Trường Đại học Khoa học tự nhiên, Đại học Quốc gia Thành phố Hồ Chí Minh

Ngành Công nghệ thông tin

Nhóm 4

Học phần: Trực quan hóa dữ liệu

Sinh viên:
Trần Nguyễn Nhật Cường (22127048)
Nguyễn Công Tuấn (22127436)
Trần Đăng Tuấn (22127438)

Giảng viên:
Bùi Tiến Lên
Võ Nhật Tân
Lê Ngọc Thành

Ngày 10 tháng 3 năm 2025

Mục lục

1 Thông tin chung

1.1 Thông tin nhóm

Họ tên	**MSSV**
Trần Nguyễn Nhật Cường	22127048
Nguyễn Công Tuấn	22127436
Trần Đăng Tuấn	22127438

Bảng 1: Thông tin các thành viên

1.2 Mức độ hoàn thành tổng thể của mỗi yêu cầu

Yêu cầu	**Mức độ hoàn thành (%)**
Thu thập dữ liệu	100%
Tiền xử lý dữ liệu	100%
Trực quan hóa dữ liệu	100%

Bảng 2: Mức độ hoàn thành tổng thể

1.3 Mức độ hoàn thành của từng thành viên

Công việc	**Thành viên phụ trách**	**Mức độ hoàn thành**
Thu thập dữ liệu	Nguyễn Công Tuấn	100%
Tiền xử lý dữ liệu	Trần Nguyễn Nhật Cường, Nguyễn Công Tuấn	100%
Phân tích cơ bản về dữ liệu	Trần Đăng Tuấn	100%
Xác định mục tiêu phân tích và lựa chọn các trường dữ liệu	Tất cả thành viên	100%
Phân tích, nhận xét và đánh giá dữ liệu trên biểu đồ thu được	Tất cả thành viên	100%
Viết báo cáo	Trần Nguyễn Nhật Cường, Nguyễn Công Tuấn	100%

Bảng 3: Mức độ hoàn thành của từng thành viên

2 Giới thiệu đồ án

2.1 Các thư viện sử dụng trong đồ án

- **csv** được dùng để xử lý tệp CSV (Comma-Separated Values) và đọc/ghi dữ liệu từ tệp CSV
- **matplotlib** được dùng để vẽ biểu đồ trong Python. Có thể dùng để vẽ các loại biểu đồ như histogram, biểu đồ tán xạ (scatter plot), biểu đồ đường (line chart) và biểu đồ cột (bar chart) ,...
- **pandas** được dùng để xử lý và phân tích dữ liệu dạng bảng và cung cấp cấu trúc dữ liệu như DataFrame và Series để dễ thao tác
- **seaborn** dùng để vẽ biểu đồ dựa trên thư viện **matplotlib** nhưng biểu đồ sẽ trông đẹp hơn và cũng hỗ trợ các loại biểu đồ thống kê như violin plot, box plot, heatmap
- **spotify** dùng để truy xuất dữ liệu từ **Spotify API** nhằm lấy thông tin bài hát, nghệ sĩ, album,... từ Spotify
- **sklearn**
  - **sklearn.cluster.KMeans** là thuật toán phân cụm K-Means giúp nhóm các nghệ sĩ dựa trên số lượng người theo dõi và độ phổ biến
  - **sklearn.preprocessing.StandardScaler** dùng để chuẩn hóa dữ liệu, giúp các đặc trưng có cùng thang đo, tránh ảnh hưởng của giá trị lớn nhỏ khác nhau
  - **sklearn.metrics.silhouette_score** dùng để đánh giá chất lượng phân cụm bằng Silhouette Score, giúp chọn số cụm tối ưu.

3 Nhiệm vụ đồ án

3.1 Thu thập dữ liệu

Dữ liệu được thu thập qua nền tảng âm nhạc **Spotify** bằng việc gọi **API** được hỗ trợ bởi **Spotify for Developers**.

Nhóm đã lấy được thông tin về những bài hát và những nghệ sĩ và được lưu vào lần lượt các file **vietnamese_songs.csv** và **artists_info.csv**.

Đối với tập dữ liệu **vietnamese_songs.csv**:

- Tập dữ liệu gồm **5995** dòng và **8** cột khi chưa được tiền xử lý
- Bảng các thuộc tính của tập dữ liệu:

Tên thuộc tính	**Kiểu dữ liệu**	**Ý nghĩa**
name	categorical	Tên bài hát
release_date	date time	Ngày bài hát được ra mắt/phát hành
album_type	categorical	Loại album mà bài hát thuộc về (ví dụ: album, single, compilation)
album_name	categorical	Tên của album mà bài hát thuộc về
artists	categorical	Tên của các nghệ sĩ tham gia trình diễn bài hát. Nếu có nhiều nghệ sĩ, tên nghệ sĩ sẽ được phân tách bằng dấu phẩy (",")
duration_ms	numerical	Độ dài (thời lượng) của bài hát tính theo đơn vị mili giây
popularity	numerical	Độ phổ biến của bài hát trên nền tảng Spotify, được tính từ 0 đến 100 (với 100 là phổ biến nhất)
spotify_url	categorical	Liên kết trực tiếp đến bài hát trên nền tảng Spotify

Bảng 4: Các thuộc tính của bài hát trên Spotify

- Mô tả dữ liệu

Statistic	Duration (ms)	Popularity
Count	5995	5994
Mean	287436.00	19.47
Std	179887.36	20.67
Min	0.00	0.00
25%	215172.00	0.00
50% (Median)	255278.00	12.00
75%	296245.50	39.00
Max	2922361.00	73.00

Bảng 5: Thống kê tóm tắt về thời lượng và mức độ phổ biến của bài hát

Nhận xét chung về Duration (ms)

- Thời lượng trung bình của một bài hát là **287.436 ms** (4,79 phút)
- Bài hát ngắn nhất có thời lượng **0 ms**, có thể là dữ liệu bị thiếu hoặc lỗi
- Bài hát dài nhất có thời lượng **2.922.361** ms (48,7 phút), cao hơn đáng kể so với trung bình
- Các phần trăm vị (**25%**, **50%**, **75%**) cho thấy hầu hết các bài hát có thời lượng từ 3,6 phút đến 4,9 phút

Nhận xét chung về Popularity

- Điểm phổ biến trung bình là **19,47** nhưng có độ lệch chuẩn là **20,67**. Điều này cho thấy điểm phổ biến có sự phân tán lớn
- Giá trị lớn nhất là **73** cho thấy ngay cả bài hát phổ biến nhất cũng không đạt mức tối đa (100)

Đối với tập dữ liệu **artists_info.csv**:

- Tập dữ liệu gồm **1473** dòng và **4** cột khi chưa được tiền xử lý
- Bảng các thuộc tính của tập dữ liệu

Tên thuộc tính	Kiểu dữ liệu	Ý nghĩa
name	categorical	Tên của nghệ sĩ
followers	numerical	Số lượng người theo dõi nghệ sĩ trên Spotify
popularity	numerical	Mức độ phổ biến của nghệ sĩ trên Spotify, dao động từ 0 đến 100 (với 100 là phổ biến nhất)
spotif_url	categorical	Liên kết trực tiếp đến trang Spotify của nghệ sĩ

Bảng 6: Các thuộc tính của nghệ sĩ trên Spotify

- Mô tả dữ liệu

Statistic	Followers	Popularity
Count	1,472	1,472
Mean	142,134.29	25.11
Std	1,135,387.08	18.21
Min	0.00	0.00
25%	161.50	7.00
50% (Median)	2,856.50	27.00
75%	30,458.75	39.00
Max	25,164,423.00	84.00

Bảng 7: Thống kê tóm tắt về số người theo dõi và mức độ phổ biến của nghệ sĩ

Nhận xét chung về Followers

- Số lượng người theo dõi trung bình là **142.134,29**, nhưng độ lệch chuẩn rất lớn (**1.135.387,08**) cho thấy có sự chênh lệch đáng kể giữa các nghệ sĩ
- Phân vị **25%** là **161,5**, trong khi phân vị **75%** là **30.458,75**, điều này cho thấy đa số nghệ sĩ có số lượng người theo dõi tương đối thấp, trong khi một số ít nghệ sĩ có lượng theo dõi rất cao

Nhận xét chung về Popularity

- Điểm phổ biến trung bình là **25,11** với độ lệch chuẩn **18,21**. Điều này cho thấy sự phân bố khá rộng về mức độ phổ biến
- Điểm phổ biến cao nhất là **84** cho thấy một số nghệ sĩ có mức độ phổ biến cao nhưng vẫn không đạt điểm tối đa (100)

3.2 Tiền xử lý dữ liệu

- **Thay đổi kiểu dữ liệu release_date**
  - Mục đích: Chuẩn hóa định dạng ngày tháng trong cột **release_date**
  - Mô tả:
    * Trước tiên, các giá trị không hợp lệ (bắt đầu bằng 0, bị thiếu hoặc dữ liệu chỉ có năm) được xác định
    * Sau đó, một hàm xử lý được áp dụng để điền thêm **-01-01** cho các giá trị chỉ có năm và loại bỏ dữ liệu lỗi
    * Cuối cùng, cột **release_date** được chuyển đổi sang định dạng **datetime**

- **Xử lý các bài hát bị trùng lặp**
  - Mục đích: Xử lý các bài hát trùng lặp trong tập dữ liệu dựa trên tên bài hát và nghệ sĩ
  - Mô tả:
    * Đầu tiên, tìm hiểu đếm xem số lượng bài hát trùng lặp
    * Sau đó, dữ liệu được sắp xếp theo **release_date** và các bài hát trùng lặp dựa trên **name**, **artists**, **album_type** và giữ lại phiên bản phát hành mới nhất
- **Xử lý các dữ liệu dòng bị thiếu dữ liệu**
  - Mục đích: Xử lý các dòng bị thiếu dữ liệu trong tập dữ liệu
  - Mô tả:
    * Đầu tiên, đếm số lượng giá trị bị thiếu (**NaN**) trong từng cột
    * Sau đó, kiểm tra số dòng bị ảnh hưởng đế loại bỏ tất các các giá trị thiếu
    * Cuối cùng, các dòng chứa **NaN** được loại bỏ hoàn toàn và dữ liệu được kiểm tra lại để đảm bảo không còn giá trị thiếu
- **Xử lý các dữ liệu bài hát bất hợp lý**
  - Mục đích: Lọc và xử lý các bài hát có năm phát hành trước 2000
  - Mô tả:
    * Đầu tiên, kiểm tra một số bài hát có **release_date** trước năm 2000
    * Sau đó, các bài hát có năm phát hành 1900 (có thể là lỗi dữ liệu) được loại bỏ hoàn toàn
    * Cuối cùng, tập dữ liệu được cập nhật và kiểm tra lại kích thước sau khi xóa các dòng không hợp lệ

3.3 Trực quan hóa dữ liệu

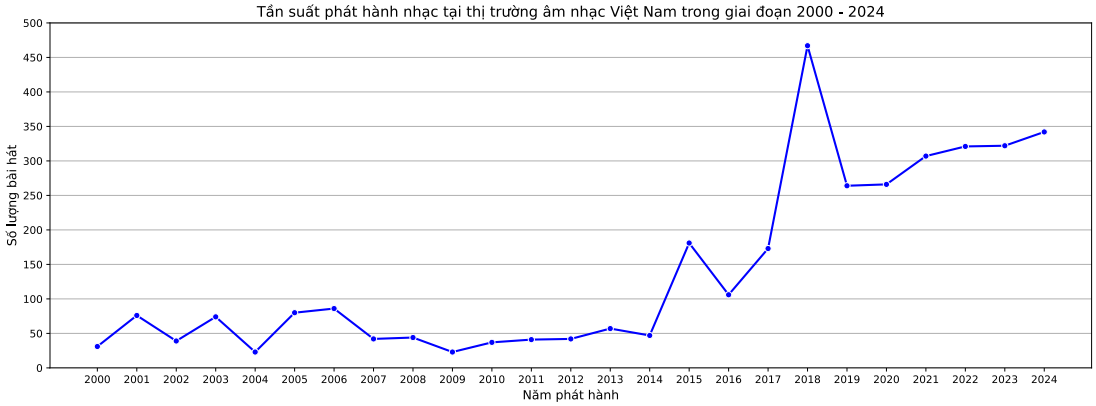
Phân tích cơ bản về dữ liệu

- **Giới thiệu về dữ liệu**: Tập dữ liệu chứa thông tin về các bài hát phát hành trên nền tảng Spotify trên thị trường âm nhạc Việt Nam bao gồm các thông tin như tên bài hát, nghệ sĩ, ngày phát hành ,... và tập dữ liệu chứa thông tin về các nghệ sĩ đã phát hành các bài hát trên Spotify bao gồm các thông tin như số người theo dõi, độ phổ biến,...
- **Cỡ mẫu và cấu trúc**: Sau khi qua bước tiền xử lý dữ liệu, tập dữ liệu đã được làm sạch và thu được **3721** dữ liệu về bài hát và **1472** dữ liệu về nghệ sĩ có ý nghĩa phân tích

Các câu hỏi

Câu 1: Tần suất phát hành nhạc tại thị trường âm nhạc Việt Nam trong 25 năm trở lại đây (từ 2000 đến 2024)?

- **Mục tiêu phân tích**: Xác định xu hướng phát hành nhạc tại thị trường Việt Nam từ năm 2000 đến 2024 để đánh giá sự thay đổi, phân phối về số lượng bài hát qua các năm, xem có sự tăng trưởng hay suy giảm nào đáng chú ý không
- **Lựa chọn trường dữ liệu**: Sử dụng 2 trường dữ liệu là **release_date** để lấy năm phát hành của bài hát và **name** để đảm bảo không có bài hát trùng lặp ảnh hưởng đến số liệu thống kê
- **Mối quan hệ giữa các trường**: Với **release_date** giúp theo dõi xu hướng phát hành theo thời gian vì nó có thể liên quan đến sự phát triển của thị trường nhạc số và xu hướng ngành công nghiệp âm nhạc
- **Loại biểu đồ sử dụng**: Biểu đồ đường (line chart)
- **Lý do chọn biểu đồ**:
  - Phù hợp để thể hiện xu hướng thay đổi theo thời gian
  - Dễ dàng quan sát sự tăng giảm của số lượng bài hát theo từng năm
- **Các bước thực hiện**:
  1. Lọc dữ liệu các bài hát phát hành từ năm 2000 đến 2024
  2. Vẽ biểu đồ đường thể hiện xu hướng phát hành nhạc theo thời gian
  3. Tùy chỉnh trục, nhãn và tiêu đề cho biểu đồ
  4. Lưu biểu đồ dưới định dạng SVG và hiển thị kết quả
  5. Viết nhận xét



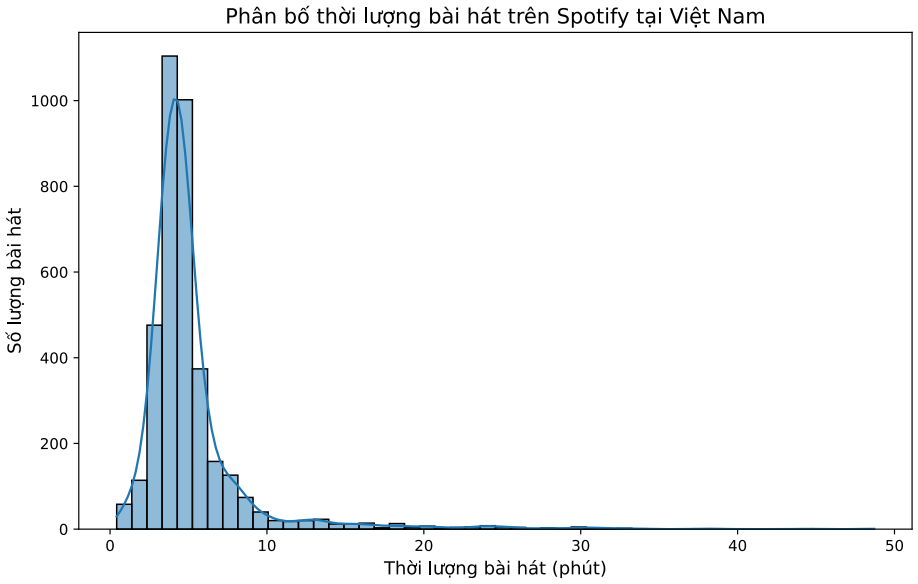
Hình 1: Tần suất phát hành nhạc tại thị trường âm nhạc Việt Nam từ 2000 đến 2024

Nhận xét

- Trong giai đoạn **2000 - 2014** thì số lượng bài hát phát hành hàng năm **khá thấp** và **không có sự tăng trưởng rõ ràng**
- Trong giai đoạn **2015 - 2018** cho thấy xu hướng phát hành nhạc **tăng đột biến** đặc biệt là **năm 2018** với số lượng bài hát đạt mức **cao nhất (450 bài)**
- **Sau năm 2018**, số lượng bài hát **giảm mạnh vào năm 2019** nhưng từ **năm 2020 trở đi**, số lượng phát hành dần **tăng trở lại** và **duy trì ổn định** ở mức cao hơn so với trước 2015
- Xu hướng phát hành nhạc tại Việt Nam đang có dấu hiệu **ổn định hơn sau giai đoạn biến động mạnh từ 2018 - 2019**

Câu 2: Phân bố thời lượng các bài hát trên Spotify tại thị trường âm nhạc Việt Nam như thế nào?

- **Mục tiêu phân tích**: Xác định sự phân bố thời lượng bài hát phổ biến nhất
- **Lựa chọn trường dữ liệu**: Sử dụng **duration_min**. Đây là thời lượng bài hát sau khi chuyển đổi từ milli giây sang phút bởi nó giúp hiểu rõ hơn về xu hướng của độ dài bài hát
- **Loại biểu đồ sử dụng**: Biểu đồ histogram kết hợp đường mật độ (KDE)
- **Lý do chọn biểu đồ**:
  - Biểu đồ histogram giúp thể hiện rõ số lượng bài hát theo từng khoảng thời gian và nó cho thấy xu hướng phân bố của dữ liệu
  - Kết hợp cùng với đường KDE sẽ giúp trực quan hóa dạng phân bố mượt hơn, tránh ảnh hưởng trực tiếp của số lượng bin trong historgram. Điều này sẽ giúp dễ dàng quan sát sự phân bố cũng như xu hướng của số lượng bài hát khi thời lượng tăng
- **Các bước thực hiện:**
  1. Chuyển đổi dữ liệu thời lượng bài hát
  2. Vẽ biểu đồ histogram thể hiện phân bố thời lượng bài hát
  3. Tùy chỉnh trục, nhãn và tiêu đề cho biểu đồ
  4. Hiển thị biểu đồ kết quả
  5. Viết nhận xét



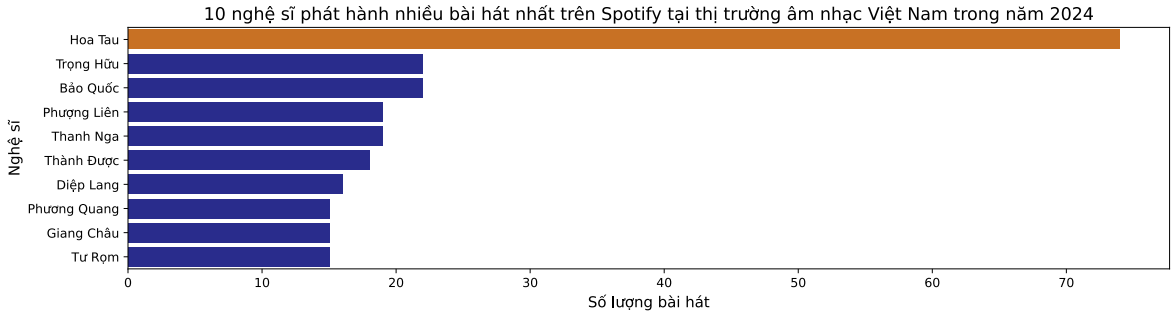
Hình 2: Phân bố thời lượng bài hát trên Spotify tại Việt Nam

Nhận xét

- **Thời lượng trung bình** của bài hát trên Spotify tại Việt Nam thường dao động trong khoảng **3 - 4 phút** và các bài hát có **thời lượng trên 10 phút là vô cùng ít**
- **Phân phối lệch phải** cho thấy phần lớn bài hát có **độ dài ngắn**, chỉ một số ít bài hát có độ dài vượt trội và xu hướng này phù hợp với xu hướng nghe nhạc trực tuyến, khi người dùng **có xu hướng ưa chuộng các bài hát có độ dài tiêu chuẩn** (3 - 4 phút) thay vì các bài hát quá dài

Câu 3: Những nghệ sĩ nào phát hành nhiều bài hát nhất trên Spotify tại thị trường âm nhạc Việt Nam trong năm 2024?

- **Mục tiêu phân tích**: Xác định nghệ sĩ nào có nhiều bài hát phát hành nhất trên Spotify trong năm 2024
- **Lựa chọn trường dữ liệu**: Sử dụng **release_date** để lọc ra các bài hát được phát hành trong năm 2024 và **artists** để thống kê số lượng bài hát theo từng nghệ sĩ
- **Mối quan hệ giữa các trường**: Với **release_date** sẽ giúp giới hạn phạm vi thời gian (chỉ xét bài hát của năm 2024) và **artists** được xử lý để đảm bảo không bỏ sót nghệ sĩ nào khi một bài hát có nhiều nghệ sĩ cùng hợp tác
- **Loại biểu đồ sử dụng**: Biểu đồ cột ngang (horizontal bar chart)
- **Lý do chọn biểu đồ**:
  - Giúp dễ dàng so sánh số lượng bài hát giữa các nghệ sĩ
  - Có thể dụng màu sắc để nhấn mạnh nghệ sĩ có số lượng bài hát cao nhất (màu cam nếu số lượng bài hát > 60, màu xanh nếu ít hơn)
  - Đơn giản và dễ đọc
- **Các bước thực hiện:**
  1. Lọc dữ liệu bài hát phát hành trong năm 2024
  2. Tách danh sách nghệ sĩ và đếm số bài hát của từng nghệ sĩ
  3. Vẽ biểu đồ barplot thể hiện nghệ sĩ có nhiều bài hát nhất
  4. Tùy chỉnh trục, nhãn và tiêu đề cho biểu đồ
  5. Lưu và hiển thị biểu đồ
  6. Viết nhận xét



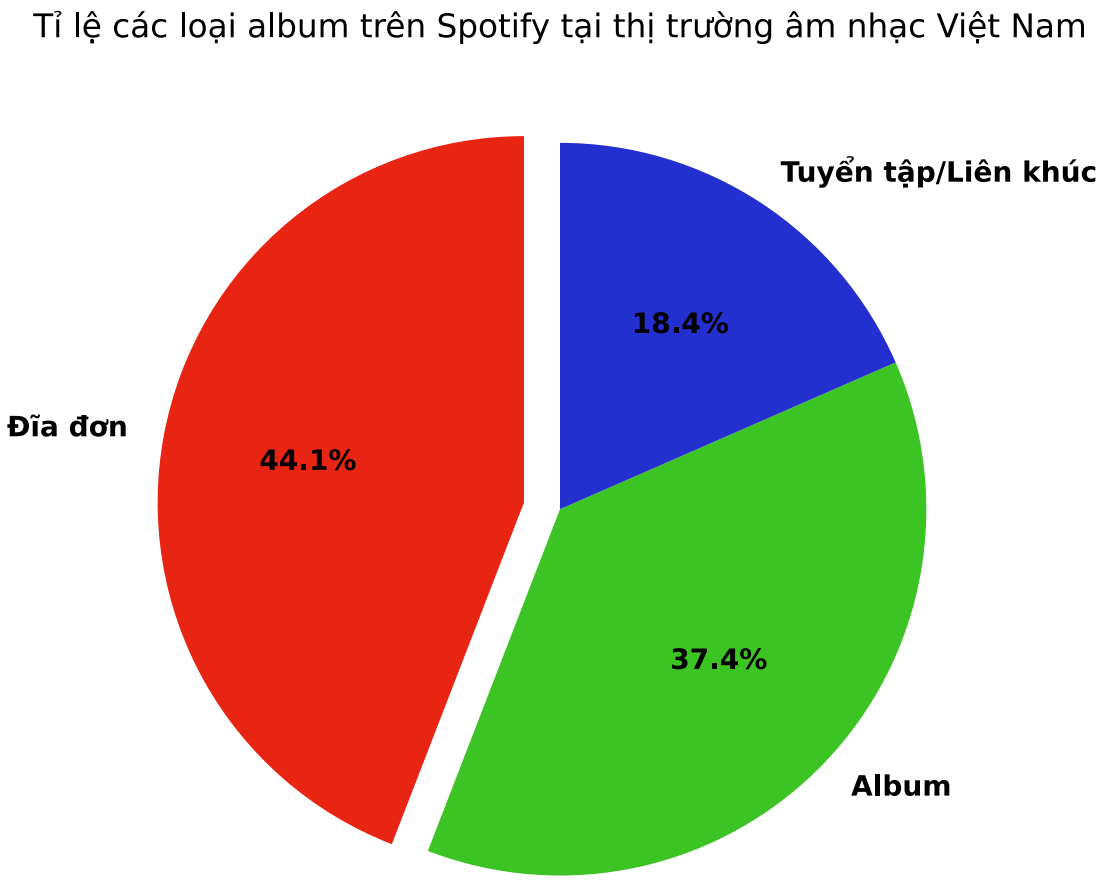
Hình 3: Top 10 nghệ sĩ phát hành nhiều bài hát nhất trên Spotify tại thị trường âm nhạc Việt Nam trong năm 2024

Nhận xét

- **Hoa Tau** có số lượng bài hát phát hành **cao nhất**, vượt xa các nghệ sĩ còn lại với **hơn 75 bài hát** và **khoảng cách giữa Hoa Tau và nghệ sĩ đứng thứ hai và ba (Trọng Hữu và Bảo Quốc) là rất lớn**
- Từ vị trí thứ 2 đến thứ 10, ố lượng bài hát **dao động trong khoảng 15 - 25 bài hát** và cũng **khá đồng đều**

Câu 4: Tỉ lệ các loại album trên Spotify tại thị trường âm nhạc Việt Nam như thế nào?

- **Mục tiêu phân tích**: Xác định xu hướng phát hành album tại Việt Nam trên nền tảng Spotify
- **Lựa chọn trường dữ liệu**: Sử dụng **album_type** chứa loại album phát hành
- **Mối quan hệ giữa các trường**: Giúp xác định loại album phổ biến nhất nhằm hiểu rõ chiến lược phát hành nhạc của các nghệ sĩ và hãng thu âm
- **Loại biểu đồ sử dụng**: Biểu đồ tròn
- **Lý do chọn biểu đồ**:
  - Phù hợp để biểu diễn tỷ lệ phần trăm của các nhóm dữ liệu
  - Dễ hiểu, trực quan khi so sánh sự phân bố giữa các loại album
- **Các bước thực hiện**:
  1. Đếm số lượng từng loại album
  2. Vẽ biểu đồ tròn (pie chart) thể hiện tỉ lệ từng loại album
  3. Tùy chỉnh tiêu đề và bố cục biểu đồ
  4. Lưu và hiển thị biểu đồ
  5. Viết nhận xét



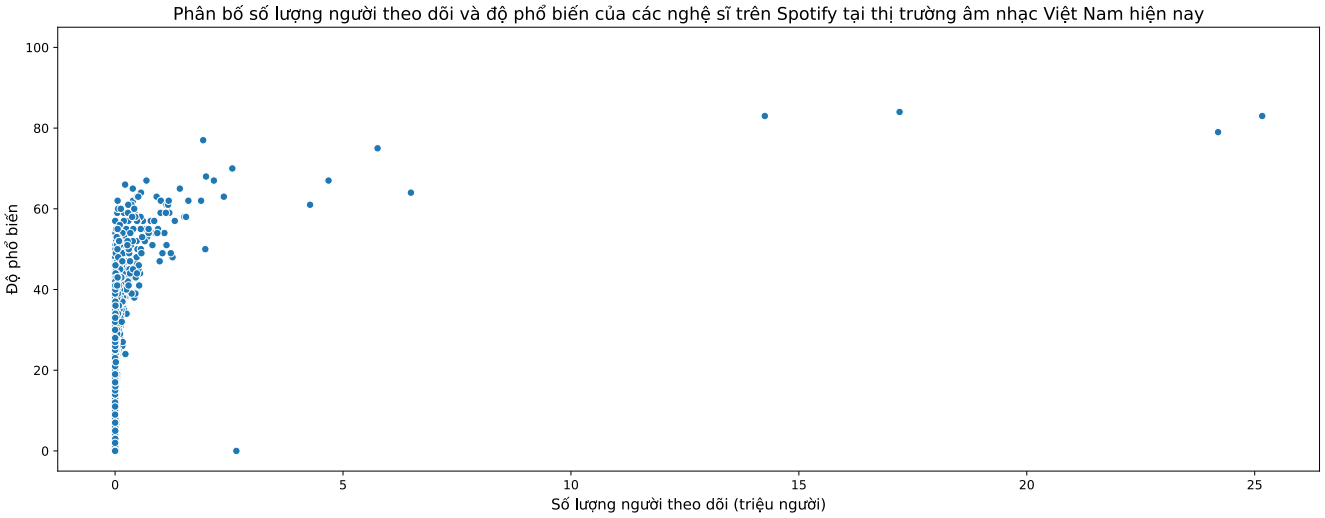
Hình 4: Tỉ lệ các loại album trên Spotify tại thị trường âm nhạc Việt Nam

Nhận xét

- **Đĩa đơn** chiếm chiếm tỷ lệ cao nhất **44.1%** phản ánh xu hướng phát hành nhạc theo từng bài hát lẻ thay vì ra mắt toàn bộ album cùng lúc. Điều này có thể do sự phát triển của nền tảng phát nhạc trực tuyến, nơi nghệ sĩ có thể tiếp cận khán giả nhanh chóng với từng ca khúc thay vì chờ hoàn thiện cả album.
- **Tuyển tập/Liên khúc** chiếm **18.4%** cho thấy tính **ít phổ biến** hơn, có thể do đặc thù của loại album này chủ yếu được phát hành theo sự kiện hoặc bởi các hãng thu âm lớn.

Câu 5: Phân bố số lượng người theo dõi và độ phổ biến của các nghệ sĩ trên Spotify tại thị trường âm nhạc Việt Nam như thế nào?

- **Mục tiêu phân tích**: Tìm hiểu xem liệu một nghệ sĩ có nhiều người theo dõi thì có đồng nghĩa với việc họ phổ biến hơn hay không
- **Lựa chọn trường dữ liệu**: Sử dụng **followers** để biết được số lượng người theo dõi của nghệ sĩ đó và **popularity** để xác định được độ phổ biến của nghệ sĩ đó
- **Mối quan hệ giữa các trường**: Đây là hai chỉ số quan trọng trong đánh giá sự thành công của một nghệ sĩ trên nền tảng nhạc số và cũng giúp nhận diện các nghệ sĩ có độ phổ biến cao nhưng lượng người theo dõi thấp, hoặc ngược lại
- **Loại biểu đồ sử dụng**: Biểu đồ tán xạ (scatter plot)
- **Lý do chọn biểu đồ**:
  - Thích hợp để phân tích mối quan hệ giữa hai biến số **followers** và **popularity**
  - Giúp phát hiện xu hướng chung và các điểm dữ liệu đặc biệt (outliers) để thấy được rõ mức độ phân bố của các nghệ sĩ trên thị trường
- **Các bước thực hiện**:
  1. Sử dụng dữ liệu **followers** và **popularity** từ bảng **artists**
  2. Vẽ biểu đồ scatter plot
  3. Tùy chỉnh biểu đồ
  4. Tinh chỉnh bố cục và lưu biểu đồ
  5. Viết nhận xét



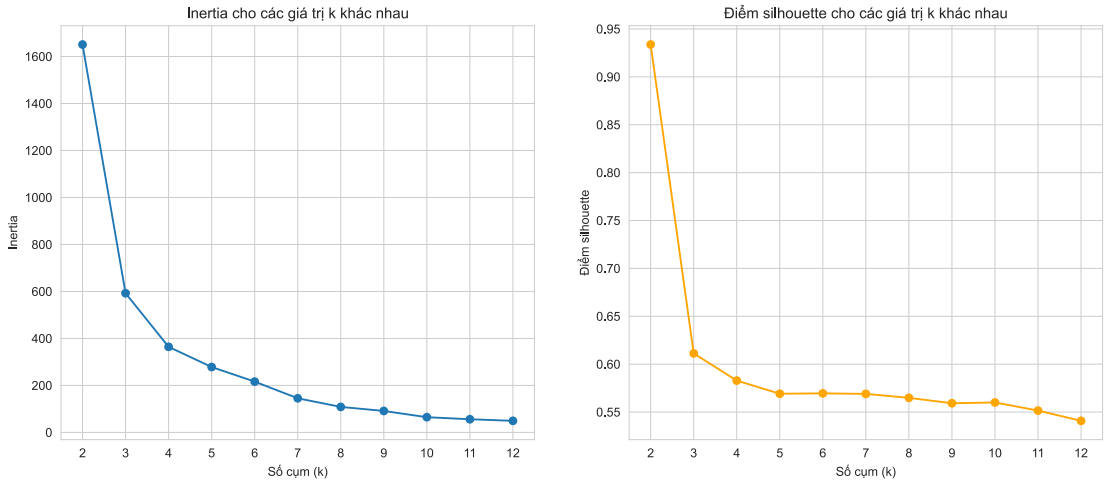
Hình 5: Phân bố số lượng người theo dõi và độ phổ biến của các nghệ sĩ trên Spotify tại thị trường âm nhạc Việt Nam

Nhận xét

- **Phần lớn nghệ sĩ** có số lượng người theo dõi **khá thấp (dưới 1 triệu)** nhưng vẫn **có độ phổ biến trung bình đến cao** bên cạnh đó là một số **ít nghệ sĩ** có lượng theo dõi **rất cao (>10 triệu)** và **độ phổ biến cũng cao**
- Có một số điểm dữ liệu **nằm rải rác** thể hiện các **nghệ sĩ có lượng người theo dõi cao** nhưng **độ phổ biến không quá nổi bật** hoặc ngược lại
- Vì thế, nhìn chung, độ phổ biến không chỉ phụ thuộc vào số người theo dõi mà còn bị ảnh hưởng bởi các yếu tố khác như tần suất ra nhạc, mức độ viral của bài hát,...

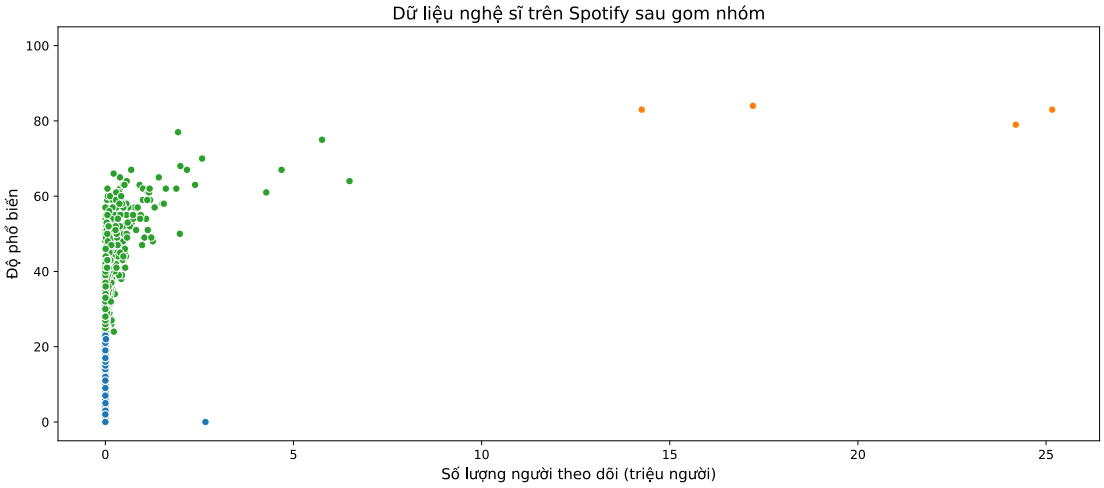
Câu 6: Làm thế nào để gom nhóm nghệ sĩ và đưa gợi ý về chiến lược phát triển cho từng nhóm

- **Mục tiêu phân tích**: Gom nhóm nghệ sĩ và đề xuất chiến lược phát triển và nhóm sử dụng thuật toán K-Means để chia nghệ sĩ thành ba nhóm dựa trên số người theo dõi và độ phổ biến
- **Loại biểu đồ sử dụng**: Biểu đồ violin plot
- **Lý do chọn biểu đồ**:
  - Giúp hiển thị không chỉ các giá trị trung vị, tứ phân vị mà còn cả hình dạng phân phối dữ liệu và điều này giúp ta dễ dàng thấy được sự khác biệt giữa các nhóm nghệ sĩ về số lượng người theo dõi và độ phổ biến
  - Có thể so sánh phân phối **followers** và **popularity** giữa các nhóm
  - Giúp nhận diện sự chồng lấn giữa các nhóm và xác định xem có nhóm nào có sự phân bố rộng hơn hay không
- **Các bước thực hiện**:
  1. Chuẩn bị dữ liệu bằng việc chọn các thuộc tính followers (số lượng người theo dõi) và popularity (độ phổ biến) để phân cụm
  2. Chuẩn hóa dữ liệu
  3. Xác định số cụm tối ưu

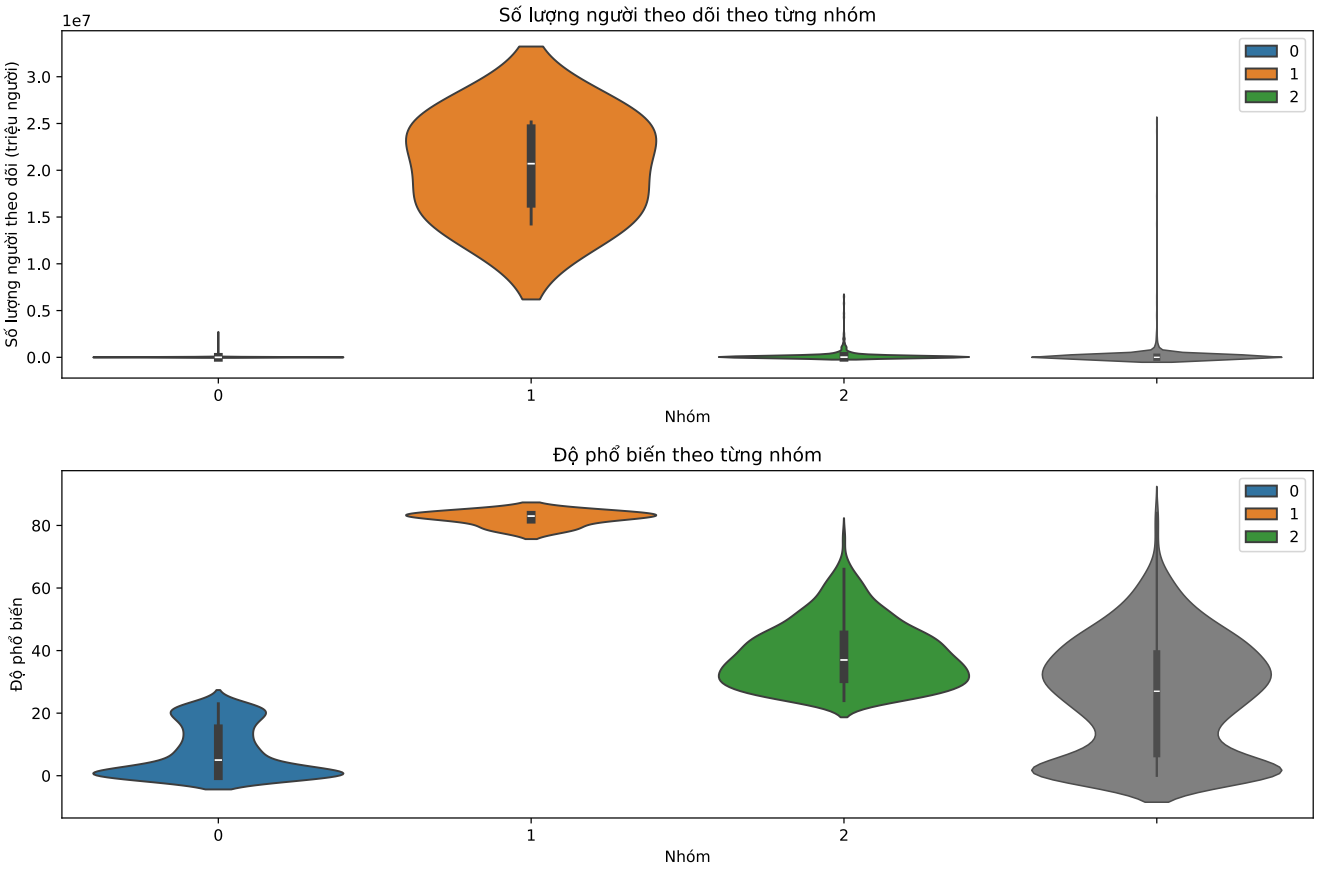


Hình 6: Kết quả inertia cho giá giá trị k khác nhau (bên trái) và kết quả điểm silhouette cho các giá trị k khác nhau (bên phải)

4. Áp dụng thuật toán K-Means với k=3 (k=3 được xác định dựa trên kết quả được thể hiện trong biểu đồ trên)
5. Minh họa phân cụm bằng biểu đồ violin plot
6. Đề xuất chiến lược phát triển cho từng nhóm



Hình 7: Dữ liệu nghệ sĩ trên Spotify sau khi gom nhóm



Hình 8: Biểu đồ gồm hai phần: Số lượng người theo dõi theo từng nhóm (trên) và Độ phổ biến theo từng nhóm (dưới)

Sau khi gom nhóm, ta chia được các nghệ sĩ vào 3 nhóm

- **Nhóm 0 (màu xanh dương)** có **số lượng người theo dõi rất thấp** phân bố tập trung gần 0 cũng như có **độ phổ biến rất thấp** phần lớn **nằm dưới 20**

  ⇒ Điều này cho thấy đây là **nhóm nghệ sĩ mới**, chưa có lượng fan đáng kể và chưa có nhiều sự chú ý từ công chúng

- **Nhóm 1 (màu cam)** sở hữu **số lượng người theo dõi rất cao**, phân bố rộng nhưng tập trung nhiều trong **khoảng từ 10 triệu đến hơn 30 triệu** người theo dõi và sở hữu **độ phổ biến rất cao**, tập trung **ở mức 80+**

  ⇒ Chứng tỏ đây là nhóm nghệ sĩ **có sức ảnh hưởng mạnh mẽ và đang dẫn đầu thị trường**

- **Nhóm 2 (màu xanh lá)** có **phân bố người theo dõi rất thấp**, một số điểm có giá trị cao đột biến nhưng đa số nằm gần 0. Điều này cho thấy nhóm này **có sự chênh lệch lớn giữa các nghệ sĩ**, có thể có **một số ít người nổi bật** nhưng **phần lớn chưa có lượng fan đáng kể**

  ⇒ Điều này cho thấy đây là **nhóm nghệ sĩ tiềm năng, một số có sức hút cao nhưng chưa đạt đến mức độ nổi tiếng như nhóm 1**

Chiến lược phát triển

- Những nghệ sĩ thuộc nhóm 1 là những nghệ sĩ nổi tiếng, có lượng fan đông đảo, độ phổ biến cao. Vì vậy có thể duy trì vị thế, mở rộng thị trường, khai thác các xu hướng toàn cầu
- Những nghệ sĩ thuộc nhóm 2 là những nghệ sĩ có tiềm năng. Dù họ có số lượng người theo dõi không quá nổi bật nhưng bù lại họ có mức độ phổ biến đa dạng. Vì thế họ có thể hợp tác với nghệ sĩ lớn, đầu tư vào quảng bá để gia tăng tầm ảnh hưởng
- Cuối cùng, những nghệ sĩ thuộc nhóm 0 là những nghệ sĩ mới do vậy họ có lượng fan ít và độ phổ biến còn thấp. Do đó, họ có thể tận dụng các nền tảng mạng xã hội để quảng cáo để thu hút sự chú ý