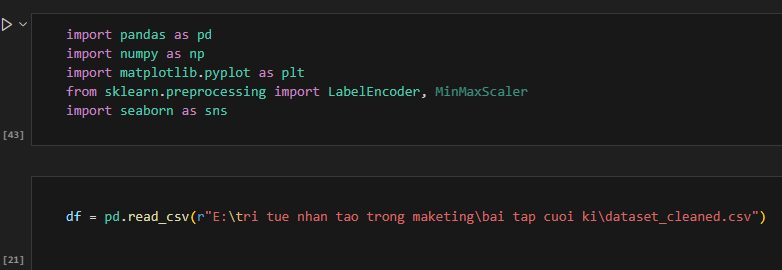
|  |
| --- |
| **BỘ GIÁO DỤC VÀ ĐÀO TẠO**  **TRƯỜNG ĐẠI HỌC MỎ ĐỊA CHẤT**  A blue circle with white text and a globe  AI-generated content may be incorrect.  **TRÍ TUỆ NHÂN TẠO TRONG MARKETING**  **BÀI TẬP TUẦN 2**  **Giảng viên hướng dẫn: Phạm Quang Hiển**  **Thực hiện**: Nguyễn Quang Nam\_2221050212  Vũ Tuấn Hoàng\_2221050766  Nguyễn Hữu Anh Minh\_2221050073  Ngô Bá Quân – 2221050396  Trần Thành Tâm - 2221050800  **HÀ NỘI – 2025** |

# Đọc và làm sạch dữ liệu

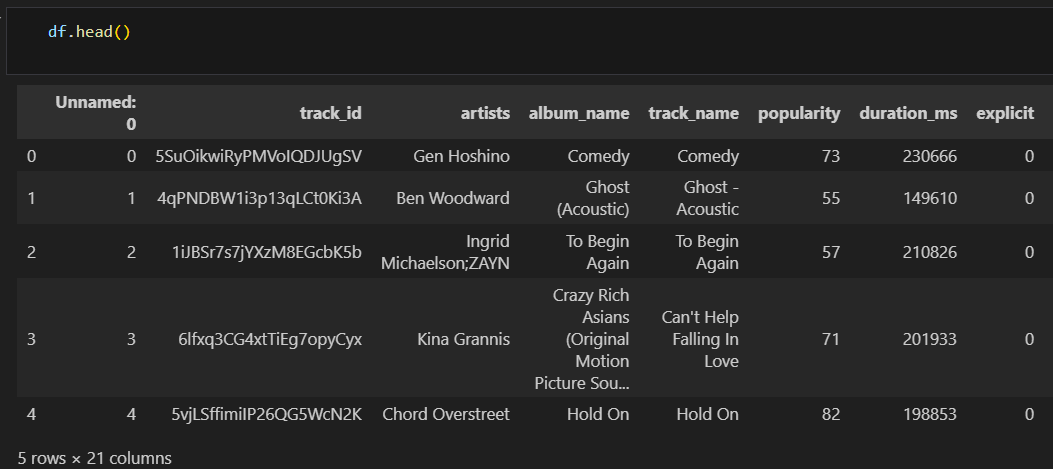
## Import các thư viện cần thiết và đọc dữ liệu



*Hình 1.1.Import thư viện và đọc dữ liệu*

* Khởi tạo các thư viện cần thiết để chuẩn bị cho khám phá, tiền xử lý dữ liệu và trực quan hóa
* Đọc dữ liệu từ file CSV vào DataFarme để kiểm tra xem có đọc được file hay có đúng đường dẫn hay chưa

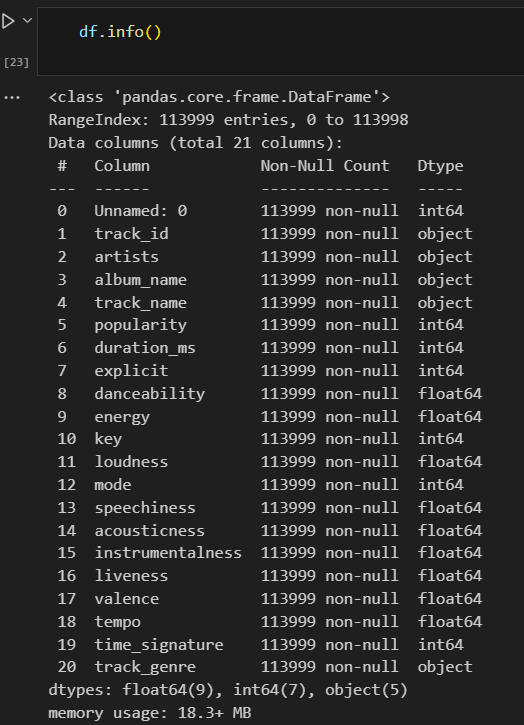
## Hiển thị 5 dòng đầu tiên



*Hình 1.2. Hiển thị 5 dòng đầu tiên của dữ liệu*

* Hiển thị 5 dòng đầu giúp hiển thị sơ bộ cấu trúc của dữ liệu

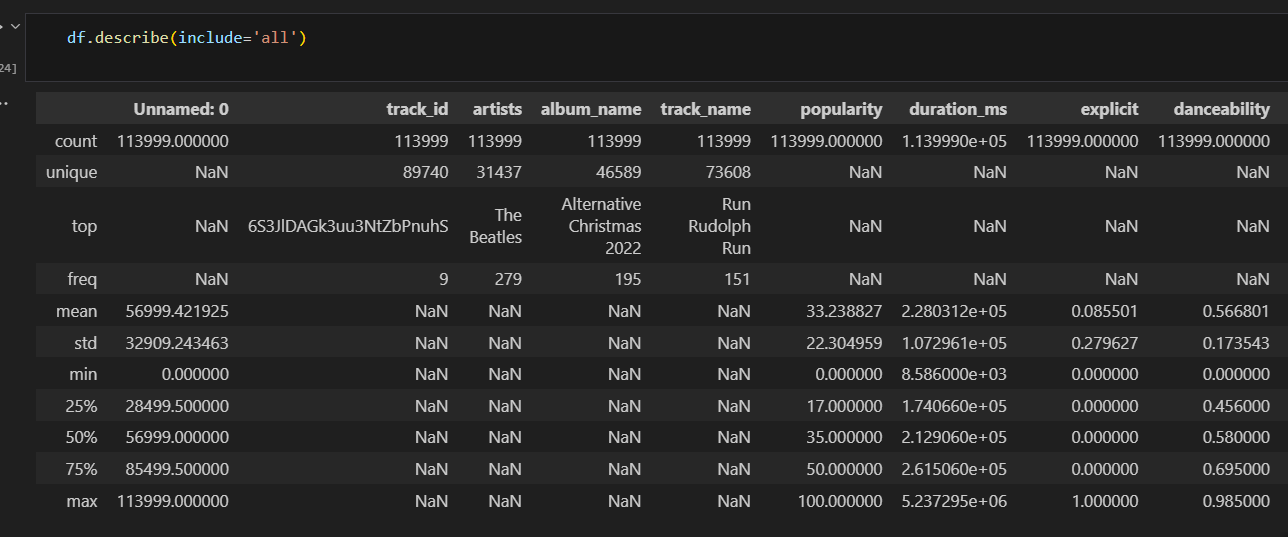
## Kiểm tra thông tin của bộ dữ liệu



*Hình 1.3. Hiển thị thông tin của bộ dữ liệu*

* Kiểm tra giá trị từng cột, xem có giá trị nào bị thiếu không
* Dữ liệu bao gồm:
* 113998 dòng
* 21 cột
* Các kiểu dữ liệu int64, object, float64
* Không có giá trị bị thiếu

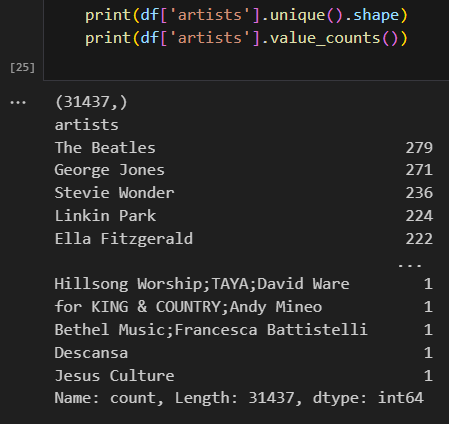
## Thống kê mô tả cho tất cả các cột



*Hình 1.4. Bảng mô tả*

* Để nắm phân bố, số lượng giá trị unique, và giá trị hay xuất hiện nhất.

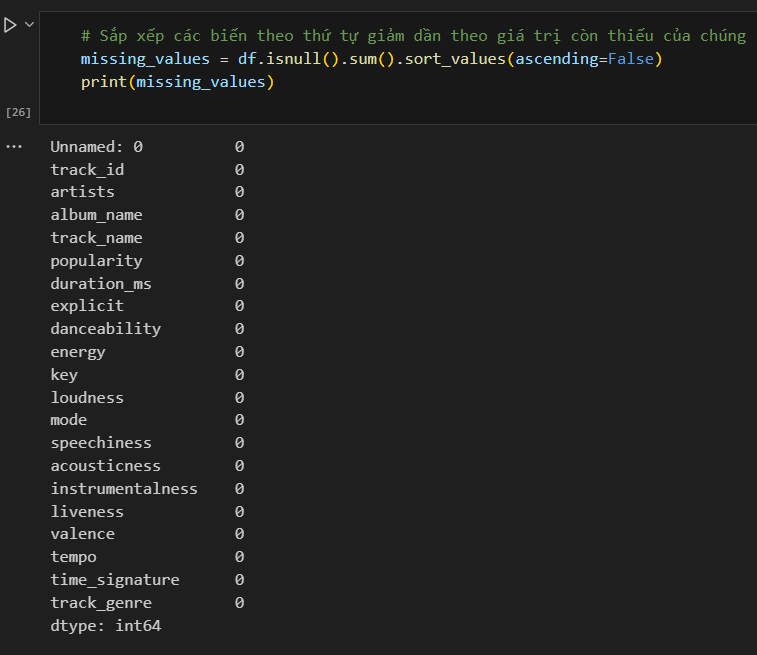
## Đếm số lượng nghệ sĩ trong tập dữ liệu



*Hình 1.5. Thống kê số lượng nghệ sĩ xuất hiện trong dataset*

* **Có 31,437 nghệ sĩ khác nhau** trong dataset
* Dataset chứa cực nhiều nghệ sĩ — phù hợp cho phân tích quy mô lớn.

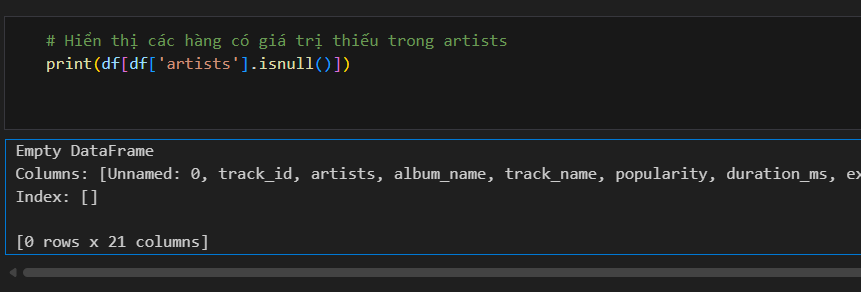
## Kiểm tra dữ liệu có missing không



*Hình 1.6. Thống kê dữ liệu missing*

* Các cột liệt kê đều có **0 missing**
* **Dataset không có giá trị thiếu**
* Dữ liệu ở trạng thái cleaned

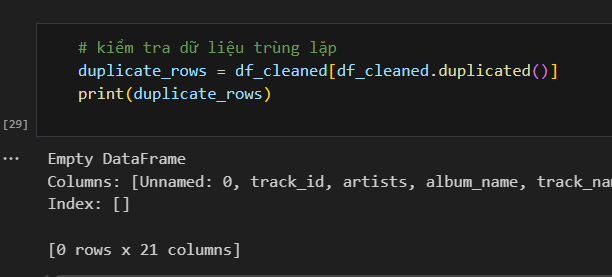
## Kiểm tra các hàng bị thiếu trong artists



*Hình 1.7. Hiển thị kết quả bị thiếu trong artists*

* Không có giá trị bị thiếu

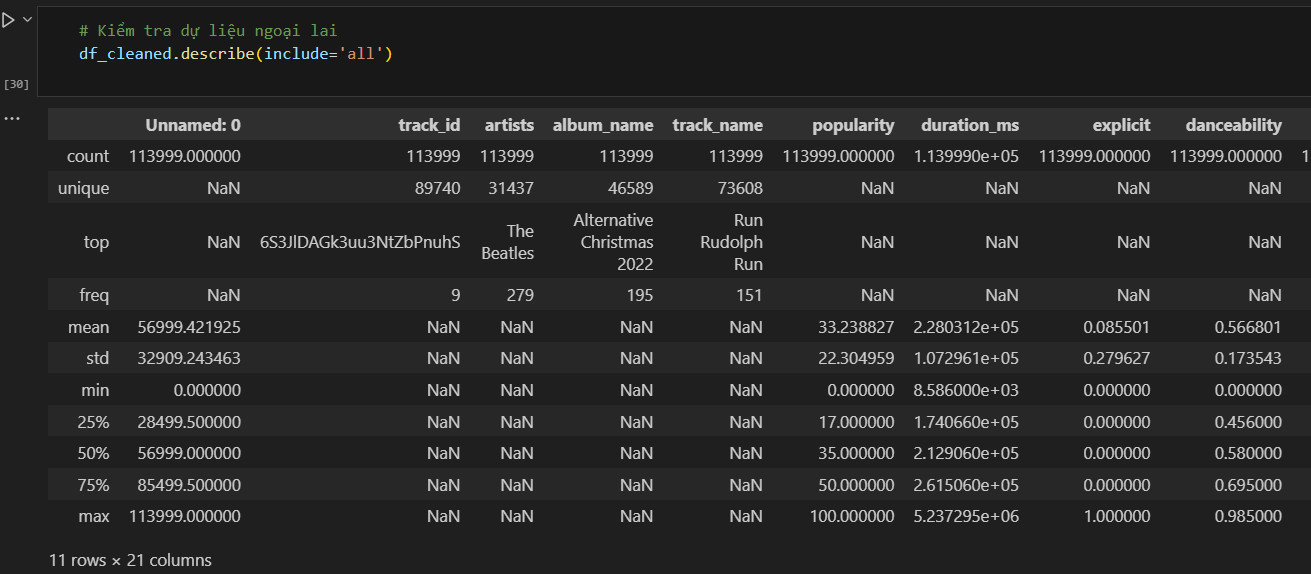
## Kiểm tra dữ liệu bị trùng lặp



*Hình 1.8. Hiển thị kết quả bị trùng lặp*

* Không có kết quả bị trùng lặp

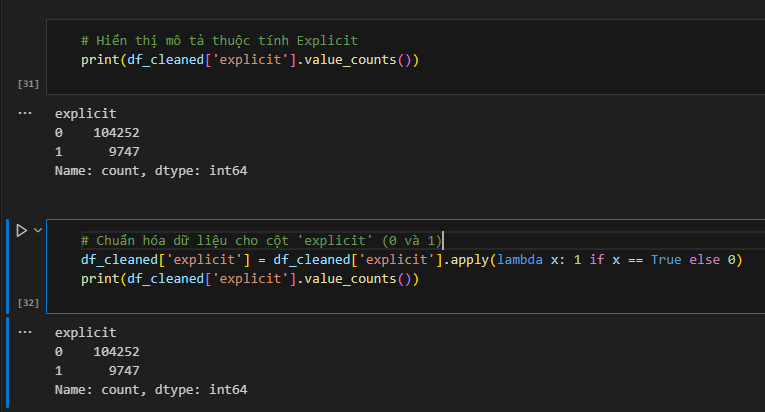
## Kiểm tra dữ liệu ngoại lai



*Hình 1.9. Hiển thị các giá trị ngoại lai*

* Không có giá trị ngoại lai

## Hiển thị mô tả và chuẩn hóa thuộc tính Explicit

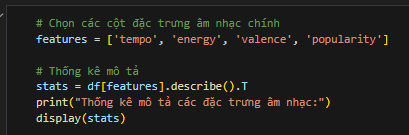


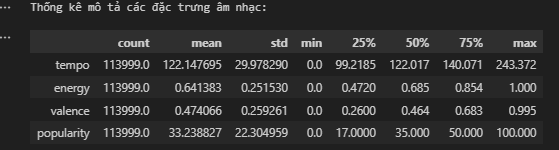
*Hình 1.10. Mô tả thuộc tính explicit*

* Đếm số lần xuất hiện của các giá trị khác nhau trong cột ecplicit
* Chuyển mọi giá trị True thành 1
* Các giá trị khác thành 0

# Phân tích, phân bố và thống kê mô tả

## Chọn các cột đặc trưng âm nhạc chính và thống kê mô tả cho các feature chính

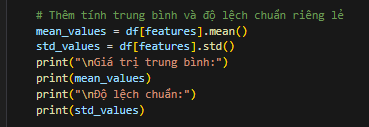


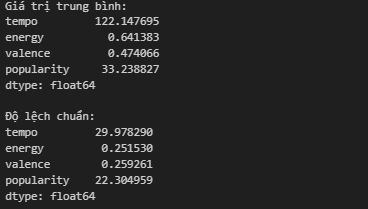


*Hình 2.1. Thống kê đặc trưng âm nhạc*

* Bảng thống kê các mô tả cho các feature chính:
  + - * Tempo:
        + **Mean ~122 BPM** — nhịp trung bình khoảng 122, phù hợp nhạc pop/dance.
        + **Std ~ 29.98** — độ phân tán lớn; có bài rất nhanh (~243 BPM) và có giá trị bằng 0 (min=0) — điểm cảnh báo (tempo = 0 rất có thể là lỗi đo/gán thiếu hoặc track instrumental không đo được).
        + **Percentiles:** 25% ~ 99, 50% ~ 122, 75% ~ 140 — phần lớn tempo nằm 99–140 BPM.
      * Energy:
        + **Mean ~ 0.64** (thang 0–1) — khá “năng lượng” (nhạc sôi động).
        + **Std ~ 0.25** — phân bố tương đối rộng. Min=0, Max=1, percentiles cho thấy nhiều bài ở khoảng trung bình-cao.
      * Valence:
        + **Mean ~ 0.47** — hơi thiên về neutral/slightly positive; không quá “vui” trung bình.
        + **Std ~ 0.26** nên phân bố trải đều.
      * Popularity:
        + **Mean ~33.24, Std ~22.30** — độ phổ biến trung bình thấp–trung bình; phổ biến có thể từ 0 đến 100. Max 100: một vài track rất nổi. Min 0: nhiều track không nổi bật.
        + Percentiles: 25% = 17, 50% = 33, 75% = 50 — nửa trên gồm track khá được nghe.

## Tính trung bình và độ lệch chuẩn

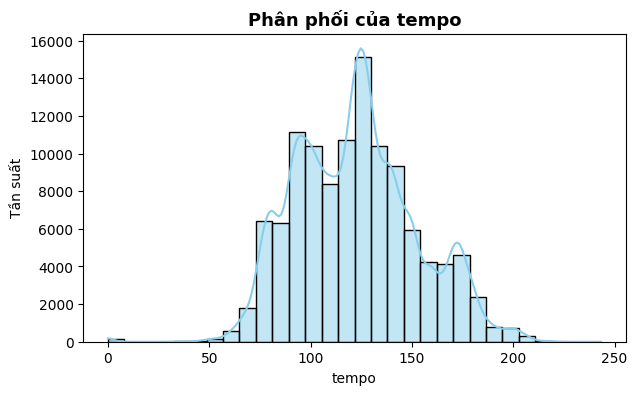




*Hình 2.2. Tính trung bình và độ lệch chuẩn*

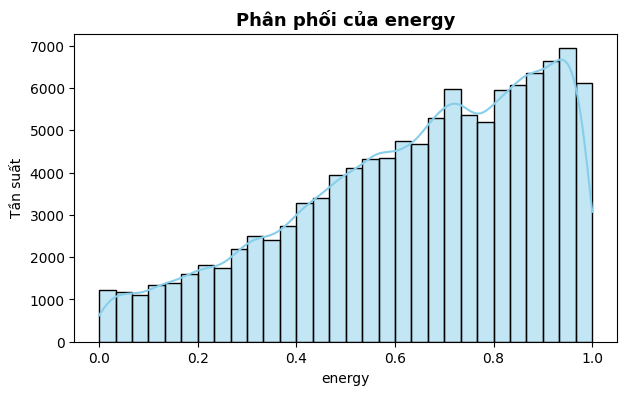
* Tempo có độ lệch chuẩn cao nên nhiều bài hát có tốc độ rất khác nhau.
* Energy và valence lệch chuẩn thấp suy ra các bài hát có năng lượng và cảm xúc khá đồng đều
* Popularity khá phân tán cho nên có bài rất nổi, có bài ít người nghe

## Phân tích chất lượng dữ liệu và vấn đề cần quan tâm



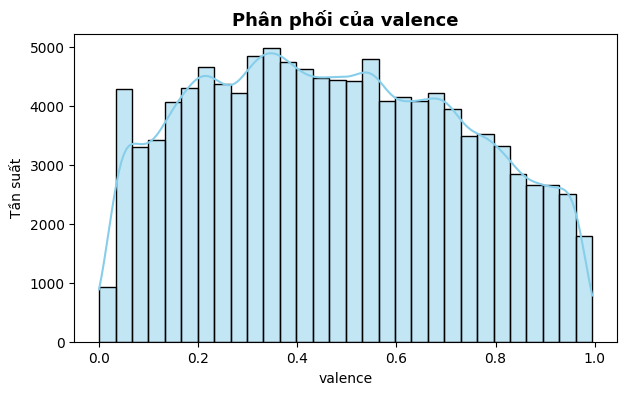
*Hình 2.3. Bảng phân phối của tempo*

* Sau khi phân phối tempo ta nhận thấy:
  + - * Có dạng **xấp xỉ phân phối chuẩn.**
      * Tâm phân phối quanh **120 bpm.**
      * Một số bài hát có tempo cao (tận 200 bpm), nhưng phần lớn nằm trong khoảng 90 đến 150 bpm.

****

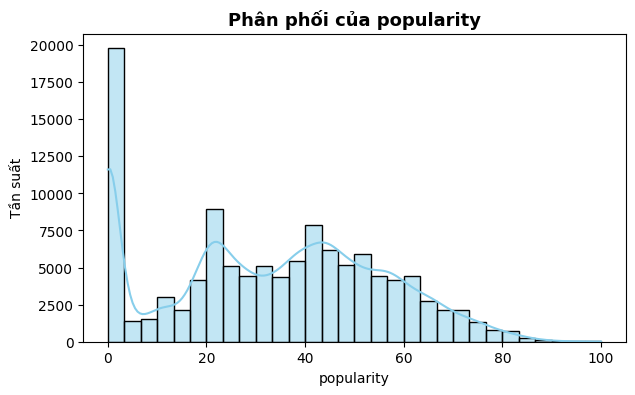
*Hình 2.4. Bảng phân phối energy*

* Bảng phân phối energy ta có thể nhận ra:
  + - * Phân phối **lệch phải** nên càng về 1thì càng nhiều bài hát.
      * Nghĩa là đa số bài trong dataset có năng lượng cao (mạnh, sôi động).

****

*Hình 2.5. Bảng phân phối valence*

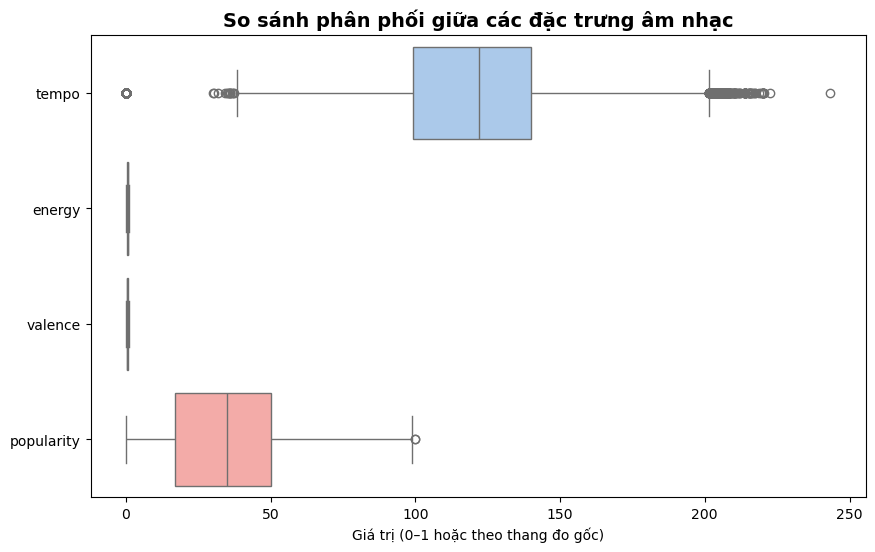
* Phân phối của valence nhận ra rằng
  + - * Phân phối **tương đối đều** (không lệch hẳn về bên nào).
      * Cho thấy tập dữ liệu có cả bài vui và bài buồn tương đối cân bằng.



*Hình 2.6. Bảng phân phối popularity*

* Phân phối của popularity cho thấy:
  + - * Phân phối **rất lệch phải**, tập trung nhiều ở giá trị thấp.
      * Nhiều bài có độ phổ biến thấp (0 - 20), ít bài đạt cao (trên 70).
* Nhận xét chung:
* Tempo: phần lớn bài hát có tempo trung bình khoảng 110 - 130 BPM nên phổ biến ở nhạc Pop/Dance.
* Energy: phân phối lệch phải suy ra đa số bài hát có năng lượng trung bình.
* Valence: phân phối rộng như vậy nhiều bài hát có cảm xúc trung lập hoặc buồn nhẹ.
* Popularity: phân phối lệch trái thì cho thấy chỉ một số ít bài hát thực sự phổ biến.

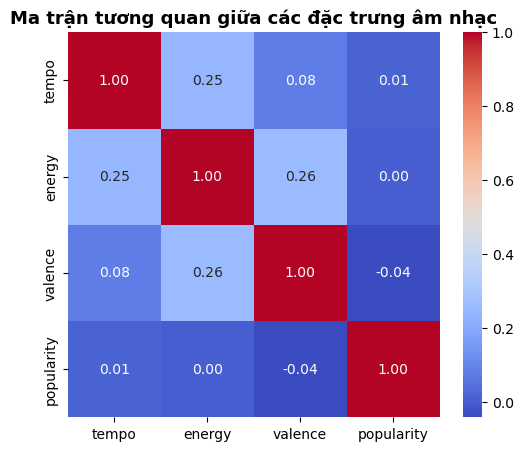
## **Bước trực quan hóa so sánh phân phối** bằng **biểu đồ hộp**



*Hình 2.7. Hình biểu đồ Boxplot*

* Biểu đồ Boxplot cho thấy:
  + - * Tempo:
        + Hộp khá rộng nên tempo có phân phối trải dài.
        + Trung vị khoảng **120 bpm**.
        + Có **nhiều điểm ngoại lệ (outliers)** ở hai đầu, biểu thị các bài có tempo cực chậm hoặc cực nhanh (bất thường so với phần lớn bài hát).
      * Energy:
        + Phân bố **tập trung ở khoảng cao (0.6–0.9)**.
        + Không có nhiều ngoại lệ → dữ liệu ổn định.
      * Valence:
        + Phân phối **trải đều từ 0 đến 1**, hộp rộng gần toàn thang đo.
        + Cho thấy sự đa dạng cảm xúc trong âm nhạc: có bài buồn, có bài vui.
      * Popularity:
        + Phân bố **rất lệch**, nhiều điểm ngoại lệ phía phải (các chấm rời rạc).
        + Hộp chính nằm ở khoảng 0–50, nghĩa là phần lớn bài không quá nổi tiếng.
        + Các **outlier phía xa** đại diện cho một số bài **rất nổi bật** (rất nhiều lượt nghe).
* Kết luận chung:
* Tempo: Nhịp độ bài hát đa dạng, nhưng tập trung quanh mức 120 bpm.
* Energy: Phần lớn bài hát có năng lượng mạnh, rất ít bài “êm dịu”.
* Valence: Dataset cân bằng giữa nhạc vui – nhạc buồn.
* Popularity: Phần lớn bài hát bình thường, chỉ vài bài cực kỳ phổ biến

## Tương quan giữa các đặc trưng của âm nhạc

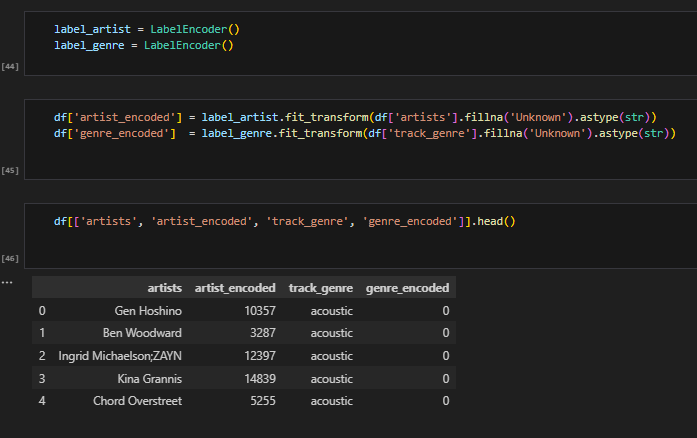


*HÌnh 2.8. Ma trận tương quan*

* Các hệ số tương quan đều nhỏ hơn 0.3, nghĩa là các đặc trưng âm nhạc hầu như độc lập với nhau.
* Không có biến nào có ảnh hưởng mạnh đến “popularity” (độ phổ biến).
* Mối quan hệ rõ nhất (dù vẫn yếu) là: energy với valence (0.26) nên nhạc mạnh thường mang cảm xúc tích cực hơn.

# Mã hóa – Chuẩn hóa dữ liệu và Phân tích mối tương quan

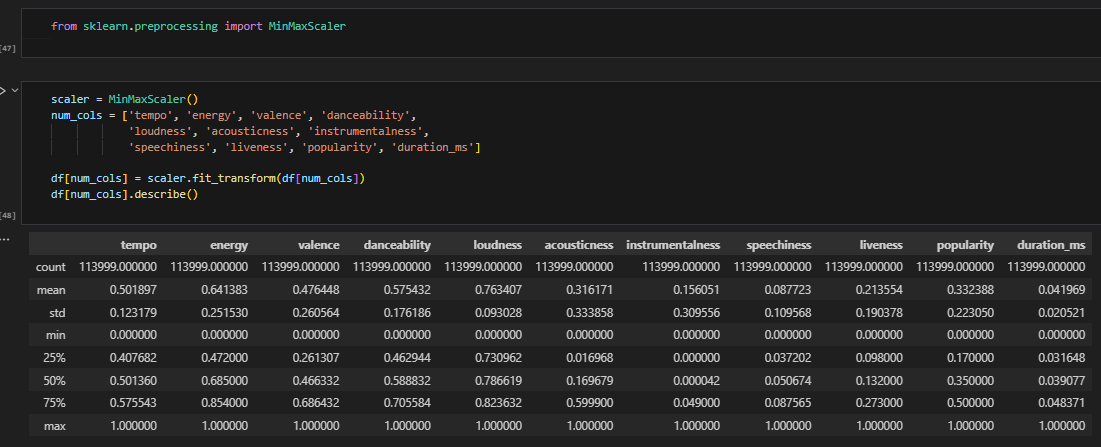
## Mã hóa dữ liệu dạng chuỗi



*Hình 3.1. Mã hóa dữ liệu*

* Nhận xét:
  + - * Cột artist\_name và genre đã được mã hóa thành dạng số (artist\_encoded, genre\_encoded) giúp mô hình dễ dàng xử lý hơn.
      * Ví dụ: “Ed Sheeran” → 105, “Pop” → 2.
      * Đây là bước quan trọng trong pipeline vì mô hình máy học không thể xử lý trực tiếp dữ liệu dạng text.

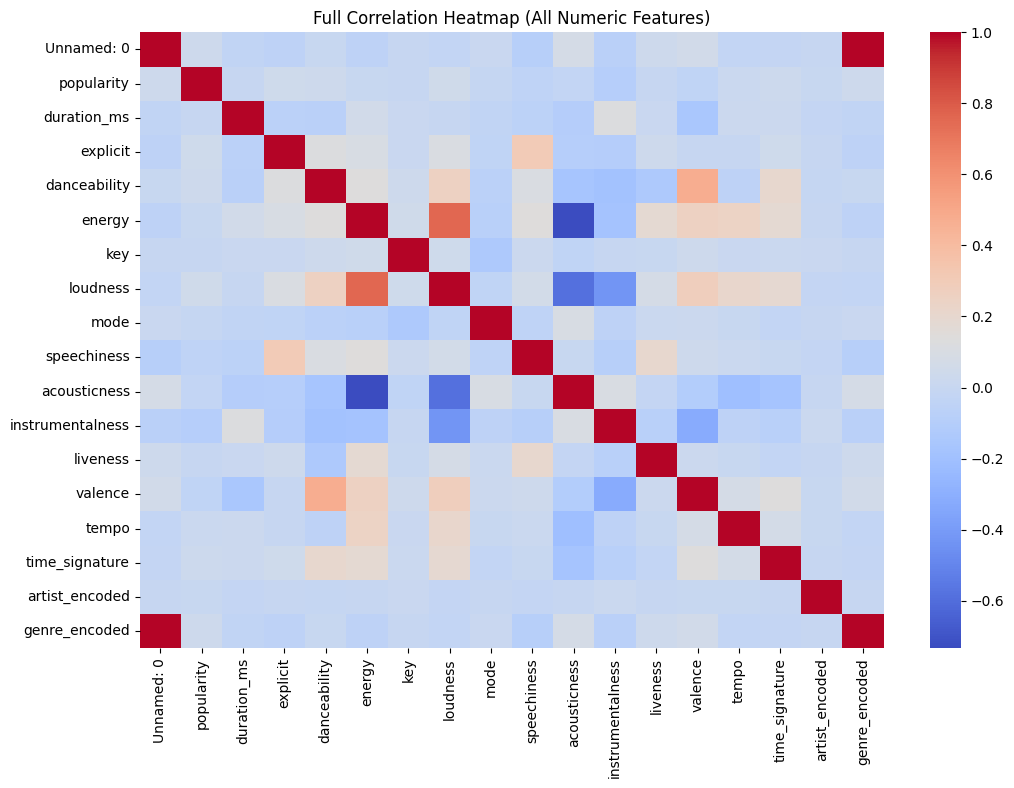
## Chuẩn hóa dữ liệu số



*Hình 3.2. Dữ liệu số được chuẩn hóa*

* Tất cả các đặc trưng âm nhạc số đã được chuẩn hóa về khoảng [0, 1], đảm bảo các đặc trưng có trọng số tương đương khi tính toán độ tương đồng.
* Điều này đặc biệt quan trọng trong Content-based filtering vì mô hình dựa vào khoảng cách giữa các điểm dữ liệu.
* Không chuẩn hóa có thể khiến các cột như duration\_ms (có giá trị lớn) chi phối kết quả tính toán.

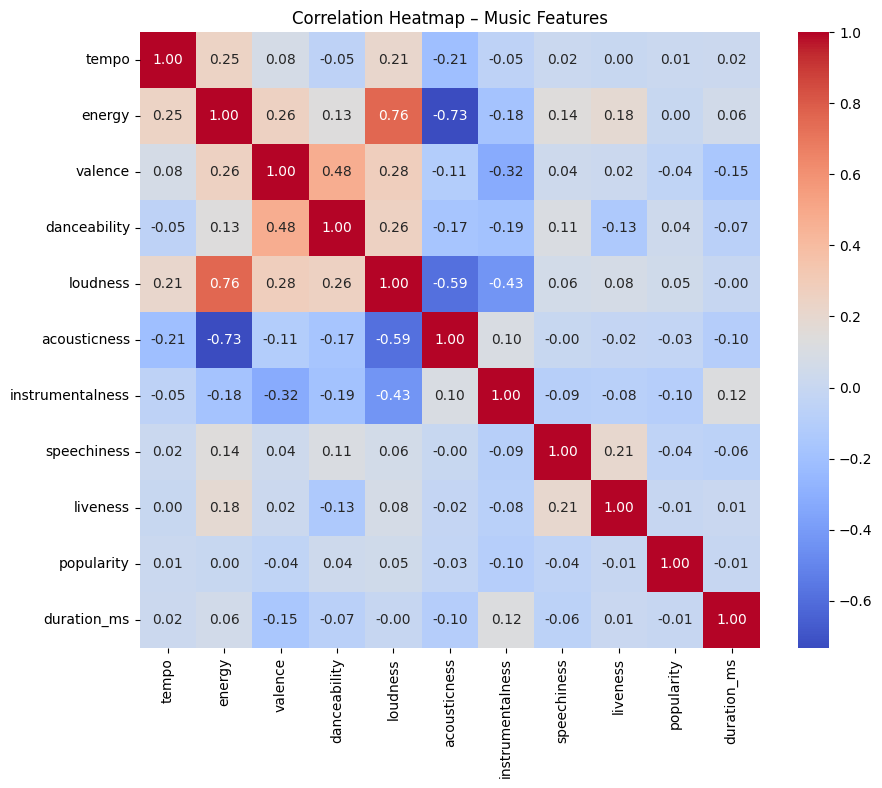
## Ma trận tương quan dữ liệu số



*Hình 3.3. Ma trận tương quan dữ liệu số*

* Ma trận tương quan tổng thể cho thấy đa số các đặc trưng có tương quan yếu đến trung bình, thể hiện bằng màu nhạt gần trung tính. Điều này cho thấy nhiều đặc trưng tương đối độc lập và có thể mang thông tin riêng biệt khi đưa vào mô hình.
* Một số cặp đặc trưng nổi bật có tương quan mạnh:
* energy ↔ loudness ≈ 0.76 (thuận mạnh): bài hát năng lượng cao thường có âm lượng lớn hơn.
* energy ↔ acousticness ≈ -0.73 (nghịch mạnh): nhạc acoustic thường có năng lượng thấp, trong khi nhạc điện tử hoặc rock có năng lượng cao hơn.
* loudness ↔ acousticness ≈ -0.59 (nghịch trung bình mạnh): nhạc acoustic thường có âm lượng nhỏ hơn.
* Các cột như artist\_encoded hay genre\_encoded chỉ là nhãn mã hóa nên hệ số tương quan của chúng không có nhiều ý nghĩa âm học.
* Biến popularity có tương quan thấp với các đặc trưng âm học, gợi ý rằng độ phổ biến của bài hát không chỉ phụ thuộc vào đặc điểm âm nhạc mà còn bị chi phối bởi yếu tố bên ngoài như thời gian phát hành, nghệ sĩ hay xu hướng thị trường.

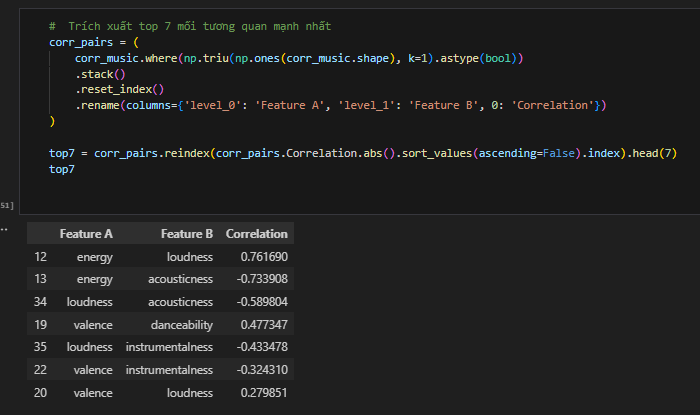
## Ma trận tương quan nhóm đặc trưng âm nhạc chính



*Hình 3.4. Ma trận tương quan nhóm đặc trưng âm nhạc chính*

* Khi chỉ tập trung vào các đặc trưng âm nhạc cốt lõi (tempo, energy, valence, danceability, loudness, acousticness, instrumentalness, speechiness, liveness, popularity, duration\_ms), các mối quan hệ thể hiện rõ ràng và có ý nghĩa hơn:
  + - * energy ↔ loudness ≈ 0.76 (thuận rất mạnh): xác nhận rằng các bài hát sôi động có âm lượng cao hơn – điều quan trọng khi mô hình hóa phong cách nhạc
      * energy ↔ acousticness ≈ -0.73 (nghịch mạnh): đặc trưng quan trọng để phân tách nhạc acoustic với nhạc điện tử/rock.
      * valence ↔ danceability ≈ 0.48 (thuận trung bình): bài hát có cảm xúc tích cực thường dễ nhảy – gợi ý tiềm năng để xây dựng hệ thống gợi ý dựa trên tâm trạng người dùng.
      * loudness ↔ instrumentalness ≈ -0.43 (nghịch trung bình): các bản nhạc không lời có xu hướng âm lượng thấp hơn.
      * valence ↔ instrumentalness ≈ -0.32 (nghịch): các bài instrumental thường có cảm xúc trung tính hoặc trầm hơn.
* Ngoài ra, biến `popularity` tiếp tục thể hiện tương quan yếu với các đặc trưng âm nhạc, cho thấy cần thêm yếu tố bổ trợ nếu muốn dự đoán độ phổ biến của bài hát.

## Trích xuất top 7 mối tương quan mạnh nhất

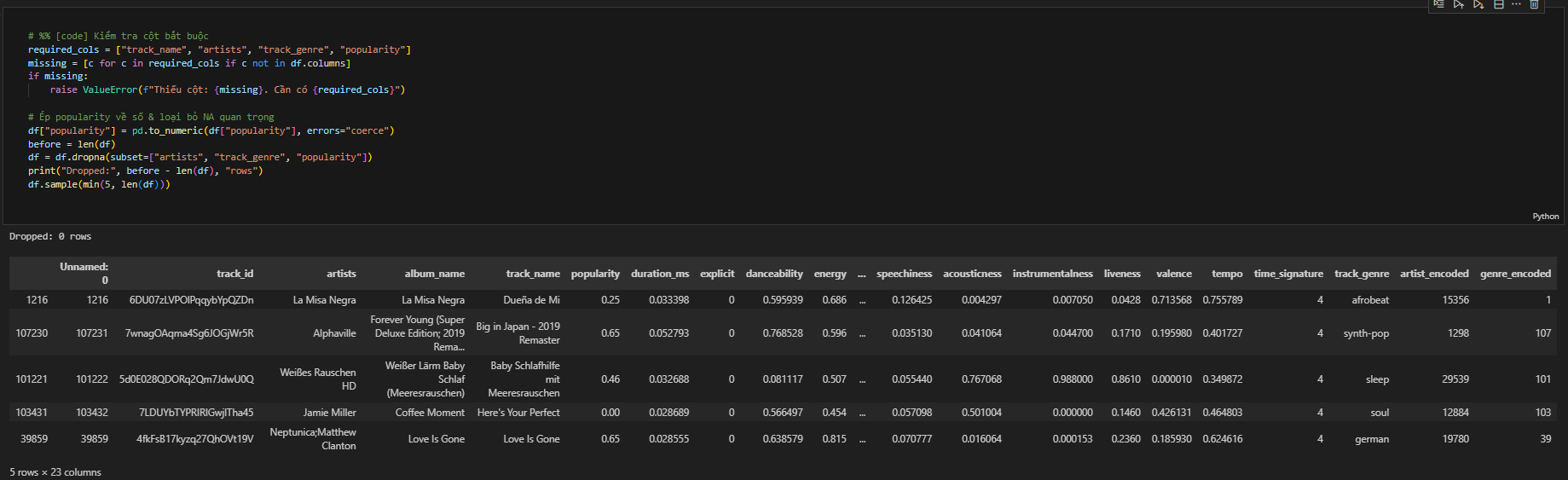


*Hình 3.5. Top 7 mối tương quan mạnh nhất*

* Nhóm đặc trưng energy – loudness – acousticness thể hiện mối liên hệ rất mạnh, phản ánh rõ rệt sự khác biệt giữa các thể loại nhạc sôi động và nhạc mộc.
* Cặp valence – danceability là đặc trưng quan trọng để gợi ý nhạc dựa trên tâm trạng và trải nghiệm người dùng.
* Các mối tương quan âm cho thấy các đặc điểm như “mức độ acoustic” hay “instrumentalness” thường đối lập với “năng lượng” và “âm lượng” – điều này có thể khai thác để mô hình phân tách các dòng nhạc khác nhau.
* Một số đặc trưng có tương quan cao (> 0.7) như energy và loudness có thể gây đa cộng tuyến, nên cân nhắc chọn lọc khi xây dựng mô hình.

# Phân tích theo thể loại và ca sĩ

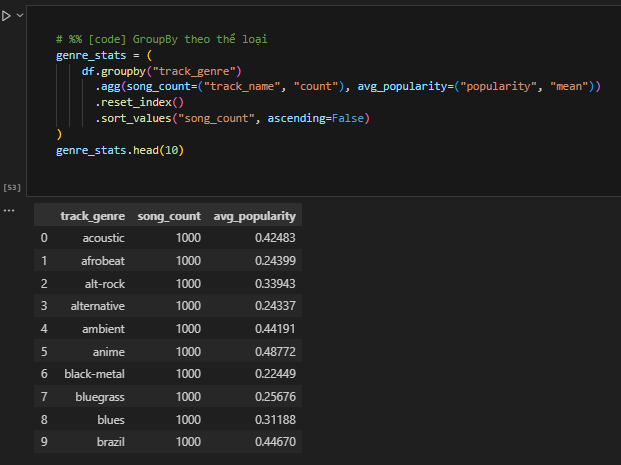
## Kiểm tra cột và tiền xử lý nhẹ



*Hình 4.1. Kiểm tra cột và tiền xử lý nhẹ*

* Dữ liệu có cấu trúc rõ ràng, mỗi bài hát kèm đầy đủ thông tin:
* Nghệ sĩ (artists)
* Thể loại (track\_genre)
* Các đặc trưng âm nhạc (energy, acousticness, valence, ...)
* Độ phổ biến (popularity) – biến mục tiêu để dự đoán hoặc phân cụm.
* Popularity là số thực nằm trong khoảng [0, 1]
  + - * Có thể đã được chuẩn hóa (normalized) từ thang điểm gốc của Spotify (0–100).
* Không có missing values ở các cột chính, nên bạn có thể yên tâm tiếp tục:
* Phân tích thống kê (describe(), groupby() theo thể loại)
* Trực quan hóa (boxplot, heatmap, barplot theo track\_genre)
* Xây dựng mô hình (K-means, Linear Regression, hoặc Recommendation System)

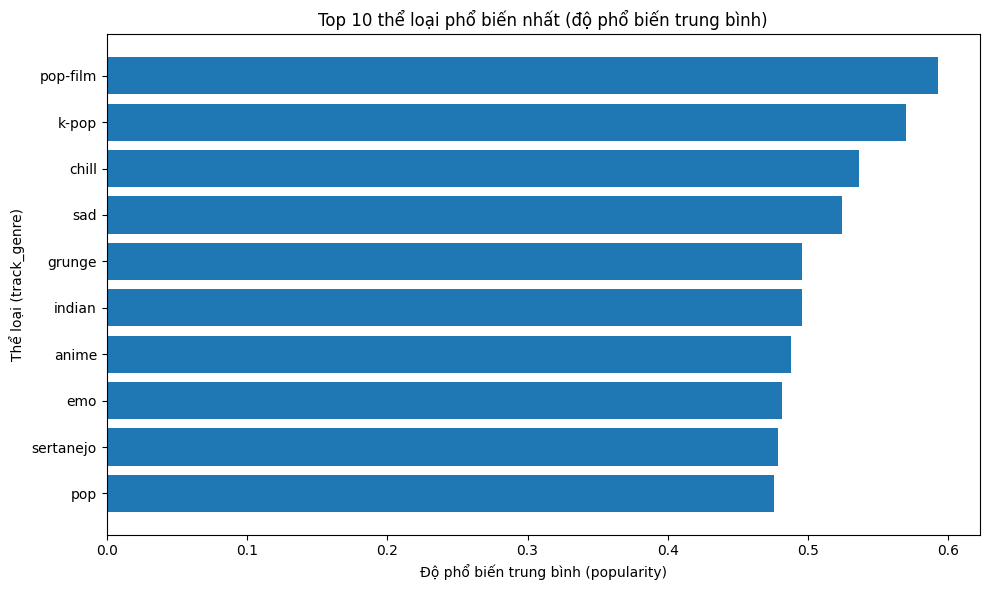
## Thống kê số lượng bài hát và độ phổ biến trung bình



*Hình 4.2. Bảng thống kê*

* Tổng hợp dữ liệu theo **thể loại nhạc**.
* Cho biết **số lượng bài hát và độ phổ biến trung bình** của từng thể loại.
* Giúp bạn dễ dàng nhận biết **thể loại nào phổ biến nhất hoặc được nghe nhiều nhất**.

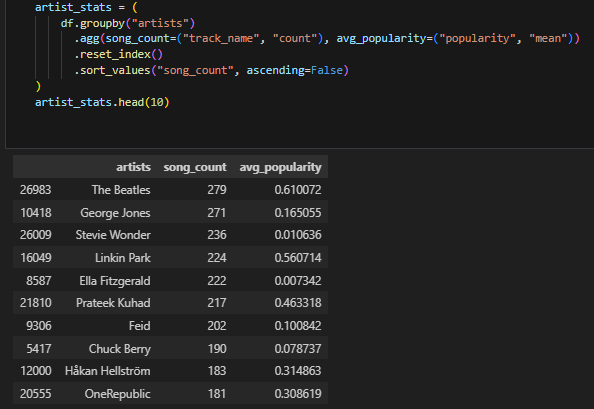
## Top 10 thể loại có độ phổ biến trung bình cao nhất



*Hình 4.3. Bảng thống kê thể loại nhạc phổ biến nhất*

* Một số track\_genre có nhiều bài hát nhất phản ánh mức độ phủ rộng của thể loại đó trong dataset.
* Độ phổ biến trung bình (avg\_popularity) cho thấy mức độ yêu thích - có thể khác thứ hạng theo số lượng.
* Nếu cần gợi ý cho người dùng mới, có thể ưu tiên các thể loại có avg\_popularity cao.

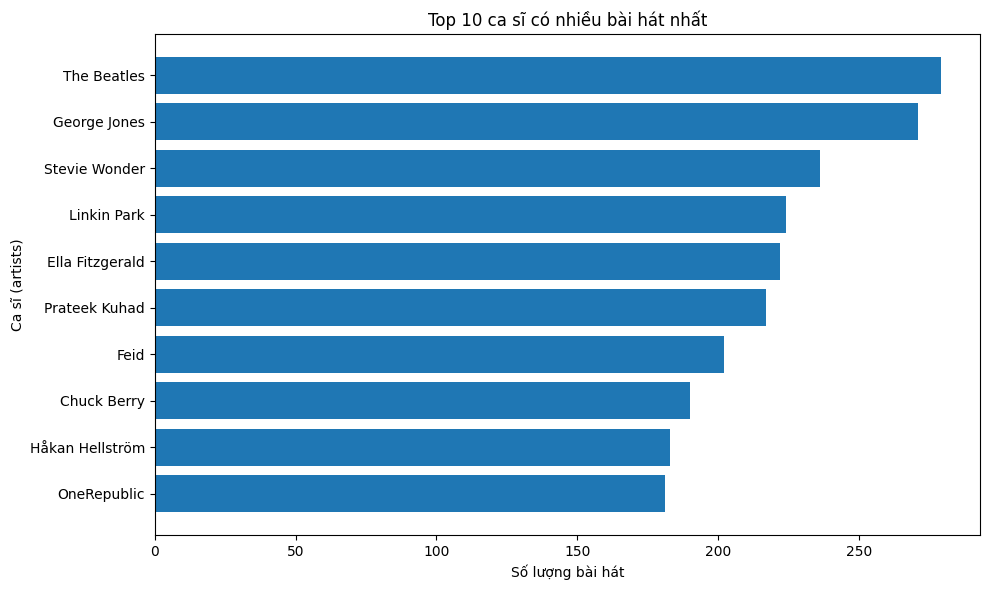
## Thống kê **theo từng nghệ sĩ**



*Hình 4.4. Bảng thống kê theo từng nghệ sĩ*

* **The Beatles** có số lượng bài hát nhiều nhất (279 bài), và cũng **có độ phổ biến trung bình cao nhất** (~ 0.61). Nghĩa là họ không chỉ sáng tác nhiều, mà các bài hát của họ cũng được yêu thích rộng rãi.
* **Linkin Park** đứng thứ 4 về số lượng bài hát (224), nhưng có **độ phổ biến cao thứ hai** (~ 0.56). Cho thấy nhạc của họ có sức ảnh hưởng mạnh.
* **George Jones** và **Stevie Wonder** có khá nhiều bài, nhưng độ phổ biến trung bình lại thấp (0.16 và 0.01). Có thể vì các bài hát của họ trong tập dữ liệu thuộc dòng nhạc cũ, ít người nghe hiện nay.

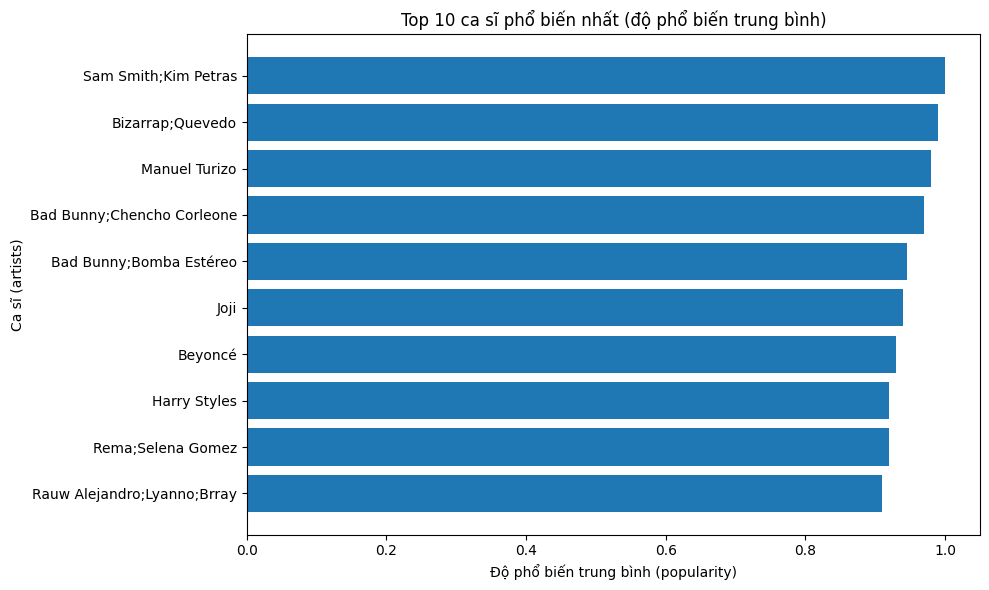
## Top 10 ca sĩ có nhiều bài nhất



*Hình 4.5. Bảng thống kê ca sĩ*

* **The Beatles và George Jones** là 2 người dẫn đầu về số lượng.

## Top 10 ca sĩ có độ phổ biến trung bình cao nhất



*Hình 4.6. Bảng thống kê ca sĩ có độ phổ biến*

* Nghệ sĩ có nhiều bài hát thường là những người phát hành nhiều hoặc xuất hiện trong nhiều album/playlist.
* Nghệ sĩ có avg\_popularity cao cho thấy mức độ yêu thích mạnh dù số bài có thể không nhiều.
* Hai chiều số lượng vs yêu thích hữu ích để xây dựng gợi ý theo nghệ sĩ.