

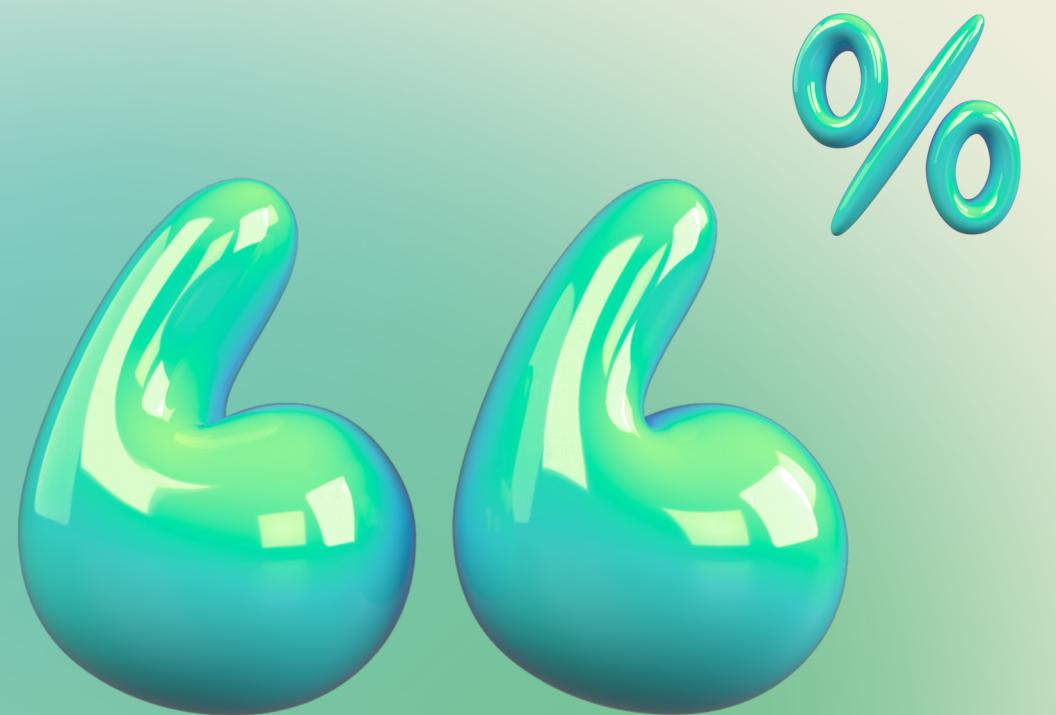


HAHA GROUP - 22KHDL1

INTRODUCTION TO DATA SCIENCE

# THÀNH VIÊN NHÓM

6 %



**ĐẶNG CHÂU ANH**

22127008



**NGUYỄN KIM ANH**

22127014



**TRẦN DỊU HUYỀN**

22127170



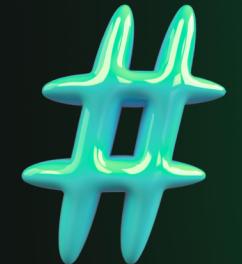
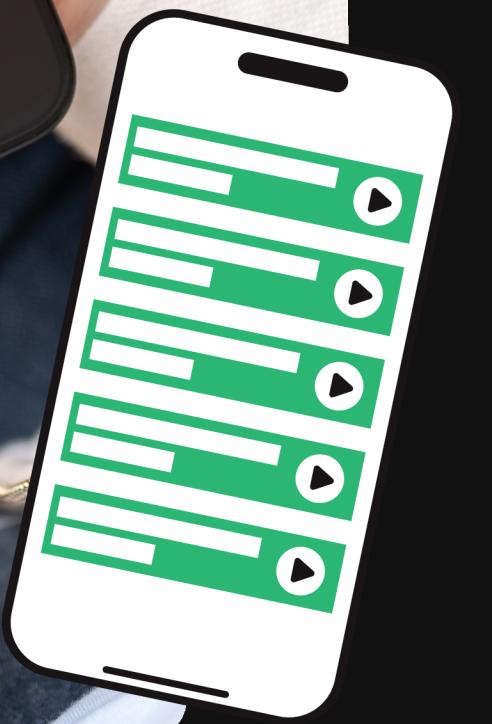
**ĐỖ MINH HUY**

22127147

# GIỚI THIỆU

## ĐỐI TƯỢNG PHÂN TÍCH

**Spotify** là một dịch vụ nghe nhạc trực tuyến (music streaming) rất phổ biến trên thế giới. Nó cũng đạt được mức độ phổ biến đáng kể tại Việt Nam. Do đó, phân tích được xu hướng trong âm nhạc là cần thiết cho đối tượng song phương: người nghe và nghệ sĩ. Đặc biệt, phân tích ra xu hướng âm nhạc trong tương lai, giúp các nghệ sĩ sáng tạo ra những sản phẩm âm nhạc phù hợp với thị hiếu của khán giả.



# NỘI DUNG



THU THẬP DỮ LIỆU

KHÁM PHÁ DỮ LIỆU

TRỰC QUAN HÓA & CÂU HỎI

MÔ HÌNH HÓA DỮ LIỆU

# NỘI DUNG



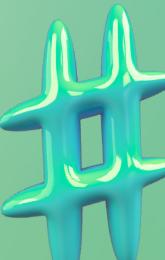
THU THẬP DỮ LIỆU



KHÁM PHÁ DỮ LIỆU

TRỰC QUAN HÓA & CÂU HỎI

MÔ HÌNH HÓA DỮ LIỆU





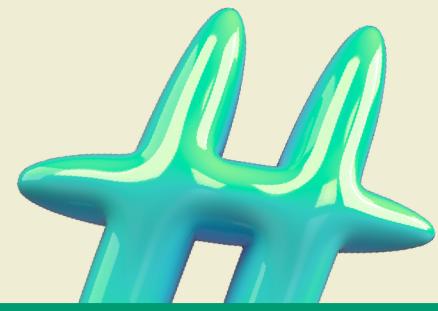
# THÔNG TIN VỀ DỮ LIỆU



<b>Thông tin</b>	<p><b>BXH:</b> Spotify Rank 200 của tháng trong năm 2024</p> <p><b>Thời gian:</b> tháng 1 - 10/2024</p> <p><b>Data source:</b> <u><a href="#">Spotify Chart</a></u>, <u><a href="#">kworb</a></u>        → Đều là trang web mã nguồn mở và không vi phạm quyền sở hữu trí tuệ của Spotify. Thông tin của chúng lấy từ API của Spotify và tổng hợp từ fan.</p>
<b>Dòng, cột</b>	2000 dòng, 10 cột

rank	uri	artist_names	track_name	source	peak_rank	previous_rank	weeks_on_chart	streams	month
0	1 spotify:track:2HRgqmZQC0MC7GeNuDIXHN	Jung Kook, Latto	Seven (feat. Latto) (Explicit Ver.)	BIGHIT MUSIC	1	1	29	1178606	1
1	2 spotify:track:4qYfRfSxsmhJ1WMaywtLyl	Wren Evans, itsnk	Tùng Quen	Universal Music Indochina	1	2	14	1087520	1
2	3 spotify:track:0X28PqBpbQhWdi4usYw0w5	Wren Evans, itsnk	Tò Te Tí	Universal Music Indochina	3	4	7	1036739	1
3	4 spotify:track:1khMN4Adfi3LrZvxOq4YM5	VSOUL, RPT MCK, Obito, Ronboogz, Boyzed	Buồn Hay Vui (feat. RPT MCK, Obito, Ronboogz &...)	12 trái lê	3	5	6	919584	1
4	5 spotify:track:1bG6Q8sR8jda7ryl365y8o	Vũ., Dear Jane	Những Lời Hứa Bỏ Quên	WM Vietnam	1	3	7	856107	1
...	...	...	...	...	...	...	...	...	...
1995	196 spotify:track:7sZgr8RsXkDwkmiQok691a	Thịnh Suy	Mai Minh Xa	InQ International	85	187	28	145969	10
1996	197 spotify:track:2M2nM9t0YQoXMTnFH8AXBu	Huỳnh Tú, Andiez	Đường Một Chiều	Huỳnh Tú	114	191	67	145160	10
1997	198 spotify:track:45Mswno1F7FoZkcmQkp7fi	Wren Evans	Thích Em Hơn Nhiều	Universal Music Indochina	1	-1	168	145003	10
1998	199 spotify:track:23ep27rDA9gklzuJ6qzRD	Dương Domic	Yêu Em 2 Ngày	DAO Entertainment	174	-1	5	143566	10
1999	200 spotify:track:2jcdws2044suWnDo8DTcKC	RAP VIỆT, CoolKid	NẮNG (feat. Coolkid)	DAT VIET VAC	200	-1	1	143534	10

2000 rows × 10 columns



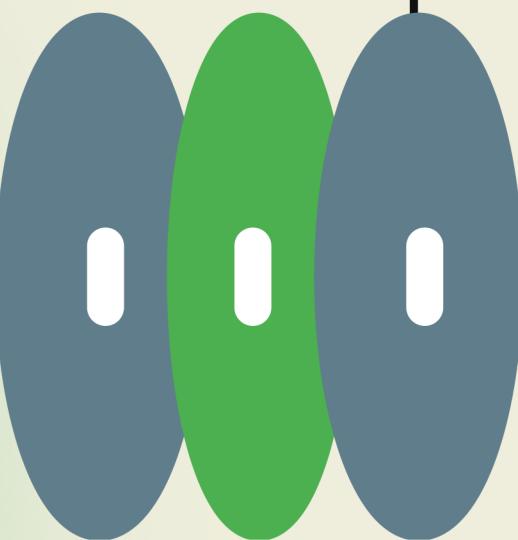
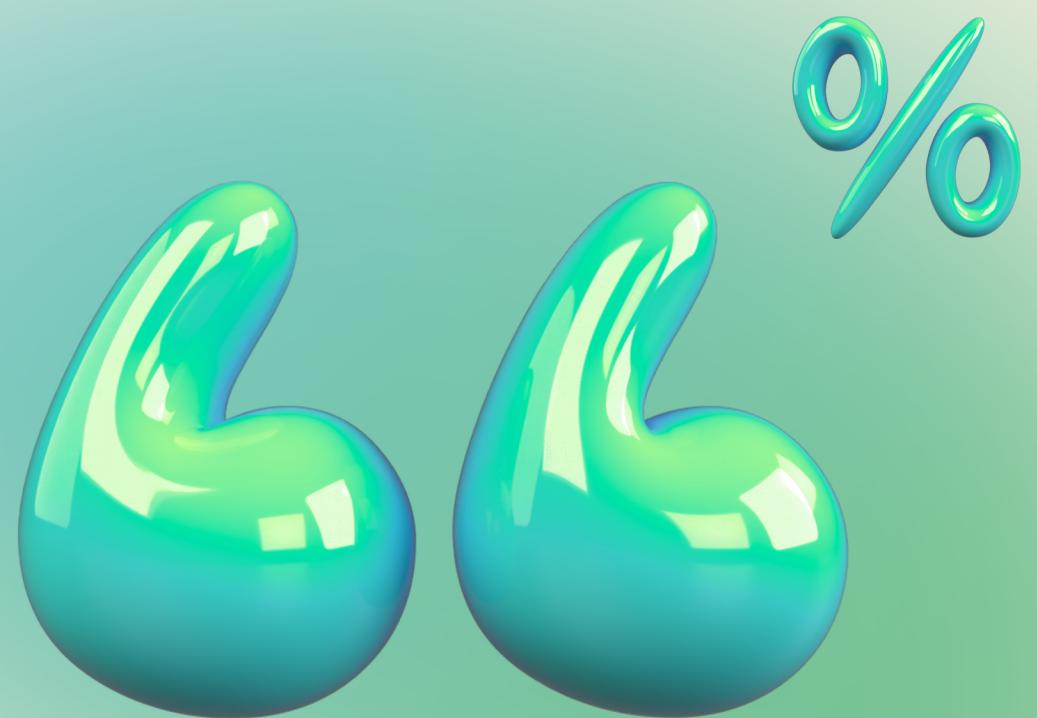
# TRƯỜNG VÀ KIỂU DỮ LIỆU

%

<b>rank</b>	Thứ hạng của bài hát	numerical
<b>URI</b>	Mã định danh duy nhất cho bài hát trên Spotify	categorical
<b>artist_names</b>	Tên nghệ sĩ	categorical
<b>track_name</b>	Tên bài hát	categorical
<b>source</b>	Hãng thu âm hoặc nhà phân phối	categorical
<b>peak_rank</b>	Vị trí cao nhất mà bài hát đạt được trên bảng xếp hạng	numerical
<b>previous_rank</b>	Vị trí trong tuần trước	numerical
<b>weeks_on_chart</b>	Số tuần bài hát nằm trên bảng xếp hạng	numerical
<b>streams</b>	Số lượt phát trực tuyến	numerical
<b>month</b>	Mã tháng	categorical

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2000 entries, 0 to 1999
Data columns (total 10 columns):
 #   Column      Non-Null Count  Dtype  
--- 
 0   rank        2000 non-null    int64  
 1   uri         2000 non-null    object  
 2   artist_names 2000 non-null    object  
 3   track_name   2000 non-null    object  
 4   source       2000 non-null    object  
 5   peak_rank    2000 non-null    int64  
 6   previous_rank 2000 non-null    int64  
 7   weeks_on_chart 2000 non-null    int64  
 8   streams      2000 non-null    int64  
 9   month        2000 non-null    object  
dtypes: int64(5), object(5)
memory usage: 156.4+ KB
```

# THU THẬP DỮ LIỆU



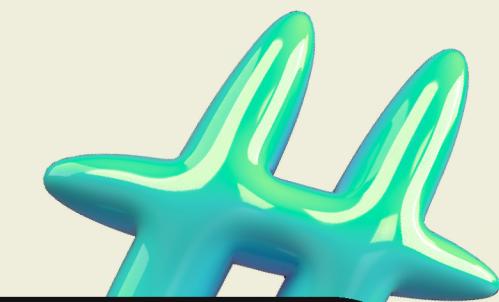
**Xác định** mục tiêu, đối tượng phân tích

**Crawl dữ liệu** BXH top 200 bài hát từ tháng 1 đến tháng 10 năm 2024

**Sử dụng thư viện** Selenium, BeautifulSoup, pandas,...

**Lưu** mỗi tháng thành một file csv và một bảng xếp hạng tổng cho 10 tháng

# CRAWL DỮ LIỆU



01

## Crawl thông tin từ trang Spotify Chart

*BeautifulSoup*: phân tích HTML và trích xuất thêm **release\_date** (ngày phát hành) và **duration** (thời lượng) từ các thẻ meta trong trang HTML

*Selenium*: tự động tìm và lấy dữ liệu **chính** từ trang web.

Trích xuất ID của bài hát từ các URI đã thu thập được ở bước trên và sử dụng thư viện `requests` để gửi yêu cầu HTTP đến trang web Spotify với URL chứa ID của bài hát.

The screenshot shows the Spotify Charts interface for the week of December 26, set to Vietnam. The chart lists three tracks:

#	TRACK	Peak	Streak	Streams	
1	Seven (feat. Latto) (Explicit Ver.) Jung Kook, Latto	1	1	76	2,528,162
2	Who Jimin	1	2	23	2,251,763
3	Mắt Kết Nối Dương Domic	3	3	5	2,123,484



# CRAWL DỮ LIỆU

02

## Crawl thông tin từ trang kworb

Thư viện **requests**: gửi yêu cầu **HTTP** đến trang web với **URL** chứa ID của bài hát.

**BeautifulSoup**: phân tích **HTML** và trích xuất thêm **rank\_history** (lịch sử xếp hạng)

Nếu request fail ở một vài bài hát  
→ request lại. Đối với các bài hát  
không có **rank\_history** thì điền  
bằng danh sách **rỗng** []

## Spotify Chart History

Title: Từng Quen

Artists: Wren Evans, itsnk

Show: **Weekly** · Daily | **Positions** · Streams · Both

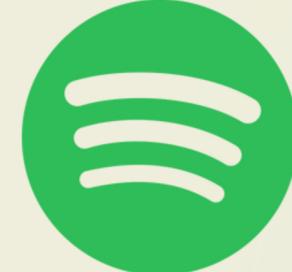
Date	VN
Peak	1
2023/11/02	2
2023/11/09	2
2023/11/16	2
2023/11/23	2
2023/11/30	2
2023/12/07	4
2023/12/14	4
2023/12/21	1
2023/12/28	5
2024/01/04	5
2024/01/11	4
2024/01/18	4
2024/01/25	2

# NỘI DUNG



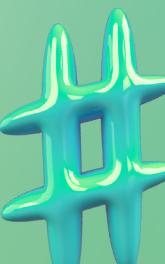
THU THẬP DỮ LIỆU

KHÁM PHÁ DỮ LIỆU



TRỰC QUAN HÓA & CÂU HỎI

MÔ HÌNH HÓA DỮ LIỆU



# KHÁM PHÁ DỮ LIỆU

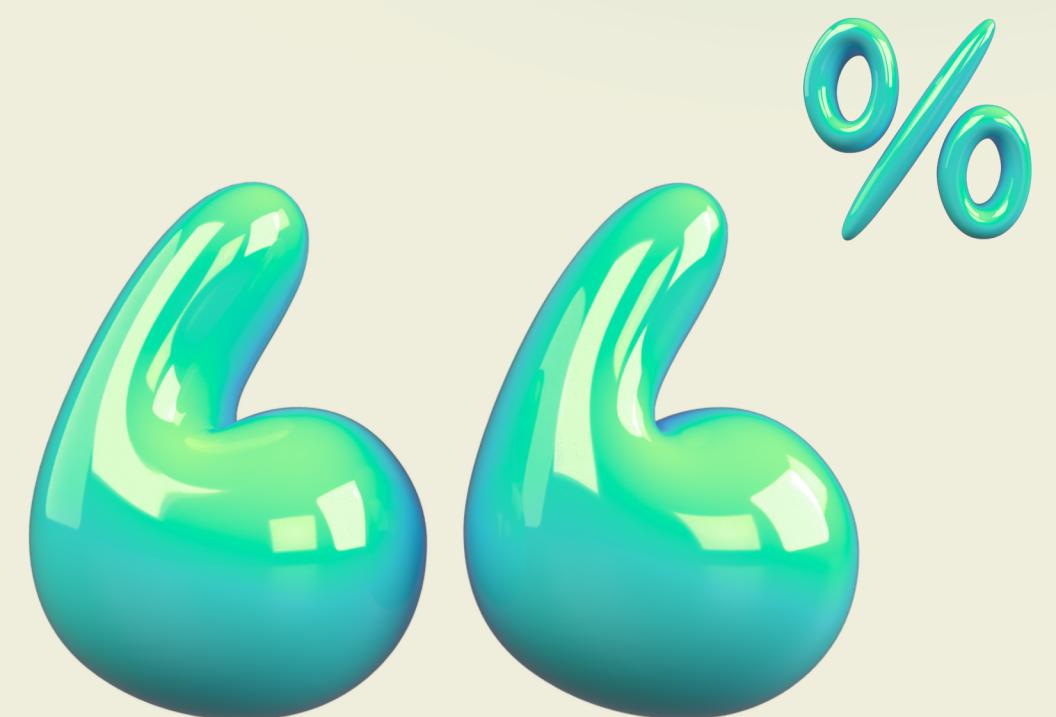
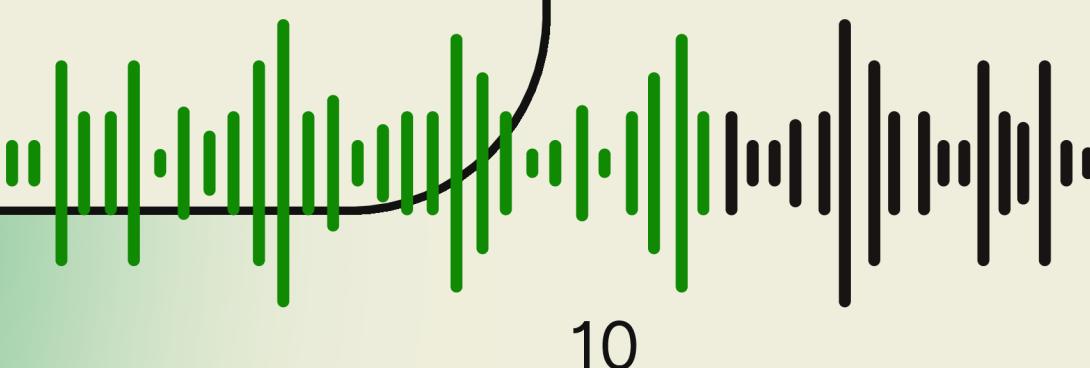
**Khám phá** về cấu trúc dataset, số dòng, số cột, nghĩa của các dòng, cột đó

**Kiểm tra** giá trị thiếu và giá trị trùng lặp

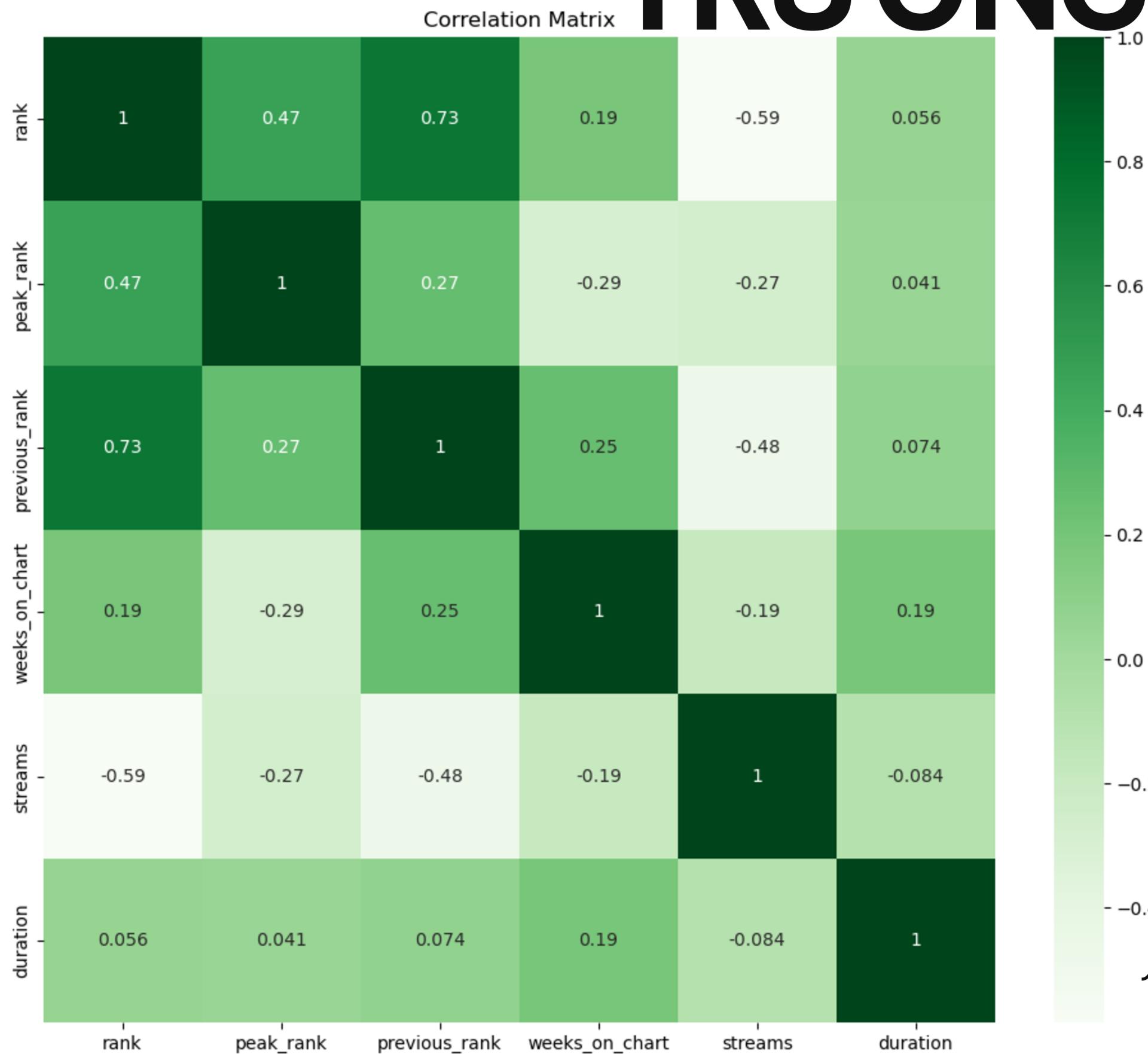
**Tính toán** các giá trị thống kê

**Khám phá** phân bố của các giá trị số và giá trị phân loại

**Kiểm tra** chất lượng dataset



# TƯƠNG QUAN GIỮA CÁC TRƯỞNG DỮ LIỆU



**rank vs. previous\_rank:** Tương quan dương

**rank vs. peak\_rank:** Tương quan dương

**rank vs. streams:** Tương quan âm

# NỘI DUNG

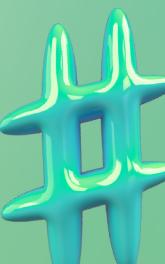


THU THẬP DỮ LIỆU

KHÁM PHÁ DỮ LIỆU

TRỰC QUAN HÓA & CÂU HỎI

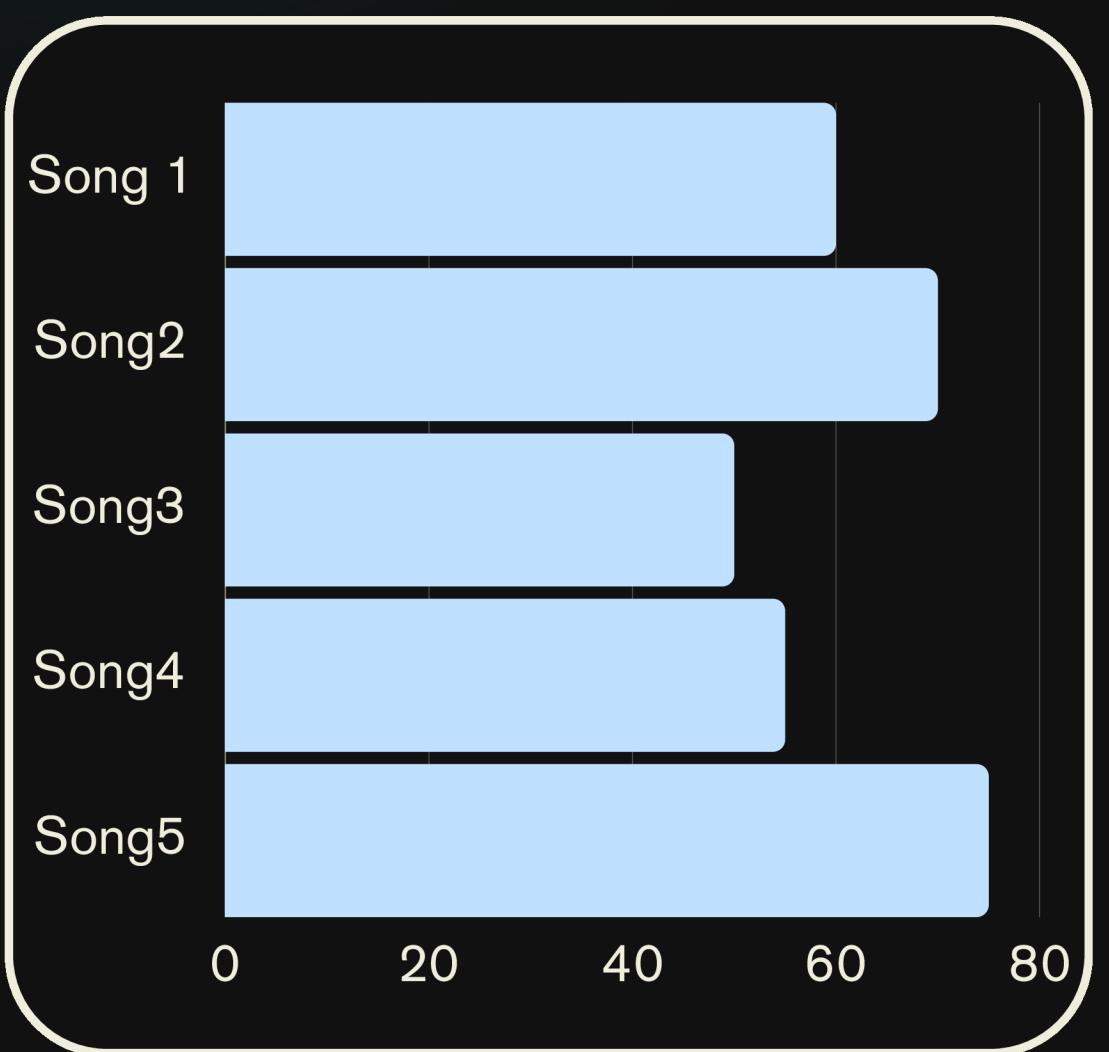
MÔ HÌNH HÓA DỮ LIỆU



# TRỰC QUAN HÓA VÀ CÂU HỎI

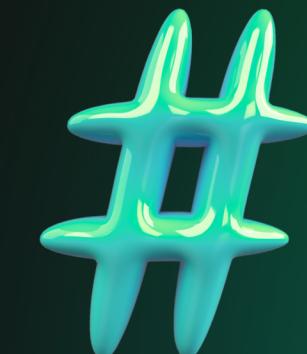


## TRỰC QUAN HÓA



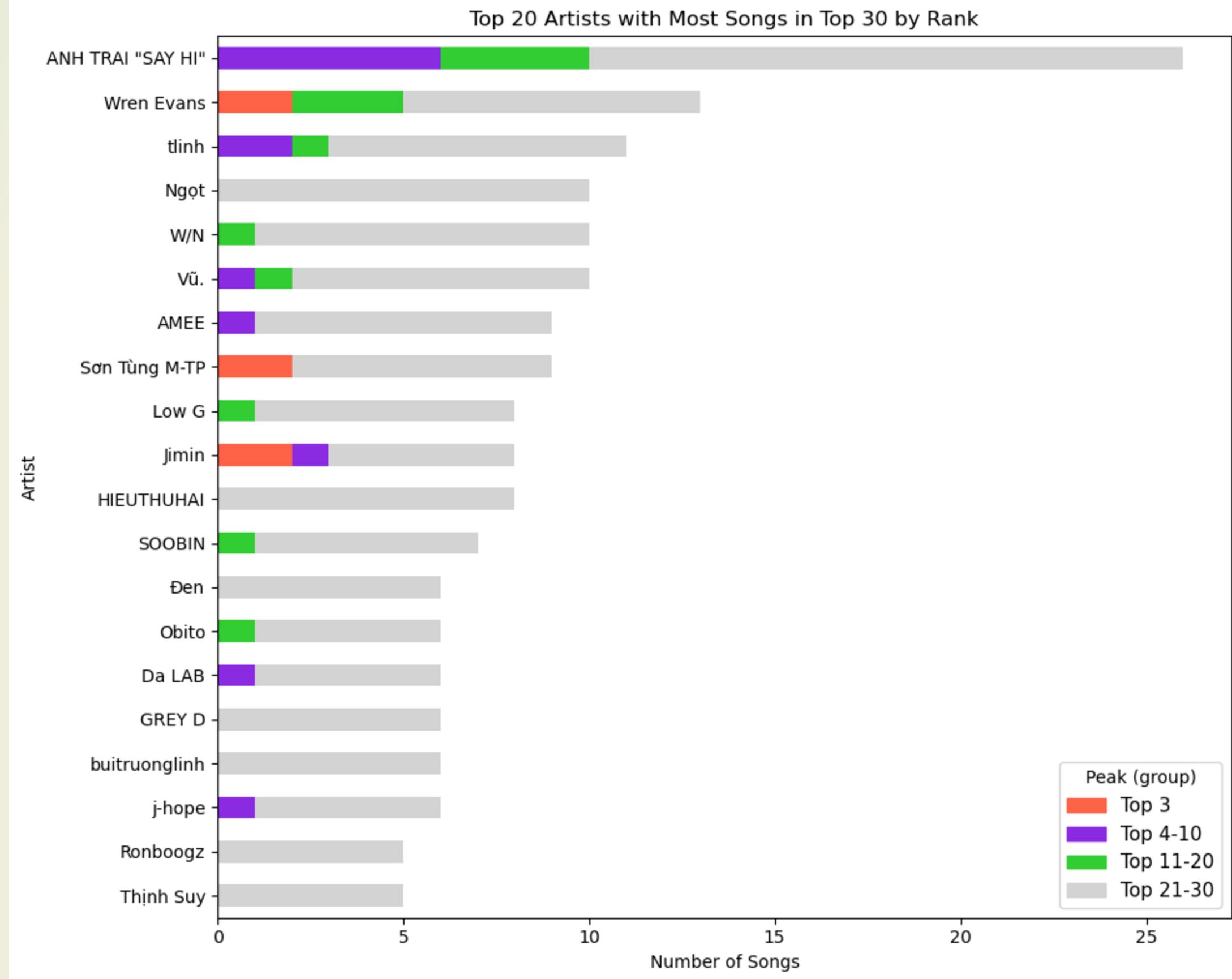
CÂU HỎI?

Ý nghĩa?

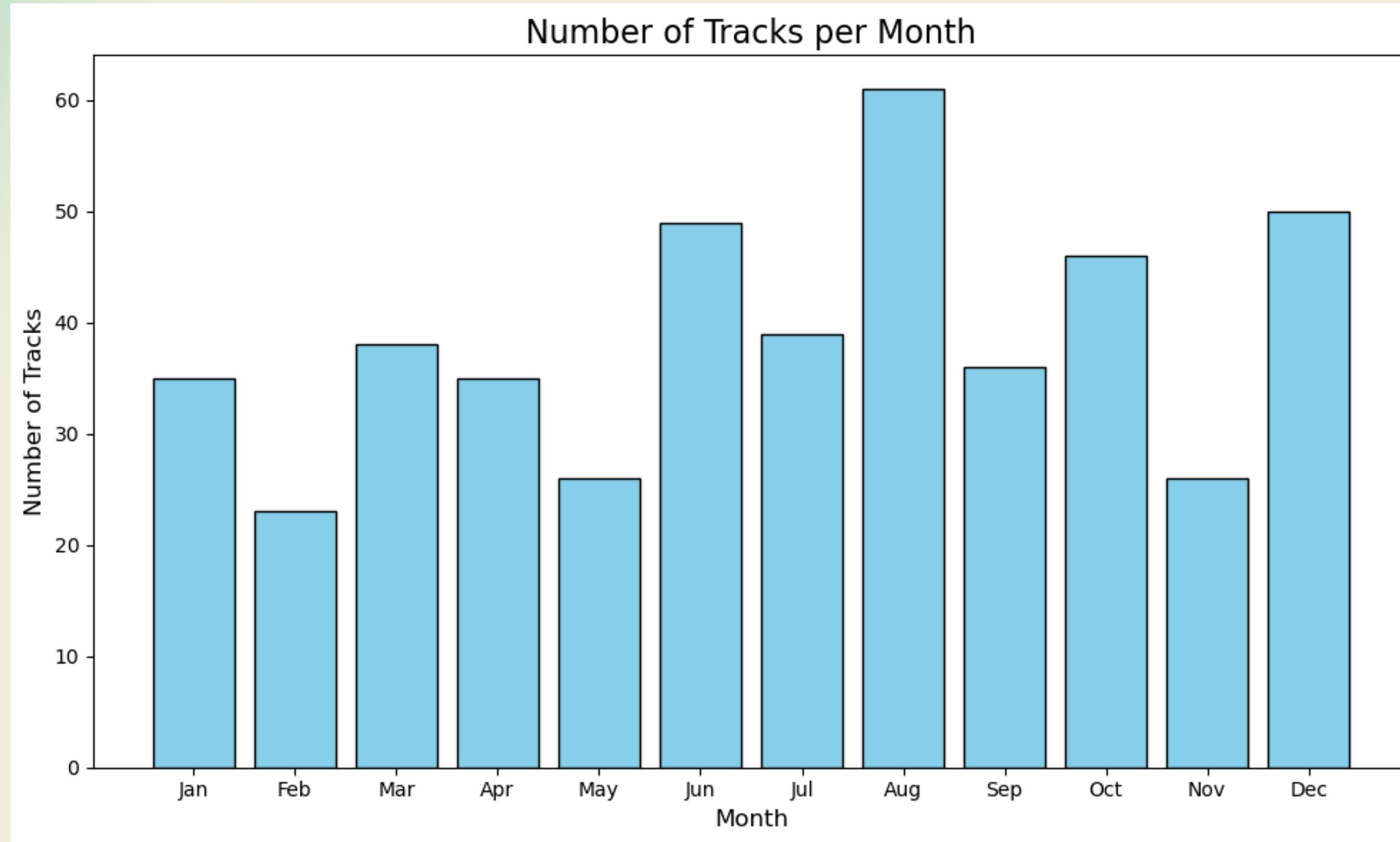




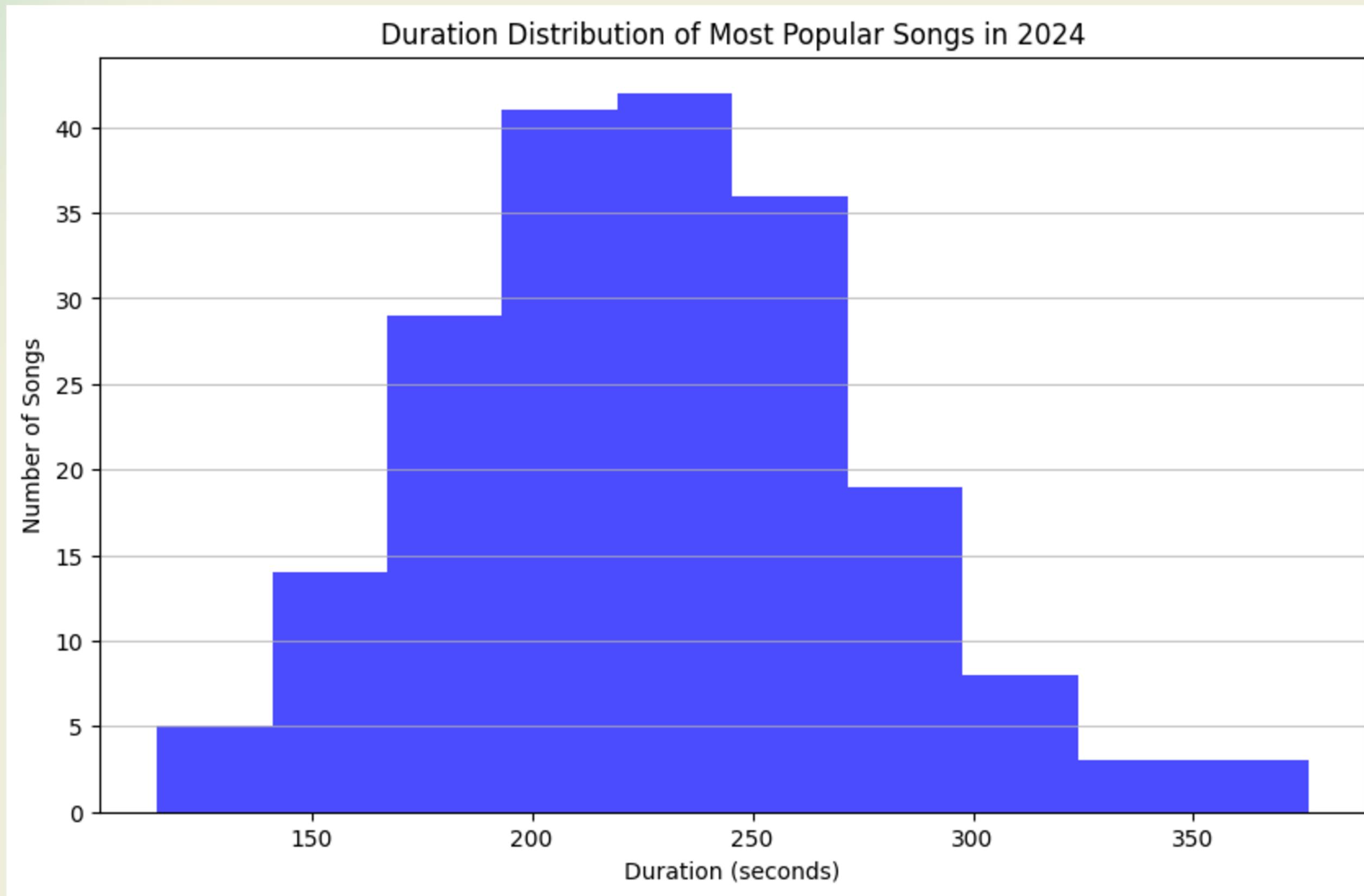
# TOP 20 NGHỆ SĨ CÓ CÁC BÀI HÁT BỨT PHÁ



# SỐ LƯỢNG BÀI HÁT PHÁT HÀNH THEO THÁNG



# THỜI LƯỢNG VÀ LƯỢT NGHE CỦA CÁC BÀI HÁT PHỔ BIẾN NHẤT NĂM 2024

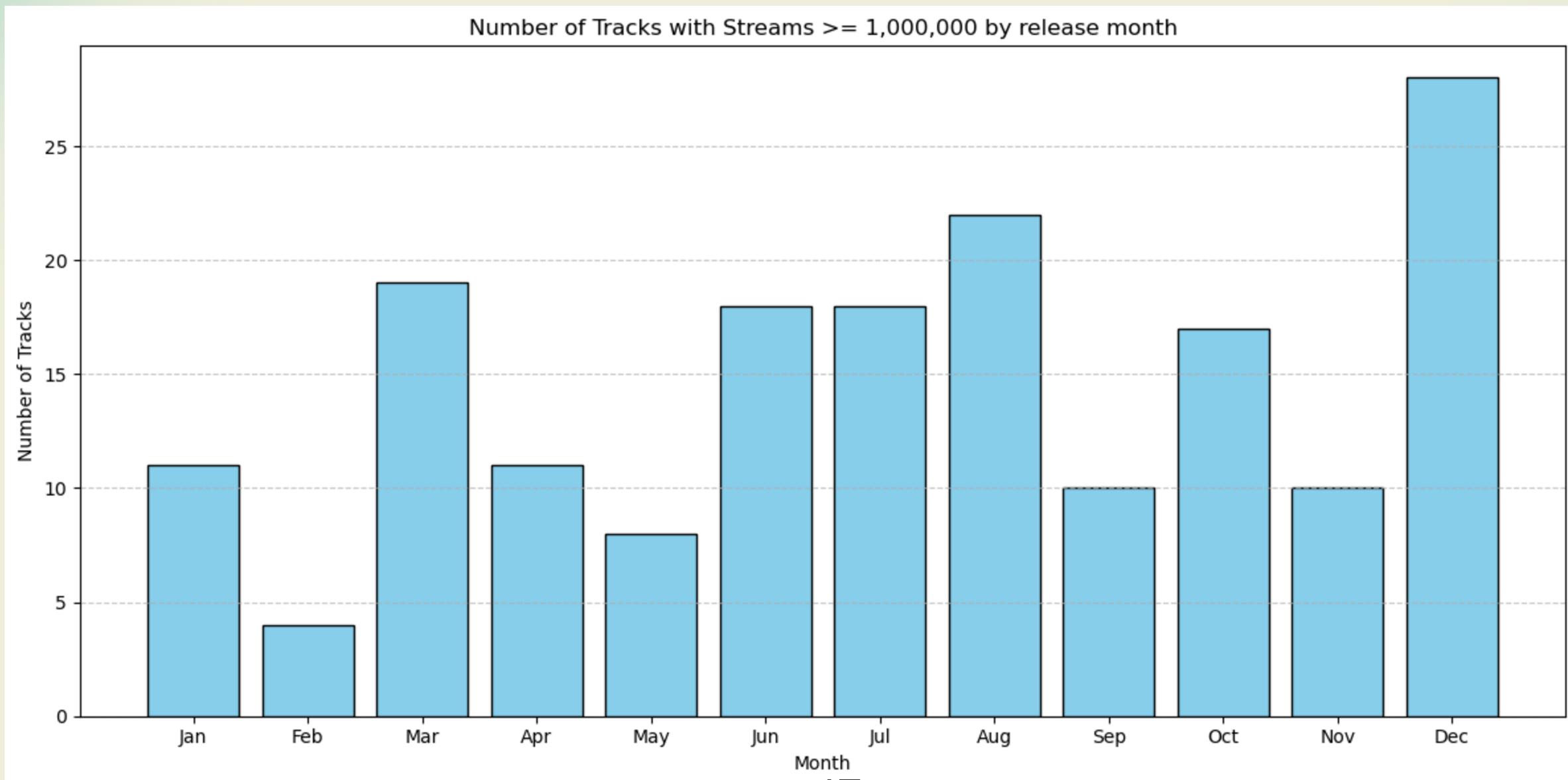


# CÂU HỎI

01

## Chiến lược tối ưu cho ra mắt sản phẩm trên Spotify năm 2024?

**Mục tiêu:** Xác định các tháng có lượng phát hành nhạc nhiều nhất trong năm 2024 để đưa ra chiến lược phát hành sản phẩm tối ưu hóa lượt stream trên Spotify.



# NHẬN XÉT

01

**Số lượng bài hát mỗi tháng:** Cho thấy sự gia tăng đáng kể vào tháng 7, tháng 8 và tháng 12 so với các tháng khác

**Số lượng bài hát đạt trên 1.000.000 lượt stream mỗi tháng:** Cũng cho thấy sự tăng vọt đáng kể vào tháng 7, tháng 8 và tháng 12

**Số lượng bài hát được phát hành cao nhất** vào tháng 8, nhưng số lượng bài hát lọt vào top cao nhất vào tháng 12

**Tháng 3** có tỷ lệ bài hát lọt vào top tương đối cao (50%), mặc dù đây không phải là tháng có số lượng phát hành cao nhất

**Tháng 2** có số lượng bài hát phát hành ít nhất và tỷ lệ bài hát lọt vào top thấp nhất (<20%)

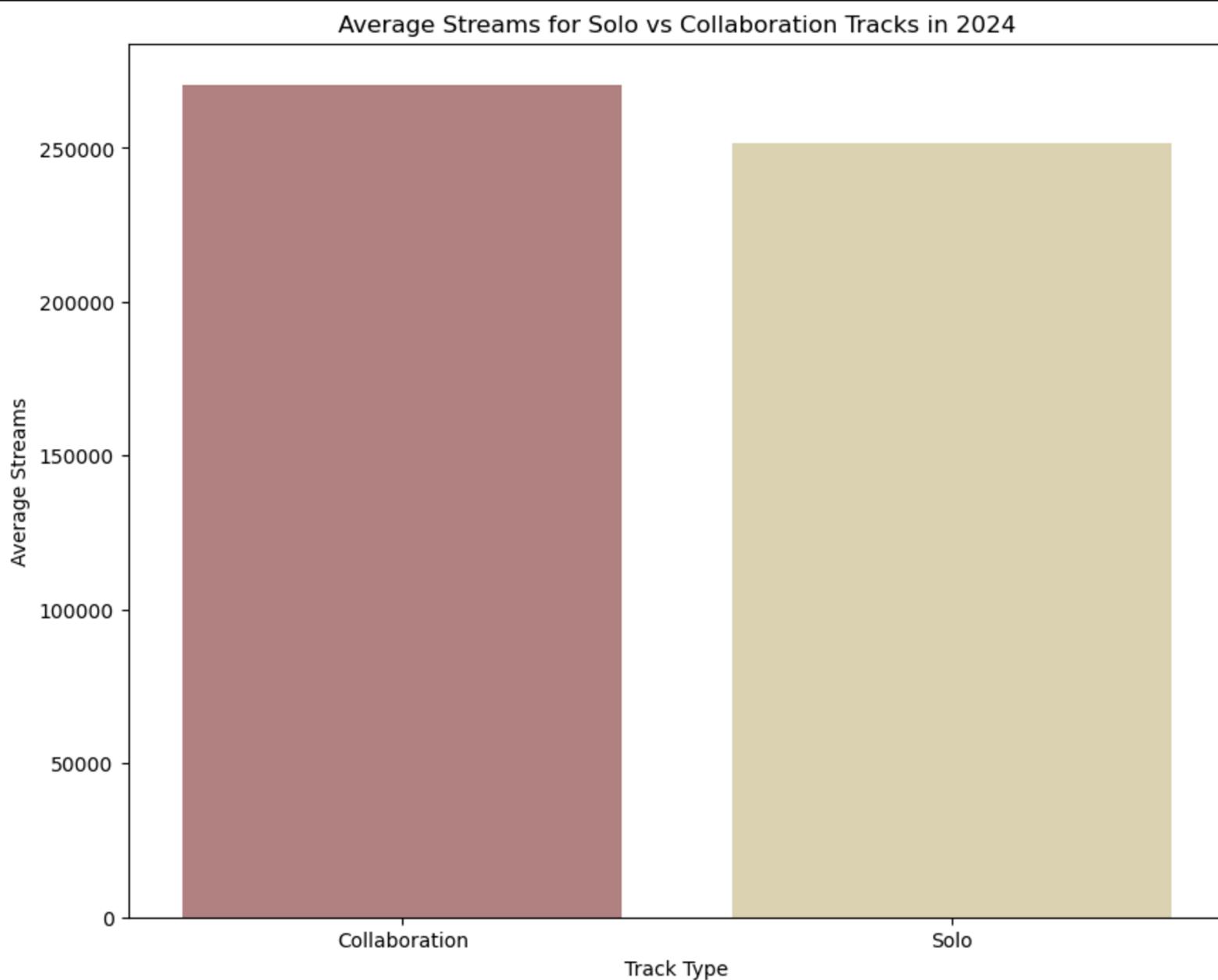
**Chiến lược phát hành sản phẩm tiềm năng:** Các nghệ sĩ có xu hướng tuân theo xu hướng thị trường, tập trung phát hành nhạc trong các tháng cao điểm như tháng 8. Tuy nhiên, tháng 12 dường như là thời điểm hiệu quả nhất để đạt được thứ hạng cao, mặc dù đây không phải là tháng có số lượng bài hát phát hành cao nhất.

# CÂU HỎI

02

**Nên chọn phát hành sản phẩm solo hay hợp tác vào năm 2024?**

**Mục tiêu:** So sánh lượt stream giữa các sản phẩm solo và hợp tác để xác định chiến lược hiệu quả hơn cho nghệ sĩ trong năm 2024.



# NHẬN XÉT

02

**Bài hát hợp tác** có xu hướng thu hút nhiều người nghe hơn do tiếp cận được lượng fan đông đảo hơn từ nhiều nghệ sĩ, được quảng bá mạnh mẽ hơn nhờ sự đóng góp của nhiều bên, và thường kết hợp nhiều phong cách âm nhạc, thu hút nhiều đối tượng khán giả.

**Tuy nhiên**, việc phát hành bài hát solo vẫn rất quan trọng để xây dựng và củng cố bản sắc riêng nên nghệ sĩ vẫn khá quan tâm đến việc ra mắt các sản phẩm solo. Bằng chứng là tỉ lệ các sản phẩm hợp tác và solo không chênh nhau quá nhiều.

# CÂU HỎI

03

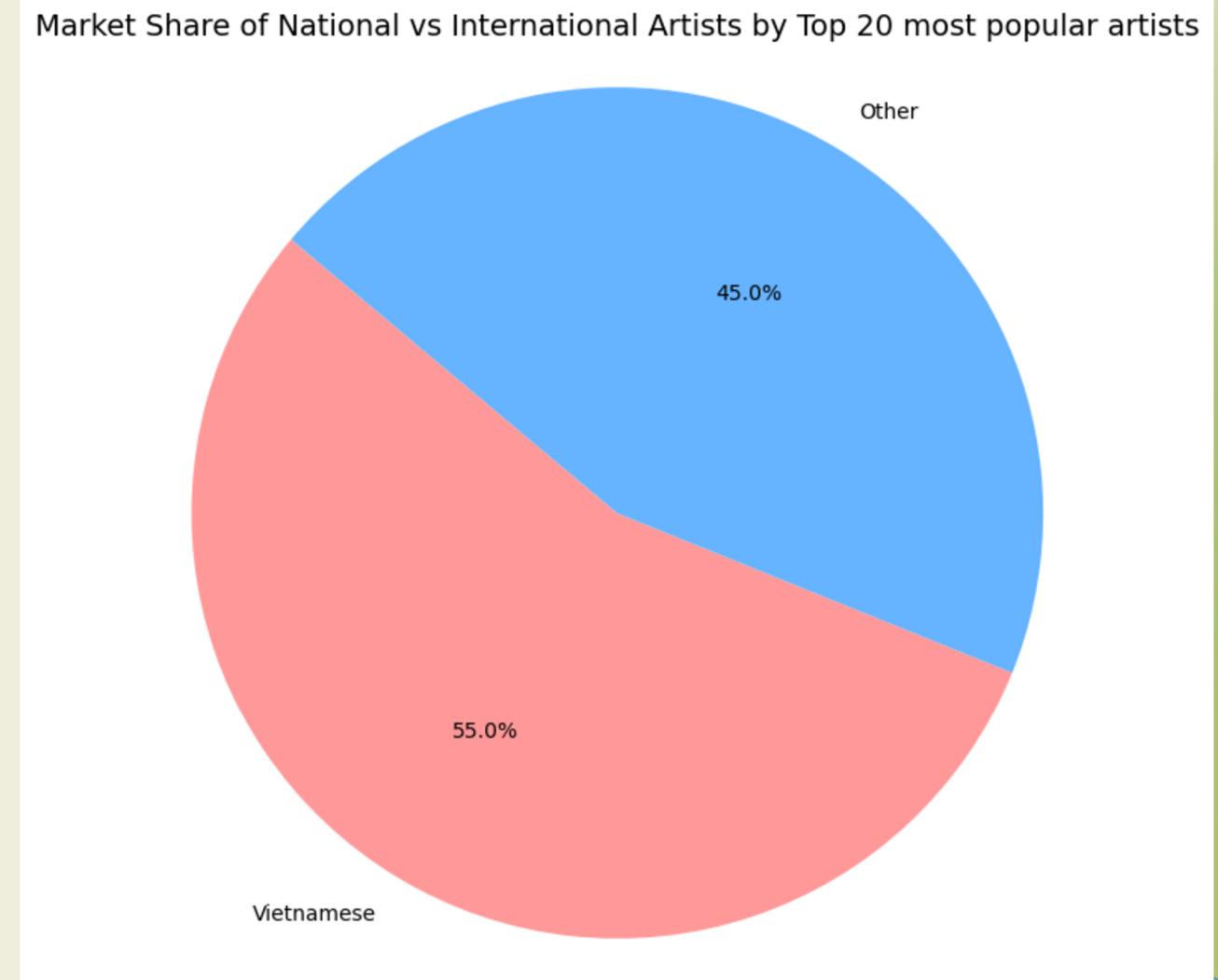
artist_names	count
tlinh	105
HIEUTHUHAI	93
GREY D	89
RPT MCK	84
W/N	70
Vũ.	66
ANH TRAI "SAY HI"	65
Sơn Tùng M-TP	63
Wren Evans	62
Ronboogz	61
Low G	60
Obito	58
Da LAB	57
Đen	53
itsnk	47
Thịnh Suy	44
RHYDER	42
JustaTee	40
Shiki	39
Andiez	39

Most Songs Artist: tlinh  
Most Songs Count: count  
nationality Vietnamese  
Name: tlinh, dtype: object  
Most Songs Nationality: Vietnamese  
Market Share Insights: nationality  
Vietnamese 11  
Other 9

105

**Thị phần của nghệ sĩ dựa trên sở thích khán giả năm 2024: nghệ sĩ Việt Nam hay quốc tế?**

**Mục tiêu:** Đưa ra nhận xét về sở thích của khán giả đối với nghệ sĩ Việt Nam hay quốc tế trong năm 2024 và thị trường âm nhạc Spotify thị phần hiện nay thế nào.



The song with the largest number of consecutive weeks in the top 10 is 'tiny love' by Thịnh Suy with 1 consecutive weeks.

# NHẬN XÉT

03

**Nghệ sĩ có nhiều bài hát nhất** trên bảng xếp hạng năm 2024 là tlinh, với tổng cộng 105 bài hát. Điều này cho thấy sự hiện diện và mức độ phổ biến đáng kể của tlinh trên bảng xếp hạng.

Đồng thời tlinh cũng là Nghệ sĩ Việt Nam

**Trong top 20**, nghệ sĩ Việt chiếm ưu thế (11/20), cho thấy khán giả năm 2024 ưu ái nghệ sĩ trong nước.

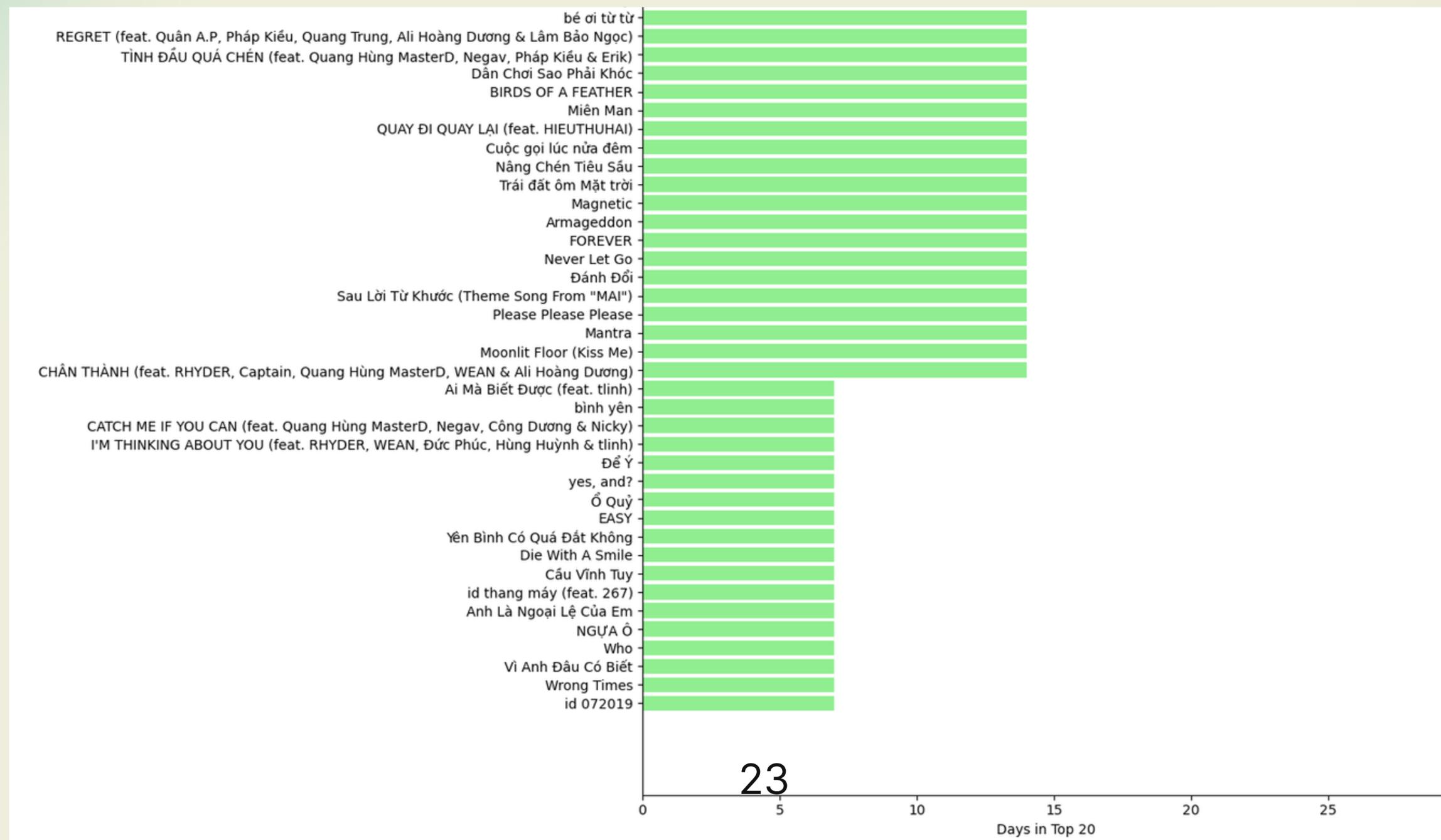
**Bài hát trụ top 20 lâu nhất** là tiny love của Thịnh Suy (1 tuần).

# CÂU HỎI

04

## Xu hướng trụ hạng trong top 20 bảng xếp hạng?

**Mục tiêu:** Xác định xu hướng loại bỏ bài hát khỏi top 20 bảng xếp hạng. Từ đó, đưa ra nhận xét về xu hướng thay đổi vị trí bài hát trên bảng xếp hạng trong năm 2024.



# NHẬN XÉT

04

**Hầu hết** các bài hát có xu hướng ở lại top 20 không quá một tháng.

**Các bài hát Việt Nam** (Bạn đời, QUAY ĐI QUAY LẠI,... - gần 25 ngày) có xu hướng ở lại top 20 lâu hơn so với các bài hát quốc tế (yes, and?, EASY - khoảng 7 ngày).

**Số lượng bài hát trong từng nhóm thời lượng** phân bố khá đồng đều, cho thấy tỷ lệ thay đổi bài hát giữa các nhóm là cân bằng.

**Tuy nhiên**, một số bài hát Việt Nam có thời gian trụ lại khá ngắn (Cầu Vĩnh Tuy, Để Ý - khoảng 7 ngày), trong khi một số bài hát quốc tế vẫn nằm trong top 20 trong một khoảng thời gian dài đáng kể (SPOT! trong 20-30 ngày).

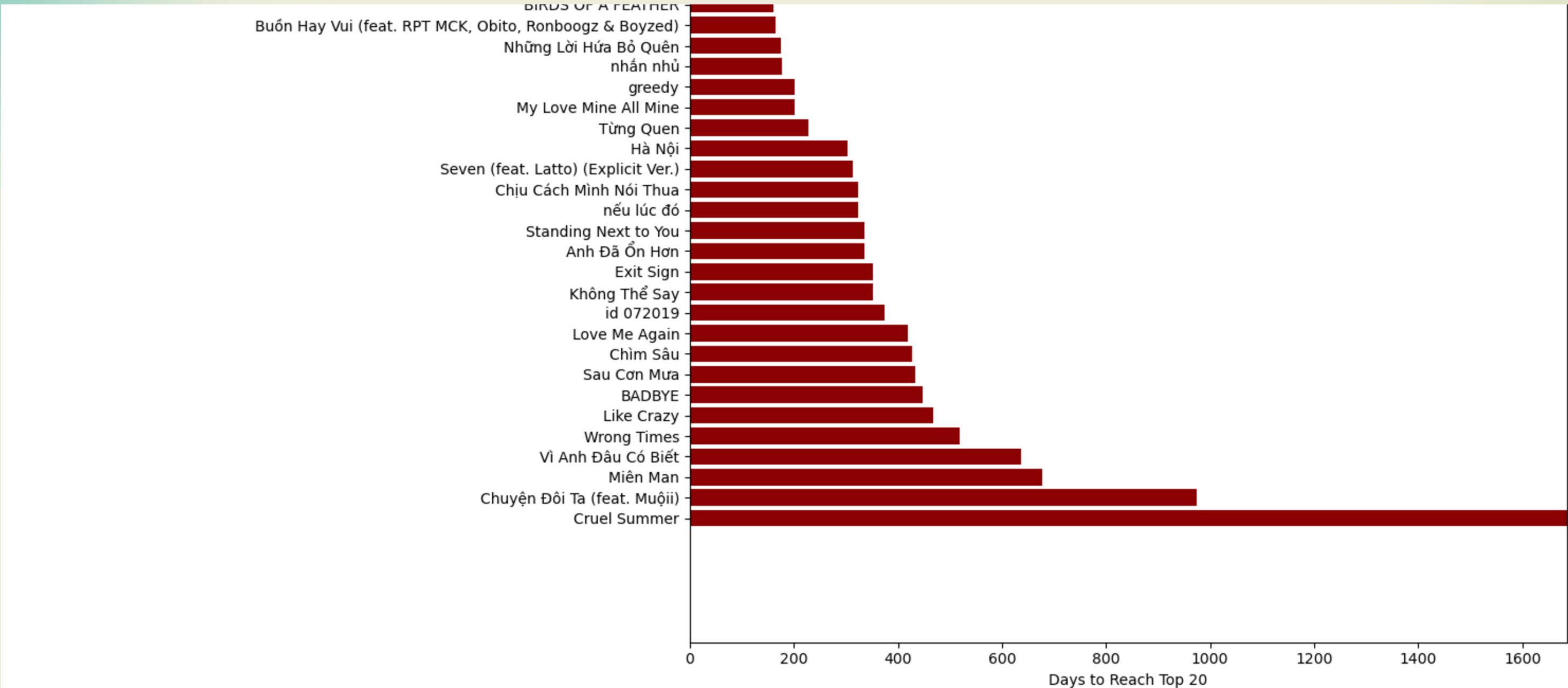
**Kết luận:** Mặc dù có một vài ngoại lệ không đáng kể, xu hướng cho thấy nhạc Việt nói chung ở lại top 20 lâu hơn, với sự phân bố thời gian giữ hạng khá đồng đều giữa các nhóm bài hát.

# CÂU HỎI

05

Tình hình để một bài hát từ top 200 vào đến top 20?

**Mục tiêu:** Có thể đưa ra thời gian cần thiết để một bài hát leo lên top 20 bảng xếp hạng trong năm 2024.



# NHẬN XÉT

05

**Thời gian để các bài hát đạt đến top 20** rất khác nhau, từ dưới 20 ngày đến hơn 1000 ngày.

**Hầu hết các bài hát Việt Nam** (Nâng chén tiêu sầu, REGRET, bình yên - khoảng 10 ngày) mất ít thời gian hơn để đạt đến top 20 so với các bài hát quốc tế (Cruel Summer - hơn 1600 ngày).

**Phân bố số lượng bài hát theo các nhóm thời gian** tăng dần đều, cho thấy hầu hết các bài hát mất ít thời gian để leo lên top 20, với một số ít mất thời gian lâu hơn đáng kể.

**Tuy nhiên**, một số bài hát Việt Nam đòi hỏi nhiều thời gian hơn để thăng hạng (Chuyện đôi ta - gần 1000 ngày), trong khi một vài bài hát quốc tế cần ít thời gian hơn (Taste, APT. - khoảng 20 ngày).

**Kết luận:** Thời gian để một bài hát đạt đến top 20 phần lớn phụ thuộc vào nguồn gốc của bài hát, và cũng bị ảnh hưởng bởi thời điểm phát hành và các yếu tố bên ngoài khác.

# NỘI DUNG

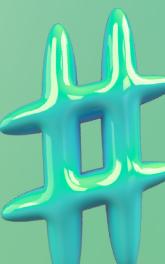


THU THẬP DỮ LIỆU

KHÁM PHÁ DỮ LIỆU

TRỰC QUAN HÓA & CÂU HỎI

MÔ HÌNH HÓA DỮ LIỆU



# TIỀN XỬ LÍ

## Làm sạch dữ liệu:

- Không có dữ liệu thiếu và dữ liệu trùng lặp

## Feature Engineering và Feature Selection:

- Chuyển đổi kiểu dữ liệu về đúng kiểu dữ liệu phù hợp.
- Thu thập thêm và nội suy ra một số trường mới:
  - Tách **release\_date** thành **release\_month** và **release\_year**
  - Tách trending\_date thành **trending\_month** và **trending\_year**
  - duration**
  - Tính toán **days\_to\_trend**
- Bỏ các trường không cần thiết: 'uri', 'release\_date', 'trending\_date', 'rank\_history', 'artist\_names', 'track\_name'

Data columns (total 12 columns):			
#	Column	Non-Null Count	Dtype
0	rank	2000 non-null	object
1	peak_rank	2000 non-null	object
2	previous_rank	2000 non-null	object
3	weeks_on_chart	2000 non-null	int64
4	streams	2000 non-null	int64
5	current_month	2000 non-null	object
6	duration	2000 non-null	float64
7	release_year	2000 non-null	int64
8	release_month	2000 non-null	int64
9	trending_year	2000 non-null	int64
10	trending_month	2000 non-null	int64
11	days_to_trend	2000 non-null	int64

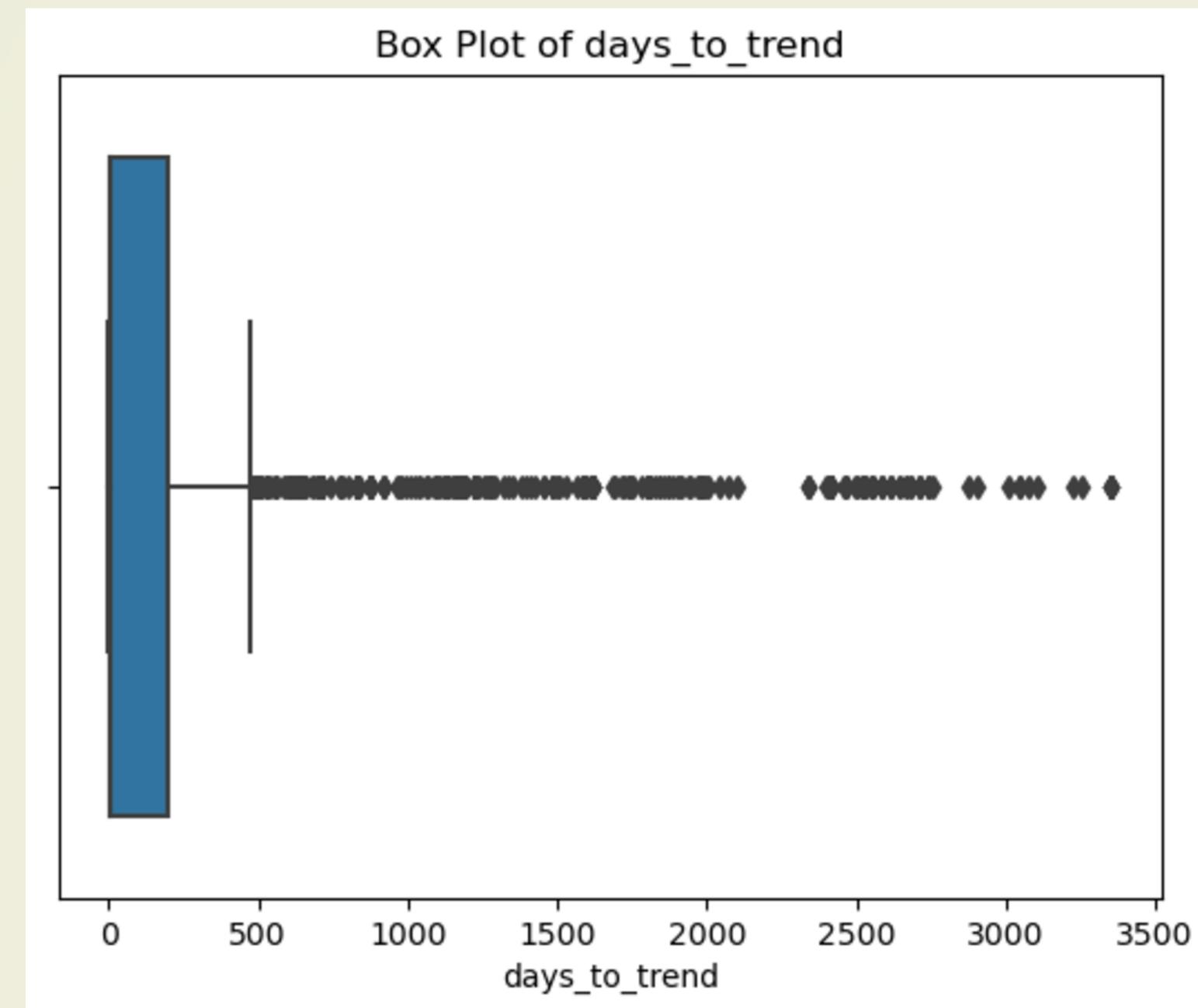
dtypes: float64(1), int64(7), object(4)

	rank	peak_rank	previous_rank	weeks_on_chart	streams	current_month	duration	release_year	release_month	trending_year	trending_month	days_to_trend
1675	76	34	87	211	241594	9	178.0	2023	7	2023	7	6
495	96	11	97	191	189214	3	235.0	2020	5	2021	7	423
1744	145	27	133	170	167555	9	237.0	2021	6	2021	7	10
383	184	26	189	177	118577	2	203.0	2020	8	2020	9	7
583	184	65	155	63	127911	3	176.0	2021	12	2021	12	20

# TIỀN XỬ LÍ

## Khảo sát giá trị ngoại lai:

- Nhiều giá trị ngoại lai đặc biệt ở cột **days\_to\_trend**
- **Xử lí giá trị ngoại lai** bằng IQR



# TIỀN XỬ LÍ

## Chuẩn hóa dữ liệu:

- Z-score để chuẩn hóa dữ liệu vì chuẩn hóa Z-score không nhạy cảm với các giá trị ngoại lệ

	weeks_on_chart	streams	duration	release_year	release_month	trending_year	trending_month	days_to_trend	rank	peak_rank	previous_rank	current_month
1993	-0.476301	-0.448737	0.218774	1.003777	-1.773145	0.889855	-1.525484	-0.405501	194	56	196	10
155	-0.911352	-0.480702	-0.179684	0.053680	1.103702	0.235790	0.861582	0.085031	156	156	159	1
321	1.800466	-0.440266	0.764033	-1.846513	-1.197775	-3.034532	1.159965	-0.088975	122	35	147	2
1742	2.699572	-0.364794	0.638204	-1.371465	0.816018	-2.380467	0.861582	-0.435330	143	1	151	9
86	-0.505305	-0.225457	-1.396029	-0.421368	-0.622406	0.235790	-0.630335	0.724711	87	65	85	1

## Phân chia dữ liệu:

- Chia dữ liệu thành các tập huấn luyện (train), kiểm tra (test) theo tỷ lệ 70% - 30%.
- Sử dụng train\_test\_split từ scikit-learn để chia dữ liệu.

Training set shape: (1400, 11)  
Testing set shape: (600, 11)

# MÔ HÌNH HÓA DỮ LIỆU

01

## Lựa chọn mô hình:

- Sử dụng các mô hình hồi quy như **LinearRegressor**, **RandomForestRegressor**, **BaggingRegressor**, hoặc **XGBoost** để dự đoán lượt phát.
- Sử dụng **RandomizedSearchCV** hoặc **GridSearchCV** để tìm kiếm các siêu tham số tốt nhất cho mô hình.



02

## Đánh giá mô hình:

- Sử dụng các độ đo như **Mean Absolute Error (MAE)**, **Mean Squared Error (MSE)**, **RMSE**, và **R-squared** để đánh giá hiệu suất của mô hình.
- So sánh với các mô hình cơ sở như **Linear Regression** hoặc **Decision Tree Regressor**.

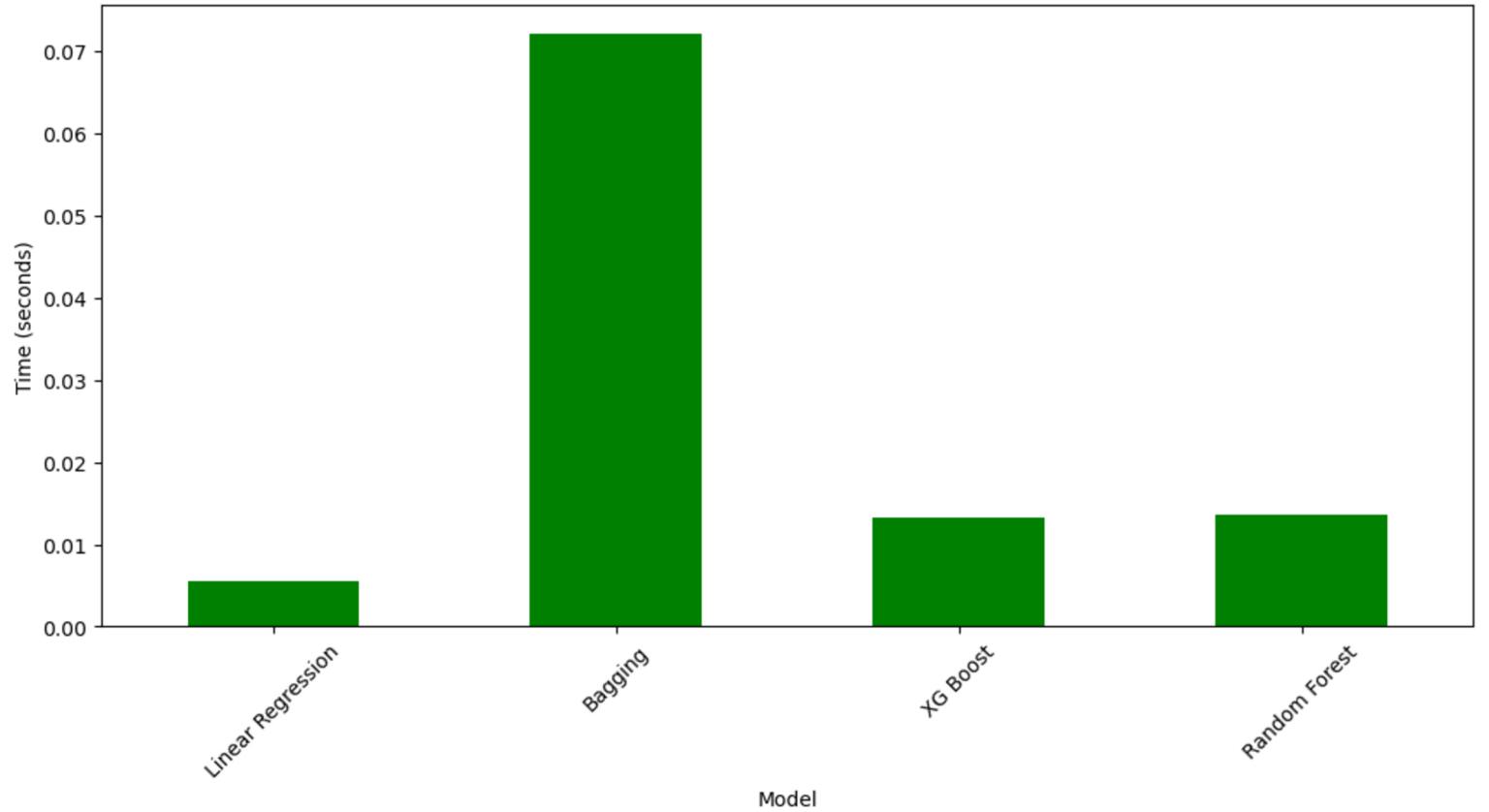


# ĐÁNH GIÁ MÔ HÌNH

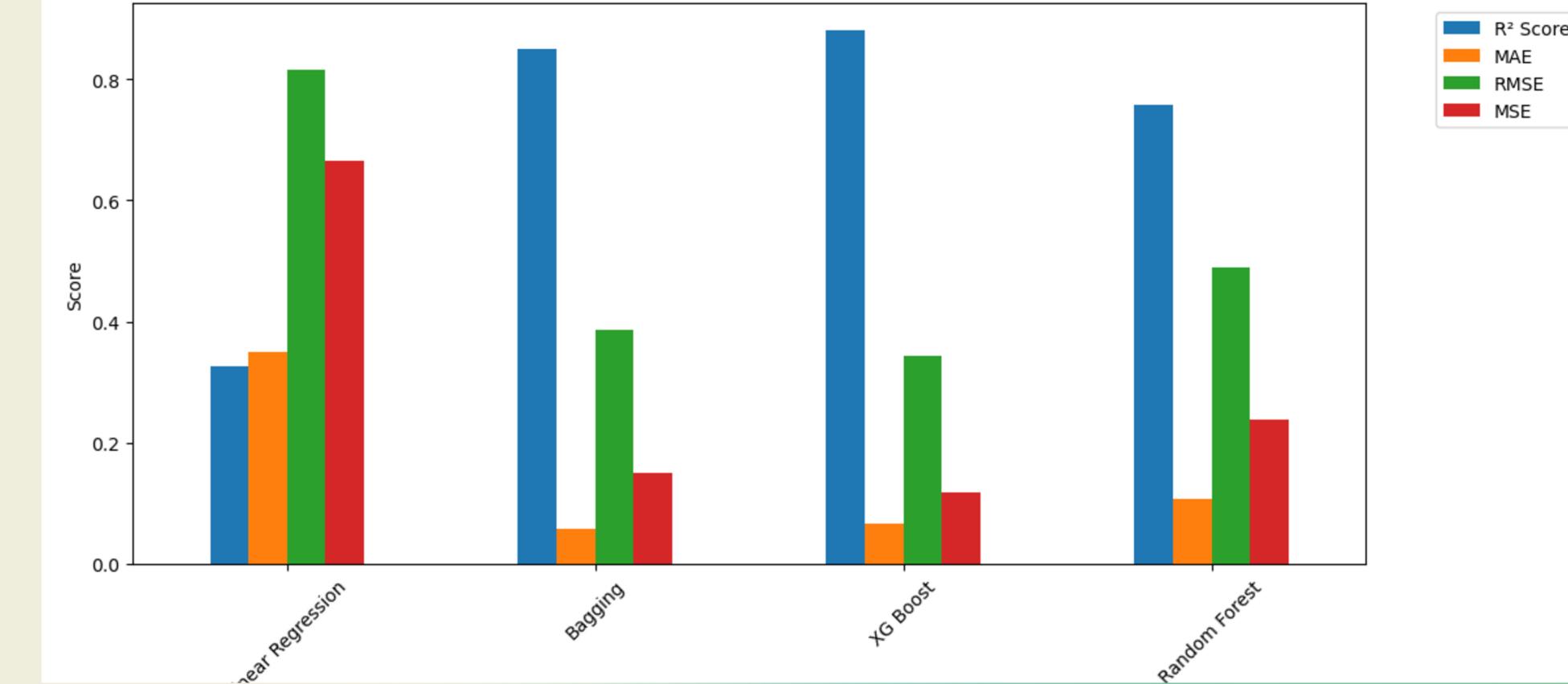
## Detailed Performance Metrics:

Model	R <sup>2</sup> Score	MAE	RMSE	MSE	Execution Time (s)
Linear Regression	0.3255	0.3494	0.8160	0.6659	0.0055
Bagging	0.8489	0.0574	0.3863	0.1492	0.0720
XG Boost	0.8802	0.0659	0.3439	0.1182	0.0133
Random Forest	0.7583	0.1070	0.4885	0.2386	0.0136

Model Execution Time Comparison



Model Performance Metrics Comparison



# MÔ HÌNH HÓA DỮ LIỆU

- **Kết quả mô hình:**
  - Về độ chính xác: XG Boost cho kết quả tốt nhất (0.88), sau đó Bagging (0.86), sau đó là Random Forest và Linear Regressor.
  - Về thời gian: XG Boost nhanh nhất, sau đó là Linear Regressor. Bagging có thời gian khá và lâu nhất là Random Forest.
- **Ý nghĩa của kết quả:**
  - **Kiến nghị sử dụng:** XG Boost và Bagging.
  - Độ chính xác cao nhất và thời gian nhanh.
- **Ứng dụng:**
  - Nghệ sĩ/Nhà sản xuất: Dự đoán xu hướng phát triển của bài hát, lên kế hoạch marketing và đánh giá hiệu quả của marketing.
  - Xác định bài hát tiềm năng để hợp tác, dự đoán phạm vi tiếp cận đối với công chúng.



FOR YOUR FOLLOWING