

DẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC BÁCH KHOA
KHOA KHOA HỌC VÀ KỸ THUẬT MÁY TÍNH



Đồ án tổng hợp - hướng kỹ thuật dữ liệu (CO3127)

Report tuần 40 Thu thập và phân tích dữ liệu của ứng dụng giải trí Spotify

Giảng viên hướng dẫn: GV.Dương Huỳnh Anh Đức

Họ tên SV	MSSV	Nhóm - Lớp
Nguyễn Minh Nhựt	2312550	2 - L02
Phạm Đình Phương Nam	2312186	2 - L02
Đoàn Mạnh Tất	2313074	2 - L02
Phạm Đức Hoài Nam	2212157	2 - L02

TP. HỒ CHÍ MINH, THÁNG 10/ 2025

Nhiệm cụ công việc

Danh sách phân công ở giai đoạn 1(tuần 37-40)

STT	Họ và tên	Tuần	Nhiệm vụ	Hoàn thành
1	Nguyễn Minh Nhựt	37	Tìm hiểu đề tài	100%
		38	Họp online, phân chia nhiệm vụ, phân tích rõ bối cảnh vấn đề của đề tài	
		39	Lấy data từ Kaggle; Dùng Spotify API để làm đầy metadata; ReccoBeats để lấy audio features; Viết report	
		40	Tiền xử lý, khám phá insight dữ liệu	
2	Phạm Đình Phương Nam	37	Tìm hiểu đề tài	100%
		38	Họp online, phân chia nhiệm vụ, phân tích rõ bối cảnh vấn đề của đề tài	
		39	Tìm hiểu nguồn data; Thu thập dữ liệu; Giải quyết vấn đề về data	
		40	Hoàn chỉnh cho thu thập, phân tích dữ liệu, demo code cho nguồn thứ 3; Viết báo cáo, chuẩn bị slide	
3	Đoàn Mạnh Tất	37	Tìm hiểu đề tài	100%
		38	Họp online, phân chia nhiệm vụ, phân tích rõ bối cảnh vấn đề của đề tài	

STT	Họ và tên	Tuần	Nhiệm vụ	Hoàn thành
		39	Tìm hiểu dữ liệu, xử lý vấn đề về dữ liệu	
4	Phạm Đức Hoài Nam	40	Viết demo phân tích dữ liệu, làm báo cáo và slide thuyết trình	
		37	Tìm hiểu đề tài	100%
		38	Hợp online, phân chia nhiệm vụ, phân tích rõ bối cảnh vấn đề của đề tài	
		39	Tiền xử lý dữ liệu; Phân tích tương quan dữ liệu	
		40	Demo code, chuẩn bị báo cáo	

Nhận xét từ giảng viên

STT	Họ và tên	Điểm số	Nhận xét
1	Nguyễn Minh Nhựt		
2	Phạm Đình Phương Nam		
3	Đoàn Mạnh Tất		
4	Phạm Đức Hoài Nam		

Mục lục

1 Giới thiệu	1
1.1 Tổng quan về Spotify	1
1.2 Vấn đề thực tế – Nhu cầu phân tích dữ liệu Spotify (xu hướng âm nhạc, gợi ý nhạc, phân tích nghệ sĩ)	2
1.2.1 Nhà nghiên cứu âm nhạc và dữ liệu	3
1.2.2 Người dùng và cộng đồng nghe nhạc:	3
1.2.3 Nghệ sĩ và hằng thu âm	3
1.2.4 Doanh nghiệp và tổ chức quảng cáo	3
1.2.5 Kỹ sư dữ liệu và nhà phát triển hệ thống	4
2 Mục tiêu đồ án	4
2.1 Thu thập và tiền xử lý dữ liệu	4
2.2 Thiết kế và xây dựng hệ thống dữ liệu	4
2.3 Xây dựng pipeline xử lý và phân tích dữ liệu	4
2.4 Phân tích và trực quan hóa dữ liệu Spotify	5
2.5 Ứng dụng và mở rộng	5
3 Tìm hiểu và phân tích đặc điểm của nguồn dữ liệu	5
3.1 Nguồn 1: Spotify Top 50 Playlist Songs	6
3.1.1 Giới thiệu dữ liệu	6
3.1.2 Bổ sung dữ liệu:	7
3.1.3 Tiền xử lý	11
3.1.4 Phân tích và tìm insight:	15
3.2 Nguồn 2:	31
3.2.1 Giới thiệu Dataset:	31
3.2.2 Data enrichment:	31
3.2.3 Data cleaning and preprocessing:	31
3.2.4 Overview and Analysis	35
3.2.5 Mục tiêu:	41
3.3 Nguồn 3: Billboard Hot 100	42

3.3.1	Giới thiệu dữ liệu	42
3.3.2	Thu thập và bổ sung dữ liệu	43
3.3.3	Phân tích dữ liệu ban đầu	44
3.3.4	Làm sạch dữ liệu (Data Cleaning)	51
3.3.5	Tạo đặc trưng (Feature Engineering)	53
3.3.6	Phân tích xu hướng (Trend Analysis)	55
3.3.7	Kết quả và Ý nghĩa	64
4	Phụ lục	65
4.1	Timeline công việc và bảng phân công nhiệm vụ	65
4.1.1	Timeline thực hiện đề tài	65
4.1.2	Phân công nhiệm vụ dự kiến sau tuần 40	67
4.2	Source code (link GitHub)	69
4.3	Tài liệu tham khảo	69

Danh sách hình vẽ

1.1	Logo Spotify	1
1.2	Spotify Wrapped	2
3.1	Đặc điểm chung của các dataset	7
3.2	Các trường sau khi thêm	9
3.3	Bảng đặc trưng	11
3.4	Đặc trưng chung	14
3.5	Track mới phát hành theo năm	15
3.6	1.2 Tuổi thọ các bài hát trên BXH	16
3.7	1.2 Độ biến động tổng quan trên BXH	17
3.8	Top 10 nghệ sĩ nổi bật nhất	18
3.9	Top 10 bài hát trụ nổi nhất	19
3.10	Top 5 bài hát trụ lâu nhất	19
3.11	Top 10 thể loại nổi bật nhất	20
3.12	Top 10 thể loại nổi bật nhất	21

3.13 Độ đa dạng theo tuần	22
3.14 Đặc trưng energy	23
3.15 Đặc trưng tempo	25
3.16 Đặc trưng valence	26
3.17 Đặc trưng danceability	27
3.18 So sánh tỷ lệ Explicit	28
3.19 So sánh tỷ lệ Explicit	30
3.20 Kiểm tra null, xác định kiểu dữ liệu và đổi các khoảng trống giữa các features	32
3.21 Kết quả từ hình 1	33
3.22 Kiểm tra, xóa các cột null, tách tên nghệ sĩ và chuẩn hóa releaseDate . . .	34
3.23 Kết quả từ hình 3	35
3.24 Ma trận tương quan giữa các features	36
3.25 Code thực thi	37
3.26 Phân phối tỉ lệ Skip rate theo Genre	38
3.27 Lượt Stream trung bình theo kiểu user và genre	38
3.28 Code thực thi	39
3.29 Top 10 nghệ sĩ theo độ phổ biến	39
3.30 Phân bố độ phổ biến của album	40
3.31 Số lượng album phát hành theo năm	40
3.32 Top 10 label phát hành nhiều album nhất	41
3.33 Biểu đồ heatmap thể hiện các giá trị thiểu trong dữ liệu	47
3.34 Biểu đồ phân bố danceability của các bài hát.	48
3.35 Top 15 thể loại âm nhạc phổ biến nhất	49
3.36 Ma trận tương quan giữa các thuộc tính âm nhạc.	50
3.37 Heatmap giá trị thiểu sau khi làm sạch dữ liệu.	52
3.38 Ma trận tương quan sau khi thêm đặc trưng	55
3.39 Số lượng bài hát mới (new entry) theo tuần trong năm 2025.	57
3.40 Phân phối số tuần tồn tại của các bài hát trên Billboard Hot 100.	58
3.41 Timeline thứ hạng của một số ca khúc đạt vị trí số 1 trên Billboard Hot 100.	59
3.42 Top 10 thể loại âm nhạc phổ biến nhất trong BXH Billboard Hot 100. . . .	60
3.43 Top 10 nghệ sĩ có nhiều ca khúc lọt vào Top 10 Billboard Hot 100 năm 2025.	61



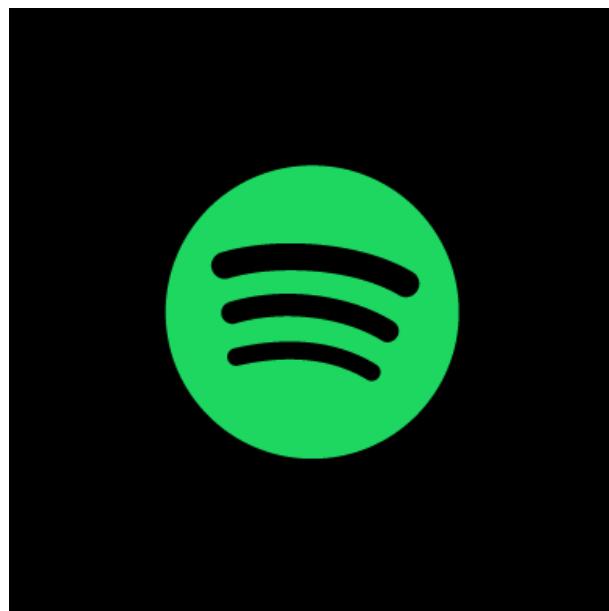
3.44 So sánh giá trị trung bình của các audio features theo mùa.	62
3.45 Mối quan hệ giữa Peak Rank và số tuần tồn tại trên BXH.	63

1 Giới thiệu

1.1 Tổng quan về Spotify

Spotify là nền tảng phát nhạc trực tuyến hàng đầu thế giới, cung cấp dịch vụ nghe nhạc, podcast và audiobook theo yêu cầu cho hàng trăm triệu người dùng trên toàn cầu. Được ra mắt lần đầu tiên vào năm 2008 tại Thụy Điển, Spotify ra đời với mục tiêu tạo ra một giải pháp hợp pháp, tiện lợi và chống lại tình trạng vi phạm bản quyền âm nhạc trong thời kỳ bùng nổ Internet. Từ đó, Spotify đã nhanh chóng phát triển, trở thành biểu tượng cho sự thay đổi cách con người tiếp cận và thưởng thức âm nhạc.

Spotify hoạt động trên nhiều nền tảng như máy tính, điện thoại thông minh, TV, và các thiết bị IoT. Ứng dụng này hiện có mặt tại hơn 180 quốc gia và vùng lãnh thổ, sở hữu hơn 100 triệu bài hát cùng hơn 5 triệu podcast. Với hơn 600 triệu người dùng hàng tháng, trong đó hơn 240 triệu thuê bao trả phí, Spotify giữ vị trí dẫn đầu trong ngành công nghiệp phát nhạc trực tuyến toàn cầu.



Hình 1.1: Logo Spotify

Không chỉ là một nền tảng âm nhạc, Spotify còn mang theo những giá trị mới mẻ về cá nhân hóa trải nghiệm. Hệ thống đề xuất thông minh dựa trên trí tuệ nhân tạo và dữ liệu lớn mang đến cho người dùng những playlist riêng biệt như Discover Weekly hay

Release Radar. Qua đó, Spotify không chỉ thỏa mãn nhu cầu giải trí mà còn kết nối nghệ sĩ với khán giả, mở rộng ảnh hưởng văn hóa và hỗ trợ sự phát triển của ngành công nghiệp âm nhạc.

Một trong những nét đặc trưng nổi bật của Spotify là các chiến dịch sáng tạo và biểu tượng văn hóa số. Ví dụ, Spotify Wrapped – bản tổng kết hàng năm – đã trở thành hiện tượng toàn cầu, nơi hàng triệu người chia sẻ thói quen nghe nhạc của mình trên mạng xã hội. Bên cạnh đó, Spotify cũng chú trọng vào trải nghiệm đồng bộ qua Spotify Connect, cho phép người dùng dễ dàng phát nhạc trên nhiều thiết bị cùng lúc.



Hình 1.2: Spotify Wrapped

Trong suốt quá trình phát triển, Spotify đã chứng kiến và góp phần tạo nên nhiều dấu ấn lịch sử của ngành nhạc số, thay đổi thói quen thưởng thức âm nhạc của cả một thế hệ. Không chỉ là nơi phát nhạc, Spotify còn là một nền tảng truyền cảm hứng, đưa âm nhạc đến gần hơn với cuộc sống hằng ngày và khẳng định sức mạnh của công nghệ trong việc kết nối con người qua âm nhạc.

1.2 Vấn đề thực tế – Nhu cầu phân tích dữ liệu Spotify (xu hướng âm nhạc, gợi ý nhạc, phân tích nghệ sĩ)

Trong kỷ nguyên số, lượng dữ liệu âm nhạc mà Spotify quản lý và tạo ra mỗi ngày là vô cùng khổng lồ: hàng tỷ lượt nghe, tìm kiếm, thêm vào playlist, chia sẻ và tương tác xã hội. Việc phân tích dữ liệu từ Spotify không chỉ phục vụ mục tiêu thương mại mà còn mang lại nhiều giá trị trong nghiên cứu, phát triển công nghệ và thậm chí là lĩnh vực văn hóa – xã hội. Một số nhóm đối tượng tiêu biểu có nhu cầu sử dụng dữ liệu này như sau:

1.2.1 Nhà nghiên cứu âm nhạc và dữ liệu

- Phân tích xu hướng nghe nhạc theo thời gian, theo khu vực địa lý hoặc theo độ tuổi.
- Nghiên cứu mối liên hệ giữa đặc điểm âm nhạc (tempo, energy, danceability) với mức độ phổ biến.
- Tạo ra các mô hình dự đoán bài hát/ nghệ sĩ có khả năng trở thành xu hướng trong tương lai.

1.2.2 Người dùng và cộng đồng nghe nhạc:

- Khám phá thói quen nghe nhạc cá nhân và so sánh với bạn bè hoặc cộng đồng.
- Tìm kiếm và gợi ý playlist phù hợp với tâm trạng, bối cảnh, hoạt động hàng ngày.
- Theo dõi các bảng xếp hạng như Top 50 Global hay Top 50 theo quốc gia để nắm bắt trào lưu âm nhạc mới.

1.2.3 Nghệ sĩ và hằng thu âm

- Phân tích dữ liệu lượt nghe để đánh giá mức độ thành công của ca khúc, album hay tour diễn.
- Nghiên cứu thị trường mục tiêu: quốc gia, độ tuổi, giới tính của người nghe.
- Tối ưu hóa chiến lược phát hành (ngày ra mắt, thể loại, hợp tác nghệ sĩ) nhằm tối đa hóa doanh thu và mức độ lan tỏa.

1.2.4 Doanh nghiệp và tổ chức quảng cáo

- Sử dụng dữ liệu hành vi người dùng để tối ưu hóa việc phân phối quảng cáo âm thanh và banner.
- Xác định nhóm khách hàng tiềm năng thông qua sở thích âm nhạc, từ đó đưa ra chiến lược tiếp thị hiệu quả hơn.

- Tận dụng dữ liệu thời gian thực để phân tích hiệu quả chiến dịch marketing (ví dụ: chiến dịch gắn liền với sự kiện âm nhạc quốc tế).

1.2.5 Kỹ sư dữ liệu và nhà phát triển hệ thống

- Nhu cầu xử lý big data với tốc độ cao, từ đó yêu cầu hạ tầng dữ liệu mạnh mẽ (Apache Kafka, Spark, Hadoop).
- Đảm bảo tính chính xác, toàn vẹn và bảo mật của dữ liệu khi có hàng trăm triệu người dùng cùng truy cập.
- Xây dựng pipeline phân tích dữ liệu streaming theo thời gian thực, phục vụ cho hệ thống gợi ý cá nhân hóa.

2 Mục tiêu đồ án

2.1 Thu thập và tiền xử lý dữ liệu

- Thu thập các dataset về Spotify từ Kaggle, API Spotify, ReccoBeats
- Làm sạch và chuẩn hóa dữ liệu: xử lý giá trị thiếu, định dạng ngày, chuẩn hóa các trường thuộc tính, loại bỏ dữ liệu dư thừa.

2.2 Thiết kế và xây dựng hệ thống dữ liệu

- Thiết kế mô hình dữ liệu quan hệ (ERD, RM) dựa trên các file CSV đã chuẩn hóa.
- Tổ chức dữ liệu vào hệ quản trị CSDL (MySQL,...) nhằm đảm bảo tính toàn vẹn, tối ưu truy vấn và dễ dàng mở rộng.
- Áp dụng các kỹ thuật Data Engineering: indexing, trigger, backup/restore, tối ưu hóa dữ liệu.

2.3 Xây dựng pipeline xử lý và phân tích dữ liệu

- Thiết kế mô hình dữ liệu quan hệ (ERD) dựa trên các file CSV đã chuẩn hóa.
- Tổ chức dữ liệu vào hệ quản trị CSDL (MySQL,...) nhằm đảm bảo tính toàn vẹn, tối ưu truy vấn và dễ dàng mở rộng.

- Áp dụng các kỹ thuật Data Engineering: indexing, trigger, backup/restore, tối ưu hóa dữ liệu.

2.4 Phân tích và trực quan hóa dữ liệu Spotify

- Phân tích xu hướng âm nhạc: sự nổi bật của ca khúc/nghệ sĩ, vòng đời của bài hát (Hot/Cold cycle), sự khác biệt giữa các quốc gia.
- Khai thác đặc trưng âm nhạc (danceability, energy, acousticness, valence, tempo, v.v.) để tìm mối liên hệ với độ phổ biến.
- Trực quan hóa dữ liệu qua dashboard (Streamlit/Matplotlib/Seaborn), cho phép người dùng tra cứu bảng xếp hạng, so sánh quốc gia, xem biểu đồ xu hướng.
- còn tìm hiểu để thay đổi hoặc mở rộng

2.5 Ứng dụng và mở rộng

- Hỗ trợ người dùng (sinh viên, nhà nghiên cứu, nghệ sĩ, doanh nghiệp) trong việc khai thác thông tin từ dữ liệu Spotify.
- Đề xuất khả năng mở rộng với Machine Learning như gợi ý bài hát, dự đoán bài hát tiềm năng sẽ vào bảng xếp hạng.
- Nâng cấp hệ thống thành mô hình phân tích dữ liệu streaming thời gian thực (Kafka + Spark) để phục vụ cập nhật liên tục.
- còn tìm hiểu để thay đổi hoặc mở rộng (chọn 1 hướng để phát triển)

3 Tìm hiểu và phân tích đặc điểm của nguồn dữ liệu

Trong quá trình thực hiện đề tài, nhóm đã tiến hành khảo sát và lựa chọn các nguồn dữ liệu liên quan trực tiếp đến lĩnh vực nghiên cứu. Mỗi nguồn dữ liệu đều có những đặc trưng riêng về cấu trúc, dung lượng, cũng như mức độ phù hợp với yêu cầu của bài toán. Việc phân tích đặc điểm của từng nguồn giúp nhóm:

- Đánh giá tính tin cậy và độ bao phủ của dữ liệu.
- Hiểu rõ cấu trúc, định dạng, và mối quan hệ giữa các thuộc tính.

- Xác định những vấn đề cần tiền xử lý (dữ liệu thiếu, dữ liệu nhiễu, trùng lặp, . . .).
- Đề ra phương án tích hợp và khai thác hiệu quả cho hệ thống.

Dưới đây là phân tích chi tiết từng nguồn dữ liệu mà nhóm sử dụng :

3.1 Nguồn 1: Spotify Top 50 Playlist Songs

- **Link dataset:** <https://www.kaggle.com/datasets/anxods/spotify-top-50-playlist-songs-anxods>
- **Link github:** <https://github.com/minhnhat273/spotify-data-project>

3.1.1 Giới thiệu dữ liệu

Tổng quan

- Bộ dữ liệu cung cấp thông tin bài hát nằm trong bảng xếp hạng Top 50 của Spotify của một số quốc gia gồm 9 nước (United States, Spain, United Kingdom, Italy, France, Mexico, Argentina, Japan, South Korea) và 1 bảng xếp hạng chung top 50 trên thế giới.
- Phạm vi dữ liệu ngoài vị trí địa lý còn về mặt thời gian bộ dataset cung cấp nằm trong khoảng 05/2023 đến 11/2024.

Một số đặc trưng ban đầu:

- **date:** Ngày thu thập dữ liệu hoặc ngày xếp hạng.
- **position:** Vị trí của bài hát trong bảng xếp hạng Top 50.
- **song:** Tên bài hát.
- **artist:** Tên nghệ sĩ hoặc nhóm nghệ sĩ chính.
- **popularity:** Chỉ số phổ biến (Spotify popularity score, 0–100).
- **duration_ms:** Độ dài bài hát tính bằng mili-giây.

- **album_type**: Loại album chứa track này (ví dụ: album, single, compilation).
- **total_tracks**: Tổng số track trong album chứa bài hát đó.
- **release_date**: Ngày phát hành album (chứa bài hát này).
- **is_explicit**: Cho biết bài hát có nội dung 18+ hay không.
- **album_cover_url**: Link ảnh bìa album.

```
Thông tin chi tiết về DataFrame:  
<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 27800 entries, 0 to 27799  
Data columns (total 11 columns):  
 #   Column           Non-Null Count Dtype  
---  --  
 0   date             27800 non-null  object  
 1   position         27800 non-null  int64  
 2   song              27800 non-null  object  
 3   artist            27800 non-null  object  
 4   popularity        27800 non-null  int64  
 5   duration_ms       27800 non-null  int64  
 6   album_type        27800 non-null  object  
 7   total_tracks      27800 non-null  int64  
 8   release_date      27800 non-null  object  
 9   is_explicit       27800 non-null  bool  
 10  album_cover_url  27800 non-null  object  
dtypes: bool(1), int64(4), object(6)  
memory usage: 2.1+ MB  
None  
  
kích thước  
(27800, 11)
```

Hình 3.1: Đặc điểm chung của các dataset

=> Kết quả có được 10 file spotify-streaming-top-50-region.csv có đặc điểm như hình trên

3.1.2 Bổ sung dữ liệu:

- Dùng Spotify API để hoàn thiện metadata

- **Giới thiệu API:** Spotify Web API là dịch vụ do Spotify cung cấp. Cho phép lập trình viên truy cập dữ liệu nhạc số trên Spotify, hoạt động qua HTTP request (RESTful API).
- **Mục đích:** Do nhóm nhận thấy các trường dữ liệu của dataset gốc chưa cung cấp đầy đủ, chi tiết về metadata của dữ liệu nên nhóm quyết định sử dụng API này để hoàn thiện nó.
- **Các trường có thể thêm gồm:**
 - **track_id:** ID duy nhất của bài hát trên Spotify.
 - **album_id:** ID duy nhất của album chứa bài hát.
 - **uri:** Định danh Spotify URI (dùng để mở trực tiếp trong Spotify).
 - **href:** Link API đến resource (track/album) trong Spotify Web API.
 - **external_url:** Link mở trên Spotify (dành cho người dùng).
 - **external_ids:** Thông tin định danh khác (ví dụ ISRC – mã nhận dạng bản ghi).
 - **disc_number:** Số đĩa (trong trường hợp album nhiều đĩa).
 - **track_number:** Vị trí bài hát trong album/dĩa.
 - **release_date_precision:** Độ chính xác của ngày phát hành (có thể là year, month, hoặc day).
 - **isPlayable:** Cho biết bài hát có thể phát được không (True/-False).
 - **linked_from:** Nếu bài hát được liên kết từ một track khác (ví dụ bản sao trong album khác).
 - **preview_url:** Link nghe thử 30 giây bài hát.
 - **restrictions:** Các giới hạn (ví dụ chỉ phát được ở một số quốc gia).

- **available_markets**: Danh sách quốc gia mà track này có sẵn.
- **genres**: Thể loại âm nhạc nghệ sĩ theo đuổi.

[5 rows x 26 columns]			
Thông tin chi tiết về DataFrame:			
<class 'pandas.core.frame.DataFrame'>			
RangeIndex: 27800 entries, 0 to 27799			
#	Column	Non-Null Count	Dtype
0	date	27800	non-null object
1	position	27800	non-null int64
2	song	27800	non-null object
3	artist	27800	non-null object
4	popularity	27800	non-null int64
5	duration_ms	27800	non-null int64
6	album_type	27800	non-null object
7	total_tracks	27800	non-null int64
8	release_date	27800	non-null object
9	is_explicit	27800	non-null bool
10	album_cover_url	27800	non-null object
11	track_id	27781	non-null object
12	album_id	27781	non-null object
13	release_date_precision	27781	non-null object
14	preview_url	0	non-null float64
15	external_url	27781	non-null object
16	is_playable	0	non-null float64
17	available_markets	27781	non-null object
18	disc_number	27781	non-null float64
19	track_number	27781	non-null float64
20	href	27781	non-null object
21	uri	27781	non-null object
22	external_ids	27781	non-null object
23	linked_from	0	non-null float64
24	restrictions	0	non-null float64
25	genres	14097	non-null object
dtypes: bool(1), float64(6), int64(4), object(15)			
memory usage: 5.3+ MB			

Hình 3.2: Các trường sau khi thêm

=> Kết quả có được các file ..with-meta.csv có đặc điểm như hình trên

- **Hạn chế**: Tuy nhiên, do một số chính sách mới ra cuối năm 2024 của Spotify dẫn đến các nhóm người không thể truy cập được một số nội dung (chart, lyric,...) trong đó có audio feature - nguồn dữ liệu mà nhóm quan tâm.
- Dùng ReccoBeats để bổ sung audio feature:

• **Giới thiệu API:** Recobeats API là một dịch vụ bên thứ ba giúp lấy dữ liệu nhạc từ Spotify (playlist, audio features, metadata) và dùng cho phân tích hoặc gợi ý nhạc. Tuy nhiên nó không phải API chính thức.

• **Các trường có thể thêm:**

- **href:** Link Spotify đến bài hát (https://open.spotify.com/track/<track_id>). Dùng để mở hoặc kiểm tra thủ công.
- **acousticness:** Mức độ “mộc” (0.0–1.0). Cao => nhạc acoustic, ít electronic.
- **danceability:** Độ dễ nhảy (0.0–1.0). Cao => dễ nhảy, dựa trên nhịp, groove, tempo.
- **energy:** Cường độ/độ bốc (0.0–1.0). Liên quan tới tempo, loudness, mật độ âm thanh.
- **instrumentalness:** Khả năng không lời (0.0–1.0). >0.5 thường là nhạc cụ/không lời.
- **key:** Tông nhạc (0–11, -1 nếu không phát hiện). Ví dụ: 0=C, 1=C#/Db, ... 11=B.
- **liveness:** Dấu hiệu biểu diễn live (0.0–1.0). >0.8 thường là bản live.
- **loudness:** Độ ồn trung bình (dB, -60 => 0). Giá trị càng gần 0 càng to.
- **mode:** Diệu thức: 1 = Major (tươi sáng), 0 = Minor (trầm buồn).
- **speechiness:** Tỷ lệ lời nói (0.0–1.0). >0.66: chủ yếu là nói; 0.33–0.66: rap/spoken; <0.33: chủ yếu là nhạc.
- **tempo:** Nhịp độ (BPM, 0–250). Nhạc phổ biến: 60–200 BPM.

- **valence**: Độ “tươi vui”/tích cực (0.0–1.0). 0.0 = buồn/tối, 1.0 = vui/sáng.

```
Thông tin chi tiết về DataFrame:  
<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 27692 entries, 0 to 27691  
Data columns (total 13 columns):  
 #   Column           Non-Null Count Dtype  
---  --  
 0   track_id        27692 non-null  object  
 1   href             26512 non-null  object  
 2   acousticness    26512 non-null  float64  
 3   danceability    26512 non-null  float64  
 4   energy           26512 non-null  float64  
 5   instrumentalness 26512 non-null  float64  
 6   key              26512 non-null  float64  
 7   liveness          26512 non-null  float64  
 8   loudness          26512 non-null  float64  
 9   mode              26512 non-null  float64  
 10  speechiness      26512 non-null  float64  
 11  tempo             26512 non-null  float64  
 12  valence           26512 non-null  float64  
 dtypes: float64(11), object(2)  
 memory usage: 2.7+ MB
```

Hình 3.3: Bảng đặc trưng

=> Kết quả có các file ..with-audio.csv có đặc điểm như hình trên

3.1.3 Tiềm xử lý

- Xử lý Metadata:

- Giữ lại một số trường cần thiết:
 - **date** — phân tích theo thời gian.
 - **position** — vị trí xếp hạng trong Top 50.
 - **song** — tên bài hát.
 - **artist** — tên nghệ sĩ.
 - **track_id** — khóa chính để join với with-audio.
 - **popularity** — đo mức độ phổ biến.

- **duration_ms** — độ dài bài hát (có thể phân tích thêm).
- **is_explicit** — xem tỷ lệ nhạc explicit.
- **album_id** — phân tích theo album (nếu cần).
- **release_date** — phân tích bài mới/cũ.
- **genres** — phân tích xu hướng thể loại.
- **Bỏ các trường không cần thiết còn lại**

- **Xử lý Missing Values (Data Cleaning):**

- **Mục tiêu:** đảm bảo dữ liệu không còn NaN ở các cột quan trọng.
- **track_id, extttalbum_id và genres :** nếu thiếu → điền "unknown_track", "unknown_album" và "unknown".
- **release_date:**
 - * Nếu chỉ có năm → ghép "YYYY-01-01".
 - * Nếu NaN → dùng "YYYY-01-01" dựa trên năm nhỏ nhất trong cột date.

- **Chuẩn hóa dữ liệu (Standardization):**

- **Mục tiêu:** đưa dữ liệu về format thống nhất để dễ phân tích.
- **date & release_date:**
 - * Convert toàn bộ sang datetime (YYYY-MM-DD).
- **genres:**
 - * Nếu có nhiều genre → lấy genre đầu tiên tạo cột main_genre.
- **is_explicit:** convert về 0/1 (boolean → int).

=> **Kết quả:** tạo được các file ...with-meta-clean.csv .

- Xử lý file audio:

• Bước 1. Kiểm tra & xử lý Missing Values:

- track_id, href: giữ nguyên (định danh và link).
- Các đặc trưng [0–1] (danceability, energy, valence, acousticness, instrumentalness, liveness, speechiness):
 - * Nếu NaN → thay bằng median (hoặc giá trị phô biến nhất).
 - * Sau đó clip về [0,1].
- tempo:
 - * Nếu NaN hoặc bất thường (< 30 hoặc > 250) → thay bằng median.
- loudness:
 - * Nếu NaN hoặc bất thường (< -60 hoặc > 0) → thay bằng median.
- mode (0/1): nếu NaN → điền giá trị chiếm đa số trong cột.
- key:
 - * Nếu -1 hoặc NaN → thay bằng giá trị gần nhất (forward/backward fill hoặc mode).

• Bước 2. Chuẩn hóa dữ liệu (Standardization):

- Các đặc trưng [0–1] (danceability, energy, valence, acousticness, instrumentalness, liveness, speechiness):
 - * Clip về [0,1], giữ nguyên cột gốc (không thêm).
- tempo:
 - * Thêm tempo_norm = (tempo - 30)/(250 - 30).
- loudness:

- * Thêm `loudness_norm = (loudness + 60)/60.`
- `key`:
 - * Thêm cột mới `key_name` (C, C#, D, ..., B) và Dảm bảo không còn "unknown" (đã thay bằng giá trị hợp lệ).

```
Thông tin chi tiết về DataFrame:  
<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 775 entries, 0 to 774  
Data columns (total 16 columns):  
 #   Column           Non-Null Count  Dtype     
---  --  
 0   track_id         775 non-null    object    
 1   href              775 non-null    object    
 2   acousticness     775 non-null    float64   
 3   danceability     775 non-null    float64   
 4   energy            775 non-null    float64   
 5   instrumentalness 775 non-null    float64   
 6   key               775 non-null    float64   
 7   liveness          775 non-null    float64   
 8   loudness          775 non-null    float64   
 9   mode              775 non-null    int64     
 10  speechiness       775 non-null    float64   
 11  tempo              775 non-null    float64   
 12  valence            775 non-null    float64   
 13  tempo_norm        775 non-null    float64   
 14  loudness_norm      775 non-null    float64   
 15  key_name          775 non-null    object    
dtypes: float64(12), int64(1), object(3)  
memory usage: 97.0+ KB
```

Hình 3.4: Đặc trưng chung

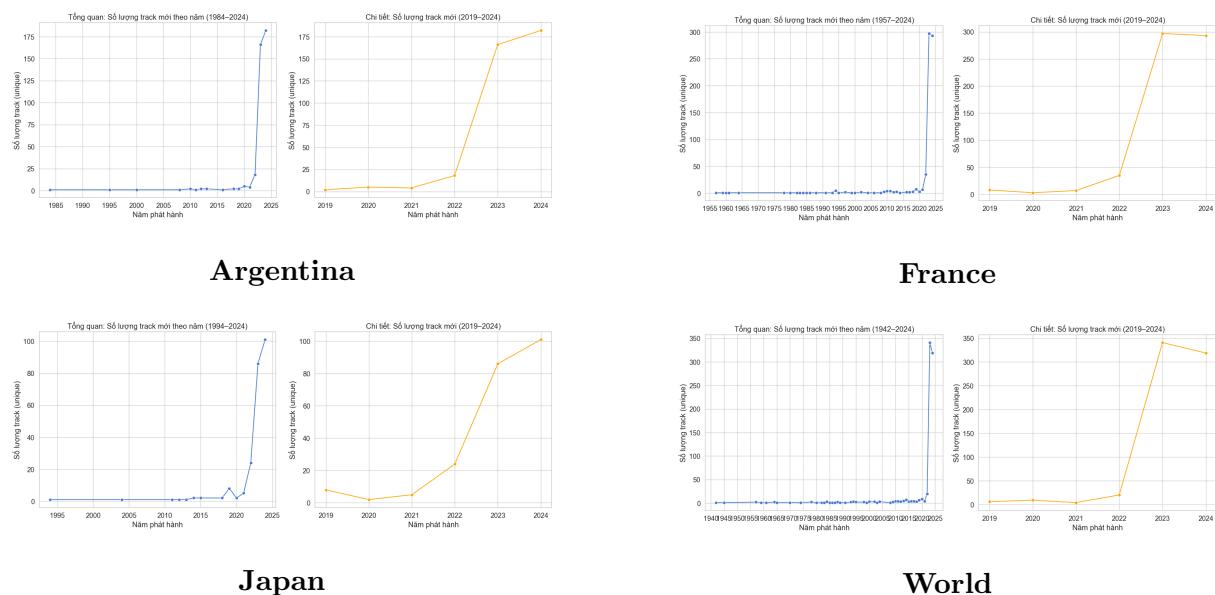
=> **Kết quả:** ghi đè vào các file ...with-audio.csv với các đặc trưng như trên.

3.1.4 Phân tích và tìm insight:

-**Công cụ khám phá:** Jupyter Notebook

1. Bức tranh toàn diện về vòng đời âm nhạc trên BXH

• 1.1 Số bài hát mới phát hành theo năm

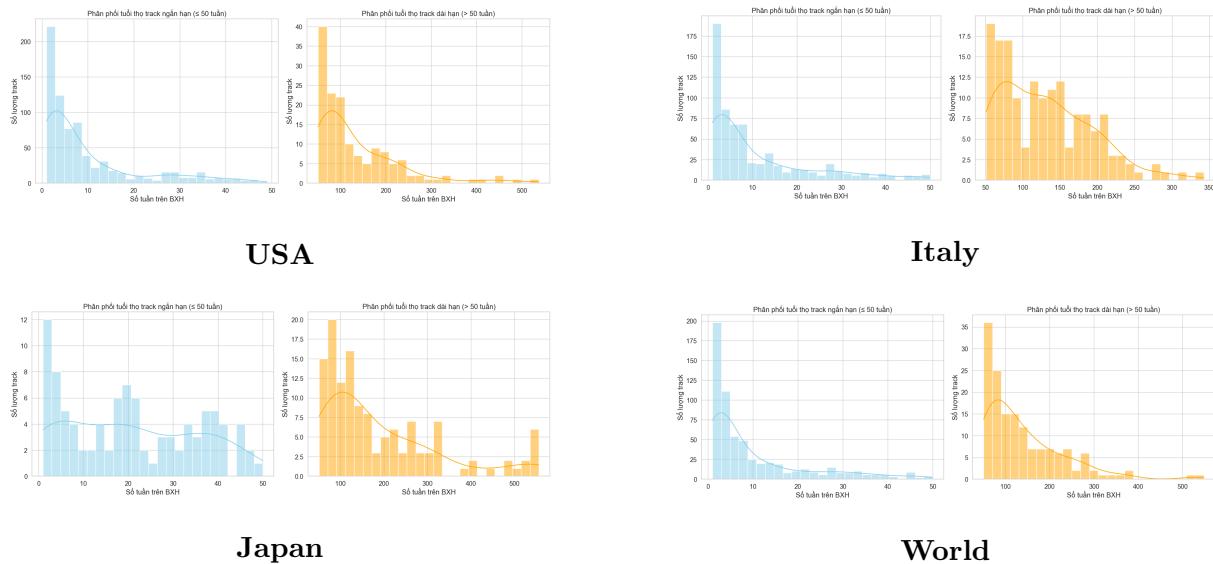


Hình 3.5: Track mới phát hành theo năm

– Kết luận:

- Trước năm 2015: số lượng track mới trên BXH rất ít, tăng trưởng chậm
- Giai đoạn 2015–2020: bắt đầu có xu hướng tăng nhưng chưa bùng nổ.
- Từ 2020 trở đi: số lượng track mới tăng vọt, đặc biệt 2023–2024 đạt đỉnh. Xuất hiện ở hầu hết quốc gia, mạnh nhất tại Mỹ & Argentina, trong khi Hàn Quốc và Nhật Bản nổi bật nhờ K-pop và J-pop.
- Streaming và toàn cầu hoá đã tạo ra “làn sóng bùng nổ” sản xuất nhạc mới sau 2020

• 1.2 Tuổi thọ bài hát trên bảng xếp hạng:

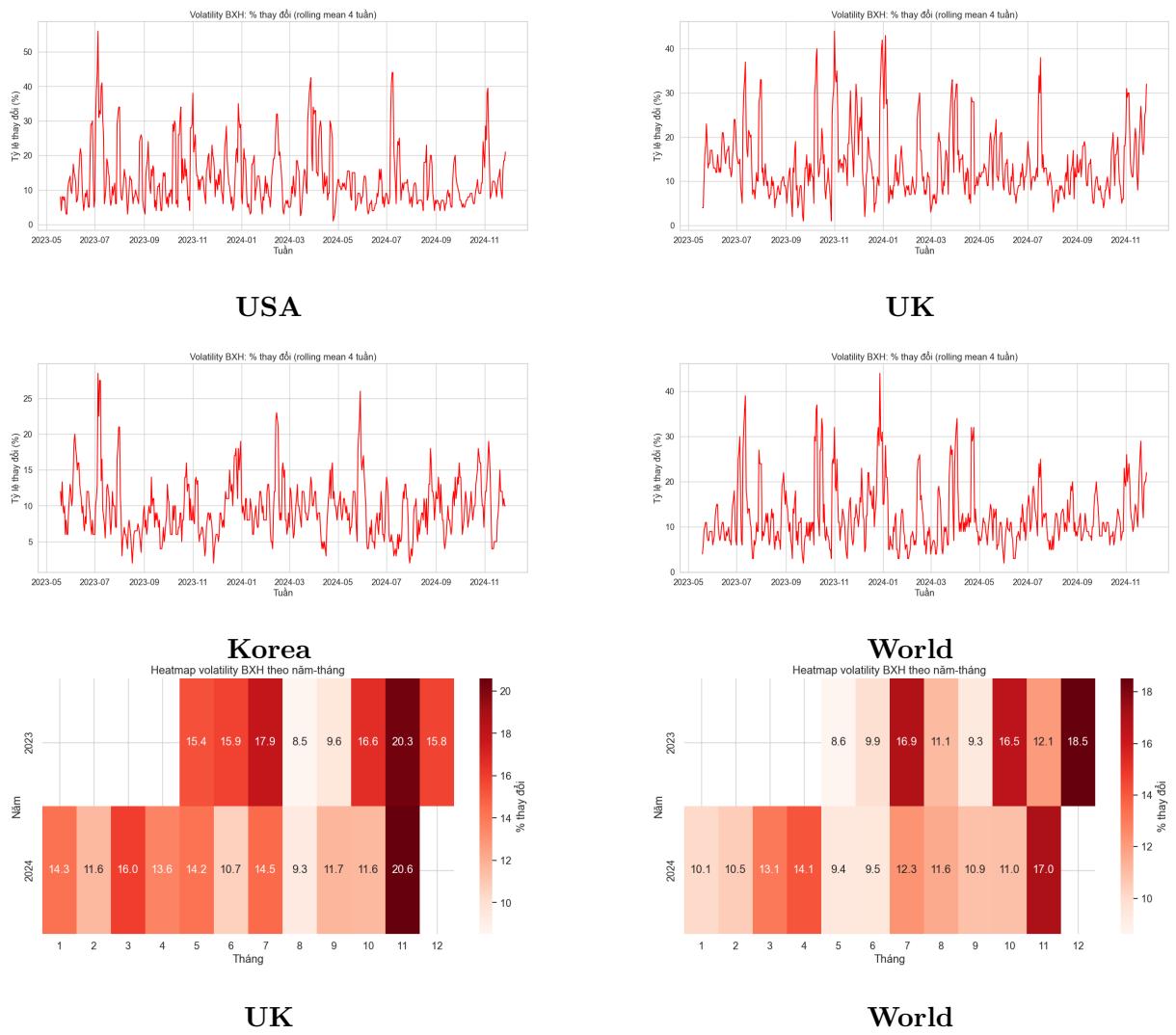


Hình 3.6: 1.2 Tuổi thọ các bài hát trên BXH

– Kết luận:

- Phần lớn bài hát có tuổi thọ ngắn, đa số bài hát chỉ trú được dưới 10 tuần trên BXH => đặc trưng chung của BXH Top 50: hit bùng nổ nhanh nhưng cũng dễ rời khỏi bảng.
- Vẫn tồn tại nhóm nhỏ có tuổi thọ dài (> 50 tuần), kéo dài đến 300–500 tuần. Những bài này thường là siêu hit toàn cầu hoặc gắn liền với văn hóa.
- So sánh giữa các nước và thế giới: Ở cấp quốc gia, tuổi thọ thường ngắn hơn => thị trường có tính “nóng hổi”. Ở cấp thế giới, xuất hiện nhiều bài hát có tuổi thọ cực dài => tính bền vững và lan tỏa rộng.

• 1.3 Độ biến động (volatility) trên BXH



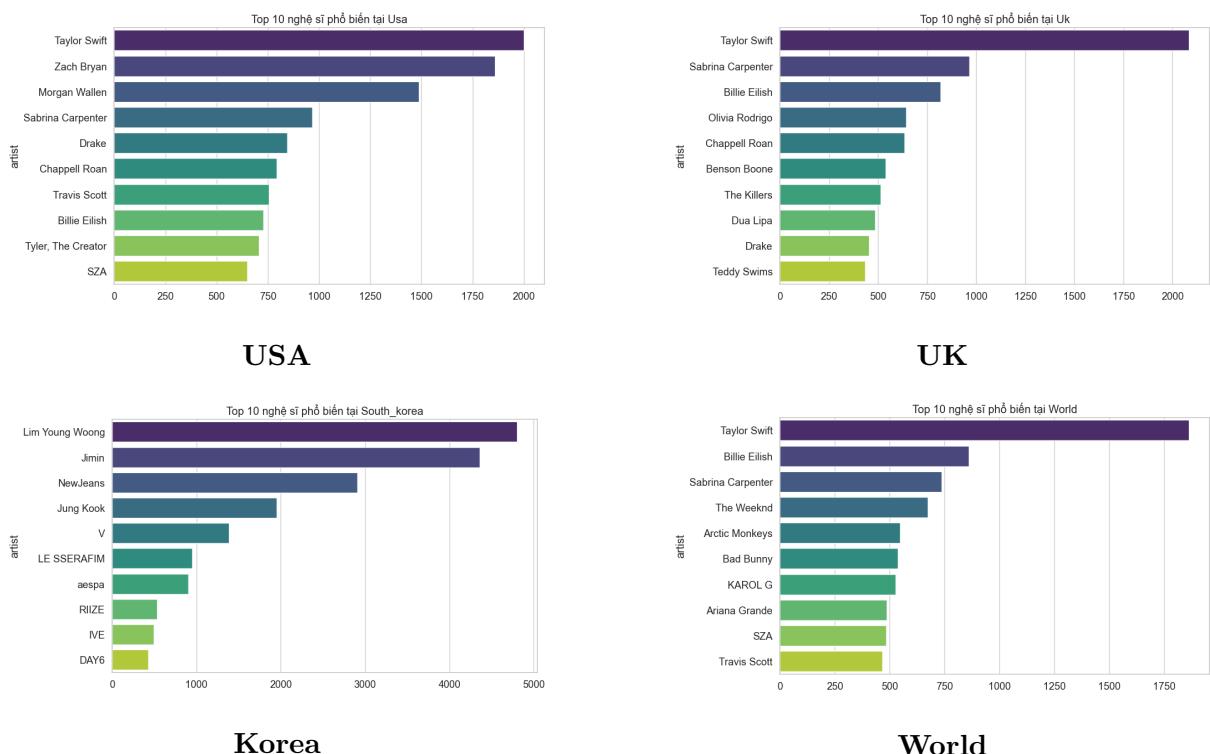
– Kết luận:

- Mức biến động cao ở nhiều thị trường lớn, Mỹ, Anh thường có volatility $> 20\text{--}30\%$ trong nhiều giai đoạn. Cho thấy cạnh tranh gay gắt, các ca khúc nhanh chóng leo hạng và rời BXH
- Một số thị trường ổn định hơn Nhật, Ý, Tây Ban Nha có volatility thấp hơn (chủ yếu $< 15\%$). Điều này phản ánh thị hiếu nghe nhạc ổn định, ít thay đổi đột ngột theo xu hướng.
- Đặc thù mùa vụ và sự kiện âm nhạc: Các đỉnh biến động thường

rơi vào dịp cuối năm, mùa lễ hội. Ví dụ: giai đoạn cuối 2023 và giữa 2024 có nhiều peak volatility (đỉnh) đồng loạt ở nhiều nước.

2. Nghệ sĩ và Bài hát

• 2.1 Top 10 nghệ sĩ phổ biến



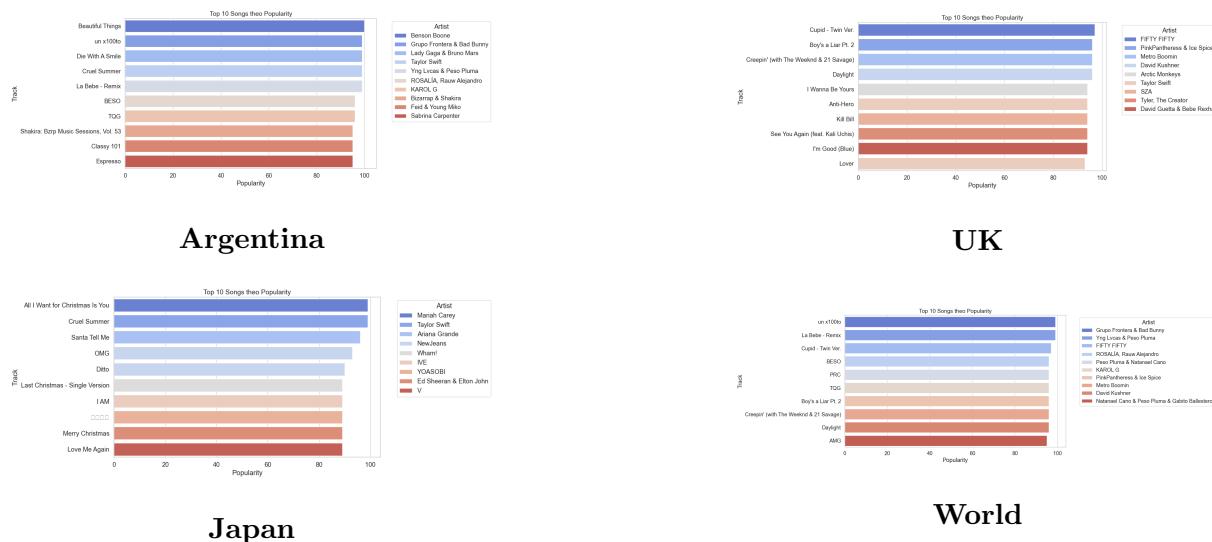
Hình 3.8: Top 10 nghệ sĩ nổi bật nhất

– Kết luận:

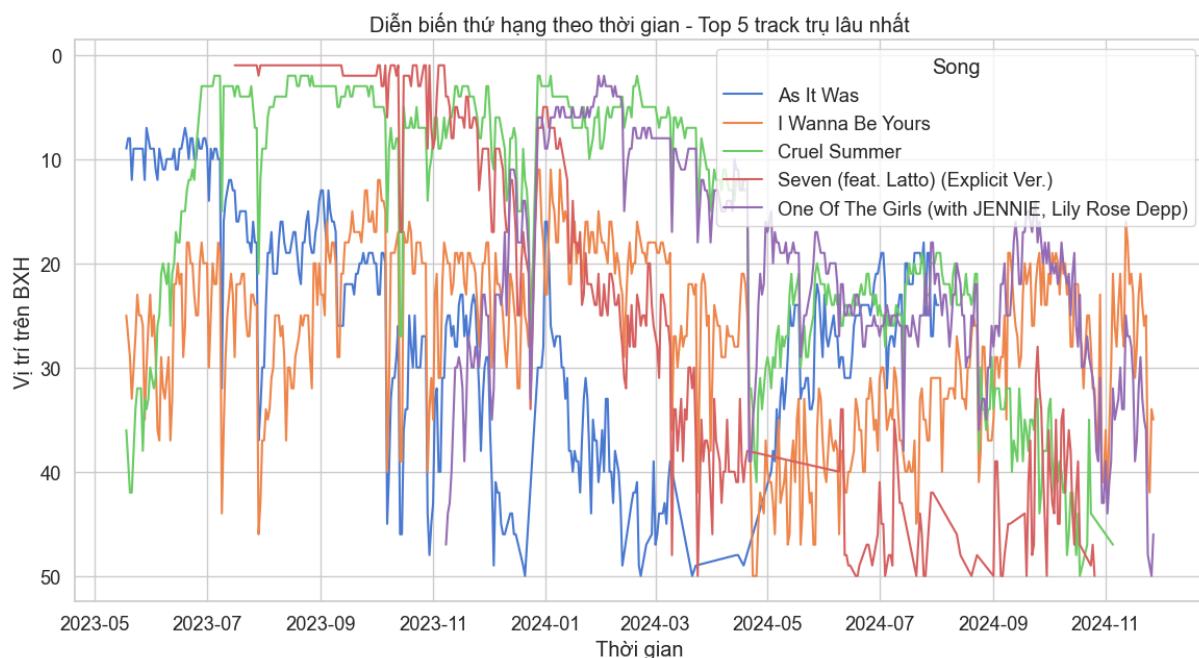
- Nhìn chung đa số quốc gia có nghệ sĩ nội địa thống trị: ví dụ Emilia, Luck Ra (Argentina), Werenoi (France), Geolier, Tedua (Italy), Mrs. GREEN APPLE, YOASOBI (Japan), Peso Pluma (Mexico), Lim Young Woong, BTS members, NewJeans (Korea).
- Riêng thị trường Âu-Mỹ vẫn nổi bật với những ngôi sao toàn cầu: Taylor Swift, Billie Eilish, Drake, Bad Bunny, KAROL G... thường xuyên góp mặt cả ở BXH quốc gia và World.

- Có sự khác biệt rõ rệt giữa thị hiếu nội địa và toàn cầu: Nhật, Hàn Quốc chủ yếu nghệ sĩ bản địa; trong khi BXH World và Mỹ/UK có nhiều nghệ sĩ đa quốc gia.

• 2.2 Bài hát theo độ hot



Hình 3.9: Top 10 bài hát trụ nỗi nhất



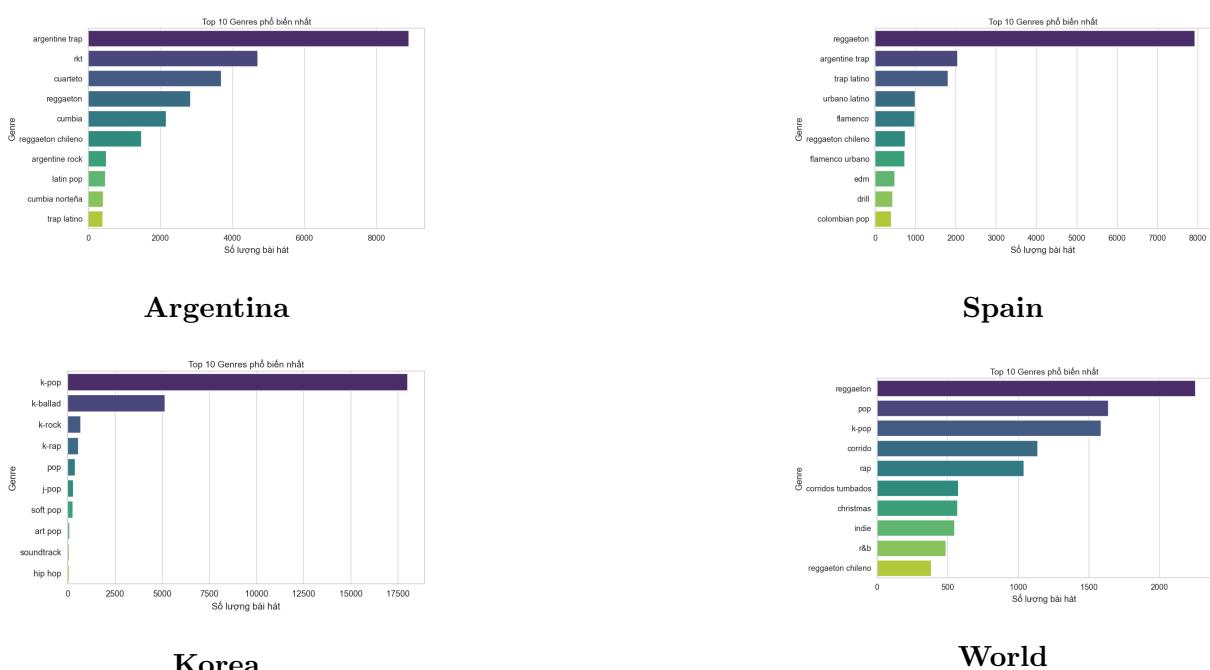
Hình 3.10: Top 5 bài hát trụ lâu nhất

– Kết luận:

- Top 10 bài hát phổ biến: mỗi quốc gia có “quốc ca nội địa” riêng , trong khi BXH World lại nổi bật với các siêu hit toàn cầu .
- Top 5 bài hát trụ lâu nhất : Cho thấy sự khác biệt giữa các thị trường: US/UK có những bản hit lâu bền, Hàn Quốc ghi dấu với loạt K-pop nổi bật, còn các nước Mỹ Latin có nhiều ca khúc trụ lâu thuộc dòng nhạc địa phương. Trên BXH World, một số siêu hit quốc tế thể hiện sức hút bền vững, duy trì thứ hạng cao trong nhiều tháng.
- So sánh quốc gia – thế giới cho thấy: bài hát nội địa thường nổi trong nước nhưng ít khi bền vững toàn cầu, ngược lại siêu hit quốc tế có sức lan tỏa rộng và trụ lâu hơn

3. Thể loại

• 3.1 Top 10 thể loại phổ biến

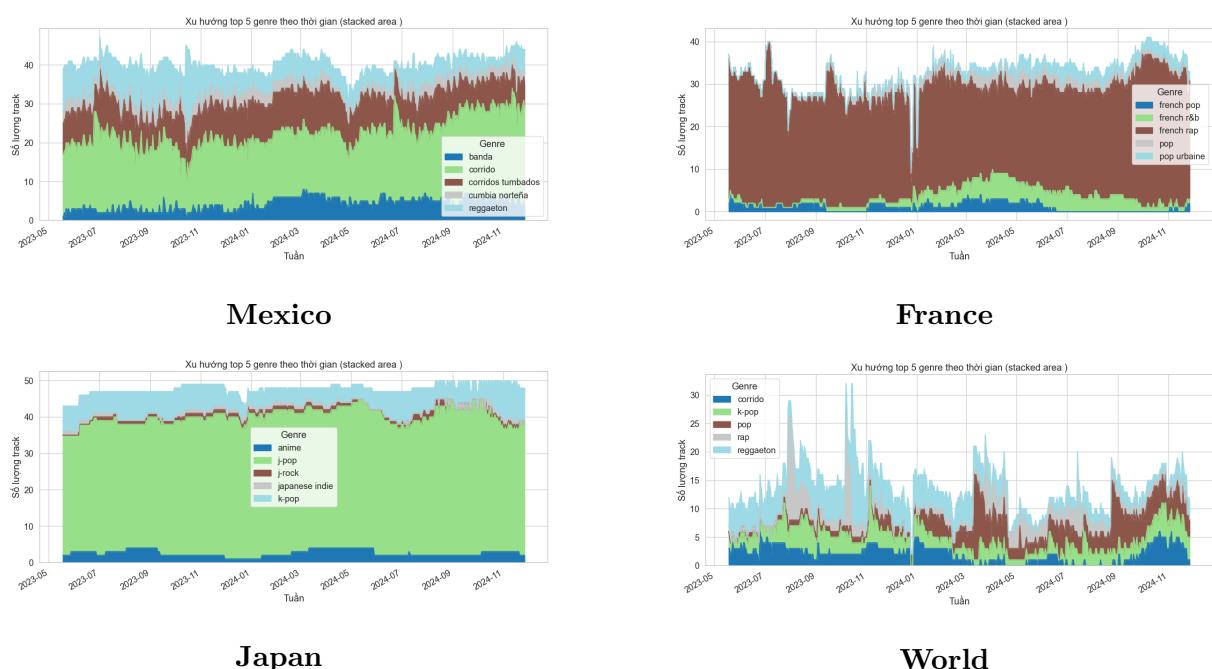


Hình 3.11: Top 10 thể loại nổi bật nhất

– Kết luận:

- Mỗi quốc gia có thể loại bản địa thống trị, ví dụ: Argentina (Argentine trap, RKT, cuarteto), Japan (J-pop, Anime, J-rock), ..
- Sự khác biệt văn hoá rõ rệt: Nhật – Hàn nghiêng về nhạc bản địa (J-pop, K-pop), Mexico – LATAM mạnh về reggaeton/corrido, còn Âu-Mỹ giữ vị thế với pop/rap.
- Xu hướng toàn cầu: Một số thể loại vượt biên giới mạnh mẽ (reggaeton, trap, pop, rap) => lan tỏa sang nhiều thị trường khác nhau.

• 3.1 Xu hướng thể loại theo thời gian

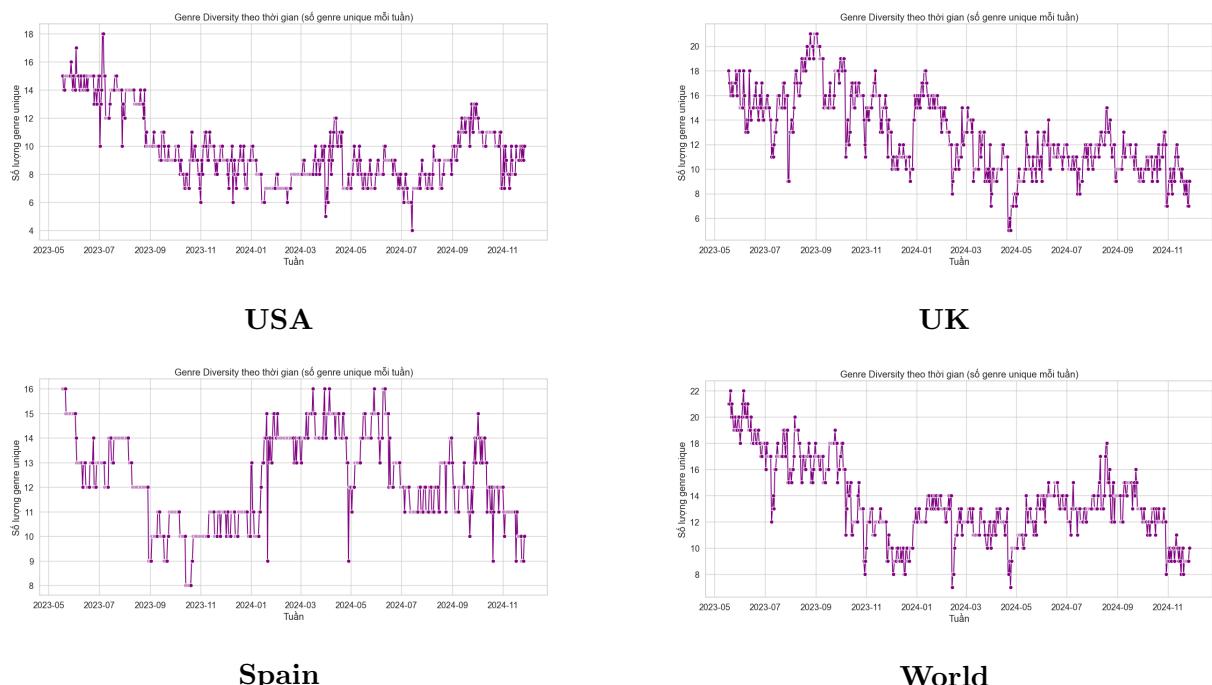


Hình 3.12: Top 10 thể loại nổi bật nhất

– Kết luận:

- Genre bản địa áp đảo ổn định: J-pop (Japan), K-pop/K-ballad (Korea), French rap (France), Italian trap (Italy).

- Đa dạng nhất: Mexico – sự kết hợp của corrido, banda, tumbados, reggaeton, cumbia norteña.
- Khép kín: Nhật và Hàn hầu như chỉ nghe nhạc nội địa.
- Thể loại Latin (reggaeton, corrido, trap latino) tăng mạnh tại Argentina, Mexico, Spain và có sức lan tỏa sang thị trường quốc tế.
- Thể loại mùa vụ như Christmas nổi bật ở US/UK, tạo các đỉnh ngắn hạn cuối năm.
- Thị trường toàn cầu (World) cho thấy sự kết hợp của reggaeton, K-pop và pop/rap, phản ánh sự giao thoa văn hóa âm nhạc xuyên biên giới.



Hình 3.13: Độ đa dạng theo tuần

– Kết luận:

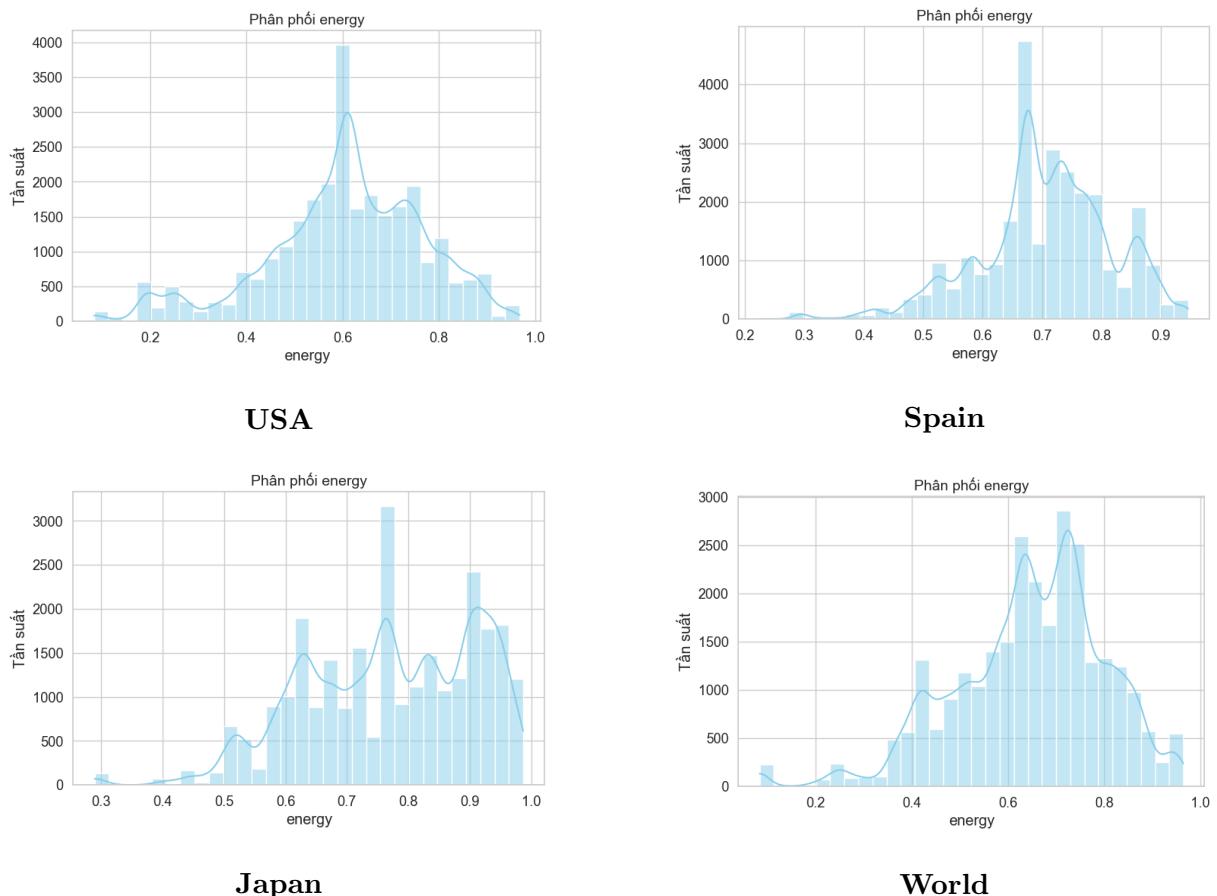
- Khác biệt thị trường: US và World đa dạng nhất (10–22 genre/-tuần), trong khi Nhật, Hàn, Ý khá tập trung (5–12 genre, chủ yếu

J-pop, K-pop, Italian trap).

- Xu hướng mùa vụ: US/UK xuất hiện nhiều genre phụ cuối năm (Christmas, country, grime), còn Nhật/Hàn ổn định quanh 5–8 genre.
- Mỹ Latin (Argentina, Mexico, Spain): trung bình 8–15 genre, xoay quanh reggaeton, corrido, trap latino => ít bùng nổ thể loại mới.
- Nhật Bản & Hàn Quốc: số genre unique thấp và ổn định (5–8/tuần), gần như bị áp đảo bởi J-pop (Japan) và K-pop/K-ballad (Korea).

4. Đặc trưng âm nhạc

- 4.1 Đặc trưng energy

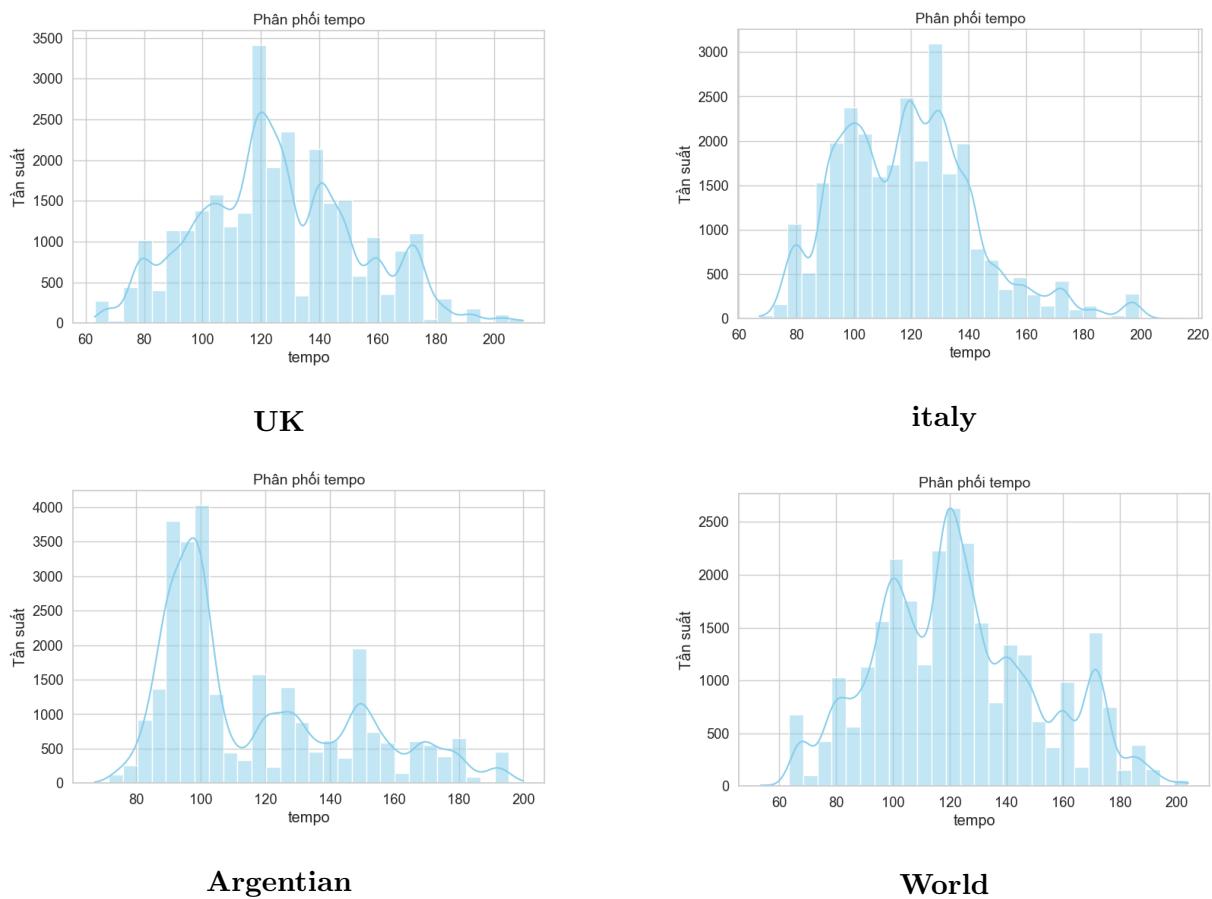


Hình 3.14: Đặc trưng energy

– **Kết luận:**

- Tồn tại sự khác biệt rõ rệt về mức năng lượng ưa chuộng giữa các thị trường, phản ánh sự đa dạng trong văn hóa nghe nhạc và đặc tính thể loại bản địa.
- Phân cực rõ ràng giữa hai nhóm thị trường:
 - * Nhóm năng lượng cao & đa dạng: Hầu hết các quốc gia như Argentina, France, UK, USA có phân phối energy. trải dài từ 0.2 đến 1.0 => thị hiếu âm nhạc tại đây rất đa dạng, chấp nhận cả những bài nhạc trầm lắng (energy thấp) lẫn những bài cực kỳ sôi động (energy cao).
 - * Nhóm năng lượng trung bình & "ôn hòa": Italy và Spain có phân phối giới hạn trong khoảng 0.2 đến 0.9 => top âm nhạc tại đây có xu hướng thiếu vắng những bài hát có mức năng lượng cực cao, phù hợp với các thể loại mang tính chất nhẹ nhàng, êm dịu hơn như Pop, Ballad, hoặc Latin truyền thống.
- Japan là trường hợp đặc biệt: Phân phối energy của Nhật bắt đầu từ 0.3 đến 1.0. Nguồn năng lượng tối thiểu cao hơn => ưa chuộng những bản nhạc có tiết tấu nhanh và sôi động ngay từ đầu, ít các bài hát có nhịp độ chậm và trầm buồn
- Thị trường toàn cầu (World) là sự kết hợp: Phân phối energy của World phủ rộng từ 0.2 đến 1.0, phản ánh chính xác việc nó là sự tổng hòa của tất cả các thị trường, thể hiện sự đa dạng và toàn diện nhất về mức năng lượng.

• **4.2 Đặc trưng tempo**



Hình 3.15: Đặc trưng tempo

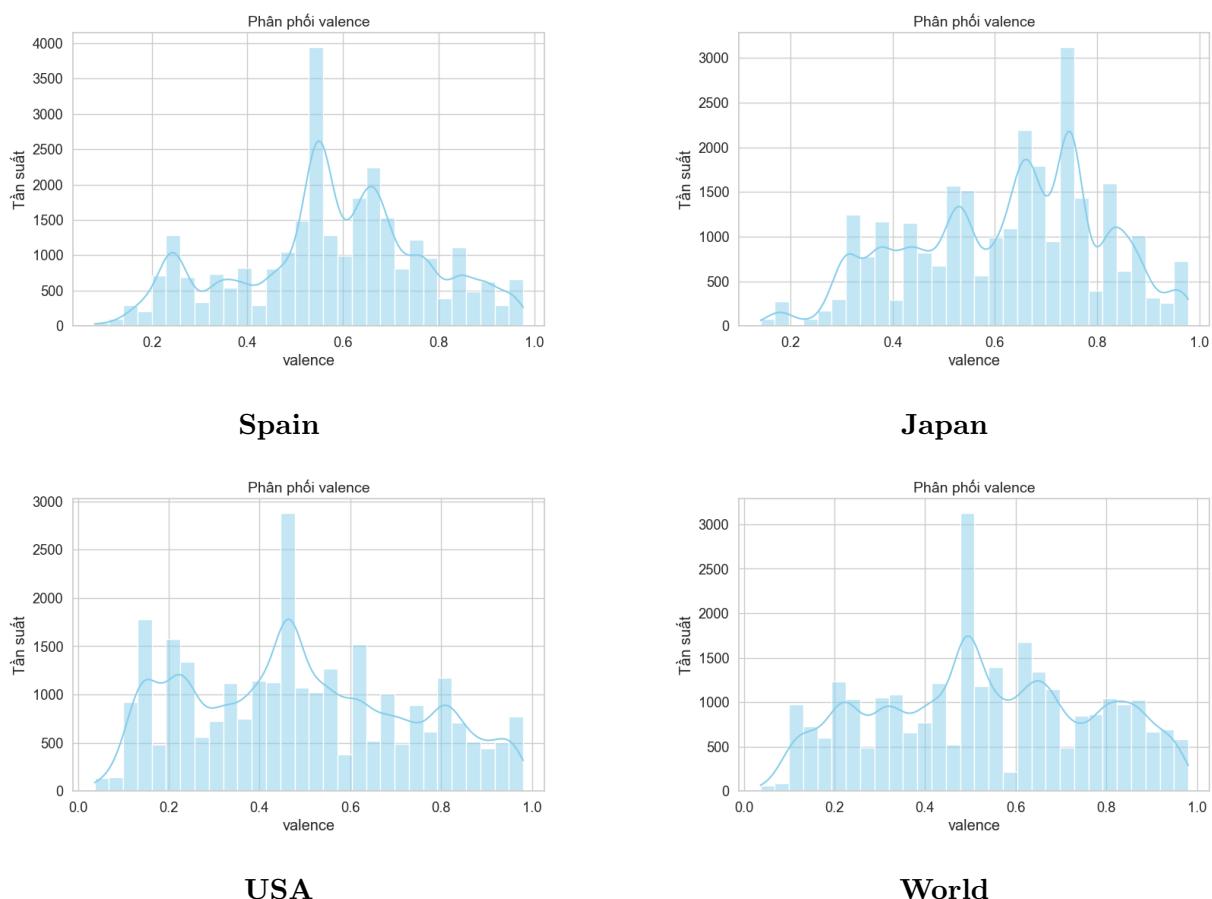
– Kết luận:

- Tồn tại sự tương đồng lớn về phân phối nhịp độ giữa các thị trường chủ chốt. Các nước như Argentina, Mexico, Nhật Bản (Japan) và Hàn Quốc (South Korea) có phân phối tempo gần như trùng khớp, tập trung cao ở khoảng 100-140 BPM. Điều này cho thấy một "công thức nhịp độ toàn cầu" cho các bản hit.
- Một số thị trường có phạm vi nhịp độ rộng hơn. Các nước như Italy, Tây Ban Nha (Spain), Pháp (France), Anh (UK), Mỹ (USA) và thị trường Thế giới (World) có phân phối trải dài từ khoảng 60 BPM đến 180-200 BPM. Sự đa dạng này phản ánh tính chất âm nhạc phong phú, bao gồm cả những bản ballad chậm rãi (tempo

thấp) lẫn nhũng bài EDM hoặc Rock sôi động (tempo cao).

- Nhịp độ trung bình (80-140 BPM) là phổ biến nhất trên toàn cầu, phù hợp với các thể loại nhạc phổ biến như Pop, Rap và Dance. Phân phối của thị trường Thế giới (World) một lần nữa đóng vai trò là trung bình chuẩn mực, tổng hòa xu hướng từ tất cả các thị trường thành phần.

• 4.3 Đặc trưng valence



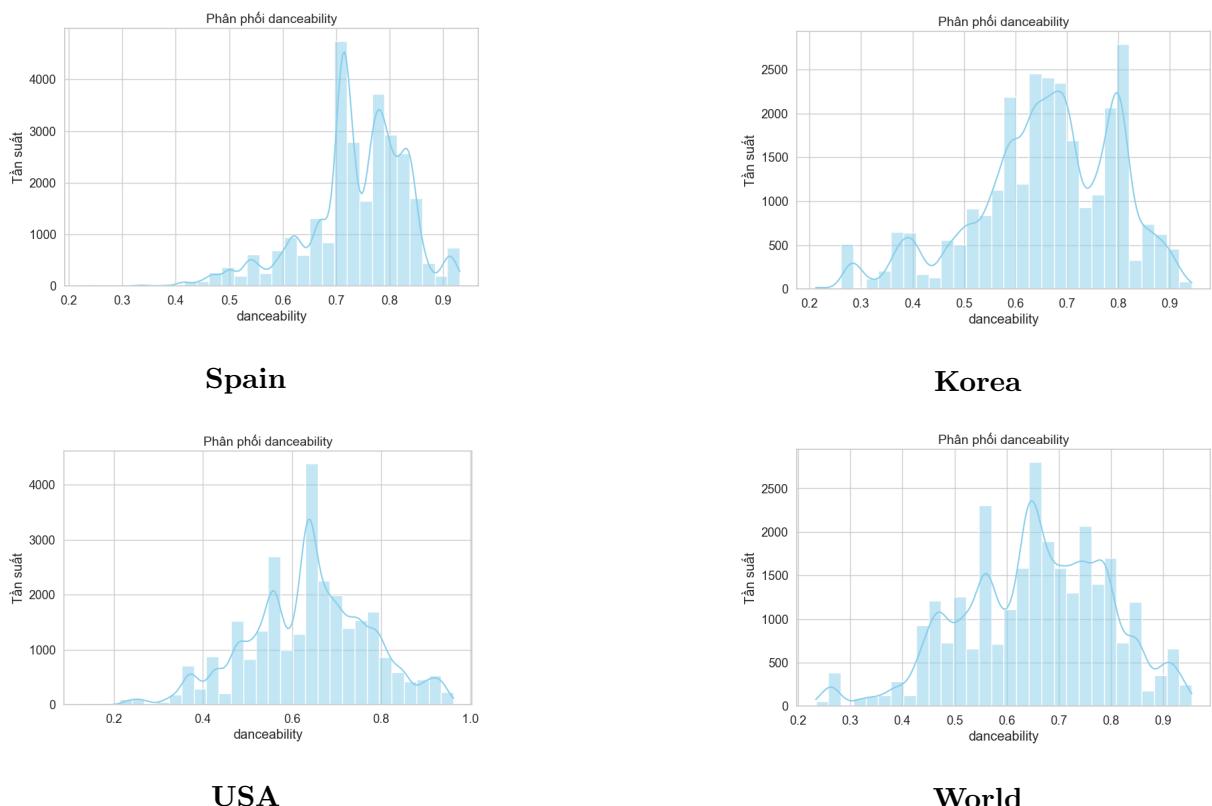
Hình 3.16: Đặc trưng valence

– Kết luận:

- Đa số các thị trường có phân phối valence tập trung ở mức trung bình đến cao (0.4 - 0.8). Điều này cho thấy xu hướng chung toàn cầu là ưa chuộng những bài hát mang cảm xúc tích cực, vui tươi.

- Tây Ban Nha (Spain) và Nhật Bản (Japan) là hai thị trường nổi bật với xu hướng yêu thích các bản nhạc cực kỳ tích cực. Phân phối valence của hai nước này lệch hẳn về phía giá trị cao (0.8 - 1.0), với tần suất cực lớn. Điều này đặc biệt phù hợp với các thể loại sôi động, tươi vui như Reggaeton (Tây Ban Nha) và J-pop (Nhật Bản).
- Các thị trường Âu-Mỹ (Pháp, Italy, Anh, Mỹ) có phân phối cân bằng và đa dạng hơn. Phân phối của họ trải rộng từ 0.0 đến 1.0, cho thấy sự chấp nhận đối với cả những bài hát có cảm xúc trầm lắng, u sầu (valence thấp) bên cạnh những bài hát vui vẻ. Điều này phản ánh thị hiếu âm nhạc đa chiều và có chiều sâu.

• 4.4 Đặc trưng danceability



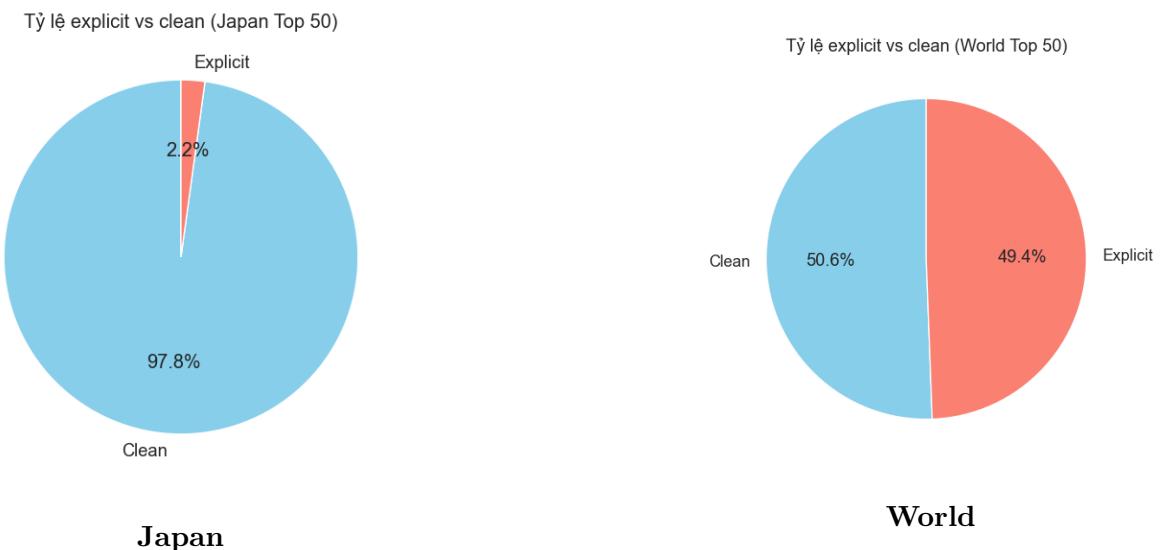
Hình 3.17: Đặc trưng danceability

– Kết luận:

- Nhóm ưa chuộng tính nhảy múa rất cao: Các quốc gia như Argentina, Italy, Japan, Mexico và Tây Ban Nha (Spain) có phân phối danceability rất hẹp, tập trung ở khoảng 0.6 - 0.9. Điều này cho thấy các bài hát ở đây gần như bắt buộc phải có nhịp điệu dễ nhảy theo, phù hợp với các thể loại như Reggaeton, Latin Pop hay EDM.
- Nhóm linh hoạt hơn: Các thị trường Pháp (France), Hàn Quốc (South Korea), Anh (UK), Mỹ (USA) và Thế giới (World) có phân phối rộng hơn, từ 0.2 đến 1.0. Mặc dù vẫn nghiêng về các bài hát dễ nhảy, họ vẫn chấp nhận những bài hát ít tính nhảy múa hơn (ví dụ: ballad, nhạc acoustic), cho thấy thị hiếu đa dạng.
- Phân phối của thị trường Thế giới (World) cho thấy sự cân bằng, nhưng vẫn nghiêng nhiều về nhóm các bài hát có danceability cao, chứng tỏ đây là một xu hướng chủ đạo.

5. So sánh nhóm

• 5.1 Tỷ lệ Explicit

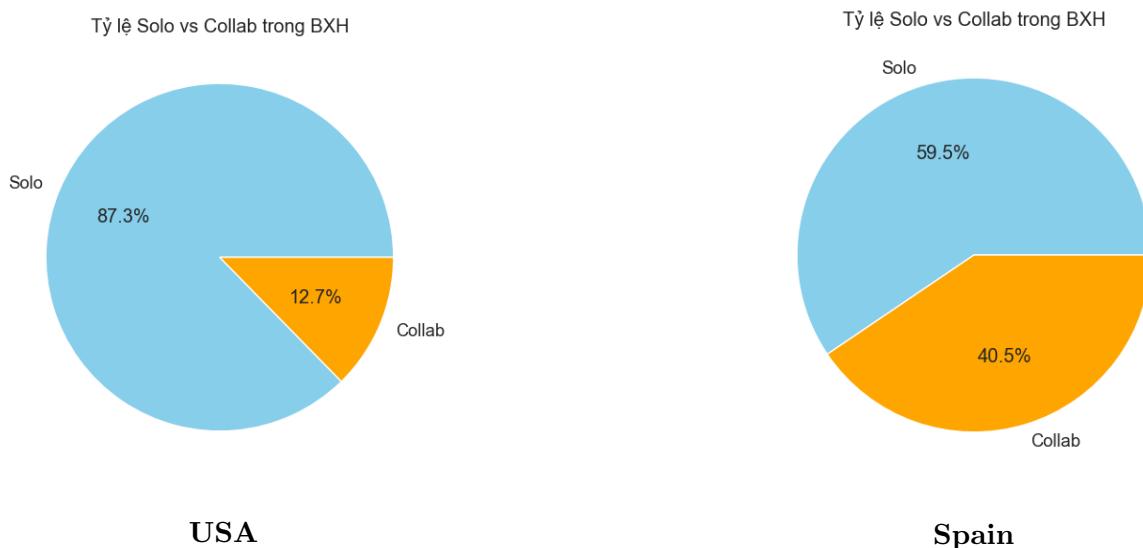


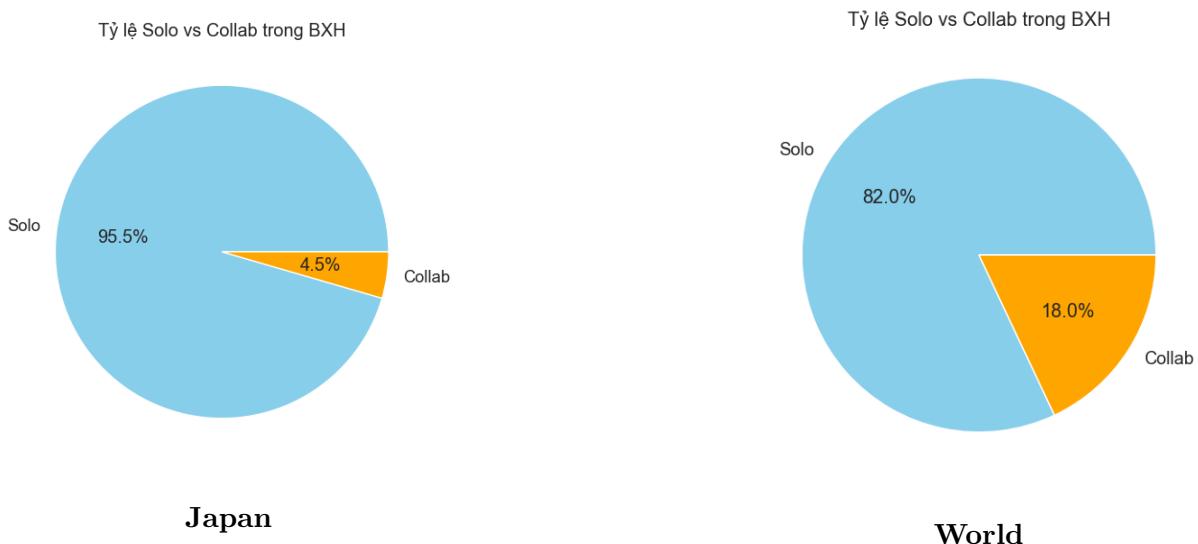
Hình 3.18: So sánh tỷ lệ Explicit

– Kết luận:

- Tồn tại sự khác biệt văn hóa cực kỳ lớn giữa các thị trường trong việc chấp nhận nội dung explicit.
- Nhật Bản và Hàn Quốc với tỷ lệ bài hát clean áp đảo tuyệt đối ($> 90\%$). Điều này cho thấy một tiêu chuẩn văn hóa và giải trí rất khắt khe, phù hợp với các thể loại nhạc đại chúng như J-pop và Anime.
- Các thị trường châu Á và châu Âu thể hiện sự phân cực rõ rệt:
- Argentina là thị trường có tỷ lệ bài hát explicit cao nhất (68.3 %), cho thấy sự thống trị của các thể loại như Argentine Trap và Reggaeton, nơi nội dung explicit là một phần của văn hóa âm nhạc.
- Các thị trường lớn còn lại (Italy, Mexico, Spain, UK, USA, World) có tỷ lệ gần như cân bằng (50/50). Điều này cho thấy sự chấp nhận và cạnh tranh song hành giữa hai dòng nhạc, phản ánh sự đa dạng và phân hóa trong thị hiếu của người nghe.

• 5.2 Solo vs Colab





Hình 3.19: So sánh tỷ lệ Explicit

– **Kết luận:**

- Tỷ lệ bài hát Solo chiếm ưu thế tuyệt đối trên toàn cầu => nghệ sĩ cá nhân vẫn có sức hút và khả năng thống trị các bảng xếp hạng lớn.
- Tồn tại sự khác biệt lớn về văn hóa âm nhạc giữa các thị trường:
 - * Nhật Bản và Hàn Quốc là thị trường đề cao tính cá nhân rõ rệt nhất, với tỷ lệ bài hát Solo chiếm tới 95.5 %. Điều này phù hợp với đặc trưng của J-pop, K-pop, nơi các nghệ sĩ solo thường được xây dựng hình tượng mạnh mẽ.
 - * Các thị trường Mỹ Latin (Argentina, Mexico, Tây Ban Nha) có tỷ lệ Collab (hợp tác) cao nhất, dao động từ 37 % đến 46%. Điều này phản ánh đặc tính cộng đồng và xu hướng kết hợp trong các thể loại âm nhạc phổ biến tại đây như Reggaeton, Latin Trap.
 - * Các thị trường Âu-Mỹ (Mỹ, Anh, Pháp, Italy) có tỷ lệ Solo rất cao (từ 69.7% đến 87.3%), cho thấy sự thống trị của các ngôi sao solo trong làng nhạc đại chúng.

- Ở Thế giới, Bài hát Solo chiếm ưu thế rõ rệt (82.0%), cho thấy sức hút và khả năng thống trị bảng xếp hạng toàn cầu của các nghệ sĩ cá nhân.

3.2 Nguồn 2:

3.2.1 Giới thiệu Dataset:

Bộ dữ liệu này ghi lại xu hướng nghe nhạc trực tuyến toàn cầu trên Spotify trong năm 2024 được mô phỏng để phản ánh các xu hướng thực tế. Nguồn tham khảo bao gồm Spotify Wrapped 2023, IFPI Global Music Report, và Chartmetric. Nó mang đến những thông tin giá trị về sở thích của người dùng tại nhiều quốc gia, các nghệ sĩ và album nổi bật, tổng số giờ nghe nhạc, cũng như hành vi của người nghe. Được xây dựng nhằm phục vụ cho các mục đích như phân tích dữ liệu chuyên sâu, phát triển mô hình học máy, và xây dựng hệ thống báo cáo – bảng điều khiển thông minh trong lĩnh vực âm nhạc và truyền thông.

Bộ dữ liệu ‘*Spotify_2024_Global_Streaming_Data.csv*’ gồm 500 dòng và 12 cột. Bao gồm các features sau:

3.2.2 Data enrichment:

Gọi Spotify Search API với album name + artist để lấy đúng album_id. Với album_id lấy được gọi Recobeats API để lấy thêm các features sau đây:

3.2.3 Data cleaning and preprocessing:

Kiểm tra các dữ liệu bị thiếu, xác minh các kiểu dữ liệu:

Column	Data type	Meaning
Country	object	Quốc gia ghi nhận dữ liệu nghe nhạc
Artist	object	Tên nghệ sĩ hoặc ban nhạc
Album	object	Tên album do nghệ sĩ phát hành
Genre	object	Thể loại âm nhạc (ví dụ: Pop, Hip-hop, R&B, Classical, ...)
Release Year	int64	Năm album được phát hành chính thức
Monthly Listeners (Million)	float64	Số lượng người nghe hàng tháng trung bình của nghệ sĩ (tính theo triệu)
Total Streams (Millions)	float64	Tổng số lượt phát (lượt nghe) toàn cầu của album (tính theo triệu)
Total Hours Streamed (Millions)	float64	Tổng thời gian người dùng đã nghe album
Avg Stream Duration (Min)	float64	Thời lượng trung bình của 1 lượt nghe album
Platform Type	object	Loại tài khoản người dùng: Free hay Premium
Streams Last 30 Days	float64	Tổng lượt nghe 30 ngày gần nhất của bài hát
Skip Rate (%)	float64	Tỉ lệ bỏ qua bài hát khi chưa nghe hết bài hát

Bảng 3.1: Thông tin features dữ liệu nghe nhạc

```

▶ import pandas as pd
# Load the data
df = pd.read_csv('/content/Spotify_2024_Global_Streaming_Data.csv')

# Kiểm tra dữ liệu trống
missing_values = df.isnull().sum()
print('Dữ liệu trống ở mỗi cột:')
print(missing_values)

df = df.dropna()
print('\nSau khi bỏ dữ liệu trống ta được, shape of DataFrame:', df.shape)

# Xác định kiểu dữ liệu
print('\nData Types:')
print(df.dtypes)

# Xóa các khoảng trắng giữa các tên column đổi thành -
df.columns = df.columns.str.strip().str.replace(" ", "_").str.replace("( ", "").str.replace(")", "").str.lower()
print('\nSau khi đổi tên cột, ta được:')
print(df.columns)

```

Hình 3.20: Kiểm tra null, xác định kiểu dữ liệu và đổi các khoảng trắng giữa các features

Feature	Type	Meaning
id	string	Reccobeats' id, id duy nhất cho mỗi album
albumType	string	Kiểu album (album, single, compilation)
artist	object	Danh sách các nghệ sĩ tham gia album, id artist, href Spotify artist
totalTracks	integer	Tổng số track trong album
href	string	Spotify URL của album
name	string	Tên của album
availableCountries	string	Danh sách các quốc gia có sẵn cho album này, định danh bằng ISO 3166 alpha-2 code, phân biệt bởi dấu phẩy
releaseDate	string	Ngày phát hành đầu tiên
releaseDateFormat	string	Định dạng của ngày phát hành (day, month, year)
isrc	string	International Standard Recording Code (Optional)
ean	string	European Article Number (Optional)
upc	string	Universal Product Code (Optional)
label	string	Nhãn gắn với album (hashtag)
poppularity	integer	Độ nổi tiếng của album. Giá trị chạy từ 0–100, với 100 là giá trị nổi tiếng nhất

Bảng 3.2: Thông tin các trường dữ liệu album từ Reccobeats/Spotify

```

Dữ liệu trống ở mỗi cột:
Country          0
Artist           0
Album            0
Genre             0
Release Year     0
Monthly Listeners (Millions) 0
Total Streams (Millions)    0
Total Hours Streamed (Millions) 0
Avg Stream Duration (Min)   0
Platform Type     0
Streams Last 30 Days (Millions) 0
Skip Rate (%)      0
dtype: int64

Sau khi bỏ dữ liệu trống ta được, shape of DataFrame: (500, 12)

Data Types:
Country          object
Artist           object
Album            object 33
Genre             object
Release Year     int64
Monthly Listeners (Millions) float64
Total Streams (Millions)  float64
Total Hours Streamed (Millions) float64
Avg Stream Duration (Min)  float64
Platform Type     object

```

Từ file dataset ban đầu ta gọi spotify api để lấy spotify_id cho mỗi album được file ‘Albums_Info_Clean.csv’, từ đó gọi reccobeats api để lấy thêm các features, tạo thành 1 file mới là:

```

import pandas as pd
import ast

df = pd.read_csv("/content/Albums_Info_from_ReccoBeats_2.csv")

# Kiểm tra null
print("Info null:")
df.info()
# xóa các cột toàn null
cols_to_drop = df.columns[df.notna().sum() == 0]
print("Các cột bị xoá:", list(cols_to_drop))

df = df.drop(columns=cols_to_drop)

#Tách tên nghệ sĩ chính từ cột artists
def extract_artist_name(x):
    try:
        artists = ast.literal_eval(x)
        if isinstance(artists, list) and len(artists) > 0:
            return artists[0]['name']
        else:
            return None
    except:
        return None

df['artist_name'] = df['artists'].apply(extract_artist_name)

#Chuẩn hóa releaseDate
df['releaseDate'] = pd.to_datetime(df['releaseDate'], errors='coerce')
df['release_year'] = df['releaseDate'].dt.year

df_clean = df.dropna(subset=['artist_name', 'name', 'release_year', 'label', 'popularity'])

output_path = "/content/Albums_Info_Clean.csv"
df_clean.to_csv(output_path, index=False)

print("\n✓ Đã tiền xử lý và lưu file mới tại:", output_path)
print("❗ Các cột còn lại:", df_clean.columns.tolist())
print(df_clean.head())

```

Hình 3.22: Kiểm tra, xóa các cột null, tách tên nghệ sĩ và chuẩn hóa releaseDate

```

Info null:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 179 entries, 0 to 178
Data columns (total 14 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   id               179 non-null    object  
 1   albumtype        179 non-null    object  
 2   artists          179 non-null    object  
 3   totalTracks      179 non-null    int64  
 4   href             179 non-null    object  
 5   name             179 non-null    object  
 6   availableCountries 179 non-null    object  
 7   releasedate     179 non-null    object  
 8   releasedateFormat 179 non-null    object  
 9   isrc             0 non-null     float64 
 10  ean              0 non-null     float64 
 11  upc              179 non-null    int64  
 12  label            179 non-null    object  
 13  popularity       179 non-null    int64  
dtypes: float64(2), int64(3), object(9)
memory usage: 19.7+ kB
các cột bị xoá: ['isrc', 'ean']

Nó tiễn xử lý và lưu file mới tại: /content/Albums_Info_clean.csv
Các cột còn lại: ['id', 'albumtype', 'artists', 'totalTracks', 'href', 'name', 'availableCountries', 'releasedate', 'releasedateFormat', 'upc', 'label', 'popularity', 'artist_name', 'release_year']
id albumtype \
0 b6d2e255-51c5-4df6-a22c-e294a48a382d album
1 4f9fe8ba-c293-4e95-ba6c-c394feef95d album
2 233a2a2a-1a0a-4a20-8a10-1a2a2a2a2a2a album
3 f7d01f2-4bf4-4e78-b528-d5a34eb9e9ad album
4 bc332ec6-8303-4ab8-af2c-c2fd6392bc9 album

artists totalTracks \
0 [{"id": "0451b6b2-8746-ad43-ab07-c355ed1e3948", "name": "The Weeknd", "total_tracks": 14}
1 [{"id": "b9fc5ca2-f233-484c-b894-3ab946a99989", "name": "Drake", "total_tracks": 16}
2 [{"id": "69a21d95-c9a4-4756-8f0a-bd02711156f4", "name": "Lil Nas X", "total_tracks": 17}
3 [{"id": "2e7ea94c5-9c78-487f-b9d4-04030905121a", "name": "Kanye West", "total_tracks": 8}
4 [{"id": "c7b330b5-a62e-42b1-bf02-943cad08746", "name": "BTS", "total_tracks": 22}

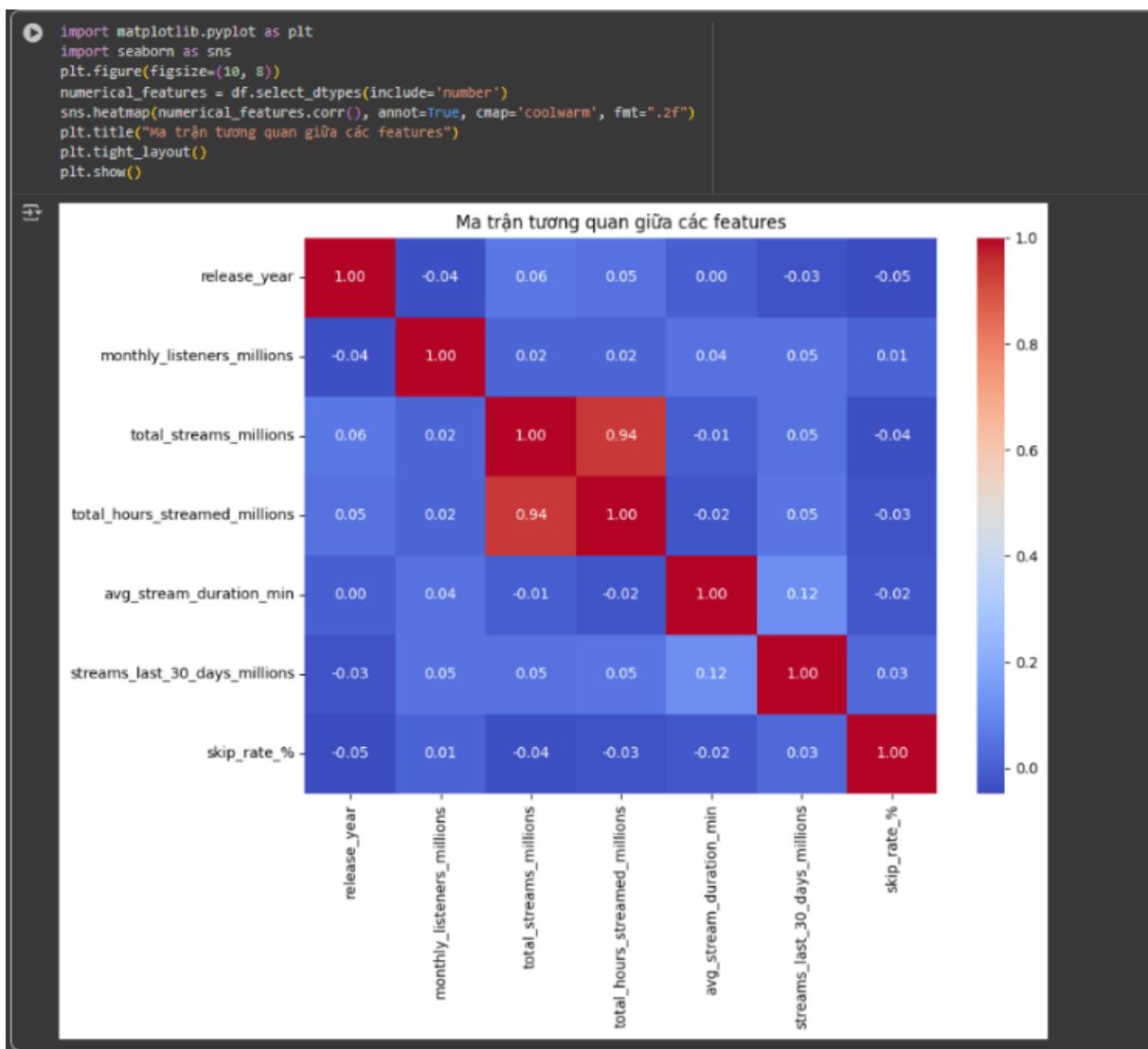
href \
0 https://open.spotify.com/album/4Yp6hdh02zPlshxJ...
1 https://open.spotify.com/album/03G01034nRkDpR...
2 https://open.spotify.com/album/10ef2f3cub6c2y5...
3 https://open.spotify.com/album/03G01034nRkDpR...
4 https://open.spotify.com/album/10ef2f3cub6c2y5...

```

Hình 3.23: Kết quả từ hình 3

3.2.4 Overview and Analysis

Dataset ban đầu Spotify_2024_Global_Streaming_Data.csv:



Hình 3.24: Ma trận tương quan giữa các features

Nhận xét:

- Giữa total_streams_millions và total_hours_streamed_millions có mối tương quan mạnh, hợp lý về mặt ý nghĩa vì khi số lượng stream tăng thì tổng giờ nghe cũng tăng tương ứng.
- Các hệ số của monthly_listeners_millions khá thấp so với các chỉ số khác, một nghệ sĩ có nhiều người nghe hàng tháng chưa chắc có tổng giờ stream cao(vì có thể mỗi người chỉ nghe 1-2 lần)

- Tỉ lệ skip rate mặc dù có nhiều giá trị âm trong bảng tương quan không ảnh hưởng quá nhiều bởi các features khác, nó có thể liên quan nhiều hơn đến chất lượng nội dung hoặc sở thích cá nhân chứ không vì release_year, nhiều nhạc cũ vẫn có thể hot nếu trở nên viral và nhạc mới ra chưa chắc đã nhiều người nghe.

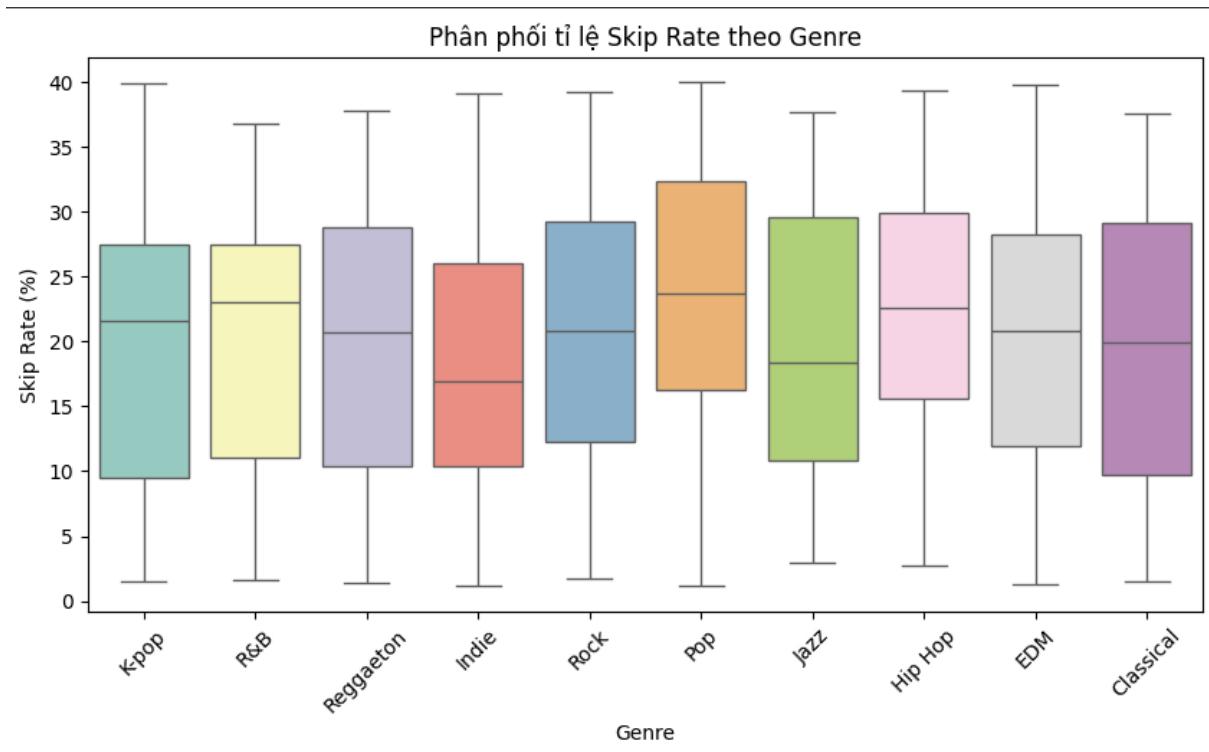
```

df = pd.read_csv("/content/Spotify_2024_Global_Streaming_Data.csv")
# Skip Rate Distribution by Genre
plt.figure(figsize=(10, 5))
sns.boxplot(data=df, x='Genre', y='Skip Rate (%)', hue='Genre', palette='Set3', legend=False)
plt.title("Phân phối tỉ lệ Skip Rate theo Genres")
plt.xlabel("Genre")
plt.ylabel("Skip Rate (%)")
plt.xticks(rotation=45)
plt.show()
#Free vs Premium Users
total_streams = df.groupby("Platform Type")["Total Streams (Millions)"].mean().reset_index()
total_streams["Genre"] = "Total"
df_for_plot = pd.concat([df[["Platform Type", "Genre", "Total Streams (Millions)"]], total_streams])

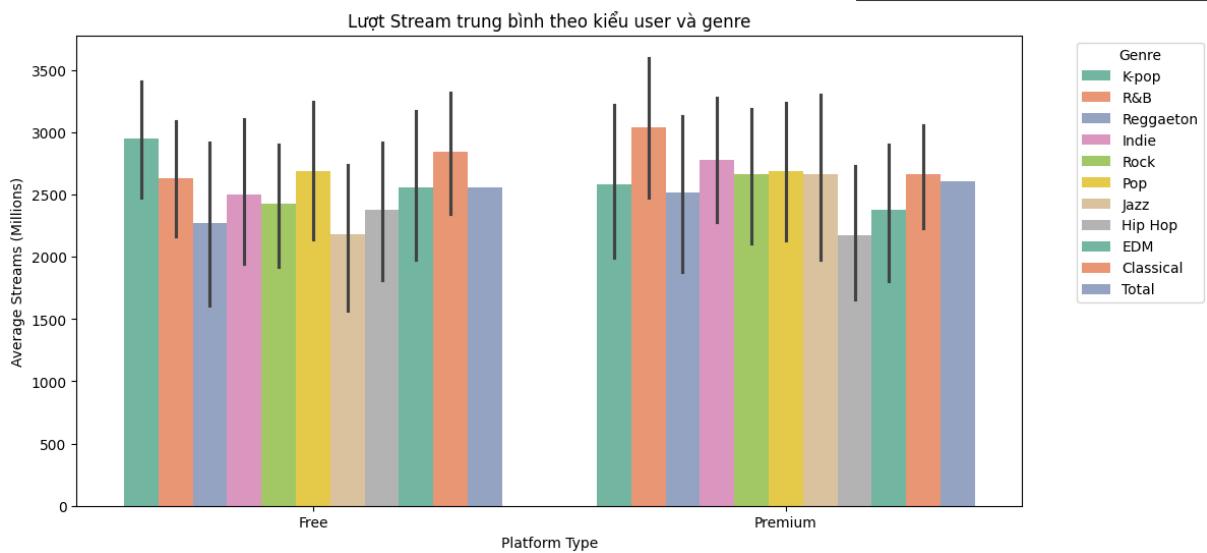
plt.figure(figsize=(12, 6))
sns.barplot(
    data=df_for_plot,
    x="Platform Type",
    y="Total Streams (Millions)",
    hue="Genre",
    palette="Set2",
    estimator="mean"
)
plt.title("Luật Stream trung bình theo kiểu user và genre")
plt.xlabel("Platform Type")
plt.ylabel("Average Streams (Millions)")
plt.legend(title="Genre", bbox_to_anchor=(1.05, 1), loc="upper left")
plt.show()

```

Hình 3.25: Code thực thi



Hình 3.26: Phân phối tỉ lệ Skip rate theo Genre



Hình 3.27: Lượt Stream trung bình theo kiểu user và genre

Albums_Info_Clean.csv

Từ file csv thu được từ reccobeats, ta có được 1 số phân tích và cái nhìn

tổng quan như sau:

```
df = pd.read_csv("/content/Albums_Info_Clean.csv")

#Top nghệ sĩ theo độ phổ biến
top_artists = df.groupby("artist_name")["popularity"].mean().sort_values(ascending=False).head(10)
print("Top 10 nghệ sĩ theo độ phổ biến (trung bình):")
print(top_artists)

plt.figure(figsize=(10,6))
top_artists.plot(kind='bar', color='purple', edgecolor='black')
plt.xlabel("Nghệ sĩ")
plt.ylabel("Popularity (mean)")
plt.xticks(rotation=45, ha='right')
plt.show()

#Popularity distribution
plt.figure(figsize=(8,5))
df['popularity'].hist(bins=30, color='skyblue', edgecolor='black')
plt.xlabel("Popularity")
plt.ylabel("Số lượng album")
plt.title("Phân bố độ phổ biến của album")
plt.show()

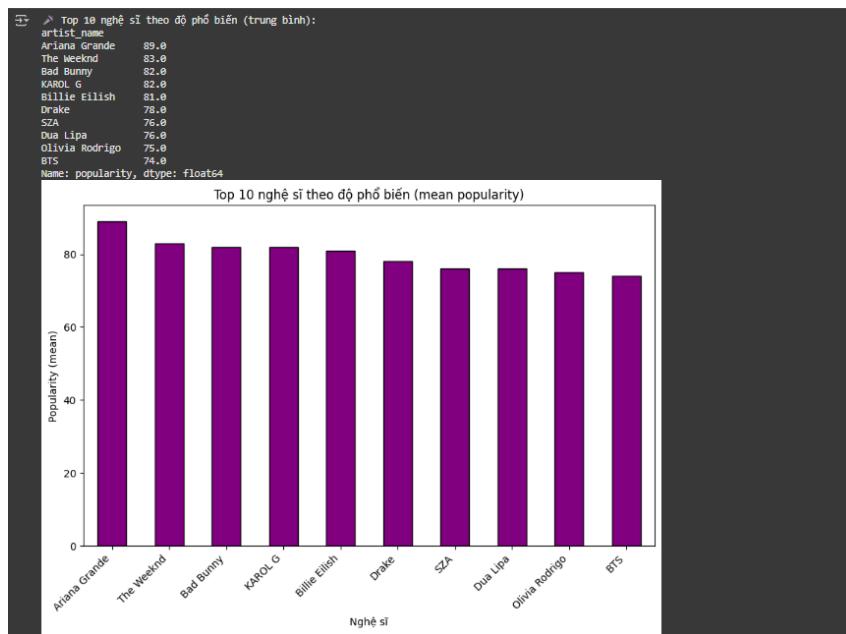
#Số lượng album phát hành theo năm
albums_per_year = df['release_year'].value_counts().sort_index()

plt.figure(figsize=(12,6))
albums_per_year.plot(kind='bar', color='orange', edgecolor='black')
plt.xlabel("Năm")
plt.ylabel("Số album")
plt.xticks(rotation=45)
plt.show()

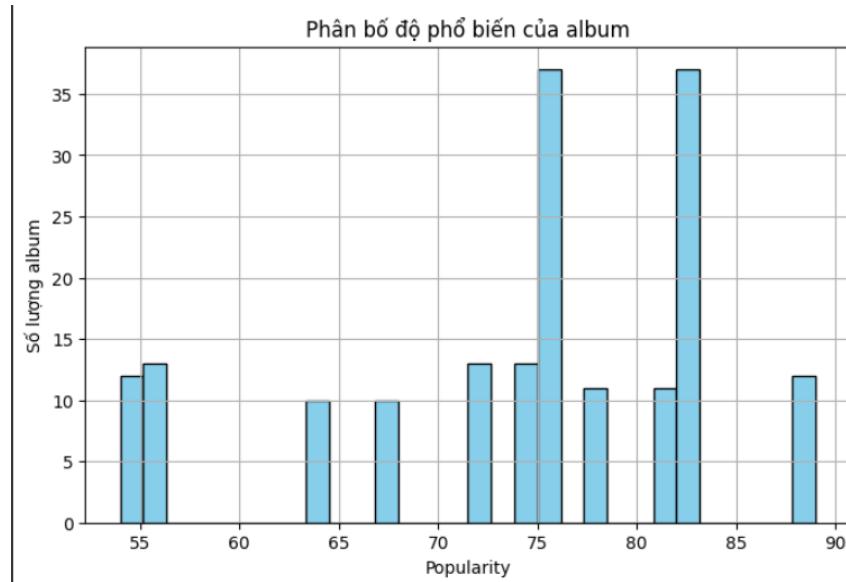
#Label nào phát hành nhiều album nhất
top_labels = df['label'].value_counts().head(10)
print("\nTop 10 label phát hành nhiều album nhất:")
print(top_labels)

plt.figure(figsize=(10,6))
top_labels.plot(kind='bar', color='green', edgecolor='black')
plt.title("Top 10 label phát hành nhiều album nhất")
plt.xlabel("Label")
plt.ylabel("Số album")
plt.xticks(rotation=45, ha='right')
plt.show()
```

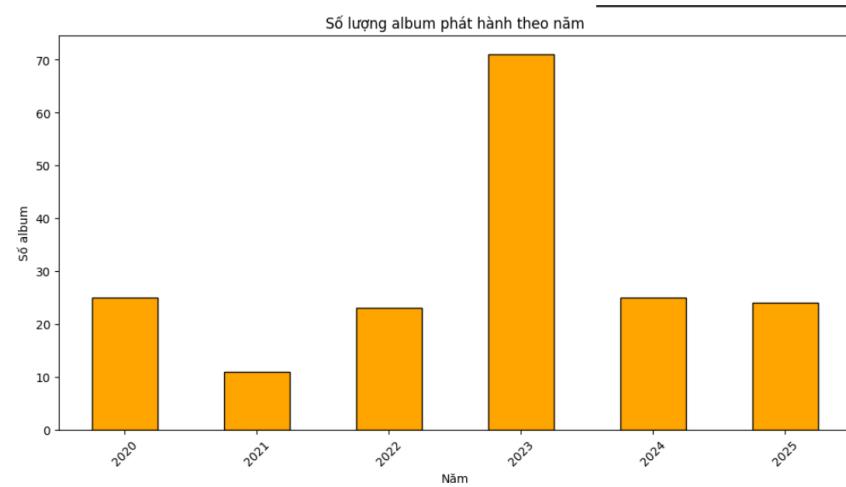
Hình 3.28: Code thực thi



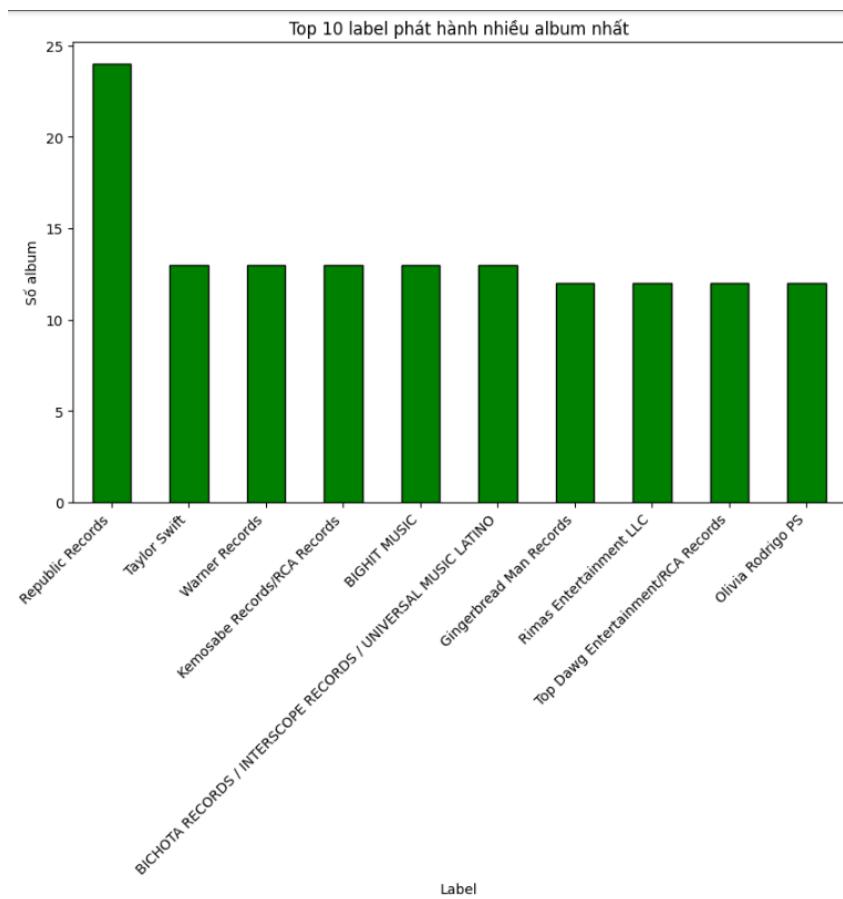
Hình 3.29: Top 10 nghệ sĩ theo độ phổ biến



Hình 3.30: Phân bố độ phổ biến của album



Hình 3.31: Số lượng album phát hành theo năm



Hình 3.32: Top 10 label phát hành nhiều album nhất

3.2.5 Mục tiêu:

- Với những gì dataset này cung cấp ta có thể thấy người dùng có tỷ lệ bỏ qua (skip rate) tăng và thời gian nghe giảm có thể được xem là rủi ro có thể rời bỏ nền tảng này hoặc không cùm nhu cầu với premium. Kết hợp với lịch sử nghe nhạc, sở thích thể loại (genre preferences) và mức độ hoạt động trên ứng dụng, Spotify có thể dự đoán và ngăn chặn rủi ro rời bỏ bằng mô hình học máy.
- Dự đoán bài hát tiềm năng trở thành hit dựa trên skip rate, thời gian nghe trung bình, số lượt stream ban đầu, mức độ lan truyền theo khu vực.

- Dự báo doanh thu dựa trên lượt stream trung bình và xu hướng theo thời gian, đề xuất nghệ sĩ hoặc thể loại nhạc phù hợp cho người dùng dựa trên hành vi nghe trước đó.
- Xác định tần suất bỏ qua (skip rate) của từng thể loại → đo lường sự hấp dẫn của nhạc.

3.3 Nguồn 3: Billboard Hot 100

3.3.1 Giới thiệu dữ liệu

Bối cảnh

- Từ năm 2025, Spotify đã ngừng công khai dữ liệu bảng xếp hạng **Global Top 200** hằng tuần. Điều này khiến việc thu thập dữ liệu gốc phục vụ cho phân tích trở nên khó khăn nếu chỉ dựa vào nguồn chính thức từ Spotify.

- Để khắc phục hạn chế này, nhóm lựa chọn sử dụng **Billboard Hot 100** làm nguồn dữ liệu thay thế. Đây là bảng xếp hạng uy tín, được cập nhật hằng tuần vào Chủ Nhật, phản ánh mức độ phổ biến của các ca khúc trên thị trường âm nhạc Mỹ và có sức ảnh hưởng toàn cầu.

Dữ liệu thu thập ban đầu

- Pipeline được xây dựng bằng Python, sử dụng thư viện **billboard** để tự động tải danh sách **Top 100 ca khúc mỗi tuần trong năm 2025**.

- Các trường thông tin thu được gồm:

- **week:** ngày Chủ Nhật của tuần (chuẩn hóa mốc thời gian).

- **rank:** thứ hạng của bài hát trong tuần.

- **title:** tên bài hát.

- **artist:** nghệ sĩ thể hiện chính.

- Bộ dữ liệu này được lưu trữ ban đầu trong tệp, làm nền tảng cho các bước **bổ sung dữ liệu (enrichment)** ở giai đoạn sau.

Hạn chế của dữ liệu gốc

- Billboard chỉ cung cấp **thông tin xếp hạng và metadata cơ bản** (tựa đề, nghệ sĩ), chưa có định danh duy nhất để liên kết với dữ liệu ngoài.

- Hoàn toàn thiếu vắng **đặc trưng âm thanh (audio features)** và **thể loại (genres)** – trong khi đây lại là các yếu tố then chốt để phân tích sâu hơn về “công thức tạo hit” và “xu hướng thể loại theo thời gian”.

3.3.2 Thu thập và bổ sung dữ liệu

Giải pháp từng bước

1. Thu thập Billboard Hot 100

- Sinh danh sách ngày Chủ Nhật trong năm 2025 và tải dữ liệu xếp hạng Hot 100 cho từng tuần.
- **Kết quả:** dữ liệu gốc đã có thứ hạng bài hát theo tuần nhưng chưa đủ thông tin để phân tích chuyên sâu.

2. Tìm Spotify track ID

- Với mỗi cặp (title, artist) từ Billboard, gọi **Spotify Search API** để tìm `spotify_id`.
- **Kết quả:** bổ sung khóa định danh duy nhất giúp liên kết với các API khác.

3. Lấy audio features từ ReccoBeats

- Gọi **ReccoBeats API** để lấy các đặc trưng âm nhạc: `danceability`, `energy`, `tempo`, `valence`, `speechiness`, `acousticness`, `instrumentalness`, `liveness`, `loudness`, `key`, `mode`.
- Do API giới hạn batch, pipeline được thiết kế để chia nhỏ các request (`BATCH_SIZE = 30`).
- **Kết quả:** dataset đã mô tả được “chất âm” của từng bài hát.

4. Lấy genres từ Spotify API

- Sử dụng `spotify_id` để gọi endpoint `track/artist` nhằm lấy thể loại (genres) của nghệ sĩ chính.
- Nếu genres trống thì gán NaN.
- **Kết quả:** dataset cuối cùng đã có thêm nhãn **thể loại âm nhạc** cho từng ca khúc.

3.3.3 Phân tích dữ liệu ban đầu

Lý do

- Sau khi đã bổ sung dữ liệu từ nhiều nguồn khác nhau, cần tiến hành phân tích dữ liệu ban đầu (Exploratory Data Analysis – EDA) để hiểu rõ cấu trúc, chất lượng cũng như các vấn đề tiềm ẩn trước khi bước vào giai đoạn làm sạch và trích xuất đặc trưng.

Giải pháp

1 Tổng quan dữ liệu

- Thống kê số dòng, số cột, tên các trường dữ liệu.
- Kiểm tra kiểu dữ liệu, số lượng giá trị bị thiếu (Null), số dòng trùng lặp.
- Tính toán các thống kê mô tả cơ bản cho dữ liệu số.

Số dòng: 3900, Số cột: 17

Column	DataType	Meaning	How to Use
week	object	Ngày Chủ Nhật của tuần trong BXH Billboard (YYYY-MM-DD).	Mốc thời gian để phân tích xu hướng theo tuần.
rank	int64	Thứ hạng bài hát (1 = cao nhất, 100 = thấp nhất).	Phân tích Top 10/50, độ độ bền hạng, vẽ timeline.
title	object	Tên bài hát.	Hiển thị trên dashboard, kết hợp với week để vẽ xu hướng.
artist	object	Tên nghệ sĩ chính.	Phân tích Top Artist, số lần vào BXH, xu hướng hợp tác.
spotify_id	object	ID duy nhất của track trên Spotify.	Làm khóa để join với Spotify API/ReccoBeats.
danceability	float64	Mức độ dễ nhảy (0–1).	Đặc trưng hit songs, cao trong Pop/Dance.

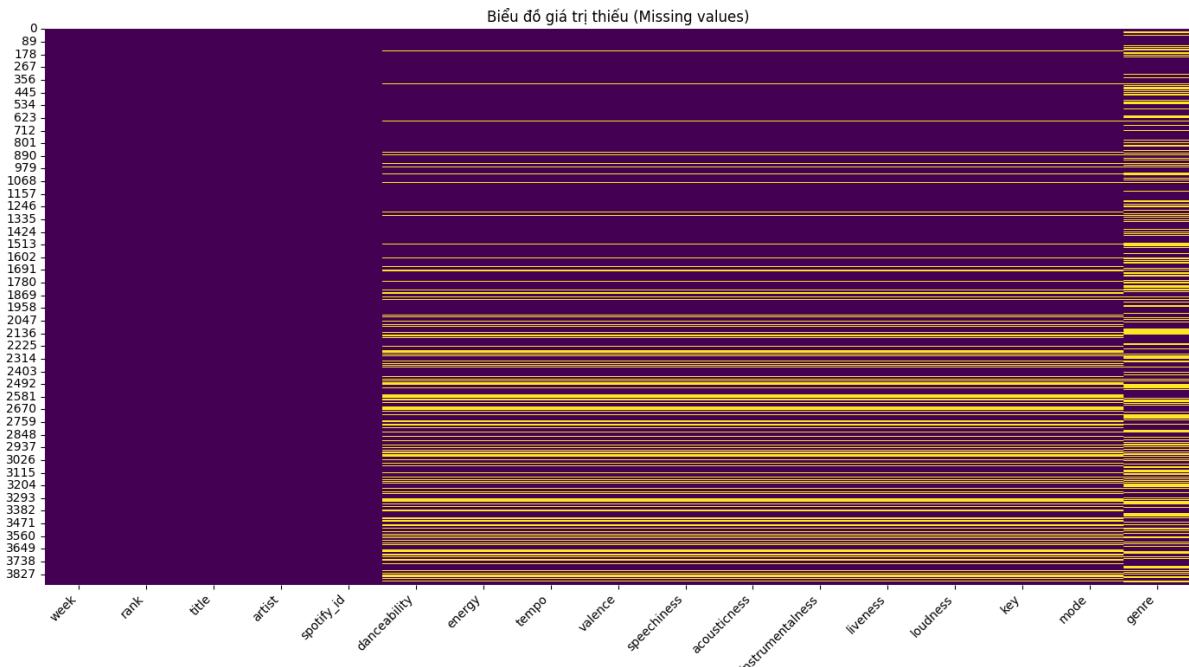
Column	DataType	Meaning	How to Use
energy	float64	Độ “năng lượng” của bài hát (0–1).	So sánh hit vs non-hit, nhận diện nhạc sôi động.
tempo	float64	Nhip độ bài hát (beats per minute – BPM)	Phân tích BPM phổ biến trong các hit.
valence	float64	Mức độ tích cực/vui vẻ (0–1).	Phân biệt nhạc vui (valence cao) với ballad (thấp).
speechiness	float64	Mức độ giọng nói (0–1).	Cao trong Rap/Hip-hop, thấp trong Pop/Ballad.
acousticness	float64	Mức độ acoustic (0–1).	Ballad/Indie cao; Pop/EDM thấp.
instrumentalness	float64	Khả năng là nhạc không lời (0–1).	EDM/nhạc nền cao, Pop vocal thấp.
liveness	float64	Mức độ biểu diễn live (0–1).	Phát hiện bài hát thu live concert.
loudness	float64	Độ lớn trung bình (dB).	Kết hợp với energy để phân tích phong cách sản xuất.
key	float64	Tông nhạc (0–11).	Phân tích xu hướng tông nhạc ưa chuộng.
mode	float64	Thang âm: 1=Major, 0=Minor.	So sánh mood Major vs Minor.

Column	DataType	Meaning	How to Use
genre	object	Thể loại chính (Spotify API).	Nhóm xu hướng theo genre, Top genres theo rankstreams.

Bảng 3.3: Bảng mô tả dữ liệu

2 Trực quan giá trị thiếu

- Vẽ heatmap để quan sát sự phân bố các giá trị bị thiếu theo cột.
- Nhận diện các cột thiếu dữ liệu nhiều, cần xử lý trong bước làm sạch.



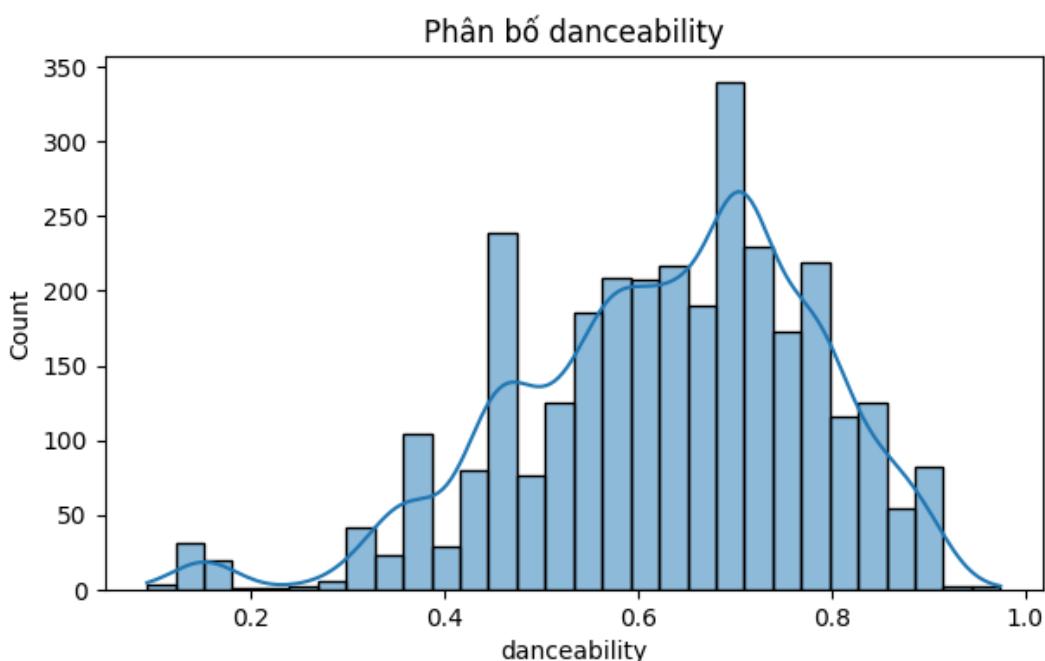
Hình 3.33: Biểu đồ heatmap thể hiện các giá trị thiếu trong dữ liệu

Biểu đồ heatmap trên cho thấy một số cột dữ liệu xuất hiện nhiều giá trị thiếu (các vạch sáng). Cụ thể, các trường liên quan đến audio

features (như danceability, energy, tempo, valence, speechiness, acousticness, instrumentalness, liveness, loudness) và trường genre có tỷ lệ thiếu khá cao. Điều này xuất phát từ việc Billboard chỉ cung cấp dữ liệu thứ hạng, tên bài hát và nghệ sĩ, trong khi các đặc trưng âm nhạc và thể loại phải được bổ sung từ API ngoài. Việc trực quan hóa bằng heatmap giúp nhận diện nhanh các cột còn thiếu nhiều giá trị, từ đó đưa ra chiến lược xử lý hợp lý trong bước làm sạch dữ liệu.

3 Phân bố các thuộc tính âm nhạc

- Vẽ histogram kèm KDE cho các thuộc tính quan trọng: danceability, energy, tempo, valence, speechiness, acousticness, instrumentalness, liveness, loudness.
- Giúp hiểu rõ phân phối của từng đặc trưng.



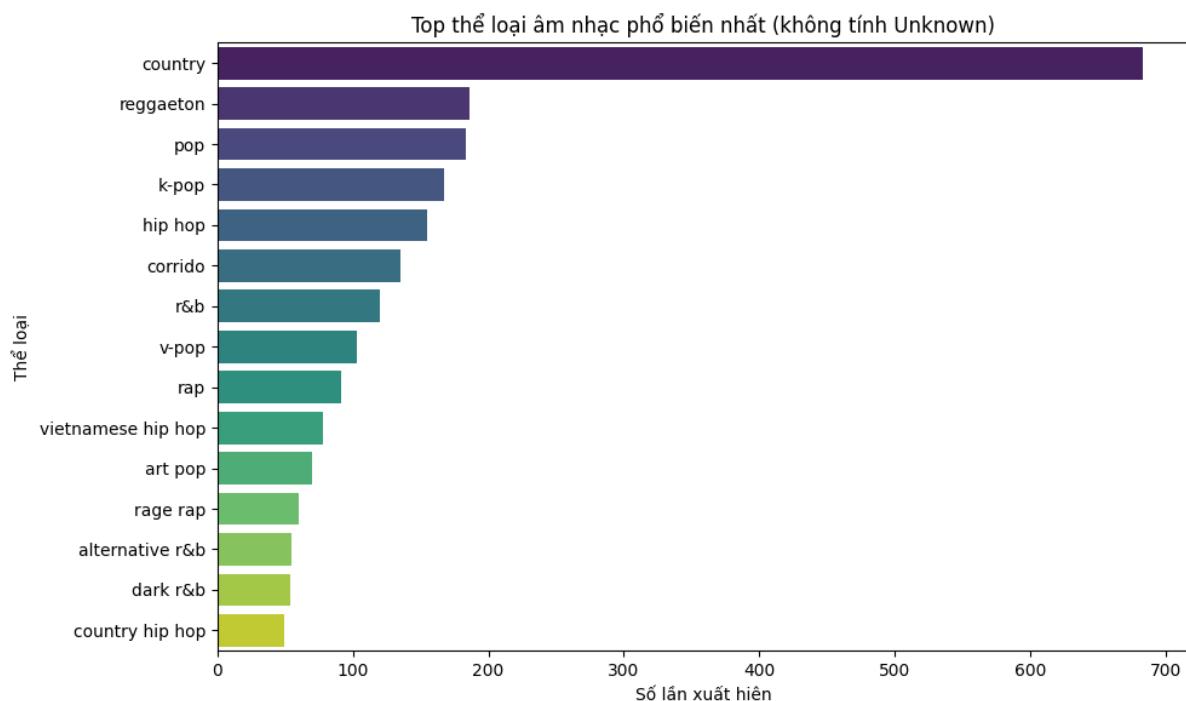
Hình 3.34: Biểu đồ phân bố danceability của các bài hát.

Biểu đồ trên cho thấy đa số các bài hát có giá trị danceability

tập trung trong khoảng 0.5 – 0.8. Điều này phản ánh rằng các ca khúc phổ biến trên thị trường âm nhạc toàn cầu thường có tiết tấu dễ nhảy, phù hợp với thị hiếu nghe nhạc của công chúng. Những bài hát có **danceability** quá thấp hoặc quá cao chỉ chiếm tỷ lệ nhỏ, cho thấy thị trường ưu tiên sự cân bằng giữa khả năng nhảy và tính giai điệu.

4 Phân bố thể loại âm nhạc

- Loại bỏ nhãn NaN, sau đó thống kê lần xuất hiện của các thể loại.



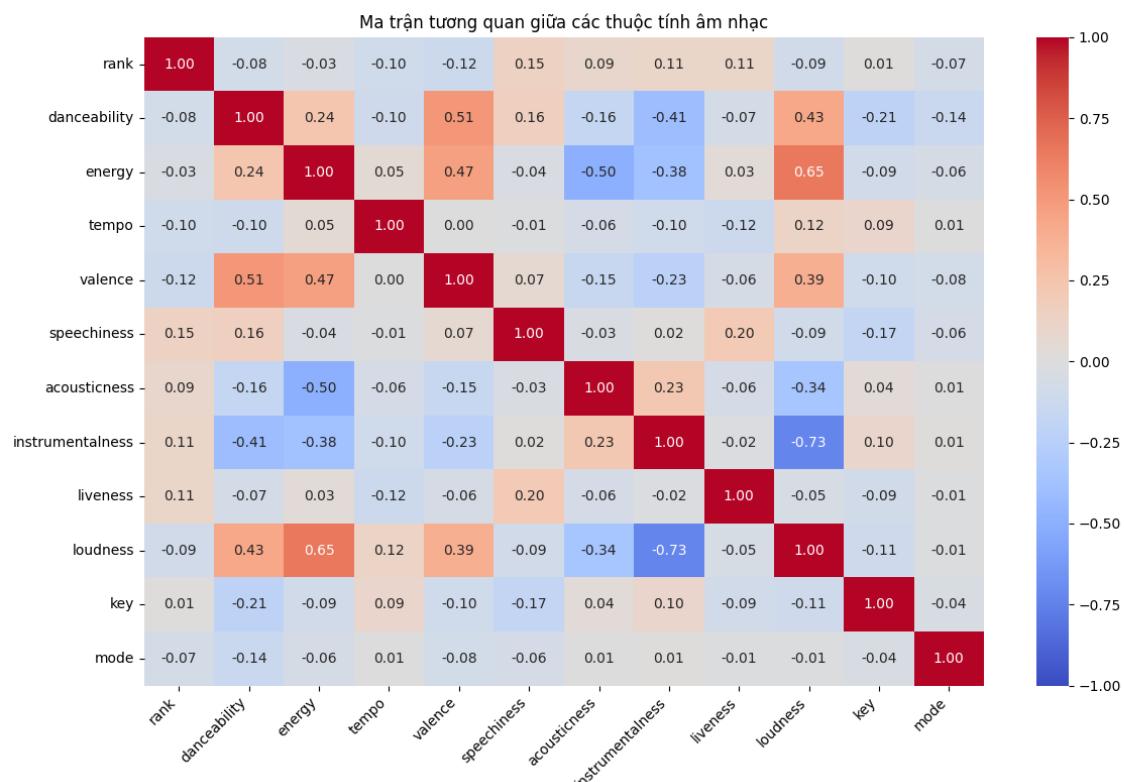
Hình 3.35: Top 15 thể loại âm nhạc phổ biến nhất .

Biểu đồ trên cho thấy thể loại **Country** chiếm ưu thế vượt trội với gần 700 lần xuất hiện, cao hơn hẳn so với các thể loại còn lại. Các dòng nhạc như **Reggaeton**, **Pop**, **K-pop**, **Hip Hop** cũng nằm trong nhóm dẫn đầu, phản ánh xu hướng toàn cầu hiện nay. Đáng chú ý,

V-Pop và **Vietnamese Hip Hop** cũng có sự hiện diện nhất định, cho thấy nhạc Việt đang dần khẳng định vị thế trong thị trường quốc tế. Sự phân bố này gợi ý rằng các doanh nghiệp và nghệ sĩ có thể ưu tiên đầu tư vào các thể loại đang thịnh hành như Country, Pop hay K-pop để tối ưu cơ hội tiếp cận khán giả rộng hơn.

5 Ma trận tương quan

- Tính hệ số tương quan giữa các biến số.
- Phát hiện mối quan hệ mạnh.



Hình 3.36: Ma trận tương quan giữa các thuộc tính âm nhạc.

Biểu đồ trên cho thấy một số mối quan hệ nổi bật giữa các đặc trưng âm nhạc. **Energy** có tương quan dương mạnh với **Loudness** (0.65), phản ánh rằng các bài hát có cường độ năng lượng cao thường đi kèm với âm lượng lớn. Ngược lại, **Acousticness**

có tương quan âm với **Energy** (-0.50), cho thấy nhạc acoustic thường có mức năng lượng thấp. **Danceability** và **Valence** cũng có mối quan hệ dương (0.51), gợi ý rằng những ca khúc dễ nhảy thường mang lại cảm xúc tích cực. Những phát hiện này hỗ trợ việc lý giải tại sao một số thuộc tính đóng vai trò quan trọng trong việc tạo nên “công thức” của các bài hit toàn cầu.

Kết quả / Ý nghĩa

- Bộ dữ liệu đã được mô tả đầy đủ về cấu trúc và chất lượng, cung cấp cái nhìn tổng quan cần thiết trước khi làm sạch.
- Phát hiện được những vấn đề cần xử lý trong bước kế tiếp (missing values, dữ liệu trùng lặp, định dạng chưa thống nhất).
- Các insight sơ bộ: nhạc dễ nhảy (danceability) chiếm ưu thế, energy và loudness cao, acousticness thấp; streams có quan hệ nghịch mạnh với rank. Đây là cơ sở định hướng cho bước làm sạch dữ liệu và xây dựng đặc trưng.

3.3.4 Làm sạch dữ liệu (Data Cleaning)

Lý do

- Sau khi phân tích dữ liệu ban đầu mặc dù không có giá trị trùng lặp, nhưng nhận thấy còn tồn tại nhiều giá trị thiếu (missing values), một số cột chưa đúng kiểu dữ liệu. - Nếu không xử lý, các vấn đề này sẽ gây sai lệch cho kết quả phân tích và trực quan hóa ở các bước tiếp theo.

Giải pháp

1 Chuẩn hóa kiểu dữ liệu

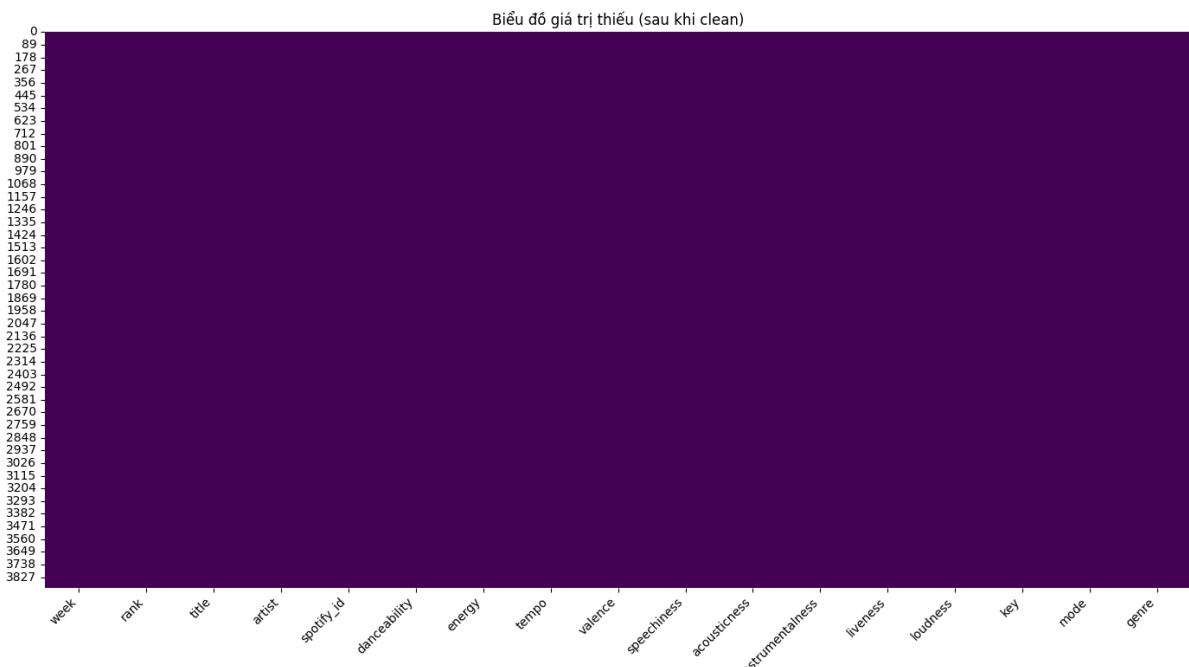
- Cột `week` được chuyển về định dạng `datetime`.
- Cột `rank` được ép kiểu về số nguyên `Int64`.

2 Xử lý missing values

- Với các cột số (`danceability`, `energy`, `tempo`, `valence`, `speechiness`, `acousticness`, `instrumentalness`, `liveness`, `loudness`, `key`, `mode`): điền giá trị trung bình (mean).
- Với cột `genre`: điền giá trị mặc định là `Unknown` nếu bị thiếu.

3 Đánh giá lại dữ liệu sau khi làm sạch

- Sử dụng heatmap để trực quan hóa, xác nhận rằng toàn bộ giá trị thiếu đã được xử lý.



Hình 3.37: Heatmap giá trị thiếu sau khi làm sạch dữ liệu.

Biểu đồ trên cho thấy toàn bộ các cột dữ liệu đều đã được xử lý đầy đủ, không còn giá trị thiếu. Điều này chứng minh quy trình

làm sạch dữ liệu đã hoàn tất, tập dữ liệu đã đồng nhất và sẵn sàng cho các bước phân tích đặc trưng và xu hướng tiếp theo.

Kết quả và Ý nghĩa

- Bộ dữ liệu đã được chuẩn hóa và không còn giá trị thiếu.
- Các cột dữ liệu quan trọng đều ở định dạng chính xác, đảm bảo độ tin cậy cho các phép phân tích.
- Việc gán giá trị trung bình cho các cột số và điền `Unknown` cho thẻ loại giúp giữ lại toàn bộ dữ liệu, không gây mất mát bản ghi.
- Dữ liệu sạch, đồng nhất, sẵn sàng cho bước tiếp theo: **Tạo đặc trưng (Feature Engineering)**.

3.3.5 Tạo đặc trưng (Feature Engineering)

Lý do - Dữ liệu sau khi làm sạch vẫn yếu mô tả thông tin cơ bản như thứ hạng, nghệ sĩ và đặc trưng âm nhạc. Tuy nhiên, để phân tích chuyên sâu hơn, cần có các biến phản ánh mức độ bền vững của bài hát trên BXH, đặc điểm nghệ sĩ (solo/hợp tác), cũng như yếu tố thời gian (mùa trong năm). - Việc tạo thêm đặc trưng mới sẽ giúp khám phá xu hướng ẩn và nâng cao giá trị phân tích.

Giải pháp

1 Đặc trưng theo bài hát (Track-level features)

- `weeks_on_chart_total`: tổng số tuần xuất hiện trên BXH.
- `weeks_on_chart_cum`: số tuần tích lũy đến thời điểm hiện tại.
- `peak_rank`: thứ hạng cao nhất mà bài hát đạt được.

- **avg_rank:** thứ hạng trung bình của bài hát.
- **is_top10:** 1 nếu bài hát lọt Top 10 trong tuần, ngược lại là 0.
- **is_new:** 1 nếu đây là tuần đầu tiên xuất hiện trên BXH, ngược lại là 0.

2 Đặc trưng theo nghệ sĩ (Artist-based features)

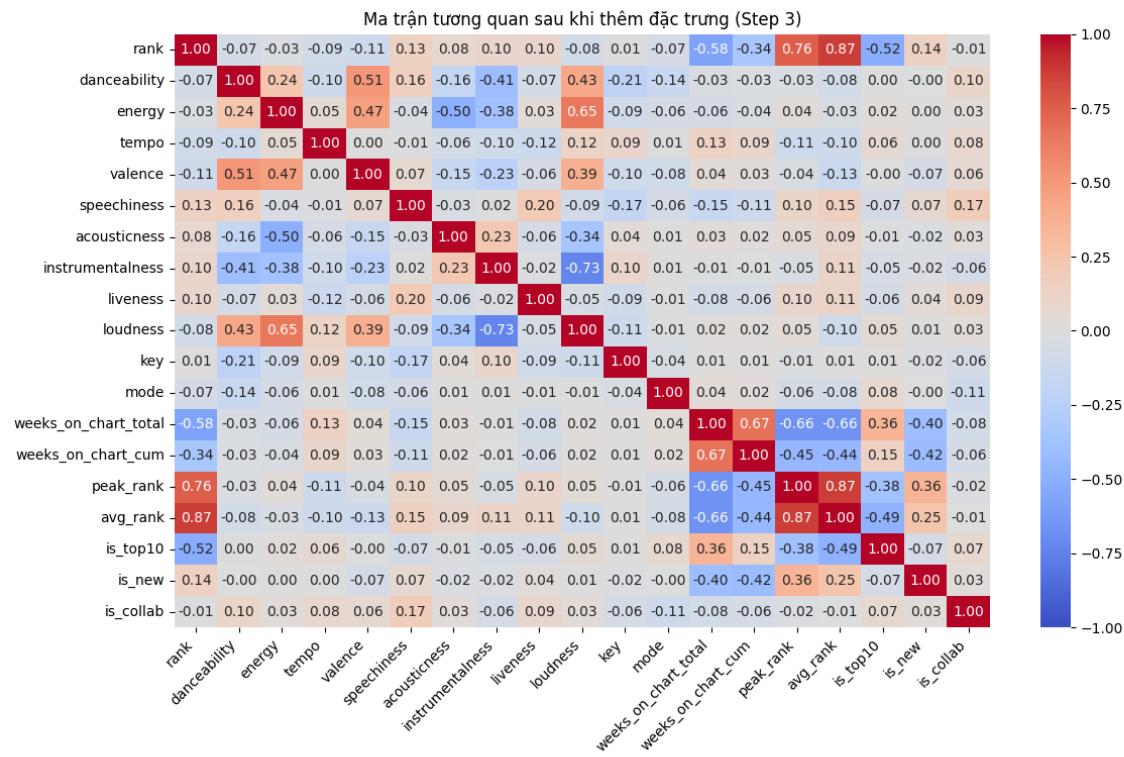
- **is_collab:** 1 nếu bài hát có sự hợp tác (nhiều nghệ sĩ, có “feat.”, “&” hoặc dấu phẩy trong tên), ngược lại là 0.

3 Đặc trưng theo thời gian (Time-based features)

- **season:** phân loại theo mùa (Spring, Summer, Fall, Winter) dựa trên tháng trong năm.
- Hỗ trợ phân tích yếu tố mùa vụ ảnh hưởng đến sự phổ biến của các thể loại nhạc.

Kết quả và Ý nghĩa

- Bộ dữ liệu sau khi thêm đặc trưng đã phản ánh đầy đủ hơn vòng đời của bài hát, đặc điểm nghệ sĩ và bối cảnh thời gian.
- Các đặc trưng như **peak_rank**, **weeks_on_chart_total** giúp đo lường mức độ bền vững của hit.
- Đặc trưng **is_collab** cho phép đánh giá vai trò của hợp tác nghệ sĩ trong việc tạo ra các ca khúc thành công.
- Đặc trưng **season** mở ra khả năng phân tích tác động của mùa vụ đến xu hướng nghe nhạc.



Hình 3.38: Ma trận tương quan sau khi thêm đặc trưng

Biểu đồ trên cho thấy các đặc trưng mới có mối quan hệ rõ rệt với các trường dữ liệu gốc: `weeks_on_chart_total` và `weeks_on_chart_cum` tương quan âm với `rank`, chứng tỏ các bài hát trụ hạng lâu thường có thứ hạng cao. `peak_rank` và `avg_rank` có tương quan mạnh với `rank`, khẳng định thứ hạng hàng tuần phản ánh hiệu suất tổng thể. Đặc trưng `is_top10` cũng có quan hệ nghịch với `rank`, phù hợp với định nghĩa lọt Top 10. Trong khi đó, `is_collab` và `is_new` có tương quan yếu, nhưng vẫn là những yếu tố bổ sung quan trọng cho phân tích xu hướng và hành vi âm nhạc.

3.3.6 Phân tích xu hướng (Trend Analysis)

Bối cảnh

Phần phân tích này tập trung vào dữ liệu **Billboard Hot 100** tại **thị trường Mỹ năm 2025**, được bổ sung thông tin từ Spotify API và ReccoBeats. Do đặc thù Billboard Hot 100 phản ánh riêng thị trường Mỹ, các insight thu được sẽ cho thấy rõ xu hướng âm nhạc chủ đạo tại Mỹ thay vì toàn cầu.

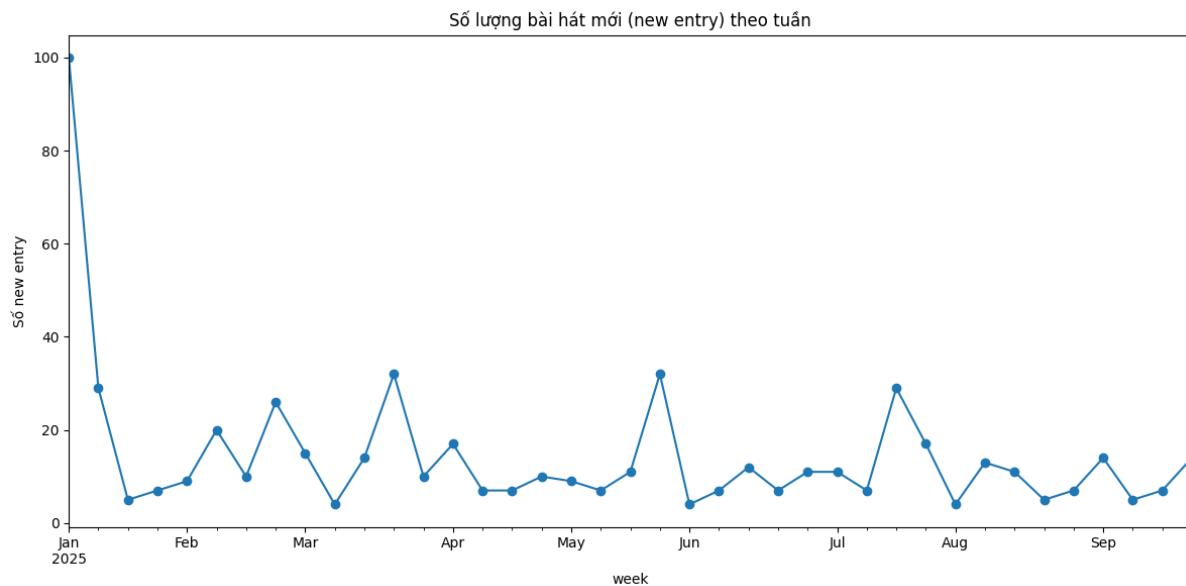
Lý do

- Sau khi dữ liệu đã được làm sạch và bổ sung đặc trưng, bước tiếp theo là tiến hành phân tích xu hướng. - Mục tiêu nhằm phát hiện các mẫu hành vi nổi bật: cách bài hát mới xuất hiện, độ bền hit, thể loại phổ biến trong Top 10, vai trò của nghệ sĩ, yếu tố mùa vụ và mối quan hệ giữa độ bền (longevity) và vị trí cao nhất (peak rank).

Giải pháp

1 Xu hướng bài hát mới theo thời gian

- Đếm số lượng `is_new` theo tuần để theo dõi số lượng bài hát mới gia nhập BXH.
- Trực quan bằng line chart cho thấy tuần nào có nhiều ca khúc debut nhất.

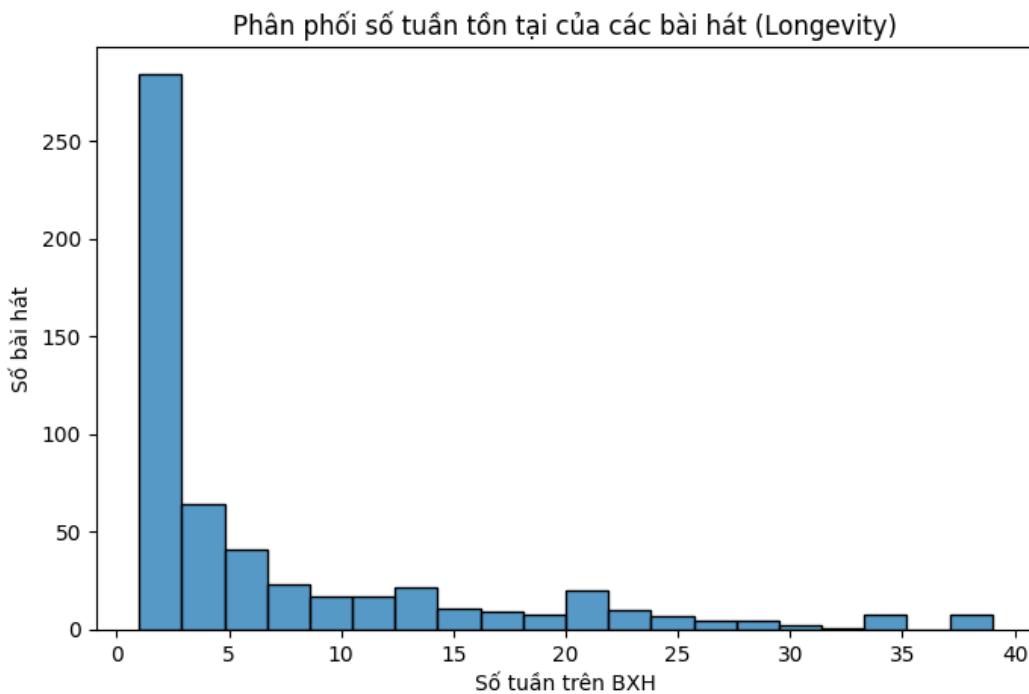


Hình 3.39: Số lượng bài hát mới (new entry) theo tuần trong năm 2025.

Biểu đồ cho thấy trong tuần đầu tiên của năm 2025 có hơn 100 bài hát mới xuất hiện trên BXH Billboard Hot 100 (hiện tượng “reset” đầu năm). Sau đó, số lượng new entry giảm mạnh và dao động quanh mức 5–30 bài mỗi tuần. Một số tuần có đột biến tăng (ví dụ cuối tháng 3, tháng 6, tháng 8), thường trùng với các đợt phát hành album lớn hoặc cao điểm mùa lễ hội. Điều này phản ánh tính cạnh tranh của thị trường âm nhạc.

2 Phân phối độ bền (longevity)

- Vẽ histogram số tuần một bài hát tồn tại trên BXH.
- Giúp phân biệt hit “chớp nhoáng” và hit bền vững.

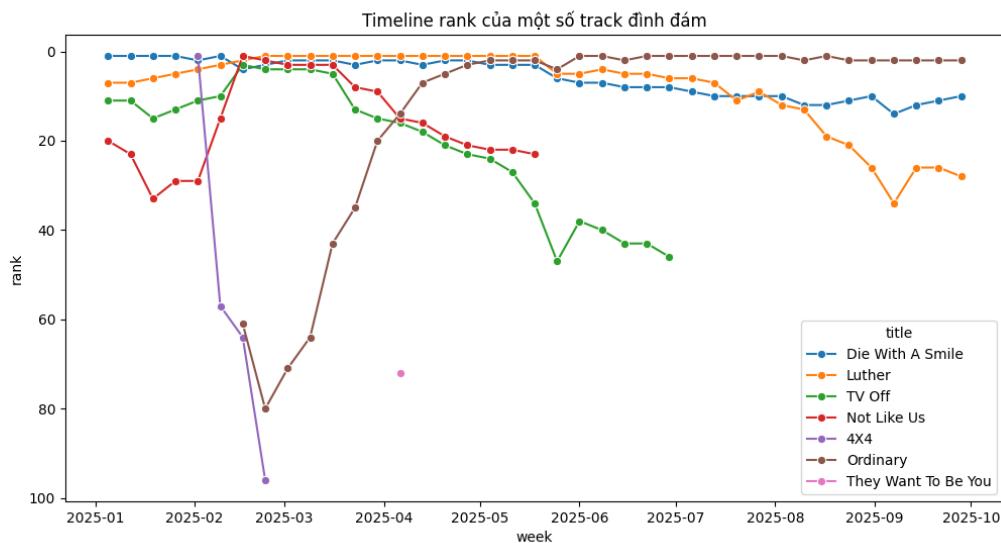


Hình 3.40: Phân phối số tuần tồn tại của các bài hát trên Billboard Hot 100.

Biểu đồ cho thấy phần lớn các bài hát chỉ xuất hiện trên BXH trong thời gian ngắn (khoảng 1–3 tuần), phản ánh đặc trưng cạnh tranh khốc liệt của thị trường âm nhạc. Tuy nhiên, vẫn có một nhóm nhỏ ca khúc duy trì được vị trí lâu dài (trên 20–30 tuần), được xem là những “hit bền vững” có sức hút mạnh mẽ. Kết quả này gợi ý rằng việc duy trì vị trí lâu dài trên BXH khó khăn hơn nhiều so với việc “đột phá” vào BXH ở tuần đầu tiên, và có thể phụ thuộc vào các yếu tố như độ nổi tiếng của nghệ sĩ, chiến dịch quảng bá, cũng như đặc điểm âm nhạc (energy, danceability, valence).

3 Timeline các hit đình đám

- Chọn ra một số bài có `peak_rank = 1`, vẽ line chart thứ hạng theo thời gian.
- Đảo ngược trục Y để thể hiện trực quan đường đi lên/xuống BXH.



Hình 3.41: Timeline thứ hạng của một số ca khúc đạt vị trí số 1 trên Billboard Hot 100.

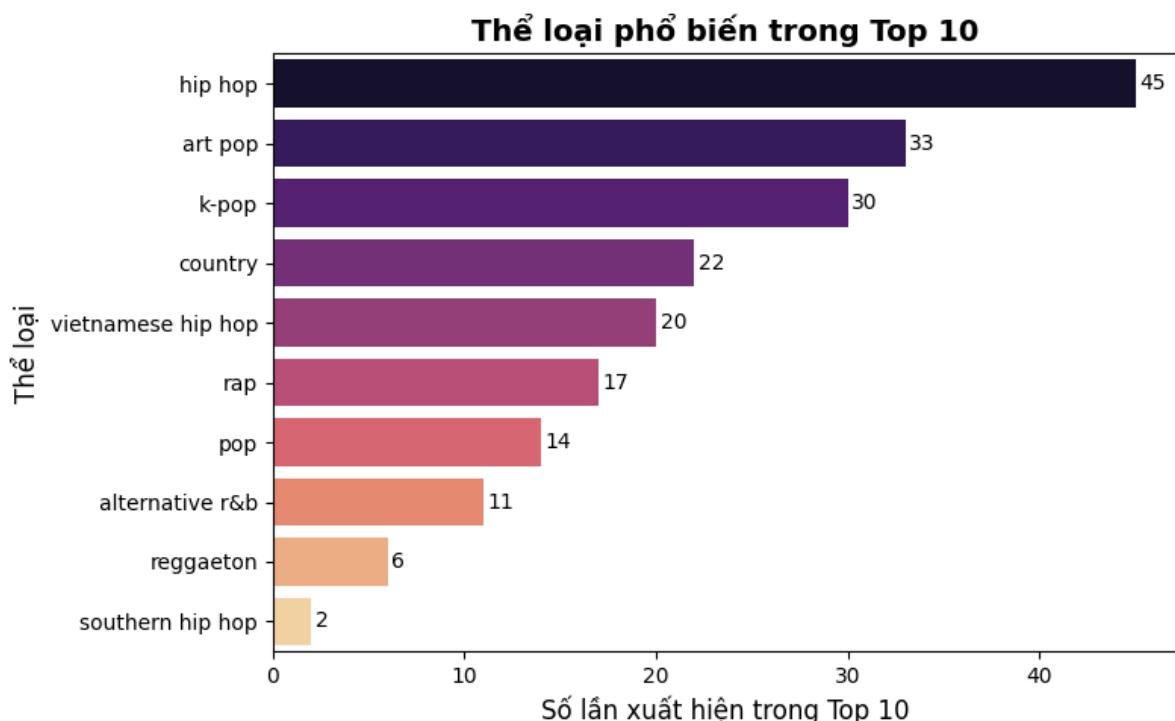
Biểu đồ trên minh họa quá trình lên xuống hạng của một số bài hit nổi bật trong năm 2025. Có thể quan sát rằng:

- Một số ca khúc như *Die With A Smile*, *Luther* giữ được vị trí cao trong thời gian dài, chứng tỏ sức hút ổn định và khả năng duy trì độ phổ biến.
- Các bài như *TV Off*, *Not Like Us* nhanh chóng leo lên Top 1 nhưng tụt hạng mạnh sau vài tuần, thể hiện đặc điểm “bung nổ ngắn hạn”.
- Một số bài khác có quỹ đạo đi lên dần (ví dụ *Ordinary*), cho thấy hiệu ứng lan tỏa muộn nhờ truyền thông hoặc viral.

Điều này phản ánh sự đa dạng trong vòng đời của một hit: có ca khúc “one-hit wonder” chỉ thịnh hành trong thời gian ngắn, trong khi một số khác có khả năng trụ vững nhiều tháng liền nhờ nền tảng fanbase và chiến lược quảng bá hiệu quả.

4 Phân tích thể loại trong Top 10

- Lọc các bài có `is_top10 = 1`, loại bỏ Unknown.
- Đếm tần suất xuất hiện của từng thể loại → Top 10 genres.



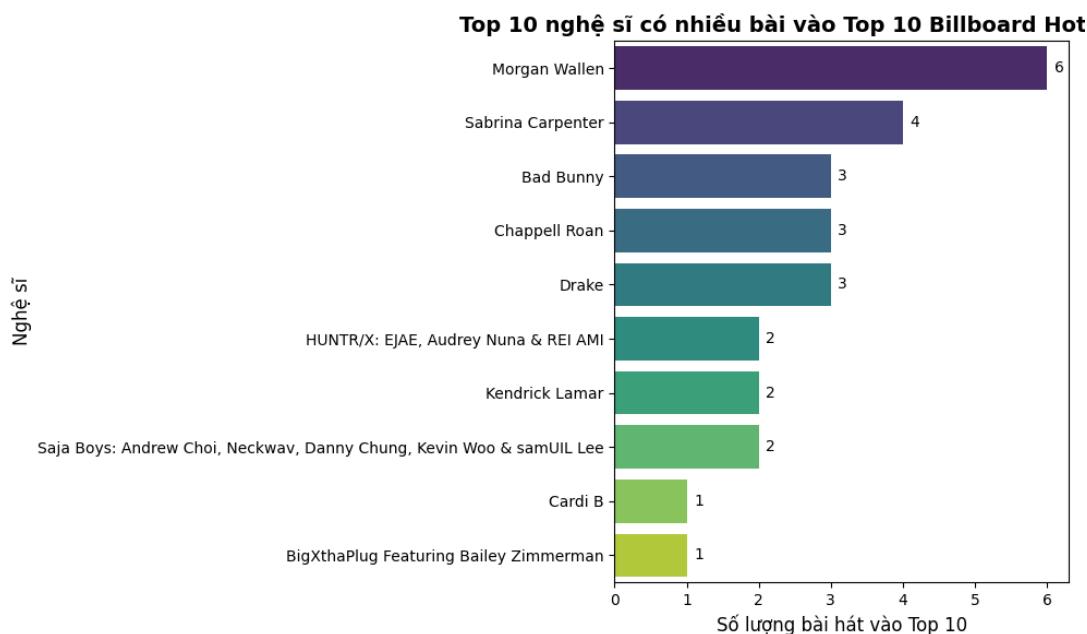
Hình 3.42: Top 10 thể loại âm nhạc phổ biến nhất trong BXH Billboard Hot 100.

Biểu đồ cho thấy **Hip Hop** là thể loại thống trị, xuất hiện tới 45 lần trong Top 10, tiếp theo là **Art Pop** và **K-pop** với lần lượt 33 và 30 lần. Các thể loại như **Country**, **Vietnamese Hip Hop**, **Rap** và **Pop** cũng giữ vị trí quan trọng nhưng ít xuất hiện hơn.

Điều này phản ánh sự đa dạng trong thị hiếu âm nhạc: Hip Hop và Pop vẫn giữ vai trò chủ đạo, nhưng những làn sóng mới như K-pop và Vietnamese Hip Hop đang có sự bứt phá mạnh mẽ, đóng góp đáng kể vào thị trường quốc tế. Ngoài ra, sự xuất hiện của thể loại nhỏ như **Southern Hip Hop** cho thấy một số “ngách” âm nhạc vẫn có thể vươn lên Top 10 trong những giai đoạn đặc biệt.

5 Top nghệ sĩ trong Top 10

- Nhóm theo nghệ sĩ, đếm số lượng ca khúc vào Top 10.
- Xếp hạng Top 10 nghệ sĩ có nhiều hit nhất.



Hình 3.43: Top 10 nghệ sĩ có nhiều ca khúc lọt vào Top 10 Billboard Hot 100 năm 2025.

Biểu đồ cho thấy **Morgan Wallen** dẫn đầu với 6 ca khúc vào Top 10, khẳng định vị thế vững chắc trên thị trường âm nhạc năm 2025. Theo sau là **Sabrina Carpenter** với 4 ca khúc, trong khi các tên tuổi đình đám như **Bad Bunny**, **Chappell Roan**, **Drake** đều có 3 bài lọt Top 10.

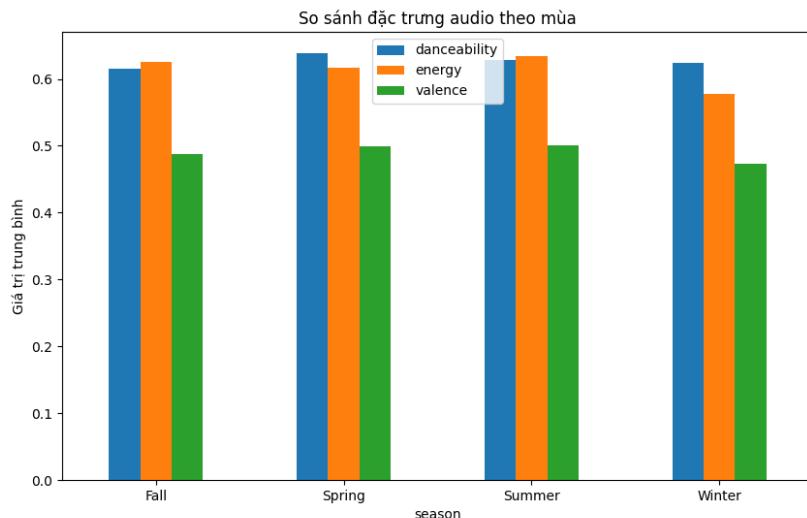
Ngoài ra, nhiều nghệ sĩ khác như **Kendrick Lamar**, nhóm **HUNTR/X**, hay các nhóm hợp tác đa nghệ sĩ (**Saja Boys**) cũng góp mặt với 2 ca khúc. Các nghệ sĩ nổi tiếng như **Cardi B** hay **Bailey Zimmerman** tuy chỉ có 1 ca khúc vào Top 10 nhưng vẫn cho thấy sức ảnh hưởng nhất định.

Điều này phản ánh rằng năm 2025 không chỉ có các tên tuổi kỳ cựu (Drake, Bad Bunny, Cardi B) mà còn chứng kiến sự vươn lên mạnh

mẽ của những gương mặt mới (Chappell Roan, Sabrina Carpenter), làm đa dạng bức tranh thị trường âm nhạc toàn cầu.

6 So sánh audio features theo mùa

- Tính trung bình danceability, energy, valence cho mỗi mùa (Spring, Summer, Fall, Winter).
- Vẽ bar chart để so sánh sự khác biệt.



Hình 3.44: So sánh giá trị trung bình của các audio features theo mùa.

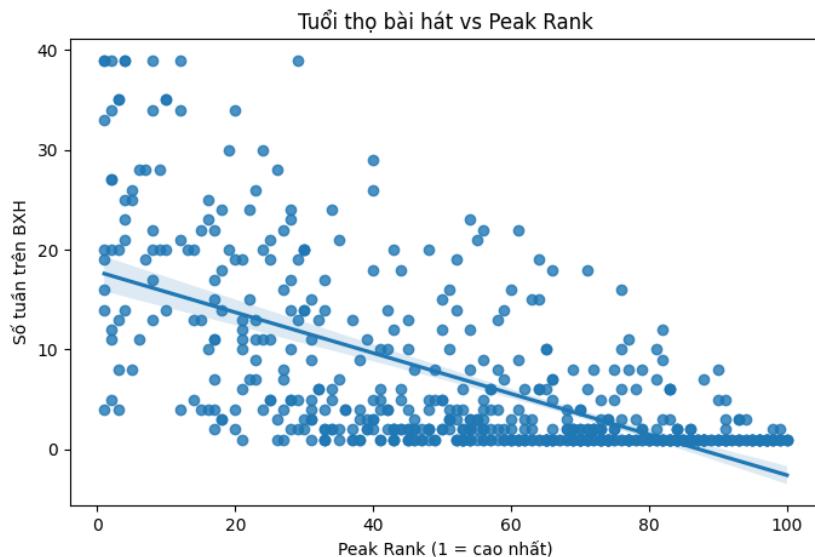
Biểu đồ cho thấy có sự khác biệt nhẹ về đặc trưng âm nhạc giữa các mùa:

- **Spring** và **Summer** có mức danceability và energy cao nhất, phản ánh xu hướng nghe nhạc sôi động, dễ nhảy trong mùa lễ hội và du lịch.
- **Fall** duy trì năng lượng cao nhưng valence thấp hơn, gợi ý sự xuất hiện nhiều hơn của các ca khúc buồn/ballad.
- **Winter** có energy và valence thấp nhất, cho thấy thiên hướng âm nhạc trầm lắng, sâu lắng phù hợp không khí cuối năm.

Điều này chứng minh yếu tố **mùa vụ** có tác động đến thị hiếu nghe nhạc: mùa Hè và Xuân ưu tiên nhạc sôi động, trong khi Thu và Đông phổ biến nhạc trữ tình, chậm rãi hơn.

7 Mối quan hệ Longevity và Peak Rank

- Vẽ scatter plot kèm regression line giữa `weeks_on_chart_total` và `peak_rank`.
- Đánh giá mối liên hệ giữa thứ hạng cao nhất và tuổi thọ bài hát.



Hình 3.45: Mối quan hệ giữa Peak Rank và số tuần tồn tại trên BXH.

Biểu đồ cho thấy có mối quan hệ nghịch giữa **Peak Rank** và **tuổi thọ bài hát** trên BXH. Cụ thể:

- Những ca khúc đạt `peak_rank` cao (từ 1–10) có xu hướng tồn tại nhiều tuần hơn trên BXH, thậm chí một số bài trụ vững hơn 30 tuần.
- Ngược lại, những bài chỉ đạt `peak` ở vị trí thấp (trên 50) thường rời BXH sau vài tuần ngắn ngủi.

- Regression line minh họa xu hướng này một cách rõ rệt: càng đạt vị trí cao, tuổi thọ trên BXH càng dài.

Điều này khẳng định rằng: **vị trí đỉnh cao nhất mà bài hát từng đạt được là một chỉ báo quan trọng cho sự bền vững của nó trên thị trường**. Những bản hit lớn không chỉ leo lên Top 1 mà còn duy trì sức hút lâu dài, trong khi các bài có thứ hạng trung bình khó giữ chân khán giả trong nhiều tuần.

3.3.7 Kết quả và Ý nghĩa

Mục tiêu bài toán

- Mục tiêu là xây dựng một **pipeline phân tích dữ liệu âm nhạc hoàn chỉnh**, giúp phát hiện các xu hướng nổi bật về hit songs, thể loại, nghệ sĩ, yếu tố mùa vụ và hành vi của thị trường.

Kết quả đạt được

- Hoàn thiện pipeline tự động với 4 bước: (i) *Thu thập dữ liệu*, (ii) *Làm sạch dữ liệu*, (iii) *Tạo đặc trưng*, (iv) *Phân tích xu hướng*.
- Bổ sung các đặc trưng quan trọng (audio features, genres, track-level, artist-based, time-based), làm cho dữ liệu đa chiều và giàu ý nghĩa hơn.
- Thực hiện EDA và trực quan hóa để đánh giá chất lượng dữ liệu, phát hiện vấn đề và xử lý kịp thời.
- Một số insight chính:
 - Các hit chủ yếu tập trung ở **Hip-hop, Pop, K-pop, Country**.

- Nghệ sĩ có nhiều hợp tác thường xuất hiện dày đặc trong Top 10.
- Ca khúc đạt **peak rank** cao (Top 1–5) thường có tuổi thọ lâu hơn trên BXH.
- Yếu tố mùa vụ ảnh hưởng đến đặc trưng nhạc: Hè/Xuân thiên về nhạc sôi động (danceability, energy cao), Thu/Đông phổ biến ballad trữ tình với valence thấp.

Ý nghĩa

- Bộ dữ liệu sau xử lý không chỉ phản ánh thứ hạng, mà còn lý giải được *nguyên nhân thành công của một ca khúc và xu hướng vận động của thị trường âm nhạc Mỹ*. - Pipeline này có thể mở rộng sang các năm khác hoặc thị trường khác, làm cơ sở cho phân tích so sánh xuyên quốc gia và ứng dụng trong **dự báo (predictive analytics)** về khả năng thành công của các ca khúc trong tương lai.

4 Phụ lục

4.1 Timeline công việc và bảng phân công nhiệm vụ

4.1.1 Timeline thực hiện đề tài

Bảng 4.1: Timeline thực hiện đề tài

Tuần	Nhóm cần làm gì
37	Giới thiệu đề tài, phân tích yêu cầu bài toán, tính cấp thiết

Tuần	Nhóm cần làm gì
38	Xác định nguồn dữ liệu, phân tích đặc điểm dữ liệu, đặt mục tiêu, phân chia công việc
39	Hoàn thiện phần phân tích yêu cầu (chuẩn bị cho L.0.1), tìm hiểu sơ bộ kiến trúc & công nghệ
40	Trình bày L.0.1 – Phân tích yêu cầu & bài toán thực tế
41	Bắt đầu thiết kế kiến trúc dữ liệu, mô hình ERD, chuẩn hóa dữ liệu
42	Hoàn thiện kiến trúc & mô hình dữ liệu, lựa chọn công nghệ, báo cáo tiến độ
43	Thi giữa kỳ, rà soát tiến độ
44	Hiện thực giải pháp: thiết kế CSDL, tạo bảng, trigger, index
45	Kiểm thử dữ liệu & hệ thống, so sánh với yêu cầu ban đầu
46	Hoàn thiện phần kiểm thử & đánh giá, chỉnh sửa hạn chế
47	Trình bày L.0.2 – Thiết kế & hiện thực giải pháp
48	Ứng dụng thực tế: giao diện, biểu đồ, trực quan hóa
49	Hoàn thiện kết luận & hướng phát triển, báo cáo tiến độ
50	Hoàn thành báo cáo + slide cuối kỳ

Tuần	Nhóm cần làm gì
51	Báo cáo cuối kỳ & nộp sản phẩm

4.1.2 Phân công nhiệm vụ dự kiến sau tuần 40

Tuần	Thành viên	Nhiệm vụ
41	Nguyễn Minh Nhựt Phạm Đình Phương Nam Đoàn Mạnh Tất Phạm Đức Hoài Nam	Thiết kế kiến trúc dữ liệu, vẽ ERD sơ bộ Tìm hiểu mô hình quan hệ, chuẩn bị ràng buộc dữ liệu Bắt đầu tiền xử lý dataset bằng Python/Pandas Khảo sát công nghệ triển khai (DB, web)
42	Nguyễn Minh Nhựt Phạm Đình Phương Nam Đoàn Mạnh Tất Phạm Đức Hoài Nam	Hoàn thiện ERD, mô hình dữ liệu quan hệ Viết phần chuẩn hóa (1NF–BCNF) Tiếp tục xử lý dữ liệu, sinh file clean Thiết kế sơ đồ kiến trúc hệ thống
43	Cả nhóm	Thi giữa kỳ, rà soát tiến độ, chỉnh sửa báo cáo
44	Nguyễn Minh Nhựt Phạm Đình Phương Nam Đoàn Mạnh Tất Phạm Đức Hoài Nam	Hỗ trợ tạo bảng trong DB Thêm ràng buộc, trigger, index Import dữ liệu vào DB Kiểm tra pipeline DB–ứng dụng
45	Nguyễn Minh Nhựt Phạm Đình Phương Nam	Viết kịch bản kiểm thử dữ liệu Kiểm thử CSDL (truy vấn, trigger)

	Đoàn Mạnh Tất Phạm Đức Hoài Nam	Kiểm thử pipeline xử lý dữ liệu Tổng hợp so sánh kết quả với yêu cầu ban đầu
46	Cả nhóm	Hoàn thiện phần kiểm thử & đánh giá, chuẩn bị L.0.2
47	Nguyễn Minh Nhựt Phạm Đình Phương Nam Đoàn Mạnh Tất Phạm Đức Hoài Nam Cả nhóm	Chuẩn bị slide kiến trúc & ERD Chuẩn bị slide CSDL Chuẩn bị slide xử lý dữ liệu Chuẩn bị slide triển khai & demo ứng dụng Trình bày L.0.2
48	Nguyễn Minh Nhựt Phạm Đình Phương Nam Đoàn Mạnh Tất Phạm Đức Hoài Nam	Viết phần ứng dụng tra cứu dữ liệu Thiết kế giao diện web/dashboard Vẽ biểu đồ trực quan hóa Tích hợp kết quả vào ứng dụng demo
49	Nguyễn Minh Nhựt Phạm Đình Phương Nam Đoàn Mạnh Tất Phạm Đức Hoài Nam	Viết kết quả đạt được Viết hạn chế còn tồn tại Viết hướng phát triển (data/ML) Hoàn thiện báo cáo tiền đề
50	Cả nhóm	Hoàn thành báo cáo, hoàn thiện slide cuối kỳ
51	Cả nhóm	Báo cáo cuối kỳ, nộp sản phẩm

4.2 Source code (link GitHub)

Toàn bộ mã nguồn của dự án được lưu trữ công khai tại GitHub: https://github.com/nam2312186/Project1_Group2

4.3 Tài liệu tham khảo

Tài liệu

- [1] Kaggle. *Datasets for Data Science and Machine Learning.* <https://www.kaggle.com/>
- [2] Spotify. *Spotify Web API Documentation.* <https://developer.spotify.com/documentation/web-api>
- [3] Billboard. *Billboard Hot 100 – Music Charts.* <https://www.billboard.com/>
- [4] ReccoBeats. *Track Audio Features API.* <https://reccobeats.com/docs/apis/get-track-audio-features>