TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

## VIỆN TRÍ TUỆ NHÂN TẠO

**-----\*\*\*-----**

## BÁO CÁO MÔN HỌC KỸ THUẬT VÀ CÔNG NGHỆ DỮ LIỆU LỚN

**ĐỀ TÀI**

**Đề Xuất Hành Động Từ Phân Tích Phân Cụm**

### Nhóm sinh viên thực hiện:1. Nguyễn Bảo Sơn -22022539

### Chu Huỳnh Đức -22022612

## HÀ NỘI, 12/2024

## MỞ ĐẦU

Big Data, hay Dữ liệu lớn, đã và đang thay đổi cách thức chúng ta nhìn nhận và xử lý thông tin trong thế giới hiện đại. Đây không chỉ đơn thuần là việc lưu trữ và quản lý một lượng lớn dữ liệu, mà còn bao gồm khả năng phân tích và rút ra những hiểu biết có giá trị từ khối dữ liệu khổng lồ này. Với sự phát triển không ngừng của công nghệ, Big Data đang mở ra những cơ hội mới trong việc giải quyết các thách thức và tối ưu hóa quy trình trong nhiều lĩnh vực của đời sống. Sự phát triển của các công nghệ phân tích tiên tiến, kết hợp với khả năng xử lý dữ liệu ngày càng mạnh mẽ, đã giúp các tổ chức và doanh nghiệp có thể khai thác tối đa tiềm năng của Big Data, từ đó đưa ra những quyết định thông minh và chiến lược hiệu quả hơn.Từ đó, chúng em đã chọn đề tài: "**Đề Xuất Hành Động từ Phân Tích Phân Cụm**" để làm báo kết thúc môn học của mình.

Báo cáo này trình bày về quá trình phân tích một bộ dataset âm nhạc, được chia thành 5 phần chính:

* Phần 1 - Xử lý và chuẩn bị dữ liệu: Tập trung vào việc làm sạch và chuẩn hóa dữ liệu thô ban đầu.
* Phần 2 - Phân tích và trực quan hóa: Tiến hành phân tích chuyên sâu dữ liệu và biểu diễn kết quả bằng các biểu đồ trực quan.
* Phần 3 - Xử lý phân cụm: Áp dụng các kỹ thuật phân cụm để nhóm các bài hát có đặc điểm tương đồng.
* Phần 4 - Hệ thống gợi ý: Xây dựng hệ thống đề xuất bài hát dựa trên kết quả phân tích và phân cụm.
* Phần 5- Tổng kết : Kết quả sau khi thực hiện và xử lý phân cụm

**Mục Lục**

[MỞ ĐẦU 2](#_bookmark0)

Phần 1: Xử lý và chuẩn bị dữ liệu 4

* 1. Chọn Data 4
  2. Khởi tạo và import thư viện 4
  3. Xử lý dữ liệu ban đầu 5
  4. Chuyển đổi và chuẩn hóa dữ liệu 6
  5. Phân tích thống kê cơ bản 8

Phần 2: Phân tích và trực quan hóa 9

* 1. Phân tích theo popularity 9
  2. Tính toán sự tương quan các tính năng 10
  3. Phân tích về nghệ sĩ album và thời gian 11
  4. Một số biểu đồ phân tích khác 13

Phần 3: Xử lý phân cụm 14

3.1 Phân cụm dữ liệu 14

3.2 Phân tích từng cụm 16

Phần 4: Hệ thống gợi ý 20

Phần 5: Tổng kết kết quả 20

1. **XỬ LÝ VÀ CHUẨN BỊ DỮ LIỆU**
   1. **Chọn data**

Link data:‘https://www.kaggle.com/datasets/maharshipandya/-spotify-tracks-dataset’

Dataset này được lấy từ Spotify- một trong những nền tảng phát nhạc trực tuyến lớn nhất thế giới.Sau khi phân tích và xử lý dữ liệu,dataset thể hiện nhiều ưu điểm nổi bật:

· Là một data đủ lớn bao gồm 114000 bài hát, tạo độ tin cậy cao trong việc phân tích.

· Chất lượng dữ liệu đáng tin cậy với các thông số kỹ thuật chi tiết về âm nhạc như: danceability, energy, loudness - những chỉ số quan trọng trong việc đánh giá một bài hát.

· Nguồn dữ liệu phong phú và đa dạng, bao gồm thông tin từ nhiều nghệ sĩ, album và thể loại âm nhạc khác nhau, tạo cơ sở tốt cho việc phân tích chuyên sâu.

· Cấu trúc dữ liệu được tổ chức khoa học, dễ dàng cho việc xử lý và phân tích, đồng thời cho phép mở rộng nghiên cứu theo nhiều hướng khác nhau.

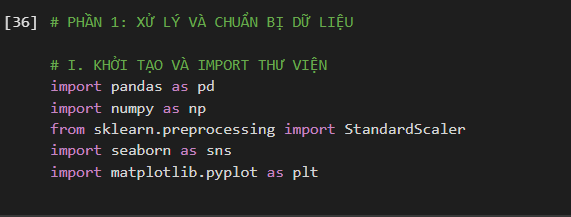
· Thông tin về độ phổ biến (popularity) của bài hát giúp hiểu rõ hơn về thị hiếu người nghe và xu hướng âm nhạc hiện đại.

· Khả năng ứng dụng cao trong việc xây dựng các mô hình đề xuất và phân tích xu hướng âm nhạc.

* 1. **Khởi tạo và import thư viện**

Ở bước đầu tiên, chúng ta tiến hành nhập các thư viện cần thiết cho quá trình phân tích:

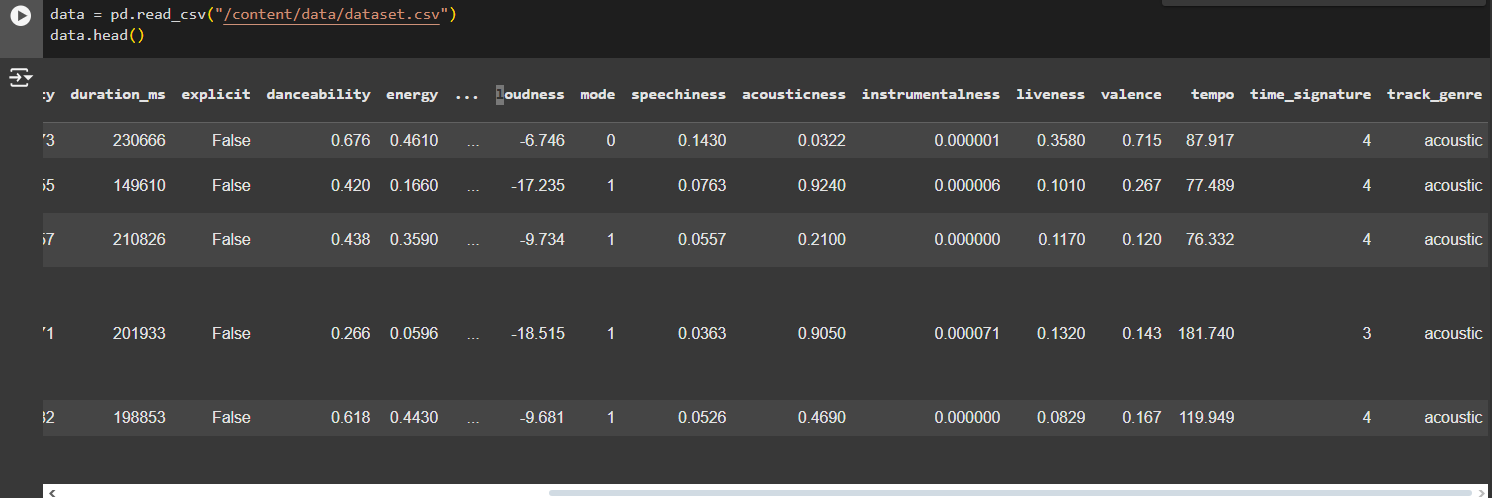
* Pandas và Numpy: phục vụ cho việc xử lý và phân tích dữ liệu
* Seaborn và Matplotlib.pyplot: dùng để tạo các biểu đồ trực quan
* StandardScaler từ thư viện Sklearn: được sử dụng để chuẩn hóa dữ liệu trong các bước tiếp theo



Việc sử dụng đồng thời các thư viện này giúp tạo nên một quy trình phân tích dữ liệu hoàn chỉnh, từ xử lý đến trực quan hóa kết quả.

* 1. **Xử lý dữ liệu ban đầu**
     1. Đọc dữ liệu

****

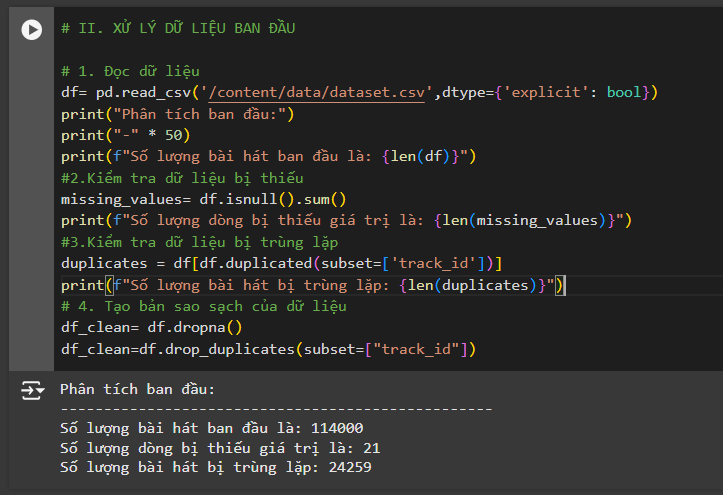
****

* + 1. Kiểm tra và loại bỏ dữ liệu còn thiếu và trùng lặp

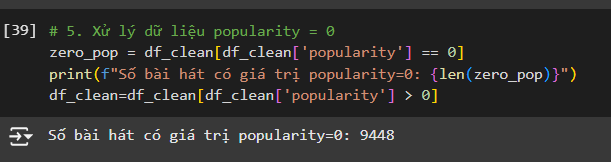
Việc tiền xử lý dữ liệu là bước đầu tiên và quan trọng nhất khi làm việc với một bộ dataset. Trong bước này, chúng ta cần tập trung vào hai nhiệm vụ chính:

* Loại bỏ các giá trị trùng lặp (duplicated values)
* Xử lý các dữ liệu thiếu hoặc null (missing values)

Quá trình làm sạch dữ liệu này giúp đảm bảo tính chính xác của kết quả phân tích sau này và tránh được các lỗi có thể phát sinh trong quá trình xử lý. Đồng thời, việc này cũng giúp tăng độ tin cậy của các kết luận và insight rút ra từ dữ liệu.

****

* + 1. Xử lý những bài hát có popularity =0

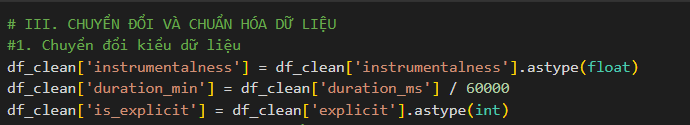


Qua quá trình phân tích các bài hát có chỉ số popularity bằng 0, chúng em nhận thấy đây không phải là những bài hát không được yêu thích, mà nguyên nhân chính đến từ việc Spotify chưa kịp cập nhật dữ liệu thống kê cho các bài hát này. Vì vậy, việc loại bỏ những bài hát có popularity = 0 khỏi tập dữ liệu là cần thiết để đảm bảo kết quả phân tích được chính xác và phản ánh đúng thực tế.

* 1. **Chuyển đổi và chuẩn hóa dữ liệu**
     1. Chuyển đổi kiểu dữ liệu

Trong quá trình chuẩn hóa dữ liệu, chúng em thực hiện một số điều chỉnh để tối ưu hóa việc phân tích:

* Chuyển đổi kiểu dữ liệu của trường instrumentalness từ dạng hiện tại sang dạng float để thuận tiện cho việc tính toán.
* Đối với thời lượng bài hát, thay vì sử dụng đơn vị millisecond (ms), chúng tôi chuyển đổi sang đơn vị phút để người đọc dễ dàng nắm bắt thông tin.
* Trường explicit được chuyển từ dạng True/False sang dạng số (0/1), giúp đơn giản hóa quá trình xử lý số liệu trong các bước phân tích tiếp theo.

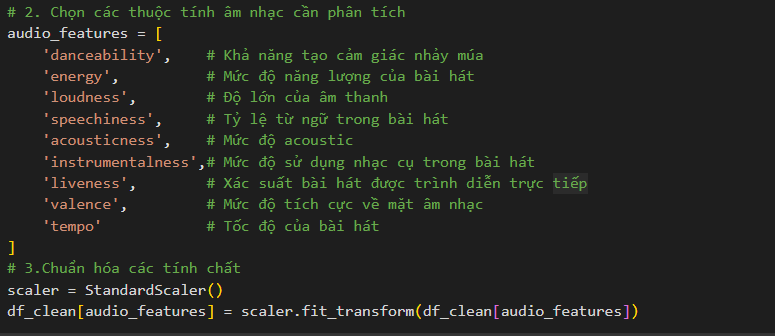


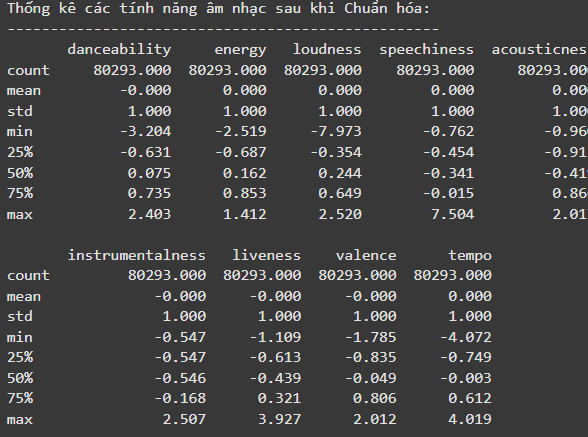
* + 1. Chọn những thuộc tính cần phân tích và chuẩn hóa nó

Sau khi phân tích tổng thể dữ liệu, chúng em đã xác định và lựa chọn các thuộc tính âm nhạc quan trọng nhất để tập trung nghiên cứu, bao gồm:

* Danceability (Khả năng tạo cảm giác nhảy múa)
* Energy (Mức độ năng lượng)
* Loudness (Độ lớn của âm thanh)
* Speechiness (Tỷ lệ lời nói)
* Acousticness (Mức độ acoustic)
* Instrumentalness (Mức độ nhạc cụ)
* Liveness (Độ tương tác trực tiếp)
* Valence (Mức độ tích cực)
* Tempo (Nhịp độ)

Để đảm bảo tính chính xác trong quá trình phân tích, các thuộc tính này được chuẩn hóa bằng phương pháp StandardScaler (Z-score normalization), giúp đưa dữ liệu về cùng một thang đo và loại bỏ sự chênh lệch về đơn vị.



****

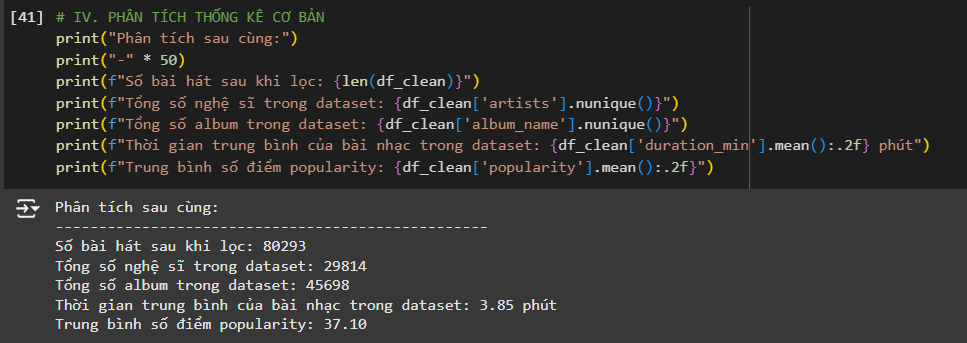
Sau chuẩn hóa, mỗi đặc tính âm nhạc đã được biến đổi để có trung bình = 0 (đặt tâm các đặc tính về cùng một điểm) và độ lệch chuẩn = 1 (đưa độ phân tán về cùng một thang đo).

* 1. **Phân tích thông kê cơ bản**

Sau quá trình làm sạch dữ liệu và loại bỏ các bài hát không hợp lệ, chúng em đúc kết được những thông tin tổng quan sau về dataset:

* Số lượng bài hát còn lại sau khi lọc
* Tổng số nghệ sĩ có trong dataset
* Tổng số album xuất hiện trong dataset
* Thời lượng trung bình của các bài hát
* Điểm popularity trung bình của các bài hát

Các thông số này giúp chúng em có cái nhìn tổng quát về bộ dữ liệu trước khi đi vào phân tích chi tiết.



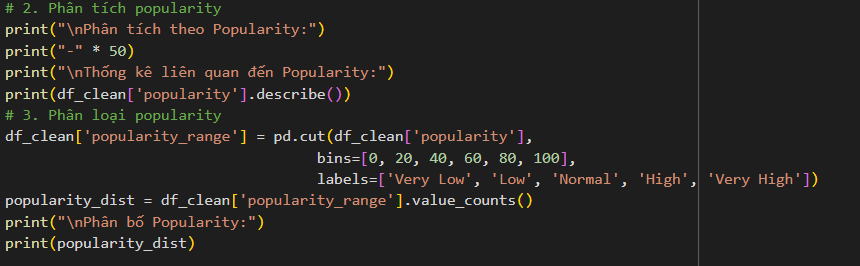
1. **PHÂN TÍCH VÀ TRỰC QUAN HÓA**

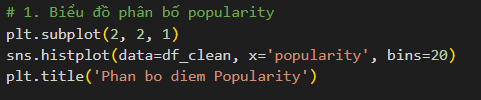
**2.1) Phân tích theo popularity**

Để thuận tiện cho việc phân tích, chúng em đã phân loại điểm popularity thành 5 mức độ:

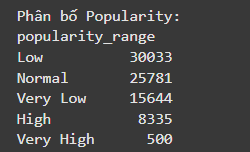
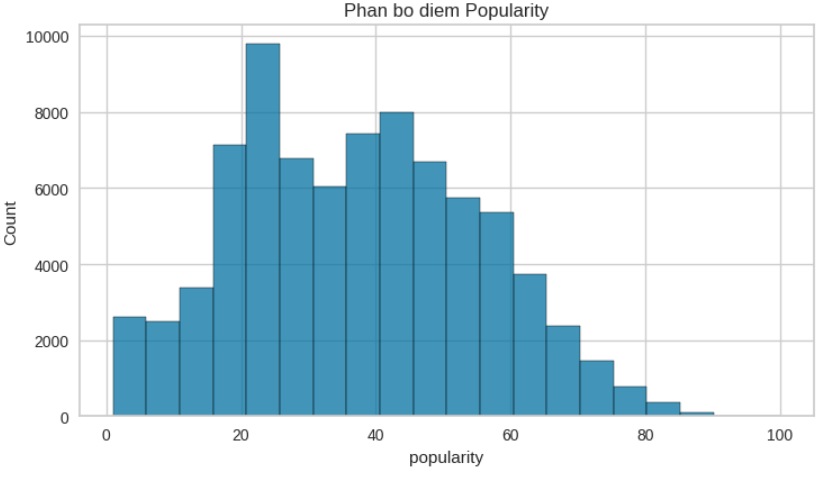
* Very Low: 0-20 điểm
* Low: 20-40 điểm
* Medium: 40-60 điểm
* High: 60-80 điểm
* Very High: 80-100 điểm

.Code:

****

****

Kết quả:

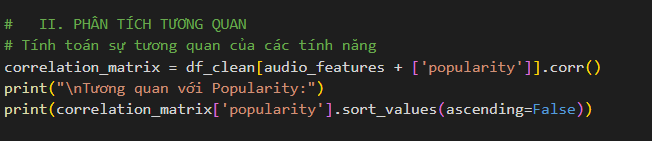
**** 

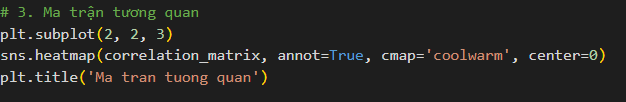
Nhận xét: Sau khi phân tích phân bố điểm popularity, chúng em nhận thấy một số đặc điểm nổi bật:

* Phần lớn các bài hát trong dataset tập trung ở mức popularity trung bình-thấp, dao động trong khoảng 20-40 điểm.
* Có sự sụt giảm đáng kể về số lượng bài hát khi mức popularity vượt quá 60 điểm.
* Tồn tại một khoảng trống nhỏ ở mức 0 điểm popularity, phản ánh việc đã loại bỏ các bài hát không có dữ liệu đánh giá.
* Số lượng bài hát có popularity lơn hơn 80 rất nhỏ với vỏn vẹn 500 bài ( chiếm khoảng 0.6% số bài hát)

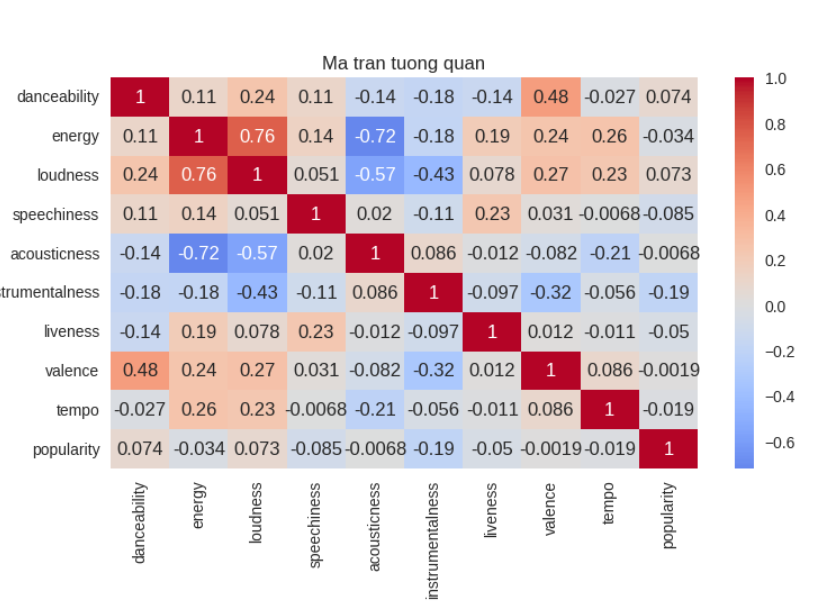
**2.2)Tính toán sự tương quan của các tính năng**

Code:





Kết quả:



Những cặp tính năng nổi bật’

· Energy và loudness: 0.76 (tương quan thuận mạnh)

· Energy và acousticness: -0.72 (tương quan nghịch mạnh)

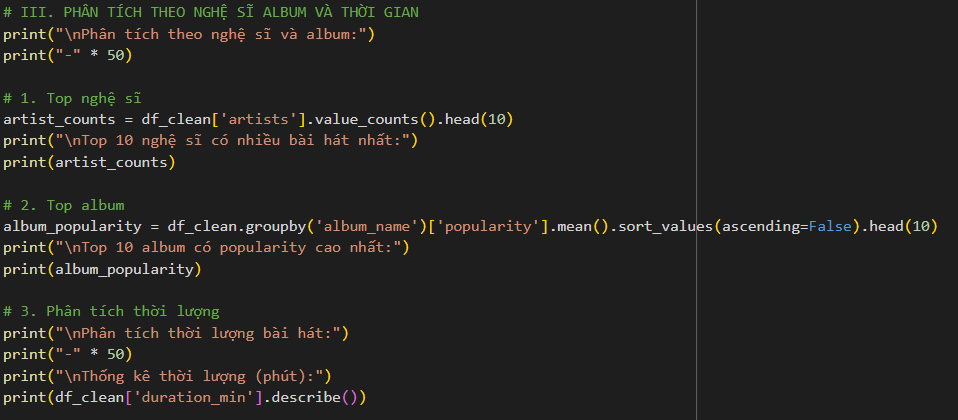
· Loudness và acousticness: -0.57 (tương quan nghịch trung bình)

Từ kết quả phân tích tương quan giữa các thuộc tính âm nhạc, chúng em rút ra được một số nhận xét quan trọng:

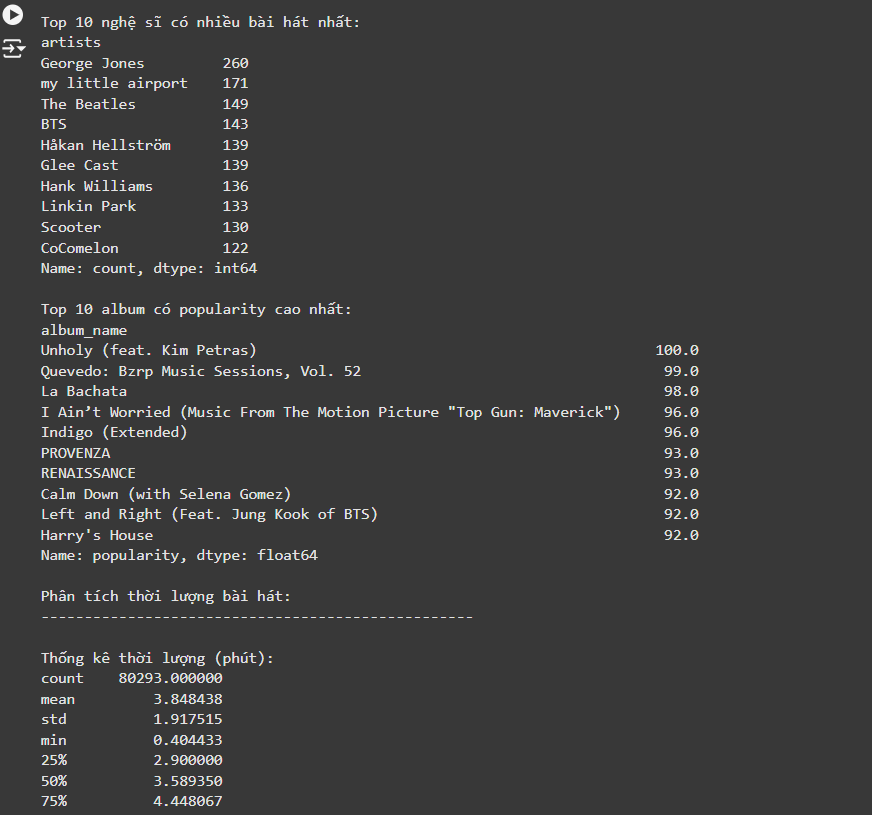
* Mối tương quan thuận chiều mạnh mẽ được tìm thấy giữa energy (năng lượng) và loudness (âm lượng) của bài hát. Điều này cho thấy những bài hát có mức năng lượng cao thường đi kèm với âm lượng lớn.
* Đáng chú ý, chỉ số popularity (độ phổ biến) không thể hiện mối tương quan đáng kể với bất kỳ đặc tính âm nhạc nào khác trong dataset, thể hiện qua các hệ số tương quan gần bằng 0. Điều này gợi ý rằng sự thành công của một bài hát không phụ thuộc vào các đặc tính kỹ thuật đơn lẻ mà có thể là kết quả của nhiều yếu tố phức tạp khác.

**2.3)Phân tích về nghệ sĩ , album và thời gian**

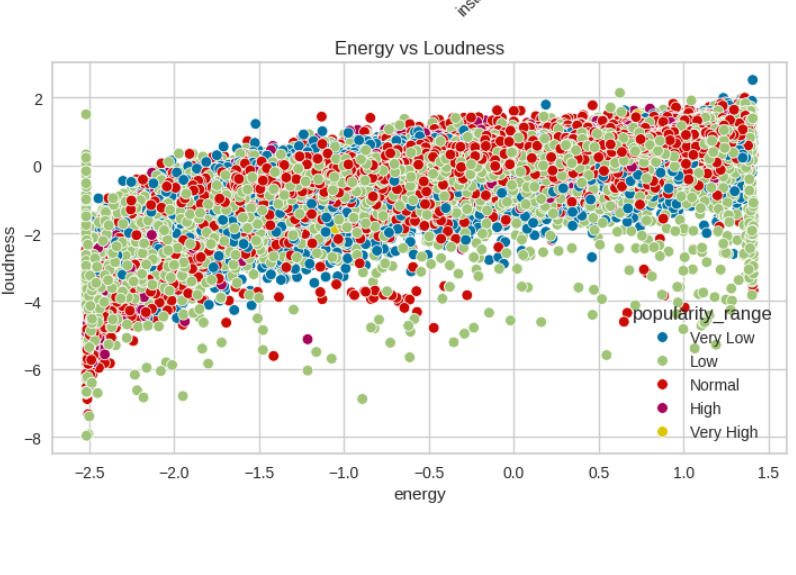
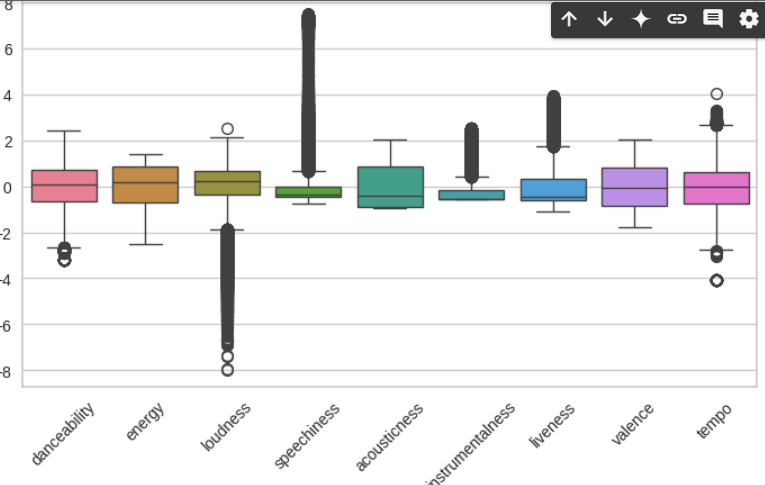
Code:



Kết quả:

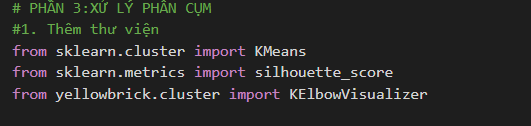


**2.4) Một số biểu đồ khác:**

1. **XỬ LÝ PHÂN CỤM**

Thêm các thư viện cần thiết cho xử lý phân cụm:

****

**3.1 Phân cụm dữ liệu**

3.1.1 Chuẩn bị dữ liệu cho phân cụm

Để tối ưu hóa quá trình phân cụm và giảm thiểu độ phức tạp trong tính toán, chúng em đã lựa chọn 6 đặc trưng âm nhạc quan trọng nhất:

* Danceability (Khả năng tạo cảm giác nhảy múa)
* Energy (Mức độ năng lượng)
* Valence (Mức độ tích cực)
* Tempo (Nhịp độ)
* Loudness (Độ lớn của âm thanh)
* Acousticness (Mức độ acoustic)

Việc tập trung vào các đặc trưng cốt lõi này không chỉ giúp giảm thời gian xử lý mà còn đảm bảo chất lượng của kết quả phân cụm.

# 1.Chuẩn bị dữ liệu cho phân cụm

features\_to\_cluster = ['danceability', 'energy', 'valence', 'tempo', 'loudness', 'acousticness']

X = df\_clean[features\_to\_cluster]

scaler = StandardScaler()

X\_scaled = scaler.fit\_transform(X)

3.1.2 Tìm số cụm tối ưu

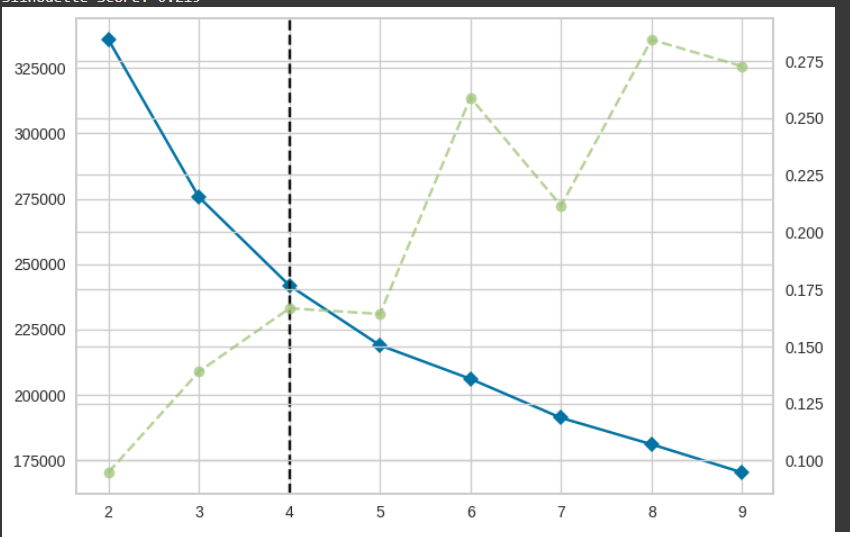
# 2.Tìm số cụm tối ưu

model = KMeans()

visualizer = KElbowVisualizer(model, k=(2,10))

visualizer.fit(X\_scaled)

optimal\_k = visualizer.elbow\_value\_



Dựa vào biểu đồ, chúng ta có thể thấy điểm khuỷu tay (elbow point) rõ ràng tại k=4, nơi đường màu xanh dương (thể hiện Within-Cluster Sum of Squares - WSS) bắt đầu giảm chậm lại đáng kể. Điều này cho thấy:

* Khi k < 4: WSS giảm mạnh, cho thấy việc tăng số cụm mang lại cải thiện đáng kể
* Khi k > 4: WSS giảm chậm lại, việc tăng số cụm không còn mang lại nhiều cải thiện

Ngoài ra, đường màu xanh lá (có thể là Silhouette Score) cũng cho thấy một mức tăng tại k=4, hỗ trợ thêm cho việc lựa chọn này.

Vì vậy, việc chọn k=4 là phù hợp, cân bằng giữa độ phức tạp của mô hình và chất lượng phân cụm.

3.1.3 Thực hiện Kmeans và kết quả

Code:

#3. Thực hiện phân cụm theo Kmeans

kmeans = KMeans(n\_clusters=optimal\_k, random\_state=42)

df\_clean['cluster'] = kmeans.fit\_predict(X\_scaled)

#4.Phân tích các cụm

cluster\_stats = df\_clean.groupby('cluster')[features\_to\_cluster].mean()

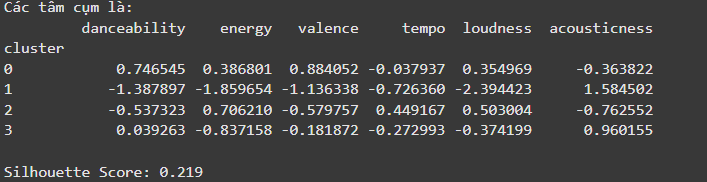
print("\nCác tâm cụm là:")

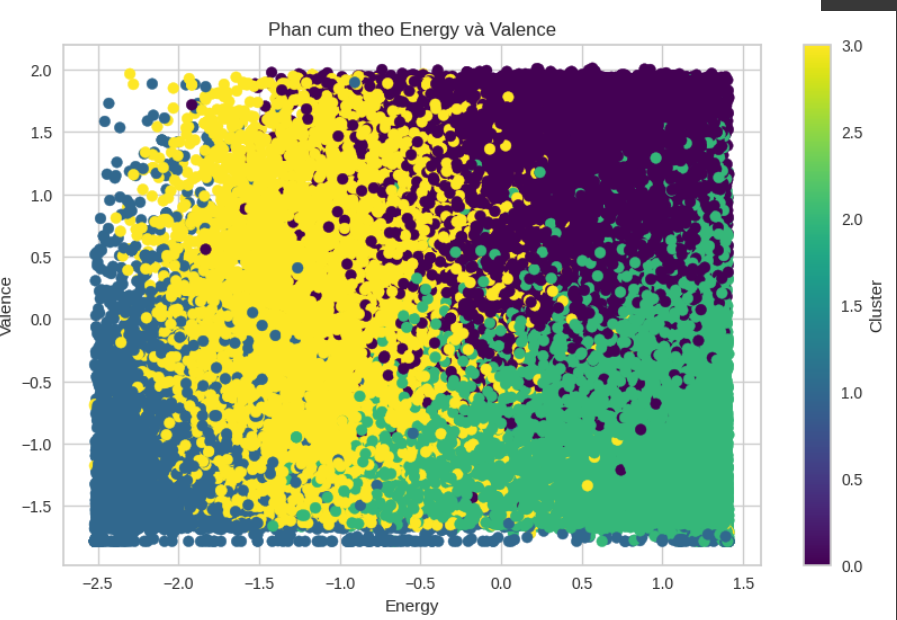
print(cluster\_stats)

silhouette\_avg = silhouette\_score(X\_scaled, df\_clean['cluster'])

print(f"\nSilhouette Score: {silhouette\_avg:.3f}")

Kết quả:





**3.2. Phân tích từng cụm**

Code:#2. Phân tích về các cụm

def analyze\_clusters(df, features):

    for cluster in df['cluster'].unique():

        cluster\_data = df[df['cluster'] == cluster]

        print(f"\nCụm {cluster}:")

        print(f"Số bài hát: {len(cluster\_data)}")

        print(f"Điểm popularity trung bình: {cluster\_data['popularity'].mean():.2f}")

        # Đặc trưng âm nhạc trung bình

        print("Đặc trưng âm nhạc trung bình:")

        print(cluster\_data[features].mean())

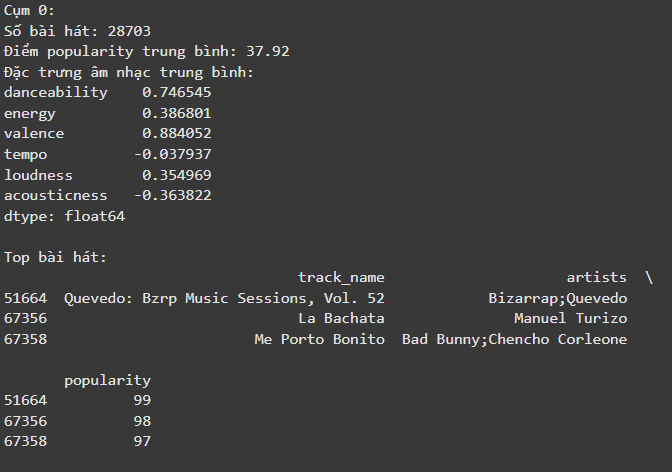
        # Top songs

        print("\nTop bài hát:")

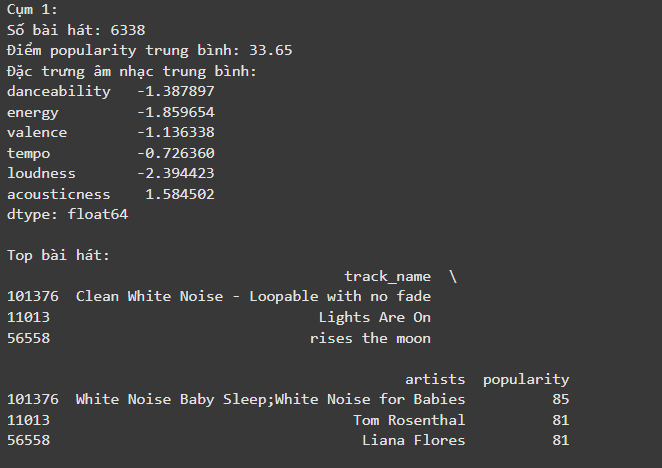
        print(cluster\_data.nlargest(3, 'popularity')[['track\_name', 'artists', 'popularity']])

Kết quả các cụm:

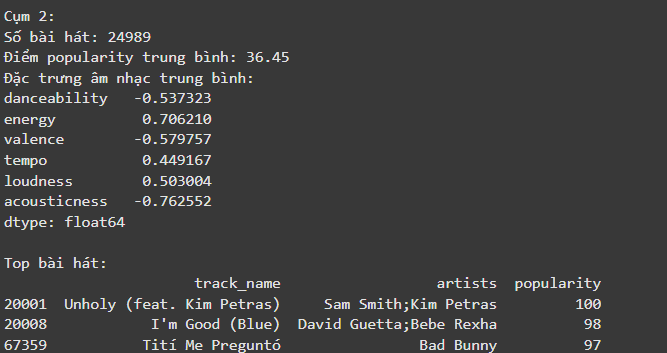
Cụm 0:



Cụm 1:

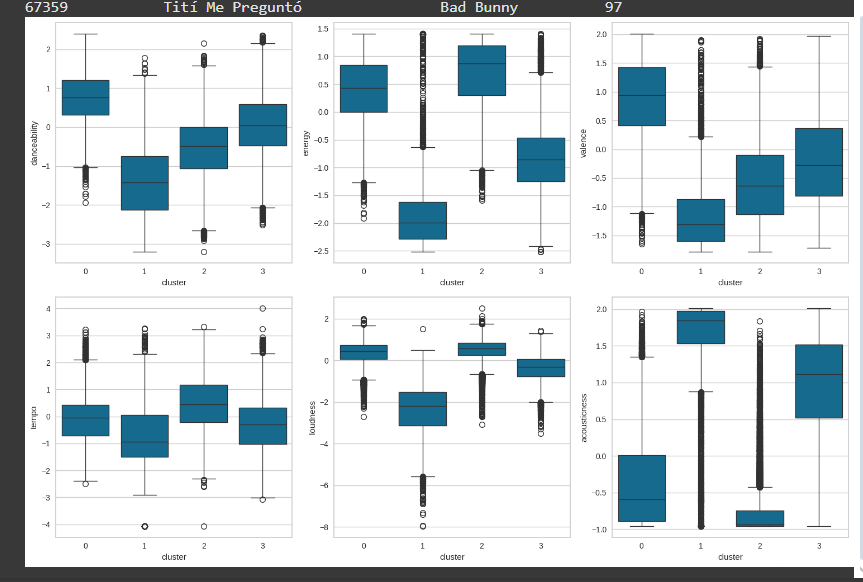


Cụm 2:



Cụm 3:





Đây là biểu đồ boxplot thể hiện đặc điểm của 6 thuộc tính âm nhạc chính trong từng cụm (cluster).Mỗi biểu đồ box thể hiện phân phối giá trị của một thuộc tính âm nhạc trong 4 cụm (từ 0 đến 3), giúp chúng ta có thể so sánh và phân biệt đặc trưng của từng cụm.

**4.HỆ THỐNG GỢI Ý (BONUS)**

Hàm recommend\_songs() được xây dựng để đề xuất những bài hát tương tự với bài hát đầu vào, hoạt động theo các bước sau:

* Nhận vào ID của bài hát cần gợi ý và số lượng bài hát muốn đề xuất (mặc định là 5 bài)
* Trích xuất các đặc trưng âm nhạc của bài hát đầu vào
* Xác định cluster mà bài hát đó thuộc về
* Tìm kiếm các bài hát cùng cluster và tính toán độ tương đồng
* Trả về danh sách các bài hát có độ tương đồng cao nhất, bao gồm tên bài hát, nghệ sĩ và điểm popularity

Cách tiếp cận này giúp đảm bảo các bài hát được đề xuất có đặc tính âm nhạc gần gũi với bài hát gốc.

def recommend\_songs(song\_id, n\_recommendations=5):

    # Lấy đặc trưng của bài hát đầu vào

    song\_features = df\_clean[df\_clean['track\_id'] == song\_id][features\_to\_cluster]

    # Dự đoán cluster của bài hát

    song\_cluster = kmeans.predict(scaler.transform(song\_features))[0]

    # Lọc các bài hát cùng cluster

    cluster\_songs = df\_clean[df\_clean['cluster'] == song\_cluster]

    # Tính độ tương tự

    similarities = cluster\_songs.apply(lambda x: np.linalg.norm(

        song\_features.values - x[features\_to\_cluster].values), axis=1)

    #Lấy n bài hát tương tự nhất

    recommendations = cluster\_songs.loc[similarities.nsmallest(n\_recommendations+1).index[1:]]

return recommendations[['track\_name', 'artists', 'popularity']]

**5.TỔNG KẾT(cho phần clustering)**

|  |  |
| --- | --- |
| Thuộc Tính | Lí do lựa chọn |
| danceability | Cho biết mức độ phù hợp để nhảy của bài hát. Quan trọng để xác định thể loại và mục đích sử dụng. |
| energy | Đo lường mức độ sôi động của bài hát. Giúp phân biệt giữa nhạc năng động và nhạc nhẹ nhàng. |
| valance | Thể hiện cảm xúc tích cực trong bài hát. Dùng để nhóm nhạc theo tâm trạng. |
| tempo | Tốc độ bài hát (BPM). Cần thiết để phân loại theo nhịp độ nhanh-chậm. |
| loudness | Độ lớn của âm thanh (dB). Giúp phân biệt các phong cách sản xuất âm nhạc. |
| acousticness | Mức độ sử dụng nhạc cụ acoustic so với nhạc điện tử. Dùng để phân loại phong cách âm nhạc. |

|  |  |
| --- | --- |
| Cụm | Tâm cụm |
| 0 | [0.74,0.38,0.88,-0.03,0.35,-0.36] |
| 1 | [-1.38,-1.86,-1.13,-0.72,-2.4,1.59] |
| 2 | [-0.54,0.71,-0.58,0.45,0.5,-0.76] |
| 3 | [0.04.-0.84,-0.18,-0.27,-0.37,0.96] |

|  |  |
| --- | --- |
| Đề xuất hành động | Lí do đề xuất |
| Cluster 0 thường dùng cho quán bar,club, các sự kiến lớn | Nhóm này có chỉ số danceability (0.74) và valence (0.88) cao nhất, phù hợp với không gian giải trí sôi động |
| Cluster 1 phù hợp cho các buổi thiền và yoga | Với acousticness cực cao (1.59) và tất cả các chỉ số khác đều rất thấp, đây là những bài hát phù hợp nhất cho hoạt động cần sự tĩnh lặng tuyệt đối |
| Cluster 2 phù hợp cho các phòng gym và hoạt động thể thao | Nhóm này có energy cao (0.71) và loudness cao (0.5), phù hợp với các hoạt động cần nhiều năng lượng |
| Cluster 3 phù hợp cho các quán bar và nhà hàng | |  | | --- | | Với acousticness cao nhưng các chỉ số khác ở mức vừa phải, phù hợp làm nhạc nền cho không gian công cộng, vừa đủ để tạo không khí nhưng không quá tĩnh lặng | |