ

DATA ANALYSIS

กราฟเส้นเปรียบเทียบจำนวนการกดถูกใจของวีดีโອ่ป์ล้อยในแต่ละปี

comparing likes of videos published per year



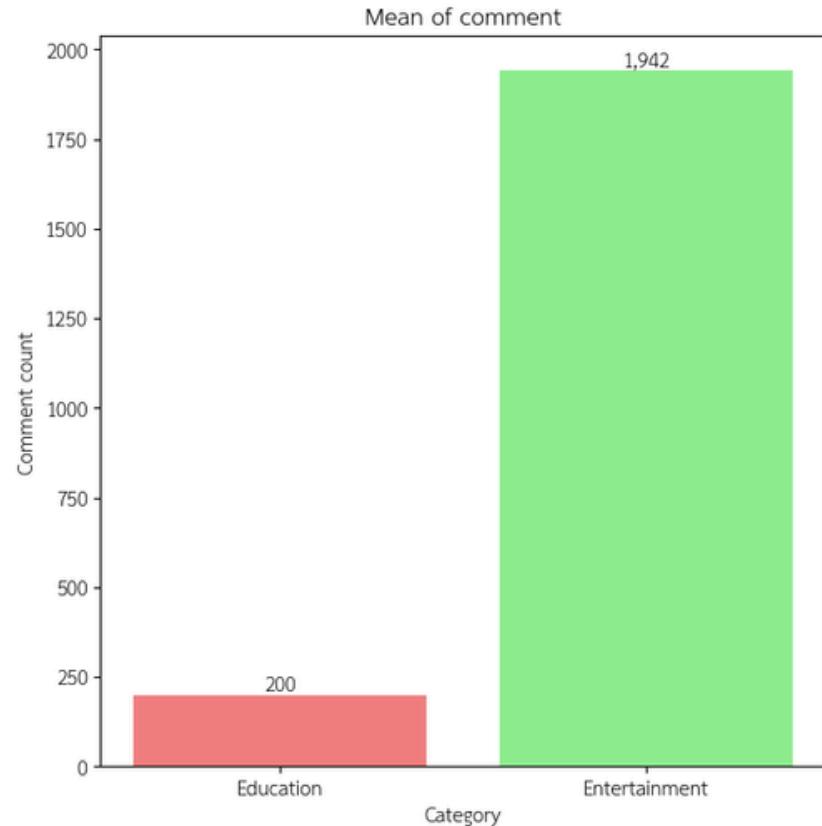
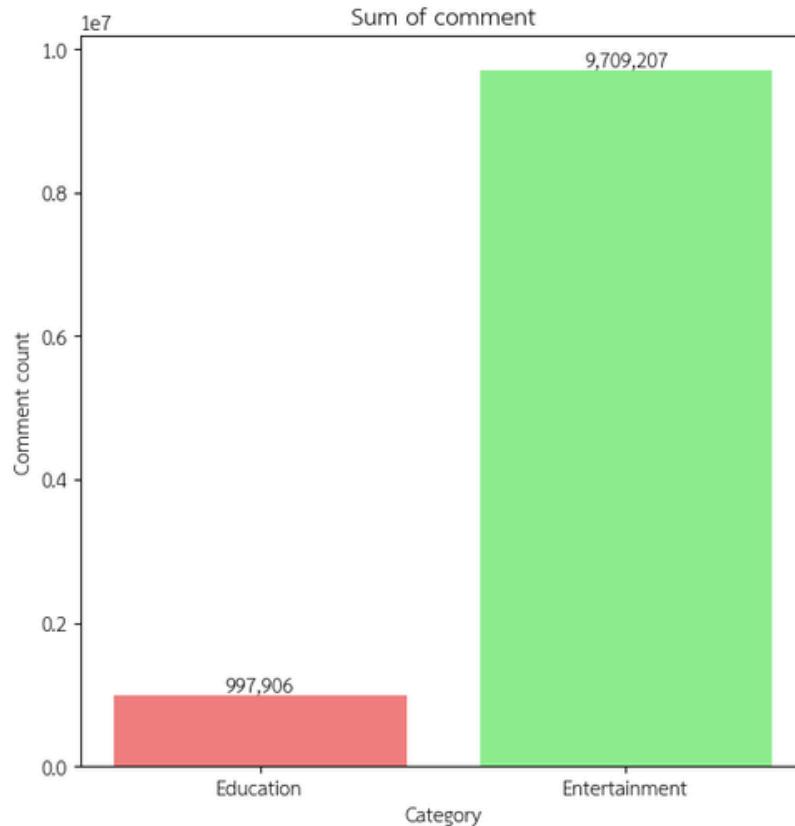
พบว่าวีดีโอมนวดหมู่การศึกษามียอดการกดถูกใจมากที่สุดในปี 2017 - 2018

วีดีโอมนวดหมู่บันเทิงมียอดการกดถูกใจมากที่สุดในปี 2017 - 2018 เช่นเดียวกัน

DATA ANALYSIS

กราฟแท่งแสดงผลรวมและเฉลี่ยของการแสดงความคิดเห็นในวีดีโอทั้ง 2 ประเภท

Comment count of video

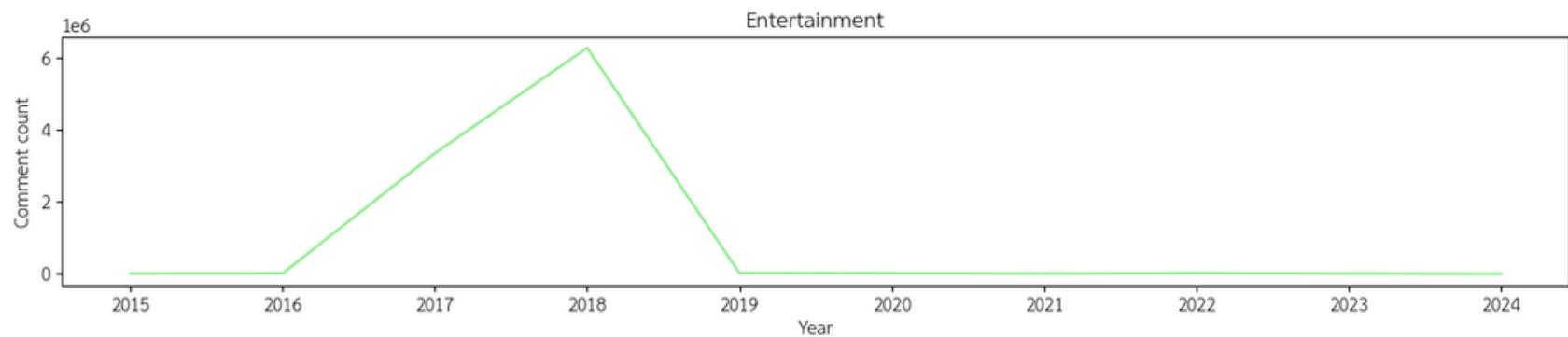
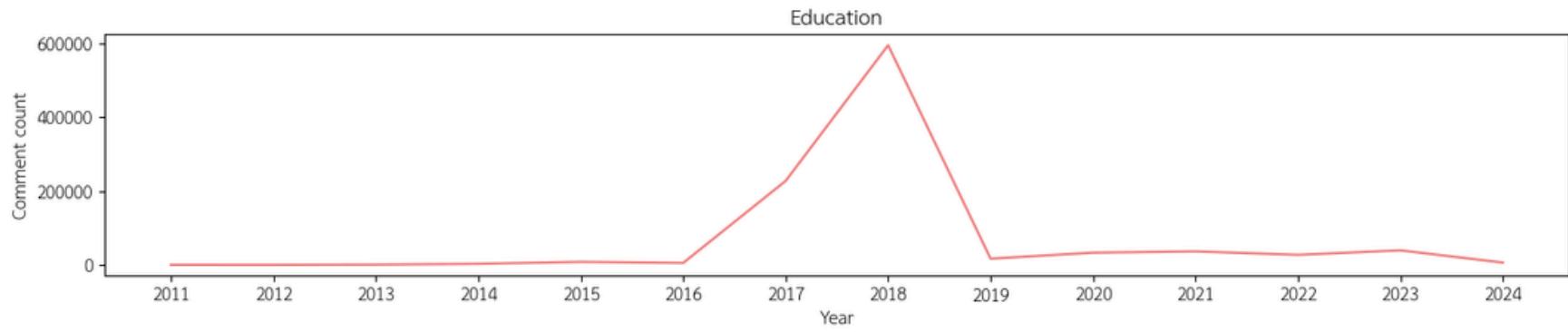


พบว่าวีดีโอมีหมวดหมู่การศึกษามียอดการแสดงความคิดเห็นน้อยกว่าวีดีโอมีหมวดหมู่บันเทิง
ค่าเฉลี่ยอยู่ที่ 200 และ 1,942 ตามลำดับ

DATA ANALYSIS

графเส้นเปรียบเทียบจำนวนการแสดงความคิดเห็นของวีดีโອ่ป์ในแต่ละปี

comparing comment count of videos published per year



พบว่าวีดีโอมีความนิยมมากที่สุดในปี 2017 - 2018
และ วีดีโอมีความนิยมมากที่สุดในปี 2017 - 2018 เช่นเดียวกัน

สรุปผล

01

เวลาช่วงเที่ยง เป็นช่วงเวลาที่มีการลงทะเบียนโอนมากที่สุด

02

ความนิยมของวีดีโอประเภทความบันเทิงสูงกว่าวีดีโอการศึกษามาก ทั้งในด้าน
ของจำนวนการรับชม จำนวนในการกดถูกใจ และจำนวนในการแสดงความคิดเห็น

03

คลิปวีดีโอที่ลงในช่วงปี 2017 - 2018 เป็นช่วงที่มีจำนวนการแสดงความคิดเห็นเป็นจำนวนมาก
กึ่งที่ไม่ได้เป็นปีที่จำนวนคลิปวีดีโอสูงที่สุด

04

ความยาวของคลิปวีดีโอประเภทการศึกษา จะยาวกว่าวีดีโอประเภทความบันเทิงโดยเฉลี่ย

05

ปี 2022 เป็นปีที่มีการลงทะเบียนโอนหมวดหมู่การศึกษามากที่สุด แต่มียอดการรับชมน้อยกว่าปี
2015 ต่างจากวีดีโอหมวดหมู่บันเทิง ที่ลงวีดีโอมากที่สุดในปี 2018 และมีจำนวนการรับชมมาก
ที่สุดในปี 2018 เช่นเดียวกัน