### Các vấn đề gặp phải trong file này:

**Dữ liệu thiếu**:

* 1. Một số cột như Track Name, Artist URI(s), Artist Name(s), Album Image URL, ISRC, Artist Genres, và Album Genres có dữ liệu thiếu.
  2. Album Genres hoàn toàn không có dữ liệu.

**Dữ liệu trùng lặp**:

* 1. File có thể chứa dữ liệu trùng lặp trong các cột như Track Name hoặc Artist Name(s).

**Định dạng ngày tháng**:

* 1. Album Release Date có thể cần được chuẩn hóa thành định dạng ngày tháng (datetime).

**Dữ liệu phân loại không phù hợp**:

* 1. Một số cột chứa dữ liệu phân loại (như Explicit, Popularity) có thể cần kiểm tra tính hợp lệ và loại bỏ giá trị ngoại lai.

**Cột dư thừa**:

* 1. Một số cột không cần thiết cho phân tích, ví dụ: Track URI, Album URI, Album Image URL.

### Hướng giải quyết:

**Xử lý dữ liệu thiếu**:

* 1. Loại bỏ các dòng chứa dữ liệu thiếu quan trọng (ví dụ: Track Name, Artist Name(s)).
  2. Điền giá trị mặc định hoặc dự đoán cho các cột ít quan trọng hơn.

**Loại bỏ trùng lặp**:

* 1. Dùng các cột quan trọng (ví dụ: Track Name, Artist Name(s)) để kiểm tra và xóa các dòng trùng lặp.

**Chuẩn hóa dữ liệu**:

* 1. Chuyển Album Release Date sang định dạng ngày tháng (datetime).
  2. Chuẩn hóa các cột dạng phân loại (ví dụ: Explicit) nếu cần thiết.

**Loại bỏ cột dư thừa**:

* 1. Giữ lại các cột liên quan đến phân tích, ví dụ:
     1. Track Name, Artist Name(s), Popularity, Danceability, Energy, Valence, Tempo, v.v.

**Phát sinh dữ liệu mới**:

* 1. Tính thêm các cột hoặc đặc trưng mới từ dữ liệu hiện có, ví dụ:
     1. Track Duration (minutes) = Track Duration (ms) / 60000.
     2. Decade từ Album Release Date.