

HO CHI MINH UNIVERSITY OF SCIENCE

**FACULTY OF INFORMATION
TECHNOLOGY**

BÁO CÁO ĐỒ ÁN CUỐI KỲ

Subject: Data Visualization

Class: 22KHDL

Group: 5

Students:

Dương Nhật Huy – 22127148

Nguyễn Đình Kiên - 22127216

Trương Thuận Kiệt – 22127224

Phạm Minh Mẫn - 22127257

Hồ Đăng Phúc - 22127492

TP. Hồ Chí Minh, tháng 4, năm 2025

Contents

1. Thông tin nhóm và phân công công việc	4
2. Giới thiệu tập dữ liệu	5
2.1. Thông tin cơ bản	5
2.1.1. Vietnamese_only_songs_cleaned	5
2.1.2. position_streams_by_time_daily_total_peak_cleaned and position_streams_by_time_weekly_total_peak_cleaned	5
2.1.3. Vietnamese_artists	6
2.2. Mô tả chi tiết	6
2.2.1. Vietnamese_only_songs_cleaned	6
2.2.2. position_streams_by_time_daily_total_peak_cleaned and position_streams_by_time_weekly_total_peak_cleaned	7
2.2.3. Vietnamese_artists	7
3. Khám phá và phân tích thống kê cho dữ liệu	8
3.1. Vietnamese_only_songs_cleaned	8
3.1.1. Dữ liệu thiếu	8
3.1.2. Phân tích thống kê	9
3.2. position_streams_by_time_daily_total_peak_cleaned and position_streams_by_time_weekly_total_peak_cleaned	10
3.2.1. Dữ liệu thiếu	10
3.2.2. Phân tích thống kê	10
3.3. Vietnamese_artists	13
3.3.1. Dữ liệu thiếu	13
3.3.2. Phân tích thống kê	13
4. Môi trường quan giữa các biến	14
4.1. Vietnamese_only_songs_cleaned	14
4.2. position_streams_by_time_daily_total_peak_cleaned and position_streams_by_time_weekly_total_peak_cleaned	16

4.3.	Vietnamese_artitsts	17
5.	Tương quan stat model.....	17
5.1.	Giới thiệu	17
5.2.	Phương pháp	18
5.3.	Stream_counts vs các biến còn lại	18
5.4.	Popularity vs các biến còn lại	20
6.	Trực quan hóa dữ liệu	21
6.1.	Tích hợp AI trong việc thiết kế dashboard	21
6.2.	Dashboard tổng quan.....	25
6.2.1.	Tổng thể dashboard	25
6.2.2.	Các Chỉ số Chính (KPIs):.....	25
6.2.3.	Xu hướng theo Thời gian:	25
6.2.4.	Phân bố:	26
6.2.5.	Bộ lọc:	26
6.2.6.	Kết luận:.....	26
6.3.	Songs dashboard	26
6.3.1.	Tổng thể dashboard	26
6.3.2.	Các nhận xét biểu đồ tổng quát:.....	27
6.3.3.	Các chi tiết về một số bài hát nổi bật (lấy các bài Dù cho tận thế, Chuyện đời ta).	27
6.3.4.	Bộ lọc:	28
6.3.5.	Kết luận:.....	28
6.4.	Artists dashboard	29
6.4.1.	Tổng thể dashboard	29
6.4.2.	So sánh tổng quát:.....	29
6.4.3.	Các phân tích chi tiết (xét các ca sĩ nổi bật như Sơn Tùng, HIEUTHUHAI):	30
6.4.4.	Bộ lọc:	31
6.4.5.	Kết luận:.....	31
6.5.	Dashboard thể loại.....	32
6.5.1.	Tổng thể dashboard	32

6.5.2.	Các Chỉ số Chính (KPIs):.....	32
6.5.3.	Phân tích chi tiết theo Thể loại:.....	32
6.5.4.	Bộ lọc:	33
6.5.5.	Kết luận:.....	33
7.	Tích hợp AI trong việc phân tích	33
7.1.	Hỏi đáp	33
7.2.	Chatbot.....	33
8.	Tích hợp AI trong việc đánh giá dashboard	34
8.1.	Dashboard Tổng quan	34
8.2.	Dashboard Bài hát	35
8.3.	Dashboard Thể loại	35
8.4.	Dashboard Nghệ sĩ.....	35
8.5.	Đánh giá tổng quát	36

1. Thông tin nhóm và phân công công việc

MSSV	Họ và Tên	Công việc	Mức độ hoàn thành
22127148	Dương Nhật Huy	<ul style="list-style-type: none"> Trực quan hóa Summary Dashboard Thiết kế, trang trí các dashboard Tích hợp AI 	100%
22127216	Nguyễn Đình Kiên	<ul style="list-style-type: none"> Trực quan hóa Genres Dashboard Viết báo cáo Làm powerpoint 	100%
22127224	Trương Thuận Kiệt	<ul style="list-style-type: none"> Thu thập dữ liệu Tiền xử lý dữ liệu 	100%

		<ul style="list-style-type: none"> • Trục quan hóa Songs Dashboard • Viết báo cáo 	
22127257	Phạm Minh Mẫn	<ul style="list-style-type: none"> • Thu thập dữ liệu • Trục quan hóa Artists Dashboard • Tích hợp AI 	100%
22127492	Hồ Đăng Phúc	<ul style="list-style-type: none"> • Trục quan hóa Genres Dashboard • Viết báo cáo • Làm powerpoint 	100%

2. Giới thiệu tập dữ liệu

2.1. Thông tin cơ bản

2.1.1. Vietnamese_only_songs_cleaned

- Nguồn gốc: Thu thập từ [Spotify API](#) dành cho developers
- Chủ đề: Thông tin chi tiết về các bài hát tiếng việt và các bài hát ở thị trường Việt Nam



2.1.2. position_streams_by_time_daily_total_peak_cleaned and position_streams_by_time_weekly_total_peak_cleaned

- Nguồn gốc: Thu thập từ [Kworb](#), một trang web về xếp hạng cũng như thông tin về 1 bài hát hoặc 1 nghệ sĩ.
- Chủ đề: Xếp hạng và số lượng phát cao nhất trong ngày và tuần.

TOP 50	NEW MUSIC	ARTISTS	CHARTS	RETRY LOG
RADIO	SPOTIFY	YOUTUBE	TRENDING	HOME



Spotify Daily Chart - Vietnam - 2025/04/05 | Totals

Pos	#	Artist and Title	Days	Pop	Streams	Stream+	Play	Play+	Total
1	=	Sống Rồi - Sonam (feat. Latou)	332	100%	342,594	5,568	1,911,310	10,313	213,581,928
2	=	Sống Rồi - 33 (feat. Jack Harlow)	556	100%	113,789	4,733	1,881,589	7,017	58,679,165
3	+1	Phải - 33 (feat. 33)	15	100%	158,851	+1,399	1,813,707	40,962	1,441,543
4	+1	Phải - Hoàng Thùy Linh (feat. 33)	305	100%	138,813	+11,267	885,918	+11,801	31,812,867
5	-2	HẾT (feat. 33) - 33 (feat. 33)	5	100%	113,806	-23,788	886,810	-115,869	396,451
6	=	SOON - 33 (feat. 33)	24	4	122,988	+8,972	811,797	+10,212	1,695,358
7	+1	Sống Rồi - 33 (feat. 33)	178	100%	118,871	+5,549	788,818	31,859	37,247,138
8	-1	Kiêu - 33 (feat. 33)	25	7	107,488	-1,138	816,700	-30,449	1,358,939
9	=	SOON - 33 (feat. 33)	178	100%	106,387	+2,387	815,444	-34,845	33,859,943
10	+2	HẾT (feat. 33) - 33 (feat. 33)	306	4	102,498	+3,777	800,016	+9,881	26,450,033
11	=	Sống Rồi - 33 (feat. 33)	152	11	98,631	-288	800,016	-11,719	13,886,787
12	+1	Phải - 33 (feat. 33)	35	100%	94,826	+12,141	817,008	-10,201	3,585,211
13	+1	Phải - 33 (feat. 33)	35	100%	91,911	+8,287	802,827	-12,318	10,100,709

2.1.3. Vietnamese_artists

- Nguồn gốc: Thu thập từ Spotify API dành cho developers
- Chủ đề: Thông tin về các nghệ sĩ Việt Nam như tên, độ phổ biến, ...

2.2. Mô tả chi tiết

2.2.1. Vietnamese_only_songs_cleaned

- Mỗi dòng diễn tả thông tin về các bài hát bao gồm: tên, thể loại và thuộc album nào.
- Ý nghĩa của các cột

Cột	Ý nghĩa
Id	Mã Spotify của mỗi bài hát.
Album	Tên album mà bài hát xuất hiện.
Album_type	Loại album. Giá trị cho phép: "album", "single", "compilation".
Album_popularity	Độ phổ biến của album. Giá trị nằm trong khoảng từ 0 đến 100, với 100 là phổ biến nhất.
Artists	Tên các nghệ sĩ biểu diễn bài hát
Disc_number	Số đĩa.
Duration_ms	Độ dài của bài hát.
Explicit	Bài hát có chứa lời bài hát tục tĩu hay không.
Name	Tên bài hát.
Popularity	Độ phổ biến của bài hát. Giá trị nằm trong khoảng từ 0 đến 100, với 100 là phổ biến nhất. Độ phổ biến của bài hát là một giá trị từ 0 đến 100, với 100 là phổ biến nhất. Độ phổ biến được tính bằng thuật

	<p>toán và chủ yếu dựa trên tổng số lần phát bài hát và thời gian gần đây của các lần phát đó.</p> <p>Nói chung, các bài hát đang được phát nhiều hiện tại sẽ có độ phổ biến cao hơn so với các bài hát được phát nhiều trong quá khứ. Các bài hát trùng lặp (ví dụ: cùng một bài hát từ một đĩa đơn và một album) được đánh giá độc lập. Độ phổ biến của nghệ sĩ và album được suy ra từ độ phổ biến của bài hát thông qua tính toán.</p> <p>Lưu ý: giá trị độ phổ biến có thể chậm hơn so với độ phổ biến thực tế vài ngày: giá trị này không được cập nhật theo thời gian thực.</p>
Track_number	Số thứ tự của bài hát trong album. Nếu album có nhiều đĩa, số thứ tự bài hát là số trên đĩa được chỉ định.
Release_date	Ngày album được phát hành lần đầu tiên.
Genres	Thể loại của bài hát.
Stream_counts	Tổng số lần bài hát được phát trên Spotify.

2.2.2. **position_streams_by_time_daily_total_peak_cleaned and position_streams_by_time_weekly_total_peak_cleaned**

- Mỗi dòng diễn tả thông tin về sự thể hiện của các bài hát trên thị trường Việt Nam.
- Ý nghĩa của các cột:

Cột	Ý nghĩa
Id	Mã Spotify của từng bài hát.
TotalStream	Tổng số lượt phát theo tuần/ngày.
PeakStream	Lượt phát cao nhất theo tuần/ngày.
PeakPosition	Vị trí cao nhất theo tuần/ngày.

2.2.3. **Vietnamese_artists**

- Mỗi dòng diễn tả thông tin về nghệ sĩ Việt Nam ở thị trường Việt Nam.
- Ý nghĩa của các cột:

Cột	Ý nghĩa
Id	Mã Spotify của từng nghệ sĩ.

Name	Tên của nghệ sĩ.
Followers	Số lượt theo dõi của nghệ sĩ.
Genres	Thể loại âm nhạc của nghệ sĩ.

3. Khám phá và phân tích thống kê cho dữ liệu

3.1.Vietnamese_only_songs_cleaned

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4846 entries, 0 to 4845
Data columns (total 14 columns):
#   Column                Non-Null Count  Dtype
---  -
0   album                 4846 non-null   object
1   album_type            4846 non-null   object
2   album_popularity      4846 non-null   int64
3   artists               4846 non-null   object
4   disc_number           4846 non-null   int64
5   duration_ms           4846 non-null   int64
6   explicit              4846 non-null   bool
7   id                    4846 non-null   object
8   name                  4846 non-null   object
9   popularity            4846 non-null   int64
10  track_number          4846 non-null   int64
11  release_date          4846 non-null   object
12  genres                4846 non-null   object
13  stream_counts         4817 non-null   float64
dtypes: bool(1), float64(1), int64(5), object(7)
memory usage: 497.0+ KB
```

- Bộ dữ liệu chứa tổng cộng 4846 dòng và 14 cột.
- Trong đó có cột stream_counts có mất mát.
- Chuyển cột release_date về datetime.

3.1.1. Dữ liệu thiếu

Cột	Số lượng thiếu	Tỷ lệ (%)
Stream_counts	29	0.598

- Ở đây, ta có thể thấy được các giá trị thiếu là vô cùng ít, đồng thời ở đây ta cũng biết được các giá trị thiếu là do khi thu thập một số bài hát có **số lượng nghe < 1000** nên sẽ không có.

- Và để giữ sự đa dạng của các thuộc tính nên nhóm sẽ giữ lại các dòng này.

3.1.2. Phân tích thống kê

	album_popularity	disc_number	duration_ms	popularity	track_number	stream_counts
count	4846.000000	4846.000000	4.846000e+03	4846.000000	4846.000000	4.817000e+03
mean	25.919934	1.003714	2.381399e+05	27.985349	3.120718	4.121705e+06
std	15.573992	0.060839	6.362212e+04	14.618266	3.771611	1.772154e+07
min	0.000000	1.000000	1.952800e+04	0.000000	1.000000	1.027000e+03
25%	14.000000	1.000000	1.990928e+05	18.000000	1.000000	1.545430e+05
50%	25.000000	1.000000	2.328505e+05	27.000000	1.000000	6.609920e+05
75%	37.000000	1.000000	2.725415e+05	38.000000	4.000000	2.770964e+06
max	83.000000	2.000000	1.813776e+06	80.000000	41.000000	8.532404e+08

- **album_popularity:**
 - Trung bình: 25.92, với độ lệch chuẩn là 15.57.
 - Giá trị dao động từ 0 đến 83.
 - Phần lớn các album có độ phổ biến thấp (25% dưới 14), nhưng có một số album rất phổ biến (max = 83).
- **disc_number:**
 - Hầu hết các bài hát nằm trên đĩa đầu tiên (median = 1, max = 2).
 - Điều này cho thấy dữ liệu chủ yếu liên quan đến các album đơn giản, không có nhiều đĩa.
- **duration_ms:**
 - Thời lượng trung bình của các bài hát là khoảng 238 giây (2.38e+05 ms).
 - Thời lượng dao động từ 19.5 giây (min = 1.95e+04 ms) đến khoảng 30 phút (max = 1.81e+06 ms).
 - Phần lớn các bài hát có thời lượng trong khoảng 199 giây đến 273 giây (25% đến 75%).
- **popularity:**
 - Trung bình: 27.98, với độ lệch chuẩn là 14.62.
 - Giá trị dao động từ 0 đến 80.
 - Phần lớn các bài hát có độ phổ biến thấp (25% dưới 18), nhưng có một số bài hát rất phổ biến (max = 80).
- **track_number:**

- Trung bình: 3.12, với độ lệch chuẩn là 3.77.
- Số thứ tự bài hát dao động từ 1 đến 41.
- Phần lớn các bài hát nằm ở vị trí đầu tiên trong album (median = 1).
- **stream_counts:**
 - Trung bình: 4.12 triệu lượt stream, nhưng độ lệch chuẩn rất lớn (17.72 triệu).
 - Giá trị dao động từ 1,027 đến hơn 853 triệu lượt stream.
 - Điều này cho thấy có một số bài hát rất phổ biến (outliers), trong khi phần lớn các bài hát có lượt stream thấp hơn (75% dưới 2.77 triệu).

3.2. position_streams_by_time_daily_total_peak_cleaned and position_streams_by_time_weekly_total_peak_cleaned

Weekly	Daily
<pre><class 'pandas.core.frame.DataFrame'> RangeIndex: 772 entries, 0 to 771 Data columns (total 4 columns): # Column Non-Null Count Dtype --- --- 0 id 772 non-null object 1 TotalStreamCounts 771 non-null float64 2 PeakStreamCounts 771 non-null float64 3 PeakPosition 771 non-null float64 dtypes: float64(3), object(1) memory usage: 24.3+ KB</pre>	<pre><class 'pandas.core.frame.DataFrame'> RangeIndex: 726 entries, 0 to 725 Data columns (total 4 columns): # Column Non-Null Count Dtype --- --- 0 id 726 non-null object 1 TotalStreamCounts 725 non-null float64 2 PeakStreamCounts 725 non-null float64 3 PeakPosition 725 non-null float64 dtypes: float64(3), object(1) memory usage: 22.8+ KB</pre>

- Đối với bộ dữ liệu hằng tuần có **772 dòng** và hằng ngày là **726 dòng**, cả hai đều có **4 cột**.
- Ở các cột như vn_position và vn_stream_counts thiếu dữ liệu.

3.2.1. Dữ liệu thiếu

Cột	Số lượng thiếu	Tỷ lệ (%)
Vn_position (Weekly)	1	0.129
Vn_stream_count (Weekly)	1	0.129
Vn_position (Daily)	1	0.138
Vn_stream_count (Daily)	1	0.138

- Các dòng thiếu dữ liệu là vô cùng ít vì thế nên nhóm vẫn sẽ giữ lại các dòng này.

3.2.2. Phân tích thống kê

- **Weekly**

	TotalStreamCounts	PeakStreamCounts	PeakPosition
count	7.710000e+02	7.710000e+02	771.000000
mean	5.851733e+06	2.680038e+05	66.584955
std	9.287612e+06	3.178791e+05	58.811201
min	4.698000e+03	4.698000e+03	1.000000
25%	2.867365e+05	9.259600e+04	12.000000
50%	1.499057e+06	1.797870e+05	51.000000
75%	6.552288e+06	3.087205e+05	111.000000
max	5.269765e+07	3.445288e+06	200.000000

- **TotalStreamCounts:**
 - Trung bình: 5.85 triệu lượt stream mỗi tuần.
 - Độ lệch chuẩn: 9.29 triệu, cho thấy sự biến động lớn giữa các tuần.
 - Min/Max: Dao động từ 4,698 đến 52.69 triệu lượt stream.
 - Phân phối:
 - 25% số tuần có lượt stream dưới 286,736.
 - 50% số tuần (median) có lượt stream khoảng 1.49 triệu.
 - 75% số tuần có lượt stream dưới 6.55 triệu.
- **PeakStreamCounts:**
 - Trung bình: 268,003 lượt stream cao nhất trong tuần.
 - Độ lệch chuẩn: 317,879, cho thấy sự biến động lớn giữa các tuần.
 - Min/Max: Dao động từ 4,698 đến 3.44 triệu lượt stream cao nhất trong tuần.
 - Phân phối:
 - 25% số tuần có lượt stream cao nhất dưới 92,596.
 - 50% số tuần có lượt stream cao nhất khoảng 179,787.
 - 75% số tuần có lượt stream cao nhất dưới 308,720.
- **PeakPosition:**
 - Trung bình: Vị trí cao nhất trung bình là 66.58.
 - Độ lệch chuẩn: 58.81, cho thấy sự phân tán lớn.
 - Min/Max: Dao động từ vị trí 1 đến 200.
 - Phân phối:
 - 25% số tuần có vị trí cao nhất dưới 12.

- 50% số tuần có vị trí cao nhất khoảng 51.
- 75% số tuần có vị trí cao nhất dưới 111.

○ **Daily**

	TotalStreamCounts	PeakStreamCounts	PeakPosition
count	7.250000e+02	725.000000	725.000000
mean	5.583042e+06	43500.244138	50.513103
std	9.160756e+06	47062.807284	45.455326
min	1.704000e+03	1364.000000	1.000000
25%	3.157060e+05	16963.000000	9.000000
50%	1.464363e+06	31830.000000	39.000000
75%	6.013064e+06	50816.000000	83.000000
max	5.294350e+07	455604.000000	189.000000

○ **TotalStreamCounts:**

- Trung bình: 5.58 triệu lượt stream mỗi ngày.
- Độ lệch chuẩn: 9.16 triệu, cho thấy sự biến động lớn giữa các ngày.
- Min/Max: Dao động từ 1,704 đến 52.94 triệu lượt stream.
- Phân phối:
- 25% số ngày có lượt stream dưới 315,706.
- 50% số ngày (median) có lượt stream khoảng 1.46 triệu.
- 75% số ngày có lượt stream dưới 6.01 triệu.

○ **PeakStreamCounts:**

- Trung bình: 43,500 lượt stream cao nhất trong ngày.
- Độ lệch chuẩn: 47,062, cho thấy sự biến động lớn giữa các ngày.
- Min/Max: Dao động từ 1,364 đến 455,604 lượt stream cao nhất trong ngày.
- Phân phối:
- 25% số ngày có lượt stream cao nhất dưới 16,963.
- 50% số ngày có lượt stream cao nhất khoảng 31,830.
- 75% số ngày có lượt stream cao nhất dưới 50,816.

○ **PeakPosition:**

- Trung bình: Vị trí cao nhất trung bình là 50.51.
- Độ lệch chuẩn: 45.45, cho thấy sự phân tán lớn.
- Min/Max: Dao động từ vị trí 1 đến 189.

- Phân phối:
- 25% số ngày có vị trí cao nhất dưới 9.
- 50% số ngày có vị trí cao nhất khoảng 39.
- 75% số ngày có vị trí cao nhất dưới 83.

3.3. Vietnamese_artists

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1115 entries, 0 to 1114
Data columns (total 6 columns):
#   Column          Non-Null Count  Dtype
---  -
0   id              1115 non-null   object
1   name            1115 non-null   object
2   popularity      1115 non-null   int64
3   followers       1115 non-null   int64
4   genres          1115 non-null   object
5   type            1115 non-null   object
dtypes: int64(2), object(4)
memory usage: 52.4+ KB
```

- Tất cả các cột đều ở đúng kiểu dữ liệu và không có bất kỳ dữ liệu thiếu nào.

3.3.1. Dữ liệu thiếu

- Không có bất kỳ dữ liệu bị thiếu nào

3.3.2. Phân tích thống kê

	popularity	followers
count	1115.000000	1.115000e+03
mean	30.951570	8.933268e+04
std	14.979501	3.273770e+05
min	0.000000	0.000000e+00
25%	20.000000	6.185000e+02
50%	32.000000	4.320000e+03
75%	42.000000	4.008950e+04
max	68.000000	6.549380e+06

- **popularity**
 - Trung bình: 30.95, với độ lệch chuẩn là 14.98.

- Min/Max: Dao động từ 0 đến 68.
- Phân phối:
 - 25% số nghệ sĩ có độ phổ biến dưới 20.
 - 50% số nghệ sĩ (median) có độ phổ biến khoảng 32.
 - 75% số nghệ sĩ có độ phổ biến dưới 42.
- Nhận xét:
 - Độ phổ biến của các nghệ sĩ khá đa dạng, nhưng phần lớn tập trung ở mức trung bình (20-42).
 - Một số nghệ sĩ có độ phổ biến rất thấp (min = 0), có thể là các nghệ sĩ ít được biết đến hoặc mới xuất hiện.

○ **followers**

- Trung bình: 89,332 người theo dõi, với độ lệch chuẩn là 327,377.
- Min/Max: Dao động từ 0 đến 6.54 triệu người theo dõi.
- Phân phối:
 - 25% số nghệ sĩ có dưới 618 người theo dõi.
 - 50% số nghệ sĩ (median) có khoảng 4,320 người theo dõi.
 - 75% số nghệ sĩ có dưới 40,089 người theo dõi.
- Nhận xét:
 - Số lượng người theo dõi có sự phân tán rất lớn, với một số nghệ sĩ có hàng triệu người theo dõi (max = 6.54 triệu), trong khi nhiều nghệ sĩ có rất ít hoặc không có người theo dõi (min = 0).
 - Điều này cho thấy sự chênh lệch lớn giữa các nghệ sĩ nổi tiếng và ít nổi tiếng.

4. Mối tương quan giữa các biến

4.1.Vietnamese_only_songs_cleaned



- Album_popularity

○ Tương quan mạnh với popularity (0.81):

- Điều này cho thấy album phổ biến thường chứa các bài hát phổ biến. Đây là mối quan hệ hợp lý vì độ phổ biến của album thường phụ thuộc vào độ phổ biến của các bài hát trong album.

○ Tương quan trung bình với track_number (0.28):

- Các bài hát ở vị trí đầu tiên trong album có thể đóng góp nhiều hơn vào độ phổ biến của album.

○ Tương quan trung bình với stream_counts (0.25):

- Album phổ biến thường có tổng lượt stream cao hơn.

- Disc_number

○ Không có mối tương quan đáng kể với các biến khác:

- Hệ số tương quan gần bằng 0 với tất cả các biến, cho thấy số đĩa trong album không ảnh hưởng nhiều đến các yếu tố như độ phổ biến hay lượt stream.

- Duration_ms

○ Tương quan âm yếu với album_popularity (-0.10) và popularity (-0.06):

- Thời lượng bài hát không ảnh hưởng nhiều đến độ phổ biến của album hoặc bài hát. Tuy nhiên, các bài hát quá dài có thể ít phổ biến hơn.

○ Không có mối tương quan đáng kể với các biến khác:

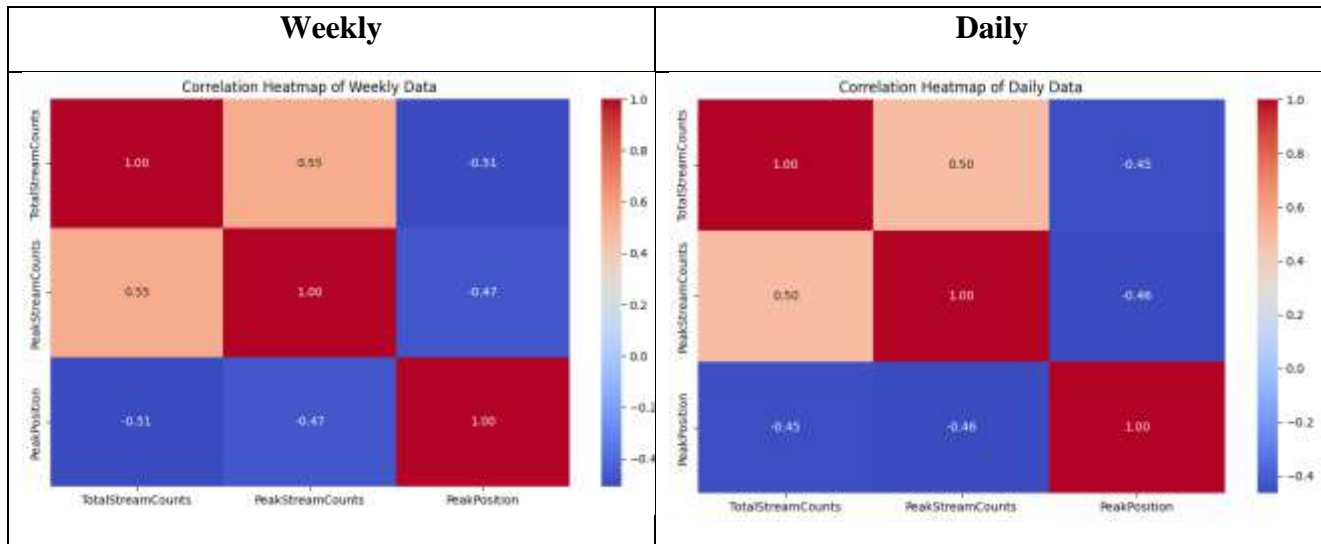
- Thời lượng bài hát không phải là yếu tố chính ảnh hưởng đến các biến khác như lượt stream hay vị trí bài hát.

- Popularity

○ Tương quan mạnh với album_popularity (0.81):

- Như đã phân tích ở trên, bài hát phổ biến thường nằm trong album phổ biến.
- Tương quan trung bình với `stream_counts` (0.33):
 - Các bài hát phổ biến thường có lượt stream cao hơn, nhưng không phải lúc nào cũng vậy. Có thể có những bài hát ít phổ biến nhưng vẫn có lượt stream cao do các yếu tố khác như quảng bá hoặc viral.
- **Track_number**
 - Tương quan trung bình với `album_popularity` (0.28):
 - Các bài hát ở vị trí đầu tiên trong album có thể đóng góp nhiều hơn vào độ phổ biến của album.
 - Không có mối tương quan đáng kể với các biến khác:
 - Vị trí bài hát trong album không ảnh hưởng nhiều đến độ phổ biến hay lượt stream.
- **Stream_counts**
 - Tương quan trung bình với `popularity` (0.33):
 - Các bài hát có lượt stream cao thường phổ biến hơn, nhưng không phải lúc nào cũng vậy.
 - Tương quan trung bình với `album_popularity` (0.25):
 - Album có tổng lượt stream cao thường phổ biến hơn.

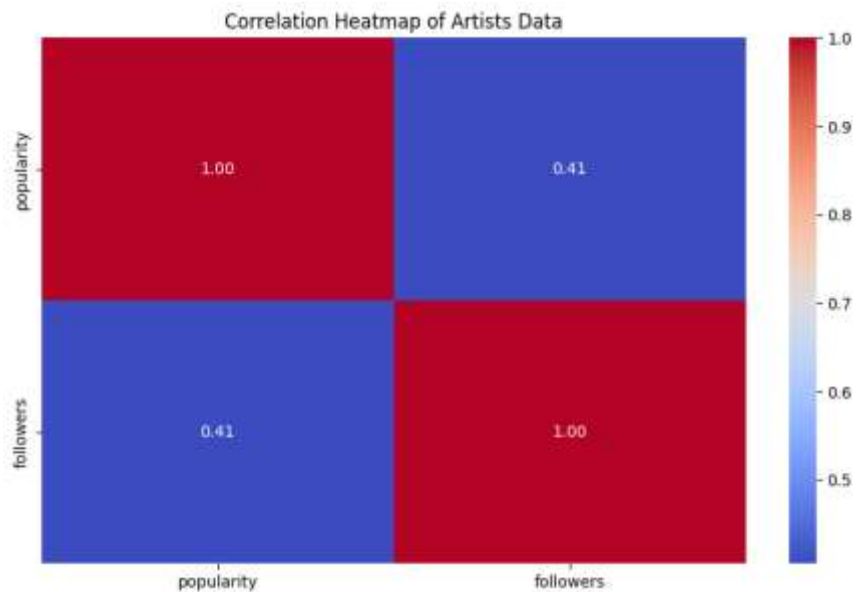
4.2. `position_streams_by_time_daily_total_peak_cleaned` and `position_streams_by_time_weekly_total_peak_cleaned`



- **Mối quan hệ giữa `TotalStreamCounts` và `PeakStreamCounts`:**
 - Weekly có hệ số tương quan cao hơn (0.55 so với 0.50), cho thấy trong dữ liệu Weekly, tổng lượt stream và lượt stream cao nhất có mối liên hệ chặt chẽ hơn.

- **Mối quan hệ giữa TotalStreamCounts và PeakPosition:**
 - o Weekly có hệ số tương quan âm mạnh hơn (-0.51 so với -0.45), cho thấy trong dữ liệu Weekly, tổng lượt stream ảnh hưởng nhiều hơn đến vị trí trên bảng xếp hạng.
- **Mối quan hệ giữa PeakStreamCounts và PeakPosition:**
 - o Weekly và Daily có hệ số tương quan tương tự (-0.47 và -0.46), cho thấy mối quan hệ giữa lượt stream cao nhất và vị trí trên bảng xếp hạng không thay đổi nhiều giữa hai tập dữ liệu.

4.3.Vietnamese_artitsts



- **Mối tương quan giữa popularity và followers**
 - o Hệ số tương quan: 0.41.
 - o Đây là mối tương quan dương trung bình, cho thấy rằng các nghệ sĩ có nhiều người theo dõi hơn thường có độ phổ biến cao hơn.
 - o Tuy nhiên, hệ số tương quan không quá mạnh, điều này có thể do:
 - Một số nghệ sĩ có nhiều người theo dõi nhưng không phổ biến (ít hoạt động hoặc không có bài hát mới).
 - Một số nghệ sĩ có độ phổ biến cao nhưng chưa thu hút được nhiều người theo dõi (nghệ sĩ mới nổi hoặc viral).
- **Mối tương quan nội tại**
 - o popularity với chính nó: Hệ số tương quan là 1.00, vì đây là mối tương quan hoàn hảo (một biến với chính nó).
 - o followers với chính nó: Hệ số tương quan là 1.00, tương tự như trên.

5. Tương quan stat model

5.1.Giới thiệu

- Trong bài toán này, nhóm sẽ phân tích mối tương quan giữa số lượng phát (stream_counts) và các biến độc lập còn lại và sự phổ biến (popularity) và các biến độc lập còn lại.

5.2. Phương pháp

- Ở đây nhóm sẽ sử dụng mô hình OLS (Ordinary Least Squares) và mô hình sẽ có dạng như sau:

$$\text{Biến Mục Tiêu} = \beta_0 + \beta_1 \times [\text{Biến Độc Lập}] + \varepsilon \quad (1)$$

- Trong đó, (β_0) là hệ số chặn, (β_1) là hệ số hồi quy cho [biến độc lập], và (ε) là sai số ngẫu nhiên.

5.3. Stream_counts vs các biến còn lại

OLS Regression Results

Dep. Variable:	stream_counts	R-squared:	0.113
Model:	OLS	Adj. R-squared:	0.112
Method:	Least Squares	F-statistic:	123.0
Date:	Tue, 08 Apr 2025	Prob (F-statistic):	6.49e-123
Time:	09:43:19	Log-Likelihood:	-86942.
No. Observations:	4817	AIC:	1.739e+05
Df Residuals:	4811	BIC:	1.739e+05
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P> t	[0.025	0.975]
const	-3.758e+06	4.14e+06	-0.908	0.364	-1.19e+07	4.36e+06
album_popularity	-6.514e+04	3.1e+04	-2.100	0.036	-1.26e+05	-4319.406
disc_number	-3.519e+05	3.96e+06	-0.089	0.929	-8.11e+06	7.41e+06
duration_ms	-12.6594	4.095	-3.092	0.002	-20.687	-4.632
popularity	4.596e+05	3.17e+04	14.518	0.000	3.98e+05	5.22e+05
track_number	1080.9226	7.5e+04	0.014	0.988	-1.46e+05	1.48e+05

Omnibus:	13601.796	Durbin-Watson:	1.699
Prob(Omnibus):	0.000	Jarque-Bera (JB):	558763657.541
Skew:	36.775	Prob(JB):	0.00
Kurtosis:	1669.898	Cond. No.	5.78e+06

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

[2] The condition number is large, 5.78e+06. This might indicate that there are strong multicollinearity or other numerical problems.

- **Tổng quan về mô hình**
 - o **R-squared (R^2): 0.113**
 - Chỉ số này cho biết mô hình giải thích được 11.3% sự biến thiên của biến phụ thuộc (stream_counts). Đây là một giá trị thấp, cho thấy mô hình chưa phù hợp để dự đoán tốt stream_counts.
 - Có thể còn nhiều yếu tố quan trọng khác chưa được đưa vào mô hình.
 - o **Adj. R-squared (R^2 điều chỉnh): 0.112**

- Giá trị này gần bằng R^2 , cho thấy số lượng biến độc lập trong mô hình không gây ra quá nhiều nhiễu.
- **F-statistic: 123.0 (p-value: 6.49e-123)**
 - Giá trị này rất nhỏ, cho thấy mô hình tổng thể có ý nghĩa thống kê. Ít nhất một biến độc lập có ảnh hưởng đáng kể đến biến phụ thuộc.
- **Ý nghĩa của các biến:**
 - **album_popularity:**
 - Hệ số: -6.514e+04.
 - Mỗi quan hệ âm: Khi độ phổ biến của album tăng thêm 1 đơn vị, số lượt stream giảm trung bình 65,140 lượt, giữ các yếu tố khác không đổi.
 - P-value: 0.036 (< 0.05), có ý nghĩa thống kê.
 - **disc_number:**
 - Hệ số: -3.519e+05.
 - Không có ý nghĩa thống kê (p-value: 0.929), cho thấy số đĩa không ảnh hưởng đáng kể đến số lượt stream.
 - **duration_ms:**
 - Hệ số: -12.6594.
 - Mỗi quan hệ âm: Khi thời lượng bài hát tăng thêm 1 ms, số lượt stream giảm trung bình 12.66 lượt.
 - P-value: 0.002 (< 0.05), có ý nghĩa thống kê.
 - **popularity:**
 - Hệ số: 4.596e+05.
 - Mỗi quan hệ dương: Khi độ phổ biến của bài hát tăng thêm 1 đơn vị, số lượt stream tăng trung bình 459,600 lượt.
 - P-value: 0.000 (< 0.05), có ý nghĩa thống kê mạnh.
 - **track_number:**
 - Hệ số: 1080.9226.
 - Không có ý nghĩa thống kê (p-value: 0.988), cho thấy vị trí bài hát trong album không ảnh hưởng đáng kể đến số lượt stream.

5.4. Popularity vs các biến còn lại

OLS Regression Results						
Dep. Variable:	popularity	R-squared:	0.737			
Model:	OLS	Adj. R-squared:	0.737			
Method:	Least Squares	F-statistic:	2699.			
Date:	Tue, 08 Apr 2025	Prob (F-statistic):	0.00			
Time:	10:01:35	Log-Likelihood:	-16501.			
No. Observations:	4817	AIC:	3.301e+04			
Df Residuals:	4811	BIC:	3.305e+04			
Df Model:	5					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
const	17.2967	1.828	9.461	0.000	13.713	20.881
album_popularity	0.8051	0.008	106.925	0.000	0.790	0.820
disc_number	-9.6738	1.759	-5.499	0.000	-13.122	-6.225
duration_ms	1.038e-05	1.82e-06	5.700	0.000	6.81e-06	1.39e-05
track_number	-1.0464	0.030	-35.095	0.000	-1.105	-0.988
stream_counts	9.132e-08	6.29e-09	14.518	0.000	7.9e-08	1.04e-07
Omnibus:	917.812	Durbin-Watson:	1.303			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2291.298			
Skew:	-1.047	Prob(JB):	0.00			
Kurtosis:	5.651	Cond. No.	4.26e+08			
Notes:						
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.						
[2] The condition number is large, 4.26e+08. This might indicate that there are strong multicollinearity or other numerical problems.						

- Tổng quan về mô hình

o R-squared (R^2): 0.737

- Mô hình giải thích được 73.7% sự biến thiên của biến phụ thuộc (popularity). Đây là một giá trị khá cao, cho thấy mô hình phù hợp để dự đoán popularity.

o Adj. R-squared (R^2 điều chỉnh): 0.737

- Giá trị này gần bằng R^2 , cho thấy số lượng biến độc lập trong mô hình không gây ra quá nhiều nhiễu.
- F-statistic: 2699 (p-value: 0.000)
- Giá trị này rất nhỏ, cho thấy mô hình tổng thể có ý nghĩa thống kê. Ít nhất một biến độc lập có ảnh hưởng đáng kể đến biến phụ thuộc.

- Ý nghĩa của các biến:

o const:

- Hệ số: 17.2967.
- Đây là giá trị trung bình của popularity khi tất cả các biến độc lập khác bằng 0.

o album_popularity:

- Hệ số: 0.8051.
- Mối quan hệ dương: Khi độ phổ biến của album tăng thêm 1 đơn vị, popularity tăng trung bình 0.8051 đơn vị.
- P-value: 0.000 (< 0.05), có ý nghĩa thống kê mạnh.
- **disc_number:**
 - Hệ số: -9.6738.
 - Mối quan hệ âm: Khi số đĩa tăng thêm 1, popularity giảm trung bình 9.6738 đơn vị.
 - P-value: 0.000 (< 0.05), có ý nghĩa thống kê mạnh.
- **duration_ms:**
 - Hệ số: 1.038e-05.
 - Mối quan hệ dương: Khi thời lượng bài hát tăng thêm 1 ms, popularity tăng trung bình 0.00001038 đơn vị.
 - P-value: 0.000 (< 0.05), có ý nghĩa thống kê mạnh.
- **track_number:**
 - Hệ số: -1.0464.
 - Mối quan hệ âm: Khi số thứ tự bài hát trong album tăng thêm 1, popularity giảm trung bình 1.0464 đơn vị.
 - P-value: 0.000 (< 0.05), có ý nghĩa thống kê mạnh.
- **stream_counts:**
 - Hệ số: 9.132e-08.
 - Mối quan hệ dương: Khi số lượt stream tăng thêm 1 đơn vị, popularity tăng trung bình 0.00000009132 đơn vị.
 - P-value: 0.000 (< 0.05), có ý nghĩa thống kê mạnh.

6. Trực quan hóa dữ liệu

6.1. Tích hợp AI trong việc thiết kế dashboard

- Ở đây nhóm đã nhờ sự giúp đỡ của [Grok](#) một chatbot AI để có thể gợi ý dashboard, trong đó:

i need to build dashboard on these dataset can you recommend how many pages should i have and each page should have what

vietnamese_artists.csv

vietnamese_songs_full.csv

Thoughts

Expand for details

To build an effective dashboard based on the provided datasets, which include artist data (e.g., name, popularity, followers, genres) and track data (e.g., album, artists, popularity, stream counts, genres), I recommend structuring it across **four distinct pages**. Each page will focus on a specific aspect of the data, providing clear insights for different user needs, such as understanding artist popularity, track performance, genre trends, and comparative analyses. Below, I outline the recommended pages and their content.

How can Grok help?



DeepSearch



Think

Grok 3



Page 1: Artist Overview

Purpose: Provide a high-level summary of artists, their popularity, and their audience reach.

Content:

- **Total Artists:** A single-value card displaying the total number of unique artists in the dataset (e.g., count of distinct `id` values in the artist dataset).
- **Top Artists by Popularity:** A bar chart or table listing the top 10 artists by `popularity`, including their `name`, `popularity` score, and `followers`. For example, artists like Sơn Tùng M-TP (popularity 64) or HIEUTHUHAI (popularity 67) could appear here.
- **Follower Distribution:** A histogram or box plot showing the distribution of `followers` across all artists, highlighting the range (e.g., from 51 for Thiên Dung to 6,425,018 for Sơn Tùng M-TP) and identifying outliers.
- **Genre Breakdown:** A pie chart or treemap displaying the frequency of genres across artists (e.g., `v-pop`, `vinahouse`, `vietnamese hip hop`), derived from the `genres` column. This helps identify dominant music styles.
- **Filters:** Interactive filters for `popularity` range, `followers` range, and `genres` to allow users to drill down into specific artist segments.

Why: This page serves as the entry point, giving users a quick snapshot of the artist landscape and enabling them to identify key players and trends.

How can Grok help?



DeepSearch



Think

Grok 3



Page 2: Track Performance

Purpose: Analyze individual tracks, their popularity, and streaming success.


Content:

- **Total Tracks:** A single-value card showing the total number of unique tracks (e.g., count of distinct `id` values in the track dataset).
- **Top Tracks by Streams:** A bar chart or table of the top 10 tracks by `stream_counts`, including `name`, `artists`, `popularity`, and `stream_counts`. For instance, "Blinding Lights" by The Weeknd with 4,293,744,430 streams would rank high.
- **Popularity vs. Streams Scatter Plot:** A scatter plot with `popularity` on the x-axis and `stream_counts` on the y-axis, with points sized by `duration_ms` or colored by `album_type` (e.g., single, album). This visualizes the relationship between popularity and streaming success.
- **Release Date Trends:** A line chart plotting the number of tracks or total `stream_counts` over time based on `release_date`, showing peaks (e.g., 2020 releases like "Dynamite" by BTS).
- **Filters:** Options to filter by `artists`, `album_type`, `release_date` range, and `genres` for focused analysis.

Why: This page focuses on track-level insights, crucial for understanding what content drives engagement and how performance varies across releases.

How can Grok help?

  DeepSearch  Think

Grok 3 

Page 3: Genre Insights

Purpose: Explore the distribution and performance of music genres across artists and tracks.

Content:

- **Genre Popularity:** A bar chart showing average `popularity` per genre, calculated by aggregating artist and track `popularity` scores linked to each genre (e.g., averaging `v-pop` artists like VU, and AMEE).
- **Genre Stream Counts:** A stacked bar chart or table displaying total `stream_counts` by genre, derived from tracks, with contributions from top tracks highlighted (e.g., `k-pop` streams from BTS's "Dynamite").
- **Artist Count per Genre:** A bar chart or heatmap showing the number of artists associated with each genre, revealing genre saturation (e.g., many artists in `vinahouse` like Orinn and Anh Quân Idol).
- **Genre Overlap:** A Sankey diagram or chord diagram illustrating how artists and tracks span multiple genres (e.g., artists with both `v-pop` and `vietnamese hip-hop`), based on the `genres` lists.
- **Filters:** Filters for specific genres and a toggle to switch between artist-based and track-based views.

Why: Genres are a key dimension in music data, and this page helps users understand their prevalence and impact, useful for market analysis or content planning.

How can Grok help?

  DeepSearch  Think

Grok 3 

Page 4: Comparative Analysis

Purpose: Enable side-by-side comparisons of artists and tracks for deeper insights.

Content:

- **Artist Comparison:** A dual-axis chart or parallel coordinates plot comparing selected artists across metrics like `popularity`, `followers`, and count of tracks (linked via the `artists` field in the track dataset). Users can select artists (e.g., Đen vs. Sơn Tùng M-TP).
- **Track Comparison:** A table or radar chart comparing selected tracks on `popularity`, `streams_count`, `duration_ms`, and `explicit` status. For example, compare "Đi Về Nhà" by Đen and JustaTee with "CUA" by MANBO and HIEUTHUAI.
- **Artist-Track Linkage:** A network graph or table showing which artists contribute to which tracks, with metrics like average track `popularity` per artist, leveraging the overlap between datasets.
- **Custom Filters:** Dropdowns to select specific artists and tracks, with dynamic updates to visuals.

Why: This page caters to users needing detailed comparisons, such as marketers or analysts benchmarking performance across entities.

Summary of Recommendation

How can Grok help?

🔍 DeepSearch - 💡 Think

Grok 3 ⌵ ⬆

Summary of Recommendation

• Number of Pages: 4

1. **Artist Overview** - Summary of artists, popularity, and genres.
2. **Track Performance** - Focus on track metrics and streaming trends.
3. **Genre Insights** - Genre distribution and performance analysis.
4. **Comparative Analysis** - Detailed artist and track comparisons.

Rationale: Four pages strike a balance between comprehensive coverage and usability, aligning with common dashboard goals: overview, detail, segmentation, and comparison. Each page leverages the datasets' key fields (`popularity`, `followers`, `genres`, `streams_count`, etc.) and supports interactivity for exploration, making the dashboard versatile for various stakeholders like music analysts, marketers, or fans.

This structure ensures clarity, avoids overwhelming users with too much data on a single page, and provides actionable insights tailored to different aspects of the music ecosystem represented in the datasets.

🔄 ⌵ ⬆

- Kể đến sẽ là phần chi tiết của các dashboard

6.2. Dashboard tổng quan

6.2.1. Tổng thể dashboard



Giao diện dashboard tổng quan

- Dashboard này cung cấp một cái nhìn tổng thể về thị trường âm nhạc Việt Nam bao gồm quy mô, các đặc điểm trung bình, xu hướng theo thời gian và phân bố của các yếu tố chính như thể loại và độ phổ biến.

6.2.2. Các Chỉ số Chính (KPIs):

- **Quy mô:** Tập dữ liệu bao gồm 1409 nghệ sĩ, 4735 bài hát, và 195 thể loại khác nhau. Tổng lượt phát (streams/plays) đạt 4.1 triệu.
- **Đặc điểm trung bình:**
 - Thời lượng trung bình của một bài hát là 4.0 phút.
 - Độ phổ biến trung bình là 28.

6.2.3. Xu hướng theo Thời gian:

- **Số lượng bài hát:** Biểu đồ đường "Số lượng bài hát qua từng năm" cho thấy một sự gia tăng đột biến và cực kỳ mạnh mẽ về số lượng bài hát được ghi nhận trong những năm gần đây (đặc biệt là khoảng sau 2015 và đỉnh điểm vào đầu những năm 2020). Giai đoạn trước năm 2000 có số lượng bài hát rất ít. Điều này có thể phản ánh sự bùng nổ của thị trường âm nhạc hoặc sự thay đổi trong phương pháp thu thập dữ liệu.
- **Thời lượng bài hát:** Biểu đồ "Thời lượng bài hát qua từng năm" có một xu hướng thú vị: thời lượng trung bình đạt đỉnh cao (khoảng 8 phút) vào đầu những năm 1980, sau đó giảm mạnh và ổn định quanh

mức 4 phút từ cuối những năm 1990 cho đến nay, với một xu hướng giảm nhẹ trong những năm gần nhất. Điều này có thể phản ánh sự thay đổi trong thị hiếu âm nhạc, từ các định dạng dài hơn sang các bài hát ngắn gọn, phù hợp với radio hoặc streaming hơn.

6.2.4. Phân bố:

- **Thể loại nhạc:** v-pop chiếm tỷ trọng lớn nhất (29.74%), theo sau là vinahouse (19.68%), vietnamese lo-fi (19.64%), vietnamese hip hop (15.8%), và vietnam indie (15.14%). Tổng cộng 5 thể loại này chiếm gần như toàn bộ thị phần trong biểu đồ này.
- **Độ phổ biến:** Phần lớn các bài hát tập trung ở mức độ phổ biến trung bình (cao nhất ở khoảng 30-40). Có ít bài hát ở mức rất thấp (0-10) hoặc rất cao (50+). Phân bố này khá chuẩn, tập trung quanh giá trị trung bình là 28.

6.2.5. Bộ lọc:

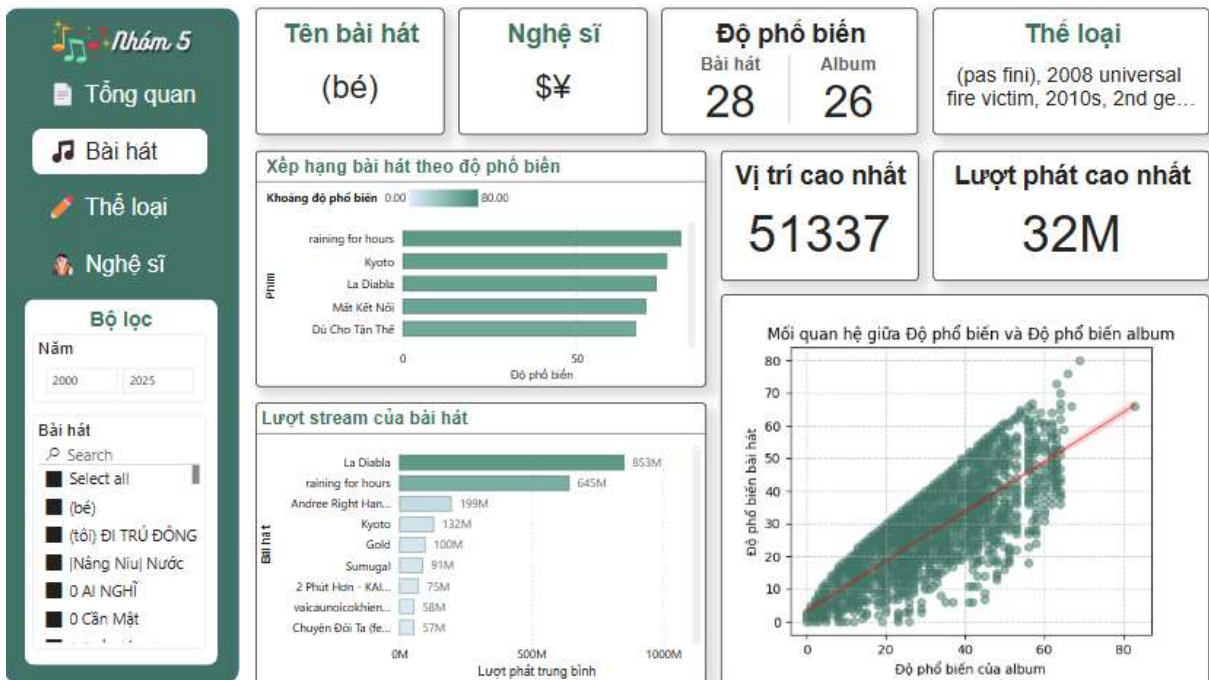
- Có thể lọc dữ liệu theo Năm, Khoảng phổ biến, Thể loại và Thời lượng.

6.2.6. Kết luận:

- Các thể loại như v-pop, vinahouse, lo-fi, hip hop và indie chiếm ưu thế áp đảo. Thời lượng bài hát trung bình đã ổn định ở mức khoảng 4 phút trong nhiều năm. Dữ liệu về độ phổ biến cho thấy hầu hết các bài hát có mức độ phổ biến trung bình.

6.3. Songs dashboard

6.3.1. Tổng thể dashboard

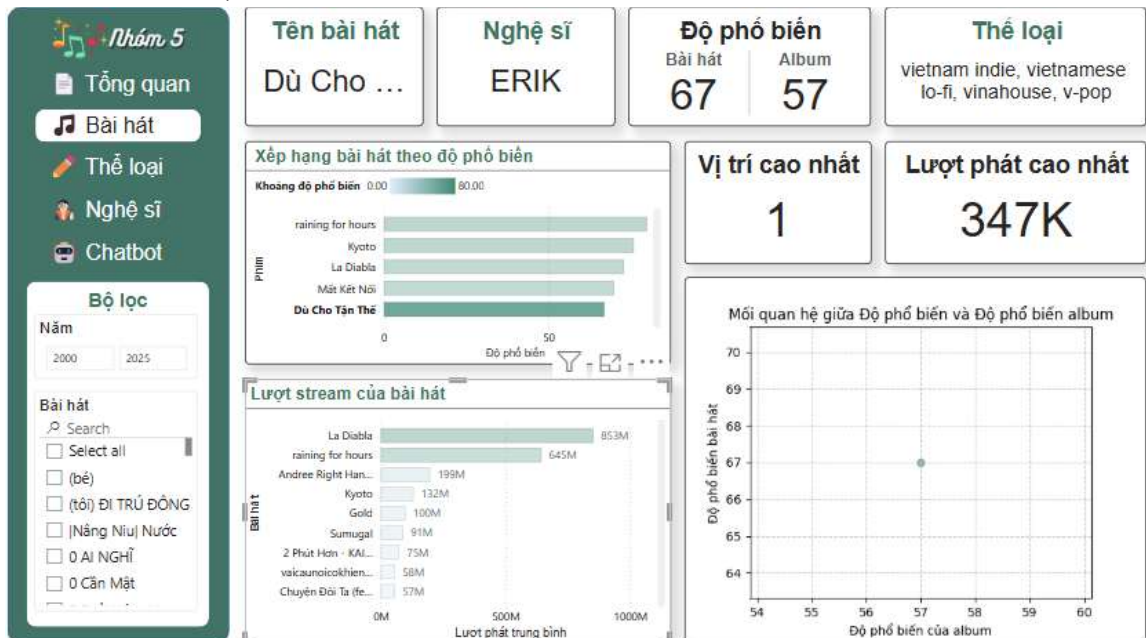


- Dashboard cung cấp thông tin về từng bài hát với các phần chính: tên bài hát, nghệ sĩ trình diễn, độ phổ biến, thể loại bài hát, vị trí cao nhất bài hát từng lên, lượt phát cao nhất mà bài hát đạt được.
- Cùng với các thông tin tổng quan về các bài hát trên thị trường như: biểu đồ xếp hạng bài hát theo độ phổ biến, biểu đồ xếp hạng lượt stream của bài hát, và biểu đồ cho thấy mối quan hệ giữa độ phổ biến với độ phổ biến album.

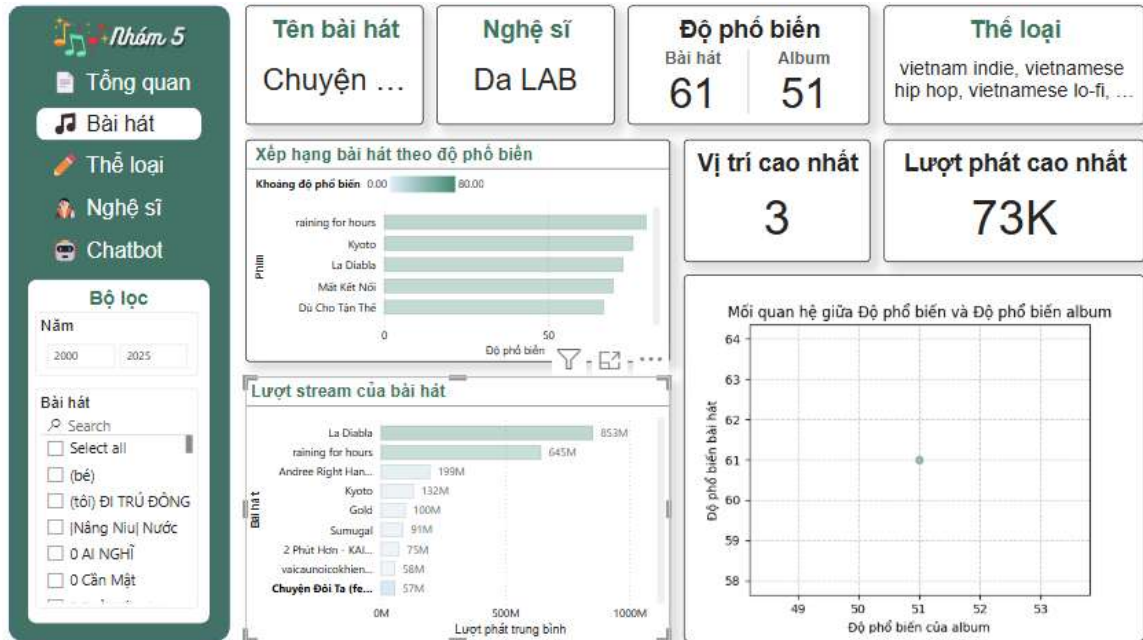
6.3.2. Các nhận xét biểu đồ tổng quát:

- Những bài hát có độ phổ biến cao nhất là: raining for hours (80), Kyoto (76), La Diabla (73), Mất kết nối (70),... Những bài hát có lượt stream cao nhất là: La Diabla (853M), raining for hours (645M), Andree Right Hand xin chào (199M),...
- Có thể thấy các bài hát có độ phổ biến cao sẽ thường có lượt stream cao và nằm trong top đầu, điều này là thực tế, vì độ phổ biến cao sẽ có nhiều người nghe dẫn tới lượt stream cao.
- **Về biểu đồ mối tương quan giữa độ phổ biến bài hát và độ phổ biến album:** Biểu đồ thể hiện mối tương quan dương tuyến tính khá rõ ràng giữa độ phổ biến của bài hát và độ phổ biến của album chứa bài hát đó. Điều này có nghĩa là khi album phổ biến, khả năng cao các bài hát trong album đó cũng phổ biến, và ngược lại.
- Dường như không có nhiều bài hát phổ biến xuất hiện trong những album ít phổ biến. Điều này phản ánh tầm quan trọng của thương hiệu và độ lan tỏa của toàn bộ album đối với thành công của từng bài hát.

6.3.3. Các chi tiết về một số bài hát nổi bật (lấy các bài Dù cho tận thế, Chuyện đời ta).



- Bài "Dù Cho Tân Thê" của ERIK đạt độ phổ biến cao với vị trí cao nhất là số 1 và lượt phát cao nhất 347K, độ phổ biến bài hát là 67 và album là 57, cho thấy sức hút mạnh mẽ của ca khúc; ERIK – một trong những nghệ sĩ nổi bật của V-pop – đã thể hiện rõ khả năng tạo hit và duy trì độ phổ biến ổn định qua nhiều bài hát.



- Tiếp theo là bài hát có lượt stream nằm trong top là "Chuyện đôi ta" của Da LAB – nhóm nhạc đình đám. Với độ phổ biến bài hát và album lần lượt là 61, 51. Bài hát đã leo tới vị trí cao nhất 3 với số lượt phát cao nhất là 73K.

6.3.4. Bộ lọc:

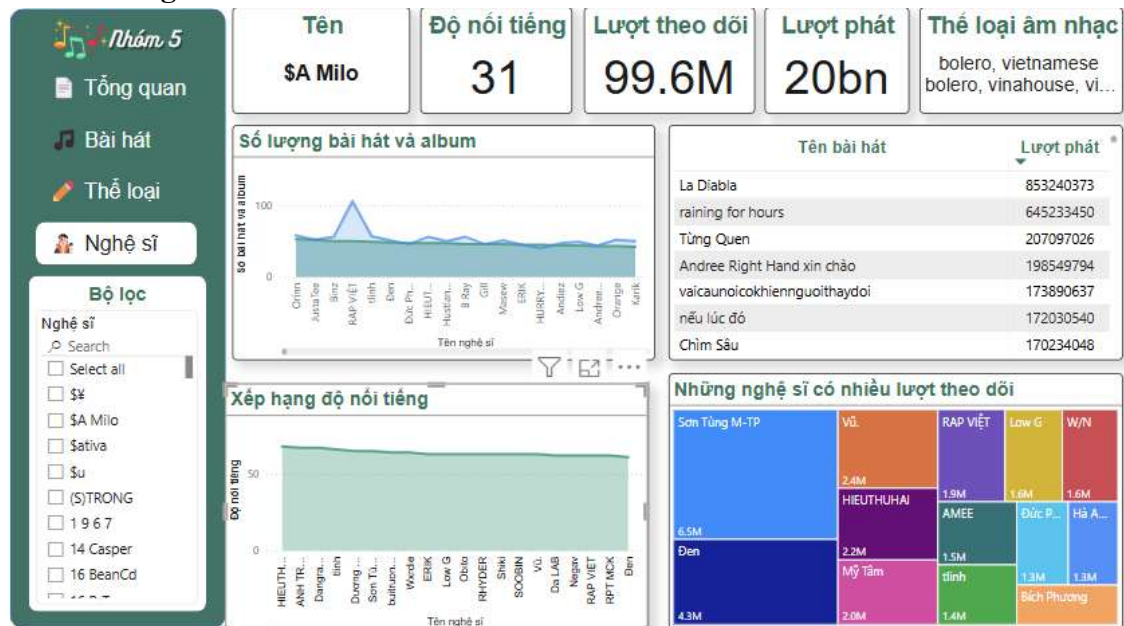
- Cho phép ta lọc theo năm, theo tên bài hát, cho phép tìm kiếm

6.3.5. Kết luận:

- Dashboard cho thấy rằng một số bài hát nổi bật như La Diabla, Raining for hours, và Kyoto đạt được độ phổ biến cao và lượt stream ấn tượng (trên hàng trăm triệu), chứng minh sức hút mạnh mẽ và tính lan truyền cao của các bản hit này. Và việc các bài hát có độ phổ biến cao sẽ thường có lượt stream nằm trong top là thực tế
- Biểu đồ tương quan giữa độ phổ biến bài hát và độ phổ biến album cho thấy mối liên hệ tuyến tính dương, tức album phổ biến thường chứa các bài hát phổ biến. Một điểm nữa đó là gần như không có nhiều bài hát phổ biến cao xuất hiện trong các album có độ phổ biến thấp, điều này cho thấy tầm quan trọng của thương hiệu

6.4. Artists dashboard

6.4.1. Tổng thể dashboard



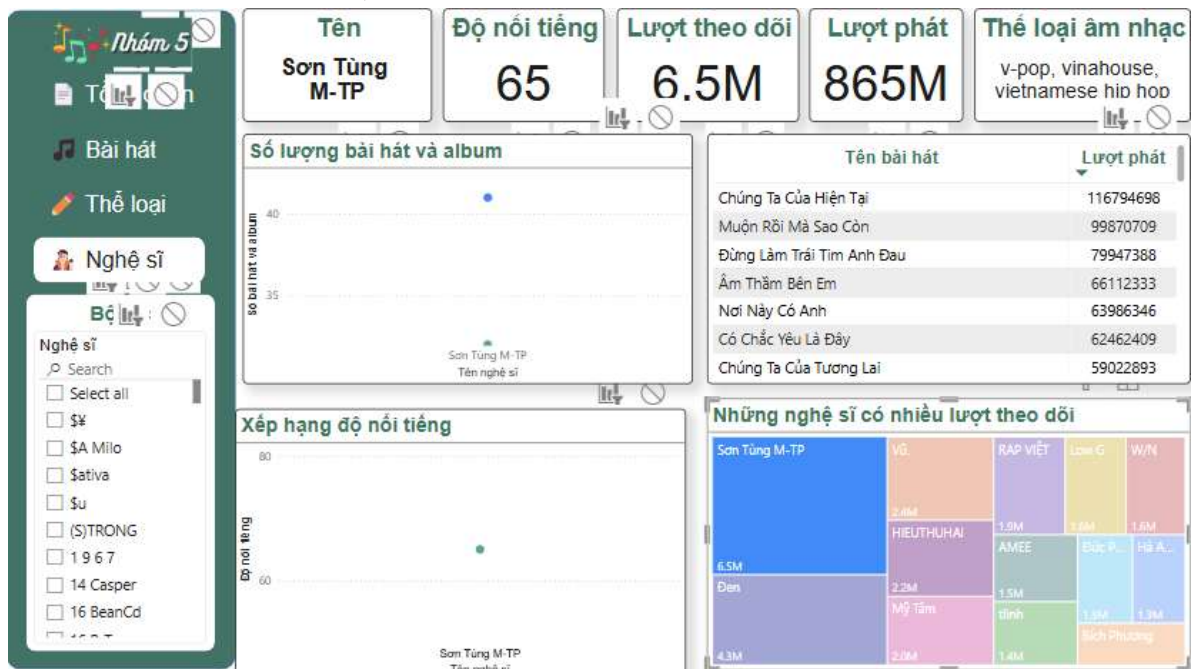
Giao diện dashboard nghệ sĩ

- Dashboard tập trung vào phân tích chi tiết từng nghệ sĩ cụ thể. Bao gồm các chỉ số như độ nổi tiếng, lượt theo dõi, tổng lượt phát, và thể loại nhạc liên quan. Ngoài ra, dashboard còn đưa ra so sánh vị trí của nghệ sĩ với các nghệ sĩ khác trên thị trường âm nhạc.

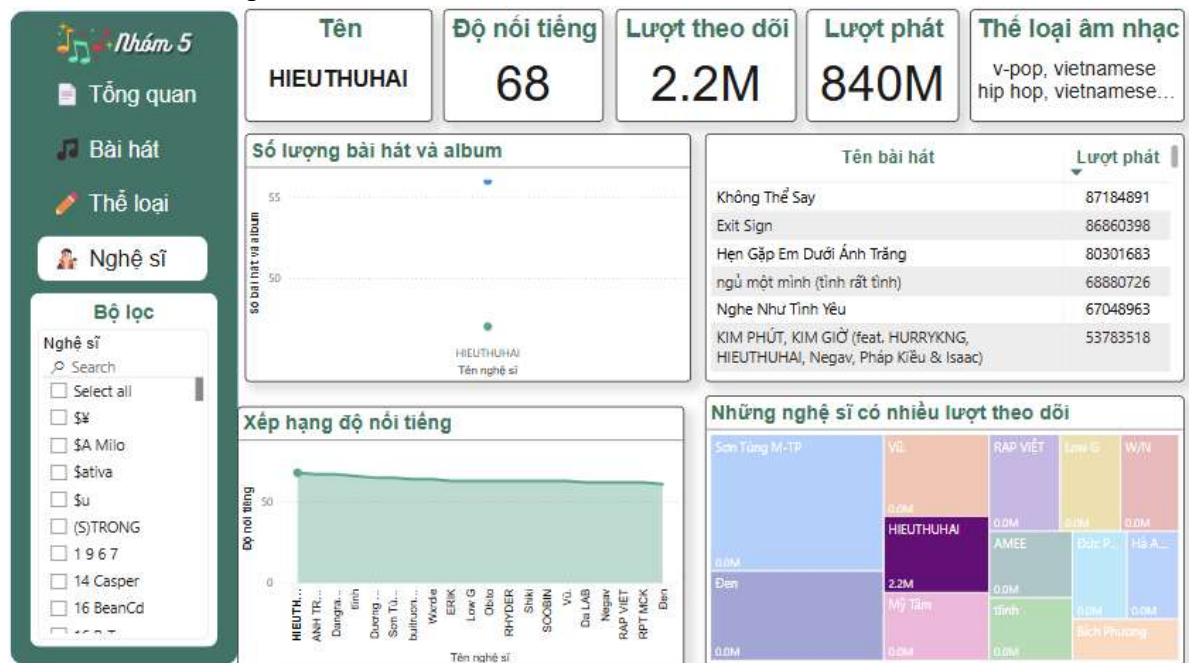
6.4.2. So sánh tổng quát:

- Tree map bên dưới so sánh các nghệ sĩ có nhiều lượt theo dõi nhất như Sơn Tùng M-TP (6.5M), Đen Vâu (4.3M), Vũ (2.4M), HIEUTHUAI (2.2M),...
- Mặc dù có nhiều lượt theo dõi nhất nhưng về độ nổi tiếng Sơn Tùng MTP (65) chỉ xếp ở vị trí thứ 6 sau HIEUTHUAI (68), Anh Trai Say Hi (67), Dangrangto (67), tlinh (66).
- Thậm chí những cái tên có nhiều lượt theo dõi khác như Đen Vâu, Vũ, Mỹ Tâm cũng có thứ hạng rất thấp trong độ nổi tiếng. Điều này cho thấy rằng các nghệ sĩ gạo cội, có hoạt động nghệ thuật lâu năm (những người đã có nhiều người theo dõi) sẽ không nổi tiếng bằng các hiện tượng mới nổi đương thời. Điều này là do khán giả trên nền tảng spotify chủ yếu là giới trẻ, sự thay đổi trong xu hướng âm nhạc. Một nghệ sĩ có thể từng rất nổi (nhiều người follow), nhưng nếu không có bài hit mới hoặc ít hoạt động gần đây, thì độ nổi tiếng hiện tại thấp.

6.4.3. Các phân tích chi tiết (xét các ca sĩ nổi bật như Sơn Tùng, HIEUTHUHAI):



- Sơn Tùng là ca sĩ có nhiều lượt theo dõi nhất với 6.5M lượt theo dõi, 865M lượt phát, 65 điểm độ nổi tiếng. Với các thể loại âm nhạc đa dạng như v-pop, vinahouse, hiphop. Cùng các bài hát quen thuộc trong giới trẻ như: Chúng ta của hiện tại, Muộn rồi mà sao còn,... đã góp phần làm nên tên tuổi của MTP.



- HIEUTHUHAI – cái tên cực kì quen thuộc trong cộng đồng người yêu vpop, vietnamese hiphop. Không bắt ngờ khi nhà sản xuất âm nhạc

này có độ nổi tiếng cao nhất, vượt mặt các cái tên đình đám như tlinh, Dangrangto,... Với những bài hit như Không thể say, Exit Sign, Ngủ một mình,... mang về cho anh 68 điểm nổi tiếng, 2.2M lượt theo dõi và 840M lượt phát.

6.4.4. Bộ lọc:

- Cho phép lọc theo tên nghệ sĩ, cho phép tìm kiếm theo tên.

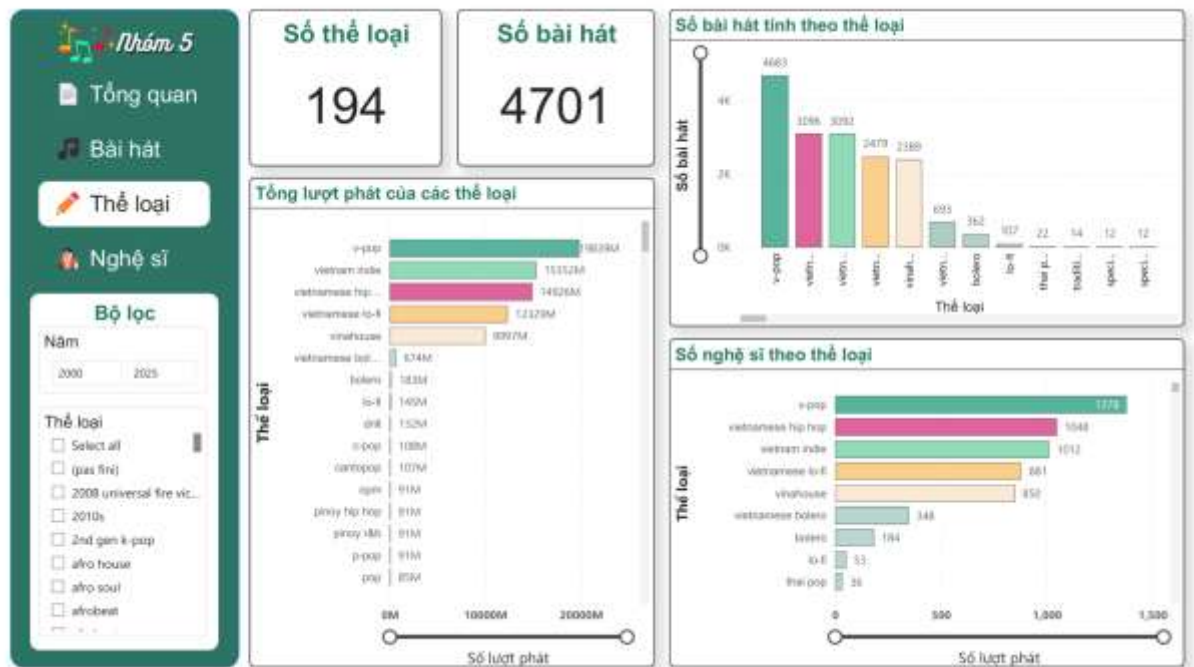
6.4.5. Kết luận:

Insight: Tại sao một nghệ sĩ có lượt theo dõi cao nhưng lại không đứng top trong bảng xếp hạng độ nổi tiếng?

- Lượt theo dõi phản ánh sự trung thành: **Lượt theo dõi (followers)** là **chỉ số tích lũy theo thời gian**, phản ánh sự trung thành của fan, mức độ được yêu thích bền vững trong quá khứ
- Độ nổi tiếng phản ánh xu hướng: **Độ nổi tiếng (popularity)** là một **chỉ số động**, có thể tính theo số lượt phát gần đây, sự viral trên mạng xã hội, vị trí trên bảng xếp hạng gần nhất,...
- Một nghệ sĩ có thể từng rất nổi (nhiều người follow), nhưng nếu không có bài hit mới hoặc ít hoạt động gần đây, thì độ nổi tiếng hiện tại thấp.
- Nghệ sĩ “huyền thoại” vs “ngôi sao hiện tại”: Một số nghệ sĩ như Mỹ Tâm, Đen Vâu, Sơn Tùng có fanbase lâu dài, ổn định, nên lượng theo dõi cao. Trong khi đó, các nghệ sĩ như HIEUTHUAI, tlinh, hay các hiện tượng có thể viral nhất thời, khiến độ nổi tiếng tăng vọt – nhưng lượng follower tích lũy chưa kịp theo kịp.

6.5. Dashboard thể loại

6.5.1. Tổng thể dashboard



Giao diện dashboard thể loại

- Dashboard này đi sâu vào phân tích các thể loại nhạc, xem xét số lượng bài hát, tổng lượt phát và số lượng nghệ sĩ liên quan đến từng thể loại.

6.5.2. Các Chỉ số Chính (KPIs):

- Bộ dữ liệu có tổng cộng 194 thể loại và 4701 bài hát.

6.5.3. Phân tích chi tiết theo Thể loại:

- **Tổng lượt phát:** Mặc dù có nhiều thể loại, lượt phát tập trung chủ yếu vào 5 thể loại hàng đầu đã thấy ở trang Tổng quan. v-pop dẫn đầu với khoảng 19.8 tỷ lượt phát (19839M), theo sau là vietnam indie (~15.3 tỷ), vietnamese hip hop (~14.9 tỷ), vietnamese lo-fi (~12.3 tỷ), và vinahouse (~10 tỷ). Các thể loại khác như vietnamese bolero, bolero, lo-fi ... có tổng lượt phát thấp hơn đáng kể.
- **Số lượng bài hát:** Số lượng của 5 thể loại chính quá khác biệt so với các thể loại còn lại: v-pop có nhiều bài hát nhất (4683), tiếp theo là vietnamese hip hop (3096), vietnam indie (3092), vietnamese lo-fi (2479), và vinahouse (2389). Các thể loại khác có số lượng bài hát ít hơn nhiều.
- **Số lượng nghệ sĩ:** V-pop liên kết với nhiều nghệ sĩ nhất (1378), theo sau là vietnamese hip hop (1048), vietnam indie (1012), vietnamese lo-

fi (881), và vinahouse (850). Điều này cho thấy sự tập trung không chỉ về bài hát mà còn nghệ sĩ trong các thể loại này.

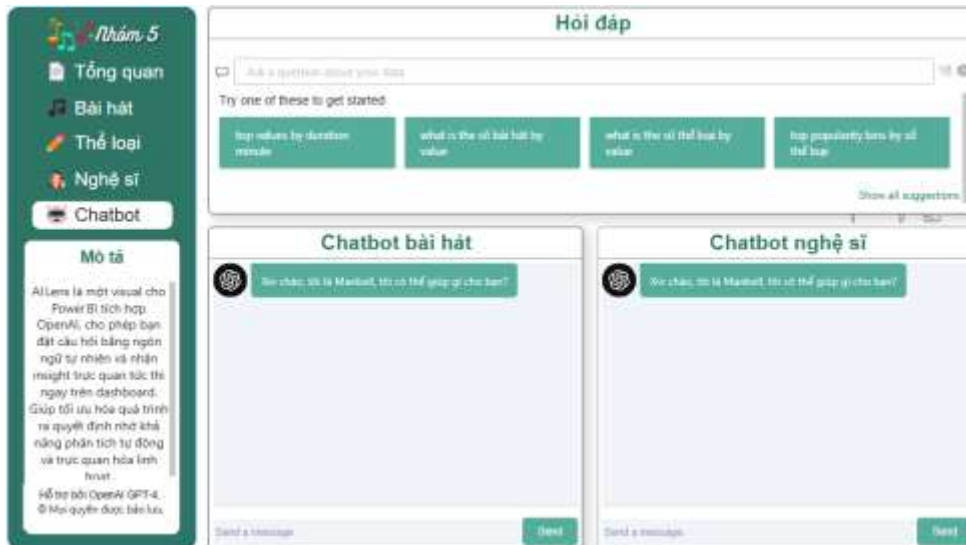
6.5.4. Bộ lọc:

- Cho phép lọc theo Năm và chọn cụ thể các Thể loại.

6.5.5. Kết luận:

- Thể hiện rõ sự thống trị của 5 thể loại nhạc Việt Nam hiện đại (v-pop, vietnamese hip hop, vietnam indie, vietnamese lo-fi, vinahouse) trên cả ba phương diện: tổng lượt phát, số lượng bài hát và số lượng nghệ sĩ. v-pop nổi bật là thể loại dẫn đầu về tổng lượt phát và cũng có nhiều bài hát/nghệ sĩ nhất. Sự chênh lệch giữa 5 thể loại hàng đầu và phần còn lại là rất lớn, cho thấy một thị trường tập trung cao độ vào các xu hướng âm nhạc hiện đại.

7. Tích hợp AI trong việc phân tích



7.1.Hỏi đáp



- Ở đây, nhóm sẽ sử dụng phần Q&A có sẵn trong PowerBI.
- Và đồng thời, chúng ta cũng có thể hỏi các câu hỏi để Q&A trả lời câu hỏi và ở trên là ví dụ.

7.2.Chatbot



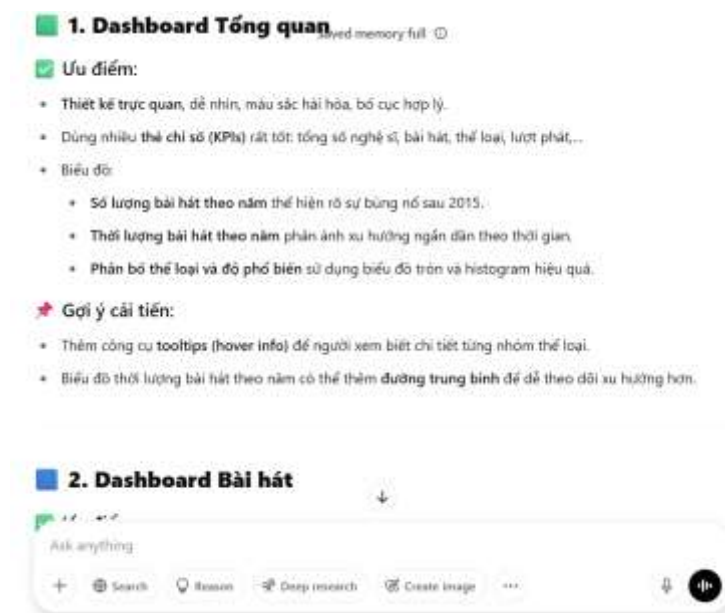
- Nhóm sử dụng một dạng visual mới là **AI Lens**: Tích hợp OpenAI cho phép đặt câu hỏi bằng ngôn ngữ tự nhiên và sẽ đưa ra insights về dữ liệu. Tuy nhiên về mặt khuyết điểm: Đây là text generation AI, vì thế sẽ bị hạn chế về mặt trực quan.

8. Tích hợp AI trong việc đánh giá dashboard

- Ở đây nhóm em sẽ sử dụng Chat GPT để có thể có được đánh giá tổng quan đồng thời để có được góc nhìn cải thiện của GPT.

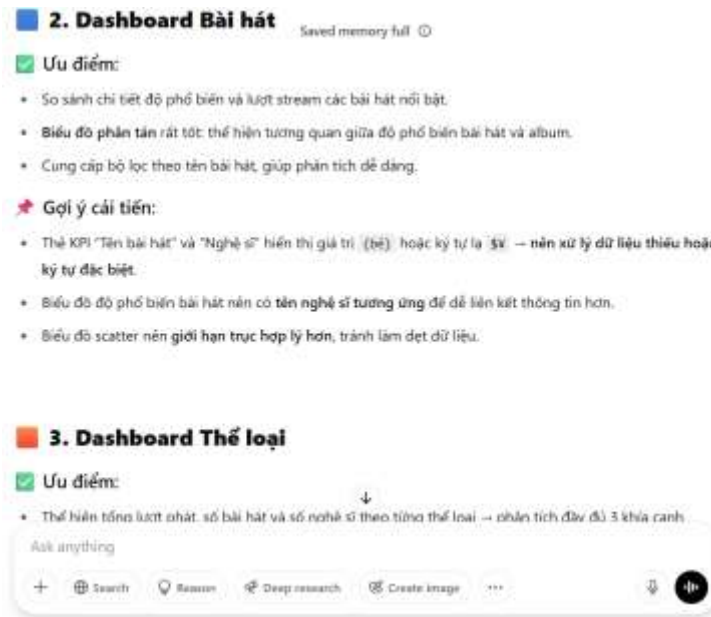
8.1. Dashboard Tổng quan

- Do ở phần bên trái, dashboard đã có filter thể loại vì thế nhóm sẽ không chỉnh sửa thêm phần tooltip.
- Và đường trung bình thời lượng, nhóm sẽ không để vào do sẽ gây khó nhìn trong lúc trực quan hóa.



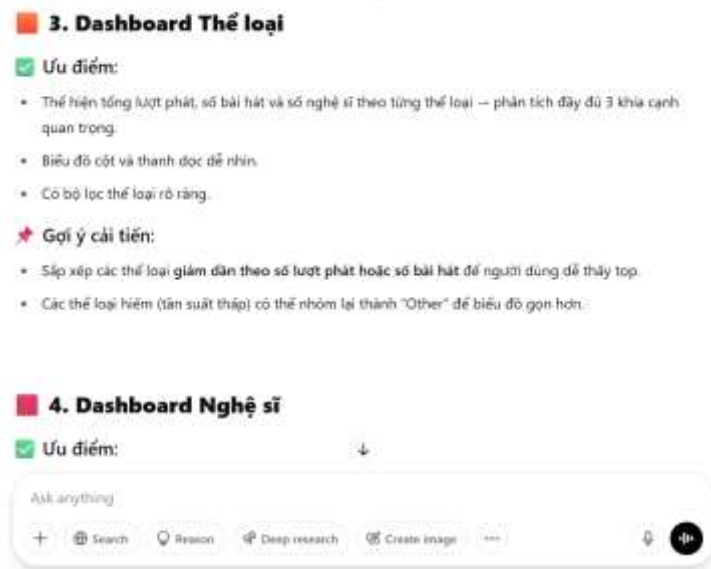
8.2. Dashboard Bài hát

- Ở các tên bài hát các ký tự đặc biệt \$ là các tên của nghệ sĩ nên nhóm em cũng sẽ không xử lý tên có ký tự đặc biệt.



8.3. Dashboard Thể loại

- Nhóm đã điều chỉnh lại sắp xếp các thể loại giảm dần theo tổng lượng phát.



8.4. Dashboard Nghệ sĩ

- Như đã đề cập phía trước tên nghệ sĩ không có ký tự nào bị lỗi vì thế nhóm em vẫn giữ nguyên tên.

- Bảng các bài hát do nhóm em muốn nhấn mạnh vào lượt phát và tên các bài hát của nghệ sĩ nên nhóm sẽ không thêm năm phát hành và thể loại để tránh cognitive overload.

4. Dashboard Nghệ sĩ

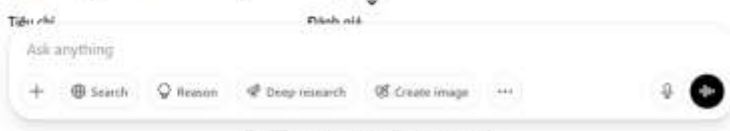
Ưu điểm:

- Cung cấp chi tiết từng nghệ sĩ: độ nổi tiếng, lượt theo dõi, lượt phát, thể loại âm nhạc liên quan.
- Biểu đồ Treemap rất hiệu quả để so sánh nghệ sĩ theo lượt theo dõi.
- Phân tích trực quan mức độ nổi tiếng từng nghệ sĩ qua biểu đồ thanh và bảng.

Gợi ý cải tiến:

- KPI "Tên nghệ sĩ" bị lỗi hiển thị (\$A Milo) → kiểm tra lại dữ liệu.
- Bảng các bài hát nên thêm cột thể loại hoặc năm phát hành để thêm chiều phân tích.
- Treemap có thể thêm tooltip có cả số liệu (hover) để cải thiện UX.

Tổng kết & Đánh giá chung:



8.5.Đánh giá tổng quát

Tổng kết & Đánh giá chung:

Tiêu chí	Đánh giá
Giao diện & màu sắc	★★★★☆ (4.5/5) - Rất tốt, đồng bộ, dễ theo dõi
Tính tương tác & bộ lọc	★★★★☆ (4/5) - Có đầy đủ lọc theo năm, thể loại, tên
Đa dạng biểu đồ	★★★★★ (5/5) - Dùng nhiều loại phù hợp: bar, donut, scatter, treemap
Thông tin & insight	★★★★☆ (4.5/5) - Có chiều sâu, đủ chi tiết
Cải thiện	Sửa lỗi dữ liệu nhỏ, tối ưu trực/trình bày

🟢 **Điểm đề xuất: 9/10** - Rất tốt cho một đồ án học thuật, có thể dùng để trình bày chuyên nghiệp.