A visual music recommender system powered by spotify data

FINAL PROJECT

Visual Analytics

Julia Othats-Dalès Gibert, NIA: 254435

Marina Castellano Blanco, NIA: 242409

2025-2026

Table of Contents

1. Problem Statement

Today's music streaming platforms provide instant access to millions of songs. While this abundance is a tremendous advantage, it also creates an overwhelming choice problem: with so much music to choose from, listeners often struggle to discover songs that genuinely match their tastes. Traditional recommendation systems often rely solely on user behavior, such as listening history or explicit ratings, which can limit discovery to familiar artists or genres and fail to capture the nuanced characteristics of the music itself.

The goal of this project is to address this challenge by leveraging Spotify's rich dataset of track, artist, and audio feature information to develop a visual music recommender system. This system aims to not only recommend similar songs based on detailed audio features but also provide insights into song popularity trends, genre evolution, and the underlying attributes that influence listener preferences. By combining data-driven analysis, interactive visualizations, and machine learning techniques (such as K-Nearest Neighbors for similarity, Random Forests and Logistic Regression for popularity and genre prediction), the system seeks to enhance music discovery and help users explore new music in a more personalized and informed way.

2. Dataset Overview

We worked with three main datasets, cleaned and merged to create a comprehensive view of songs, artists, and their acoustic properties.

2.1 Spotify Tracks and Artists (spotify_data1 + spotify_data2)

This dataset provides song and artist-level metadata:

- **Tracks:** track_name, track_id, track_popularity, duration
- **Artists:** artist_name, artist_popularity, followers, genres
- **Albums:** album_name, release_date, track_number

After preprocessing, the final combined dataset contains 114,000 tracks.
Source link: *https://www.kaggle.com/datasets/wardabilal/spotify-global-music-dataset-20092025*

2.2 Artist Metadata (artist_country.csv)

Adds country of origin for each artist, enabling geographical and demographic analysis.
Source link: https://www.kaggle.com/datasets/jackharding/spotify-artist-metadata-top-10k

2.3. Audio Features (spotify_audio_features1 + spotify_audio_features2)

Contains detailed acoustic features for 2,688 popular songs released between 2010 and 2019.

Includes variables such as: bpm, energy, danceability, valence, loudness (dB), acousticness, speechiness, duration, live likelihood, popularity.
Source link: https://www.datacamp.com/datalab/datasets/dataset-python-spotify-music

3. Objectives and Insights

The primary objective of this project is to build a comprehensive music recommender system that leverages Spotify's track, artist, and audio feature data to provide personalized song recommendations. By analyzing the characteristics of songs, such as energy, danceability, valence, and acousticness, alongside genre and artist metadata, the system can identify tracks that are similar in style and mood, helping users discover new music that aligns with their tastes.

We also wanted to explore what factors drive popularity and genre: Do songs with higher energy or louder production tend to perform better? Is danceability a reliable signal of mainstream appeal? And to what extent can popularity be explained based on audio features alone?

Finally, the project emphasizes interactive visual analytics, both in the Streamlit app developed and in the Tableau Story, to make complex patterns in the data easily interpretable. From temporal trends and genre evolution to feature correlations and popularity distributions, these visualizations provide actionable insights for both users and analysts, supporting informed music discovery and a deeper appreciation of the factors shaping today's streaming landscape.

4. Methodology

Github repository: https://github.com/othats/SpotifyAnalytics2025

Data cleaning and preparation were essential to ensure consistency and reliability across the multiple datasets. For the track and artist datasets (spotify_data1 and spotify_data2), song durations were standardized to seconds, missing critical values were removed, and duplicate tracks were discarded. Artist names were cleaned to remove variations such as "feat." or "ft.", ensuring accurate aggregation and merging. Country metadata was incorporated, and temporal features, release year and month, were extracted from album release dates. The audio features dataset underwent similar preparation: duplicates were removed, genres were consolidated into broader categories, and extremely rare genres were excluded to improve analytical robustness.

With the datasets cleaned and integrated, we performed a comprehensive exploratory data analysis (**EDA**) to uncover patterns and relationships. Popularity distributions indicated that most songs fall into low to medium ranges, and a high number of artist followers did not always correlate with highly popular tracks. Genre analysis revealed clear temporal shifts: earlier decades were dominated by Rock, while more recent years show a marked rise in Pop, reflecting changes in listener habits in the streaming era. Temporal trends also highlighted an increase in song releases over time and a gradual decrease in average track duration. Analysis of audio features showed moderate correlations with popularity, with loudness, danceability, and energy exhibiting the strongest relationships.

Building on these insights, a content-based music recommender system was developed using K-Nearest Neighbors (KNN). Audio features were standardized, and genres were one-hot encoded to create a unified feature space. The KNN model, using Euclidean distance and ten neighbors, identifies songs with similar acoustic profiles. Functions were implemented to generate recommendations based on either a specific track or an artist's catalog. The preprocessing pipeline and trained model were saved for integration into a **Streamlit interface**.

To **predict track popularity**, the original 0–100 score was transformed into a binary classification task distinguishing low versus high popularity tracks. Logistic Regression, Decision Tree, and Naive Bayes classifiers were evaluated, with Logistic Regression achieving the highest performance (~70–75% accuracy). SHAP (SHapley Additive exPlanations) was applied to interpret the model, revealing that loudness, danceability, and energy contributed most to predictions, alongside genre effects.

A third machine learning model focused on **genre classification** using only audio features. A multi-class Logistic Regression model was trained and achieved an F1-score between 50% and 60%, indicating that while acoustic features provide useful signals, musical genre is influenced by additional factors such as lyrical content, production style, and audience perception.

All models were deployed in a Streamlit application for interactive exploration, which includes a welcome page, EDA visualizations, a music recommendation tool, popularity prediction with SHAP explanations, and a genre prediction interface. Complementary **Tableau dashboards** were created to visually communicate insights, covering temporal trends, genre evolution, feature correlations, geographic distribution of artists, and track-level acoustic profiles. Together, these tools provide a data-driven framework for personalized music discovery and exploration.

5. Conclusion

This project demonstrates how data-driven analysis and machine learning can enhance music discovery in the era of streaming, and regardless of user-behaviour "pollution". By cleaning and integrating Spotify datasets and audio feature datasets, we were able to uncover temporal trends, genre shifts, and relationships between acoustic features and song popularity. Our exploratory analysis highlighted that while Rock dominated earlier decades, Pop has become increasingly prevalent in recent years.

The machine learning models implemented provide practical tools for recommendation and prediction. The K-Nearest Neighbors model enables personalized song recommendations based on audio similarity, while the popularity prediction model identifies key features that drive listener engagement. Genre classification, though moderate in accuracy, reinforces the complexity of musical style and the multifaceted factors that define it.

Overall, this project illustrates the power of combining visual analytics, predictive modeling, and interactive tools to support personalized music exploration. The Streamlit application and Tableau dashboards provide an accessible and engaging way to interact with the data, empowering users to discover new music, understand trends, and gain actionable insights into what shapes listener preferences today.