



Trường ĐH Khoa Học Tự Nhiên

# CÁC KÊNH YOUTUBE VỀ KHOA HỌC DỮ LIỆU

Trình bày: Nhóm 20



# LÝ DO CHỌN ĐỀ TÀI



## Nhu cầu?



Với tư cách là các học sinh khoa học dữ liệu, chúng em mong muốn sẽ được tiếp cận đến những kiến thức chất lượng, cập nhập những xu hướng mới trong lĩnh vực này hằng ngày

## Tại sao là YouTube?



Youtube là một nền tảng cực kì phổ biến hiện nay, nó xuất hiện trong rất nhiều hoạt động hằng ngày như là học tập, giải trí, cập nhập tin tức, .... Cũng như nó chứa rất nhiều nội dung chất lượng đến từ các chuyên gia trong nhiều lĩnh vực



## Quyết định

Chúng em quyết định sẽ phân tích, tìm hiểu về cách những kênh YouTube làm về chủ đề KHL, hiểu cách chúng hoạt động, nội dung nổi bật của từng kênh sau đó tiến hành so sánh toàn diện sao đó có thể chọn ra những kênh phù hợp với nhu cầu bản thân

# THU THẬP DỮ LIỆU

Chọn kênh

Lấy video

Lấy comment

Giới hạn quota

# THU THẬP DỮ LIỆU

Những kênh này được thu thập id bằng cách tra cứu trên mạng

Chọn kênh

Lấy video

Lấy comment

Giới hạn quota

01 DeepLearningAI

02 3Blue1Brown

03 sentdex

04 Joma Tech

05 DataCamp

06 CS Dojo

07 StatQuest with  
Josh Starmer

08 Tech With Tim

09 365 Data Science

10 Data Professor

11 Data Science Dojo

12 codebasics

13 Two Minute Papers

14 TheAiGrid

15 AI News

16 Abhishek Thakur

17 IBM Technology

18 PRO ROBOTS

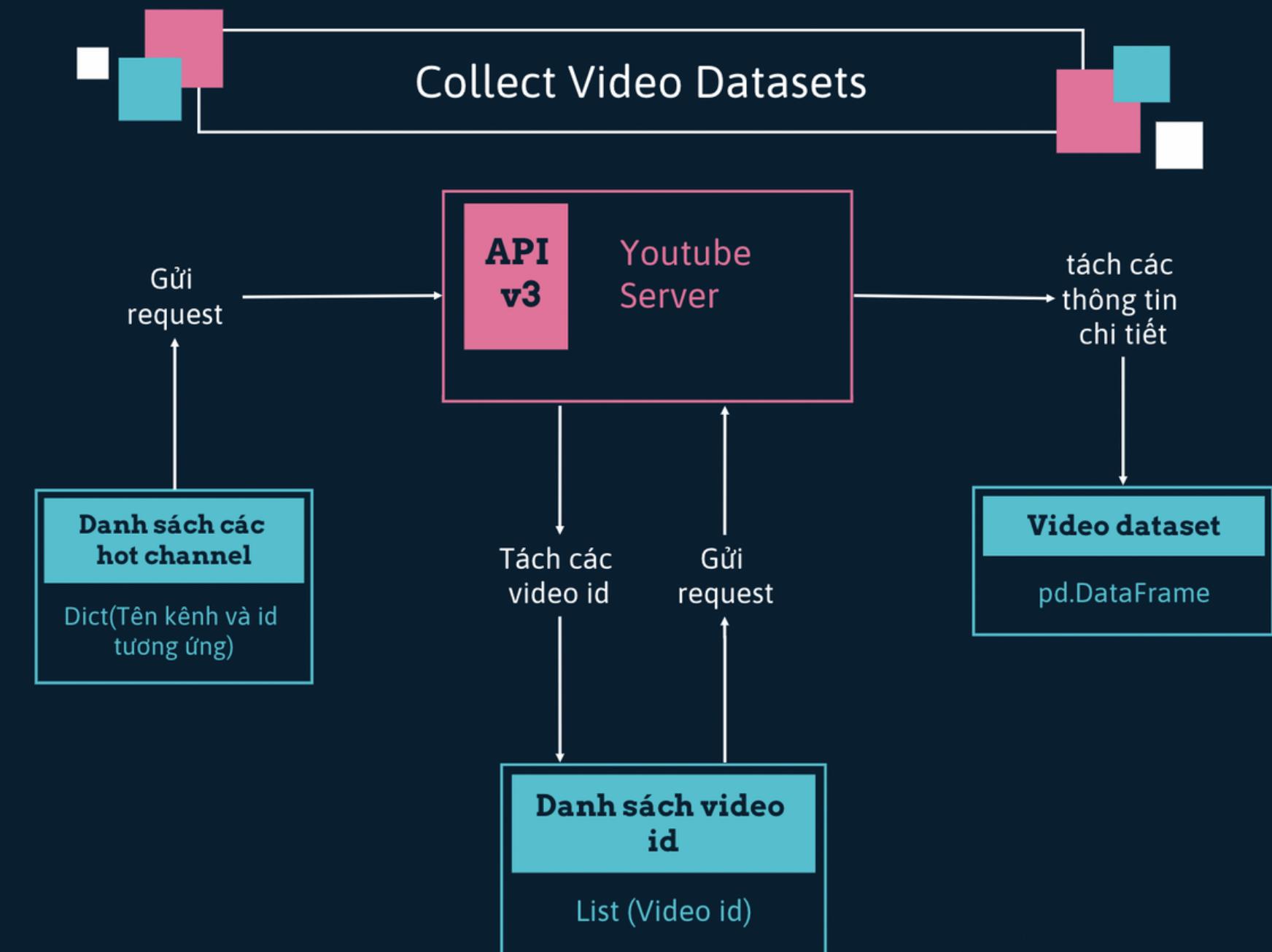
# THU THẬP DỮ LIỆU

Chọn kênh

Lấy video

Lấy comment

Giới hạn quota



# THU THẬP DỮ LIỆU

Chọn kênh

Lấy video

Lấy comment

Giới hạn quota

5879 DÒNG  
13 CỘT

- video\_id
- title
- published
- view\_count
- like\_count
- comment\_count
- duration
- definition (hd, sd ...)
- tags
- default\_audio\_language
- madeforkid (true or false)
- channel\_id
- channelTitle

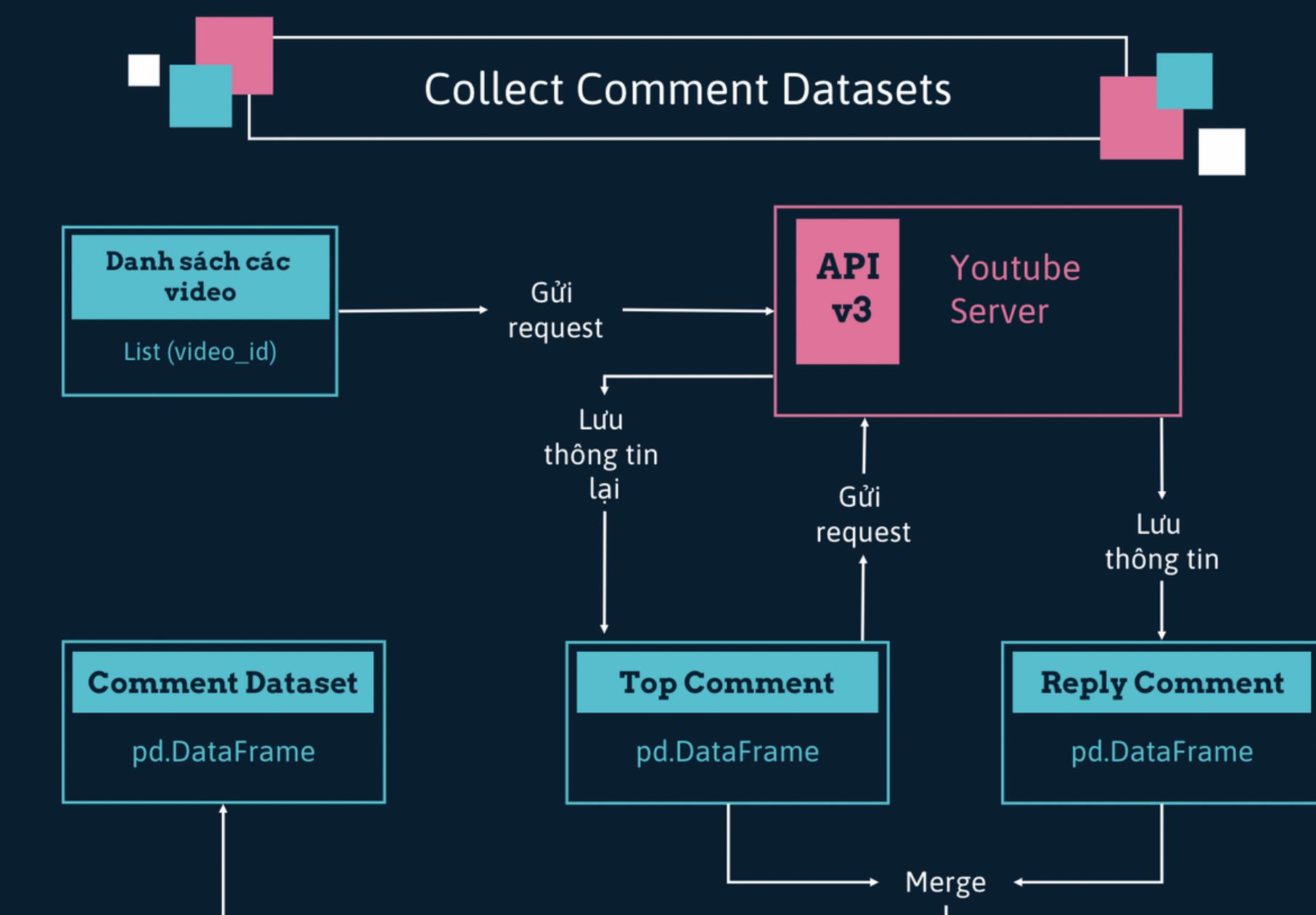
# THU THẬP DỮ LIỆU

Chọn kênh

Lấy video

Lấy comment

Giới hạn quota



# THU THẬP DỮ LIỆU

Chọn kênh

Lấy video

Lấy comment

Giới hạn quota

135996 DÒNG  
10 CỘT

- Comment\_id.
- author
- Reply\_for
- Type (top-level comment (1) or reply(2))
- video\_id
- total\_reply
- like\_count
- published\_at
- textdisplay
- updatedat

# THU THẬP DỮ LIỆU

Chọn kênh

Lấy video

Lấy comment

Giới hạn quota

Youtube chỉ cung cấp 10000 quota /ngày .Đối với request thu thập dữ liệu thì 1 request ứng với 1 quota ,nên tối ta người dùng chỉ được 10000 request trên ngày.

- Nhóm thực hiện thu thập dữ liệu theo từng kênh
- Mỗi kênh sẽ có một video\_df và một comment\_df
- Nối dữ liệu của từng kênh lại để được bảng cuối cùng

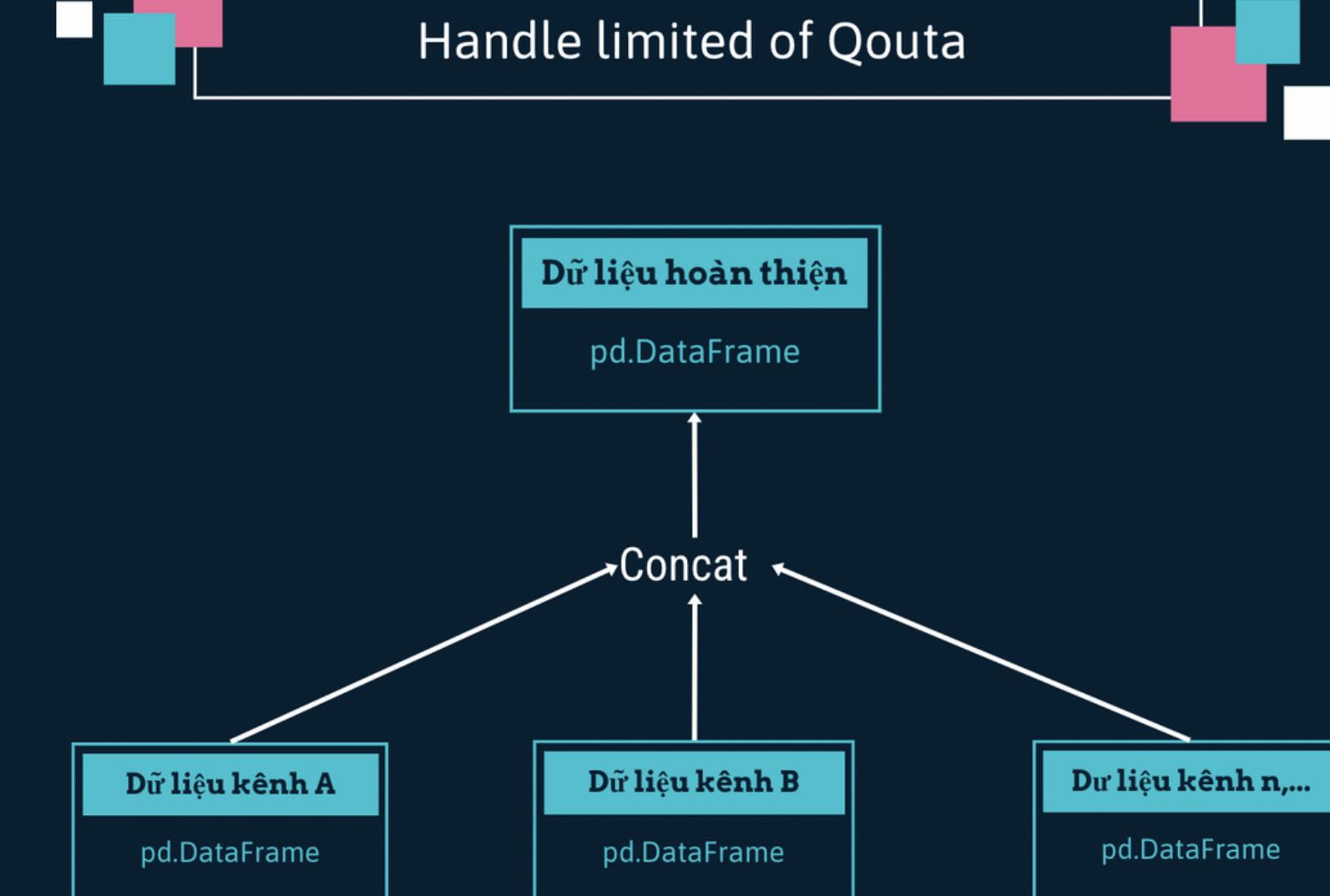
# THU THẬP DỮ LIỆU

Chọn kênh

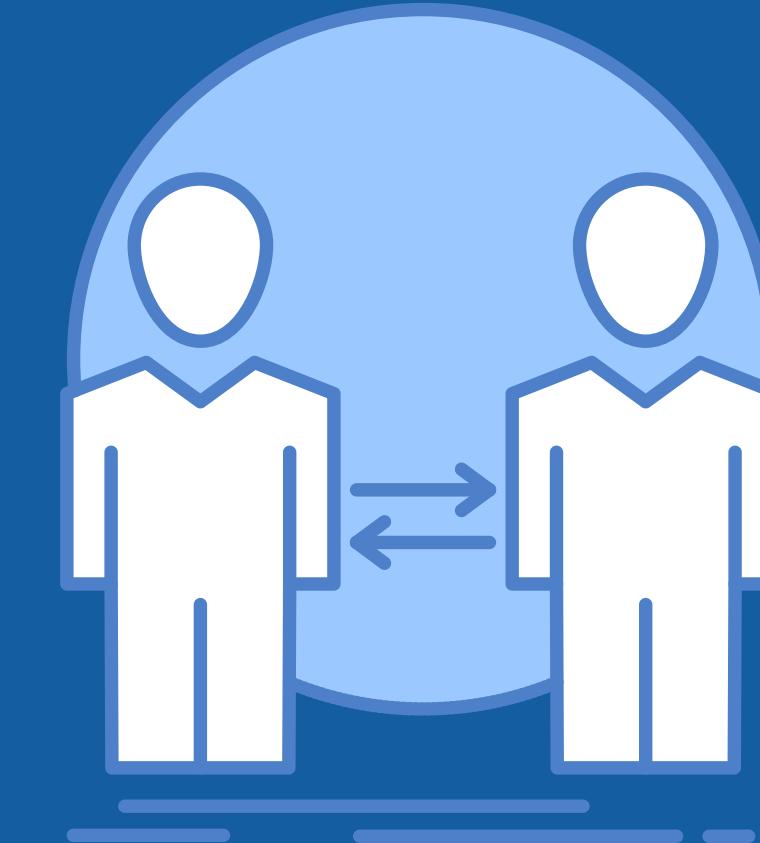
Lấy video

Lấy comment

Giới hạn quota



# LÀM SẠCH DỮ LIỆU



## Duplicates

Xuất hiện do một vài video và comment bị thu thập nhiều lần

- Xóa các dòng trùng trong videos\_df
- Xóa các dòng trùng trong comments\_df



## Datatypes

- Chuyển ngày tháng về datetime (sử dụng hàm của pandas)
- Chuyển duration về timedelta (sử dụng hàm của pandas)
- Chuyển tags về list of strings (apply hàm của thư viện ast cho từng dòng)

# LÀM SẠCH DỮ LIỆU

## videos\_df:

- like\_count (41 records)
- comment\_count(5 records)
- tags (720 records)
- default\_audio\_language (358 records)

## comments\_df

- author (2 records)
- Reply\_for (85293 records)



- mean (like\_count/ tổng view\_count) \* view\_count
  - fill bằng 0
  - giữ nguyên
  - fill bằng en
- 
- giữ nguyên
  - drop

# LÀM SẠCH DỮ LIỆU

Sau khi drop các cột dữ liệu thừa ( channel\_id, Comment\_id, Reply\_for)



VIDEO\_DF  
**5472 DÒNG**  
**12 CỘT**



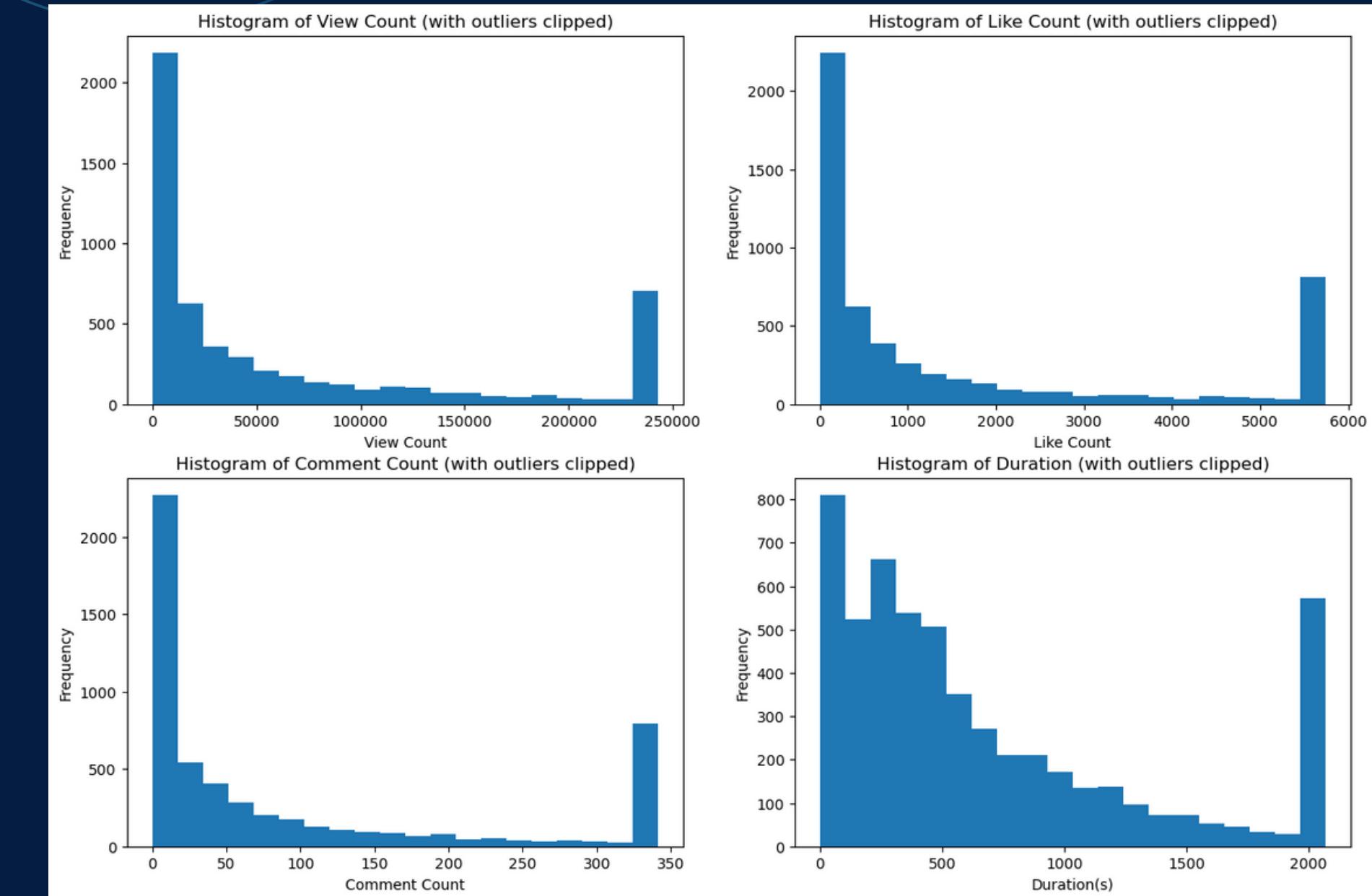
COMMENT\_DF  
**110409 DÒNG**  
**8 CỘT**



# EXPLORATORY DATA ANALYSIS

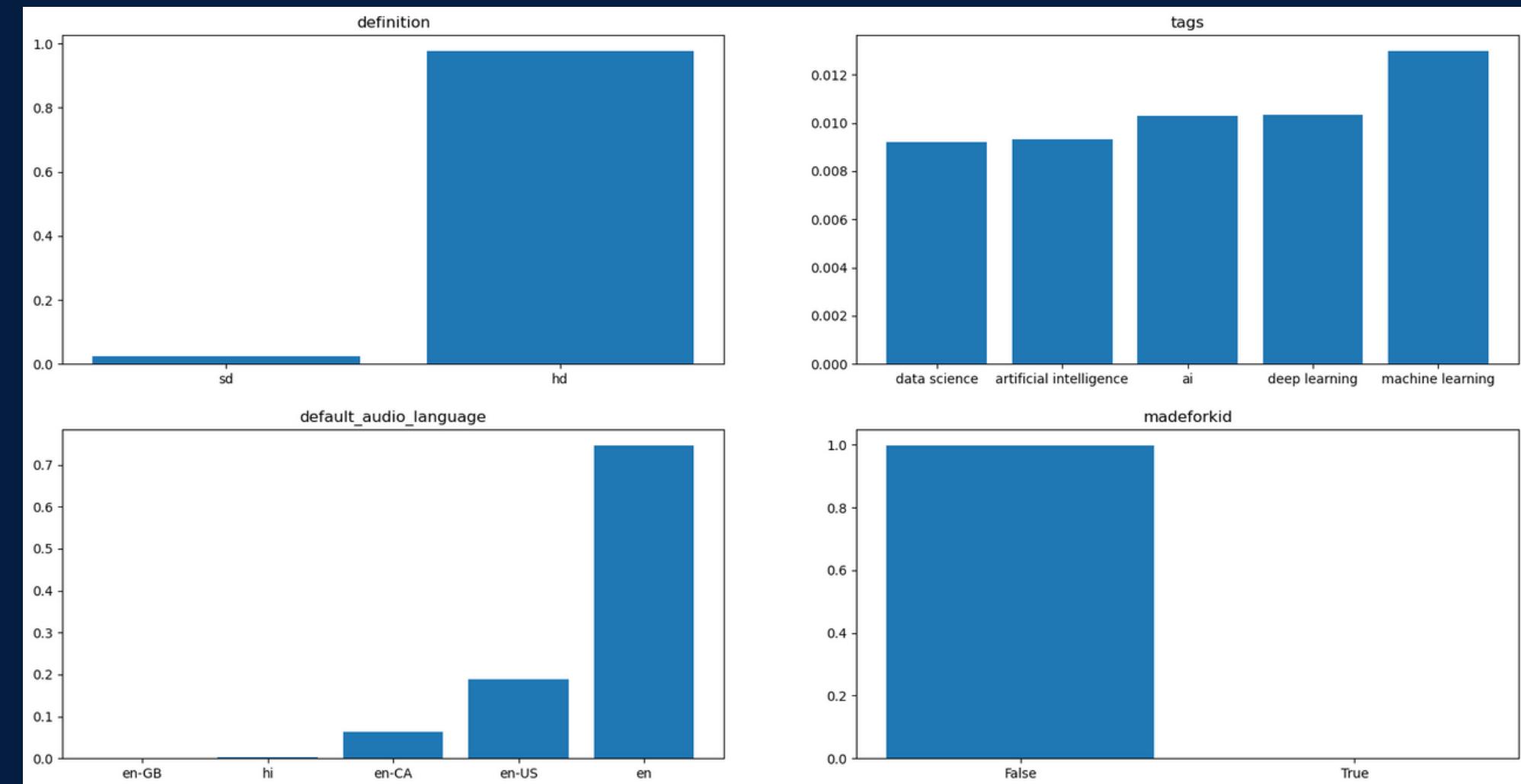
# Numerical

- Đa phần các video trong các kênh không phải là các video nổi tiếng nên các histogram của view, like, comment count có thiên hướng lệch trái
- Durations: phần lớn các kênh có xu hướng làm video ngắn (từ 0-10 phút)

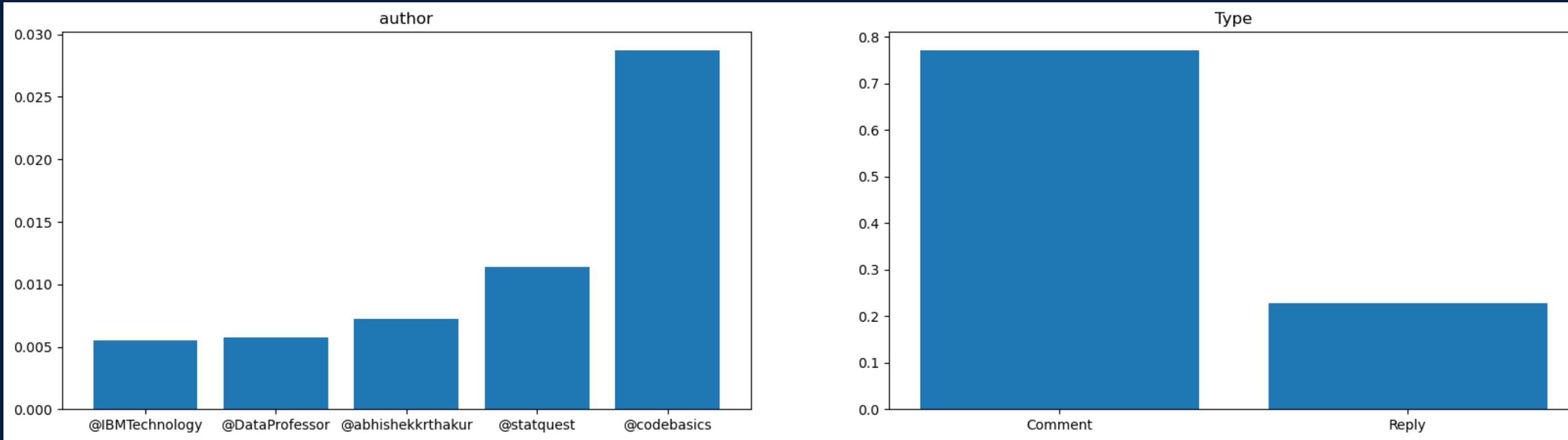


# Non-numerical

- Các video đa phần đều có chất lượng cao .
- Tags: machine learning chiếm khá cao , điều đó chứng minh machine learning là từ khoá được rất nhiều quan tâm trong khoa học dữ liệu.
- Default audio language: Chủ yếu là tiếng anh
- Các video đa phần là dành cho người lớn

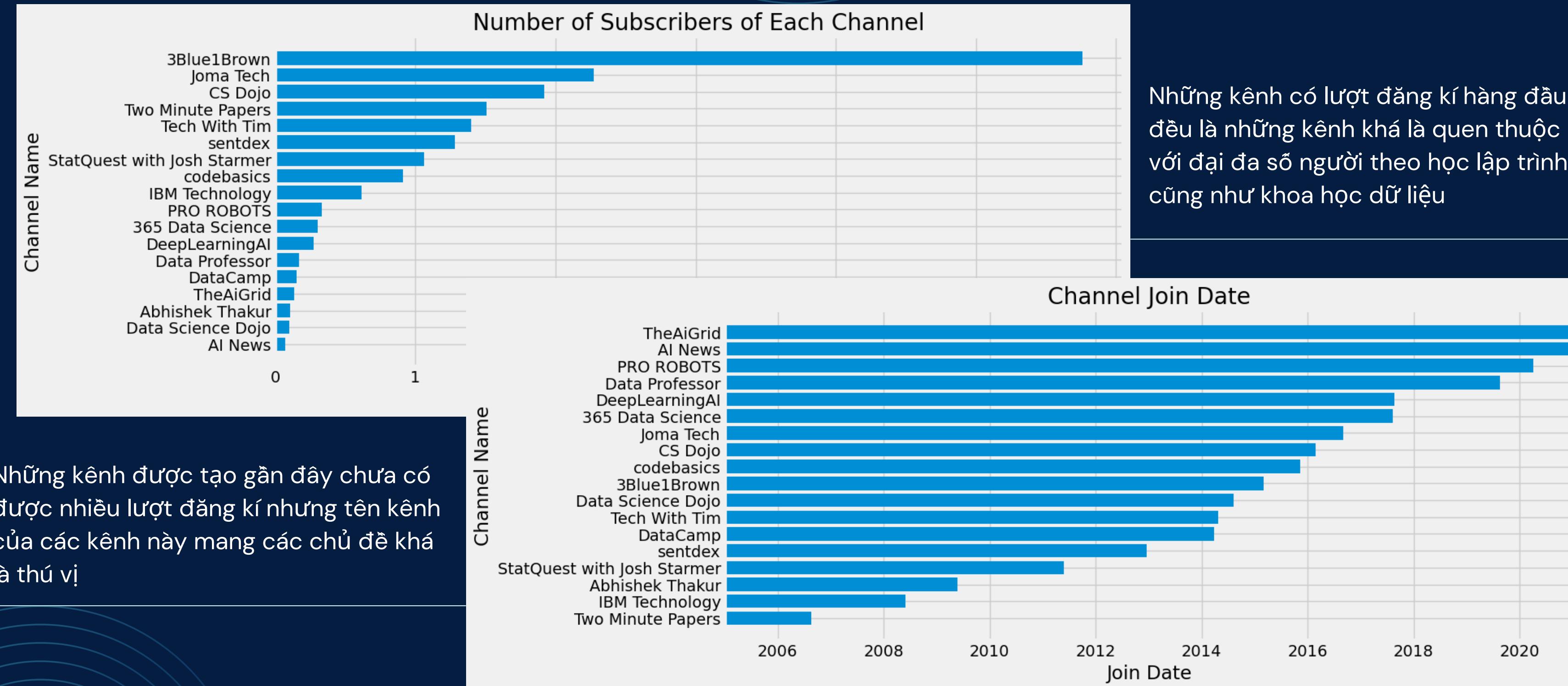


# Non-numerical



- Code basics có thể là kênh mà quan tâm ,tương tác với cộng đồng người xem tốt nhất.
- Comment top chiếm đa số hơn comment reply ,có thể hầu hết các comment là các câu cảm thán nên không quá nhiều reply.

# Thông tin chung về các kênh

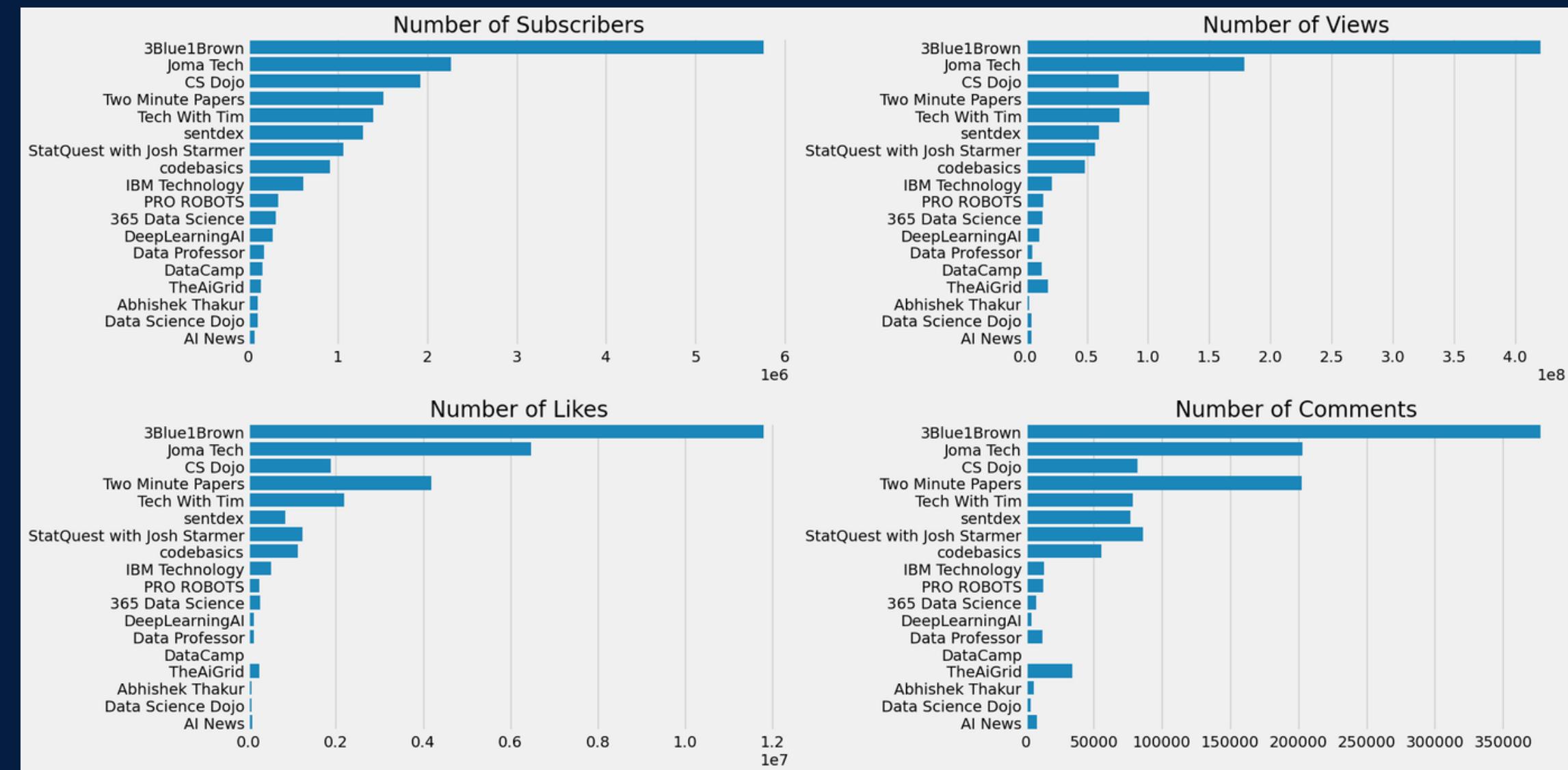


# Liệu kênh có nhiều lượt đăng kí thì sẽ nổi tiếng hơn?

Rõ ràng không hẳn là như vậy

Điển hình như kênh CS Dojo dù lượt  
đăng kí rất cao nhưng lại không nhận  
được nhiều lượt xem, like, comment

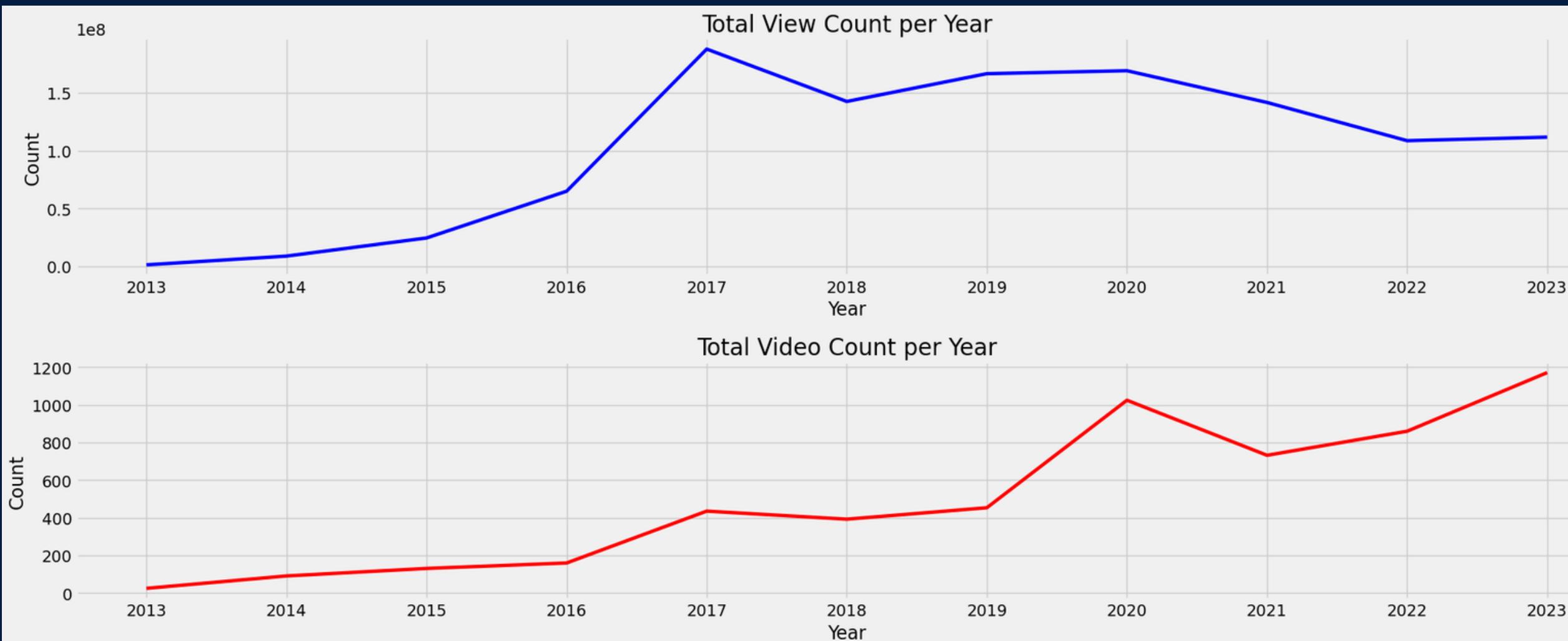
Ngược lại là kênh TheAIGrid, mặc dù  
lượt đăng kí năm ở top dưới cùng  
nhưng lại nhận được rất nhiều sự  
quan tâm



# Lượt xem của các kênh

Lượt xem đạt đỉnh điểm vào năm 2017 sau đó giảm dần

Lượng video xuất bản đạt đỉnh ở năm 2020 và 2023 nhưng lượt xem ở 2 năm này lại giảm so với năm trước đó



# Lượt xem của các kênh

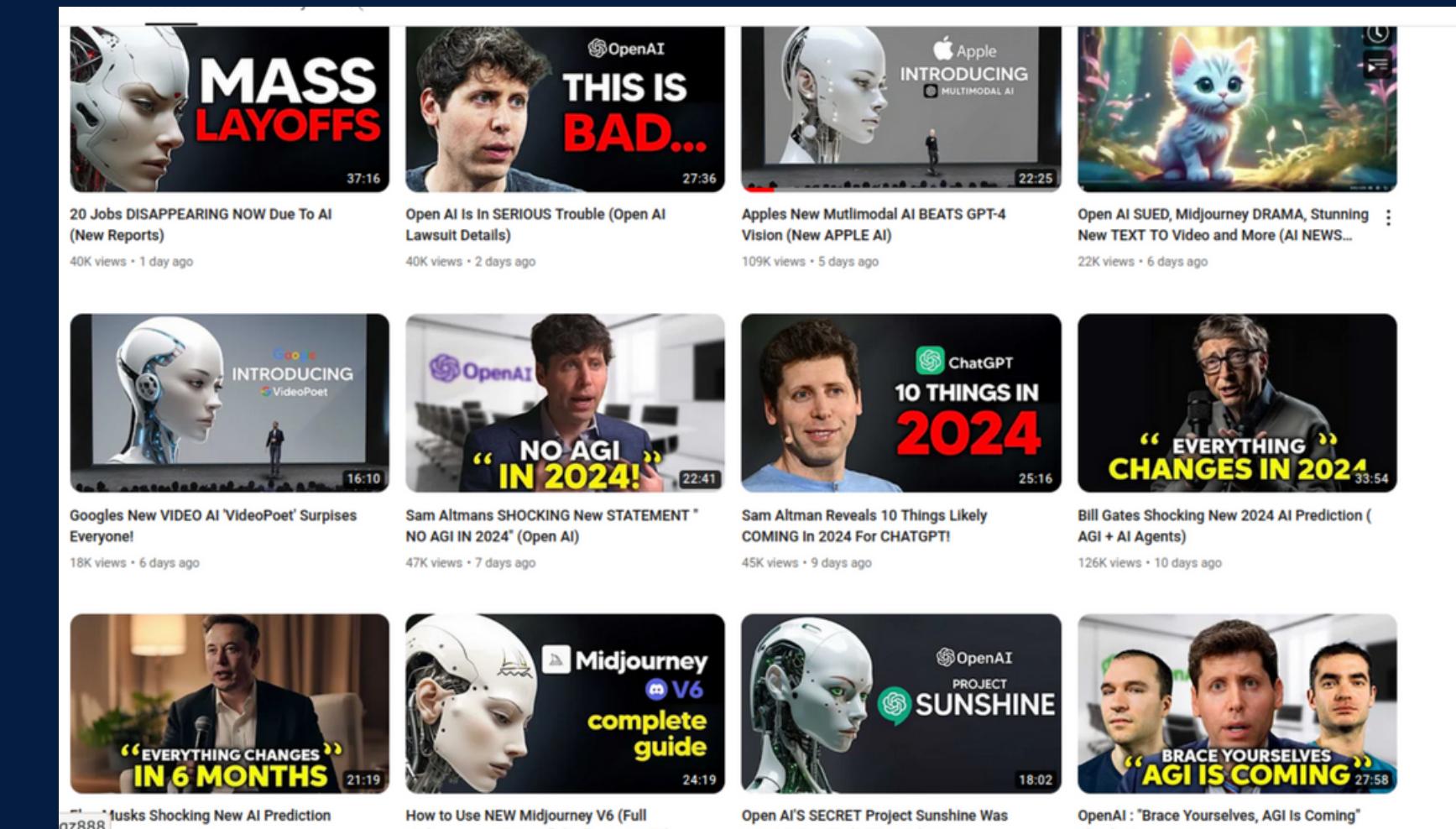
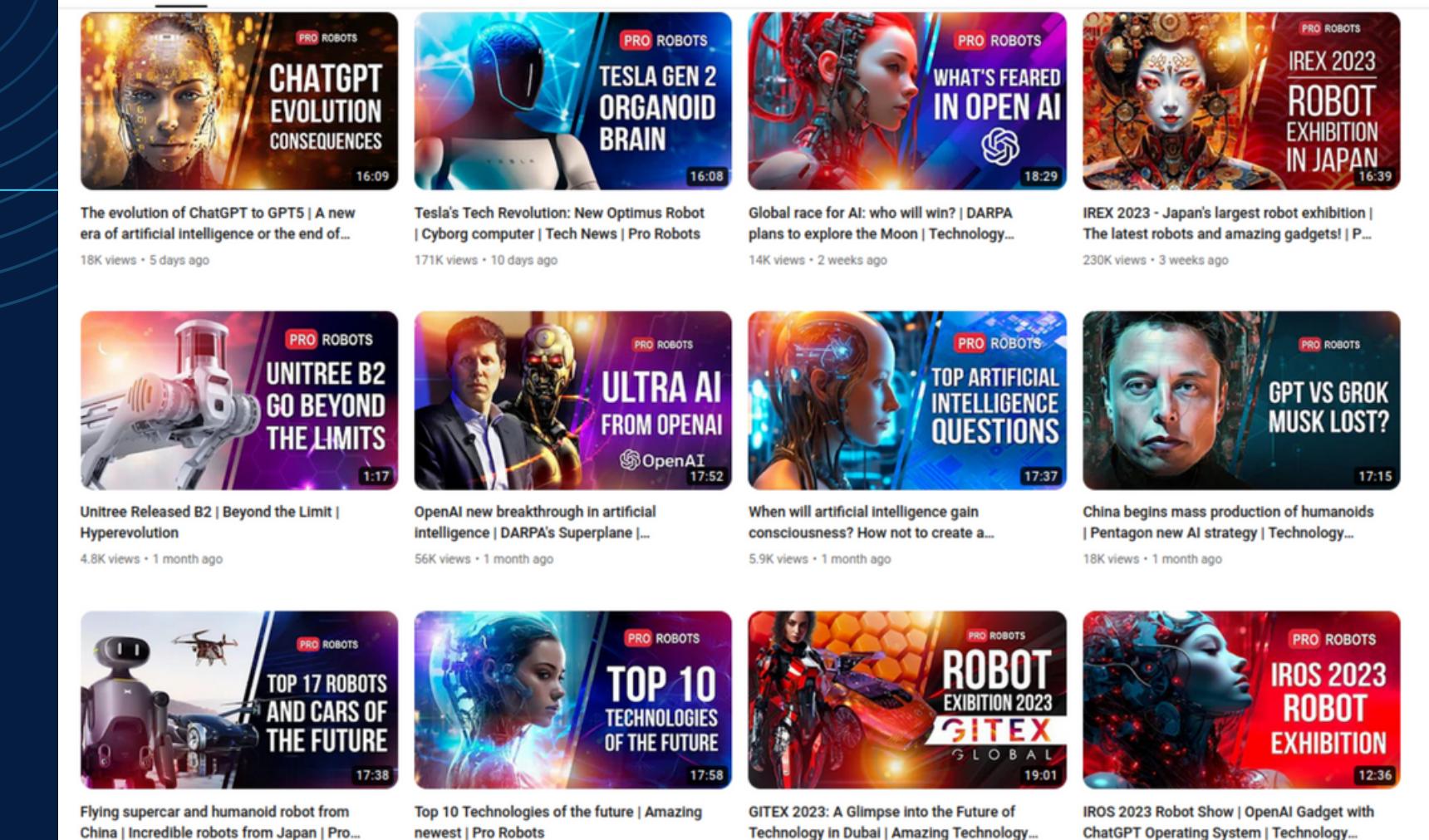
Hầu hết các kênh khác có một xu hướng chỉ tăng mạnh lượt xem ở một vài năm nhất định sau đó giảm dần

Đáng chú ý, kênh 'CS Dojo' ghi nhận không có lượt xem nào trong năm 2023. Tuy nhiên, những video cũ từ kênh này vẫn mang lại nhiều giá trị cho người xem.

# Lượt xem của các kênh

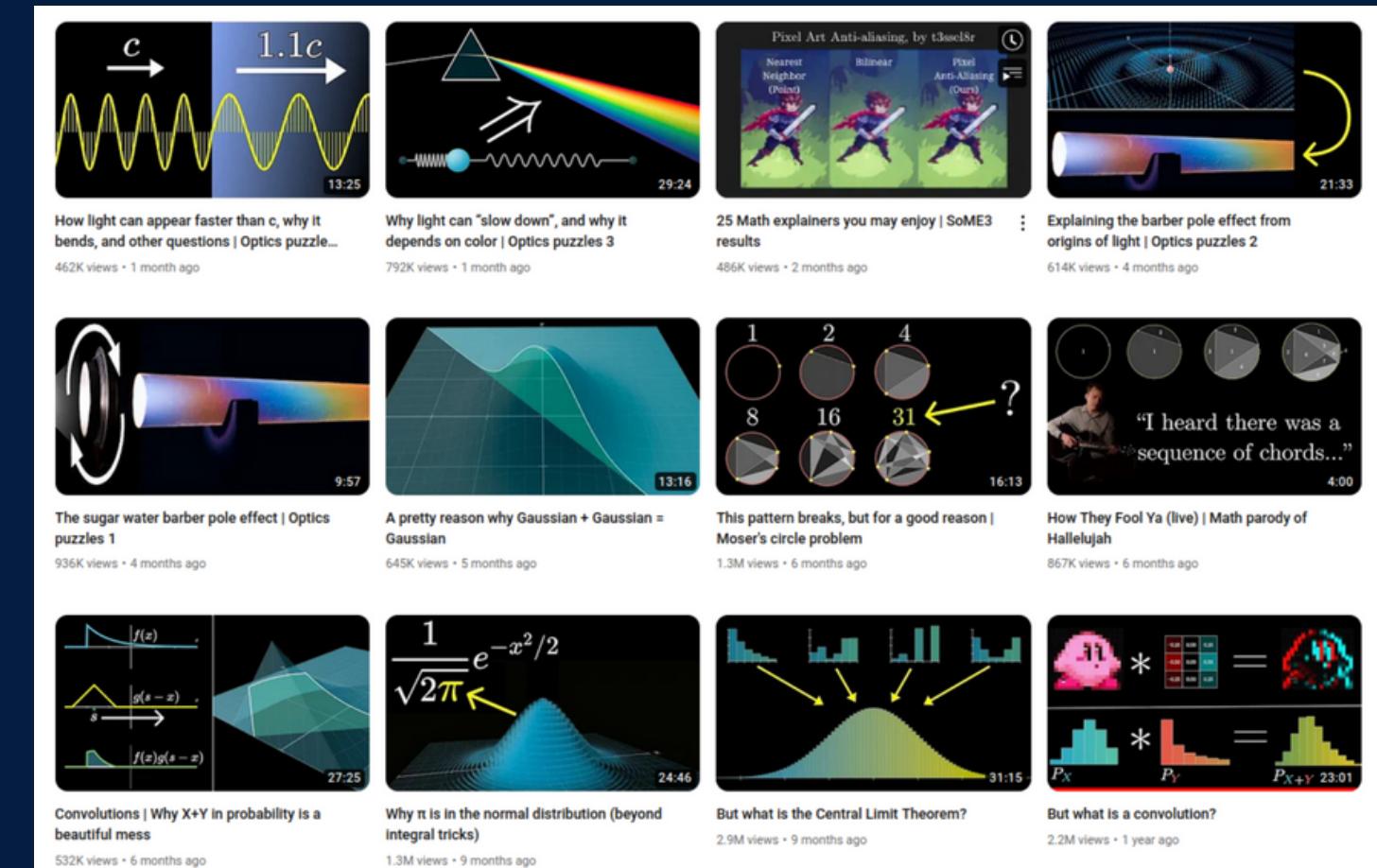
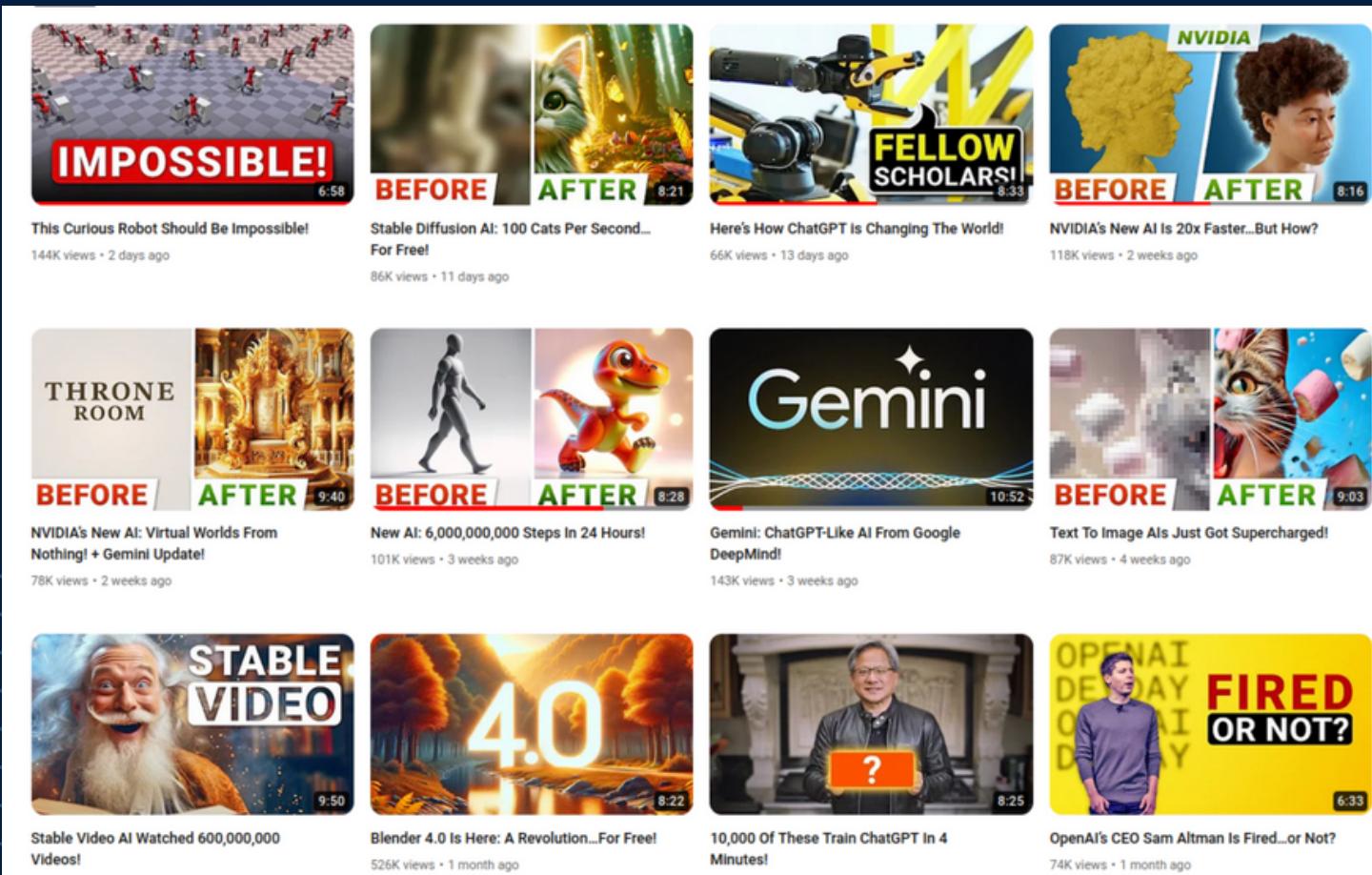
Một số kênh như 'PRO ROBOTS', 'AI News', và 'TheAiGrid' đã đạt được một lượng lớn lượt xem vào năm 2023 mặc dù mới thành lập gần đây. Điều này cho thấy nội dung hứa hẹn cho tương lai.

Các kênh này chủ yếu làm về các công nghệ, các chủ đề nóng hỏi trong lĩnh vực AI, PROBOTS, ...



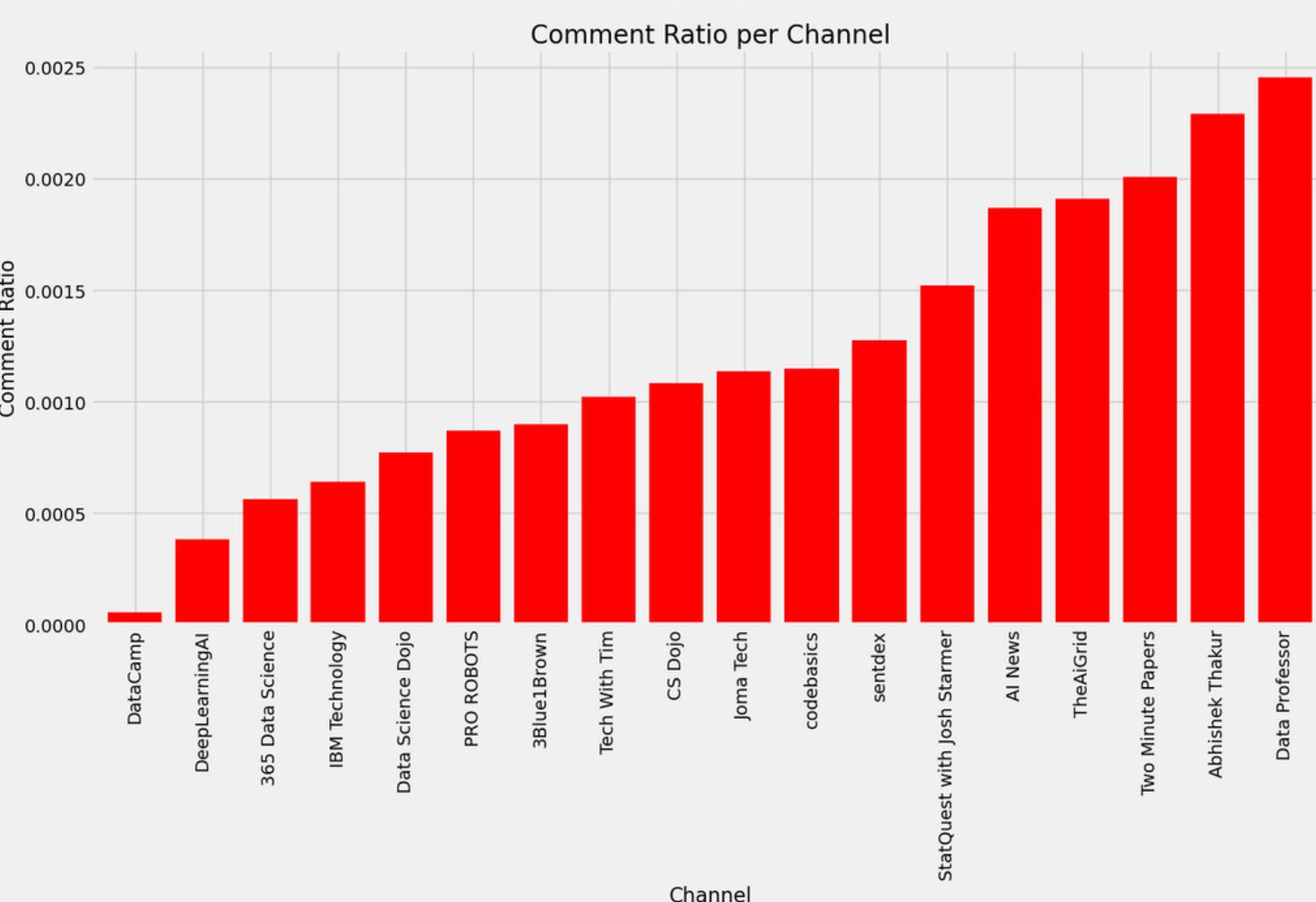
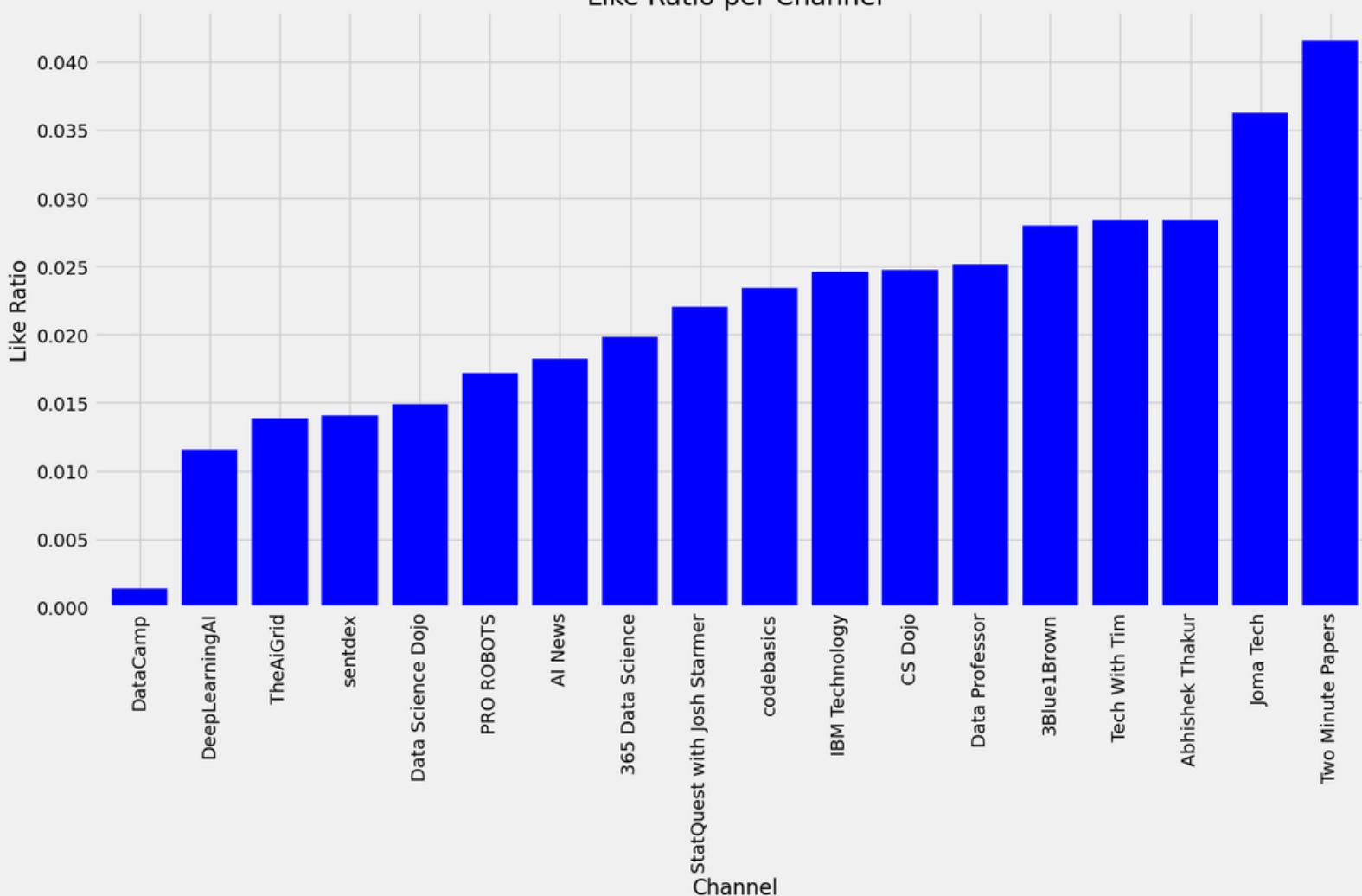
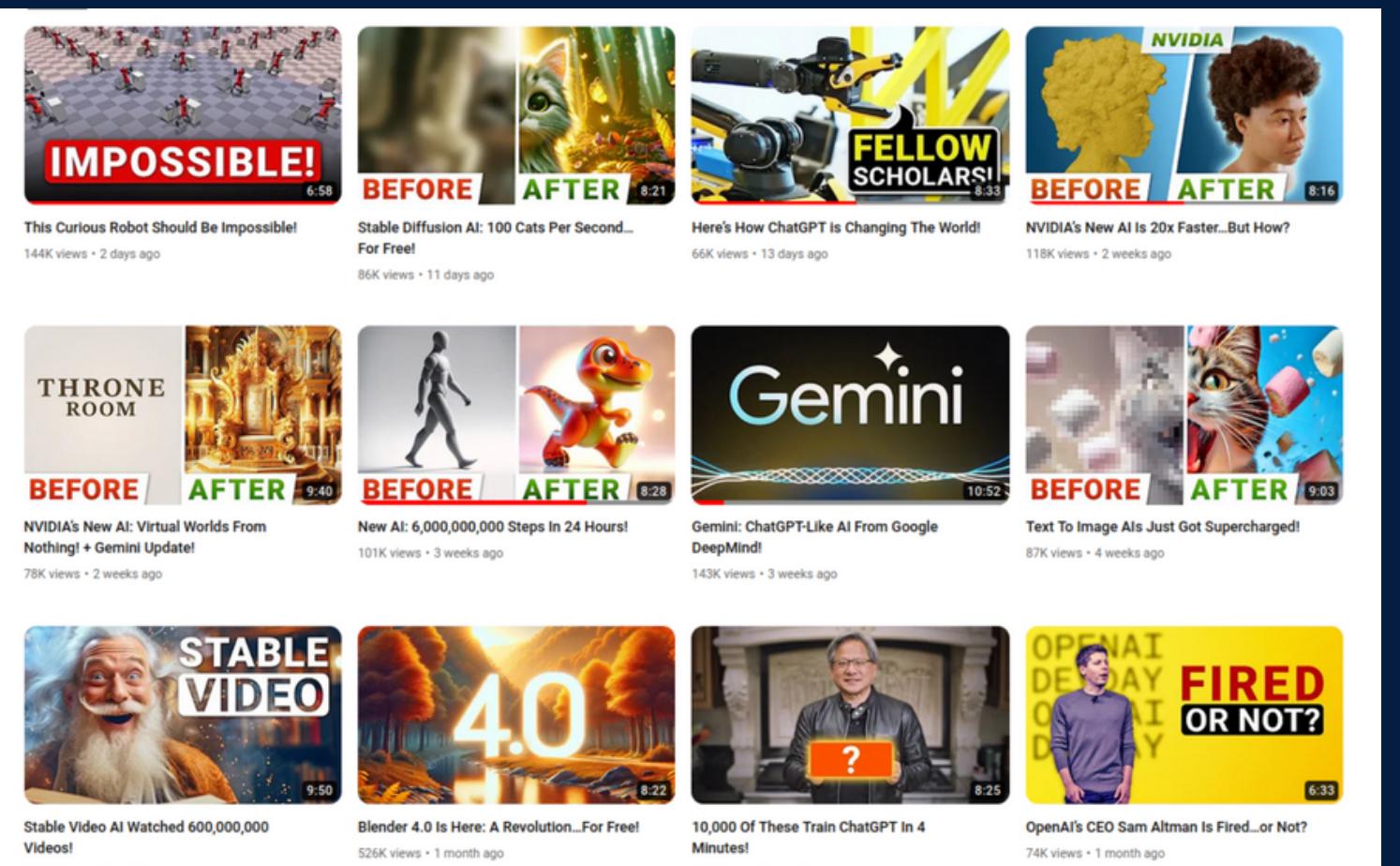
# Lượt xem của các kênh

Hai kênh, 'Two Minute Papers' và '3Blue1Brown,' liên tục duy trì một lượng lớn lượt xem đáng kể qua các năm. Điều này có lẽ do phong cách làm video của 2 kênh này cực kì cuốn hút và các chủ đề cũng rất hấp dẫn.



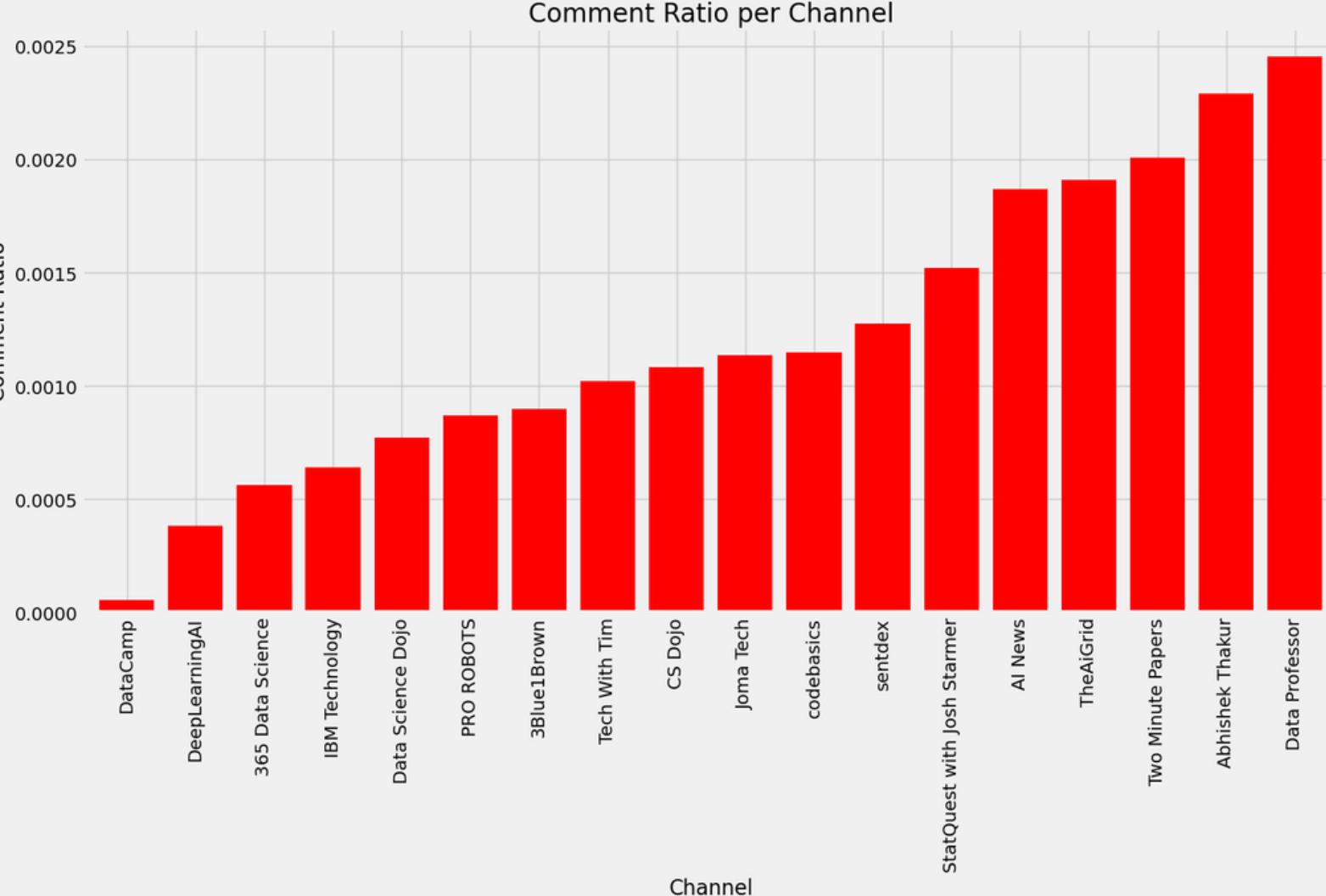
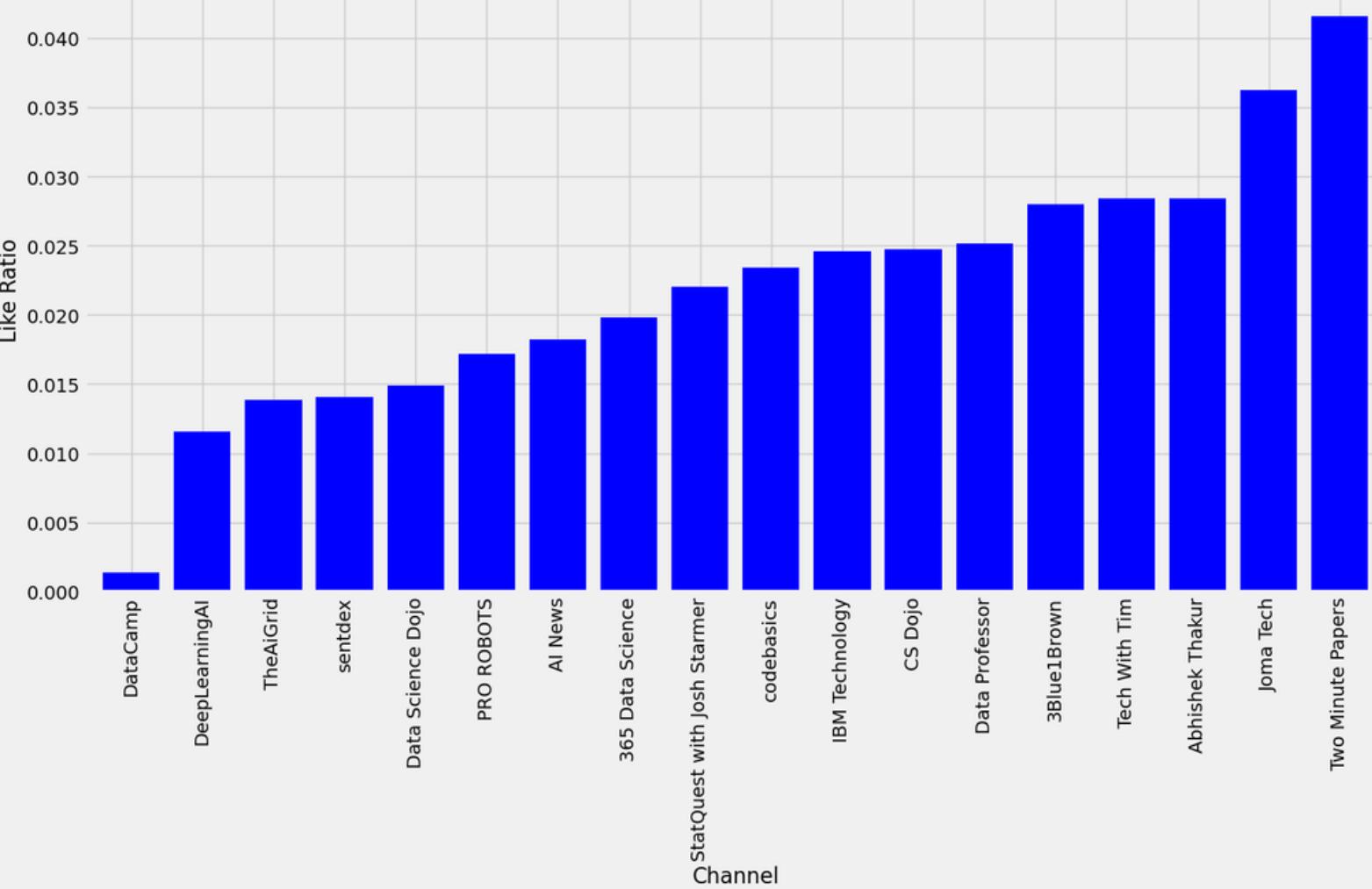
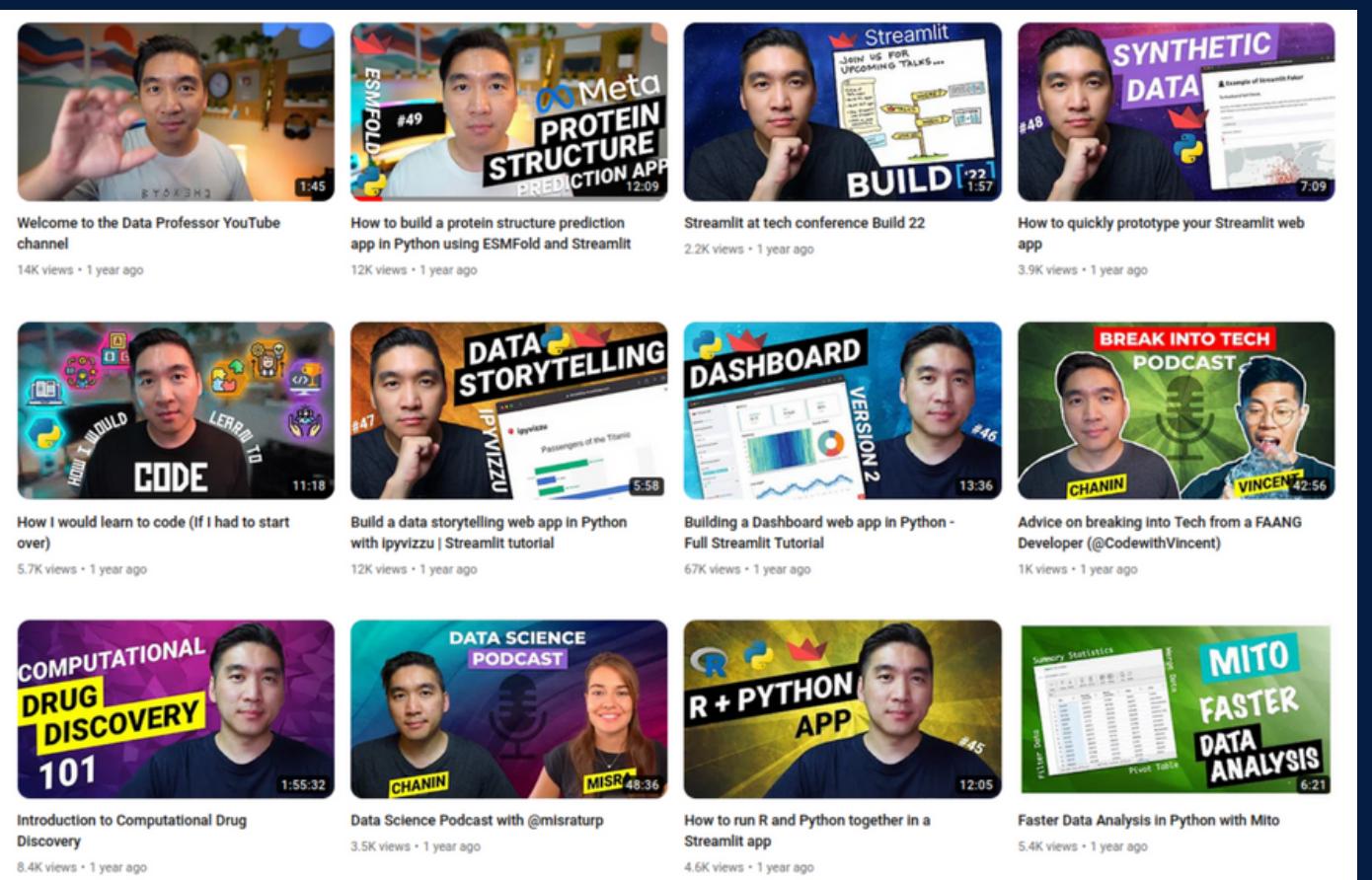
# Đâu là kênh thú vị nhất

- Không có gì ngạc nhiên khi Two Minute Papers có tỷ lệ lượt thích cao nhất.
- Chủ đề của kênh này thường là về các công trình nghiên cứu cực kì hữu ích về AI, machine learning
- Nhờ vào cách diễn đạt cô đọng, dễ hiểu mà kênh này đã nhận được rất nhiều lượt like và theo dõi



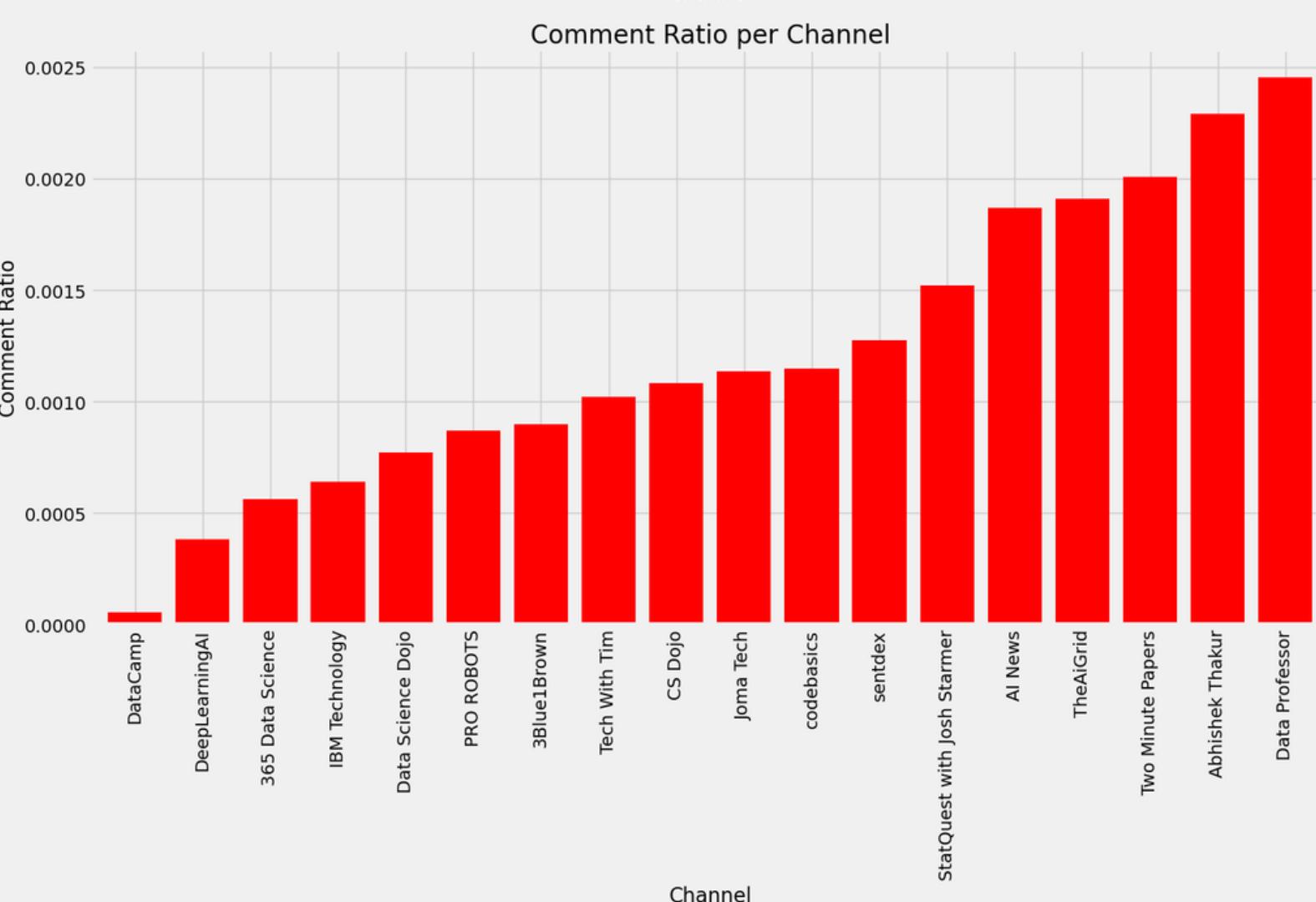
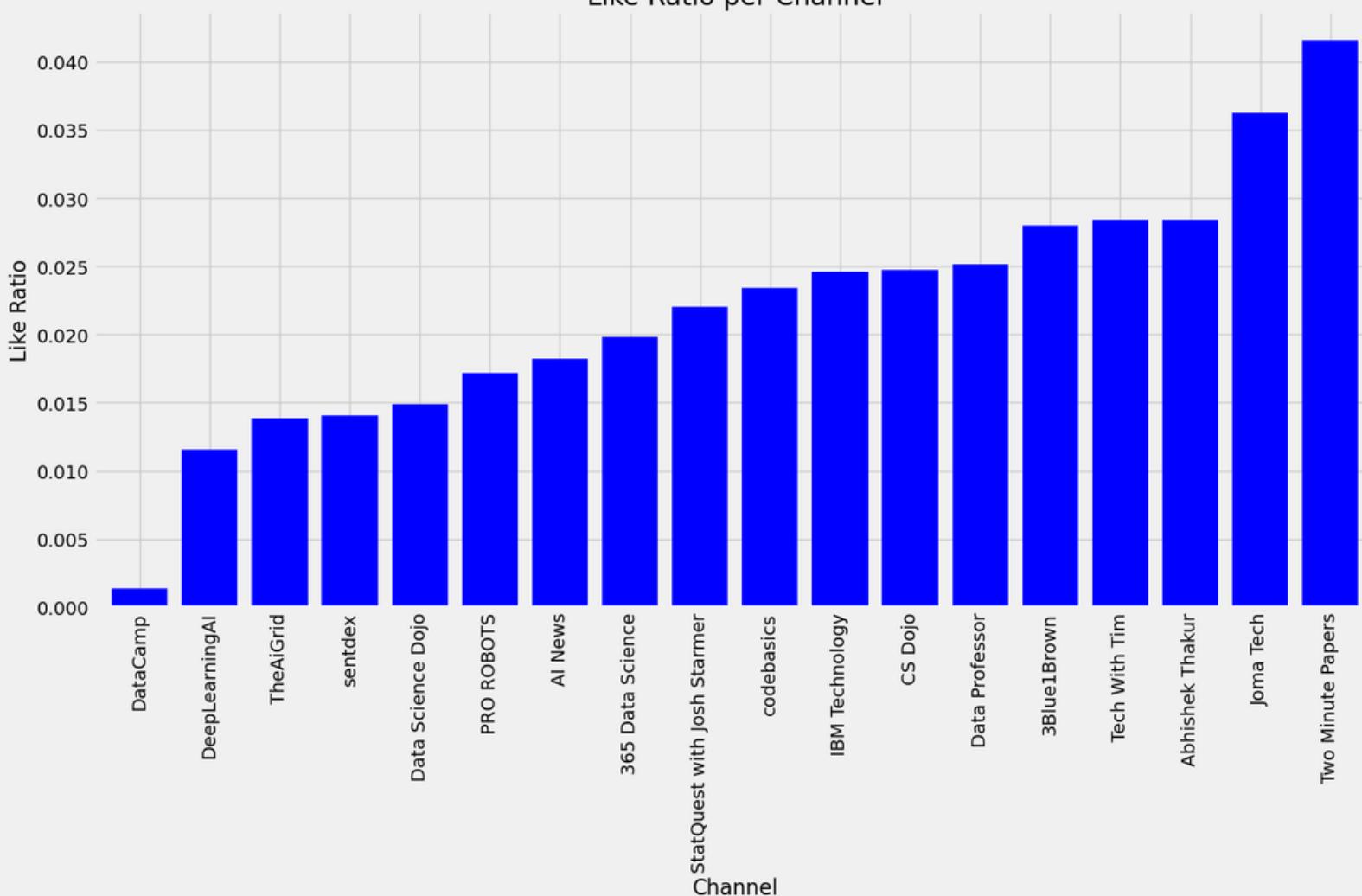
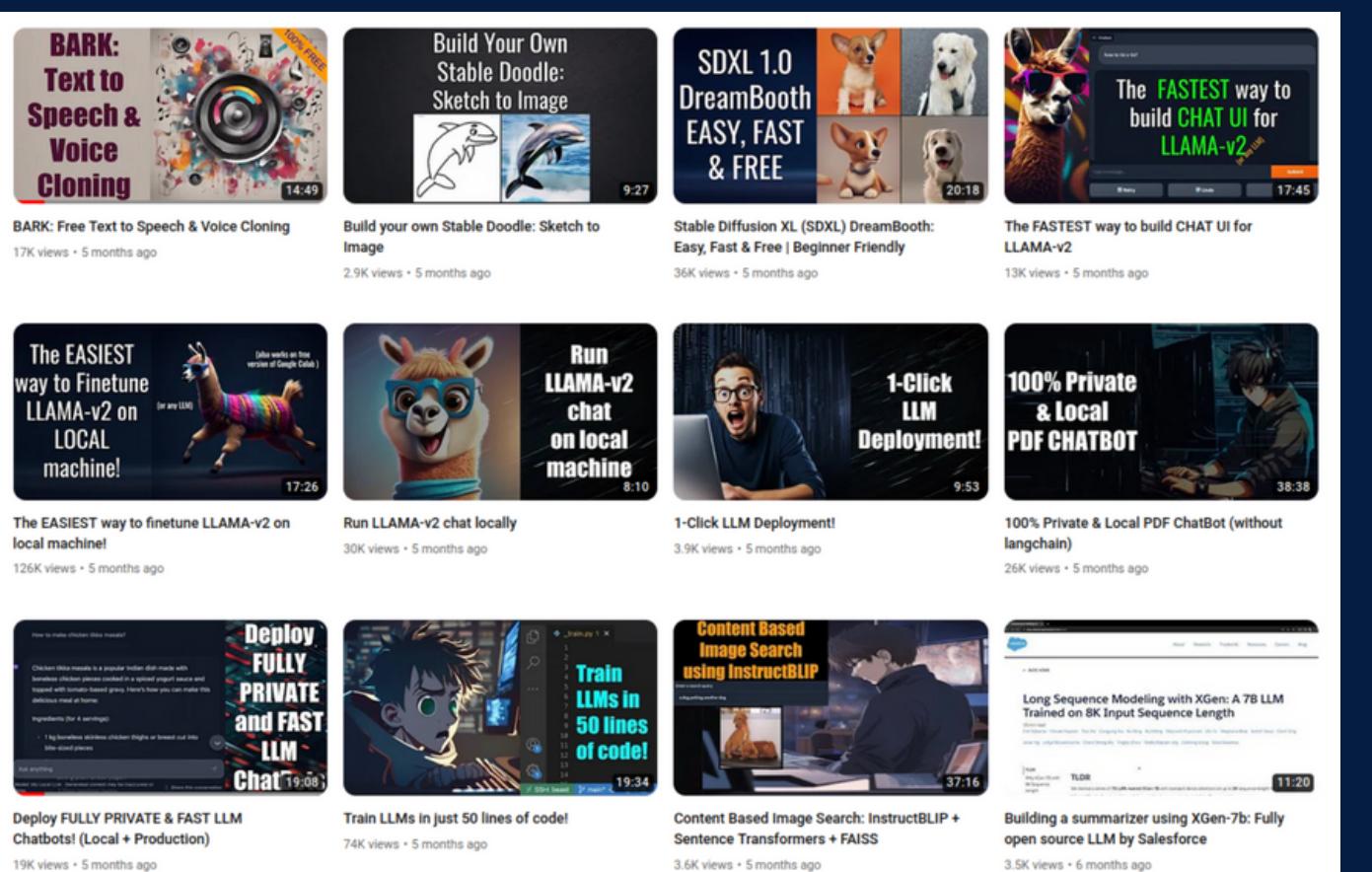
# Đâu là kênh thú vị nhất

- Khá ngạc nhiên khi Data Professor kênh có tỉ lệ comment cao nhất.
- Kênh này do một phó giáo sư tin sinh học tại một Đại học Nghiên cứu .
- Đa số video của ông tập trung vào việc giảng dạy các công cụ mới và hữu ích trong lĩnh vực khoa học dữ liệu
- Vì thế dù lượt xem không cao nhưng những comment nhận xét tích cực cho kênh này lại rất nhiều

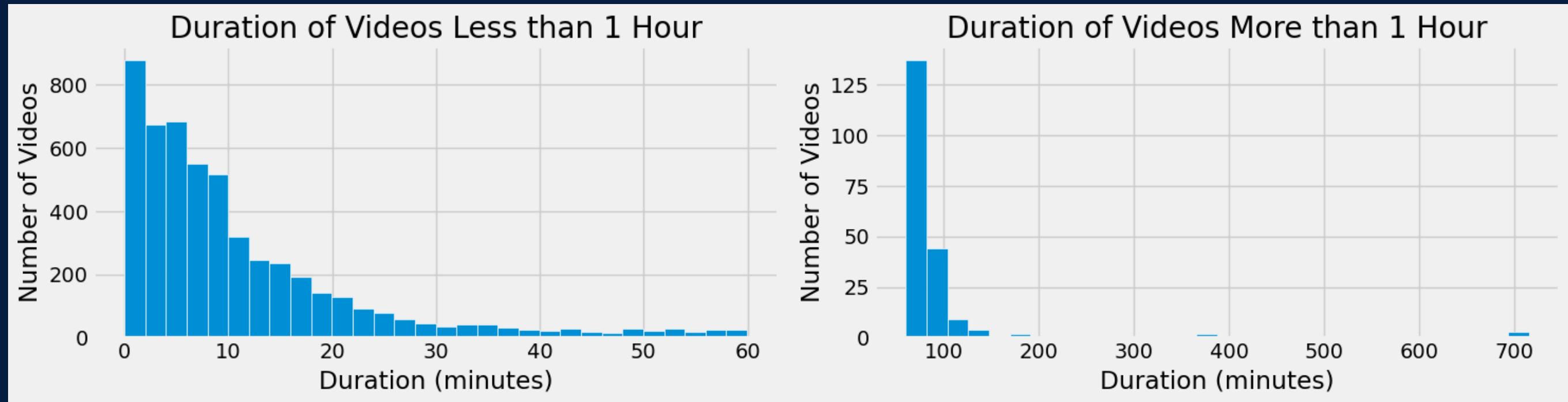


# Đâu là kênh thú vị nhất

- Một kênh đáng chú ý khác là Abhishek Thakur có tỷ lệ lượt thích và bình luận cực kỳ cao, dù cho chỉ là một kênh với lượng đăng ký không cao
- Chủ kênh là người đầu tiên đạt danh hiệu Quadruple Grand Master trên Kaggle.
- Các video chủ yếu tập trung vào việc lập trình các dự án, mô hình thú vị và sử dụng thư viện mới, từ đó tạo ra các chủ đề hấp dẫn để mọi người có thể cùng thảo luận



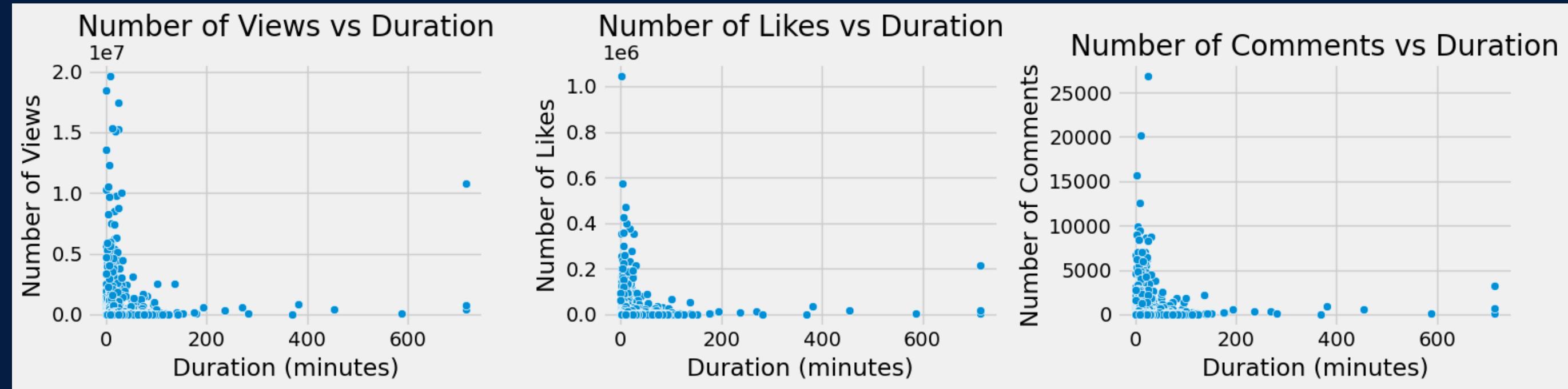
# THỜI LƯỢNG CỦA VIDEO



## HƠN 60%

Video có thời lượng dưới 10 phút

# THỜI LƯỢNG CỦA VIDEO



Đây có lẽ là xu thế làm video gần đây, các video ngắn dường như sẽ dễ được người dùng xem hơn, thường có lượt view, like, comment cao hơn (đặt biệt là shorts)

# NHỮNG VIDEO NÀO CÓ LƯỢT LIKE, COMMENT CAO NHẤT

	title	like_count	channelTitle
	I'm still astounded this is true	1043942.0	3Blue1Brown
	If Programming Was An Anime	572789.0	Joma Tech
a day in the life of an engineer working from home		469868.0	Joma Tech
	why you NEED math for programming	427596.0	Joma Tech
	The hardest problem on the hardest test	399799.0	3Blue1Brown

	title	comment_count	channelTitle
	But how does bitcoin actually work?	26833.0	3Blue1Brown
	The hardest problem on the hardest test	20199.0	3Blue1Brown
	If Programming Was An Anime	15632.0	Joma Tech
a day in the life of an engineer working from home		12569.0	Joma Tech
	why you NEED math for programming	9952.0	Joma Tech

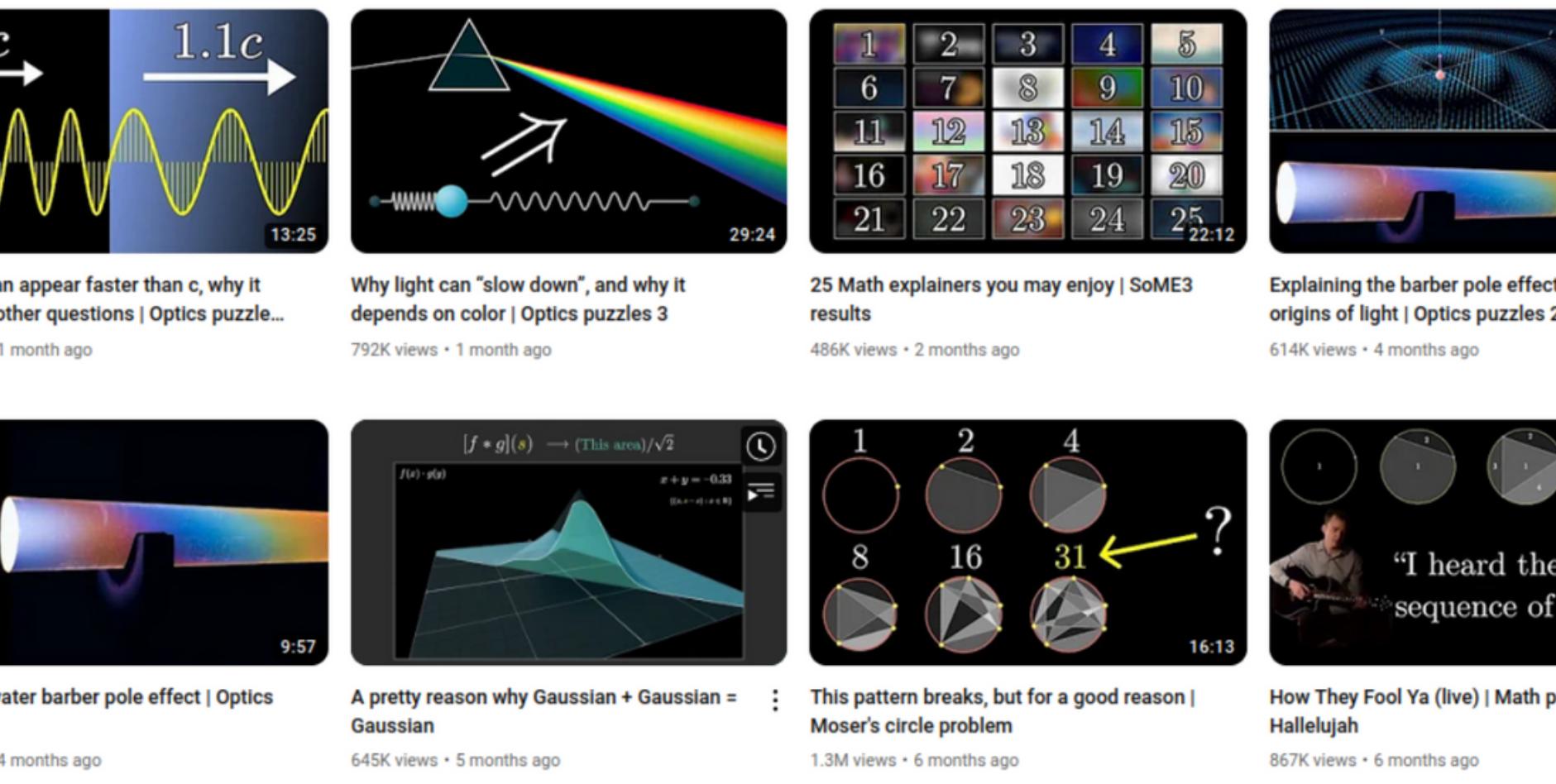
Không có gì là lạ khi top 5 đều thuộc về 2 channel là 3Blue1Brown và Joma Tech. Đây đều là những channel cực kì nổi tiếng và phổ biến

Những video có lượt like cao nhất

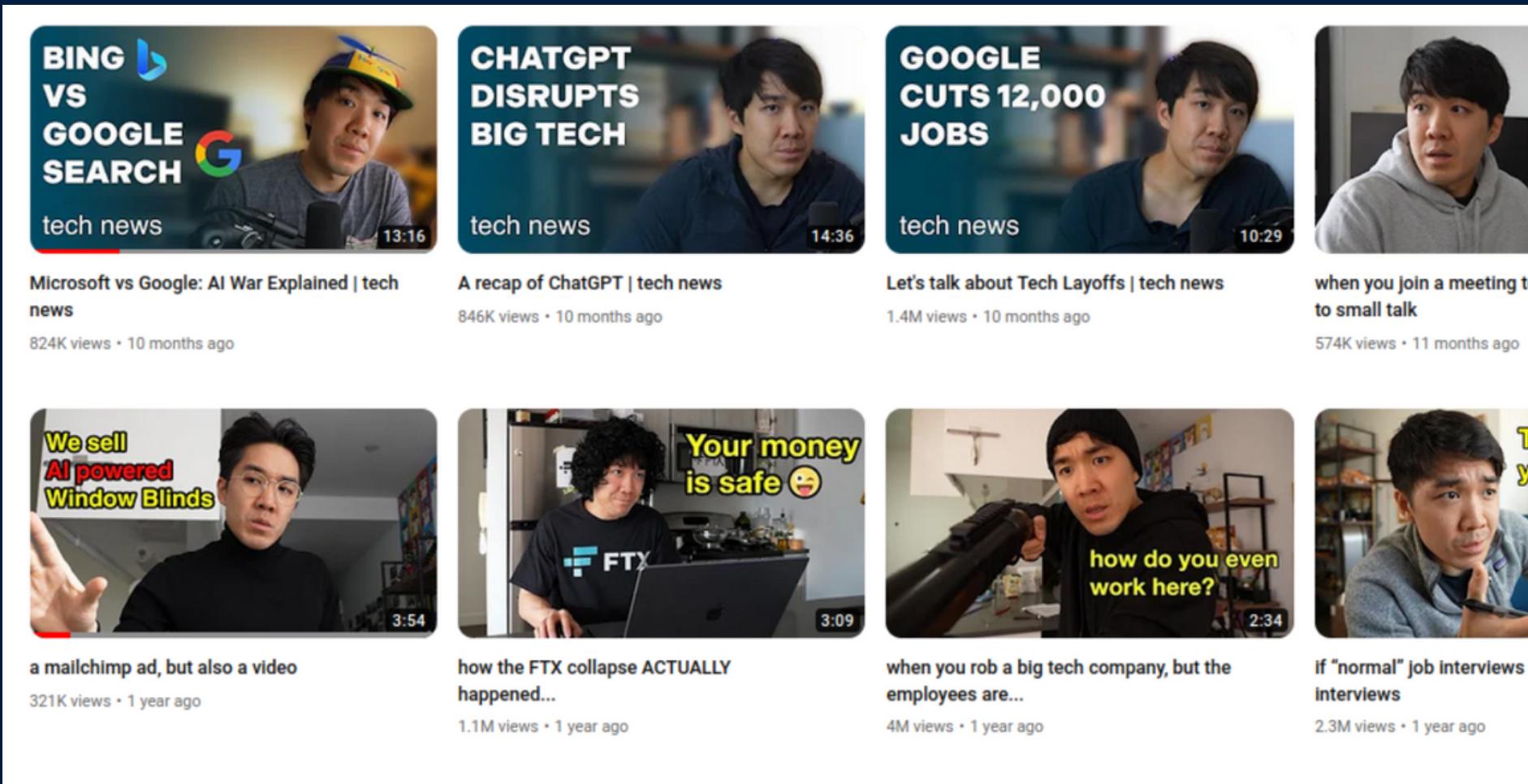
Những video có lượt comment cao nhất



# NHỮNG VIDEO NÀO CÓ LƯỢT LIKE, COMMENT CAO NHẤT



- 3Blue1Brown gần như là một channel phải xem cho những ai có hứng thú với toán học. Toán học được giảng dạy bằng hình ảnh trực quan chứ không còn đơn giản chỉ là những công thức nhảm chán
- Jome Tech là một kênh cực kì giải trí sử dụng chủ đề là cuộc sống của lập trình viên hằng ngày nhưng cũng không hề thiếu những kiến thức hữu ích cho lập trình viên



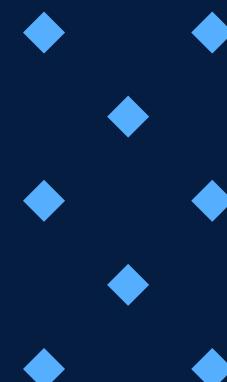
# NHỮNG VIDEO NÀO TỈ LỆ LIKE, COMMENT CAO NHẤT

	title	comment_ratio	channelTitle
	What tech topics are YOU interested in?	0.044248	IBM Technology
	NVIDIA GTC21 (The AI Conference) is FREE + Course Giveaway	0.043753	Data Professor
1 Year on YouTube as the Data Professor (Data Science YouTube Channel)		0.036600	Data Professor
	StatQuest: 10,000 Subscriber Milestone	0.033241	StatQuest with Josh Starmer
	NVIDIA GTC 21 & a Giveaway !	0.032083	codebasics

Những video có tỉ lệ lượt like cao nhất

	title	like_ratio	channelTitle
	IBM Tech Now: IBM Wazi as a Service, IBM Spectrum Sentinel, and the G2 Summer Reports	0.194553	IBM Technology
	8 Portals for Datasets 😊 #codebasics #shorts #dataanalysis #data	0.160255	codebasics
	IBM Tech Now: IBM Cloud Prep, the IBM Consulting Cloud Accelerator, and the Gartner Magic Quadrant	0.137203	IBM Technology
	15 Design Rules for BI Dashboard! 📈⭐ #codebasics #shorts #dataanalysis #data	0.122257	codebasics
	Omdena can help you get an Interview Call! ☎️😊 #codebasics #shorts #dataanalysis #data	0.117681	codebasics

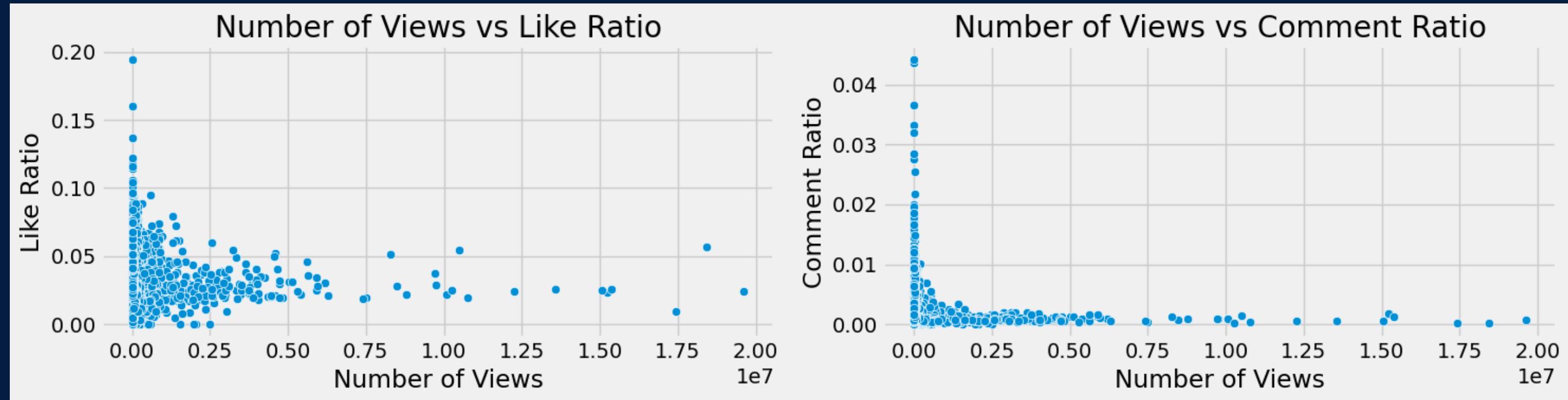
Những video tỉ lệ like, comment cao hầu như là làm về các công nghệ mới hoặc các tips trong lập trình và interview



# NHỮNG VIDEO NÀO TỈ LỆ LIKE, COMMENT CAO NHẤT

Có thể thấy các kênh có lượng like, comment cao gần như biến mất khỏi top các video có tỉ lệ like comment.

Điều này một phần là do thói quen like, comment của người dùng còn hạn chế dẫn tới việc lượt xem tăng cao kéo theo tỉ lệ like, comment giảm xuống

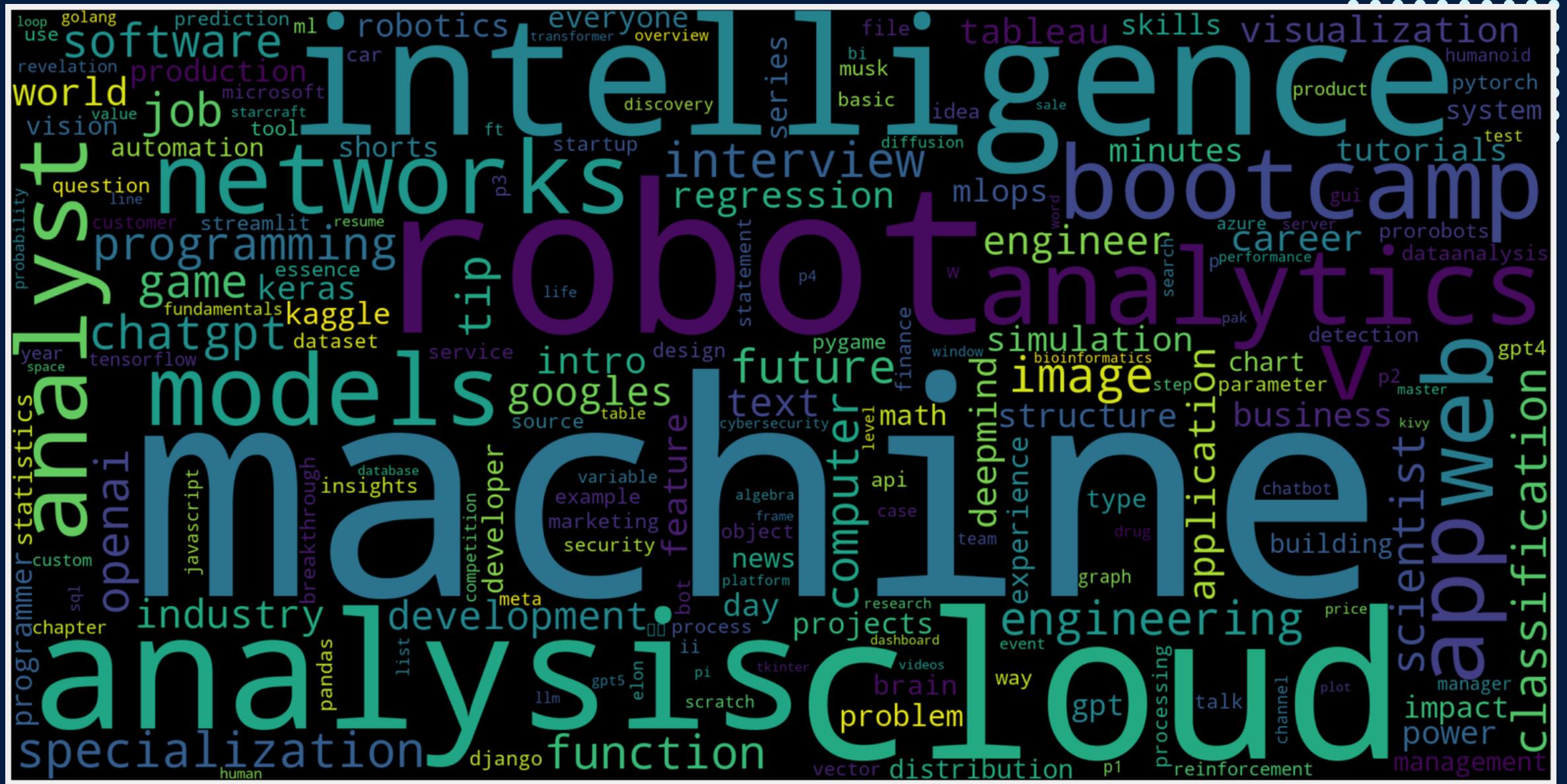


# CÁC KÊNH THƯỜNG ĐẶT TIÊU ĐỀ NHƯ THẾ NÀO

## CÁC BƯỚC TẠO WORD CLOUD:

- 01 Loại bỏ các kí tự đặc biệt
- 02 Loại bỏ tên của kênh xuất hiện trong tiêu đề
- 03 Sử dụng thư viện nltk đánh tag cho từng từ trong tiêu đề
- 04 Chỉ lấy ra danh từ (ứng với tag NN, NNS, NNP, NNPS)
- 05 Loại đi 1 vài từ quá thông dụng (ai, python, ...)
- 06 Dùng thư viện word cloud để vẽ

# CÁC KÊNH THƯỜNG ĐĂT TIÊU ĐỂ NHƯ THẾ NÀO



- Những chủ đề thường xuyên được xuất hiện là robot, machine, ananlysic, cloud, .. đều là những chủ đề hết sức quen thuộc.
  - Cho thấy các kênh thường chọn các chủ đề an toàn, hầu như đã được phổ biến rộng rãi

# CÁC VIDEO THƯỜNG ĐƯỢC PHẢN ỨNG NHƯ THẾ NÀO?

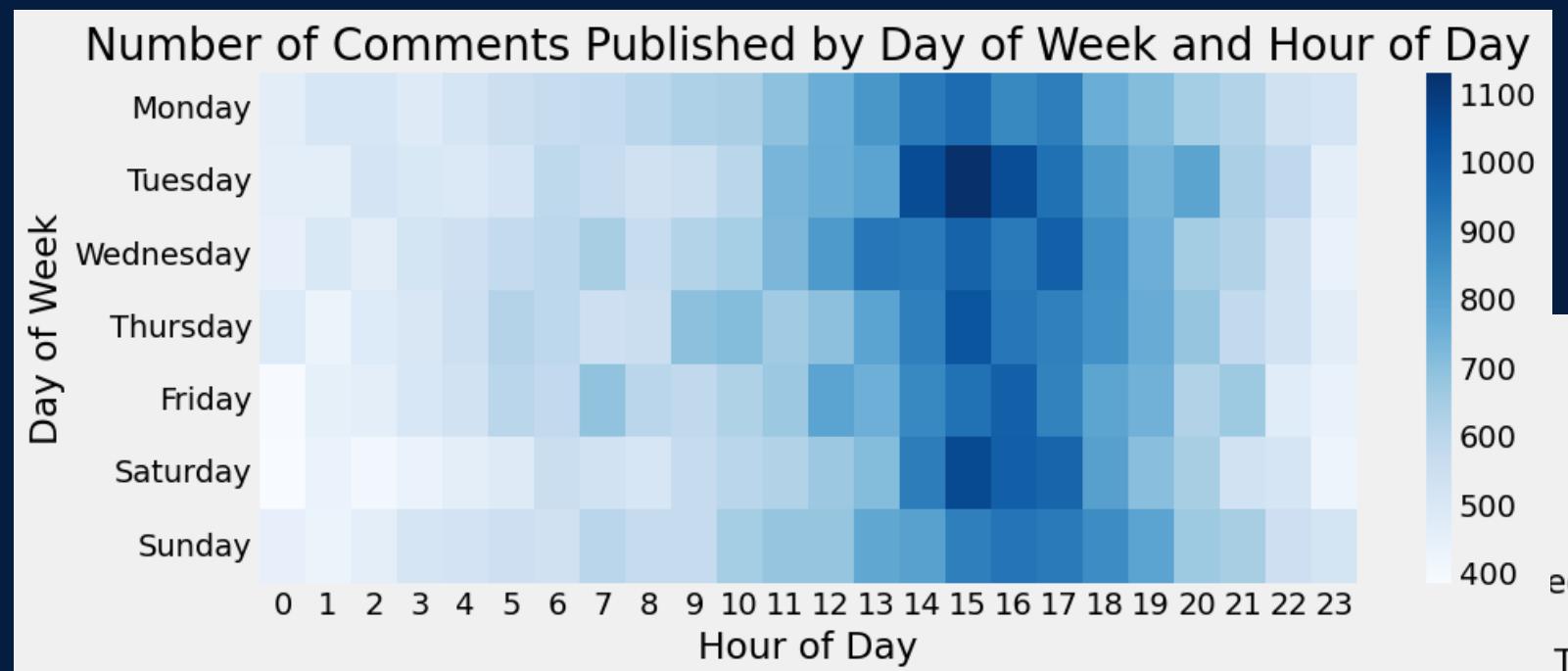
Tương tự như wordcloud cho title nhưng lúc này, ta chỉ chọn ra các tính từ có tag JJ

Hầu hết các comment đều để lại nhận xét tốt. Do hầu hết các video là về giáo dục, công đồng người xem phần lớn đều có văn hóa ứng xử tốt.

Bên cạnh đó một vài nhận xét không tốt như bad, wrong cũng không quá nặng nề và người sáng tạo hoàn toàn có thể lắng nghe để phát triển kênh

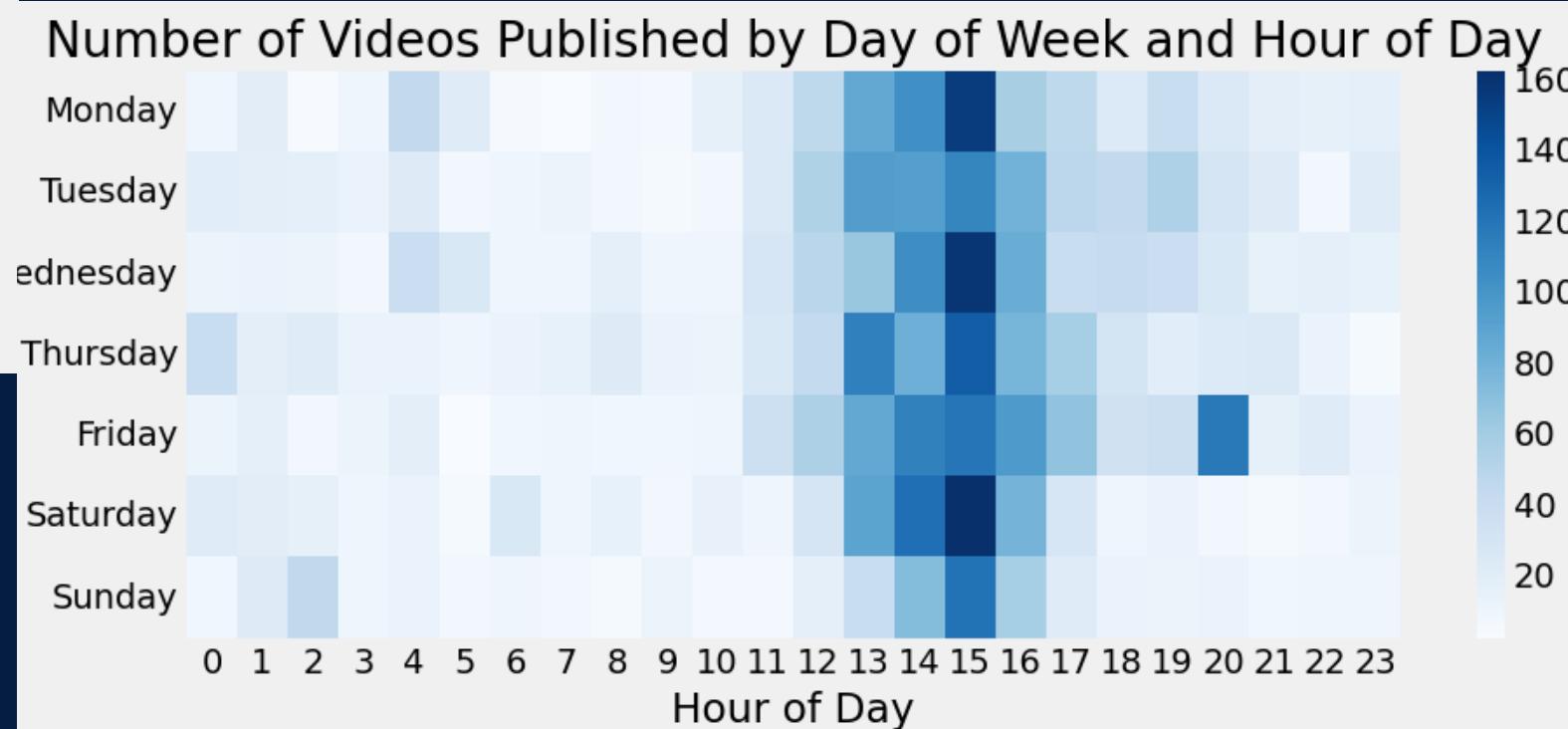


# MỐI QUAN HỆ GIỮ THỜI GIAN XUẤT BẢN VIDEO VÀ THỜI GIAN NGƯỜI XEM COMMENT



Điều này chứng tỏ hầu hết những người comment video là những người có follow kênh, thường vào xem và comment ngay sau khi video được publish

Thời gian người dùng comment khá là tương đồng với thời gian video được publish



# **SENTIMENT ANALYSIS**

**SỬ DỤNG PRE-TRAIN MODEL -  
BERTWEET-BASE-SENTIMENT-ANALYSIS  
ĐỂ PHÂN LOẠI COMMENT THÀNH 3 LOẠI**

neutral 52656 comments

positive 39701 comments

negative 18052 comments

# SENTIMENT ANALYSIS

video_id	NEG	NEU	POS		title	channelTitle
ER2It2mlagI	1.0	17.0	82.0	Neural Network Simply Explained   Deep Learning Tutorial 4 (Tensorflow2.0, Keras & Python)		codebasics
UnVyNh6P6FQ	1.0	17.0	82.0		The END of the Journey	codebasics
zfiSAzpy9NM	8.0	16.0	76.0	Simple explanation of convolutional neural network   Deep Learning Tutorial 23 (Tensorflow & Pyt...)		codebasics
VhRtaziEWd4	2.0	22.0	76.0		What is a neuron?   Deep Learning Tutorial 3 (Tensorflow Tutorial, Keras & Python)	codebasics
cdDD5t9r98c	5.0	20.0	75.0		How my health struggle inspired me to start YouTube Channel?	codebasics

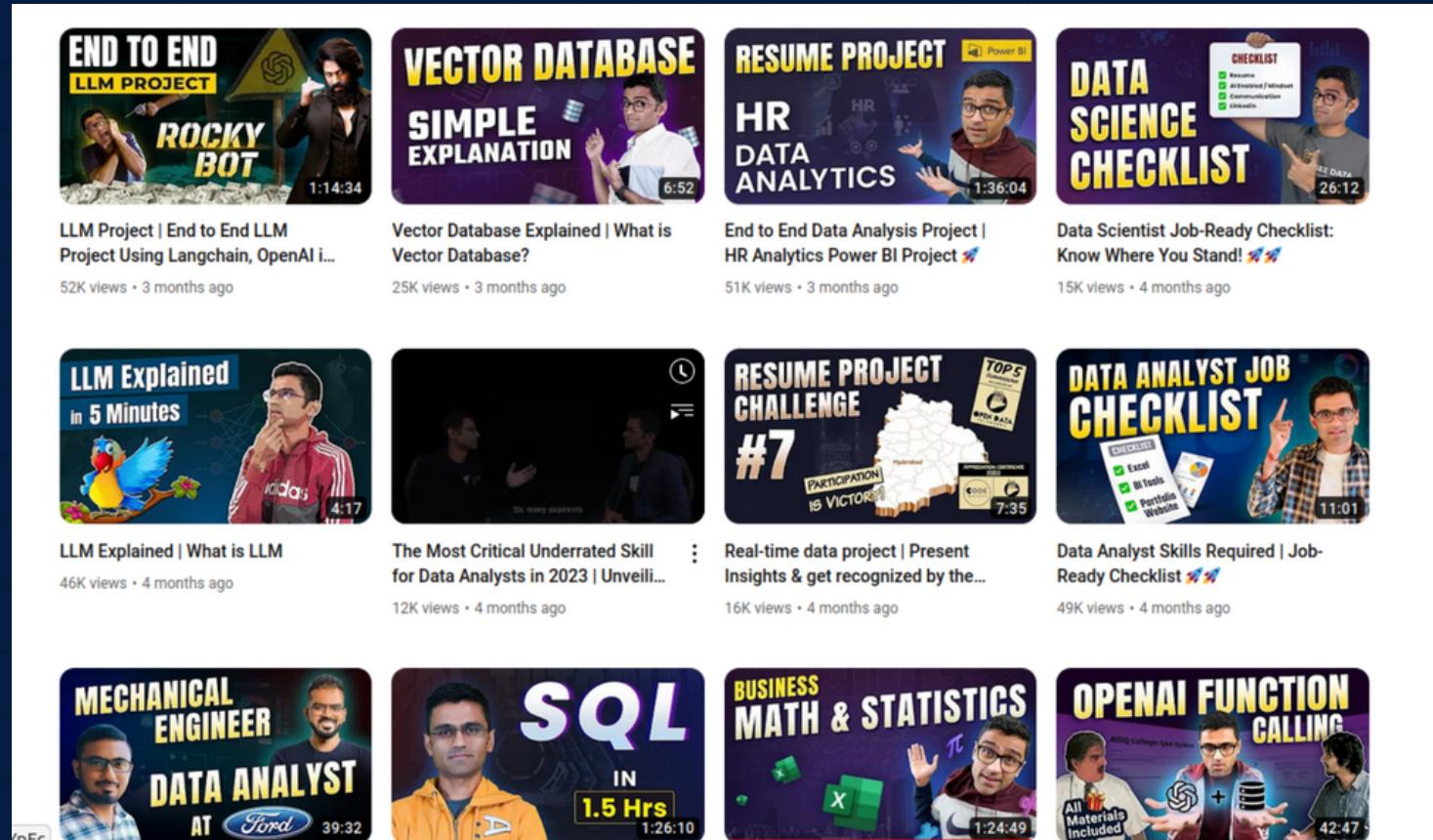
Những video với nhiều comment tích cực nhất

Năm video có số lượng bình luận tích cực cao nhất đều thuộc về 'codebasics'. Xem xét kĩ các video này hầu hết các bình luận chủ yếu là "simple and intuitive," "Thank you," and "Awesome explainer", ...

# SENTIMENT ANALYSIS

Điều này có lẽ là do phong cách dạy cực kì sáng tạo, dễ hiểu của chủ kênh. Cùng với đó là ông cực kì chịu khó tương tác với người xem thông qua comment. Việc này đã để lại thiện cảm lớn cho nhiều người xem

Codebasic chuyên dạy về ngôn ngữ lập trình cơ bản, đặc biệt là trong lĩnh vực khoa học dữ liệu, machine learning. Đây có thể là một lựa chọn tốt cho những ai đang tìm hiểu các hướng dẫn để học công nghệ mới cho mình.



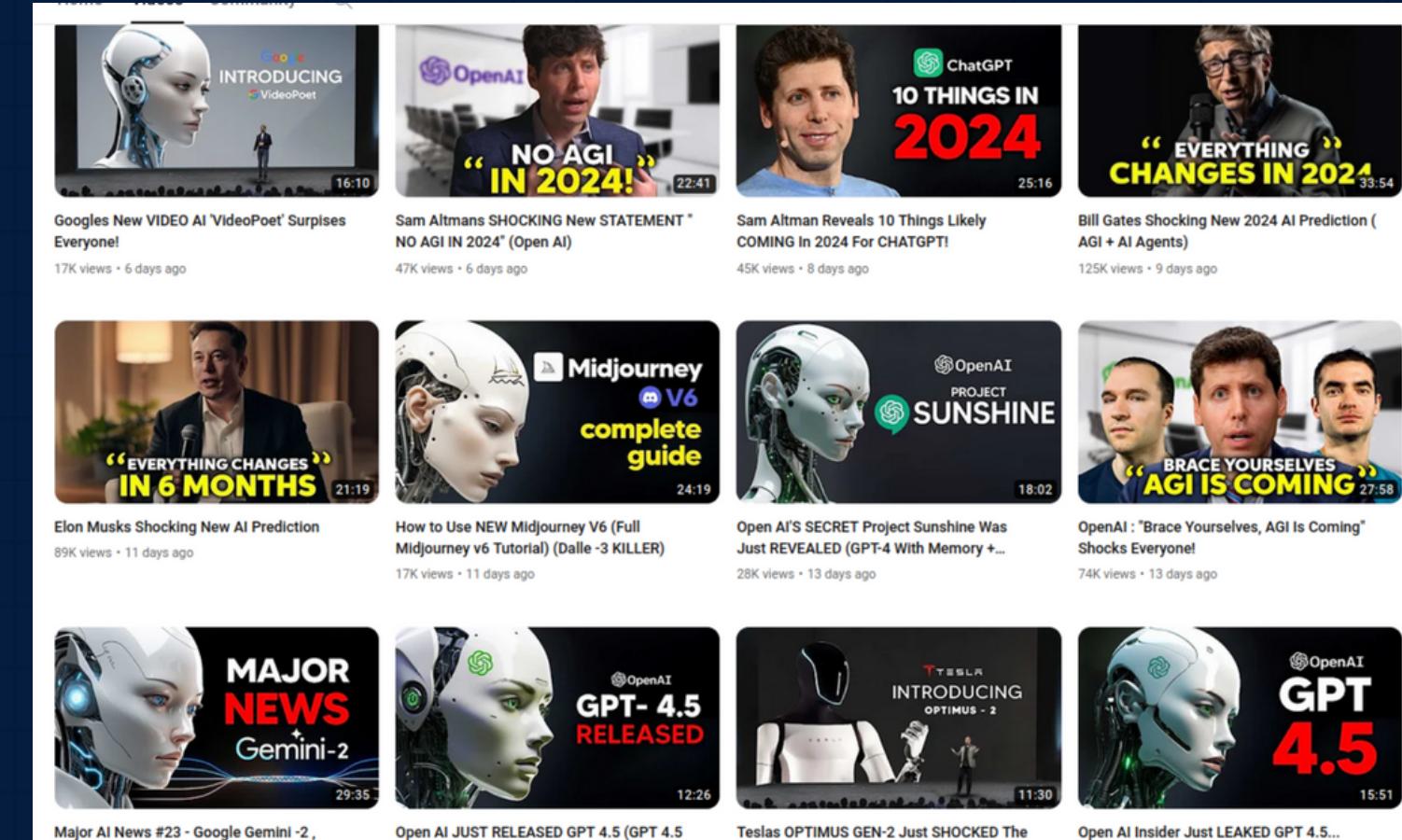
# SENTIMENT ANALYSIS

Những video với nhiều lượt đánh giá tiêu cực nhất

Top 5 đều thuộc về TheAiGrid, một kênh youtube khá mới và đang được nhận khá là nhiều sự quan tâm

video_id	NEG	NEU	POS		title	channelTitle
89yPjWjss7Y	61.0	23.0	16.0	AI Drone Kills Operator, GPT 4.2 Leaks, Bard Gets SUPERCHARGED, And Much More [AI NEWS #5]		TheAiGrid
ucp49z5pQ2s	52.0	36.0	12.0	Open AI CEO STUNS Everyone With Statements On GPT 5(GPT-5 Update)		TheAiGrid
r5JvhWpNbqM	51.0	42.0	7.0	MICROSOFTS New AGI JARVIS SHOCKS The Entire Industry! (FINALLY ANNOUNCED!)		TheAiGrid
XTzOxMq-Qk0	50.0	36.0	14.0	Open AI Team STUNS Everyone With NEW Statement On Artificial Superintelligence		TheAiGrid
Wf-s9C9uf7U	49.0	34.0	17.0	6 Minutes Ago: Godfather Of AI Shared Terrifying Message About Artificial Intelligence		TheAiGrid

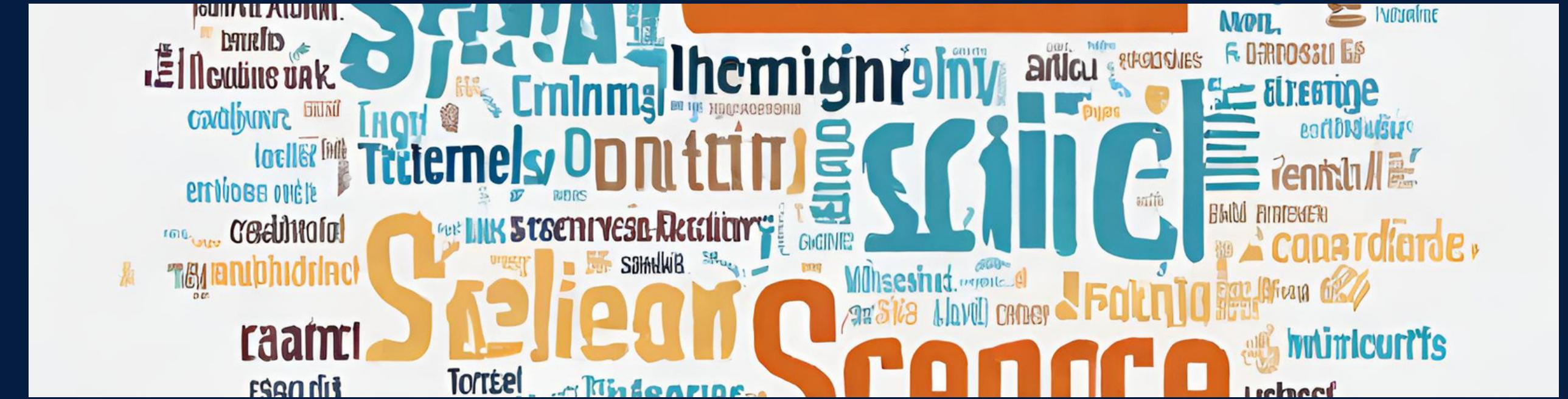
# SENTIMENT ANALYSIS



Điều này có thể lý giải là do kênh này thường đưa ra các chủ đề khá mới và còn gây nhiều tranh cãi

Nhưng điều này hoàn toàn không phải là một tín hiệu xấu. Các chủ đề gây tranh cãi thường rất thú vị và thu hút người xem. Bên cạnh đó các chủ đề như thế này cũng khá là mới lạ, độc đáo và có thể xem như một lợi thế để đưa kênh này phát triển hơn

# ĐÂU LÀ CÁC HOT KEYWORD XUẤT HIỆN QUA CÁC NĂM?



01

02

03

Tạo một file các keywords liên quan đến data science, AI (trừ các từ mang tính tổng quát)

Tính toán tần suất của các keywords này

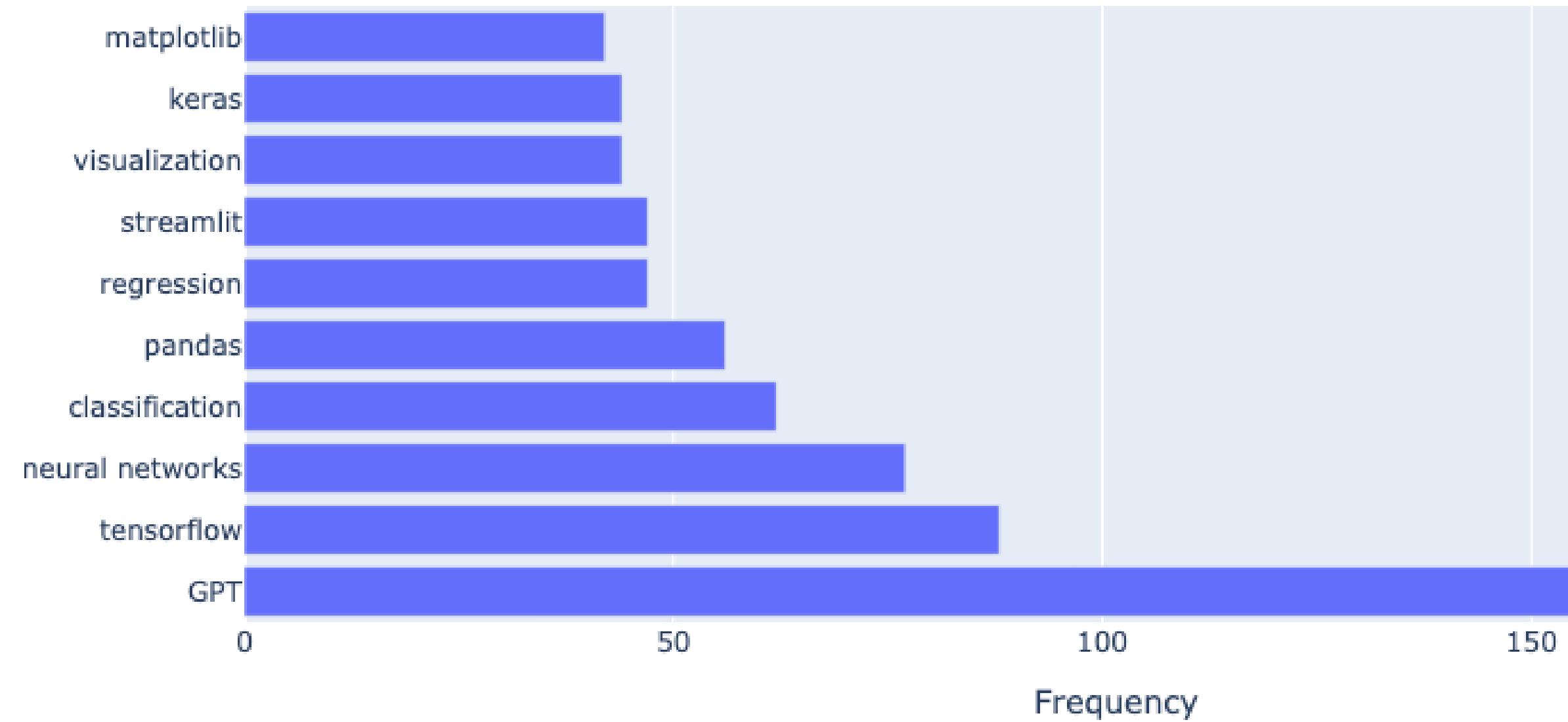
**Chọn ra top 5 keywords để đánh giá trend**

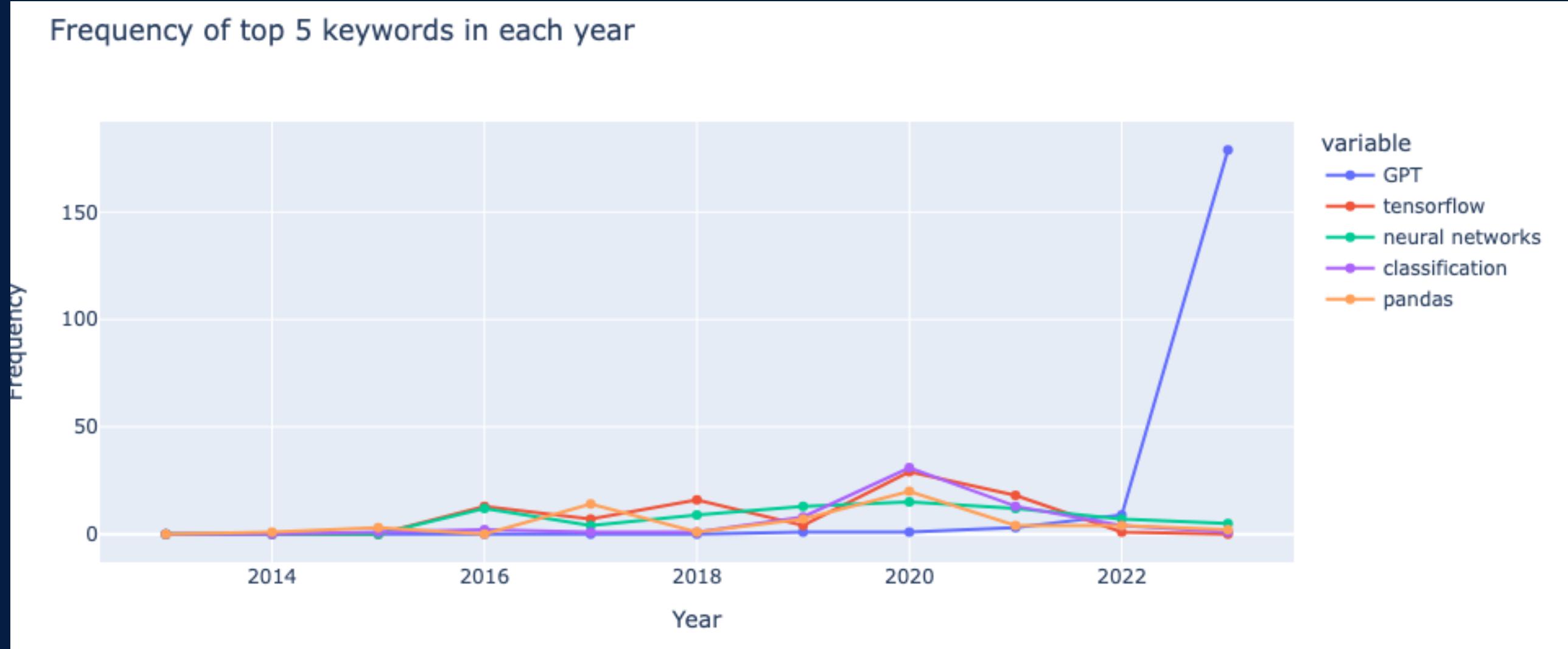
# ĐÂU LÀ CÁC HOT KEYWORDS XUẤT HIỆN QUA CÁC NĂM?

Các từ khóa phổ biến thường liên quan đến machine learning hoặc deep learning (tensorflow, neural networks) và thư viện Python hỗ trợ lĩnh vực AI và dữ liệu.

GPT là từ khóa được sử dụng nhiều nhất với khoảng 200 lần xuất hiện trong tiêu đề video, chỉ ra rằng GPT là chủ đề tuy mới nhưng lại được rất nhiều kênh quan tâm đến.

Top 10 keywords





# XU HƯỚNG CÁC HOT KEYWORDS

- Trong số 5 từ khóa hot, hầu hết đều dao động thường xuyên. Tuy nhiên, vào năm 2020, các từ khóa tăng về tần suất xuất hiện trong tiêu đề.
- GPT có tốc độ tăng trưởng nhanh chóng vào năm 2022 (từ khoảng 9 lần vào năm 2022 lên đến 180 lần vào năm 2023). Với đầu tư mạnh mẽ hiện tại, trong tương lai GPT có thể tiếp tục giữ vững vị trí cao về từ khóa xuất hiện trên YouTube.

# **ĐÁNH GIÁ CÁC TIÊU CHÍ ẢNH HƯỚNG ĐÈN LƯỢT XEM VÀ LIKE**

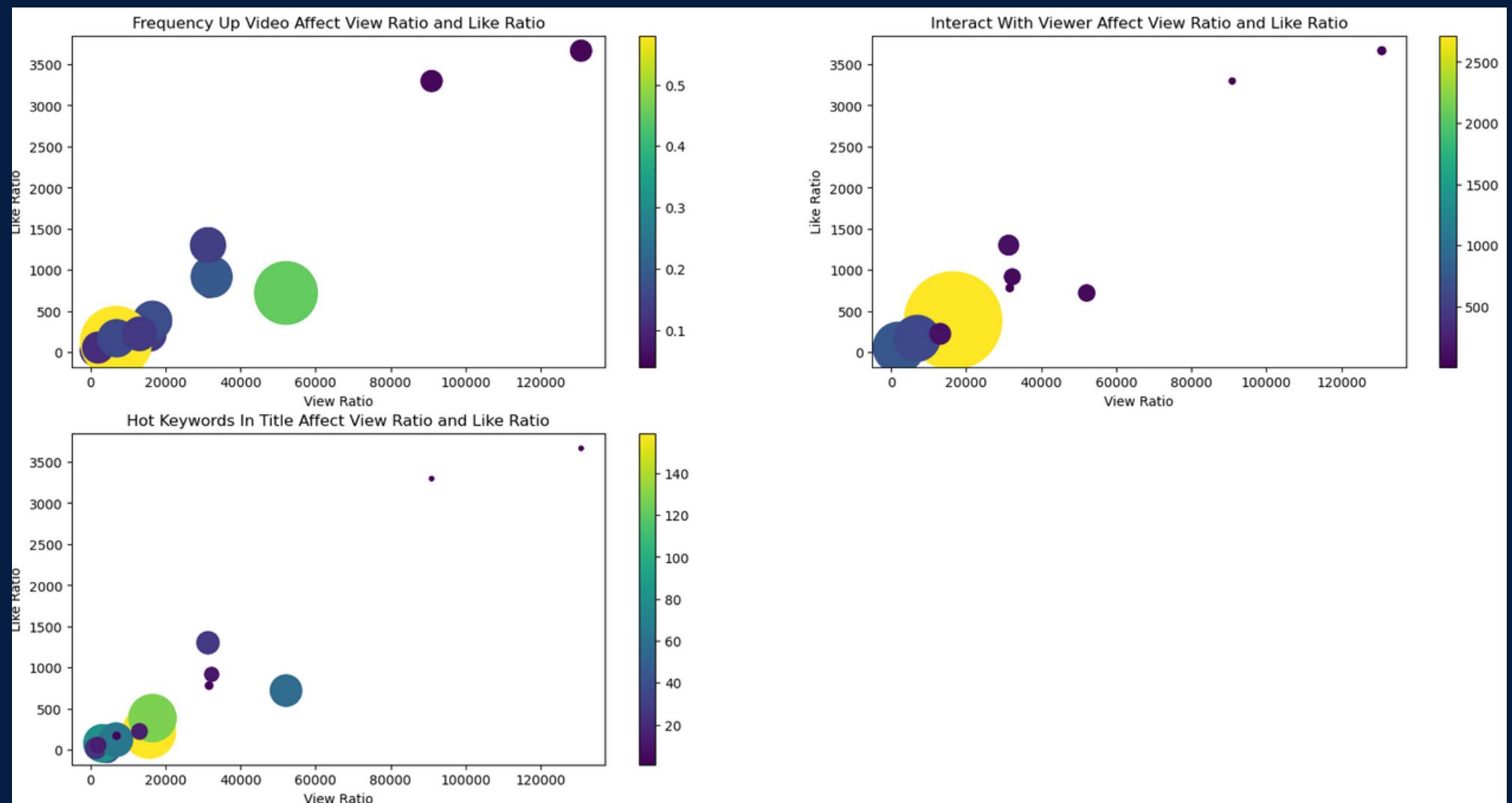
## **PHÂN TÍCH 3 YẾU TỐ :**

- Tần suất ra video .
- Sự tương tác với khán giả ( thông qua số lượt phản hồi các comment của người xem) .
- Mức độ sử dụng các hot keyword trong tiêu đề.

# ĐÁNH GIÁ CÁC TIÊU CHÍ ẢNH HƯỞNG ĐẾN LƯỢT XEM VÀ LIKE

## Tần suất ra video

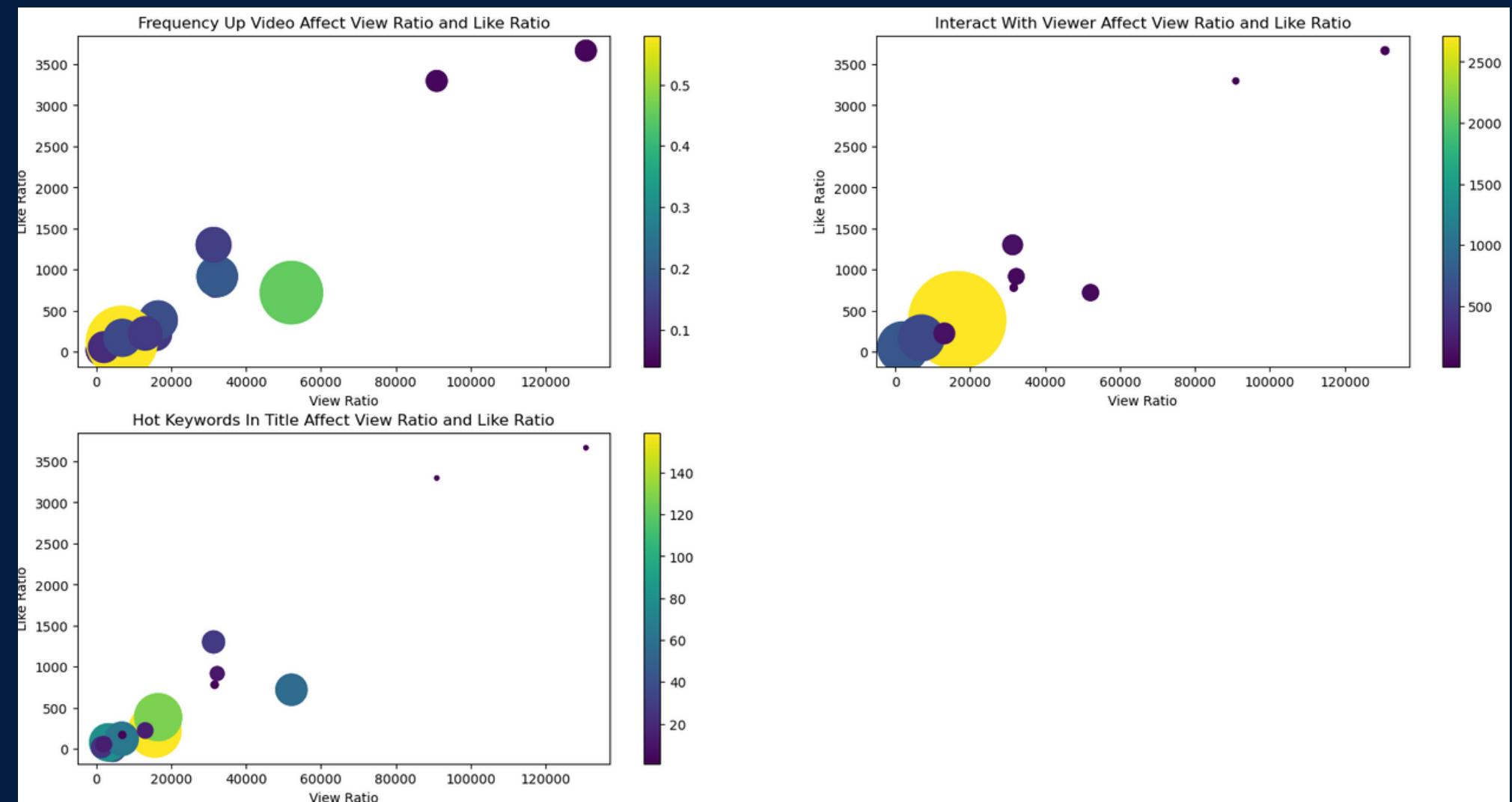
- Ta thấy rằng, nếu chúng ta đăng video quá thường xuyên hoặc quá ít, lượng lượt xem và thích sẽ không đạt được mức cao.
- Tuy nhiên, có những kênh chỉ đăng ít video nhưng vẫn thu hút một lượng lớn người xem, làm chứng tỏ chất lượng nội dung vẫn quan trọng hơn số lượng.



# ĐÁNH GIÁ CÁC TIÊU CHÍ ẢNH HƯỞNG ĐẾN LƯỢT XEM VÀ LIKE

## Sự tương tác với khán giả

- Tương tự, khi có quá ít tương tác, kênh sẽ khó thu hút nhiều người theo dõi. Tuy nhiên, một số kênh vẫn đứng đầu về lượt xem mặc dù có ít tương tác.
- NOTE:** Do giới hạn quota ,mỗi video chỉ có maximum 100 comment, nên kết quả có ảnh hưởng đôi chút.

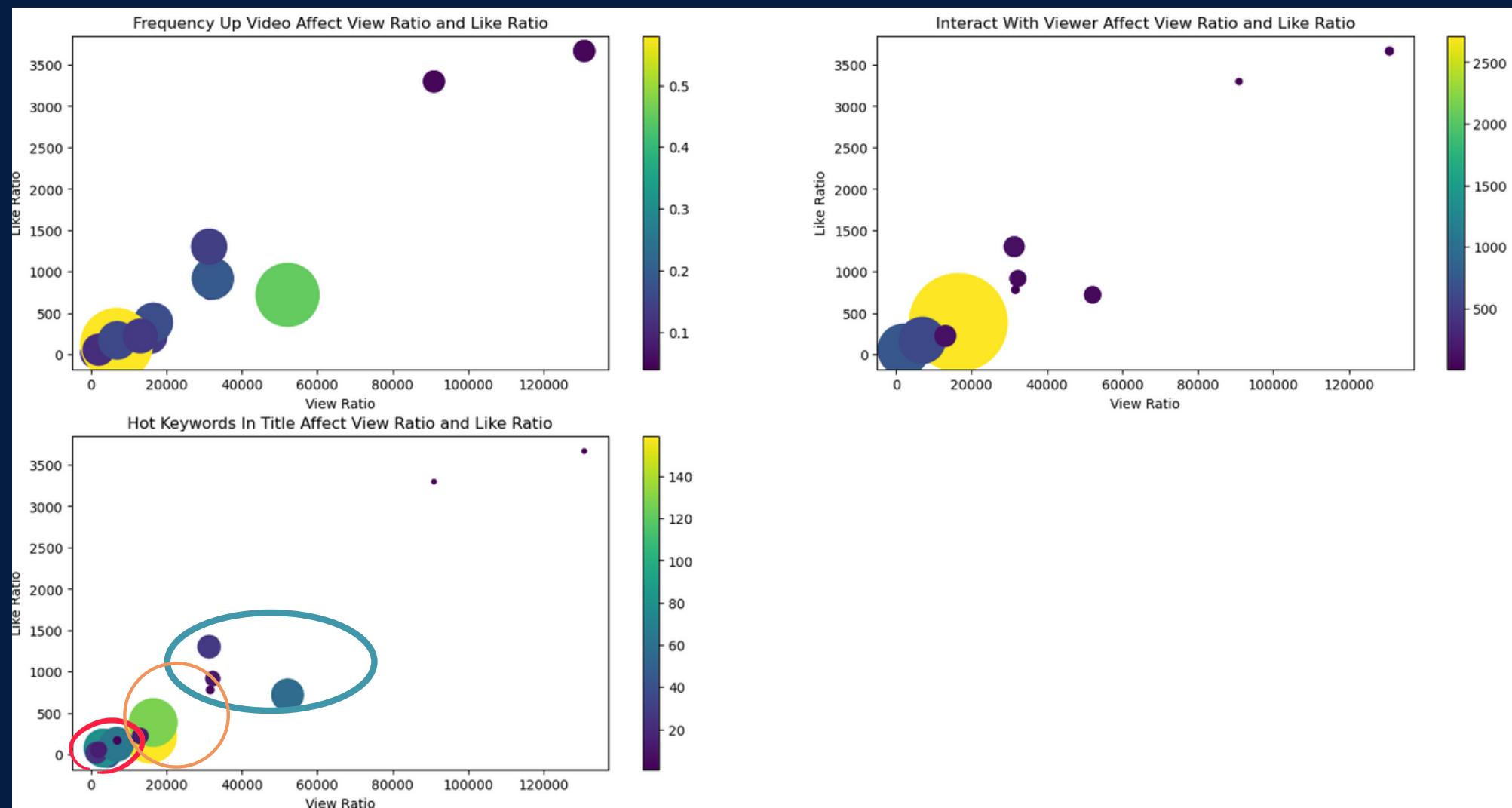


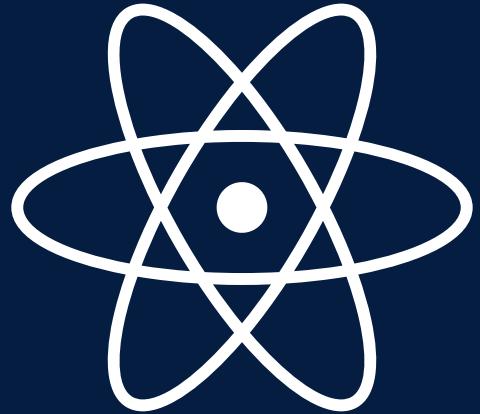
# ĐÁNH GIÁ CÁC TIÊU CHÍ ẢNH HƯỞNG ĐẾN LƯỢT XEM VÀ LIKE

## Sử dụng hot keyword trong tiêu đề

Có thể thấy có 3 nhóm chính :

- Nhóm ít sử dụng (đỏ).
- Nhóm sử dụng nhiều (cam).
- Nhóm có thể quan tâm tới chất lượng hơn (xanh).





# MÔ HÌNH HÓA DỮ LIỆU

Fact_Sales
salesDate
salesTypeID
customerID
salesmanID
itemID
zipCode
qty
unitPrice
discount
netPrice



# Phát biểu bài toán

Từ những thuộc tính trong dữ liệu các videos, làm thế nào để dự đoán số views cho các videos đó cũng như các video sắp được thực hiện.

Input

- Thời lượng
- Tag

Output:

- Số lượng view

# Kỹ thuật huấn luyện mô hình

Feature engineering

- |                                     |                          |                             |
|-------------------------------------|--------------------------|-----------------------------|
| 1. Chuẩn hóa thời gian              | 2. Tách các tag từ chuỗi | 3. Làm sạch các tag đã tách |
| 4. Chọn các tag xuất hiện nhiều lần | 5. Khử nhiễu và outlier  | 6. Label encode cho tag     |

# Kỹ thuật huấn luyện mô hình

Mô hình

Mô hình: Random Forest Regressor

Cross validation and hyperparameter tuning: GridSearchCV



# Đánh giá mô hình

---

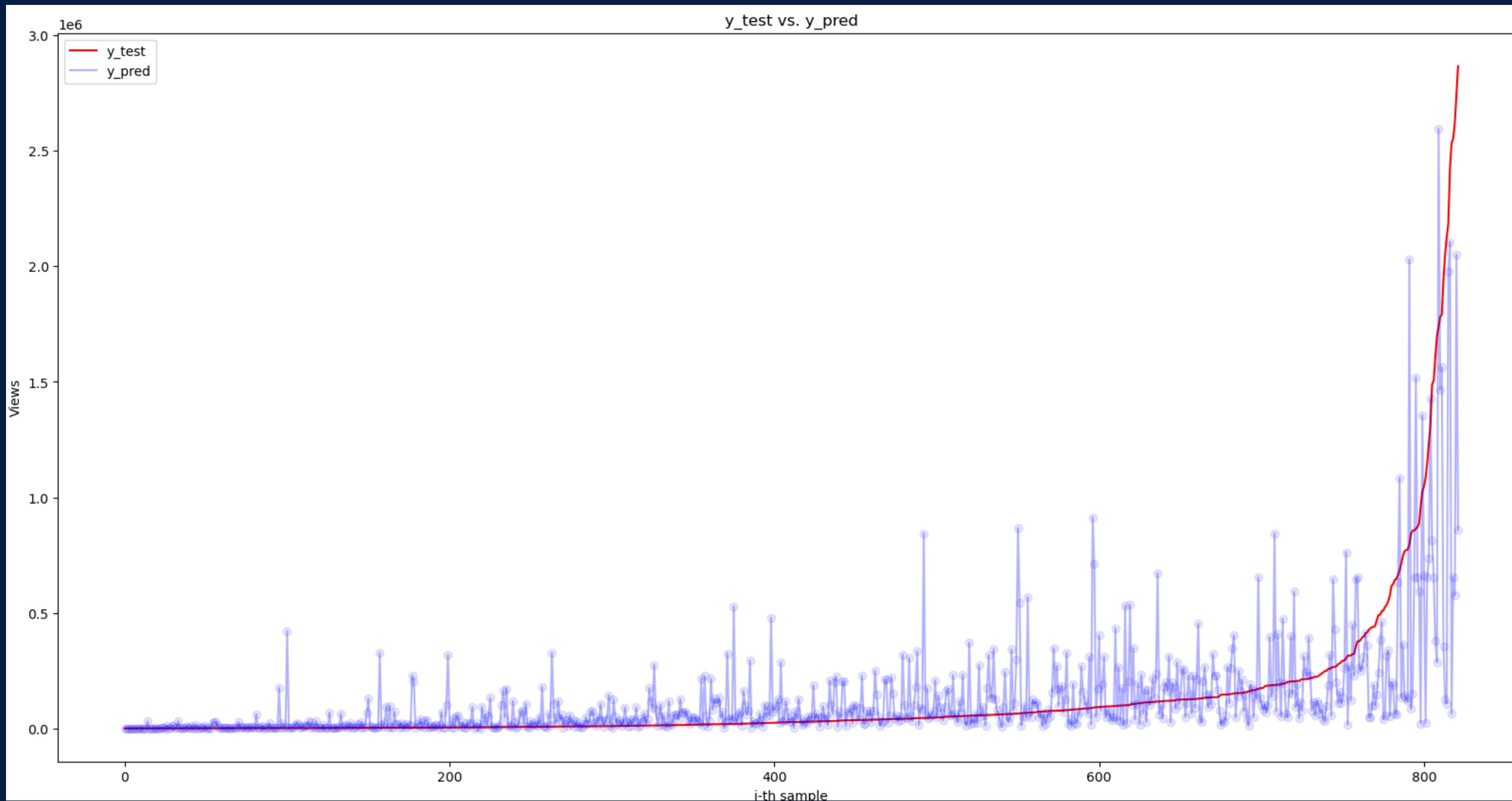
Mean Absolute Error (MAE): 105733.69

Mean Squared Error (MSE): 67879270447.39

Root Mean Squared Error (RMSE): 260536.51



# Kết quả dự đoán



# Triển khai mô hình

## Flask

Flask là một web framework nhẹ được viết bằng Python. Không có database abstraction layer, form validation, hoặc bất kỳ components nơi thư viện bên thứ 3 cung cấp những hàm phổ biến.

## Flow hoạt động

1. Load model
2. Nhận dữ liệu input
3. Dự đoán kết quả
4. Phản hồi kết quả dự đoán

# Demo

## Views Prediction

Duration (second)

1st tag

None



2nd tag

None



3rd tag

None



Submit

# GIÁ TRỊ ĐẠT ĐƯỢC

TỪ QUÁ TRÌNH TÌM HIỂU, CHÚNG EM CÓ THỂ ĐÚT KẾT ĐƯỢC CHO MÌNH NHỮNG  
KÊNH YOUTUBE SAU

Two Minute Paper,  
TheAIGrid: dùng để  
cập nhập thông tin các  
tin tức, công nghệ mới

3Blue1Brown: dùng  
hiểu tìm hiểu các khái  
niệm toán khó nhăn  
tăng hiểu biết với lĩnh  
vực này

Abhishek Thakur, Data  
Professor: dùng để cập  
nhập các kĩ thuật, thư  
viện phục vụ quá trình  
lập trình

Joma Tech: và cuối  
cùng không thể thiếu  
chính là những giây  
phút thư giãn cũng với  
kênh này

Codebasics: dùng để  
học chi tiết một kĩ  
thuật nào đó, xem các  
video về tips tricks  
trong công việc

# BƯỚC TIẾP THEO

Tăng số lượng các kênh phân tích, các comment thu thập

Xây dựng model với độ chính xác cao hơn bằng cách kết hợp thêm các mô hình ngôn ngữ để phân tích chính xác title và tag

Xây dựng các mô hình ngôn ngữ để phân tích các tiêu đề, comment chính xác hơn

Mở rộng video ra nhiều lĩnh vực để xem sức ảnh hưởng của lĩnh vực khoa học dữ liệu

Xây dựng hệ thống gợi ý video dựa trên dữ liệu động

# CÁC KHÓ KHĂN

Việc giới hạn về quota làm việc thu thập dữ liệu gấp nhiều khó khăn:

- Chia ra nhiều luồng để thu thập phát sinh nhiều lõi.
- Dữ liệu lõi muốn thu thập lại phản chờ đến ngày hôm sau gây cản trở quá trình làm việc

Dữ liệu trả về có khá nhiều text, nhóm chưa đủ kiến thức về các mô hình ngôn ngữ làm giảm hiệu suất các đánh giá, mô hình

# Thank's For Watching

