**ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

bdôca



**PHÂN TÍCH DỮ LIỆU ÂM NHẠC TRÊN NỀN TẢNG SPOTIFY**

|  |  |  |  |
| --- | --- | --- | --- |
| **Nhóm 7** | | | |
| Sinh viên thực hiện: | | | |
| STT | Họ tên | MSSV | Ngành |
| 1 | Cao Tiến Trung | 22521553 | KHDL |
| 2 | Trần Khôi Nguyên | 22520987 | KHDL |
| 3 | Thái Minh Lâm | 22520745 | KHDL |
| 4 | Nguyễn Duy Đức | 22520269 | KHDL |

**TP. HỒ CHÍ MINH – 12/2024**

# GIỚI THIỆU

Top Spotify Songs in 73 Countries (Daily Updated) [1] là bộ dữ liệu Top 50 bài hát thịnh hành trên mỗi quốc gia (gồm 72 quốc gia và thế giới) vẫn còn cập nhật liên tục qua từng ngày. Bộ dữ liệu được tham khảo tại Kaggle do tác giả thu thập từ Spotify API. Đề tài này tiến hành phân tích các khía cạnh của các bài hát để có cái nhìn sâu sắc hơn về xu hướng âm nhạc thay đổi theo thời gian và các bài hát phổ biến trên từng quốc gia hoặc toàn cầu. Để bổ sung thêm thuộc tính thể loại bài hát, nhóm tiến hành thu thập dữ liệu trên nền tảng Spotify. Đề tài do nhóm lên ý tưởng và thực hiện, không dùng trong các môn học khác và có tham khảo một số nguồn tài liệu liên quan cho quá trình xây dựng mô hình phân cụm và thư viện cào dữ liệu Spotipy, cụ thể nêu rõ ở phần tài liệu tham khảo.

Trong quá trình thực hiện đề tài, nhóm sử dụng một số công cụ thu thập dữ liệu trên nền tảng Spotify như Spotify API, Spotipy; công cụ hỗ trợ xử lý dữ liệu như Numpy, Pandas; công cụ trực quan hóa dữ liệu như Matplotlib, Seaborn, NetworkX và Plotly; công cụ hỗ trợ kiểm định giả thuyết như Scipy, Statsmodels và công cụ hỗ trợ các mô hình như Sklearn. Đồng thời, dựa vào kiến thức nền tảng và suy luận từ các hình ảnh trực quan dữ liệu để đưa ra kết luận. Sau khi hoàn thiện đề tài, nhóm rút ra được các đặc trưng xu hướng âm nhạc của người dùng trên Spotify, những kiến thức cơ bản về âm nhạc và sự tác động của đặc trưng lên mô hình.

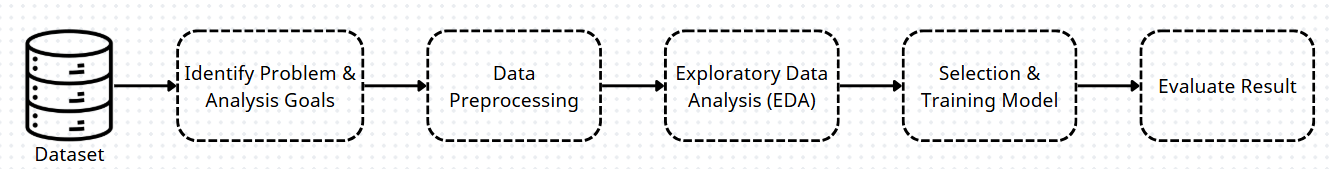
# MÔ TẢ BỘ DỮ LIỆU

Bộ dữ liệu phân tích Top Spotify Songs in 73 Countries (Daily Updated) [1] được tham khảo tại Kaggle được cập nhật mỗi ngày. Bộ dữ liệu cung cấp thông tin của nhiều bài hát lọt vào top thịnh hành theo ngày thu thập dữ liệu trên nền tảng Spotify trên từng quốc gia. Nhóm tiến hành lấy version thứ 417 (cập nhật gần nhất vào ngày 27/11/2024) để thực hiện phân tích. Ngoài ra, nhóm tự thu thập thêm thuộc tính genres nhằm hỗ trợ phân tích một cách đa dạng.

Cấu trúc bộ dữ liệu: 26 thuộc tính (cột) và 1468130 mẫu dữ liệu (dòng), trong đó có 15 biến phân loại và 11 biến số.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **STT** | **Thuộc tính** | **Kiểu dữ liệu** | **Số lượng khuyết** | **Mô tả** |
| 1 | spotify\_id | Category | 0 | Mã bài hát |
| 2 | name | Category | 29 | Tên bài hát |
| 3 | artists | Category | 28 | Tên các ca sĩ sáng tác bài hát (cách nhau bởi ‘, ’) |
| 4 | daily\_rank | Category | 0 | Xếp hạng bài hát vào ngày thu thập (giá trị từ 0 đến 50) |
| 5 | daily\_movement | Category | 0 | Số bậc xếp hạng thay đổi so với ngày trước |
| 6 | weekly\_movement | Category | 0 | Số bậc xếp hạng thay đổi so với tuần trước |
| 7 | country | Category | 20,007 | ISO code của quốc gia (Null tương ứng với Global) |
| 8 | snapshot\_date | Category | 0 | Ngày thu thập dữ liệu từ Spotify API |
| 9 | popularity | Numeric | 0 | Độ phổ biến bài hát trên nền tảng Spotify (từ 0 đến 100) |
| 10 | is\_explicit | Category | 0 | Lời tục tiễu trong bài hát |
| 11 | duration\_ms | Numeric | 0 | Thời lượng bài hát (ms) |
| 12 | album\_name | Category | 670 | Tên album |
| 13 | album\_release\_date | Category | 507 | Ngày ra mắt bài hát |
| 14 | danceability | Numeric | 0 | Chỉ số phù hợp của bài hát để khiêu vũ (từ 0 đến 1) |
| 15 | energy | Numeric | 0 | Chỉ số sôi động của bài hát (từ 0 đến 1) |
| 16 | key | Category | 0 | Hợp âm chủ đạo (từ 0 đến 11) |
| 17 | loudness | Numeric | 0 | Chỉ số âm lượng của bài hát (từ 0 đến 1) |
| 18 | mode | Category | 0 | Âm giai của bài hát (1: giai trưởng; 0: giai thứ) |
| 19 | speechiness | Numeric | 0 | Chỉ số hiện diện của lời nói trong bài hát (từ 0 đến 1) |
| 20 | acousticness | Numeric | 0 | Chất lượng âm thanh từ nhạc cụ mộc (từ 0 to 1) |
| 21 | instrumentalness | Numeric | 0 | Chỉ số bài hát không chứa lời (từ 0 đến 1) |
| 22 | liveness | Numeric | 0 | Chỉ số bài hát có yếu tố trực tiếp (từ 0 đến 1) |
| 23 | valence | Numeric | 0 | Chỉ số cảm xúc tích cực bài hát truyền tải (từ 0 đến 1) |
| 24 | tempo | Numeric | 0 | Nhịp độ bài hát (tính bằng BPM) |
| 25 | time\_signature | Category | 0 | Số chỉ nhịp ước tính của tổng thể bài hát |
| 26 | genres | Category | 181,072 | Thể loại bài hát (giá trị dạng list) |

# PHƯƠNG PHÁP PHÂN TÍCH

*Hình 1. Quy trình phân tích dữ liệu*

## Xác định vấn đề và mục tiêu phân tích

Nhóm tiến hành xác định mục tiêu cần đạt được và các vấn đề cần giải quyết khi phân tích bộ dữ liệu bài hát được chọn lọc trên Kaggle. Để củng cố thêm thông tin trong quá trình phân tích dữ liệu bài hát, nhóm tiến hành thu thập thể loại âm nhạc của các bài hát trên Spotify bằng Spotify API.

## Tiền xử lý dữ liệu

Nhóm tiến hành quan sát và đánh giá vấn đề nhằm tìm ra phương pháp làm sạch tốt nhất, bao gồm các bước:

* **Xử lý dữ liệu bị thiếu**: Quan sát dữ liệu và đưa ra phương pháp giải quyết.
* **Chuyển kiểu dữ liệu:** Chuyển các đặc trưng về đúng kiểu dữ liệu.
* **Chuyển đổi / Chuẩn hóa dữ liệu:** Thay thế các giá trị phân loại kiểu số thành mô tả cụ thể giúp quan sát dễ dàng trong quá trình phân tích. Thu hẹp phạm vi giá trị hoặc xác định phân phối của các thuộc tính để đồng bộ với các thuộc tính khác.

## Phân tích thăm dò dữ liệu

Nhóm tiến hành phân chia dữ liệu thành nhiều khía cạnh, sau đó tiến hành trực quan hóa dữ liệu và thực hiện kiểm định giả thuyết để tìm ra mối quan hệ tìm ẩn giữa các thuộc tính và các mẫu dữ liệu làm cơ sở phát triển mô hình.

## Lựa chọn mô hình và huấn luyện

Dựa vào bộ dữ liệu, đặc điểm của bài toán và các phân tích trong quá trình trước đó để lựa chọn các mô hình phù hợp.

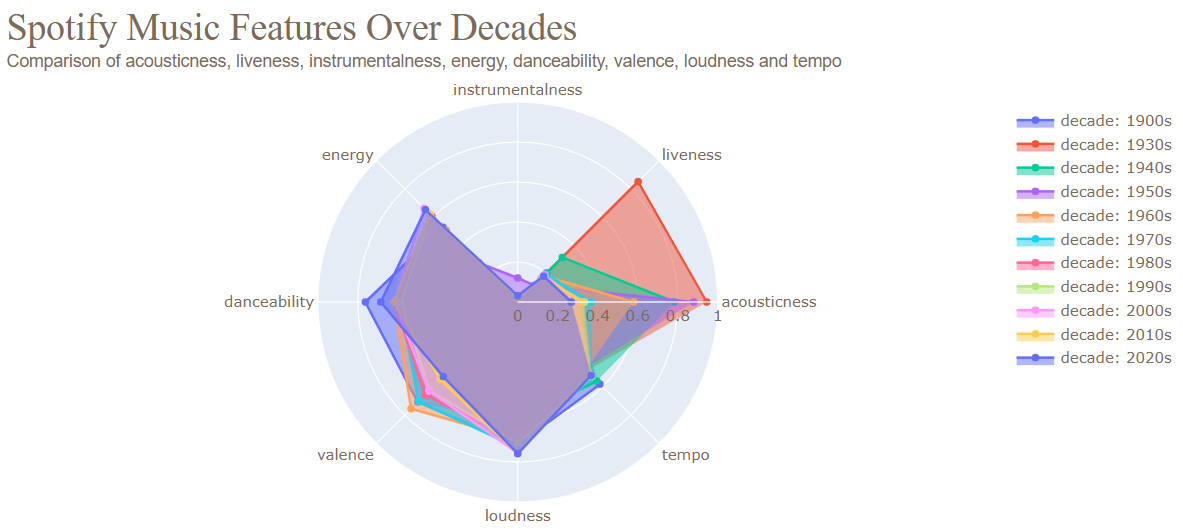
## Đánh giá kết quả

Tìm ra các đặc trưng của dữ liệu tác động đến kết quả của mô hình.

# PHÂN TÍCH THĂM DÒ/SƠ BỘ

Nhóm tiến hành phân chia thuộc tính bộ dữ liệu thành nhiều khía cạnh khác nhau để phân tích nhằm rút trích ra được những tiềm ẩn sâu sắc của bộ dữ liệu.

## Xu hướng nghe nhạc của người dùng dựa trên thuộc tính âm nhạc

*Hình 2: Giá trị trung bình thuộc tính âm nhạc qua từng thập kỷ*

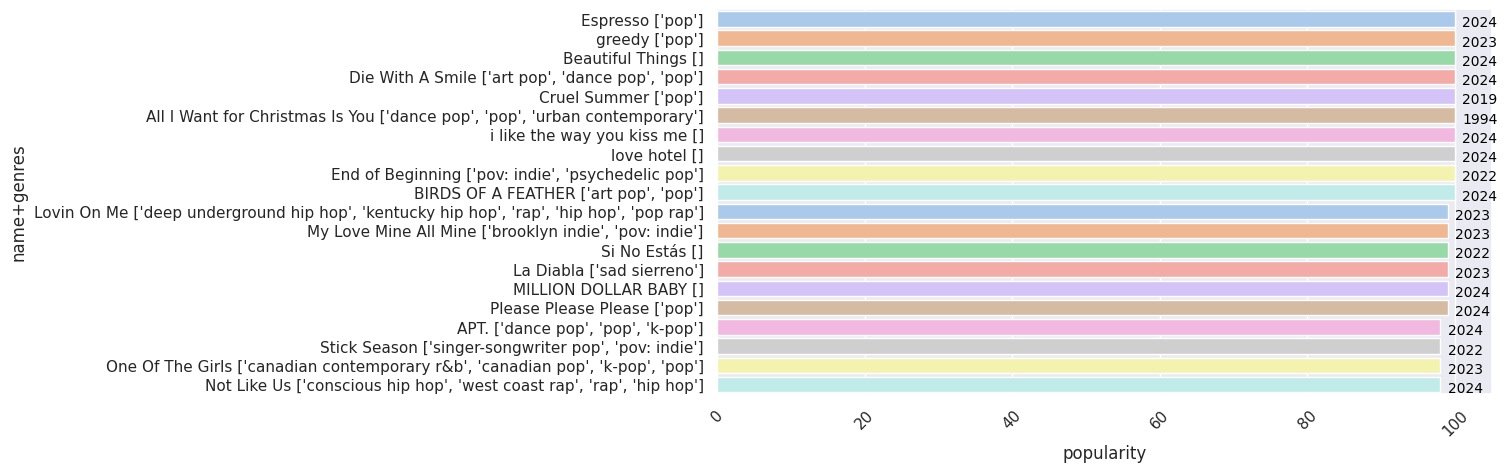
Trong giai đoạn 1930s – 1950s, các bài hát có chỉ số acousticness cao vượt trội hơn so với các bài hát ở thập kỉ khác, điều này cho thấy thị trường âm nhạc vào giai đoạn này phổ biến với các bản nhạc được sản xuất bằng nhạc cụ tự nhiên (guitar, trống truyền thống, ...). Đặc biệt, chỉ số liveness của các bài hát trong những năm 1930 vượt trội hoàn toàn, điều này cho thấy bản nhạc sáng tác được trình diễn trực tiếp trước khán giả, ít sử dụng phòng thu.

Trong giai đoạn 1900s và 1960s cho đến hiện nay, các bài hát có chỉ số energy, danceability cao hơn so với giai đoạn trên, điều này cho thấy xu hướng âm nhạc sôi động, phù hợp khiêu vũ đã dần quay trở lại khi các nhạc cụ điện tử dần ra đời và phát triển thịnh hành như ngày nay. Chỉ số valence của các bài hát giảm dần trong giai đoạn này dẫn tới xu hướng thị trường âm nhạc dần xuất hiện các bản nhạc u sầu, buồn bã (ít tích cực về cảm xúc).

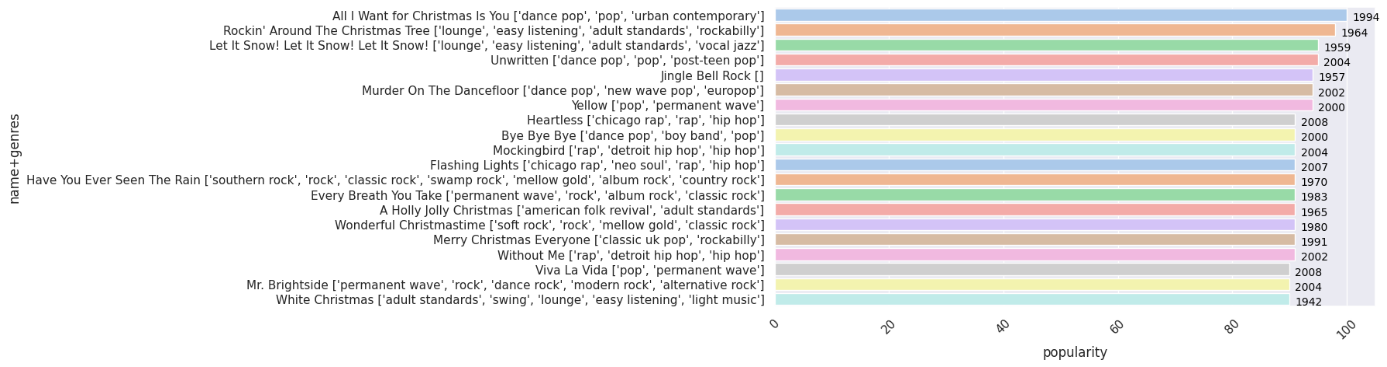
Trong xuyên suốt các thập kỷ cho tới nay, hầu như các bản nhạc được sáng tác theo hướng có lời nhạc được người dùng ưu thích (chỉ số instrumentalness không có sự khác biệt đáng kể) và chỉ số nhịp độ, âm lượng bài hát không có sự khác biệt quá lớn.

**Kết luận**: Cho đến hiện nay, người dùng vẫn ưu thích các dòng nhạc cổ điển, trình diễn trực tiếp bằng nhạc cụ truyền thống (các bài hát sáng tác trong giai đoạn 1930s – 1950s). Đa số người dùng trên nền tảng Spotify được tiếp cận bởi giới trẻ nên các bài hát mang đậm chất sôi động, khiêu vũ trở nên thịnh hành. Các bài hát ít lời hoặc không lời ít được tiếp cận bởi người dùng trên nền tảng Spotify.

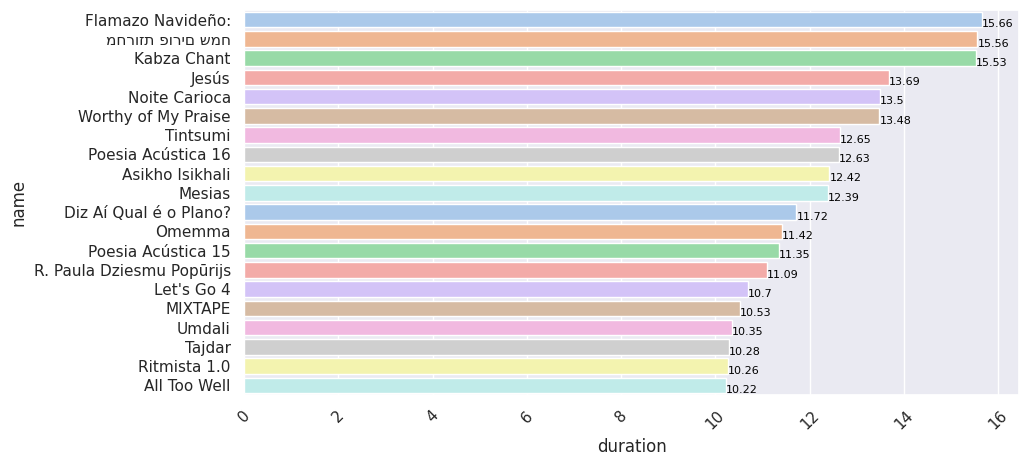
## Thể loại của bài hát phổ biến với người dùng và thời lượng bài hát

*Hình 3: Thể loại các bài hát được chia sẻ nhiều nhất cho đến nay*

Hầu hết các bài hát được chia sẻ nhiều nhất (chỉ số popularity cao nhất) đều được sáng tác trong những năm gần đây và hầu hết đều là thể loại nhạc pop và hiphop. Điều này cho thấy pop và hip hop là thể loại nhạc xu hướng của thời điểm hiện tại.

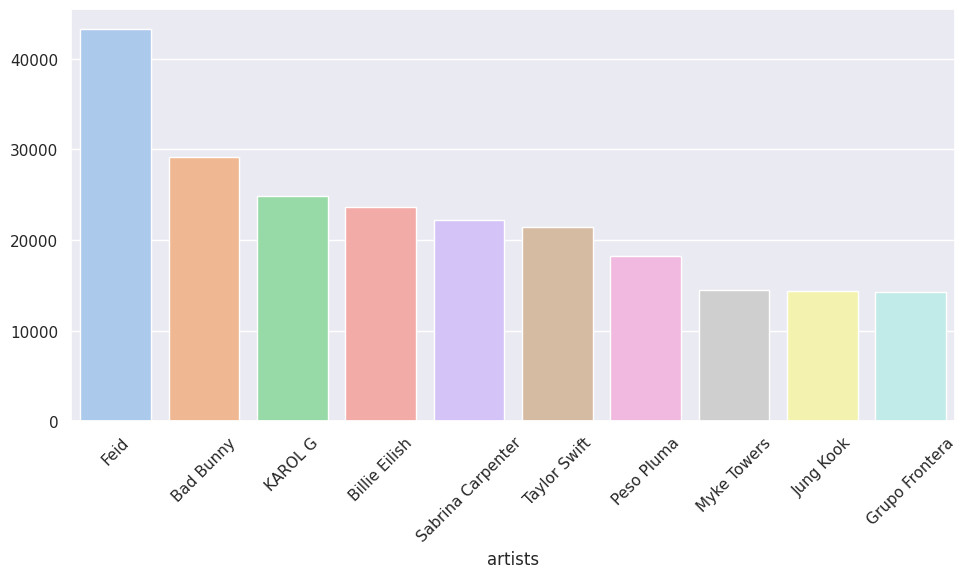
*Hình 4: Thể loại các bài hát được chia sẻ nhiều nhất trước năm 2010*

Trước năm 2010, những bài hát được chia sẻ nhiều nhất không phải các bài hát được sáng tác tại thời điểm đó mà là các bài hát những năm 2000 và các bài hát được sáng tác cách đó rất lâu. Một điều đặc biệt hơn, hầu hết các bài hát này đều liên quan đến chủ đề Giáng Sinh mặc dù bộ dữ liệu được thu thập xuyên suốt từ năm ngoái đến tận năm nay. Điều này cho thấy dù không phải dịp lễ Giáng Sinh nhưng các bài hát liên quan đến chủ đề này rất được ưa chuộng.

*Hình 5: Top 20 bài hát có thời lượng dài nhất*

Mặc dù các bài hát có thời lượng khá dài nhưng vẫn lọt vào bảng xếp hạng các bài hát thịnh hành trên các quốc gia. Các bài hát này chủ yếu là các bài hát mang màu sắc tươi mới, vui vẻ giúp người nghe thư giản, nâng cao tinh thần cho người nghe và xuất hiện bài hát ít lời.

## Độ nổi tiếng của ca sĩ sáng tác

*Hình 6: Top 10 ca sĩ lọt vào bảng xếp hạng nhiều nhất*

Để xem xét độ nổi tiếng của ca sĩ, nhóm tiến hành tính số lần lọt vào bảng xếp hạng của tất cả các bài hát của các ca sĩ. Những ca sĩ này nổi tiếng không chỉ bởi số lượng bài hát sáng tác (nhiều bài hát lọt vào bảng xếp hạng thịnh hành) mà còn là chất lượng của bài hát (bài hát trụ trong bảng xếp hạng khá lâu). Đặc điểm chung của các ca sĩ này đều sáng tác dòng nhạc chủ đạo là pop, rap, trap và hiphop - một dòng nhạc sôi động thịnh hành với đa số giới trẻ hiện nay. Bên cạnh đó, bài hát của những ca sĩ này tạo nên xu hướng trên các nền tảng mạng xã hội lớn như, điều này dẫn đến nhiều người biết đến ca sĩ và các sáng tác của họ.

## Mối liên hệ giữa giọng nhạc và các thuộc tính âm nhạc

A graph of different colored bars

Description automatically generated

*Hình 7: Tần suất giọng nhạc được dùng trong các bài hát*

Giọng C# – D♭ trưởng tuy là một giọng khá khó để chơi trên một số nhạc cụ (do có nhiều dấu hóa), song lại phù hợp với quãng giọng của đại đa số ca sĩ. Giọng B thứ phổ biến nhất vì dù đem lại cùng cảm giác, giọng này lại mang đến sự mới lạ so với giọng A thứ mà người nghe đã quen thuộc.

A graph of different colored lines

Description automatically generated

*Hình 8: Phân phối thuộc tính của bài hát theo giọng trưởng*

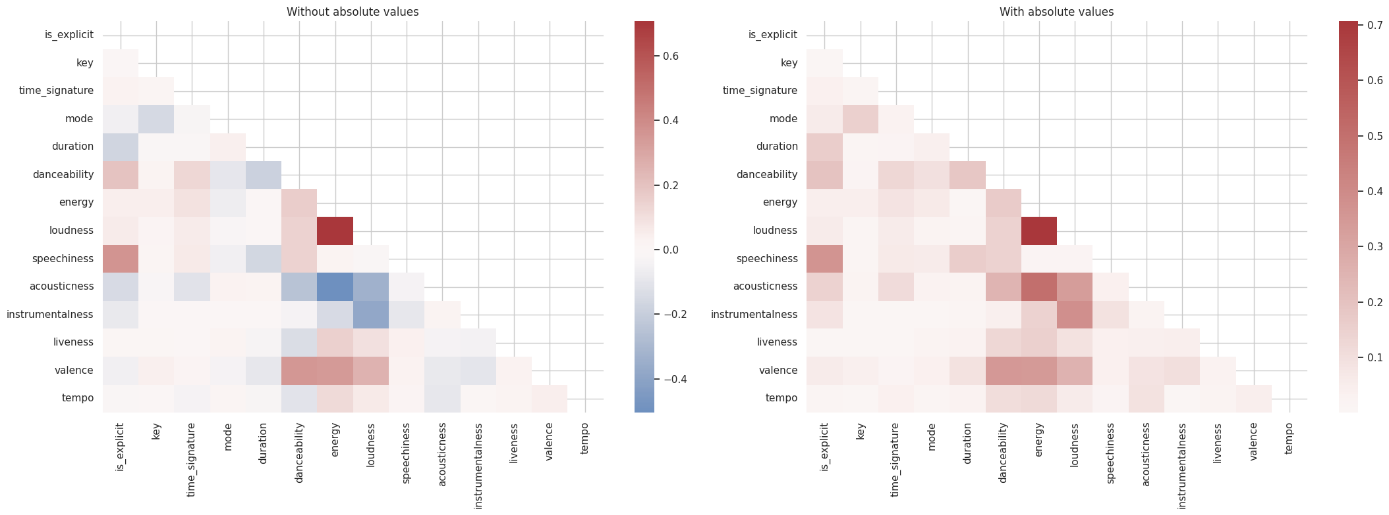
A graph of different colored lines

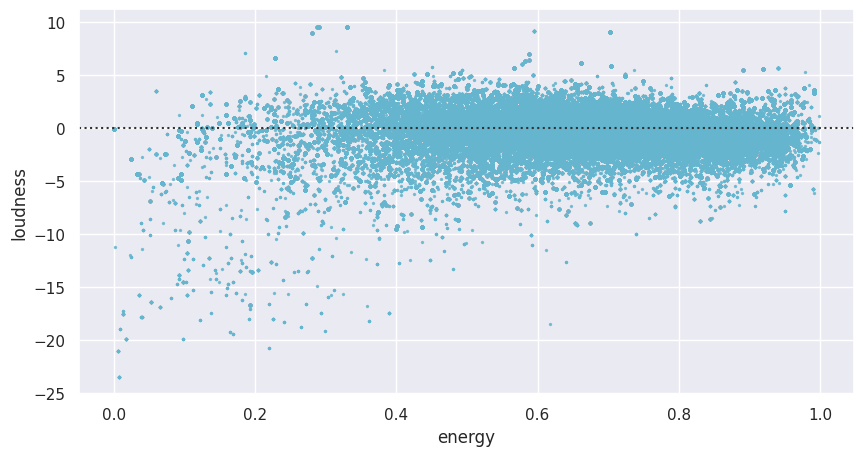
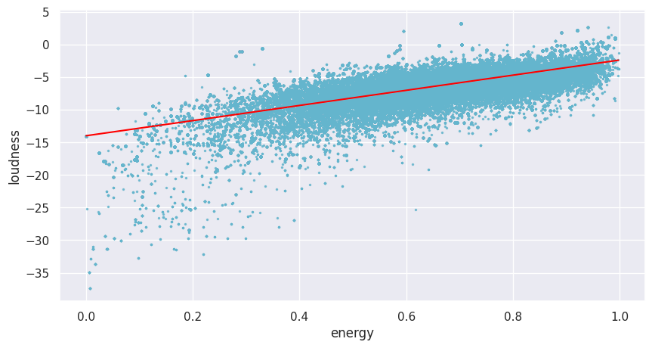
Description automatically generated

*Hình 9: Phân phối thuộc tính của bài hát theo giọng thứ*

Về mặt lý thuyết, các bài hát mang giọng trưởng thường cho cảm xúc tích cực hơn, do đó sẽ có các chỉ số cảm xúc tích cực như danceability, energy cao và chỉ số valence thấp; các bài hát giọng thứ sẽ mang các chỉ số ngược lại ở trên. Dù vậy, trong bộ dữ liệu này, các bài nhạc giọng thứ lại có chỉ số danceability và energy cao hơn các bài giọng trưởng. Điều này có thể bắt nguồn từ việc các bài này sử dụng một vòng hòa thanh có nhiều hợp âm mang năng lượng vui tươi.

## Tương quan giữa các thuộc tính trong bộ dữ liệu

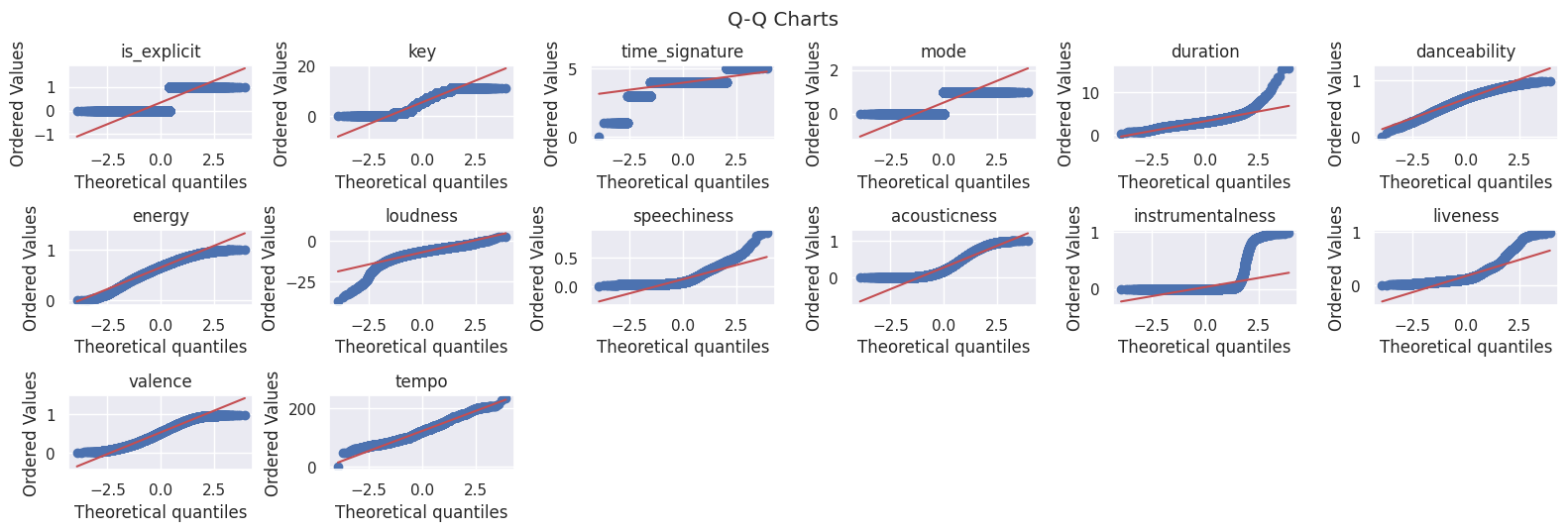
*Hình 10: Tương quan tuyến tính giữa các thuộc tính*



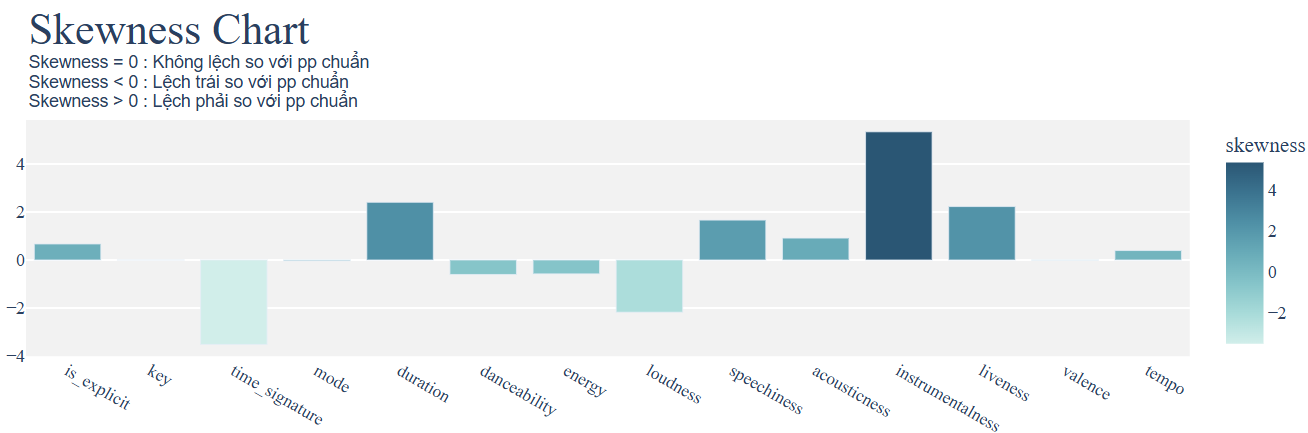
*Hình 11: Redidplot và regplot của energy đối với loudness*

Cặp thuộc tính (loudness, energy) có khả năng tương quan tuyến tính với nhau.

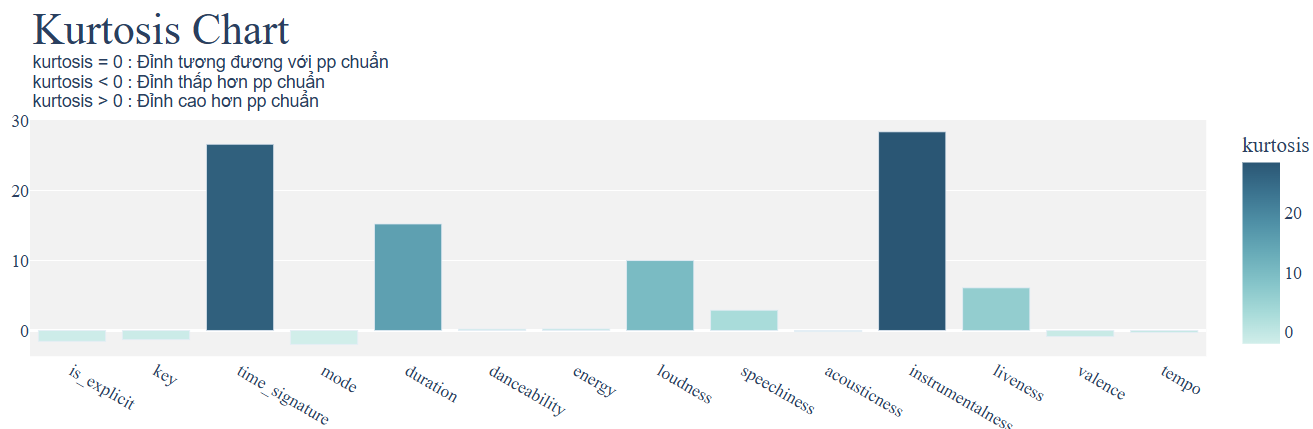
## Phân phối của các thuộc tính trong bộ dữ liệu

*Hình 12: Biểu đồ Q-Q của các thuộc tính*

Chỉ có các thuộc tính tempo, danceability, energy là tương đối gần phân phối chuẩn. Các thuộc tính còn lại không tuân theo phân phối chuẩn.

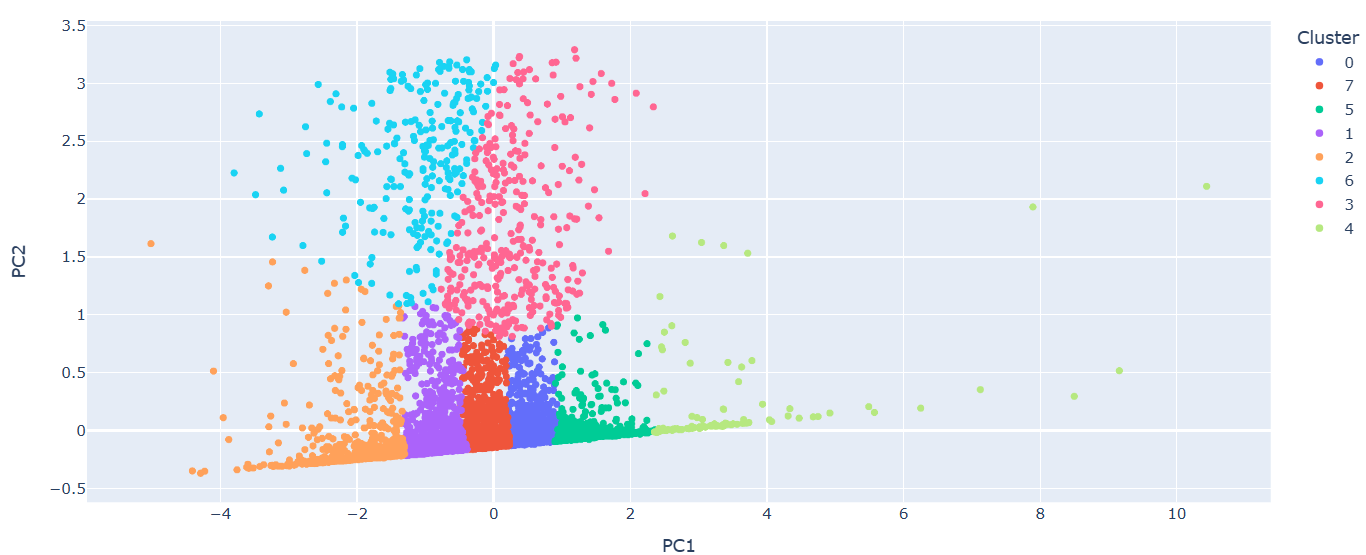
*Hình 13: Skewness của các thuộc tính*

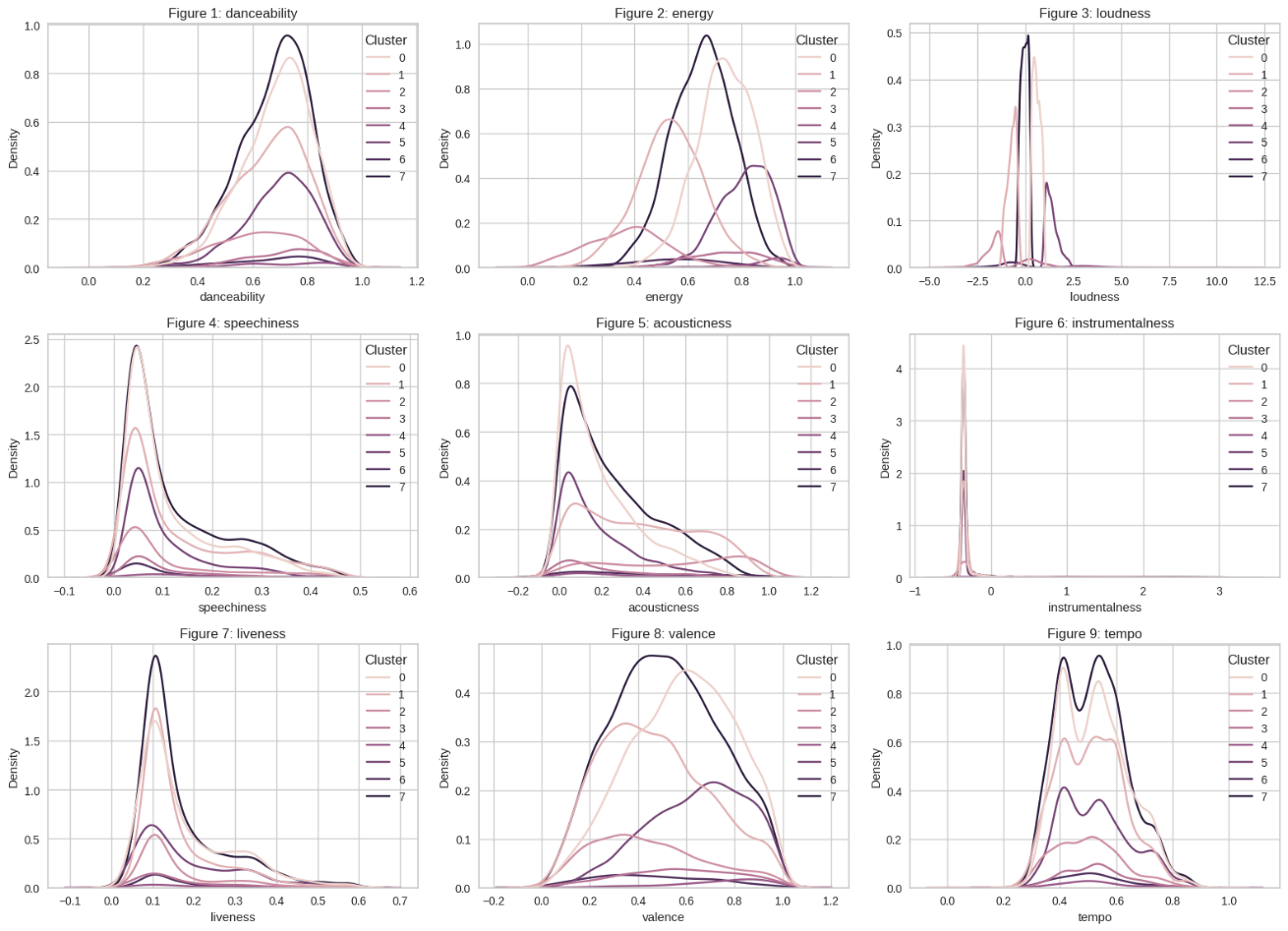
Thuộc tính time\_signature, instrumentalness, liveness, duration, loudness lệch nhiều so với phân phối chuẩn.

*Hình 14: Kurtosis của các thuộc tính*

Thuộc tính time\_signature, instrumentalness, duration, loudness có đỉnh cao hơn, nhọn hơn so với phân phối chuẩn.

# KẾT QUẢ PHÂN TÍCH

*Hình 15: Trực quan kết quả phân cụm trong không gian hai chiều*



*Hình 16: Phân phối thuộc tính theo từng cụm*

Sau khi tiến hành quá trình tiền xử lý, thực hiện loại bỏ các giá trị ngoại lai đa biến, đơn biến; cũng như dùng các phép biến đổi dữ liệu đối với các thuộc tính lệch hơn nhiều, dẹt hơn nhiều hay nhọn hơn nhiều so với phân phối chuẩn thì nhóm tiến hành quá trình phân cụm. Sau khi thu được kết quả phân cụm, nhóm tiến hành phân tích những thuộc tính tác động mạnh mẽ đến kết quả phân cụm. Theo như quan sát, thuộc tính valence và energy tác động mạnh mẽ đến kết quả của mô hình phân cụm Kmeans.

# HỆ THỐNG GỢI Ý NHẠC

Cuối cùng, nhóm tiến hành xây dựng một hệ thống gợi ý bài hát âm nhạc đơn giản nhằm mang lại trải nghiệm nghe nhạc cá nhân hóa và hiệu quả hơn cho người dùng. Hệ thống được thiết kế dựa trên các tri thức mà nhóm đã khám phá, kết hợp với mô hình phân cụm mà nhóm đã xây dựng. Bằng cách phân tích dữ liệu các thuộc tính âm nhạc, hệ thống giúp đề xuất các bài hát phù hợp với sở thích cá nhân của người dùng.

# KẾT LUẬN

Trong quá trình làm việc, nhóm đã xây dựng thành công một bộ dữ liệu về âm nhạc trên nền tảng Spotify bằng cách thu thập và tích hợp dữ liệu từ nhiều nguồn khác nhau, đồng thời thực hiện tiền xử lý nhằm khắc phục các vấn đề còn tồn đọng, đảm bảo tính chính xác và đáng tin cậy cho việc phân tích. Bên cạnh đó, nhóm đã tiến hành nghiên cứu và phân tích sâu các xu hướng âm nhạc qua từng giai đoạn cũng như phân phối của từng yếu tố, từ đó khám phá ra những giá trị tri thức ẩn chứa trong bộ dữ liệu. Dựa trên những phát hiện này, nhóm đã triển khai thành công phương pháp phân cụm các bài hát một cách hiệu quả, mở ra tiềm năng phát triển một hệ thống gợi ý âm nhạc thông minh dành cho người dùng.

**TÀI LIỆU THAM KHẢO**

1. <https://www.kaggle.com/datasets/asaniczka/top-spotify-songs-in-73-countries-daily-updated/data/version/417>
2. <https://www.kaggle.com/code/gireeshs/complete-guide-to-clustering-techniques>
3. <https://www.kaggle.com/code/javigallego/outliers-eda-clustering-tutorial>
4. <https://spotipy.readthedocs.io/en/2.24.0/>

**PHỤ LỤC PHÂN CÔNG NHIỆM VỤ**

|  |  |  |
| --- | --- | --- |
| **STT** | **Thành viên** | **Nhiệm vụ** |
| 1 | Thái Minh Lâm | Làm sạch dữ liệu, Phân tích mô hình |
| 2 | Cao Tiến Trung | Phân tích EDA, Crawl dữ liệu |
| 3 | Trần Khôi Nguyên | Viết báo cáo, Xử lý ngoại lai |
| 4 | Nguyễn Duy Đức | Phân tích EDA, thiết kế slides |