

Final Project - Introduction to Data Science

Phân tích sở thích của người xem thông qua dữ liệu của Youtube từ các kênh sản xuất nội dung trong lĩnh vực: DS - DA - AI - ML

Group 04

Võ Duy Anh - 21127221
Nguyễn Mậu Gia Bảo - 21127583
Lê Mỹ Khánh Quỳnh - 21127681
Vũ Minh Phát - 21127739

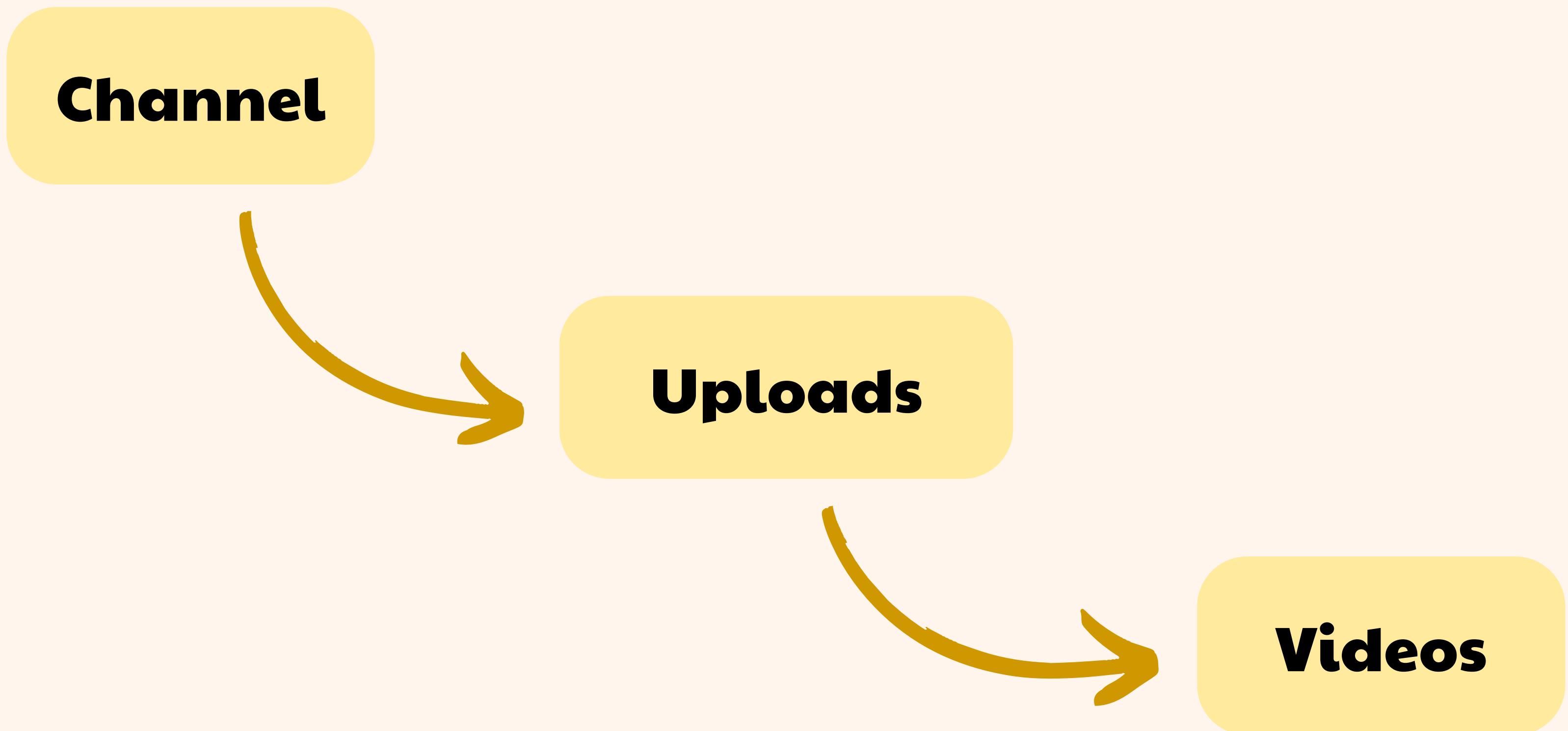
Github: https://github.com/BaoDias2505/FinalProject_Introduction-to-data-science/tree/main

Trello: <https://trello.com/b/N5UME2gN/final-project-intro2ds>

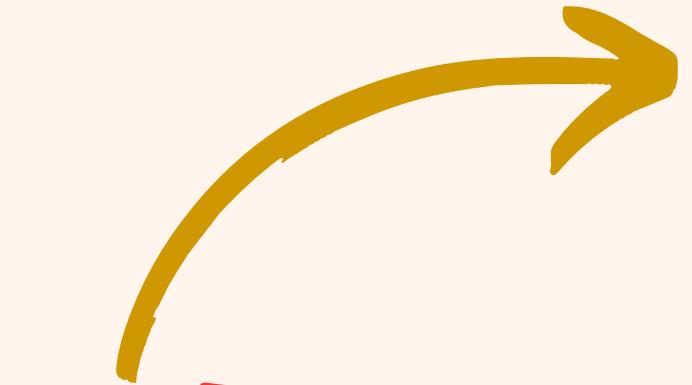
Các nội dung chính

- I. Thu thập dữ liệu thông qua API
- II. Tiền xử lý và khám phá dữ liệu
- III. Phân tích khám phá dữ liệu (EDA)
- IV. Mô hình hóa dữ liệu
- V. Triển khai ứng dụng với mô hình học máy
- VI. Tài liệu tham khảo

Thu thập dữ liệu thông qua API



① Bước 1: Lấy channel_id của các kênh youtube mà ta quan tâm

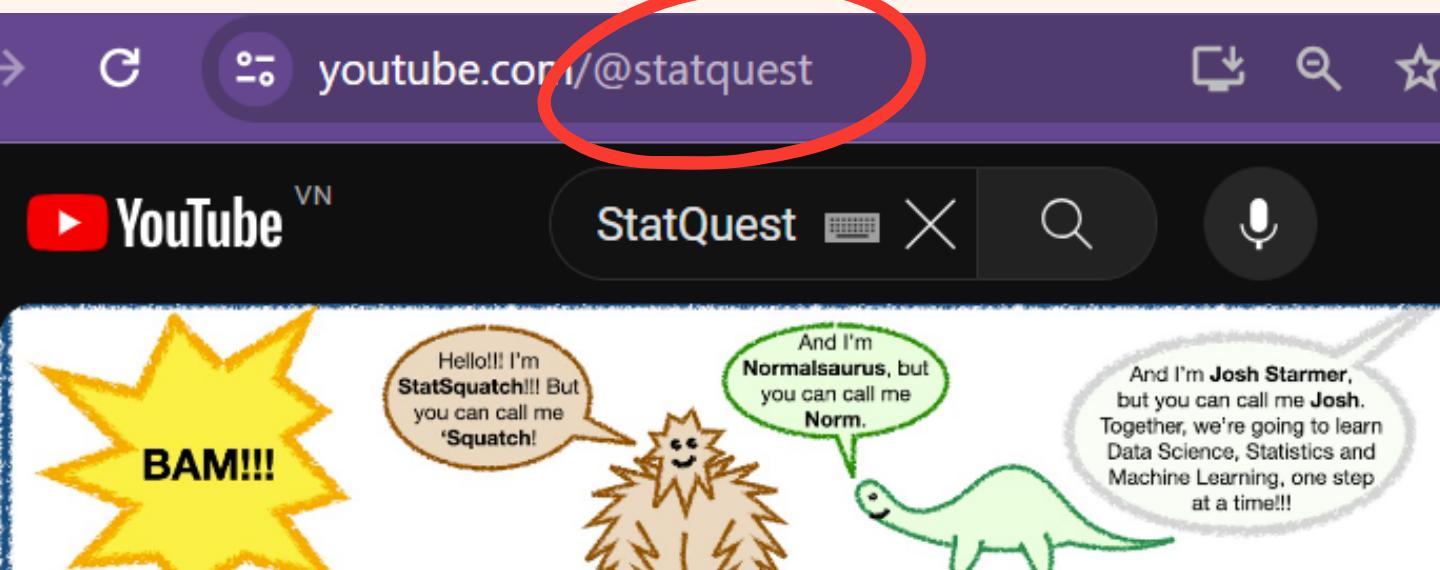


YouTube Username: /@ statquest

Convert YouTube Username to ID

YouTube Channel ID: UCtYLUTtgS3k1Fg4y5tAhLbw

[Link](#)

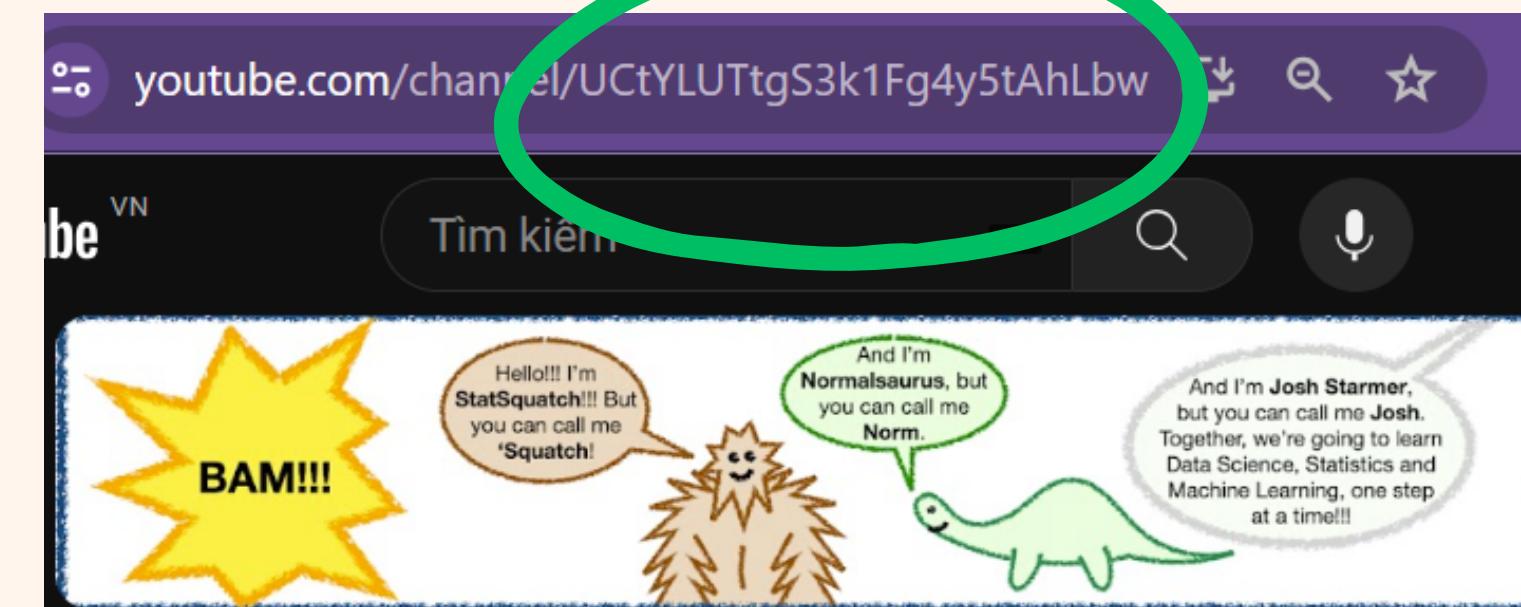


StatQuest with Josh
Starmer •

@statquest · 1,05 Tr người đăng ký · 266 video

Statistics, Machine Learning and Data Science can so

patreon.com/statquest và 4 đường liên kết khác



StatQuest with Josh
Starmer •

@statquest · 1,05 Tr người đăng ký · 266 video

Statistics, Machine Learning and Data Science can so

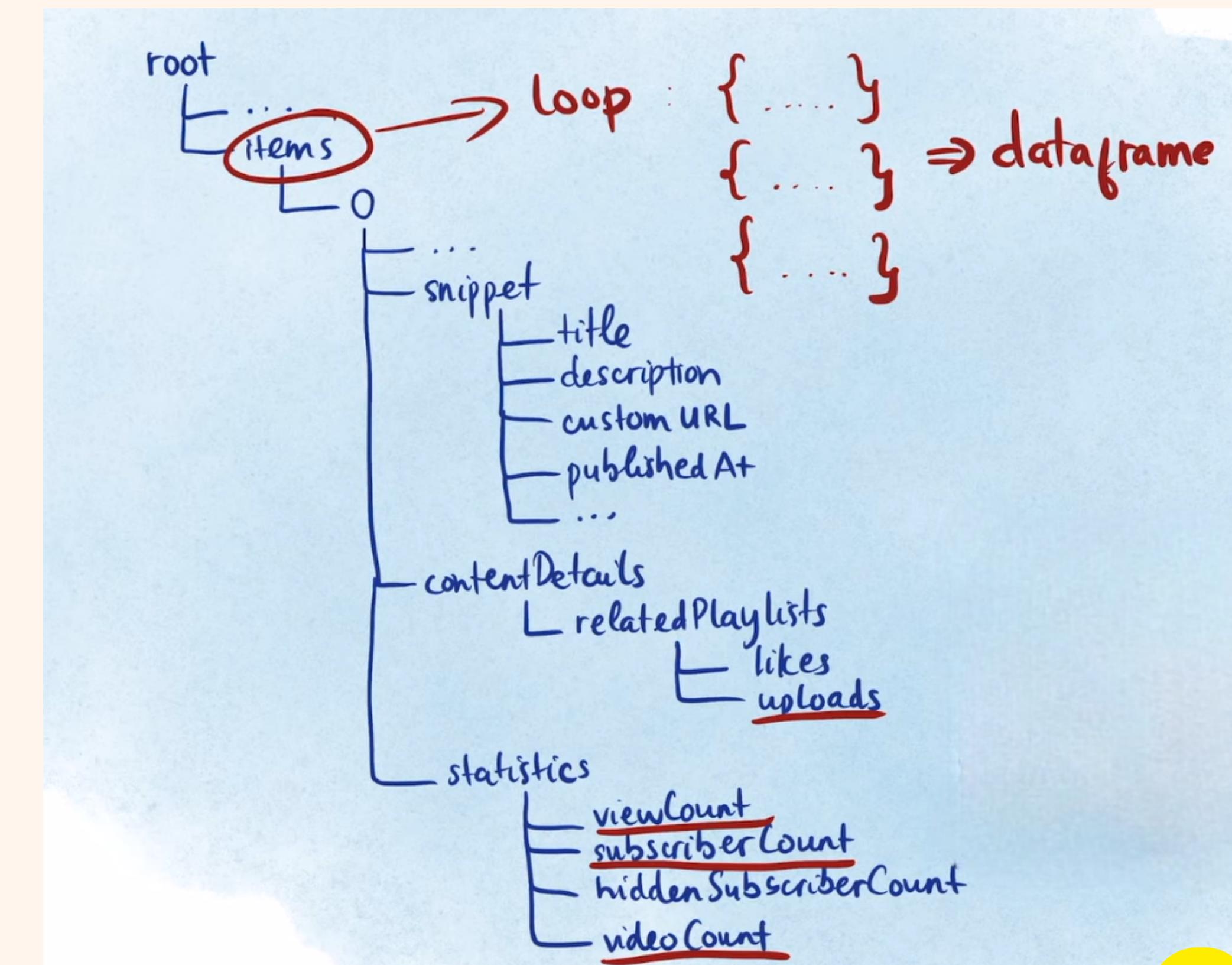
patreon.com/statquest và 4 đường liên kết

① Bước 2: Lấy playlist_id ứng với từng channel_id mà ta có

get_channel_stats(...)

request

response



[Image source](#)

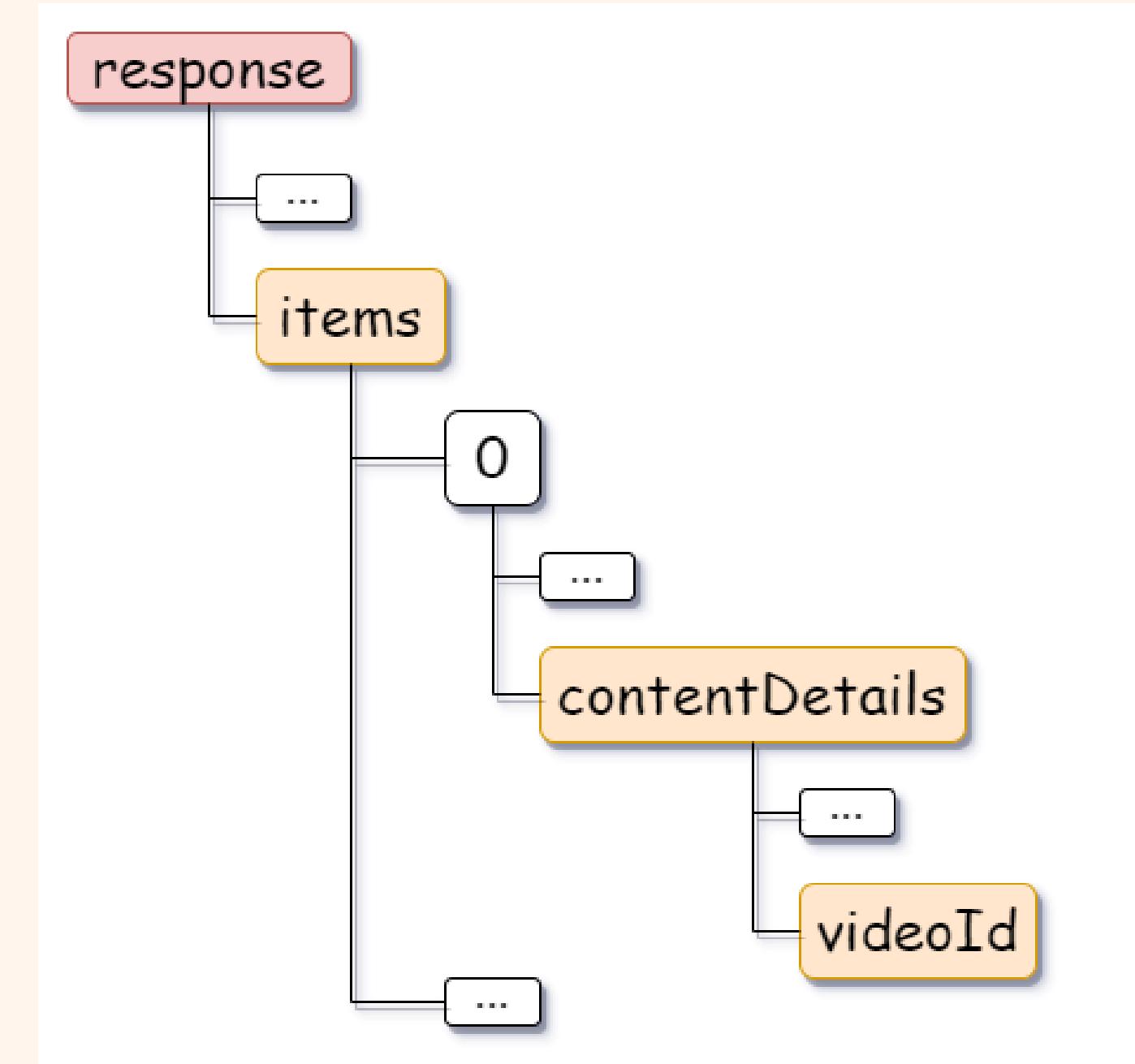
④ Bước 3: Lấy các video_id trong playlist_id tương ứng

`get_video_ids(...)`

request

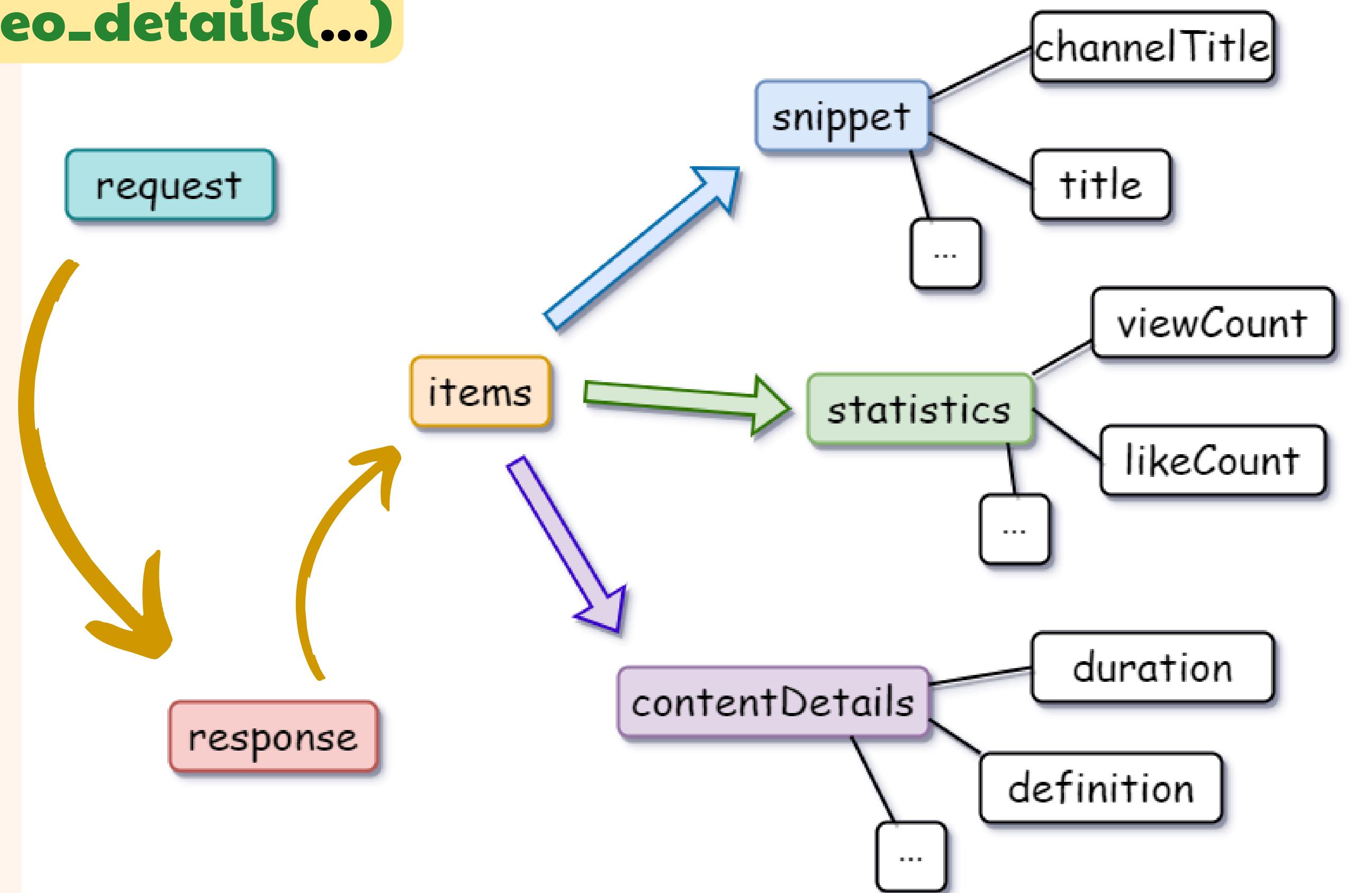


response



④ Bước 4: Lấy thông tin của video ứng với video_id

get_video_details(...)



Tiền xử lí và khám phá dữ liệu

◎ Tổng quan về dữ liệu

Columns	Meaning
video_id	Unique identifier for the YouTube video.
channelTitle	The name of the channel that uploaded the video.
title	Title of the video.
description	Description or summary of the video content.
tags	Keywords or tags associated with the video. ...
publishedAt	Date and time when the video was published.
viewCount	Number of views the video has accumulated.
likeCount	Count of likes received by the video.
favoriteCount	Deprecated; used to track how many times viewers added the video to their favorites.
commentCount	Number of comments posted on the video.
duration	Length of the video.
definition	Video resolution or quality (e.g., HD, SD).
caption	Indicates whether closed captions are available for the video.

60032 dòng
13 cột

Không có dòng
trùng lặp

○ Kiểu dữ liệu của các cột

- **Object:**

video_id , channelTitle, title,
description, tags, publishedAt,
duration, defination



Convert publishedAt sang
datetime và duration sang số
nguyên biểu diễn tổng số giây

- **Float64:**

viewCount, likeCount,
commentCount, favouriteCount

- **Bool:**

caption

○ Numerical columns

	publishedAt	viewCount	likeCount	favouriteCount	commentCount	duration
missing_ratio	0.0	6.663113e-03	0.404784	100.0	1.279318	0.0
min	2006-10-25 10:28:09+00:00	0.000000e+00	0.000000	NaN	0.000000	0.0
lower_quartile	2019-02-18 04:02:18+00:00	6.220000e+02	11.000000	NaN	0.000000	293.0
median	2020-12-22 09:26:05.500000+00:00	3.059500e+03	55.000000	NaN	5.000000	695.0
upper_quartile	2022-07-05 06:04:02.500000+00:00	1.602175e+04	334.000000	NaN	27.000000	1786.0
max	2023-11-24 13:04:35+00:00	3.447645e+07	571358.000000	NaN	60054.000000	92218.0

- **favouriteCount** không có bất kì dữ liệu nào -> **xóa bỏ** thuộc tính
- **viewCount**, **likeCount** và **commentCount** có một số missing values -> fill bằng giá trị **median** và convert sang kiểu **integer**

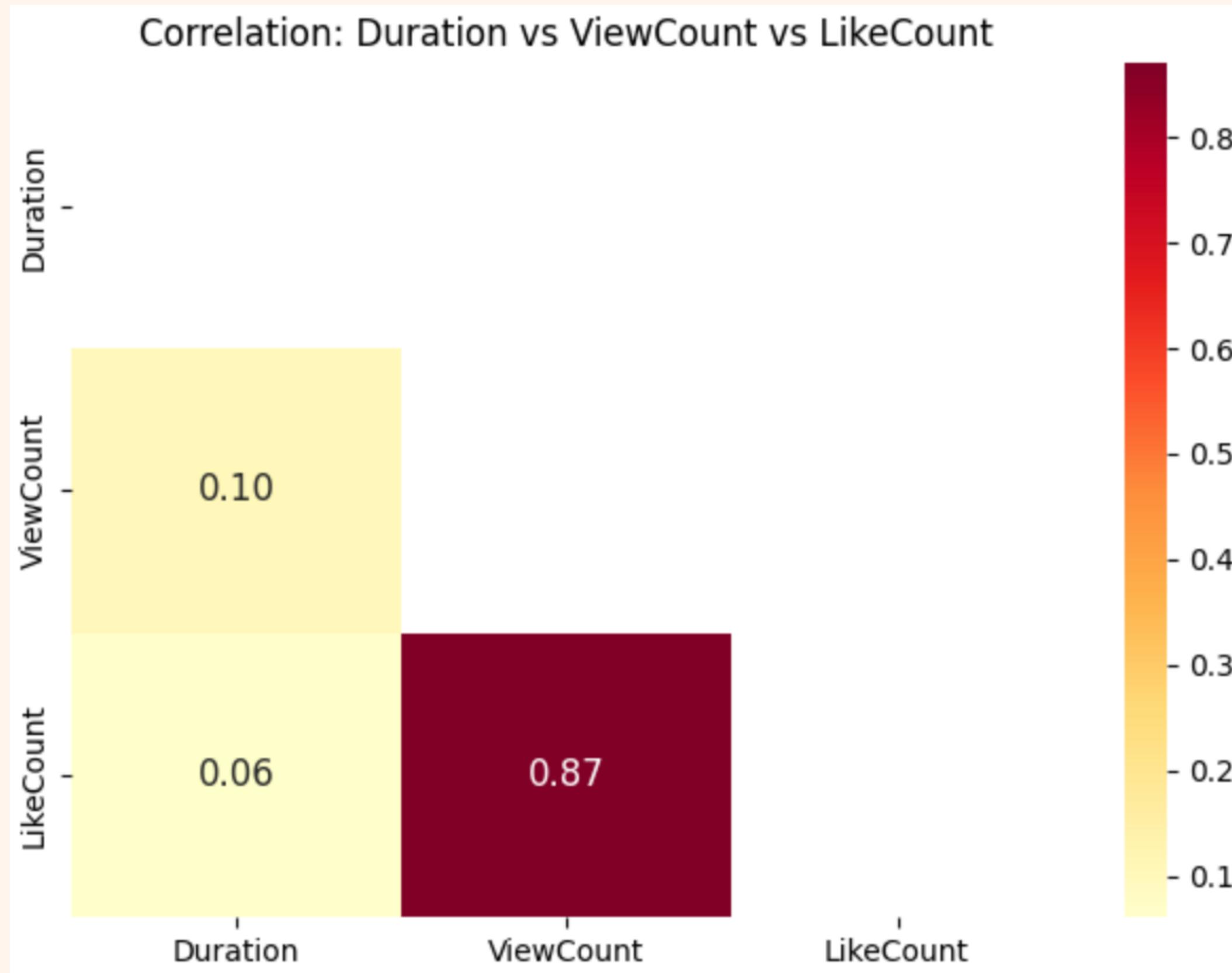
○ Categorical columns

	video_id	channelTitle	title	description	tags	definition	caption	hour	day	day_of_week	month	year	
missing_ratio	0.0	0.0	0.0	2.803505	18.956557	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
num_values	60032	160	59644	53351	35545	2	2	24	31	7	12	17	

- **description** và **tags** -> điền chuỗi ‘(nodescription)’ và ‘(notag)’ vào các vị trí trống dữ liệu.
- Format **tags**: ‘[A,B,C]’ -> ‘A|B|C’

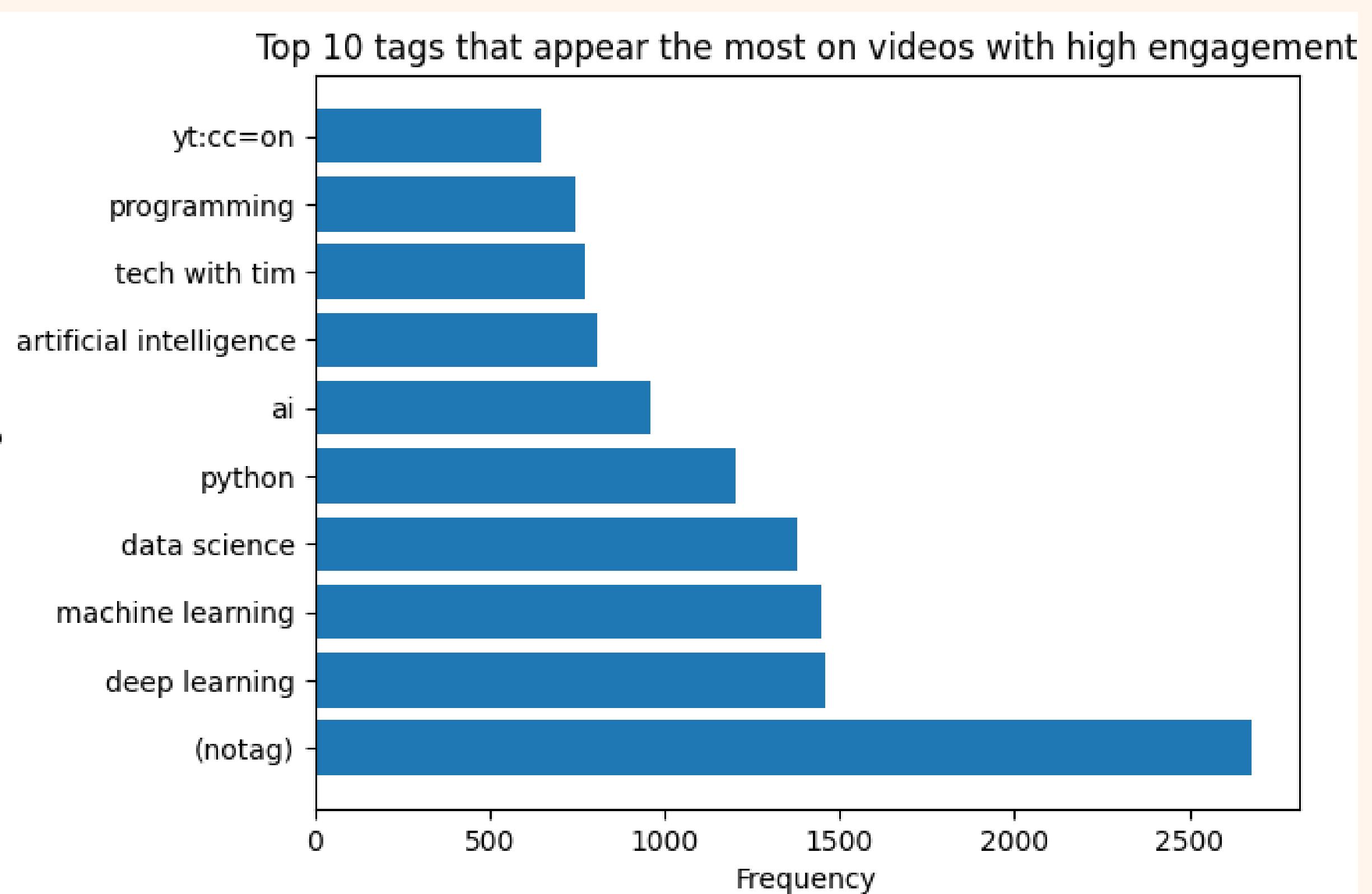
EDA (Exploratory Data Analysis)

Câu hỏi 1: Có mối tương quan giữa thời lượng của video và số lượt xem hoặc lượt thích không?



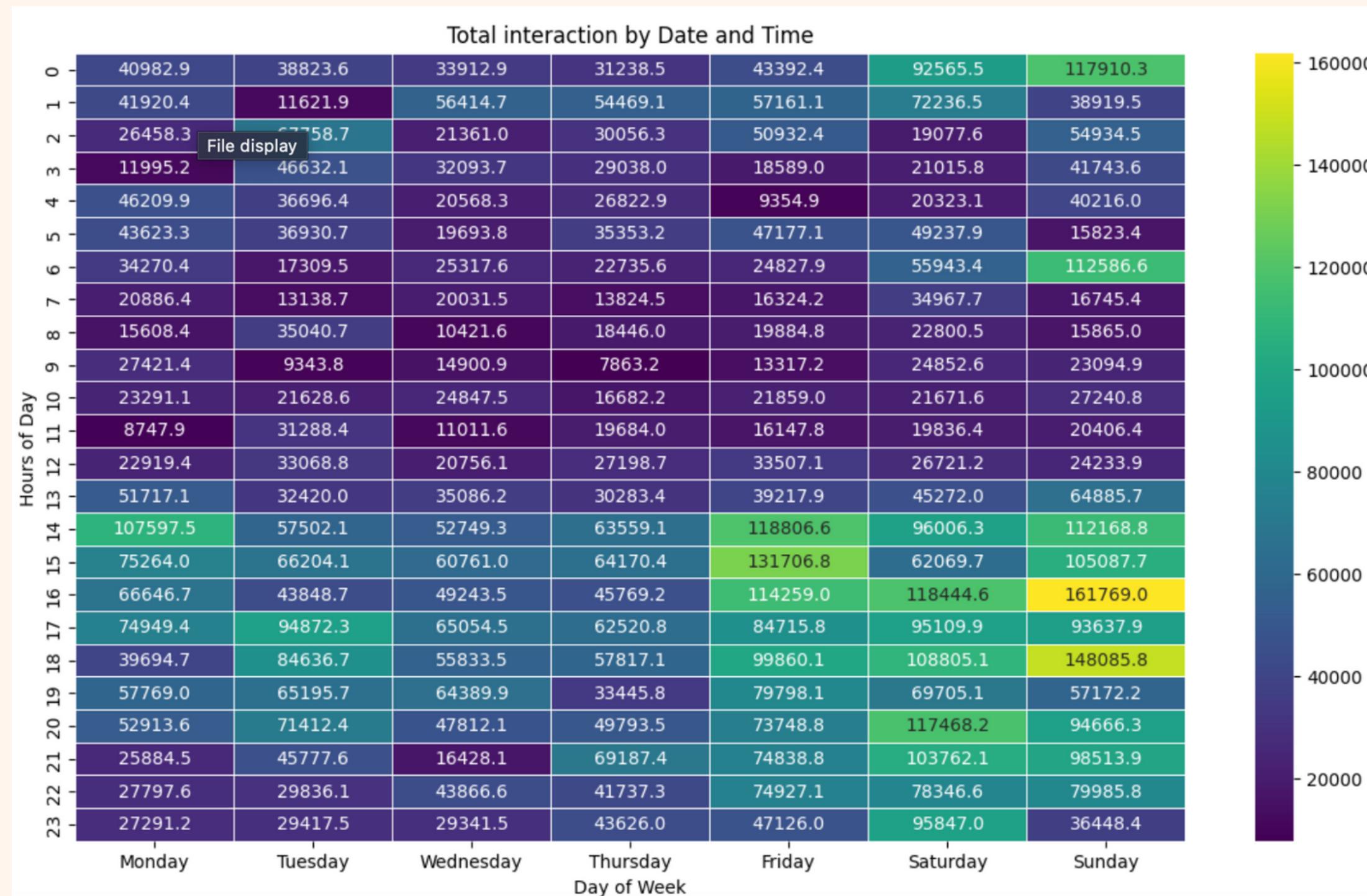
- Ý nghĩa: xem xét sự ảnh hưởng của thời lượng video đến chỉ số tương tác
- Nhận xét: Sự tương quan giữa **duration** và **viewCount** và **likeCount** là tương quan dương tuy nhiên độ tương quan rất thấp -> **duration** không có nhiều sự tác động hay ảnh hưởng đến sự tương tác mà video nhận được

Câu hỏi 2: Có tag cụ thể nào xuất hiện thường xuyên hơn trong video có số liệu tương tác cao hơn không?



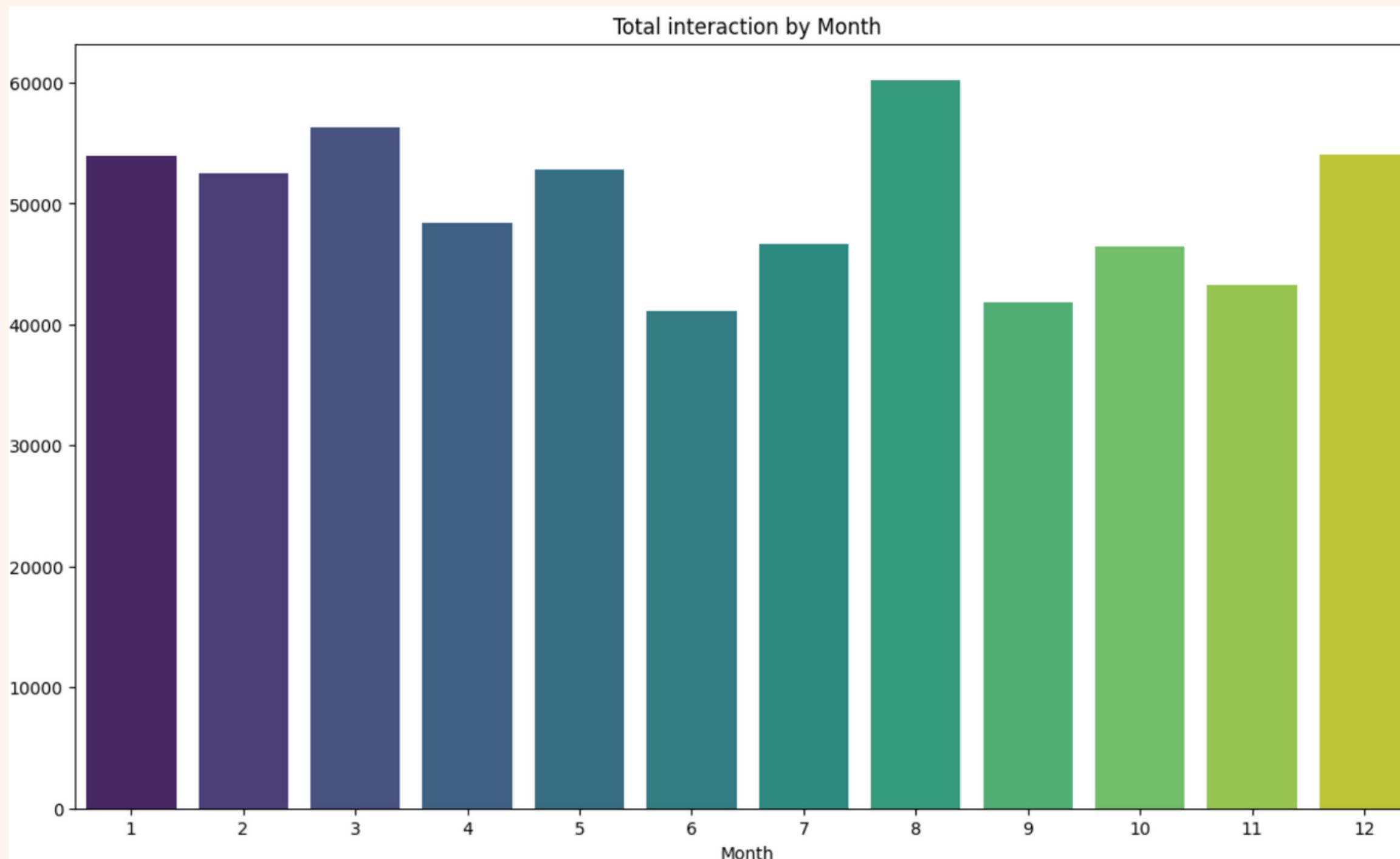
- Ý nghĩa: việc sử dụng tags cũng là 1 yếu tố quan trọng để thu hút tương tác trên các nền tảng xã hội -> rút ra các tags xuất hiện nhiều trong các video xu hướng.
- Nhận xét:
 - Các xu hướng thường cho thấy rằng việc **không sử dụng thẻ hoặc chọn cách tiếp cận tối giản** có thể là một chiến lược để thu hút sự chú ý.
 - Mặt khác, các thẻ liên quan đến chủ đề **AI, Machine Learning, Deep Learning, Data Science** hoặc **khoa học máy tính** nói chung có xu hướng thu hút nhiều sự quan tâm. Đặc điểm chung của các thẻ này là chúng đều có vẻ ngắn gọn và đầy đủ.

Câu hỏi 3: Có thời điểm cụ thể nào trong ngày, ngày trong tuần hoặc tháng mà video có xu hướng nhận được nhiều lượt xem hoặc mức độ tương tác hơn không?



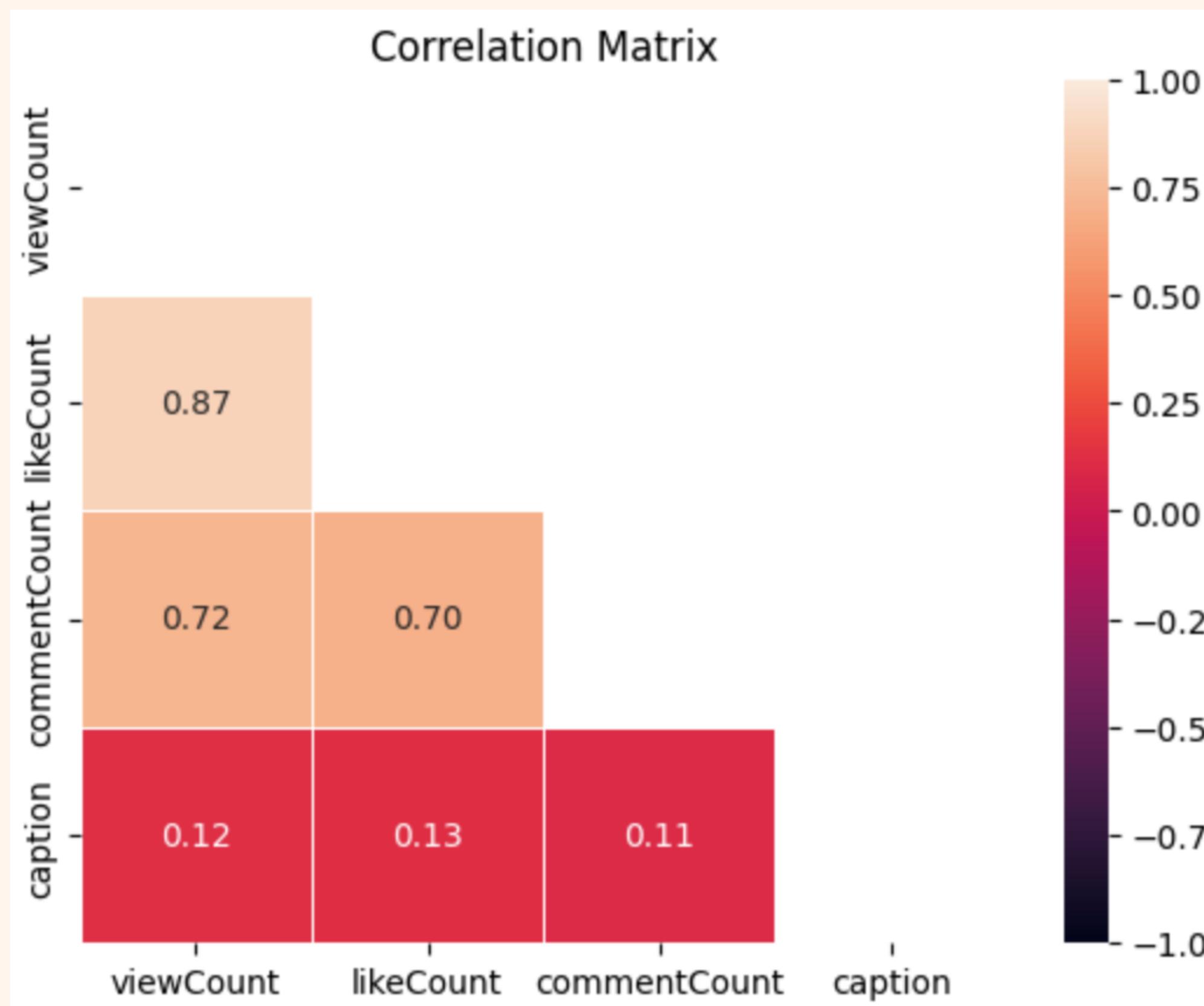
- Ý nghĩa: bằng cách xác định thời điểm nào trong ngày, tuần hoặc tháng mà video của bạn có nhiều khả năng thu hút nhiều người xem nhất, bạn có thể tối ưu hóa thời gian phát sóng hoặc xuất bản nội dung của mình. Điều này làm tăng cơ hội video của bạn được hiển thị cho nhiều khán giả và có thể tạo ra sự tương tác tích cực.
- Nhận xét:
 - Có thể thấy rằng các video được xuất bản từ **14 giờ đến 18 giờ** sẽ có lượt xem và tương tác cao nhất trong ngày và cuối tuần, đặc biệt **thứ sáu, thứ bảy và chủ nhật** sẽ có lượt xem và tương tác cao nhất, đặc biệt là **16 giờ ngày chủ nhật**

Câu hỏi 3: Có thời điểm cụ thể nào trong ngày, ngày trong tuần hoặc tháng mà video có xu hướng nhận được nhiều lượt xem hoặc mức độ tương tác hơn không?



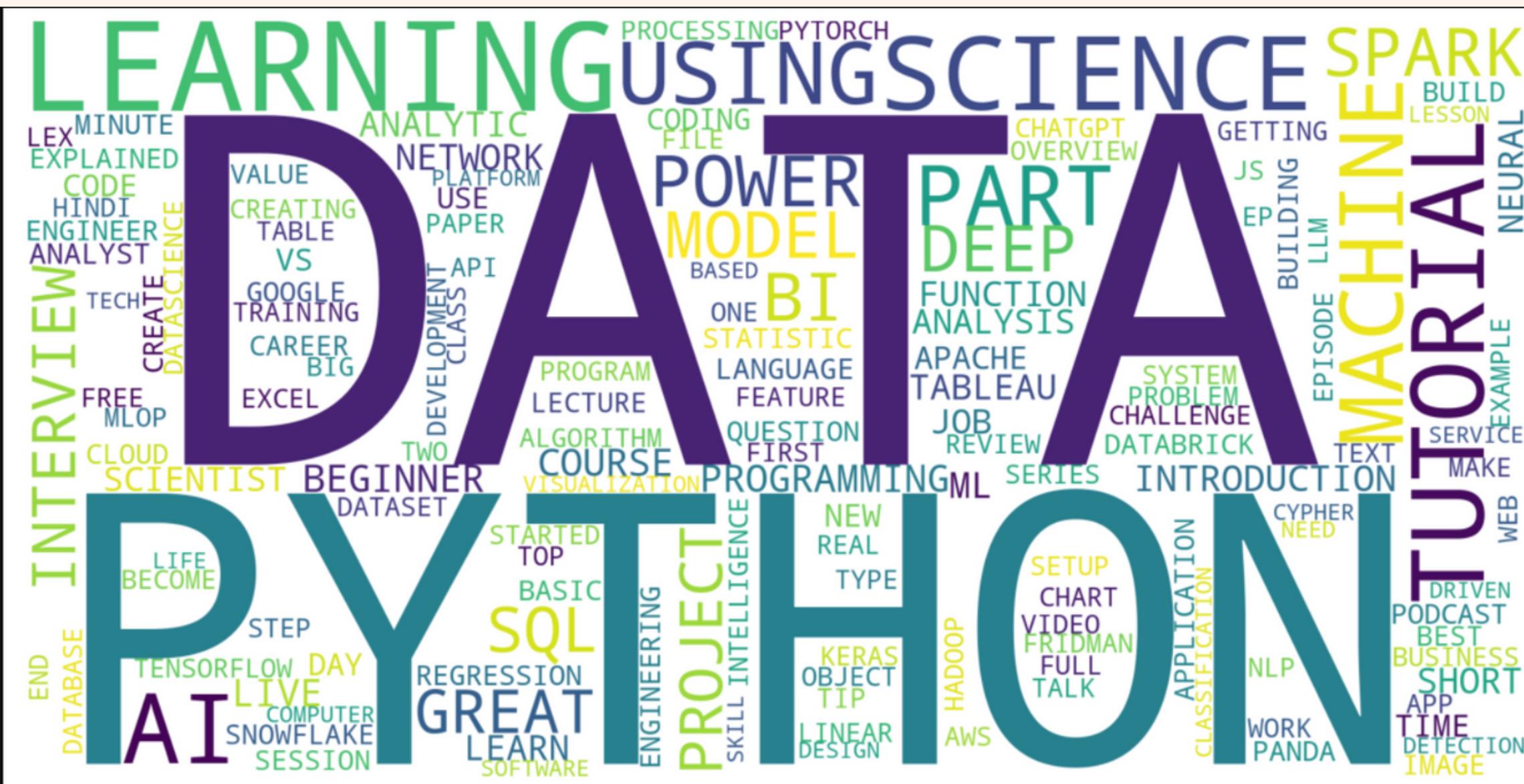
- Ý nghĩa: bằng cách xác định thời điểm nào trong ngày, tuần hoặc tháng mà video của bạn có nhiều khả năng thu hút nhiều người xem nhất, bạn có thể tối ưu hóa thời gian phát sóng hoặc xuất bản nội dung của mình. Điều này làm tăng cơ hội video của bạn được hiển thị cho nhiều khán giả và có thể tạo ra sự tương tác tích cực.
- Nhận xét:
 - Những video ra mắt vào **tháng 8** sẽ có lượt xem, lượt thích và bình luận cao nhất. Điều ngược lại đúng với **tháng 6 và tháng 9**

Câu hỏi 4: Chú thích ảnh hưởng như thế nào đến lượt xem và tương tác (lượt thích, bình luận)?



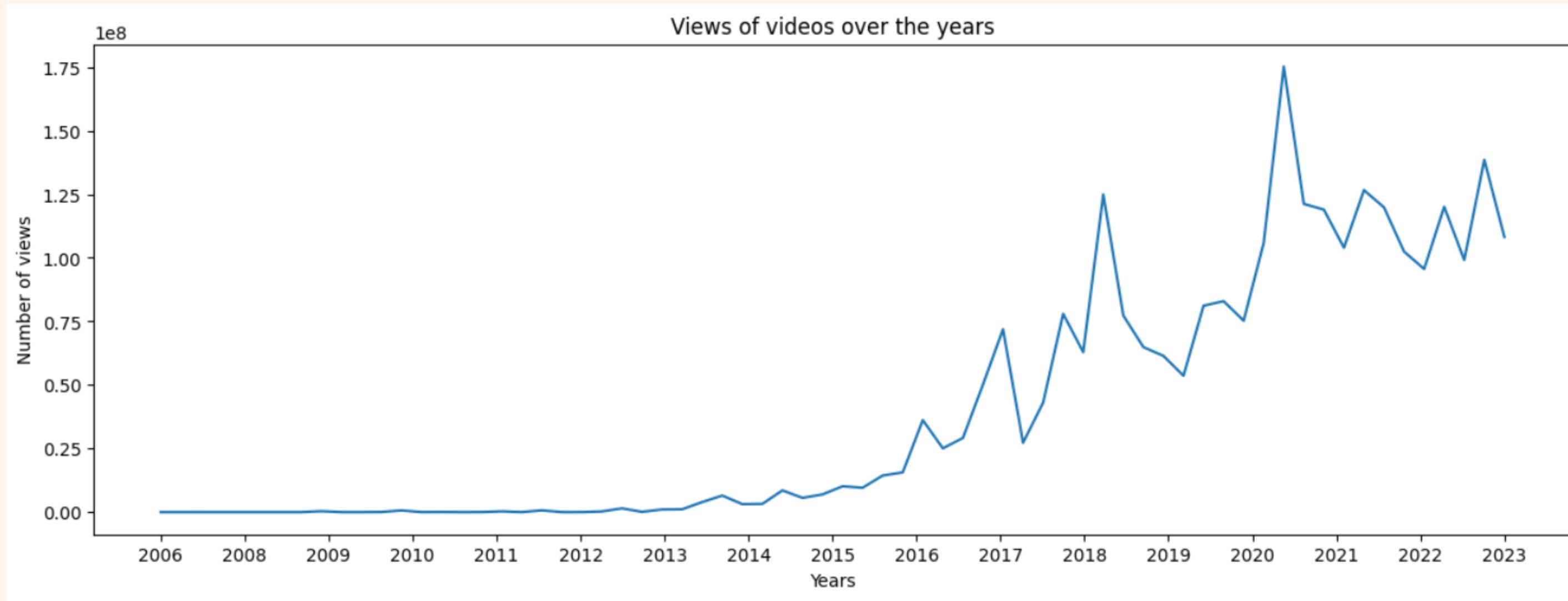
- Ý nghĩa: xem xét sự ảnh hưởng của phụ đề đến chỉ số tương tác
- Nhận xét: Sự tương quan giữa **caption**, **viewCount**, **likeCount** và **commentCount** là tương quan dương tuy nhiên độ tương quan rất thấp -> **caption** không có nhiều sự tác động hay ảnh hưởng đến sự tương tác mà video nhận được

Câu hỏi 5: Những từ nào thường được sử dụng trong tiêu đề video?



- Ý nghĩa: Việc đặt tiêu đề cho video cũng là một yếu tố thu hút người xem. Việc trả lời câu hỏi này giúp chúng ta nhận biết những từ nào được sử dụng phổ biến, người xem quan tâm. Từ đó, có thể cải thiện video bằng cách đặt tiêu đề cho video.
 - Nhận xét:
 - Các từ được dùng phổ biến đều có sự **liên quan đến nội dung** về dữ liệu.
 - Kích thước của các từ cũng tỉ lệ với sự phổ biến của các từ đó. Dễ dàng nhận biết được mật độ các từ được sử dụng.

Câu hỏi 6: Mức độ quan tâm của người dùng đến video về data thay đổi như thế nào qua từng năm?



- Ý nghĩa: Biết được mức độ quan tâm của người dùng đến các video cùng nội dung có thể giúp các nhà sáng tạo nội dung dự đoán xu hướng, tiềm năng của lĩnh vực. Từ đó đưa ra phương hướng, chiến lược phát triển phù hợp.

- Nhận xét:
 - Người xem có xu hướng **tăng sự quan tâm** đến các video về data qua các năm.
 - Tuy nhiên, sự tăng trưởng này **không ổn định**, biến động nhiều.

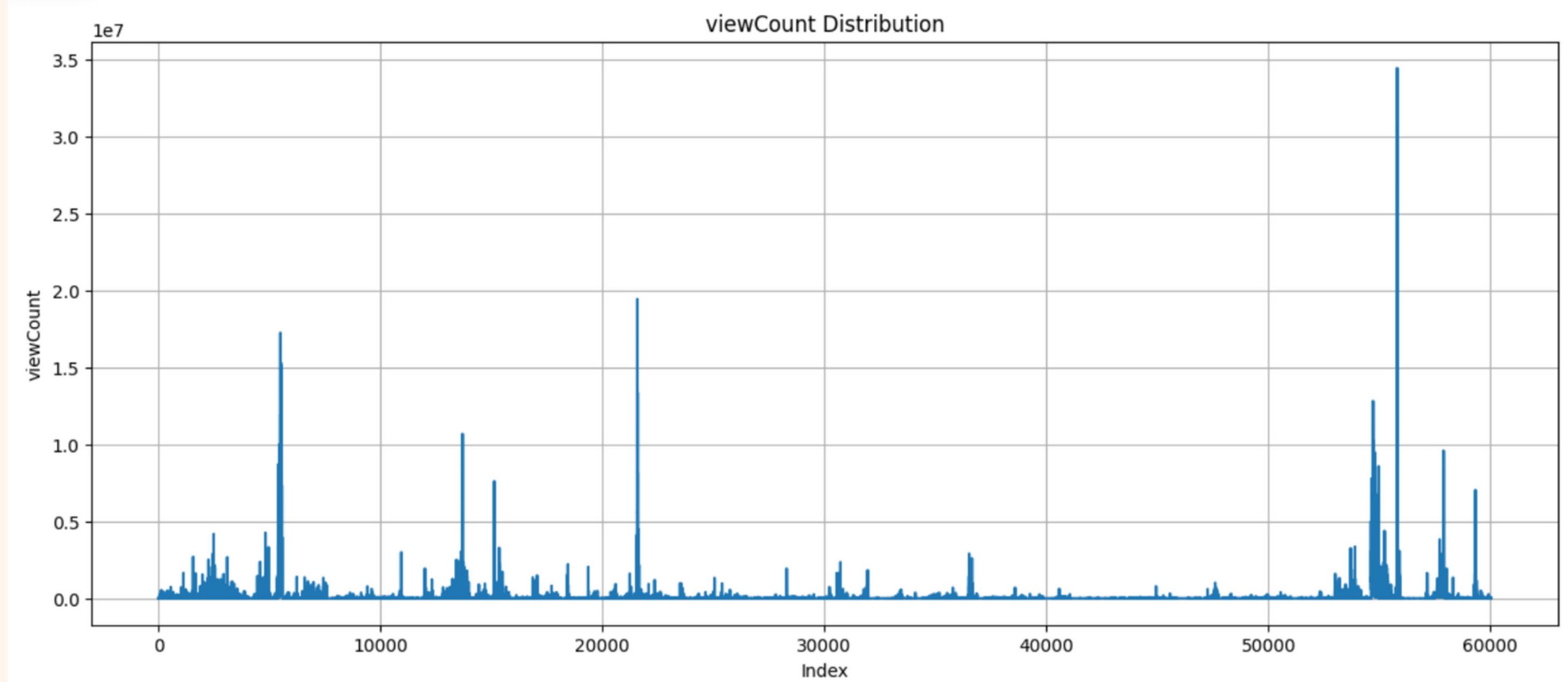
Modeling

◎ **Bài toán đặt ra:** Phân loại video thành mức độ thịnh hành, mục tiêu là dự đoán liệu một video có khả năng trở thành xu hướng hay không.

◎ **Ý nghĩa:**

- Dự đoán xu hướng video có thể mang lại lợi ích đáng kể cho nền tảng video trực tuyến, người quản lý nội dung, nhà quảng cáo và nâng cao trải nghiệm người dùng.
- Nó hữu ích trong việc dự đoán xu hướng và giúp người sáng tạo nội dung tối ưu hóa chiến lược phát hành video của họ.

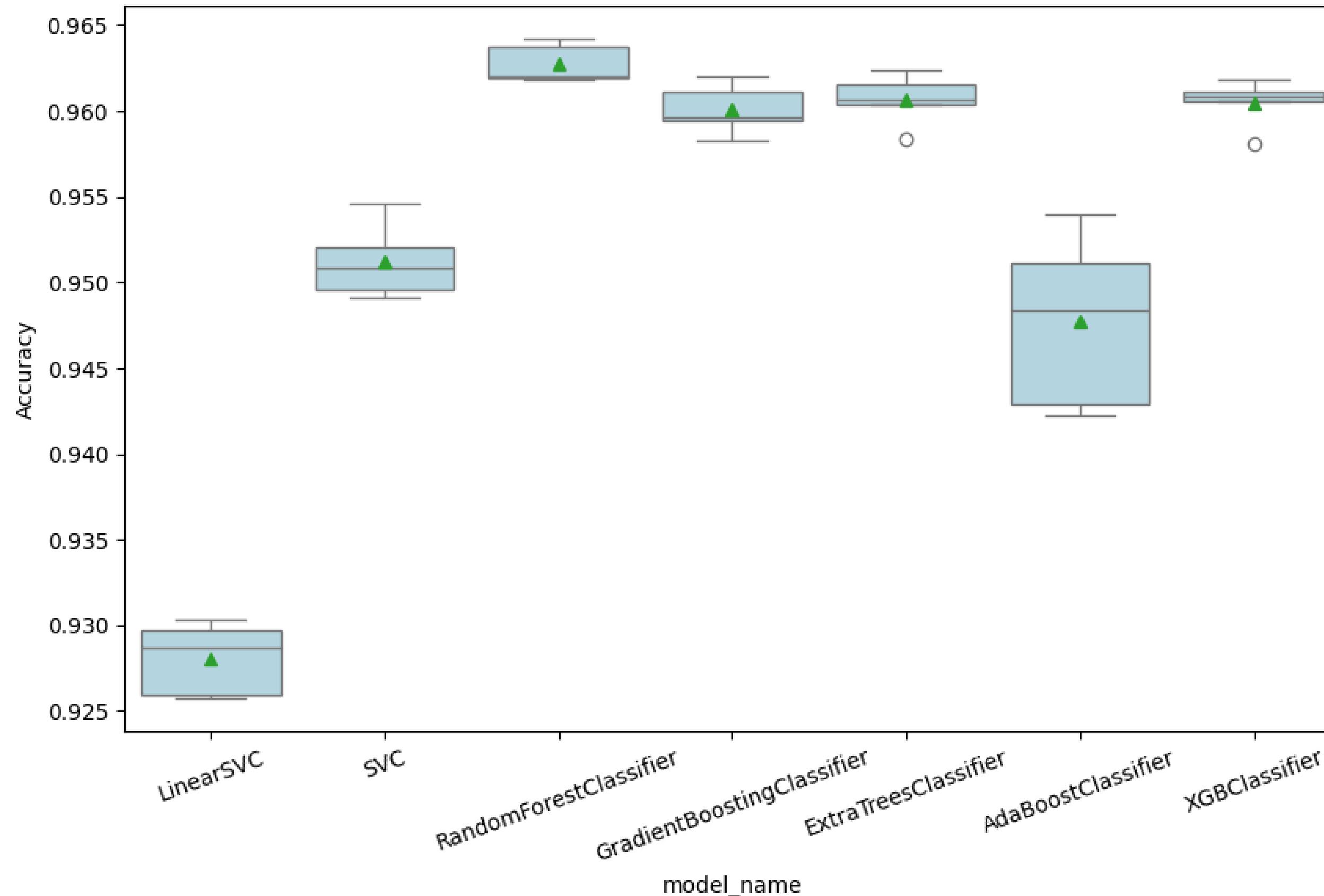
Feature engineering



- Biến mục tiêu sẽ là **Trending**
- Gán nhãn bằng cách dựa vào sự phân bố giá trị, lấy giá trị trung bình làm trung tâm
- Giả định: **low**: 0, **medium**: 1, **high**: 2

So sánh các baseline model

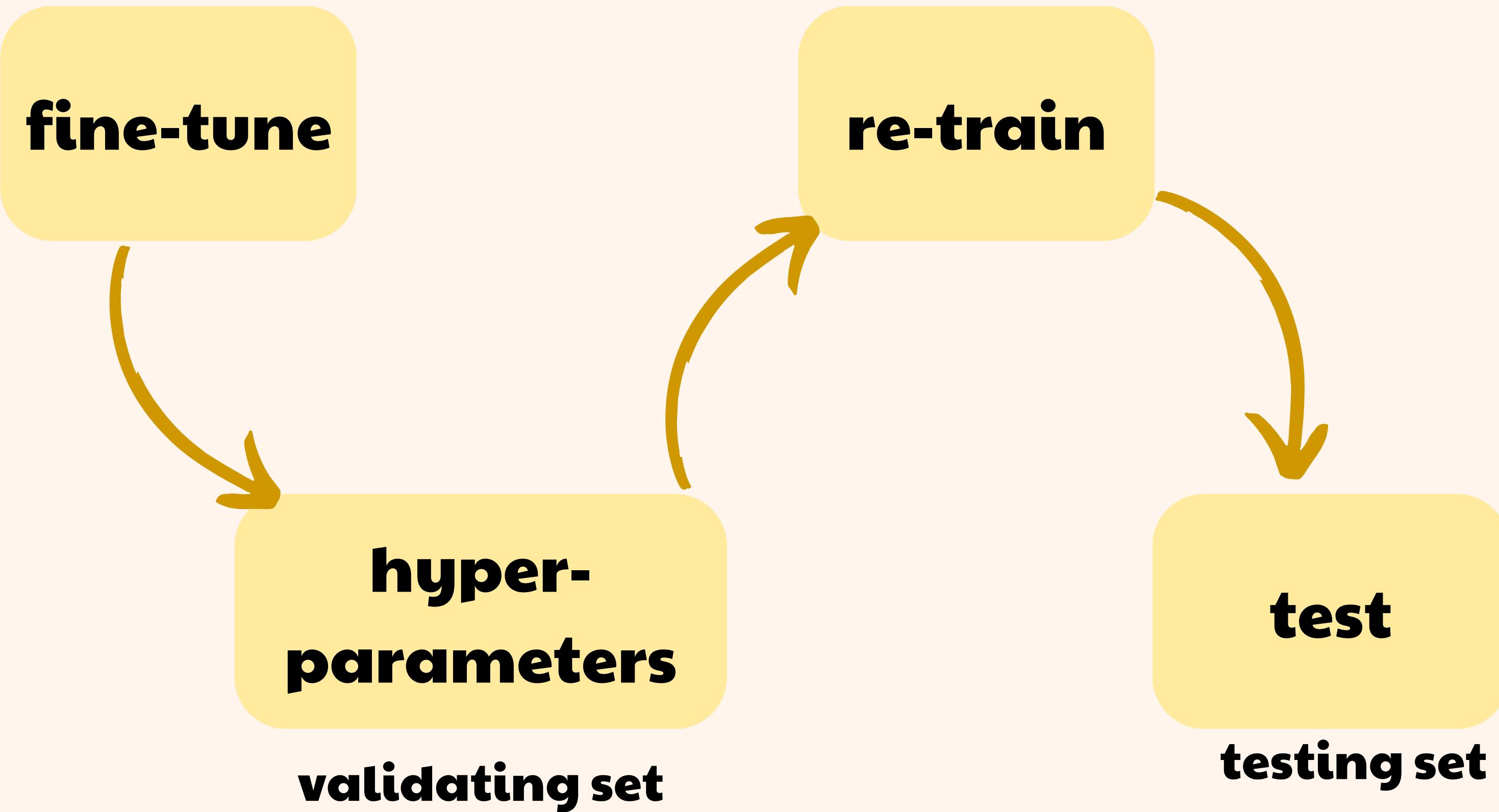
Boxplot of Base-Line Model Accuracy using 5-fold cross-validation



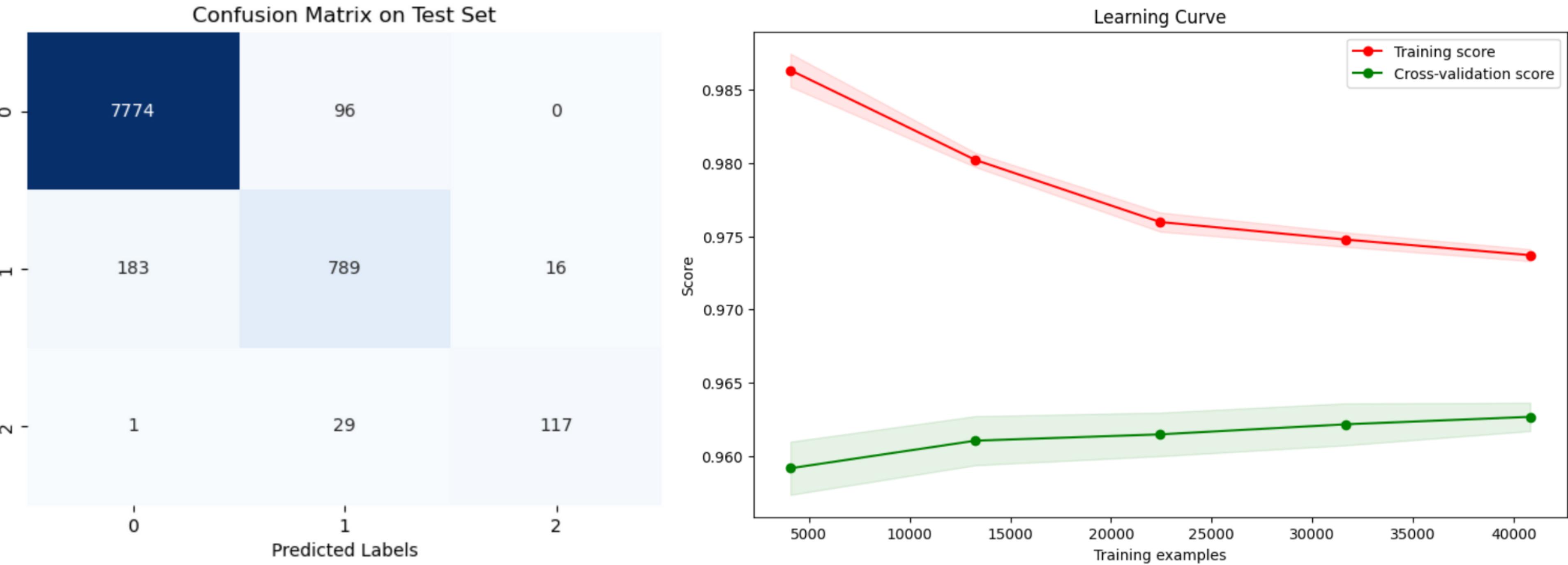
Process

GridSearchCV

training + validating set.



Trực quan hóa model



Nhận xét kết quả

○ Nhận xét:

- Đã chọn được mô hình phù hợp với yêu cầu bài toán, đưa ra độ chính xác tương đối cao.
- Quá trình fine-tune mô hình có thể giúp cải thiện accuracy, nhưng sẽ không đáng kể khi accuracy ban đầu đã khá cao.
- Có thể dẫn đến overfitting hoặc gặp khó khăn do đã đạt giới hạn tối đa hiệu suất.

○ Những cải tiến trong tương lai:

- Tăng số lượng dữ liệu đầu vào có thể giúp mô hình hội tụ tốt hơn.
- Tìm thêm các siêu tham số có thể giúp mô hình chính xác hơn, nhưng có thể làm giảm hiệu suất.

ỨNG DỤNG ĐỀ XUẤT VIDEO

○ Mô tả bài toán

- Sử dụng dữ liệu thu thập để tạo hệ thống đề xuất video thông qua một đoạn văn mà người dùng nhập vào công cụ tìm kiếm.
- Giải quyết bài toán này sẽ giúp ích cho cả người dùng và nền tảng phát sóng:
 - Phía người dùng: Có vô số video về nội dung mà người dùng quan tâm, vậy họ nên chọn video nào? Khi này, hệ thống đề xuất của chúng ta sẽ gợi ý những video mà người dùng có thể muốn xem nhất, giúp họ tiết kiệm thời gian...
 - Phía nền tảng phát sóng: Việc đề xuất những video mà người dùng muốn xem sẽ giúp khách hàng sử dụng dịch vụ của họ nhiều hơn. Và có càng nhiều người dùng thì họ sẽ có càng nhiều doanh thu...

⌚ Feature engineering - Chuẩn bị dữ liệu

- Sử dụng các thuộc tính sẵn có (như lượt xem, số lượng like và comment, v.v.) để tạo một thuộc tính **score** giúp “xếp hạng” các video thông qua một hàm “bí mật doanh nghiệp”.
- Sau đó ta sẽ “làm sạch” tiêu đề của các video: loại bỏ các icon, stopword, v.v. thông qua hàm: **preprocess_text(...)**
- Vì mô hình học máy thường chỉ hiểu dữ liệu dạng số nên ta cần khởi tạo **TfidfVectorizer** giúp chuyển hóa tiêu đề thành một ma trận thưa (**sparse matrix**)
TfidfVectorizer(max_features=20_000, dtype=np.float32, ...)
- Ta đồng thời chia dữ liệu về tiêu đề thành ba tập: train, valid và test để sử dụng ở các bước tiếp theo thông qua hàm: **train_test_split(...)**

○ Lựa chọn mô hình và phương pháp đánh giá

- Để giải quyết bài toán, ta cần chọn các mô hình có khả năng phân lớp dữ liệu dựa vào tiêu đề của video sau khi được chuyển thành “ma trận thưa”.
- Tuy nhiên không nhiều mô hình có khả năng này, và thư viện `sklearn` chỉ cung cấp cho chúng ta ba lựa chọn:

`KMeans(...)`

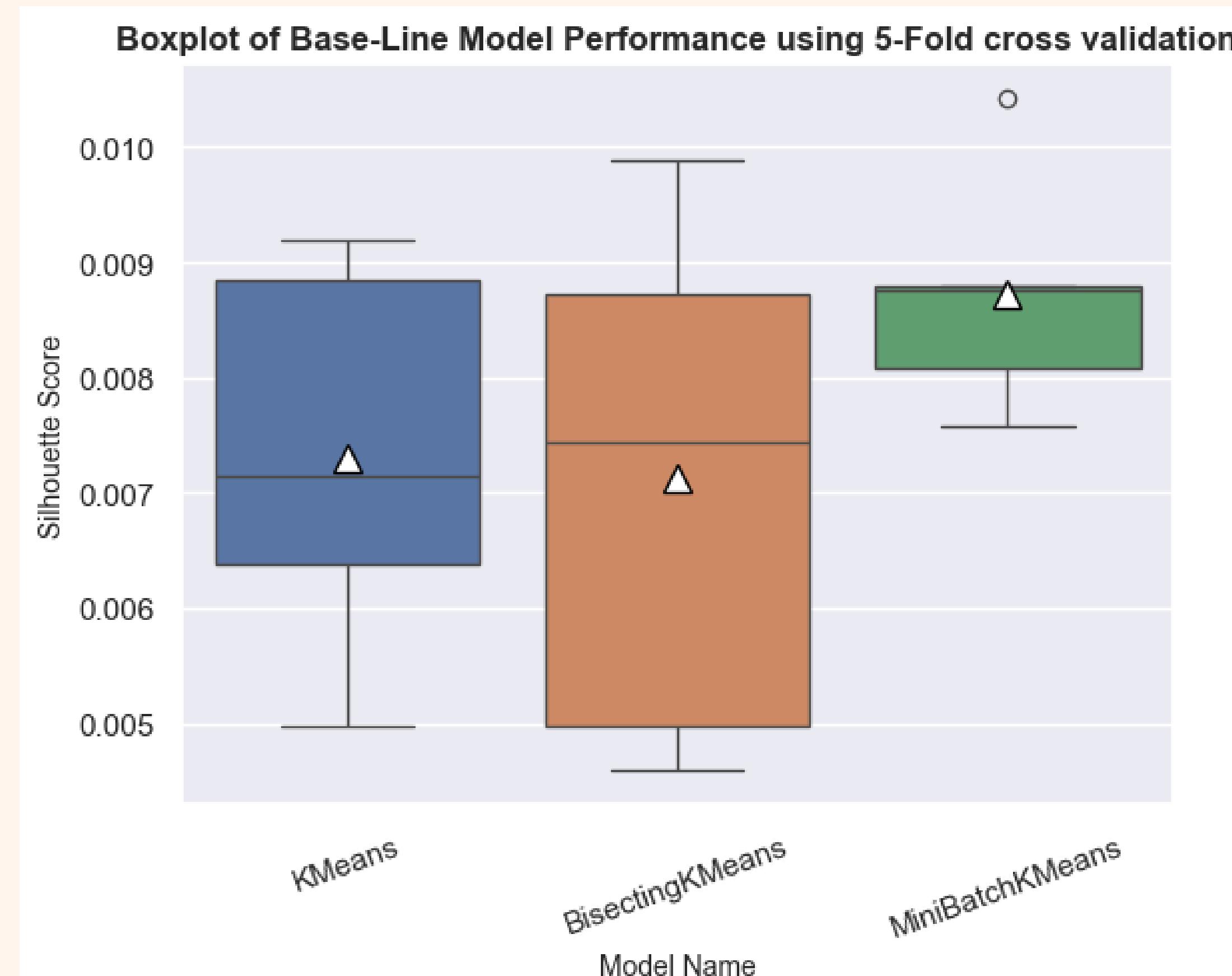
`BisectingKMeans(...)`

`MiniBatchKMeans(...)`

- Việc lựa chọn `metric` để đánh giá các mô hình phân loại khá khó khăn. Đặc biệt là khi phải tính toán trên “ma trận thưa” thì càng khó khăn hơn nữa.
- Nhưng may mắn là `sklearn` có cung cấp cho chúng ta một metric đáp ứng tất cả yêu cầu trên là:

`silhouette_score(...)`

○ So sánh các baseline models



◎ Xác định “số lượng cụm” tối ưu cho từng mô hình

- Ta sẽ sử dụng phương pháp “**Elbow method**” (phương pháp khuỷu tay) để tìm các giá trị “**n_clusters**” mà ta nghi ngờ là tối ưu cho từng mô hình.
- Việc làm này nhằm thu hẹp miền giá trị của tham số “**n_clusters**”, giúp chúng ta tiết kiệm rất nhiều thời gian khi thực hiện fine-tune từng model.

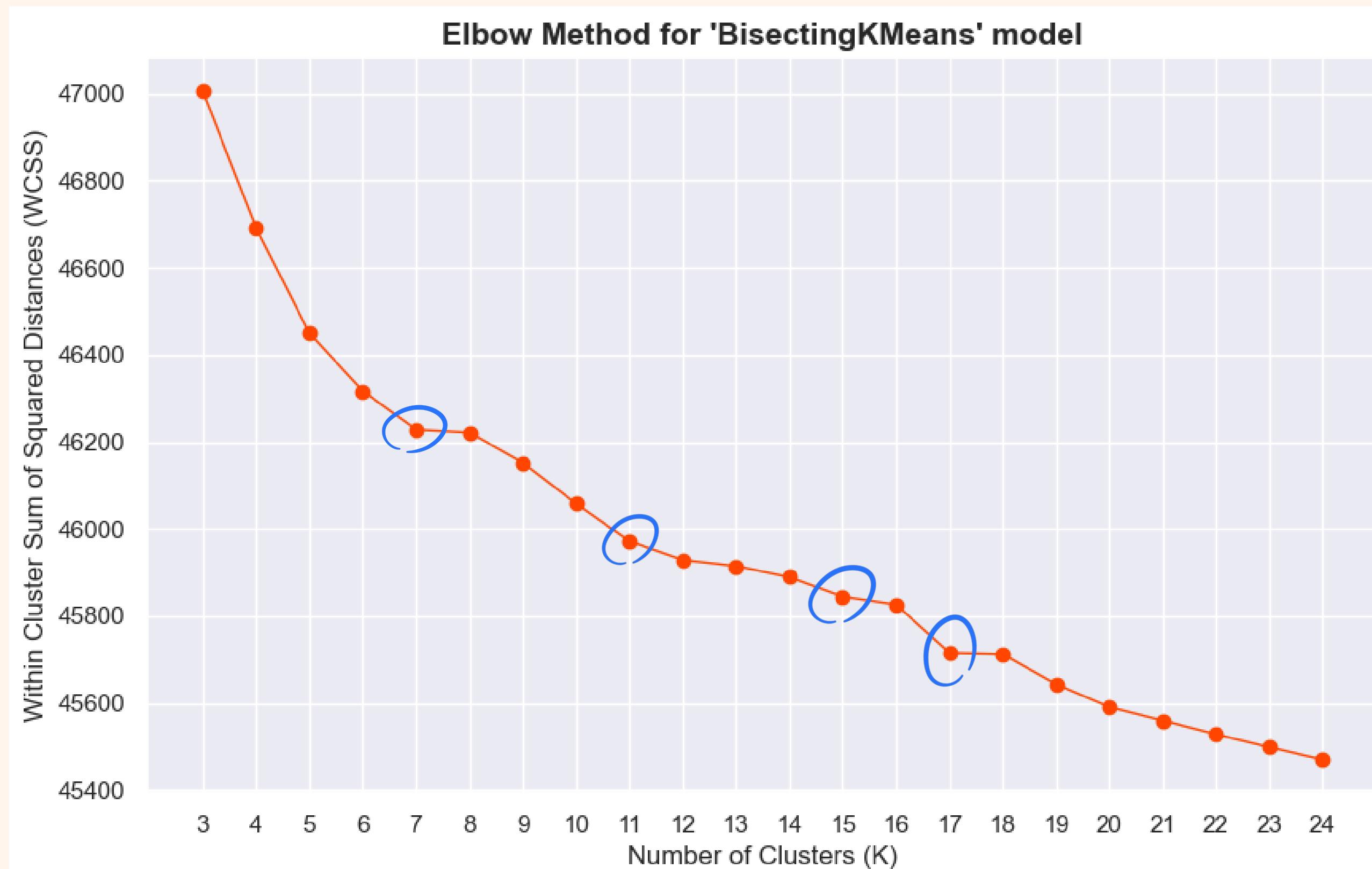
○ Xác định “số lượng cụm” tối ưu cho từng mô hình

- Kết quả của KMeans:



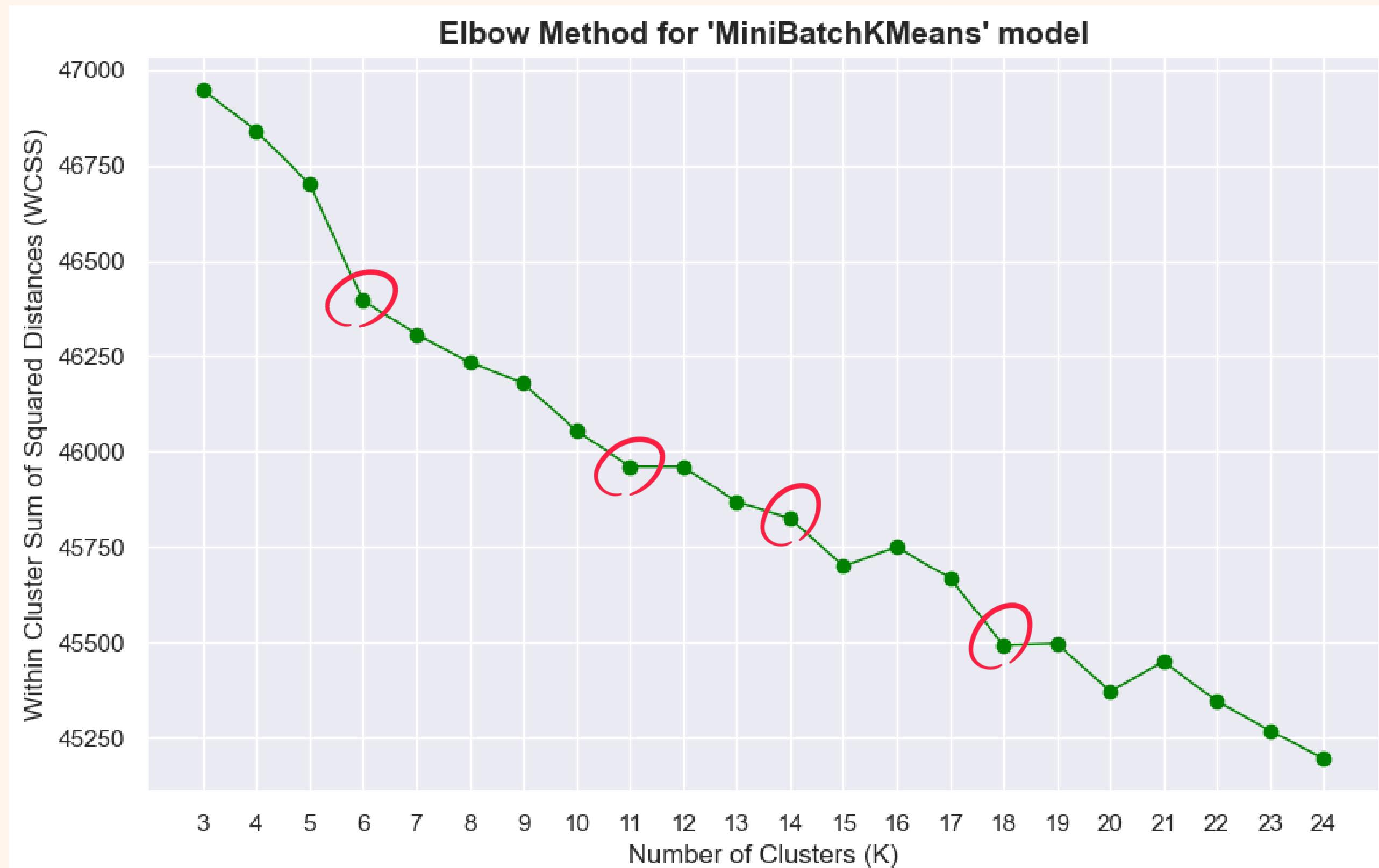
○ Xác định “số lượng cụm” tối ưu cho từng mô hình

- Kết quả của BisectingKMeans:



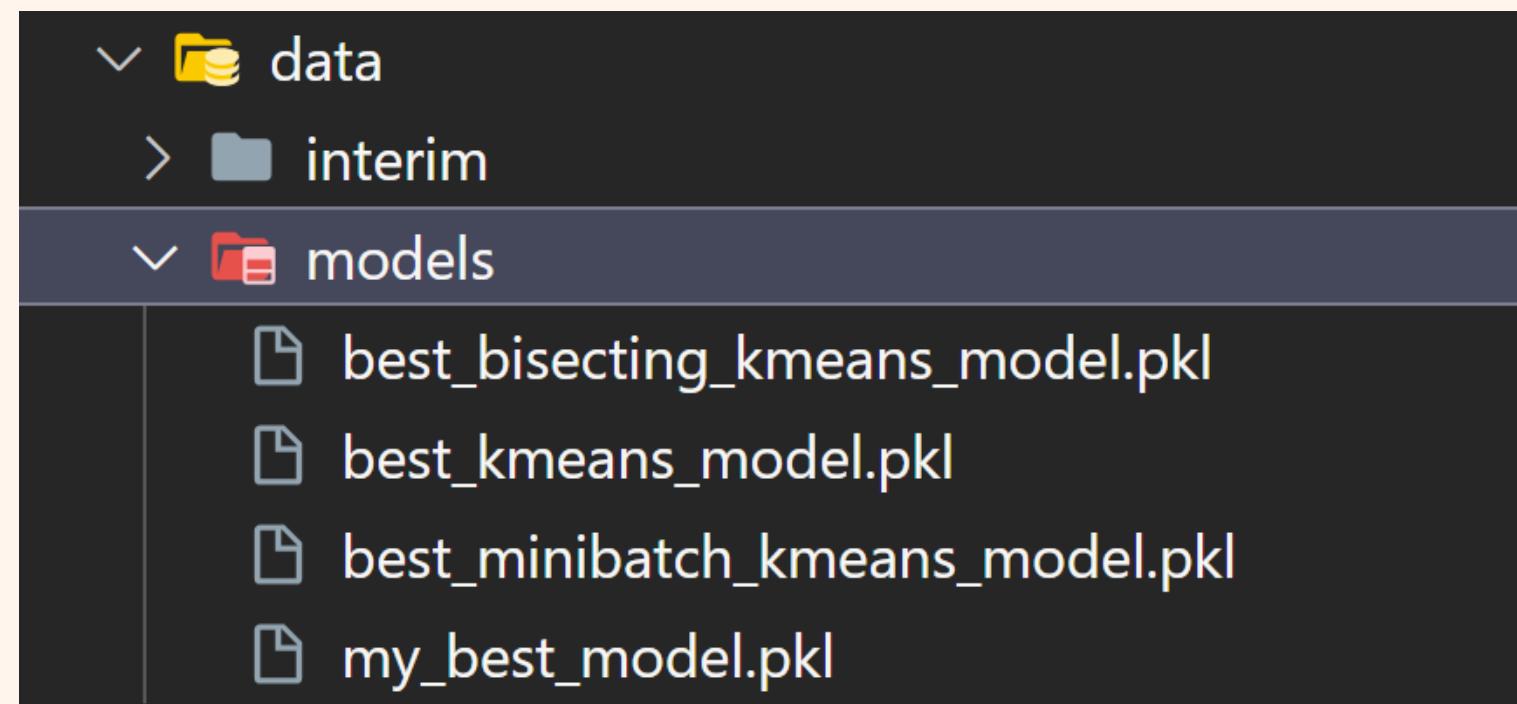
○ Xác định “số lượng cụm” tối ưu cho từng mô hình

- Kết quả của MiniBatchKMeans:



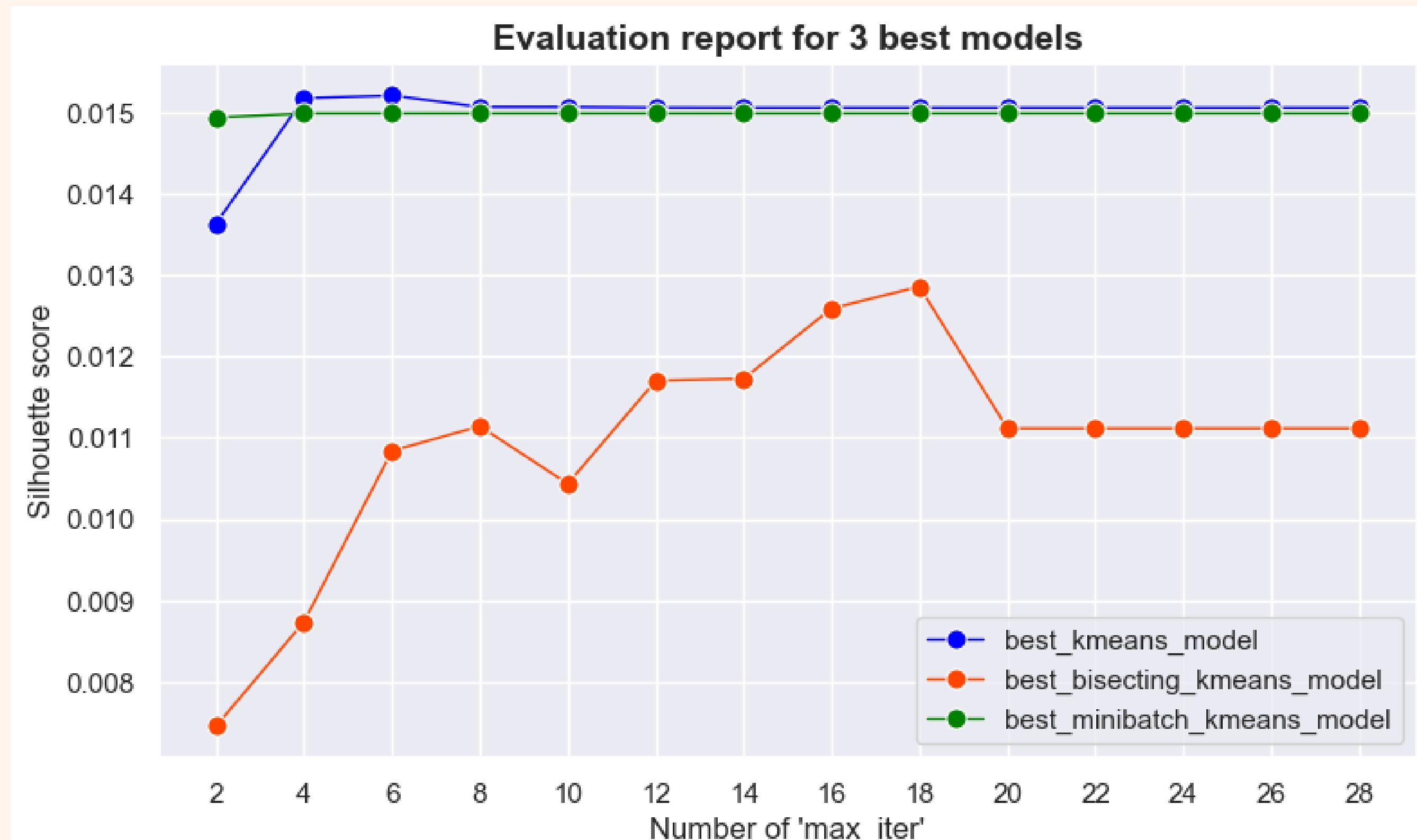
◎ Tinh chỉnh (fine-tune) các mô hình

- Từ kết quả vừa tìm được, kết hợp với các nguồn tài liệu khác, ta sẽ tạo ra các tổ hợp siêu tham số khác nhau cho từng model.
- Ứng với mỗi bộ siêu tham số, ta sẽ huấn luyện mô hình trên tập **train** và đánh giá kết quả trên tập **validation** để xác định bộ siêu tham số tốt nhất cho từng model.
- Vì quá trình này **tốn rất nhiều thời gian** nên ta không thể thường xuyên chạy lại quy trình này để tìm các giá trị tối ưu. Do đó, kết quả ở trên sẽ được lưu lại trong một file ***.pkl** để tiện sử dụng:



◎ Huấn luyện (lại) và đánh giá các mô hình

- Đánh giá trên cả ba mô hình:



◎ **Triển khai mô hình - Mô hình hóa Ứng dụng**

- Ta sẽ sử dụng mô hình tốt nhất vừa tìm được để phân cụm cho toàn bộ dữ liệu, khi này mỗi video trong tập dữ liệu sẽ thuộc vào một cụm nhất định.
- Với mỗi đoạn văn bản do người dùng nhập vào, ta lần lượt thực hiện các bước tiền xử lý và dự đoán `cluster_id` cho đoạn văn đó.
- Khi này, ứng dụng của chúng ta sẽ chạy một thuật toán giúp đề xuất các video trong cụm tương ứng. Đó có thể là các video mà người dùng có khả năng xem cao nhất.

Triển khai mô hình - Minh họa ứng dụng

#OriginI
@21127739

What are you looking for?

Focus here to search

Recommended Tags

"business intelligence", "data analyst skills", "data analytics", "data analyst job", "google data analyst"

Recommended Videos

Channel	Video title	URL
1 Luke Barousse	Become a DATA ANALYST with NO degree?!? The Google Data Analytics Professional Certificate	https://www.youtube.com/watch?v=fmLPS6FBbac
2 Shashank Kalanithi	Day in the Life of a Data Analyst - SurveyMonkey Data Transformation	https://www.youtube.com/watch?v=pKvWD0f18Pc
3 Stefanovic	FASTEAST Way to Become a Data Analyst and ACTUALLY Get a Job	https://www.youtube.com/watch?v=AYWLZ1IES6g
4 Alex The Analyst	Data Analyst Portfolio Project SQL Data Exploration Project 1/4	https://www.youtube.com/watch?v=qfyynHBFOsM
5 Applied AI Course	Akash Singh Joined Myntra as Data Analyst Data Analyst Interview Applied Ai Course Review	https://www.youtube.com/watch?v=Ccy1UuhPY4A
6 Learn with Lukas	Learn 50 Data Analyst Concepts In 6 Minutes	https://www.youtube.com/watch?v=LBSN34puYcQ
7 Anthony Smoak	Data Visualization: The Must-Have Skill for Every Data Analyst	https://www.youtube.com/watch?v=FbpI5Wsjo04
8 Data With Mo	How I explore data using Python as a Data Analyst	https://www.youtube.com/watch?v=hA0qyW-w3pQ
9 codebasics	Data Analyst Project For Beginners (HR Analytics): 5 - Dashboarding	https://www.youtube.com/watch?v=q0-XClu0fSc
10 Krish Naik	Live- Conversation With Amit Bose-Transition From Commerce Background To Data Analyst	https://www.youtube.com/watch?v=edQA8nvTLbg

◎ Nhận xét kết quả đã đạt được

- Chỉ với những nguồn tài nguyên ít ỏi, ta đã thành công xây dựng một hệ thống để xuất **video** và **tags** đơn giản, giải quyết được bài toán mà ta đã đề ra ban đầu.
- Tuy nhiên, những hạn chế nhất định về phần cứng (**RAM**) làm cho từ điển của chúng ta có sức chứa chưa đủ lớn để đáp ứng nhu cầu sử dụng trong thực tế.
- Những điều có thể cải thiện trong tương lai:
 - Nếu có nguồn tài nguyên máy tính tốt hơn thì ta có thể mở rộng phạm vi của từ điển, giúp nó chứa được nhiều từ hơn. Khi nhận một đoạn văn do người dùng nhập vào thì ta sẽ có thêm nhiều thông tin, giúp mô hình phân cụm tốt hơn.
 - Ngoài ra, ta sẽ tìm hiểu thêm các thuật toán đánh giá điểm số cho một video, hiện tại ta chỉ sử dụng các số liệu thô một cách đơn giản. Việc có một thang đánh giá "tổng quát hơn" sẽ cải thiện hiệu suất hoạt động của hệ thống lên đáng kể...

Tài liệu tham khảo

- ◎ Tất cả tài liệu tham khảo đều được đề cập tại file [list_references.md](#) trong folder `./src`