Comprehensive Analysis of Music Data

Sherali Ozodov

ISTA 322: Data Engineering, Spring 2024

University of Arizona

May 1, 2024

Introduction

Objective: Briefly restate the project's objective to develop a full data engineering pipeline that includes extracting data from Spotify's API, transforming it for analysis, and loading it into a SQL database.
Purpose: Explain the intention to analyze music trends to aid stakeholders in making informed decisions.

Data Sources and Data Integration

Data Sources

Spotify API:
Link: https://developer.spotify.com/documentation/web-api/
Data Available: Album, artist, and track details including names, IDs, release dates, and track features like danceability and energy.
Kaggle (Additional Track Features):
Link: https://www.kaggle.com/datasets/rodolfofigueroa/spotify-12m-songs
Data Available: Track IDs, popularity scores, key, mode, and acoustic features.

Data Integration

How Data Was Integrated:

- Extracted album and track information using the Spotify API.
- Merged track details with additional features from the Kaggle dataset based on track IDs.

Data Utilization:

- Used album and artist IDs to create relationships in the database.
- Integrated track features from both sources to enrich the dataset with comprehensive details for analysis.

Database Schema

Tables Description

Albums Table: Contains album ID, title, release date, and total tracks.

Artists Table: Stores artist ID and name.

Tracks Table: Includes track ID, album ID, track name, duration, and various musical features like tempo and loudness.

Album_Artist Table: Relational table linking artists to albums.

Table Definitions and Keys

1. Albums Table

- Columns:
  - album_id (PK): A unique identifier for each album.
  - title: The title of the album.
  - release_date: The release date of the album, stored in DATE format.
  - total_tracks: The total number of tracks on the album.
- Primary Key: album_id
- Purpose: Stores information about each album, including its release date and total number of tracks.

2. Artists Table

- Columns:
  - artist_id (PK): A unique identifier for each artist.
  - name: The name of the artist.
- Primary Key: artist_id
- Purpose: Maintains artist details that can be linked to albums and tracks.

3. Tracks Table

- Columns:
  - track_id (PK): A unique identifier for each track.
  - album_id (FK): The identifier for the album the track appears on, linking to album_id in the Albums table.
  - name: The name of the track.

- duration_ms: The duration of the track in milliseconds.
- explicit: Boolean value indicating if the track has explicit content.
- danceability, energy, key, loudness, speechiness, acousticness, instrumentalness, liveness, valence, tempo: Various musical features of the track extracted from the Spotify API and Kaggle dataset.

- Primary Key: track_id
- Foreign Key: album_id references album_id in the Albums table.
- Purpose: Contains detailed attributes for each track, including a reference to its album, facilitating detailed track-level analysis.

4. Album_Artist Table

- Columns:
  - album_id (FK): The identifier for the album, linking to album_id in the Albums table.
  - artist_id (FK): The identifier for the artist, linking to artist_id in the Artists table.
- Primary Key: Composite key (album_id, artist_id)
- Foreign Keys:
  - album_id references album_id in the Albums table.
  - artist_id references artist_id in the Artists table.
- Purpose: Creates a many-to-many relationship between artists and albums. This table allows albums to have multiple artists and vice versa, reflecting collaborations and various artist contributions to single albums.

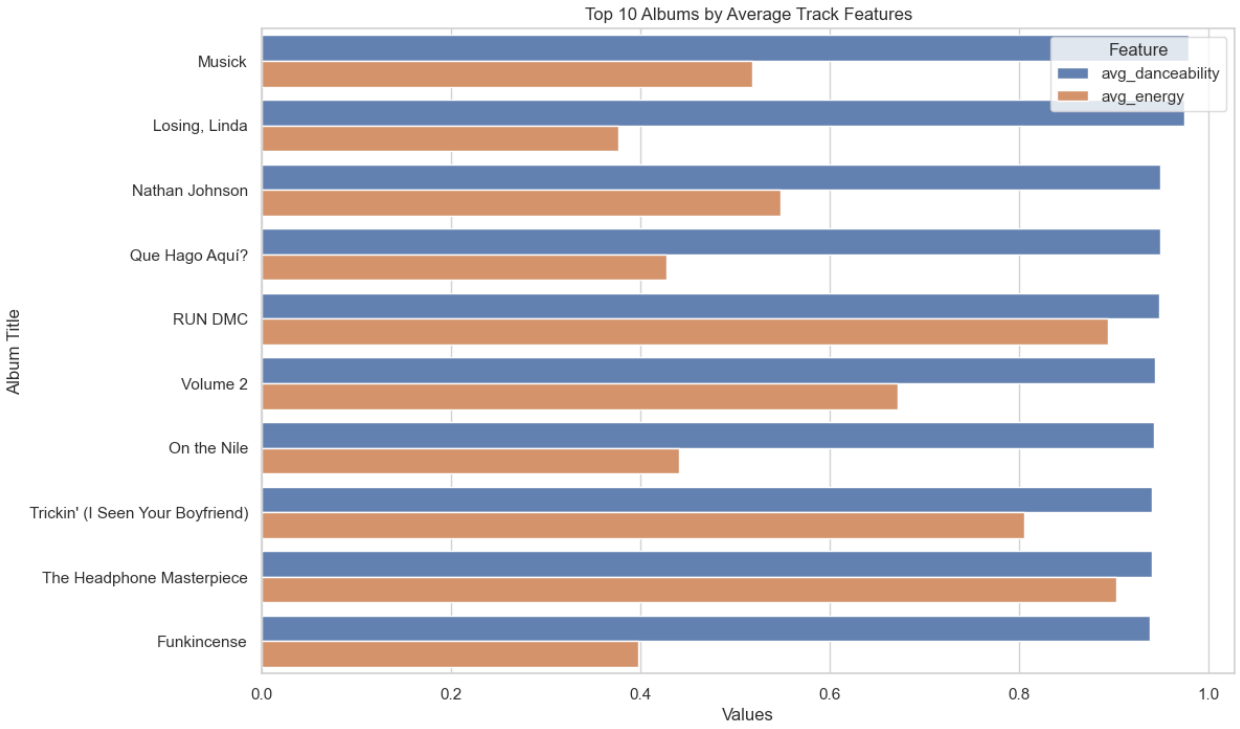

Album_Artist
album_id varchar
artist_id varchar
Artists
artist_id varchar
name varchar
Albums
album_id varchar
title varchar
release_date date
total_tracks integer
Tracks
track_id varchar
album_id varchar
name varchar
duration_ms integer
explicit boolean
danceability double
energy double
key integer
loudness double
speechiness double
acousticness double
instrumentalness double
liveness double
valence double
tempo double
dbdiagram.io

Queries and plots

Average Track Features by Album

Query to find the top 10 albums by average danceability and average energy

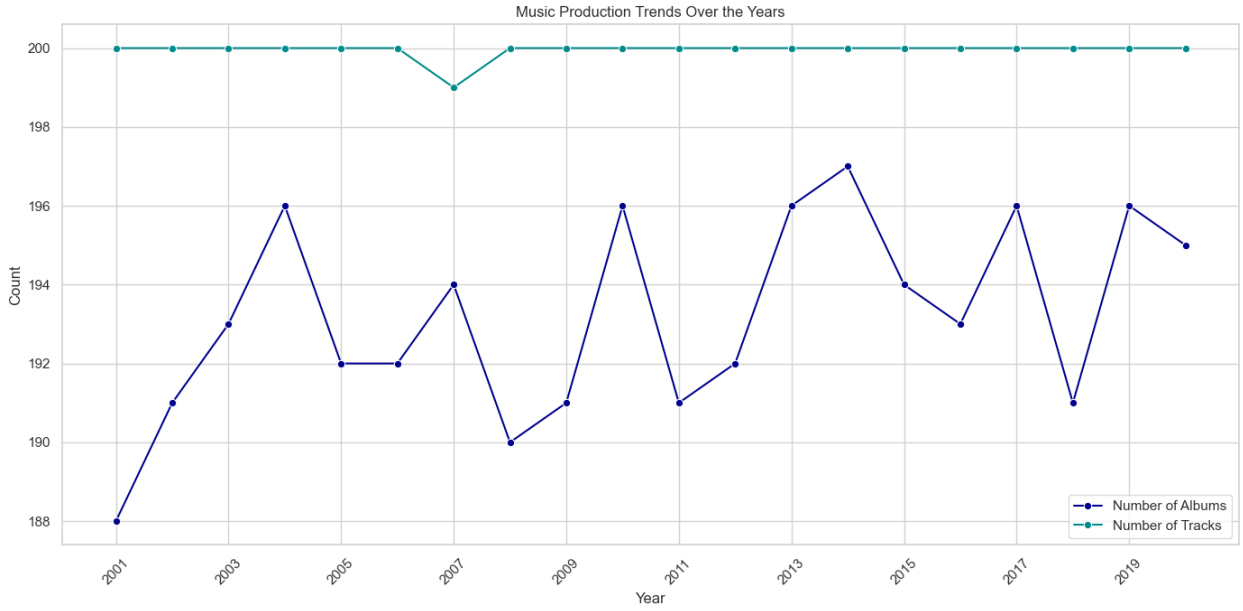
```
run_query("""
SELECT albums.title as album_title, AVG(tracks.danceability) AS avg_danceability,
AVG(tracks.energy) AS avg_energy
FROM tracks
JOIN albums ON tracks.album_id = albums.album_id
GROUP BY albums.title
ORDER BY avg_danceability DESC LIMIT 10;
""")
```



Count of Albums and Tracks per Year

Query to count the number of albums and tracks produced each year

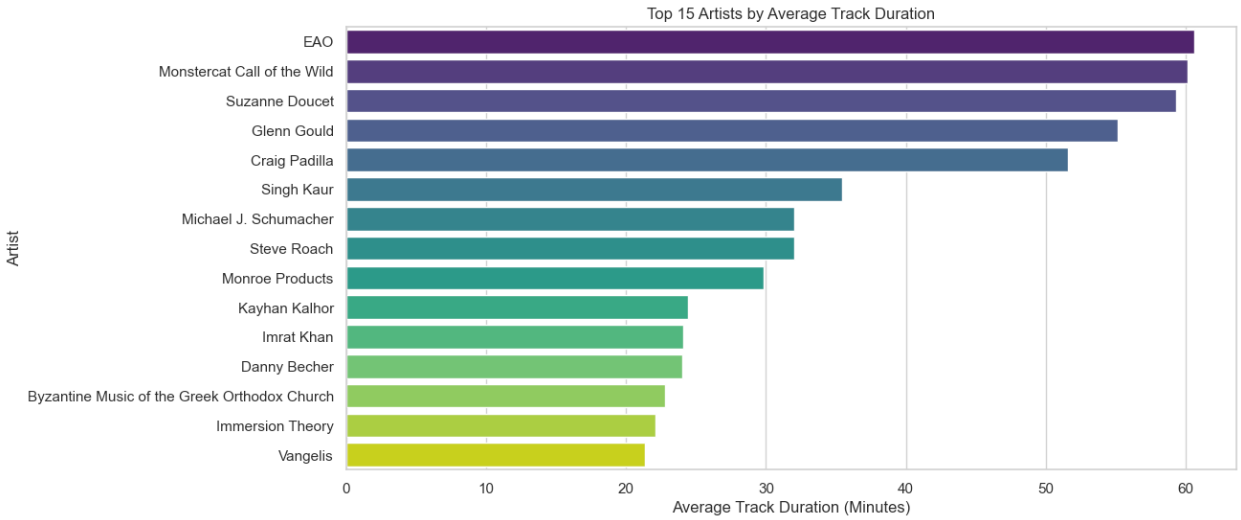
```
run_query("""
SELECT EXTRACT(YEAR FROM a.release_date) AS release_year,
       COUNT(DISTINCT a.album_id) AS num_albums,
       COUNT(t.track_id) AS num_tracks
FROM albums a
LEFT JOIN tracks t ON a.album_id = t.album_id
GROUP BY release_year
ORDER BY num_albums DESC LIMIT 20;
""")
```



Average Track Duration by Artist

Query to calculate the average duration of tracks by each artist, in minutes

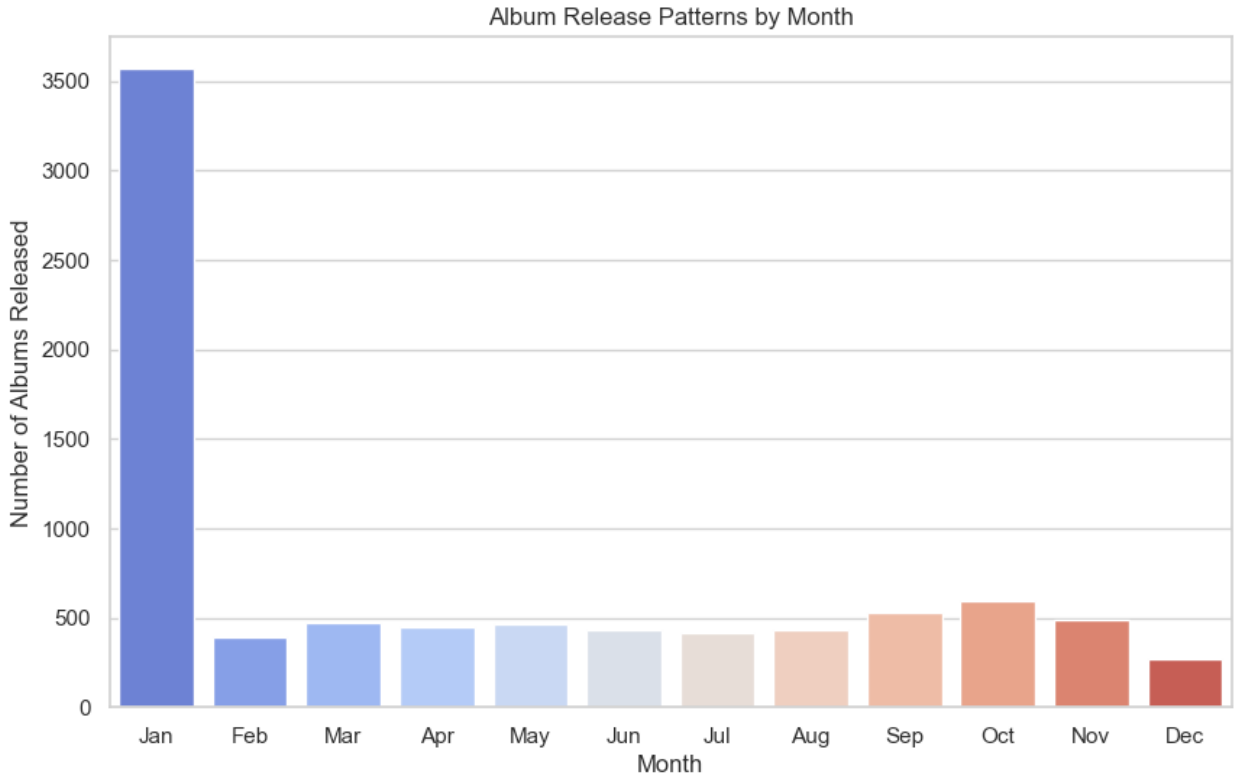
```
run_query("""
SELECT artists.name AS artist_name, AVG(tracks.duration_ms) / 60000 AS
avg_duration_minutes
FROM artists
JOIN album_artist ON artists.artist_id = album_artist.artist_id
JOIN tracks ON album_artist.album_id = tracks.album_id
GROUP BY artists.name
ORDER BY avg_duration_minutes DESC LIMIT 15;
""")
```



Album Release Patterns by Month

Query to count the number of albums released each month, aggregated over all years

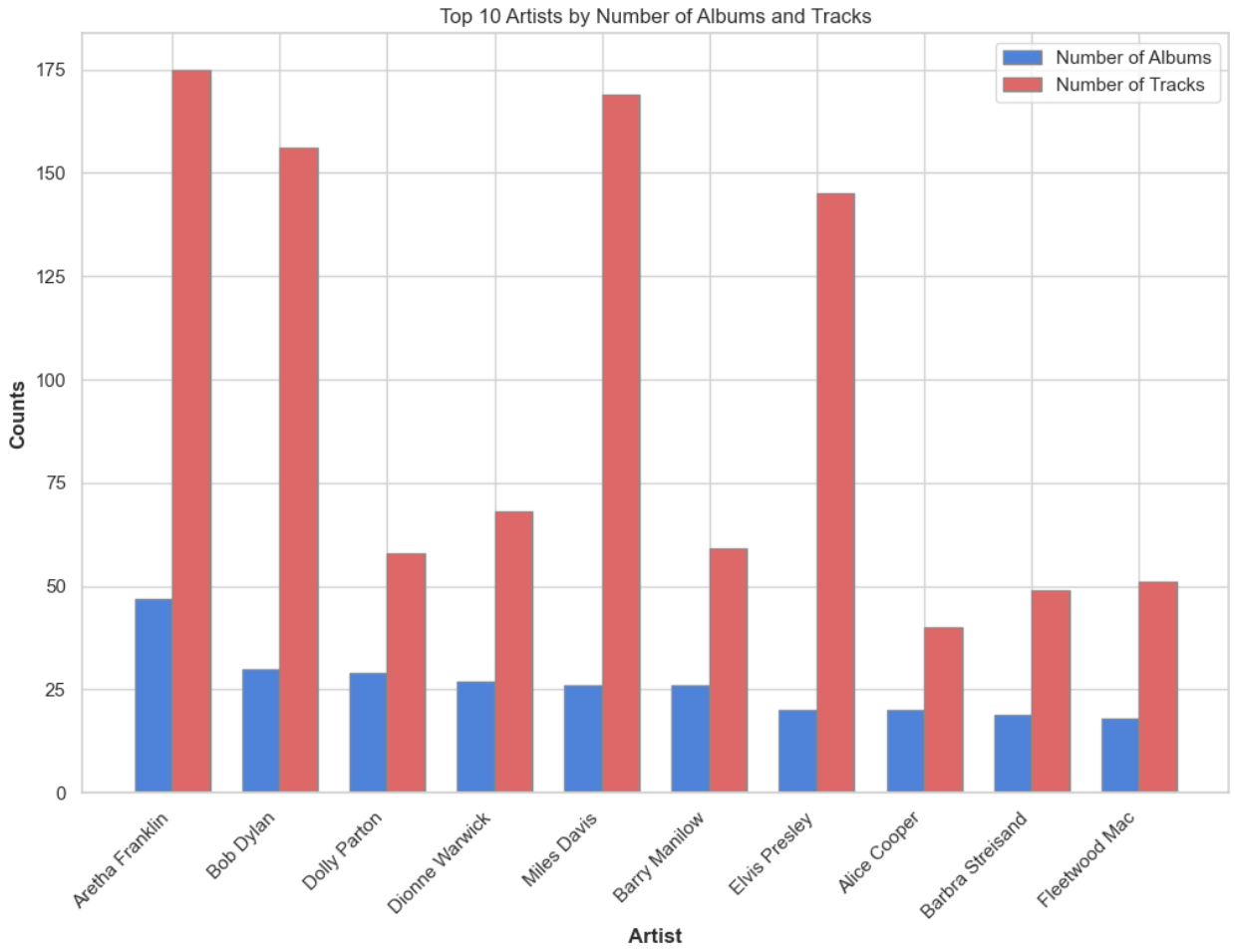
```
run_query("""
SELECT EXTRACT(MONTH FROM release_date) AS month, COUNT(*) AS num_albums
FROM albums
GROUP BY month
ORDER BY num_albums DESC LIMIT 12;
""")
```



Detailed Artist Analysis

Query to rank artists by the number of albums and tracks they are associated with

```
run_query("""
SELECT artists.name, COUNT(DISTINCT albums.album_id) AS num_albums,
COUNT(tracks.track_id) AS num_tracks
FROM artists
JOIN album_artist ON artists.artist_id = album_artist.artist_id
JOIN albums ON album_artist.album_id = albums.album_id
JOIN tracks ON albums.album_id = tracks.album_id
GROUP BY artists.name
ORDER BY num_albums DESC, num_tracks DESC LIMIT 10;
""")
```


Top 10 Artists by Number of Albums and Tracks
Number of Albums
Number of Tracks
Counts
0
25
50
75
100
125
150
175
Aretha Franklin
Bob Dylan
Dolly Parton
Dionne Warwick
Miles Davis
Barry Manilow
Elvis Presley
Alice Cooper
Barbra Streisand
Fleetwood Mac
Artist