



PHÂN TÍCH DỮ LIỆU WEB

PHÂN TÍCH DỮ LIỆU VIDEO TRENDING TIKTOK



**GVHD: ĐẶNG NHÂN CÁCH
ENTER TEAM**

ENTER TEAM



ĐINH THỊ QUỲNH NHƯ *

K184111402



NGUYỄN THỊ NGUYÊN

K184111398



NGUYỄN HẢI LY

K184111384



TRẦN ANH THƠ

K184111424



NGHIÊM THỊ CẨM THÙY

K184111425

NỘI DUNG



01

TỔNG QUAN
ĐỀ TÀI

02

CƠ SỞ LÍ THUYẾT

03

TRIỂN KHAI
PHÂN TÍCH
DỮ LIỆU

04

KẾT LUẬN VÀ
ĐÁNH GIÁ

01

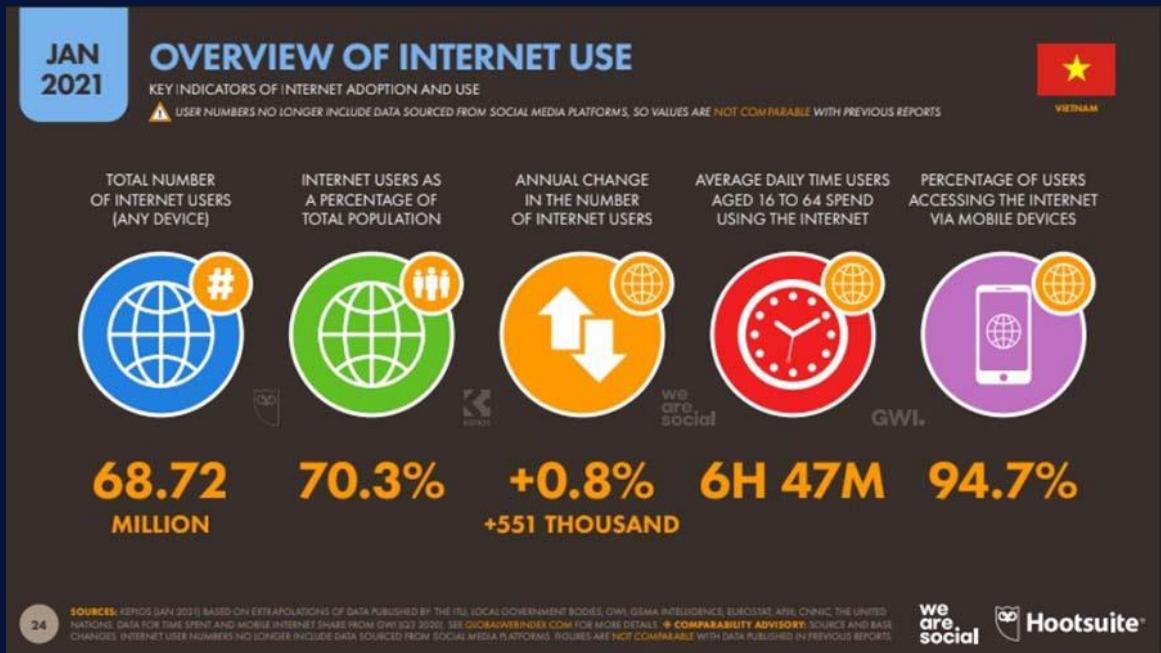
TỔNG QUAN ĐỀ TÀI



1.1 Bối cảnh

Dân số Việt Nam 97,8 triệu người (T1/2021)

- Người dùng Internet là 68,72M (70.3%.)
- Người dùng Internet tại Việt Nam tăng 551 nghìn trong giai đoạn từ 2020-2021.



Theo báo cáo Digital Vietnam in 2021

Tiềm năng Tiktok



- TikTok hiện có sẵn ở 150 quốc gia với hơn 75 ngôn ngữ (*Theo Apptrance 2019*)
- Lọt top 9 trang web MXH vượt cả Twitter, LinkedIn, Pinterest và Snapchat
- TikTok cũng trở thành ứng dụng phổ biến nhất được tải xuống trên toàn cầu năm 2020 với 850 triệu lượt tải về (*Theo Sensor Tower, 2020*)
- Tiktok vượt qua Facebook về số lượt download ở Việt Nam, 12 triệu người sử dụng (3/2020)



1.2 Mục tiêu đề tài



Áp dụng những kiến thức của môn học vào đề tài



Bộ dữ liệu đầy đủ, đáng tin cậy từ các nguồn uy tín



Đào được dữ liệu từ TikTok về để tiến hành phân tích, báo cáo

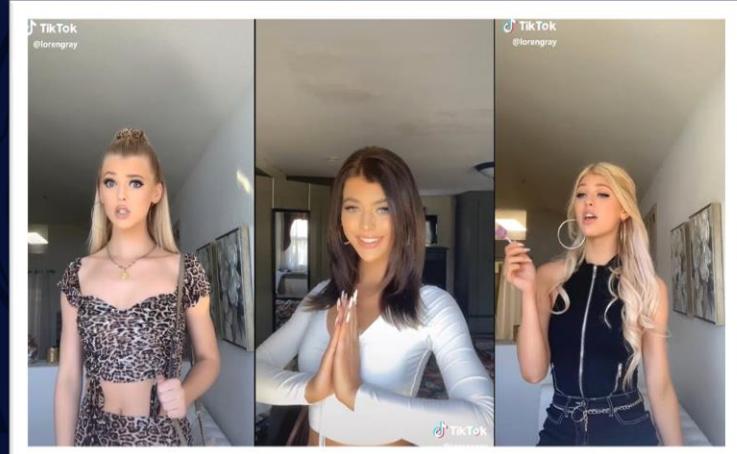
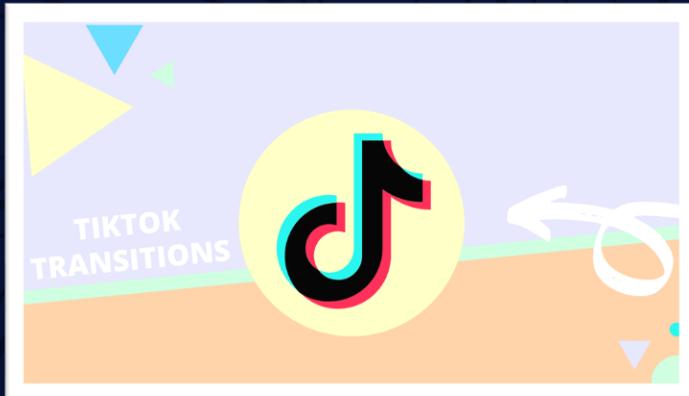


Phát triển đề tài có tính thực tiễn

1.3 Phạm vi và đối tượng nghiên cứu



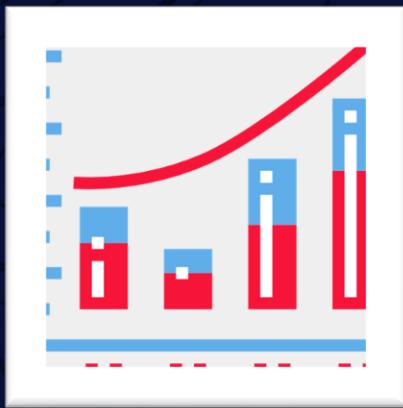
- Mạng xã hội TikTok
- 1000 Video
- 10 ngày



Các video Trending Tiktok



1.4 Phương pháp nghiên cứu



PHƯƠNG PHÁP THỐNG KÊ, SO SÁNH



PHƯƠNG PHÁP PHÂN TÍCH, TỔNG HỢP

1.5 Ý nghĩa nghiên cứu



- **ĐỐI VỚI NGƯỜI DÙNG:**

Bộ tài liệu cho các doanh nghiệp nghiên cứu, ra quyết định cho việc tiếp cận khách hàng, xây dựng chiến lược Marketing, quảng bá thương hiệu thông qua mạng xã hội Tiktok.

- **ĐỐI VỚI THÀNH VIÊN NHÓM:**

Hiểu rõ hơn về ngôn ngữ Python, về cách lấy và phân tích dữ liệu để phục vụ cho công việc sau này.

02

CƠ SỞ LÍ THUYẾT



2.1.1 Ngôn ngữ lập trình



- Python là một ngôn ngữ lập trình cấp cao, hướng đối tượng, được giải thích với ngữ nghĩa động.
- Phù hợp cho việc phát triển ứng dụng nhanh
- Cú pháp đơn giản, dễ học nhẫn mạnh khả năng đọc và do đó giảm chi phí bảo trì chương trình

2.1.2 Các thư viện



matplotlib



Seaborn



2.2 Môi trường sử dụng



PYCHARM

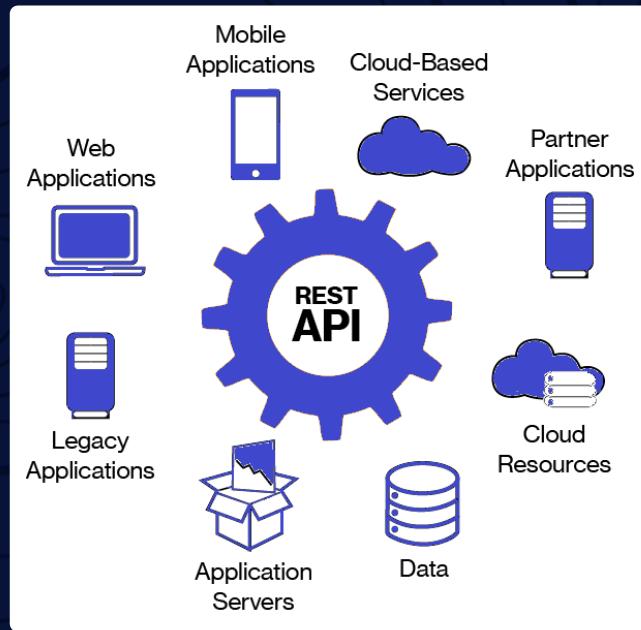


GOOGLE ANACONDA

2.3 TikTok API



RESTful API



RESTful API là một tiêu chuẩn dùng trong việc thiết kế các **API** cho các ứng dụng web để quản lý các resource.

RESTful là một trong những kiểu thiết kế API được sử dụng phổ biến ngày nay để cho các ứng dụng (web, mobile, web service...) khác nhau giao tiếp với nhau.



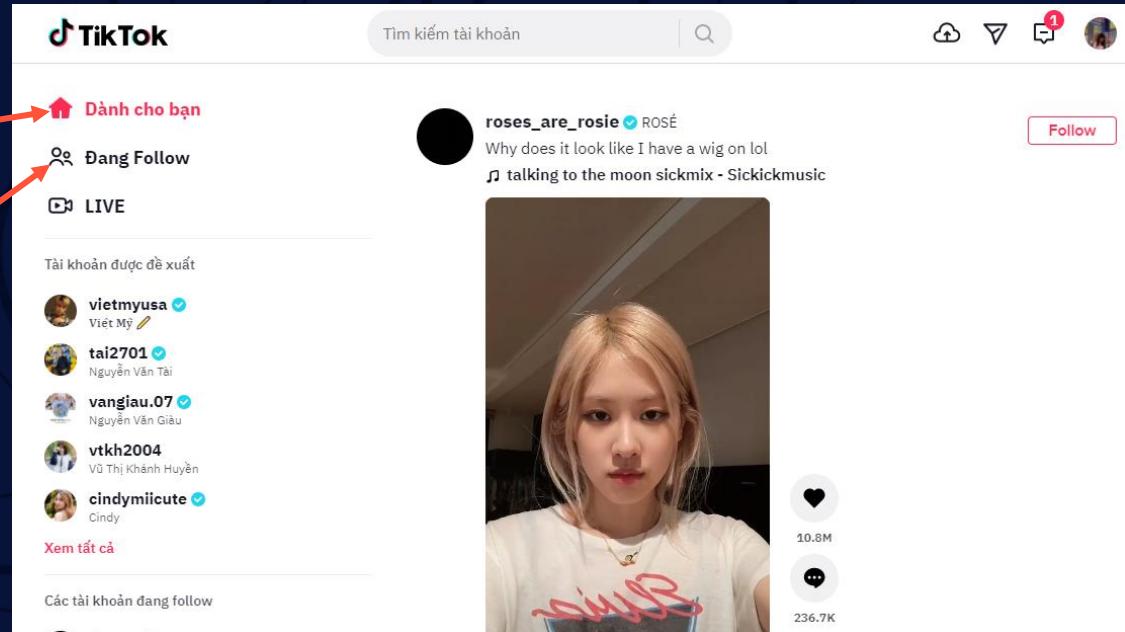
03

TRIỂN KHAI PHÂN TÍCH DỮ LIỆU

3.1 Đặt vấn đề



Dành cho bạn
Đang Follow



Nhóm tiến hành lấy 1000 video ở nguồn cấp “Dành cho bạn” trong thời gian 10 ngày để phân tích và dự đoán các đặc tính để thuật toán trí tuệ nhân tạo của TikTok chọn lựa hiển thị đề xuất trên trang chủ cho người dùng

3.2 Môi trường thực nghiệm

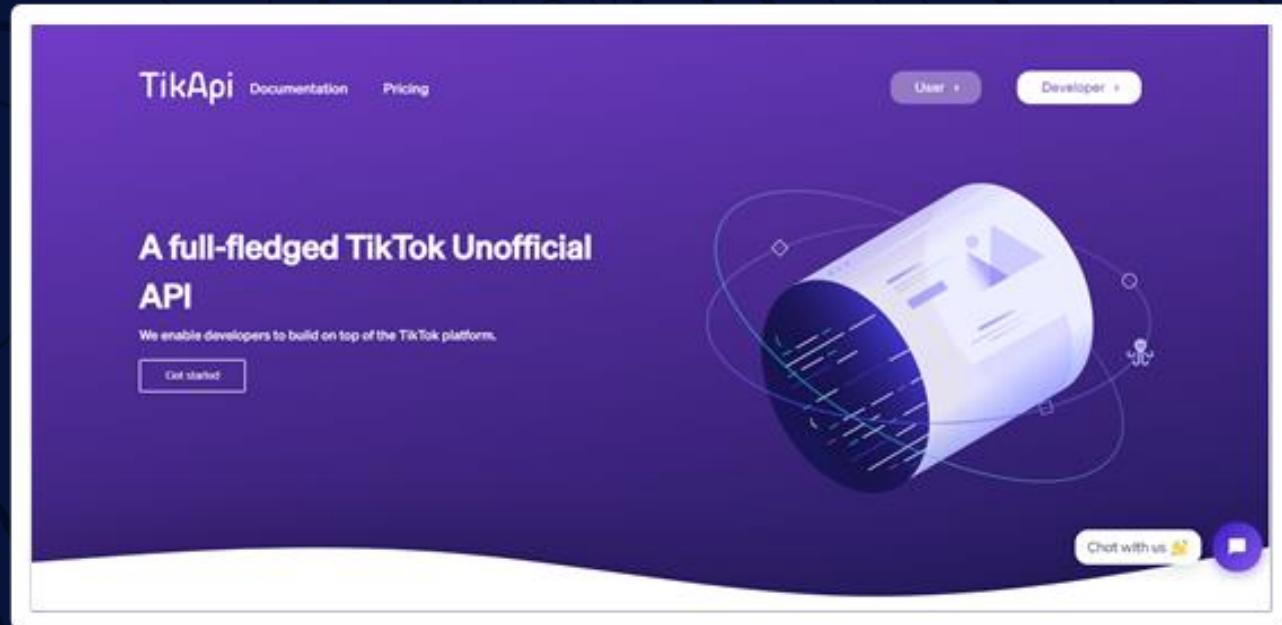


Máy tính	ASUS VivoBook S15 S530FN-BQ141T
Hệ điều hành	Windows 10 Education
Processor	Intel(R) Core (TM) i7-8565U CPU @ 1.80GHz 1.99 GHz
RAM	8.00 GB
Ngôn ngữ khai phá dữ liệu	Python
Miền dữ liệu	Dữ liệu video tiktok trending với các thông tin liên quan
Nguồn dữ liệu	https://www.tiktok.com/vi-VN/
Phần mềm sử dụng	Pycharm
Số lượng dữ liệu thu thập	1000

3.3 Quy trình thực hiện



Bước 1: Cài đặt API



TikTok Api

3.3 Quy trình thực hiện

Bước 1: Cài đặt API

```
Terminal: Local +  
Microsoft Windows [Version 10.0.19042.985]  
(c) Microsoft Corporation. All rights reserved.  
  
(venv) C:\Users\nhudt\PycharmProjects\pythonProject>pip install TikTokApi  
Requirement already satisfied: TikTokApi in c:\users\nhudt\pycharmprojects\pythonproject\venv\lib\site-packages (3.9.5)  
Requirement already satisfied: requests in c:\users\nhudt\pycharmprojects\pythonproject\venv\lib\site-packages (from TikTokApi) (2.25.1)  
Requirement already satisfied: playwright in c:\users\nhudt\pycharmprojects\pythonproject\venv\lib\site-packages (from TikTokApi) (1.10.0)  
Requirement already satisfied: selenium_stealth in c:\users\nhudt\pycharmprojects\pythonproject\venv\lib\site-packages (from TikTokApi) (1.0.6)  
Requirement already satisfied: selenium in c:\users\nhudt\pycharmprojects\pythonproject\venv\lib\site-packages (from TikTokApi) (3.141.0)  
Requirement already satisfied: greenlet==1.0.0 in c:\users\nhudt\pycharmprojects\pythonproject\venv\lib\site-packages (from playwright->TikTokApi) (1.0.0)  
Requirement already satisfied: pyee>=8.0.1 in c:\users\nhudt\pycharmprojects\pythonproject\venv\lib\site-packages (from playwright->TikTokApi) (8.1.0)  
  
TODO Problems Terminal Python Packages Python Console  
  
Terminal: Local +  
Requirement already satisfied: pyee>=8.0.1 in c:\users\nhudt\pycharmprojects\pythonproject\venv\lib\site-packages (from playwright->TikTokApi) (8.1.0)  
Requirement already satisfied: idna<3,>=2.5 in c:\users\nhudt\pycharmprojects\pythonproject\venv\lib\site-packages (from requests->TikTokApi) (2.10)  
Requirement already satisfied: certifi>=2017.4.17 in c:\users\nhudt\pycharmprojects\pythonproject\venv\lib\site-packages (from requests->TikTokApi) (2020.12.5)  
Requirement already satisfied: urllib3<1.27,>=1.21.1 in c:\users\nhudt\pycharmprojects\pythonproject\venv\lib\site-packages (from requests->TikTokApi) (1.26.4)  
Requirement already satisfied: chardet<5,>=3.0.2 in c:\users\nhudt\pycharmprojects\pythonproject\venv\lib\site-packages (from requests->TikTokApi) (4.0.0)  
WARNING: You are using pip version 21.1; however, version 21.1.1 is available.  
You should consider upgrading via the 'c:\users\nhudt\pycharmprojects\pythonproject\venv\scripts\python.exe -m pip install --upgrade pip' command.  
  
(venv) C:\Users\nhudt\PycharmProjects\pythonProject>python -m playwright install  
  
(venv) C:\Users\nhudt\PycharmProjects\pythonProject>  
  
TODO Problems Terminal Python Packages Python Console
```



3.3 Quy trình thực hiện

Bước 2: Chỉnh sửa mã để lấy video vùng Việt Nam

Nhóm nhận thấy hai biến [region] và [language] xuất hiện trong rất nhiều hàm và truy vấn

Vì vậy, tiến hành chỉnh sửa để ưu tiên lấy các video tại Việt Nam với ngôn ngữ là Tiếng Việt

```
def __extract_tag_contents(self, html):
    nonce_start = '<head nonce="'
    nonce_end = '">'
    nonce = html.split(nonce_start)[1].split(nonce_end)[0]
    j_raw = html.split(
        '<script id="__NEXT_DATA__" type="application/json" nonce="%s" crossorigin="anonymous">' % nonce
    )[1].split("</script>")[0]
    return j_raw

# Process the kwargs
def __process_kwargs__(self, kwargs):
    region = kwargs.get("region", "VN")
    language = kwargs.get("language", "vi")
    proxy = kwargs.get("proxy", None)
    maxCount = kwargs.get("maxCount", 35)

    if kwargs.get("custom_did", None) != None:
        did = kwargs.get("custom_did")
    else:
        if self.custom_did != None:
            did = self.custom_did
        else:
            did = "".join(random.choice(string.digits) for num in range(19))
    return region, language, proxy, maxCount, did
```

3.3 Quy trình thực hiện



Bước 3: Crawl video Tiktok

```
1 from TikTokApi import TikTokApi #từ thư viện Tiktok API lấy ra đối tượng Tiktok API
2
3 import pprint
4
5
6 api = TikTokApi.get_instance() #Khai báo API
7
8
9 results = 100 #Lưu số video cần lấy
10
11
12 trending = api.trending(count=results, custom_verifyFp="") #Lưu tất cả các json crawl về được, với mỗi json biểu diễn cho 1 video
13 dict_out={} #lưu file json là 1 từ điển chứa các file json nhỏ
14
15 for tiktok in trending:
16     video_id = tiktok['id'] #gọi key (id)
17     dict_out[video_id] = tiktok
18
19
20 with open('raw_data_day1.json', 'w', encoding='UTF8') as f:
21     json.dump(dict_out, f, ensure_ascii=False, indent=4)
```

Mỗi video được lưu trong biến tiktok dưới dạng 1 file json con. Để có thể lưu tất cả dữ liệu của tất cả video crawl được thành chung một file json lớn, tạo một từ điển (dictionary) dict_out lớn chứa các file json nhỏ.

3.3 Quy trình thực hiện



Bước 3: Crawl video TikTok

The screenshot shows the PyCharm IDE interface with the following details:

- Project Structure:** The project is named "pythonProject" located at "C:\Users\nhadt\PycharmProjects\pythonProject". It contains a "venv" folder, a "library root" folder, and several Python files: "convert_format_trending.py", "count_data.py", "crawled_data.csv", "crawled_data.json", "data_1week.json", "data_10days.json", "data_10days_test.json", "data_day2.json", "data_day3.json", "data_day4.json", "data_day5.json", "data_day6.json", "data_day7.json", "data_day8.json", "data_day9.json", "data_day10.json", "data_tiktok_10days.json", "raw_data_10days.json", "raw_data_day1.json", "raw_data_day2.json", and "raw_data_day3.json".
- Code Editor:** The code in "crawl_tiktok.py" is as follows:

```
from TikTokApi import TikTokApi #từ thư viện Tiktok API lấy ra đối tượng Tiktok API
import pprint
import json

api = TikTokApi.get_instance() #Khai báo API

results = 100 #Lưu số video cần lấy
trending = api.trending(count=results, custom_verifyFp="") #Lưu tất cả các json crawl về được, với mỗi json biểu diễn cho 1 video

dict_out={} #lưu file json là 1 từ điển chứa các file json nhỏ

for tiktok in trending:
    video_id = tiktok['id'] #goi key (id)
    dict_out[video_id] = tiktok

print(tiktok.keys())
```

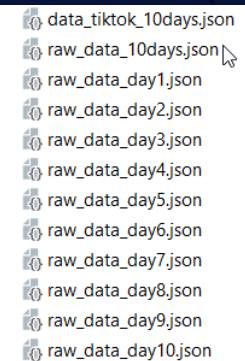
- Run Tab:** The run configuration is set to "crawl_tiktok" with the command "C:\Users\nhadt\PycharmProjects\pythonProject\venv\Scripts\python.exe C:\Users\nhadt\PycharmProjects\pythonProject\crawl_tiktok.py". The output shows the keys of the first TikTok object.
- Output:** The terminal output is:

```
C:\Users\nhadt\PycharmProjects\pythonProject\venv\Scripts\python.exe C:\Users\nhadt\PycharmProjects\pythonProject\crawl_tiktok.py
dict_keys(['id', 'desc', 'createTime', 'video', 'author', 'music', 'challenges', 'stats', 'duetInfo', 'originalItem', 'officialItem', 'textExtra', 'secret', 'forFriend', 'digged', 'itemCommentStatus', 'showNotPass', 'v11', 'itemMute', 'authorStats', 'privateItem', 'duetEnabled', 'stitchEnabled', 'shareEnabled', 'isAd'])
```

Process finished with exit code 0

3.3 Quy trình thực hiện

Bước 4: Định dạng cấu trúc dữ liệu



Dữ liệu thu được gồm 10 file json
raw_data được crawl trong 10 ngày với
số lượng mỗi ngày là 100 video và tổng
số video là 1000. Sau khi có được tập
dữ liệu thô, định dạng lại cấu trúc phù
hợp với hướng phân tích

```
convert_format_trending.py
```

```
import json

json_input = './raw_data_day10.json'
with open(json_input, 'r', encoding='utf8') as f:
    data = json.load(f)

list_video = []

for key in data:
    curr_vid = data[key]
    after_format = {}
    after_format['id'] = curr_vid['id']
    after_format['text'] = curr_vid['desc']
    after_format['createTime'] = curr_vid['createTime']
    after_format['authorMeta'] = {
        "id": curr_vid['author']['id'],
        "secUid": curr_vid['author']['secUid'],
        "name": curr_vid['author']['nickname'],
        "nickName": curr_vid['author']['nickname'],
        "verified": curr_vid['author']['verified'],
        "signature": curr_vid['author']['signature'],
    }
    list_video.append(after_format)
```



3.3 Quy trình thực hiện

Bước 4: Định dạng cấu trúc dữ liệu

Định dạng lại file json thành 1 từ điển sử dụng 1 key **collector** chứa list các video đã crawl

```
[{"collector": [ { "id": "6950151666976951557", "text": "Follow my YouTube channel #foodgood", "createTime": 1618208297, "authorMeta": { "id": "6812963291129562118", "secUid": "MS4wLjQABAAAABbfY3StRRWK8Jrs87MiPkgp9CulzWVnykeiFNYlSU_uAjrmavWhBhhfuA_N_3FhL", "name": "Khalid El Mahi", "nickName": "Khalid El Mahi", "verified": false, "signature": "CEO of BISMILLAH\nYouTube ➡️ foodgood", "avatar": "https://p16-sign-va.tiktokcdn.com/tos-maliva-avt-0068/1706de6dc569856c676ae0f5c45d0aa7~c5_1080x1080.jpeg?x-expires=1620280800&x-signature=LbQsaf75xlsBrA9cn5111rvpr0I%3D" }, "musicMeta": { "musicId": "6669854674113317638", "musicName": "BREAKFAST CHALLENGE", "musicAuthor": "Spence", "musicOriginal": false, "playUrl": "https://sf16-ies-music-sg.tiktokcdn.com/obj/tiktok-obj/f9e5e0793535f51c2dea59a06b7b30bc.mp3", "coverThumb": "https://p9-sg.tiktokcdn.com/aweme/100x100/tos-alisg-i-0000/cce7511aaafb41d595f56d5f7c677514.jpeg", "coverMedium": "https://p9-sg.tiktokcdn.com/aweme/200x200/tos-alisg-i-0000/cce7511aaafb41d595f56d5f7c677514.jpeg", "coverLarge": "https://p9-sg.tiktokcdn.com/aweme/720x720/tos-alisg-i-0000/cce7511aaafb41d595f56d5f7c677514.jpeg" }, "covers": { "default": "https://p16-sign-va.tiktokcdn.com/obj/tos-maliva-p-0068/dda3d8fa1394402e8b340b0acf0d14b5_1618208298?x-expires=1620216000&x-signature=JI2vPiESKbpR6xDdIdV3IA6mWxs%3D", "origin": "https://p16-sign-va.tiktokcdn.com/obj/tos-maliva-p-0068/47132d605c0f4b6c936955dbf0b5b3bd_1618208299?x-expires=1620216000&x-signature=2I2FQn5bABAf%2FTdE9KSGnCBYT1Xk%3D", "dynamic": "https://p16-sign-va.tiktokcdn.com/obj/tos-maliva-p-0068/41ee40fc473c483d88e2febfb9c632164_1618208298?x-expires=1620216000&x-signature=%2BZ1013LkaW8cFv8RNb91m7RzU83D" }, "webVideoUrl": "https://v16.tiktokcdn.com/1d8262c43ce9c8ad1706da9dfd97e6bf/609289e8/video/tos/useast2a/tos-useast2a-ve-0068c001/b62c2934c60b4b7980ea2903ed38e086/?" } ] }
```



3.3 Quy trình thực hiện

Bước 4: Định dạng cấu trúc dữ liệu

Giới thiệu cấu trúc dữ liệu

Tên trường	Ý nghĩa
<u>id</u>	Số nhận dạng duy nhất của video
<u>text</u>	Văn bản bên dưới của video
<u>createTime</u>	Dấu thời gian của ngày giờ khi video được tạo
<u>authorMeta</u>	Thông tin chi tiết về tác giả
<u>musicMeta</u>	Thông tin chi tiết về âm nhạc được sử dụng với video
<u>covers</u>	Một đối tượng chứa tất cả các cover của video
<u>webVideoUrl</u>	Liên kết đến video TikTok



3.3 Quy trình thực hiện

Bước 4: Định dạng cấu trúc dữ liệu

Giới thiệu cấu trúc dữ liệu

Tên trường	Ý nghĩa
<u>videoUrl</u>	Liên kết chính xác đến video TikTok (không thể truy cập trực tiếp)
<u>videoUrlNoWaterMark</u>	URL của video không có hình mờ
<u>videoMeta</u>	Một đối tượng chứa kích thước và thời lượng của video
<u>diggCount</u>	Lượng thích
<u>shareCount</u>	Video đã được chia sẻ bao nhiêu lần
<u>playCount</u>	Số lần video đã được xem
<u>commentCount</u>	Lượng bình luận
<u>hashtags</u>	Danh sách các thẻ bắt đầu bằng # được sử dụng trong video

3.4 Phân tích dữ liệu và làm sạch sơ bộ



Bước 1: Ta tiến hành cài đặt các thư viện cần thiết để thực hiện đề tài như: Pandas, Numpy, Matplotlib, Seaborn, Os, Json v.v

```
import os
import json
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.express as px
import plotly.graph_objects as go
from plotly.offline import init_notebook_mode, iplot, plot
from plotly.subplots import make_subplots
```

Bước 2: Đọc dữ liệu từ file json và đưa vào Dataframe của Pandas

```
[ ] file = open('data_tiktok_10days.json', encoding="utf8")
raw_data = json.load(file)
file.close()
```

```
len(raw_data['collector'])
```

```
1000
```

Dòng dữ liệu 1000 dòng

3.4 Phân tích dữ liệu và làm sạch sơ bộ



Bước 3: Tạo cấu trúc dữ liệu dạng DataFrame

```
trending_videos_list = raw_data['collector']
```

Khởi tạo biến trending_videos_list chứa list collector các video

```
df_tiktok_dataset = pd.DataFrame(trending_videos_list)
```

Tạo DataFrame

	id	text	createTime	authorMeta	musicMeta	covers
0	6950151666978951557	Follow my YouTube channel #foodqood	1618208297	{'id': '6812963291129562118', 'secUid': 'MS4wL...'}	{'musicId': '6669854674113317638', 'musicName':...}	{'default': 'https://p16-sign-va.tiktokcdn.com...}
1	6958464277044792577	👉	1620143720	{'id': '6626092899037134849', 'secUid': 'MS4wL...'}	{'musicId': '6586947002464996102', 'musicName':...}	{'default': 'https://p16-sign-sg.tiktokcdn.com...}
2	6957312513742540033	Makeup đã giúp thăng hạng nhan sắc thế nào?#Ch...	1619875553	{'id': '6924800214197240833', 'secUid': 'MS4wL...'}	{'musicId': '6777243018756884481', 'musicName':...}	{'default': 'https://p16-sign-sg.tiktokcdn.com...}
3	6955841401103994118	Fratello la semplicità è sempre una buona scel...	1619533035	{'id': '127905465618821121', 'secUid': 'MS4wLj...'}	{'musicId': '6955841269407042309', 'musicName':...}	{'default': 'https://p16-sign-va.tiktokcdn.com...}
4	6949465089267289346	👉👉 Watch Full Video 👉👉 #experiment #VS #car ...	1618048432	{'id': '6940526321517446149', 'secUid': 'MS4wL...'}	{'musicId': '6949465067868015361', 'musicName':...}	{'default': 'https://p16-sign-sg.tiktokcdn.com...}
...
995	6960521749687586053	I have a treat today.please be happy.ok ?#gest...	1620622768	{'id': '6960489852706767877', 'secUid': 'MS4wL...'}	{'musicId': '6960521396153764614', 'musicName':...}	{'default': 'https://p16-sign-va.tiktokcdn.com...}
996	6957311085053545730	hỏi đáp tâm sự đி các bạn oiiii Nhớ follow và ...	1619875220	{'id': '6557557841481121794', 'secUid': 'MS4wL...'}	{'musicId': '6952836017820928770', 'musicName':...}	{'default': 'https://p16-sign-sg.tiktokcdn.com...}
997	6961390399755685121	Hành trình kiếm tìm niềm vui #fyp #tiktokvn #fo...	1620825010	{'id': '6854977182419698689', 'secUid': 'MS4wL...'}	{'musicId': '6956990112127585029', 'musicName':...}	{'default': 'https://p16-sign-sg.tiktokcdn.com...}
998	6951366101380893957	Easy	1618491048	{'id': '127905465618821121', 'secUid': 'MS4wLj...'}	{'musicId': '6951365854600776453', 'musicName':...}	{'default': 'https://p16-sign-va.tiktokcdn.com...}
999	695232106405568450	Nếu không biết đau thì không thể đứng lên khi ...	1618713392	{'id': '6713690382875952130', 'secUid': 'MS4wL...'}	{'musicId': '6941381532636465922', 'musicName':...}	{'default': 'https://p16-sign-sg.tiktokcdn.com...}

1000 rows x 14 columns



3.4 Phân tích dữ liệu và làm sạch sơ bộ

Bước 3: Tạo cấu trúc dữ liệu dạng DataFrame

DataFrame

webVideoUrl	videoUrl	videoMeta	diggCount	shareCount	playCount	commentCount	hashtags
https://v16.tiktokcdn.com/1d8262c43ce9c8ad1706...	https://v16.tiktokcdn.com/1d8262c43ce9c8ad1706...	{"height": 1024, "width": 576, "duration": 13}	4600000	319000	63500000	24300	[foodgood]
https://v9-vn.tiktokcdn.com/ea51f030882c5b6d54...	https://v9-vn.tiktokcdn.com/ea51f030882c5b6d54...	{"height": 540, "width": 960, "duration": 11}	6349	40	254600	59	[]
https://v9-vn.tiktokcdn.com/bfe5c3b81e38c7616b...	https://v9-vn.tiktokcdn.com/bfe5c3b81e38c7616b...	{"height": 960, "width": 540, "duration": 95}	11000	52	268500	49	[changminmakeup, goclamedep]
https://v16.tiktokcdn.com/04c786e6c2ed7f92eba9...	https://v16.tiktokcdn.com/04c786e6c2ed7f92eba9...	{"height": 1024, "width": 576, "duration": 28}	8300000	55500	77800000	60700	[learnwithtiktok, learnfromkhabi]
https://v9-vn.tiktokcdn.com/37e738fd71678ea367...	https://v9-vn.tiktokcdn.com/37e738fd71678ea367...	{"height": 1024, "width": 576, "duration": 28}	63600	86	1300000	185	[experiment, vs, car, crunchy, crushing, crush...]
...
https://v16.tiktokcdn.com/4514408fe4ab8ace967b...	https://v16.tiktokcdn.com/4514408fe4ab8ace967b...	{"height": 1024, "width": 576, "duration": 23}	76000	1625	346100	703	[gesture, gesturedance, dancer, happy]
https://v16.tiktokcdn.com/c060cf0719d3798d84ef...	https://v16.tiktokcdn.com/c060cf0719d3798d84ef...	{"height": 1024, "width": 576, "duration": 30}	24400	54	458100	270	[]
https://v16.tiktokcdn.com/14aae0f6bb7e49b6e681...	https://v16.tiktokcdn.com/14aae0f6bb7e49b6e681...	{"height": 960, "width": 540, "duration": 12}	55600	1475	679000	502	[fyp, tiktokvn, foryou]
https://v16.tiktokcdn.com/6aa3b2faf59289e48fad...	https://v16.tiktokcdn.com/6aa3b2faf59289e48fad...	{"height": 1024, "width": 576, "duration": 25}	9400000	100800	89500000	66900	[]
https://v16.tiktokcdn.com/da8e049595c7a37dbfeb...	https://v16.tiktokcdn.com/da8e049595c7a37dbfeb...	{"height": 854, "width": 480, "duration": 11}	12200	67	120600	0	[]

Có thể thấy một số trường dữ liệu đang chứa các từ điển nhỏ hơn bên trong nó: authorMeta, musicMeta, cover, videoMeta. Ta tiến hành mở rộng các trường này thành các cột riêng biệt và xóa đi các cột gốc ban đầu chứa từ điển để có được cấu trúc dữ liệu chi tiết hơn.



3.4 Phân tích dữ liệu và làm sạch sơ bộ

Bước 3: Tạo cấu trúc dữ liệu dạng DataFrame

DataFrame

	id	text	createTime	webVideoUrl	videoUrl	diggCount	shareCount	playCount	commentCount
0	6950151666976951557	Follow my YouTube channel #foodgood	1618208297	https://v16.tiktokcdn.com/1d8262c43ce9c8ad1706...	https://v16.tiktokcdn.com/1d8262c43ce9c8ad1706...	4600000	319000	63500000	24300
1	6958464277044792577	🟡	1620143720	https://v9-vn.tiktokcdn.com/ea51f030882c5b6d54...	https://v9-vn.tiktokcdn.com/ea51f030882c5b6d54...	6349	40	254600	59
2	6957312513742540033	Makeup đã giúp tháng hang nhận sác thê nào?">#Ch...	1619875553	https://v9-vn.tiktokcdn.com/bfe5c3b81e38c7616b...	https://v9-vn.tiktokcdn.com/bfe5c3b81e38c7616b...	11000	52	268500	49
3	6955841401103994118	Fratello la semplicità è sempre una buona scel...	1619533035	https://v16.tiktokcdn.com/04c786e6c2ed7f92eba9...	https://v16.tiktokcdn.com/04c786e6c2ed7f92eba9...	8300000	55500	77800000	60700

musicMeta.coverMedium	musicMeta.coverLarge	cover.default	cover.origin	cover.dynamic	videoMeta.height	videoMeta.width	videoMeta.duration
https://p9-sg.tiktokcdn.com/aweme/200x200/tos... sg.tiktokcdn.com/aweme/720x720/tos...	https://p9-sg.tiktokcdn.com/aweme/720x720/tos... sg.tiktokcdn.com/obj/tos-mali...	https://p16-sign-va.tiktokcdn.com/obj/tos-mali...	https://p16-sign-va.tiktokcdn.com/obj/tos-mali...	https://p16-sign-va.tiktokcdn.com/obj/tos-mali...	1024	576	13
https://p9-sg.tiktokcdn.com/aweme/200x200/tos... sg.tiktokcdn.com/aweme/720x720/tos...	https://p9-sg.tiktokcdn.com/aweme/720x720/tos... sg.tiktokcdn.com/obj/tos-mali...	https://p16-sign-sg.tiktokcdn.com/obj/tos-mali...	https://p16-sign-sg.tiktokcdn.com/obj/tos-mali...	https://p16-sign-sg.tiktokcdn.com/obj/tos-mali...	540	960	11
https://p9-sg.tiktokcdn.com/aweme/200x200/tik... sg.tiktokcdn.com/aweme/720x720/tik...	https://p9-sg.tiktokcdn.com/aweme/720x720/tik... sg.tiktokcdn.com/tos-alis-p...	https://p16-sign-sg.tiktokcdn.com/tos-alis-p...	https://p16-sign-sg.tiktokcdn.com/obj/tos-mali...	https://p16-sign-sg.tiktokcdn.com/obj/tos-mali...	960	540	95
https://p16-sign-va.tiktokcdn.com/tos-mali... va.tiktokcdn.com/obj/tos-mali-a...	https://p16-sign-va.tiktokcdn.com/tos-mali-a... va.tiktokcdn.com/obj/tos-mali...	https://p16-sign-va.tiktokcdn.com/obj/tos-mali...	https://p16-sign-va.tiktokcdn.com/obj/tos-mali...	https://p16-sign-va.tiktokcdn.com/obj/tos-mali...	1024	576	28

3.4 Phân tích dữ liệu và làm sạch sơ bộ



Bước 3: Tạo cấu trúc dữ liệu dạng DataFrame

1000 rows × 31 columns

DataFrame

hashtags	authorMeta.id	authorMeta.secUid	authorMeta.name	authorMeta.nickName	authorMeta.verified	authorMeta.signature
[foodgood]	6812963291129562118	MS4wLjABAAAAAbfY3StRRWK8JrS87MiPkgp9CulzWVnyke...	Khalid El Mahi	Khalid El Mahi	False	CEO of BISMILLAHI YouTube foodgood
[]	6626092899037134849	MS4wLjABAAAALR89pAaBQRmINLvYP5r839aFKlLgREIXcV...	Trung Quang	Trung Quang	False	
[changminmakeup, goclanddep]	6924800214197240833	MS4wLjABAAAA4Sg2sJphaJXN9vw79W63ZyCc-IJWTmgf6a...	Changminmakeup	Changminmakeup	False	For work👉 IG: Changminmakeup Cá sán phẩm MAKE...
[learnwithtiktok, learnfromkhabi]	127905465618821121	MS4wLjABAAA AwAg0rSzO65WQfz4RzQgGv2Xdv108BgPXhR...	Khabane lame	Khabane lame	True	Se volete ridere siete nel posto giusto 😊SNIT

authorMeta.avatar	musicMeta.musicId	musicMeta.musicName	musicMeta.musicAuthor	musicMeta.musicOriginal	musicMeta.playUrl	musicMeta.coverThumb
https://p16-sign-va.tiktokcdn.com/tos-maliva-a...	6669854674113317638	BREAKFAST CHALLENGE	Spence	False	https://sf16-ies-music.tiktokcdn.com/obj/tik...	https://p9-sign-va.tiktokcdn.com/aweme/100x100/tos...
https://p16-sign-sg.tiktokcdn.com/aweme/1080x1...	6586947002464996102	Oh No	Kreepa	False	https://sf9-ies-music.tiktokcdn.com/obj/tik...	https://p9-sign-va.tiktokcdn.com/aweme/100x100/tos...
https://p16-sign-sg.tiktokcdn.com/aweme/1080x1...	6777243018756884481	Waiting For Heartache	BLVKSHP	False	https://sf9-ies-music.tiktokcdn.com/obj/tik...	https://p9-sign-va.tiktokcdn.com/aweme/100x100/tik...
https://p16-sign-va.tiktokcdn.com/tos-maliva-a...	6955841269407042309	suono originale	Khabane lame	True	https://sf9-ies-music.tiktokcdn.com/obj/mus...	https://p77-sign-va.tiktokcdn.com/tos-maliva-a...



3.5 Phân tích cụ thể và trực quan hóa

3.5.1 Phân tích dữ liệu trùng lặp

```
count_duplicated = pd.crosstab(index=df_tiktok_dataset['id'], columns='duplicates')  
count_duplicated
```

KẾT QUẢ

Trong số **1000 video** thu thập được, chỉ có **578 video** là duy nhất. Vì vậy, có rất nhiều video được tiktok đề xuất để xuất hiện lặp lại nhiều lần ở top 100 trong 10 ngày mà nhóm thu thập dữ liệu. Để có thể phân tích rõ hơn, tiến hành lọc ra top 10 các video có tần suất xuất hiện nhiều nhất.

col_0	duplicates
id	
6925559746128907526	1
6925773613698256133	1
6925866034490543366	1
6926508267598531845	2
6926583800365370629	3
...	...
6961642113620593922	1
6961664991611129090	1
6961702011028950273	1
6961714462088891649	1
6961733424352709890	1
578 rows × 1 columns	

3.5.1 Phân tích dữ liệu trùng lặp

```
duplicated = pd.crosstab(index=[df_tiktok_dataset['id'], df_tiktok_dataset['text'], df_tiktok_dataset['authorMeta.name'], df_tiktok_dataset['authorMeta.verified'], df_tiktok_dataset['authorMeta.signature'], df_tiktok_dataset['musicMeta.musicName']], columns='duplicates')

df_duplicate = duplicated.sort_values(by='duplicates', ascending=False)
df_duplicate.head(10)
```

Sắp xếp và trích xuất 10 video trùng lặp nhiều nhất

							col_0 duplicates
	id	text	authorMeta.name	authorMeta.verified	authorMeta.signature	musicMeta.musicName	
6955841401103994118	Fratello la semplicità è sempre una buona scelta! 😊 Bro simplicity is always a good choice! 😊 #LearnWithTikTok #LearnFromKhabi	Khabane lame	True	Se volete ridere siete nel posto giusto 😊 SNIT	suono originale		8
6942051158437448962	Mẹ có thể đem lại hy vọng cứu rỗi cho cả những linh hồn tuyệt vọng nhất. #trang_sức_thiết_kế #xuhuongtiktok	trang sức gỗ già dá phong thủy	False	Để đặt hàng nhanh nhất bám dưới đây 📲	原聲 - 錄音吉娃娃		8
6937351021383388418	Hay cảm nhận nó đang siết chặt #tips #tiktokphilippines #funnytiktok #viral_video #khoavegas #fypdonnnnnnnn #aladintelecom #tutorials #PaperCut #wow	Khoa Vegas	False	Hello Everybody ♡ I'm Vietnamese ♡ I wish you the happy video	오리지널 사운드 - Glass(유리)		7
6944331615870995718	Easy breakfast sandwich 🍔!!! #рекомендации (via miaobianhishen douyin)	foodgod	True	My name was JONATHAN.. now it's FOODGOD (legally)	La Vie En Rose		7
6950151666976951557	Follow my YouTube channel #foodqood	Khalid El Mahi	False	CEO of BISMILLAHin YouTube foodqood	BREAKFAST CHALLENGE		7
6957703300816981253	Non c'è bisogno di Didascalia 😊 —No Caption Needed. 😊 #learnfromkhaby #LearnWithTikTok #ImparaConTikTok #nocaptionneeded	Khabane lame	True	Se volete ridere siete nel posto giusto 😊 SNIT	suono originale		7
6935150455886564610	Cá mè rô #fish #mièntây #interesting #tik Tok	Nguyễn Hạ Vi	False		nhạc nền - Nguyễn Hạ Vi		7
6956939472286043397	Khabynho du Brazil brasil Ritenta bro sarai più fortunato- Try again bro you be luckier next time. 😊 #learnfromkhaby	Khabane lame	True	Se volete ridere siete nel posto giusto 😊 SNIT	suono originale		7
6934205812751518977	#zodiacsigns #zodiac #virgo #fyp #chocolate	Jenny Nguyen	False		original sound		7
6945064874519055618	Hoi lag tý 😊 😊 😊 #funny #tiktokvn	TrungSaffron	False		nhạc nền - TrungSaffron		7

Top 10 video có tần suất lặp lại nhiều nhất

3.5.2 Biểu thị một số giá trị thống kê



Vì dữ liệu có khá nhiều video bị trùng lặp, với những video xuất hiện nhiều lần, nhóm tiến hành lấy lần xuất hiện gần nhất để lọc ra tập dữ liệu chứa các video duy nhất.

	id	text	createTime	webVideoUrl	videoUrl	diggCount	shareCount	playCount	commentCount	hashtags	authorMeta.id
1	6958464277044792577	😊	1620143720	https://v9-vn.tiktokcdn.com/ea51f030882c5b6d54...	https://v9-vn.tiktokcdn.com/ea51f030882c5b6d54...	6349	40	254600	59	6626092899037134849	M
8	695387247888534785	Son môi xinh 🌹	1620125784	https://v9-vn.tiktokcdn.com/b35f087a734fb8877...	https://v9-vn.tiktokcdn.com/b35f087a734fb8877...	10800	29	137800	41	695375917210616833	M
12	6925773613698256133	Pure love#animals #lion	1612532346	https://v16.tiktokcdn.com/dd0c2c5f3960b5ce8d24...	https://v16.tiktokcdn.com/dd0c2c5f3960b5ce8d24...	30100000	770300	210200000	374400	[animals, lion]	6900065928828224518 M
13	6954977915901660417	Con dâu kho nhe 1 nỗi că mang về nỗi ăn, ông b...	1619331988	https://v9-vn.tiktokcdn.com/3284c113bcf6a53f94...	https://v9-vn.tiktokcdn.com/3284c113bcf6a53f94...	2854	13	82200	321	[nâunongcungtiktok]	66538775383 M
14	6956157274800803073	Bạn sẽ chọn ai #ceotruongnguyen #hltruongng...	1619606578	https://v9-vn.tiktokcdn.com/60956c5c8dafde611f...	https://v9-vn.tiktokcdn.com/60956c5c8dafde611f...	5910	53	124000	196	[ceotruongnguyen, hltruongnguyen, bandothan...	6946550081106396161 M
...
995	6960521749687586053	I have a treat today,please be happy,ok? #gest...	1620622768	https://v16.tiktokcdn.com/4514408fe4ab8ace967b...	https://v16.tiktokcdn.com/4514408fe4ab8ace967b...	76000	1625	346100	703	[gesture, gesturedance, dancer, happy]	6960489852706767877 M
996	6957311085053545730	hỏi đáp tâm sự đì các bạn oiiii Nhớ follow và ...	1619875220	https://v16.tiktokcdn.com/c060cf0719d3798d84ef...	https://v16.tiktokcdn.com/c060cf0719d3798d84ef...	24400	54	458100	270	6557557841481121794	M
997	6961390399755685121	Hành trình kiếm tìm nu cười #typ #tiktokvn #fo...	1620825010	https://v16.tiktokcdn.com/14aae0f6bb7e49b6e681...	https://v16.tiktokcdn.com/14aae0f6bb7e49b6e681...	55600	1475	679000	502	[typ, tiktokvn, foryou]	6854977182419698689 M
998	6951366101380893957	Easy	1618491048	https://v16.tiktokcdn.com/6aa3b2faf59289e48fad...	https://v16.tiktokcdn.com/6aa3b2faf59289e48fad...	9400000	100800	89500000	66900	127905465618821121	M
999	6952321064055688450	Nếu không biết đau thi không thể đứng lên khi ...	1618713392	https://v16.tiktokcdn.com/da8e049595c7a37dbfeb...	https://v16.tiktokcdn.com/da8e049595c7a37dbfeb...	12200	67	120600	0	6713690382875952130	M

578 rows x 31 columns

Kết quả DataFrame với 578 dòng dữ liệu là duy nhất



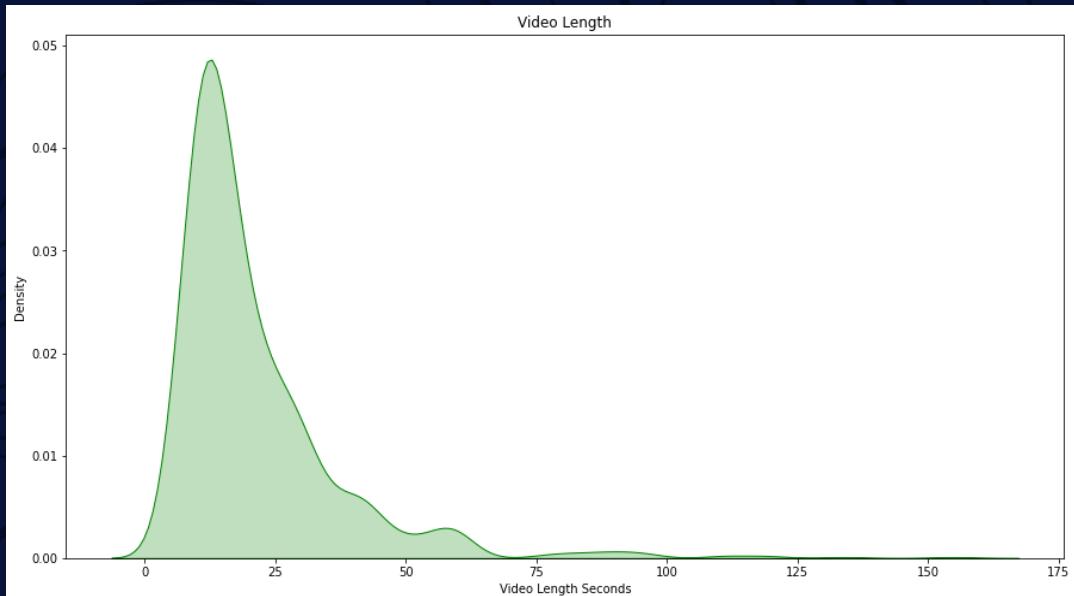
3.5.2 Biểu thị một số giá trị thống kê

Nhóm tiến hành thể hiện một số giá trị của các thuộc tính có dữ liệu dạng số liệu

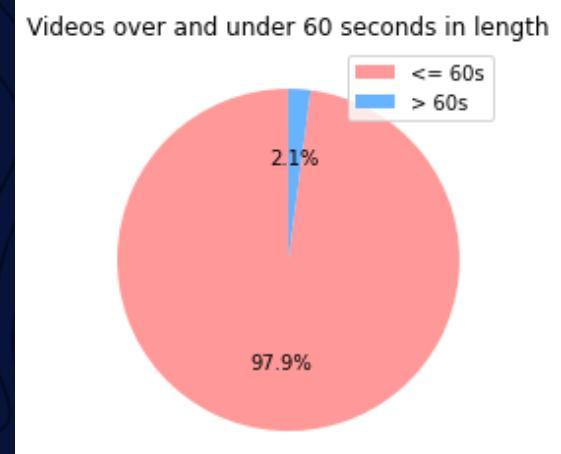
	createTime	diggCount	shareCount	playCount	commentCount	videoMeta.height	videoMeta.width	videoMeta.duration
count	5.780000e+02	5.780000e+02	578.000000	5.780000e+02	578.000000	578.000000	578.000000	578.000000
mean	1.618587e+09	1.309351e+06	23205.332180	1.326312e+07	12170.013841	956.124567	581.214533	21.451557
std	2.129832e+06	3.914255e+06	75401.955162	3.163351e+07	41966.140486	138.180445	91.657030	17.744145
min	1.612483e+09	4.300000e+01	0.000000	2.999000e+03	0.000000	464.000000	320.000000	6.000000
25%	1.617410e+09	8.423000e+03	46.500000	1.533500e+05	90.250000	960.000000	540.000000	11.000000
50%	1.619376e+09	4.350000e+04	351.000000	6.411500e+05	447.000000	1024.000000	576.000000	15.000000
75%	1.620214e+09	6.466000e+05	13000.000000	1.000000e+07	4892.750000	1024.000000	576.000000	25.000000
max	1.620905e+09	3.330000e+07	930400.000000	2.214000e+08	417100.000000	1280.000000	1280.000000	155.000000

Kết quả sau khi thu thập giá trị thống kê

3.5.3 Biểu thị thời lượng video



Kết quả biểu thị thời lượng video dạng biểu đồ ước lượng mật độ (KDE)



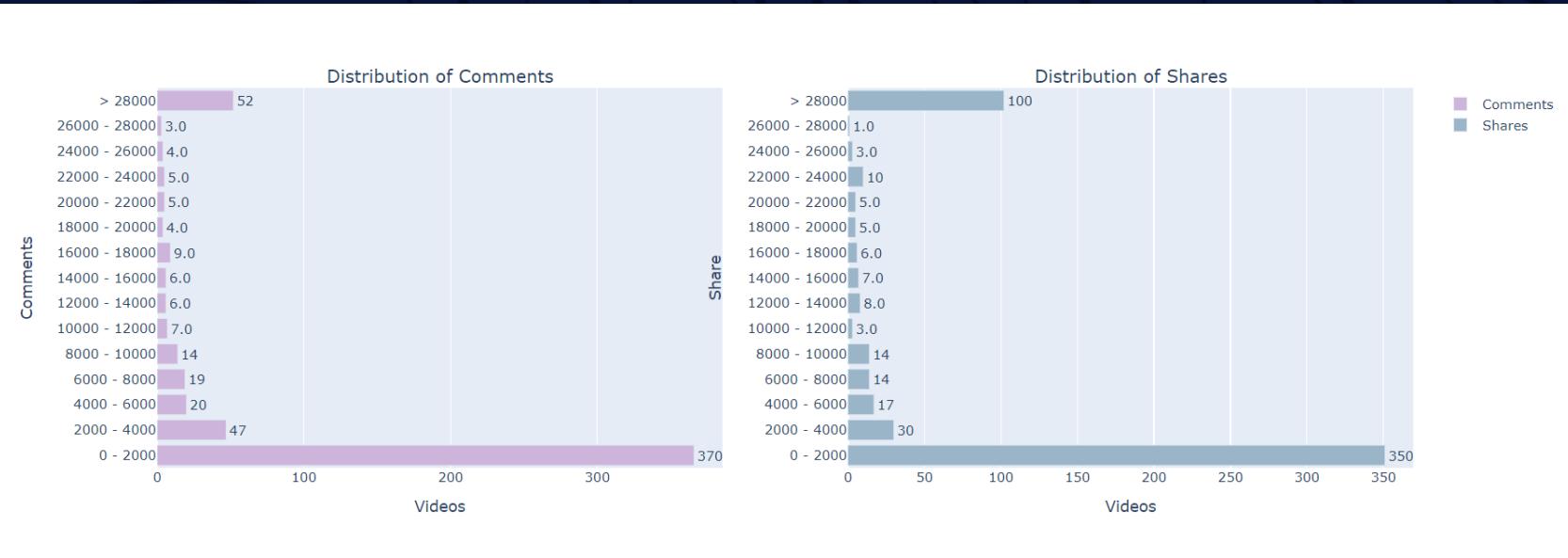
Kết quả biểu thị thời lượng video dưới và trên 60s bằng biểu đồ tròn

3.5.4 Biểu thị lượt play, like, share và comment theo từng khoảng



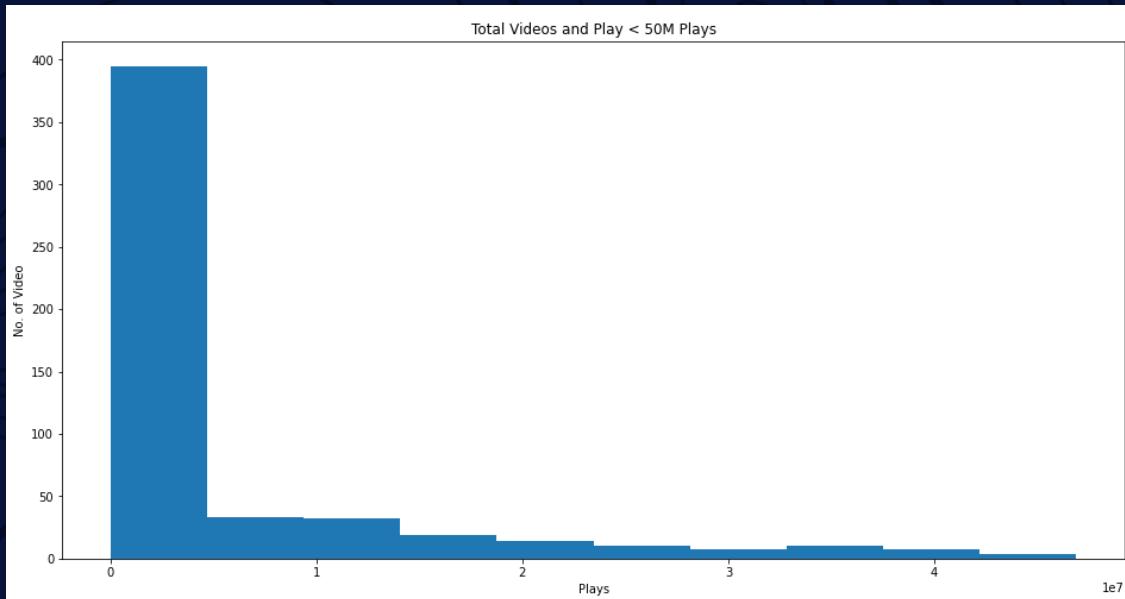
Kết quả biểu thị lượt play, like theo khoảng dạng biểu đồ thanh

3.5.4 Biểu thị lượt play, like, share và comment theo từng khoảng

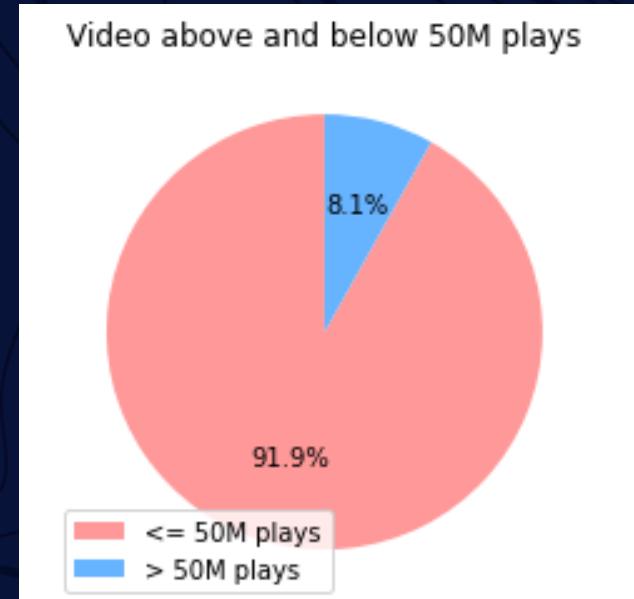


Kết quả biểu thị lượt share, comment theo khoảng dạng biểu đồ thanh

3.5.4 Biểu thị Biểu thị lượt play, like, share và comment theo từng khoảng



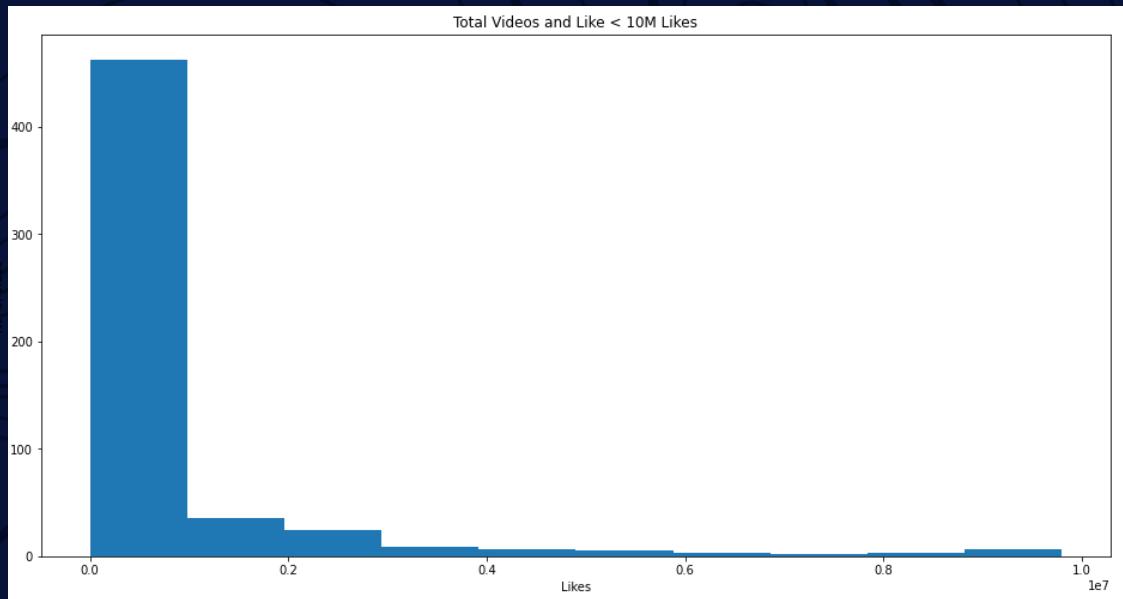
Biểu đồ cột hiển thị số lượng video có lượt plays dưới 50.000.000



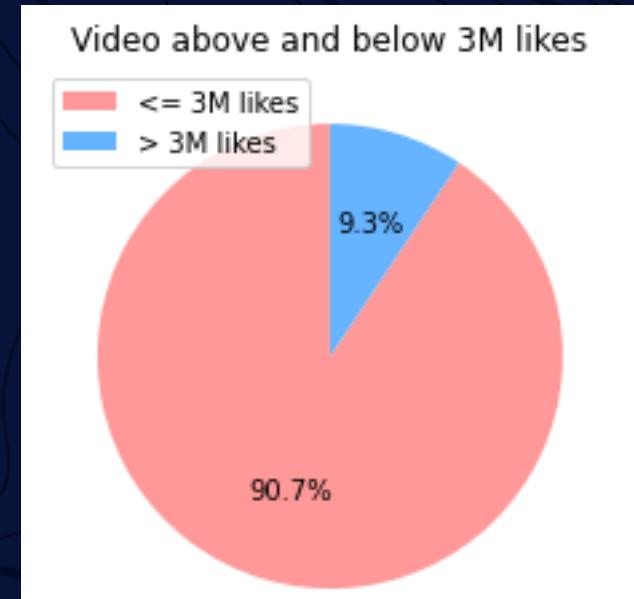
Biểu đồ tròn hiển thị số lượng video có lượt play dưới và trên 50 triệu



3.5.4 Biểu thị lượt play, like, share và comment theo từng khoảng



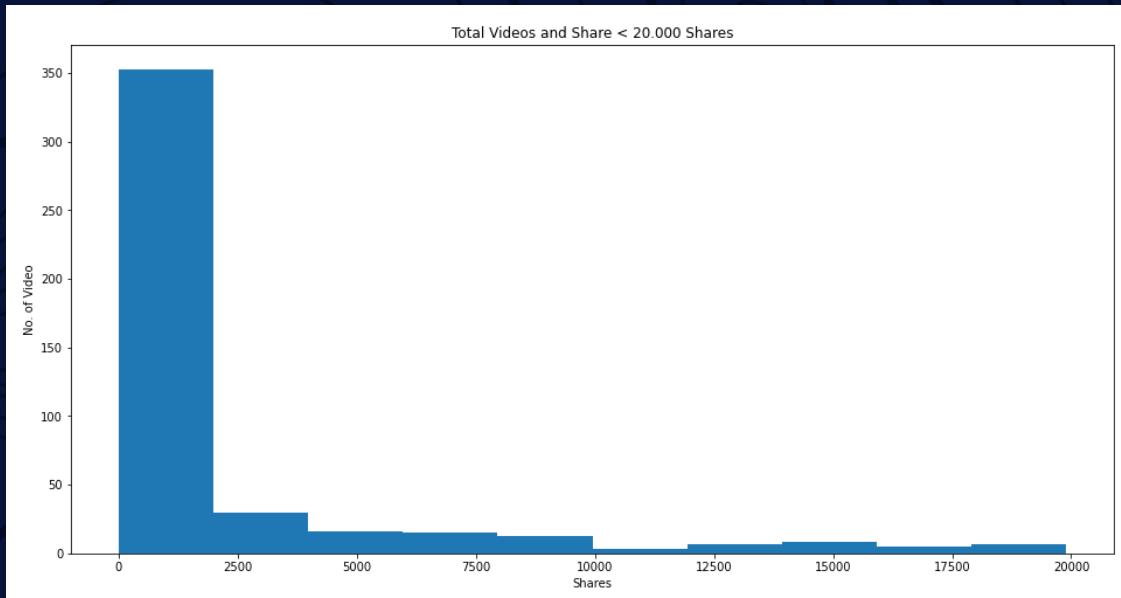
Biểu đồ cột hiển thị số lượng like dưới 10.000.000



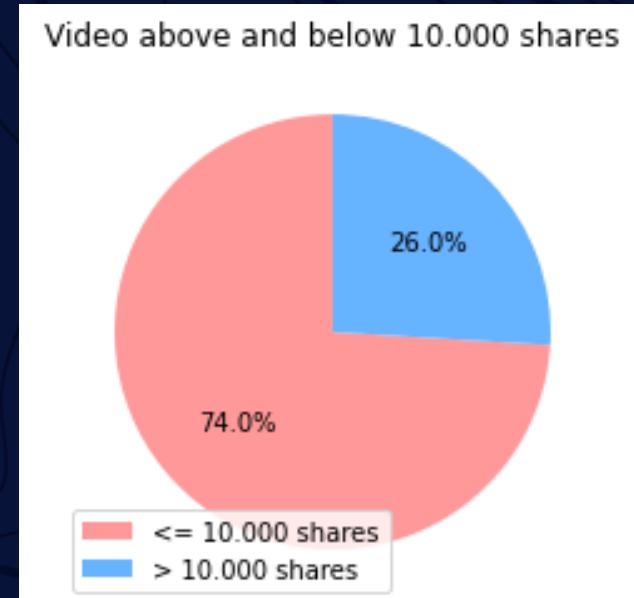
Biểu đồ tròn hiển thị số lượng video có lượt like trên và dưới 3.000.000



3.5.4 Biểu thị lượt play, like, share và comment theo từng khoảng



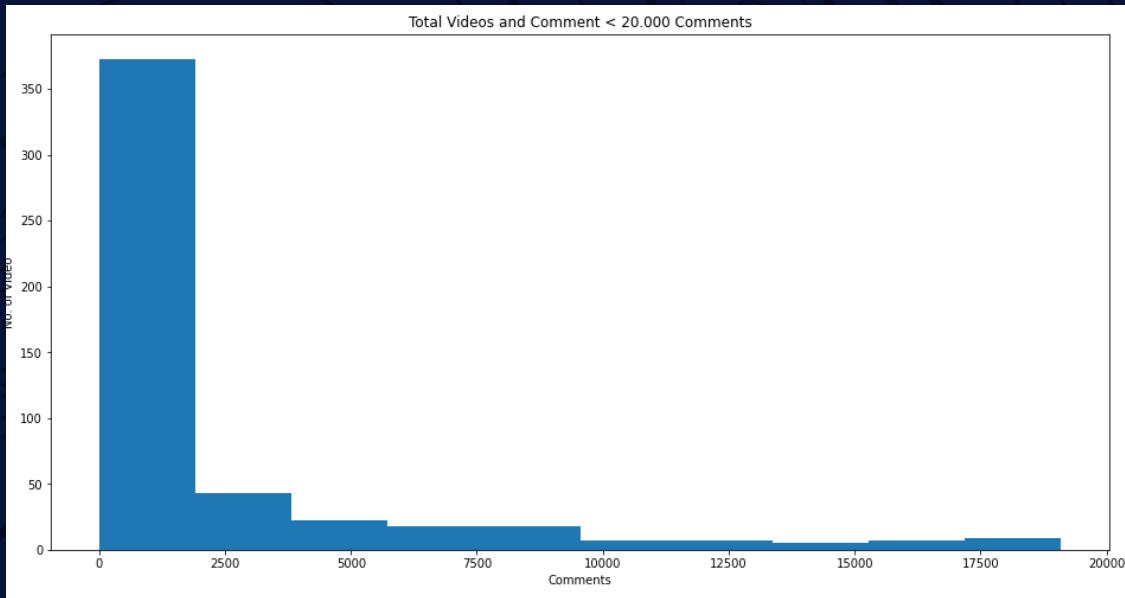
Biểu đồ cột hiển thị số lượng share dưới 20.000



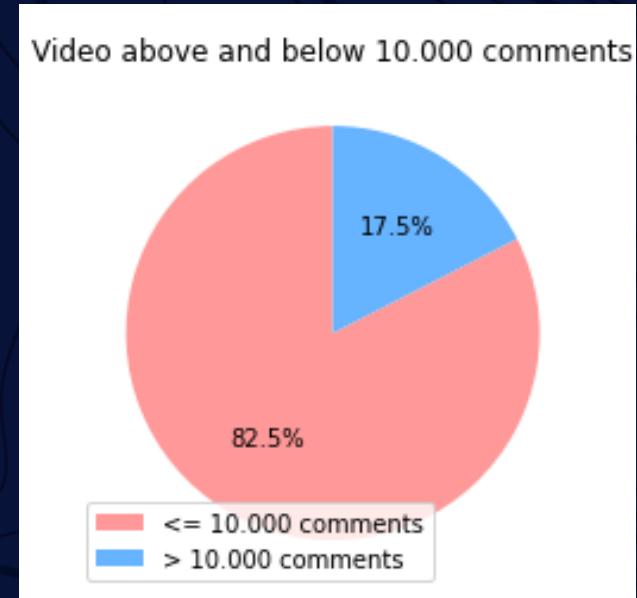
Biểu đồ tròn hiển thị số lượng video có lượt share trên và dưới 10.000



3.5.4 Biểu thị lượt play, like, share và comment theo từng khoảng

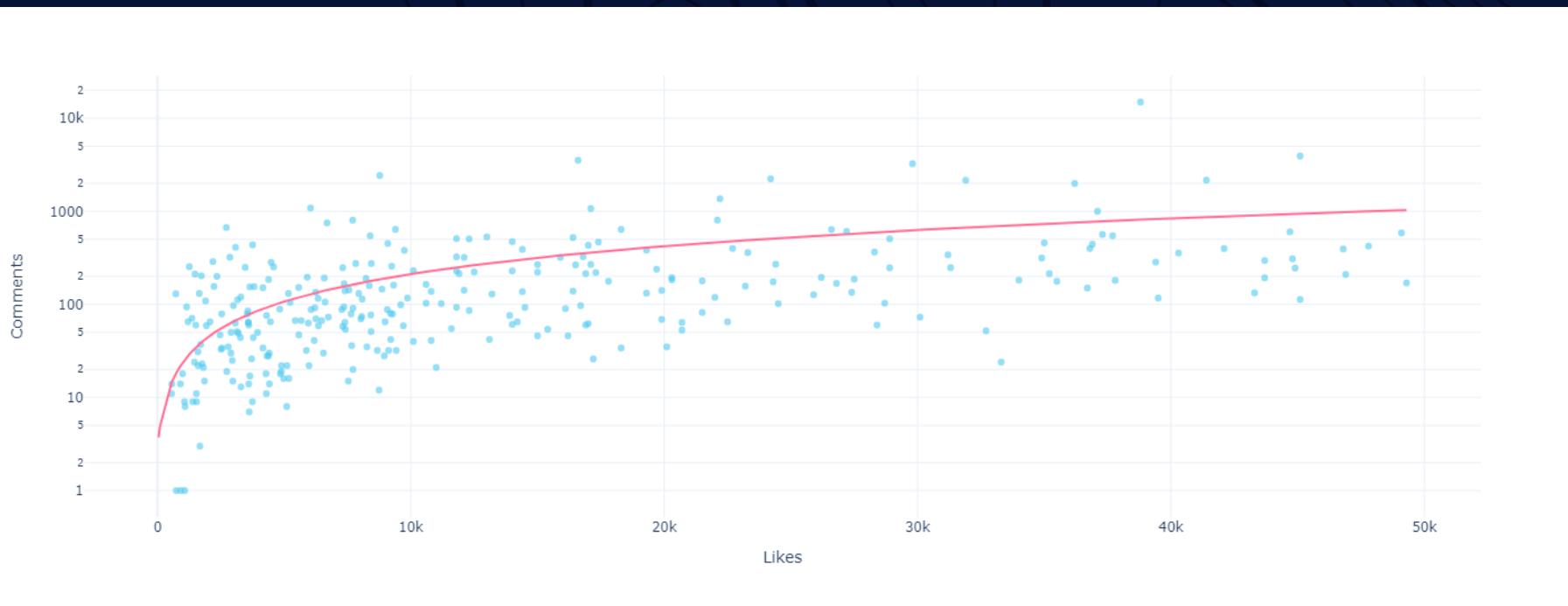


Biểu đồ cột hiển thị số lượng comment dưới 20.000



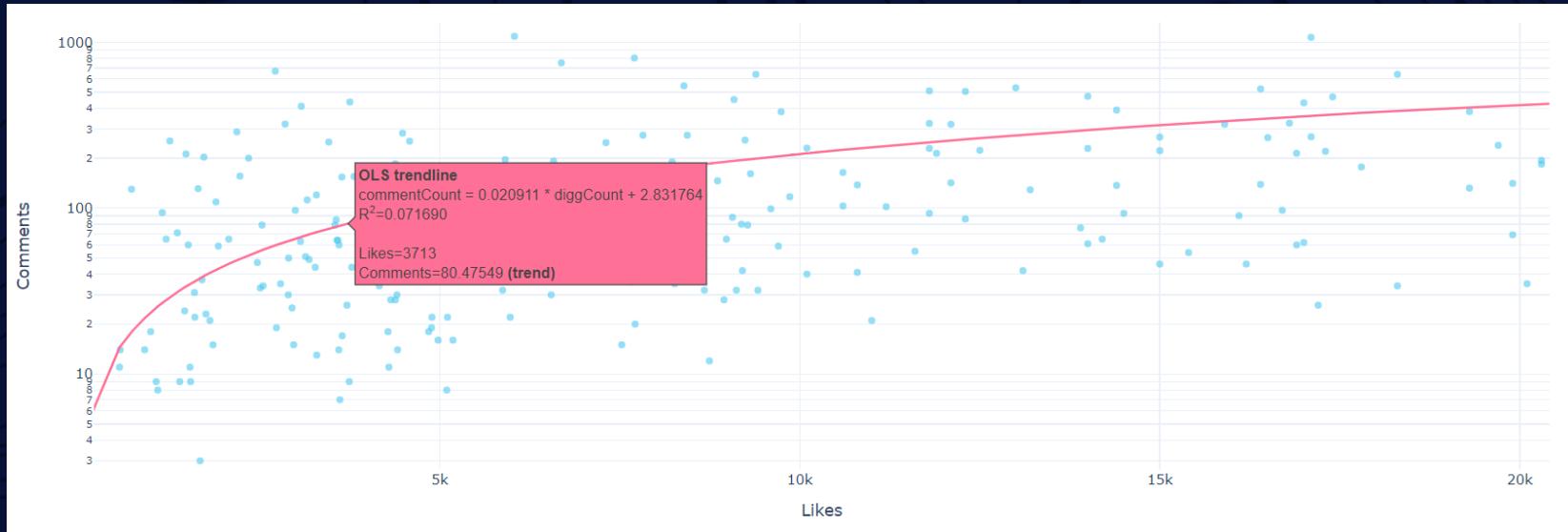
Biểu đồ tròn hiển thị số lượng video có lượt comment trên và dưới 10.000

3.5.5 Biểu thị sự tương quan giữa lượt Like và Comment



Kết quả biểu thị sự tương quan giữa lượt like và comment cùng với đường Trendline

3.5.5 Biểu thị sự tương quan giữa lượt Like và Comment

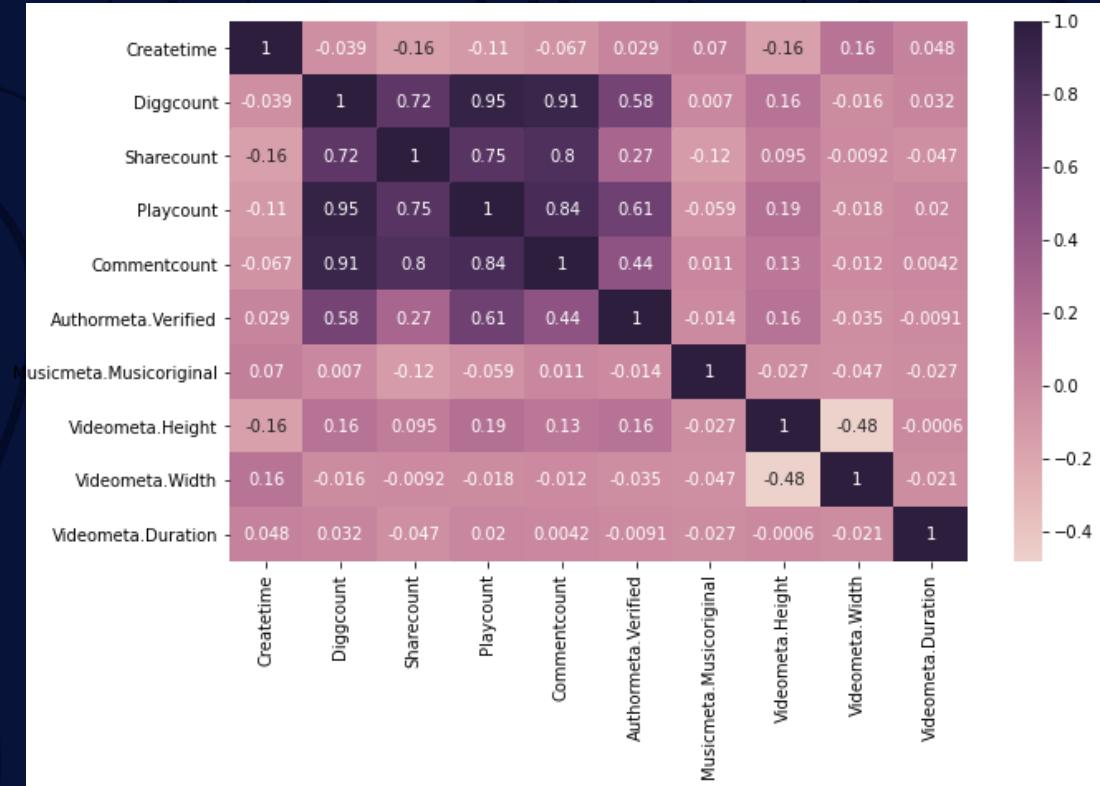


Đường trendline theo phương pháp bình phương nhỏ nhất OLS

Phương trình: $\text{commentCount} = 0.020911 * \text{diggCount} + 2.831764$

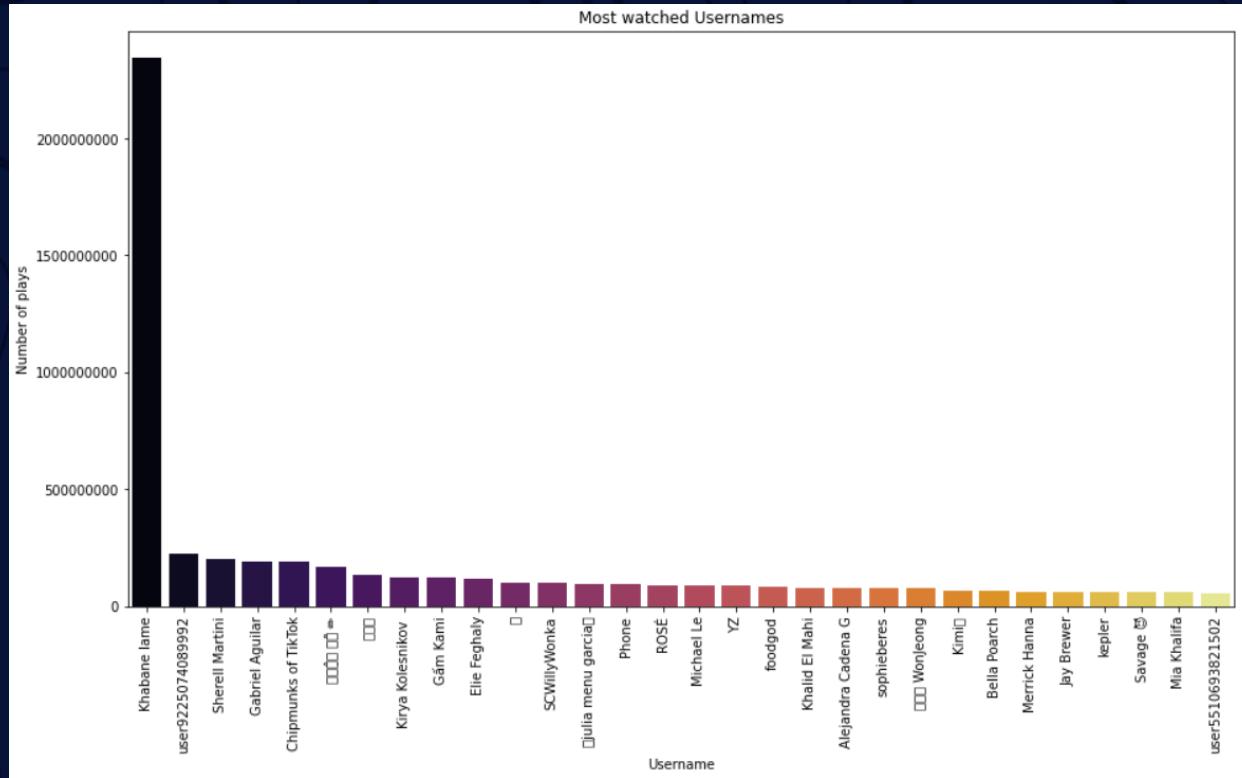
Đồ thị trên biểu thị sự tương quan của biến độc lập là lượt like và biến phụ thuộc là comment của video trending với 0.020911 là hệ số hồi quy và 2.831761 là sai số ngẫu nhiên

3.5.5 Biểu thị sự tương quan giữa các biến



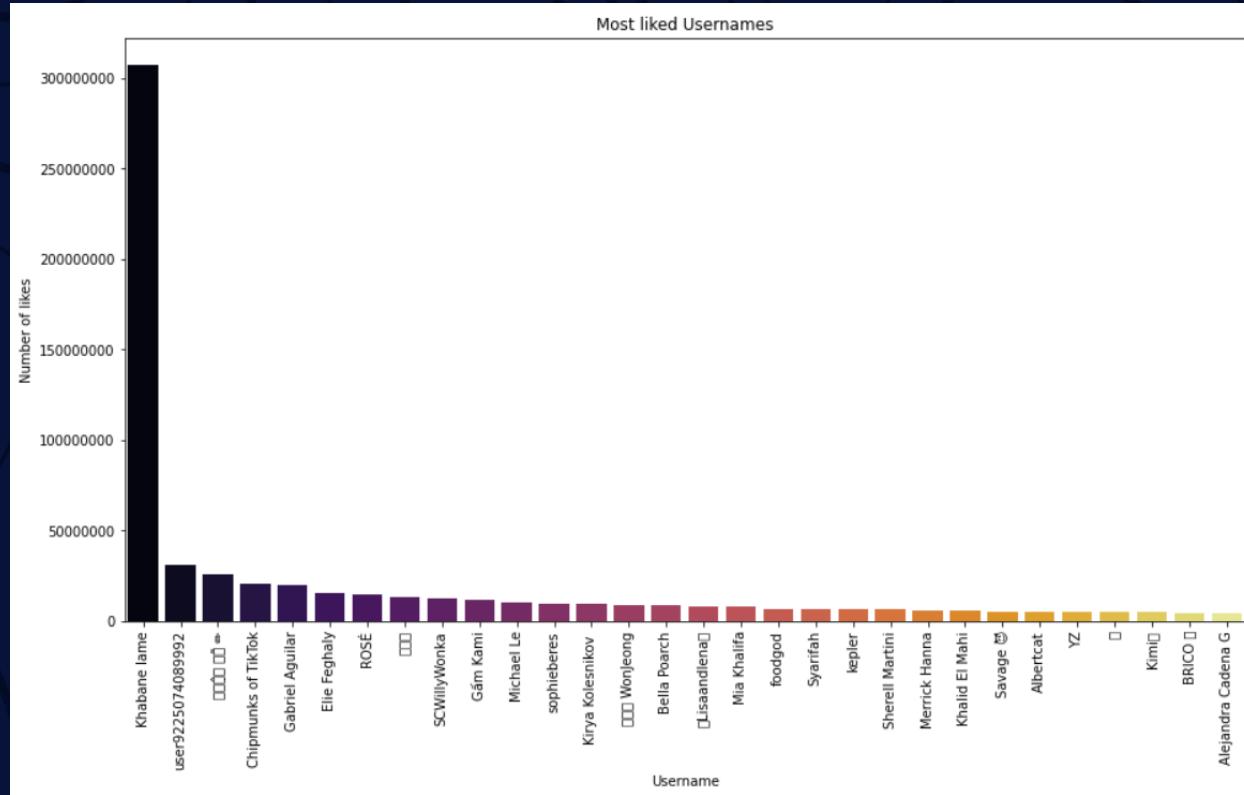
Bản đồ nhiệt thể hiện sự tương quan giữa các biến

3.5.6 Người dùng nổi bật trên TikTok



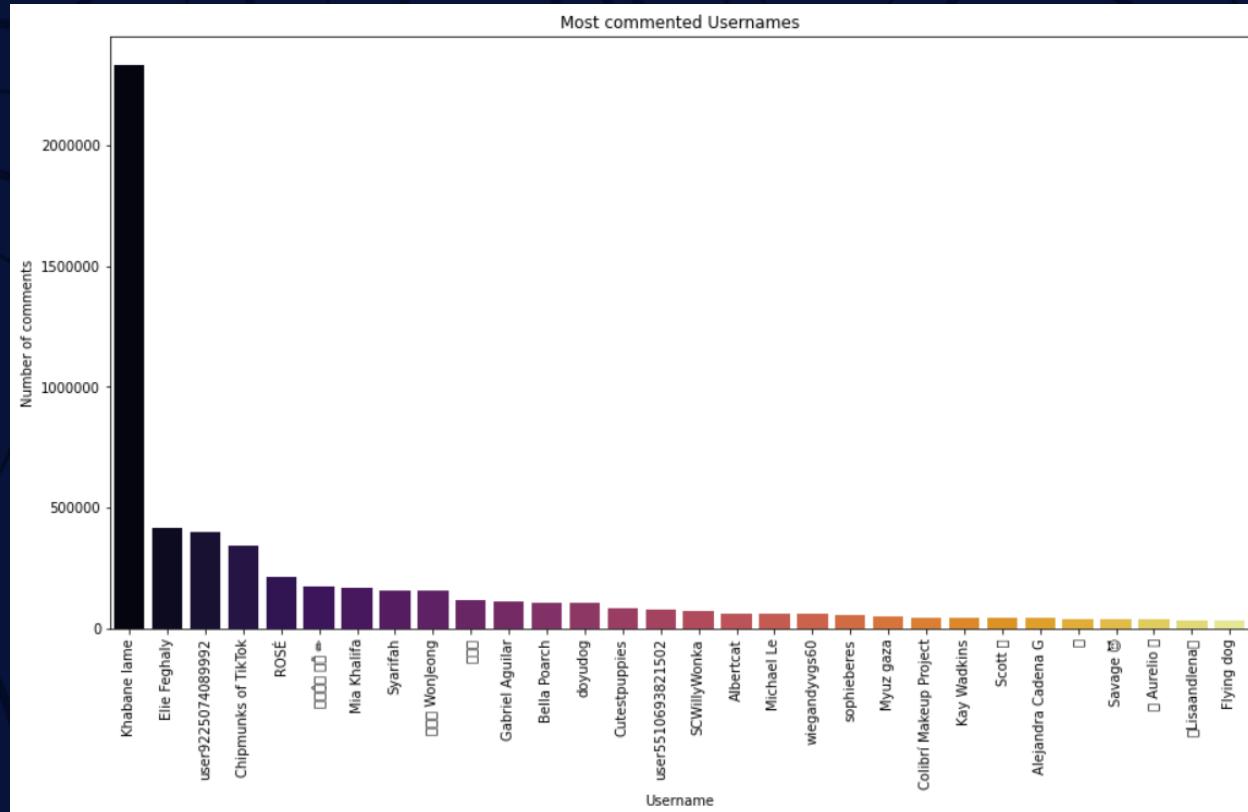
Kết quả biểu thị top người dùng được xem nhiều nhất

3.5.6 Người dùng nổi bật trên TikTok



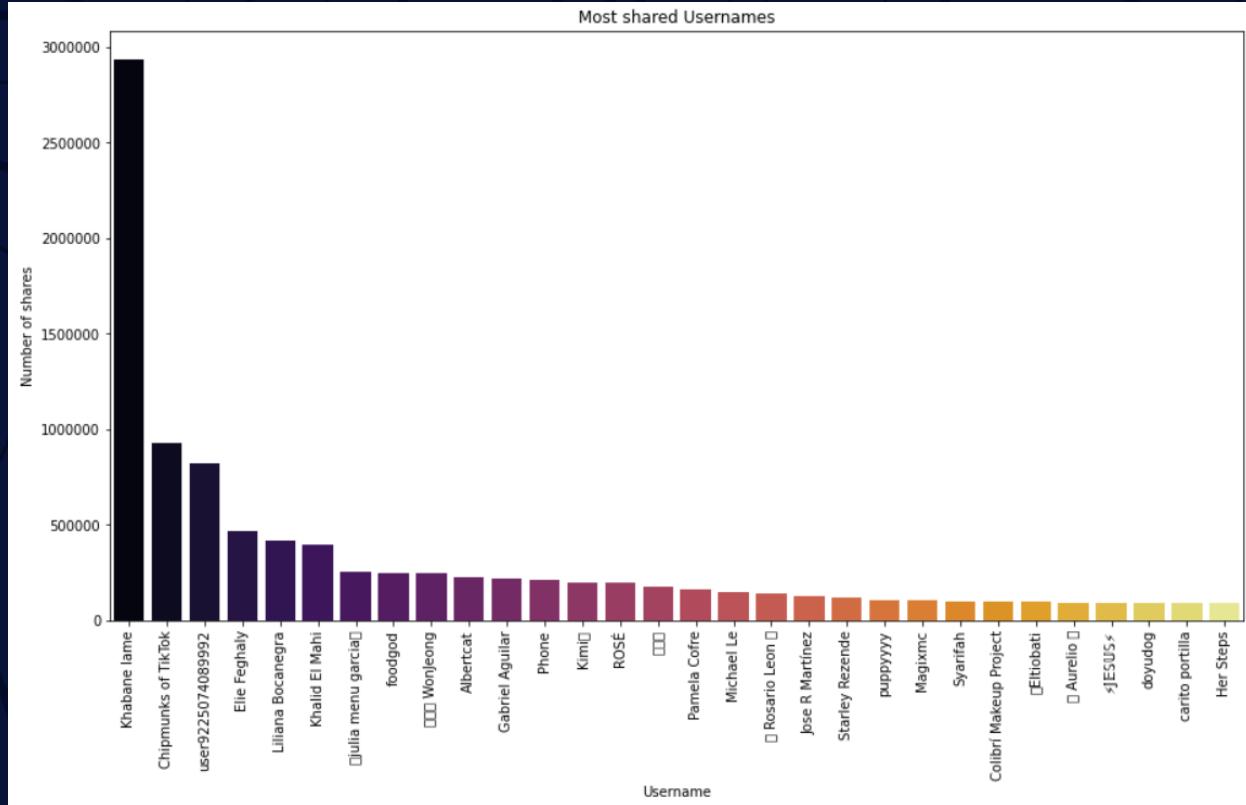
Kết quả biểu thị top người dùng được LIKE nhiều nhất

3.5.6 Người dùng nổi bật trên TikTok



Kết quả biểu thị top người dùng được COMMENT nhiều nhất

3.5.6 Người dùng nổi bật trên TikTok



Kết quả biểu thị top người dùng được SHARE nhiều nhất

3.5.6 Người dùng nổi bật trên TikTok



 **khaby.lame** 
Khabane lame

Follow

43 Đang Follow 49.5M Follower 693M Thích

Se volete ridere siete nel posto giusto 😂SNIT
[🔗 youtube.com/channel/UC86s...](https://youtube.com/channel/UC86s...)

Video



All right, just let me brush my teeth. 😊 Ok fatemi... Have a good weekend guys 😊 Buon weekend... Qualsiasi lavoro fai, devi sempre pulire, ma con...
▷ 19.9M 22giờ trước ▷ 14.9M 1ngày trước ▷ 79.8M 2ngày trước

Qua 4 biểu đồ thể hiện top người dùng nhận được lượt tương tác nhiều nhất TikTok, một cái tên nổi bật được cộng đồng TikTok Việt Nam quan tâm nửa đầu tháng 5, 2021 chính là Khabane Lame - hot TikToker với nhiều thành tích khủng



3.5.6 Người dùng nổi bật trên TikTok



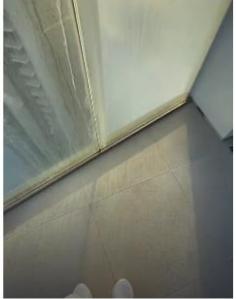
eliefeghaly7
Elie Feghaly

Follow

47 Đang Follow 1.5M Follower 18.5M Thích

Elie Feghaly
📍 Dubai ❤️
🏡 Real Estate Agent
Follow me on IG 

Video Đã thích



@pubity #eliefeghaly7 #dubai #mydubai... #eliefeghaly7 #intheend #dubai #viral #explore... #eliefeghaly7 #world #realestate #dubai #uae...

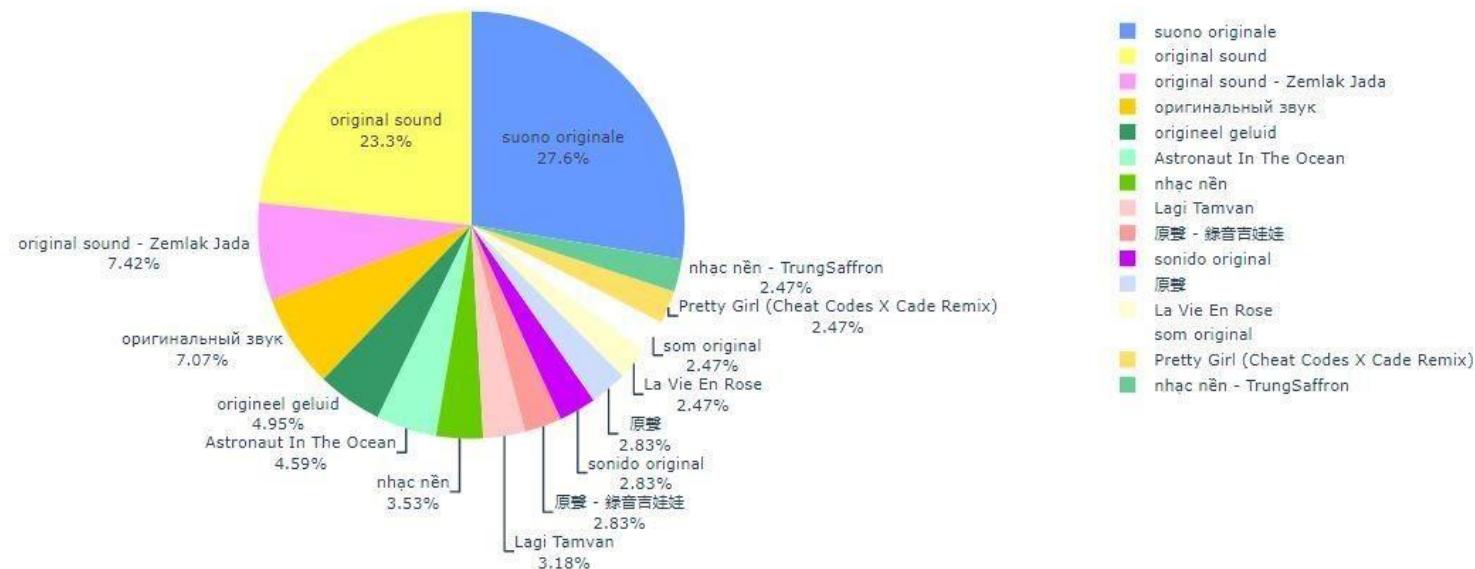
▷ 335.4K 18 giờ trước ▷ 4298 1 ngày trước ▷ 3691 1 ngày trước

Đối với lượt comment, ngoài Khabane Lame một cái tên nổi trội không kém là Elie Feghaly. Với công việc môi giới bất động sản, anh gây được chú ý của cộng đồng TikTok với những video cực lung linh, hào nhoáng tại những tòa nhà cao tầng ở Dubai



Có thể thấy TikTok không chỉ đơn thuần có những video hài hước, thuần giải trí, mang lại tiếng cười cho người xem, mà chúng ta hoàn toàn có thể phát triển nội dung mang tính định hướng thị hiếu của người xem, khéo léo đặt vào yếu tố quảng bá thương hiệu, từ đó tiếp cận được tập khách hàng mục tiêu và tăng độ nhận diện thương hiệu.

3.5.6 Trực quan dữ liệu nhạc Original



Biểu đồ Pie Chart

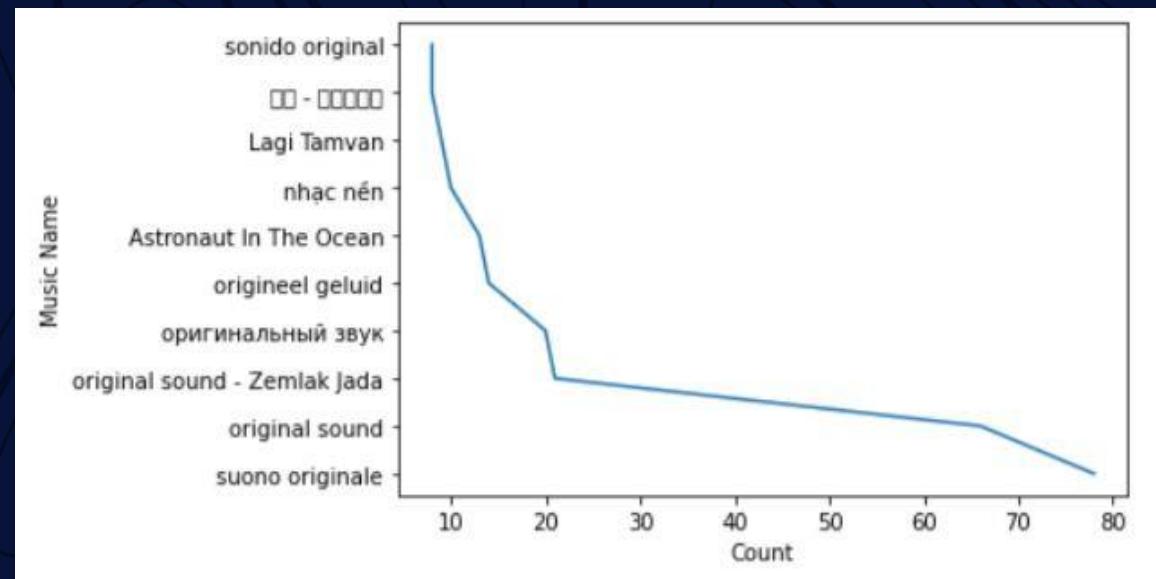
3.5.6 Trực quan dữ liệu nhạc Original



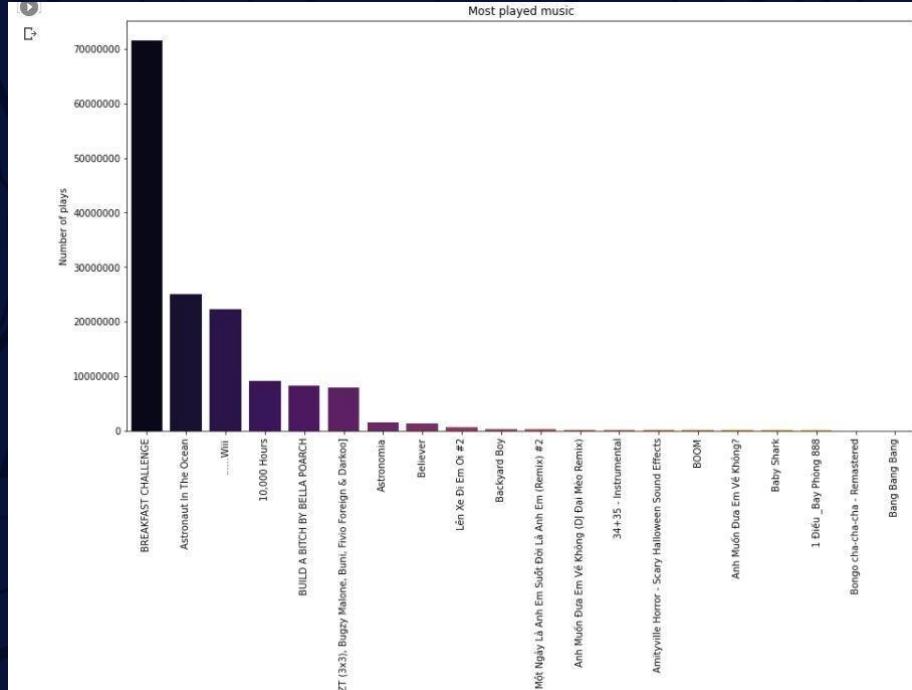
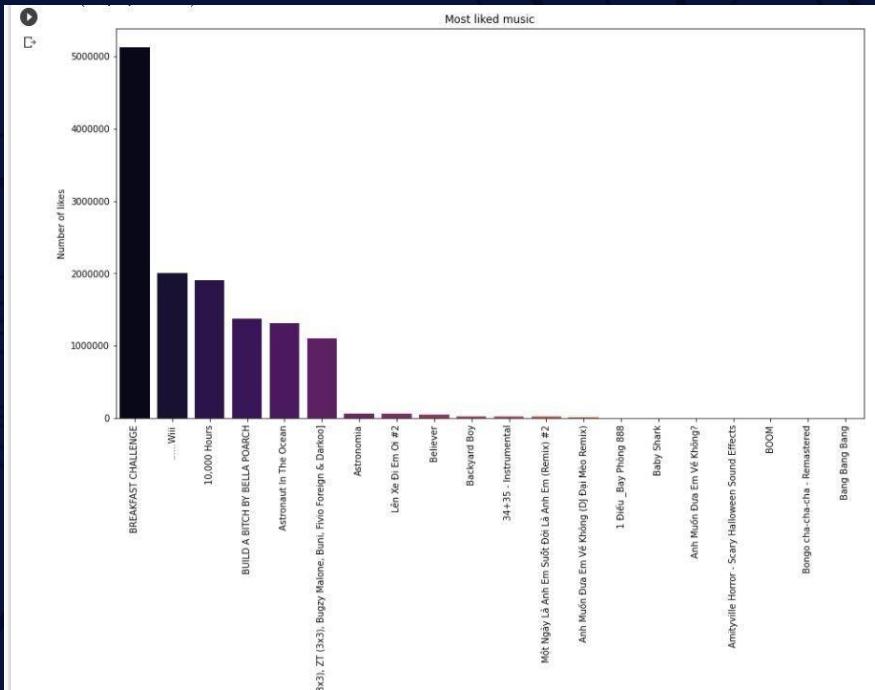
musicMeta.head(10)

	musicMeta.musicName	count
392	suono originale	78
366	original sound	66
383	original sound - Zemlak Jada	21
397	оригинальный звук	20
386	origineel geluid	14
9	Astronaut In The Ocean	13
150	nhạc nền	10
55	Lagi Tamvan	9
409	原聲 - 錄音吉娃娃	8
391	sonido original	8

Top 10 bài hát trong Data

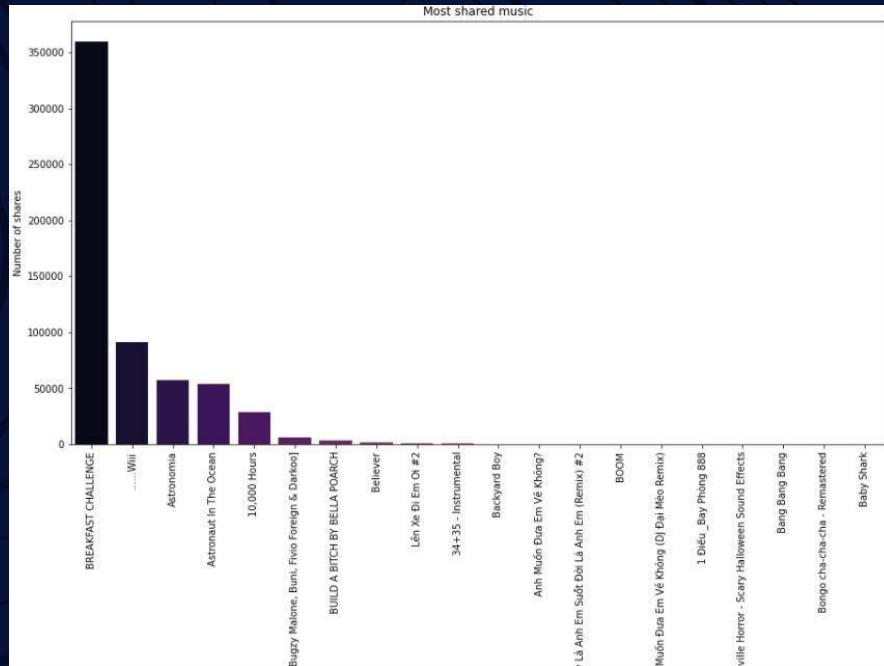
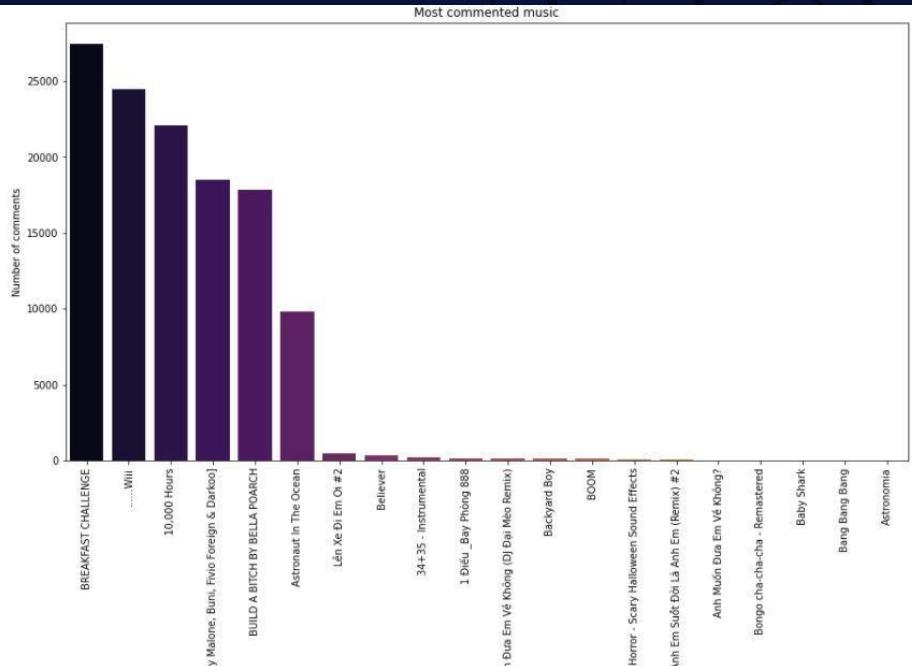


3.5.6 Trực quan dữ liệu nhạc Original



Biểu đồ cột thể hiện lượt Like, Play các bài hát

3.5.6 Trực quan dữ liệu nhạc Original



Biểu đồ cột thể hiện lượt Comment, Share các bài hát

3.5.7 Trực quan dữ liệu Hashtags



Mở rộng Hashtags thành Dataframe

	hashtag_0	hashtag_1	hashtag_2	hashtag_3	hashtag_4	hashtag_5	hashtag_6	hashtag_7	hashtag_8	hashtag_9	hashtag_10	hashtag_11	hashtag_12	hashtag_13	hashtag_14	hashtag_15	hashtag_16	hashtag_17
1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
8	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
12	animals	lion	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
13	nâunngoncungtiktok	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
14	ceotruongnguyen	hlvtruongnguyen	bandothanhhcong	vuotnguong	NaN	NaN	NaN	NaN	NaN	NaN	NaN							
...	
995	gesture	gesturedance	dancer	happy	NaN	NaN	NaN	NaN	NaN	NaN	NaN							
996	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
997	fyp	tiktokvn	foryou	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
998	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
999	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	

578 rows × 18 columns

Dòng lệnh và kết quả lọc các giá trị null

```
hashtags.isnull().sum()
```

hashtag_0	144
hashtag_1	242
hashtag_2	326
hashtag_3	384
hashtag_4	435
hashtag_5	475
hashtag_6	496
hashtag_7	510
hashtag_8	537
hashtag_9	542
hashtag_10	552
hashtag_11	561
hashtag_12	566
hashtag_13	569
hashtag_14	572
hashtag_15	574
hashtag_16	575
hashtag_17	576

`dtype: int64`

3.5.7 Trực quan dữ liệu Hashtags



Kết quả WordCloud Hashtags



04

KẾT LUÂN VÀ ĐÁNH GIÁ

4.1 Kết quả thu được



Nhóm đã thực hiện thu thập dữ liệu tiktok trong vòng 10 ngày (từ ngày 04/05/2021 đến ngày 14/05/2021). Số lượng video nhóm thu thập mỗi ngày là 100 và tổng số video trong tập dữ liệu là 1000.

- Trong số 1000 video dữ liệu đầu vào, chỉ có 578 video là duy nhất
- Chỉ có khoảng 8% video có số lượt play trên 50 triệu. Hơn 9% video TikTok lọt top trending có trên 3 triệu like. Ngoài ra, có 74% video có lượt share dưới 10.000 và 82% video có lượt comment dưới 10.000.
- Về kích thước, chiều dài lớn nhất và chiều rộng lớn nhất của video đều là 1280px. Có 97.9% video có độ dài dưới 60 giây và mật độ cao nhất là khoảng 14 giây.
- Lượt like và comment có mối liên hệ đồng biến với nhau, khi lượng like tăng lên 1 đơn vị thì lượng comment tăng lên 0,020911.



4.1 Kết quả thu được

- Các biến số liệu có sự tương quan với nhau. Tương quan tích cực cao giữa lượt xem và lượt thích, lượt thích và số bình luận, số xem và bình luận, và giữa số chia sẻ và số bình luận. Tương quan tiêu cực giữa thời lượng của video với các biến số khác như: lượt chia sẻ (khi ghi nhận giá trị âm -0.047), và giá trị rất thấp, tiệm cận 0 với lượt xem, thích, comment.
- Có gần 25% video không gắn hashtag. Số hashtag được gắn nhiều nhất là 18. Số lượng hashtag được gắn phổ biến nằm trong khung từ 1-6 hashtag, hashtag được sử dụng phổ biến nhất là: #fyp #xuhuong #foryou #foryoupage #tiktok #funny #xuhuongtiktok.
- Người dùng nhận được được lượt tương tác nhiều nhất TikTok là Khabane Lame.

4.2 Kết luận và đánh giá



Đề tài “Phân tích dữ liệu Tiktok” sử dụng RESTful TikAPI với ngôn ngữ Python được thực hiện lấy dữ liệu từ website Tiktok trong vòng 10 ngày. Sau khi có được tập dữ liệu thô, nhóm tiến hành làm sạch và định dạng lại dữ liệu cho phù hợp. Khi đã có được tập dữ liệu mong muốn, nhóm đã tiến hành thống kê, trực quan hóa, phân tích và rút ra một số kết luận về xu hướng cũng như thuật toán đề xuất của Tiktok.

4.3 Hướng phát triển



Tiếp tục xây dựng một bộ dữ liệu lớn hơn, cụ thể hơn để có được kết quả phân tích mang tính chính xác cao hơn, thuyết phục hơn. Từ đó, đưa ra bộ dữ liệu mẫu tham khảo và đề xuất tiêu chí video thịnh hành, lọt vào thuật toán đề xuất của tiktok cho người dùng và doanh nghiệp.

Thực hiện phân tích chuyên sâu và mở rộng đề tài liên quan tới một số vấn đề như: đề xuất giải pháp kinh doanh, marketing, ... trên Tiktok cho các Doanh nghiệp.



THANKS

Cảm ơn thầy và các bạn đã chú ý lắng nghe