SAPIENZA
UNIVERSITÀ DI ROMA

SPOTYVIS

Final project of Visual Analytics
A.Y. 2025/2026

NAME OF PROJECT:
SpotyVis

PRESENTED BY:
Masciotta Bruno 1989718
Maucione Mattia 2007209

PRESENTED TO:
Giuseppe Santucci

Agenda

Introduction

Music is usually an emotional experience, but beneath the sound, it is a complex dataset. Every song we stream is actually a collection of high-dimensional data points, our project was born from a simple question: Can we visualize the **DNA** of a hit song?

SpotyVis is a dashboard to explore visually the features characterizing the songs and study the layer of audio to reveal the hidden statistical patterns, trends, and correlations that define the music industry.

The final tool is an interactive visualization built in **D3.js** where different charts work together, in order to give the user a complete overview that allows users to detect meaningful patterns and outliers instantly.

Objectives and Goals

Goal # 1

Provides an interactive visual analytics tool to explore the dataset.

Goal # 2

Study the evolution of people’s musical tastes.

Goal # 3

Study the features of hit songs.

Related Works

[1] Decoding Spotify Hits: Statistical and Predictive Analysis of Track Features Driving Song Popularity (2025))

[2] Exploring Music Rankings with Interactive Visualization (2017)

[3] Dashboard Design: Interactive and Visual Exploration of Spotify Songs (2024))

Dataset

The dataset was created by joining and manipulating **Spotify Tracks Dataset**, **The Hot 100 Billboard Dataset** and by a webscraping on **whosampled.com.**

- **Normalize** all the features to a common 0-100 scale
- Filter the **Spotify Tracks** by keeping only those present in the **Hot 100 Billboard**, in this way we avoided irrelevant data
- Apply the **webscraping** on the resulting dataset in order to find:
    a. **Year** of publication
    b. Number of **covered/sampled/remixed** by
    c. Wheter it is a **cover/remix/sample** of

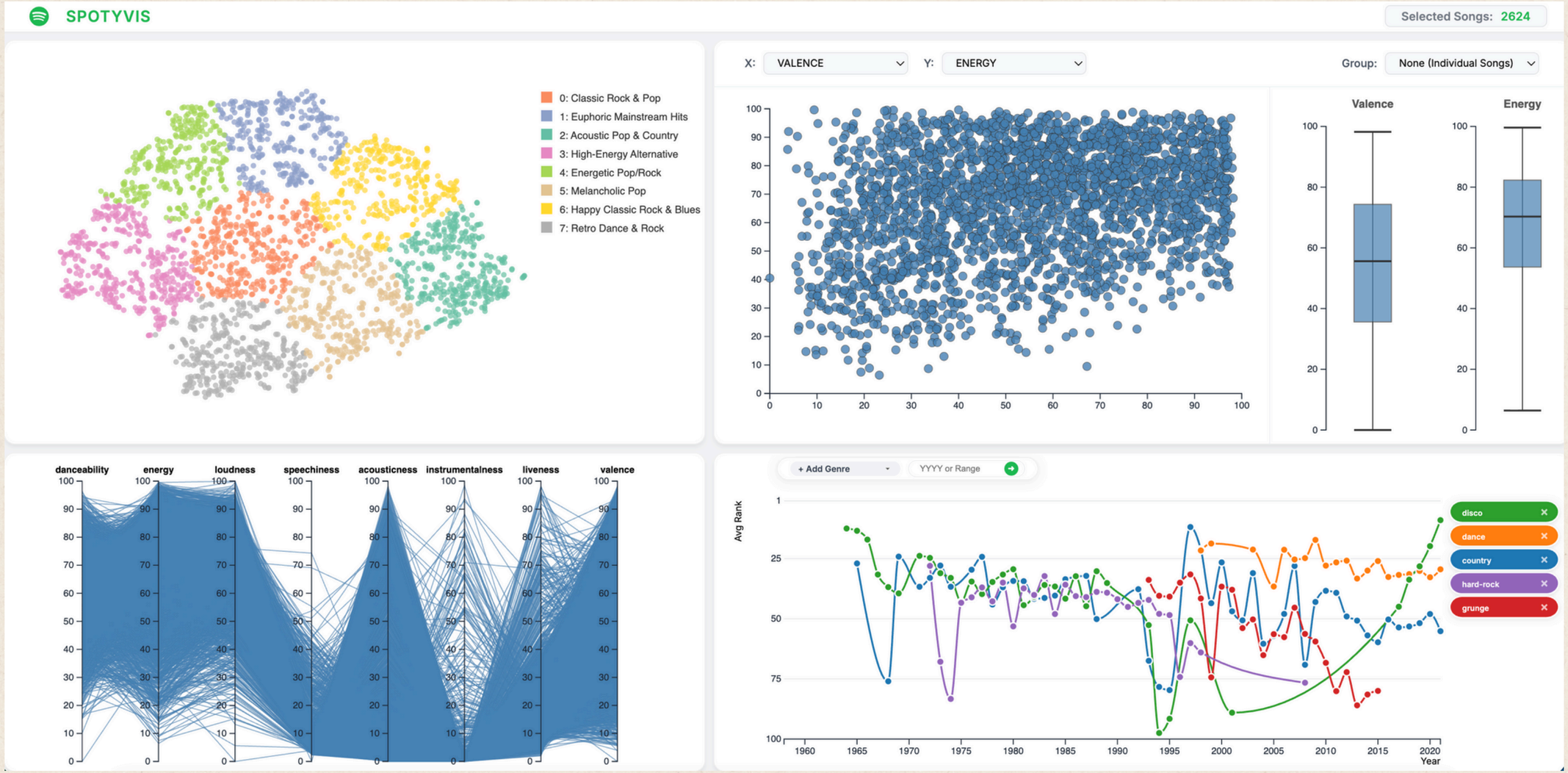
Dataset

artists: *Adele*
track_name: *Easy On Me*
Year: *2021*
track_id: *0gplL1WMoJ6iYaPgMCL0gX*
popularity: *85*
duration_ms: *224694*
danceability: *60.4*
energy: *36.6*
speechiness: *2.82*
acousticness: *57.8*
instrumentalness: *0.0*

loudness: *70.4*
liveness: *13.3*
valence: *13.0*
track_genre: *british*
Sampled_By: *3*
Is_Sample: *0*
Covered_By: *34*
Is_Cover: *0*
Remixed_By: *0*
Is_Remix: *0*

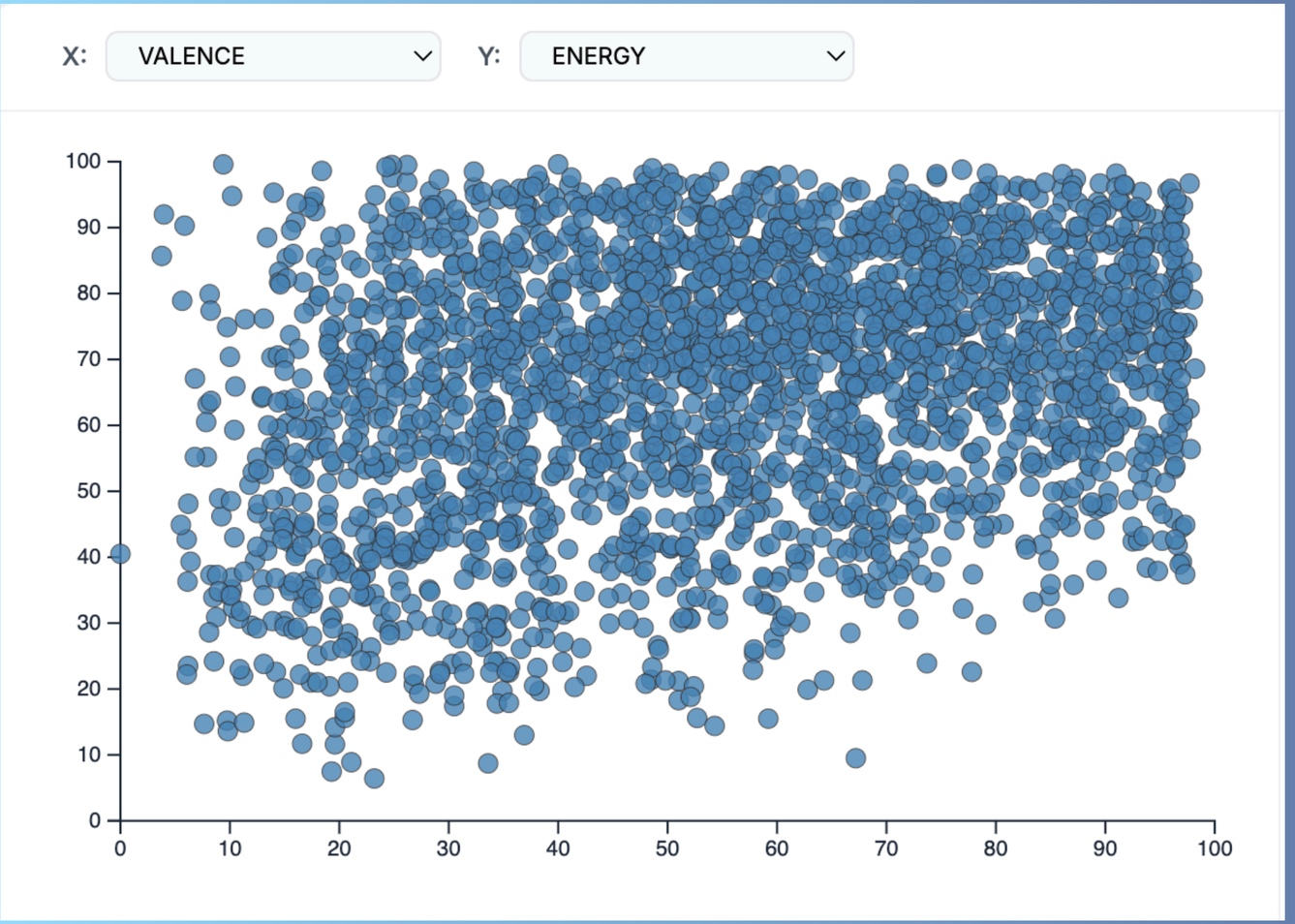
In addition to this master file, we maintain a **longitudinal dataset** of 46000 rows from the Billboard Hot 100 (filtered for Spotify compatibility).

Visualization



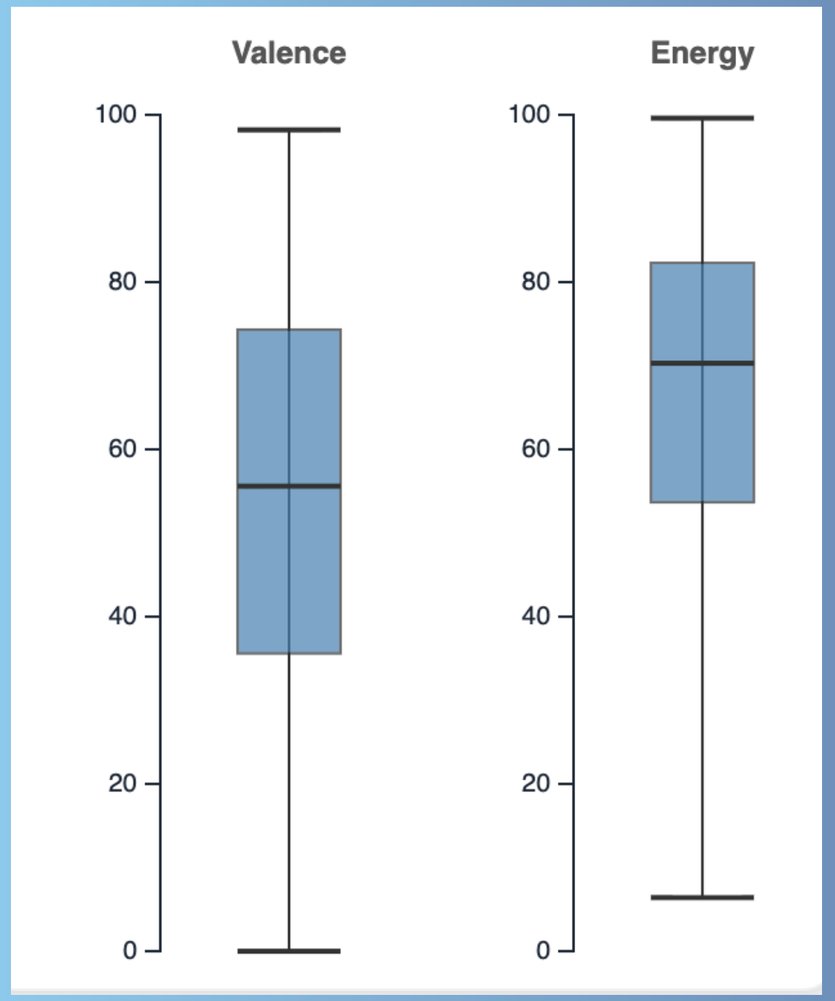
Bubbleplot

- Displays two continuous variables simultaneously within a two-dimensional Cartesian coordinate system.
- Each song is a circle of fixed dimension.
- Possibility of change axis's value and compare different features.
- Possibility of use the 'Group by' functionality.



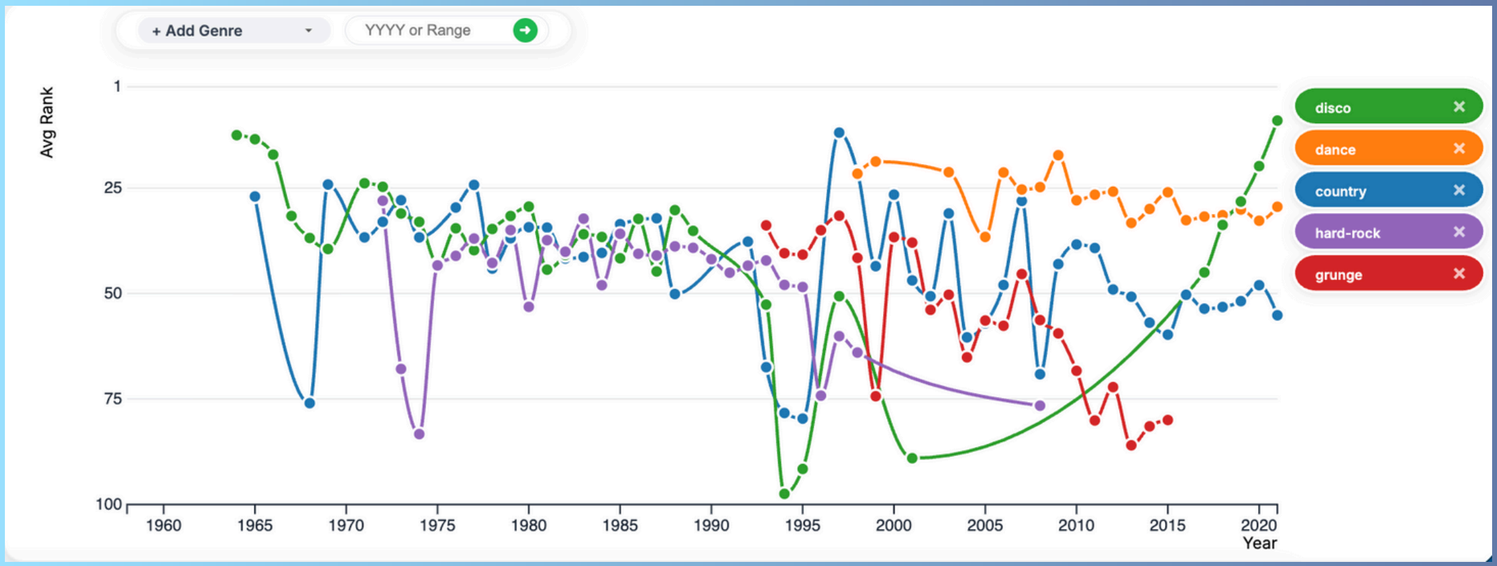
Boxplots

- Visualizes the distribution of continuous numerical data through a standardized five-number summary (minimum, first quartile, median, third quartile, and maximum).
- The central rectangle spans the Interquartile Range (IQR), effectively highlighting where the middle 50% of the dataset is concentrated.
- The values to show can be modified together with the bubble plot axis.
- The chart change if we select some bubble or drag on the parallel coordinates



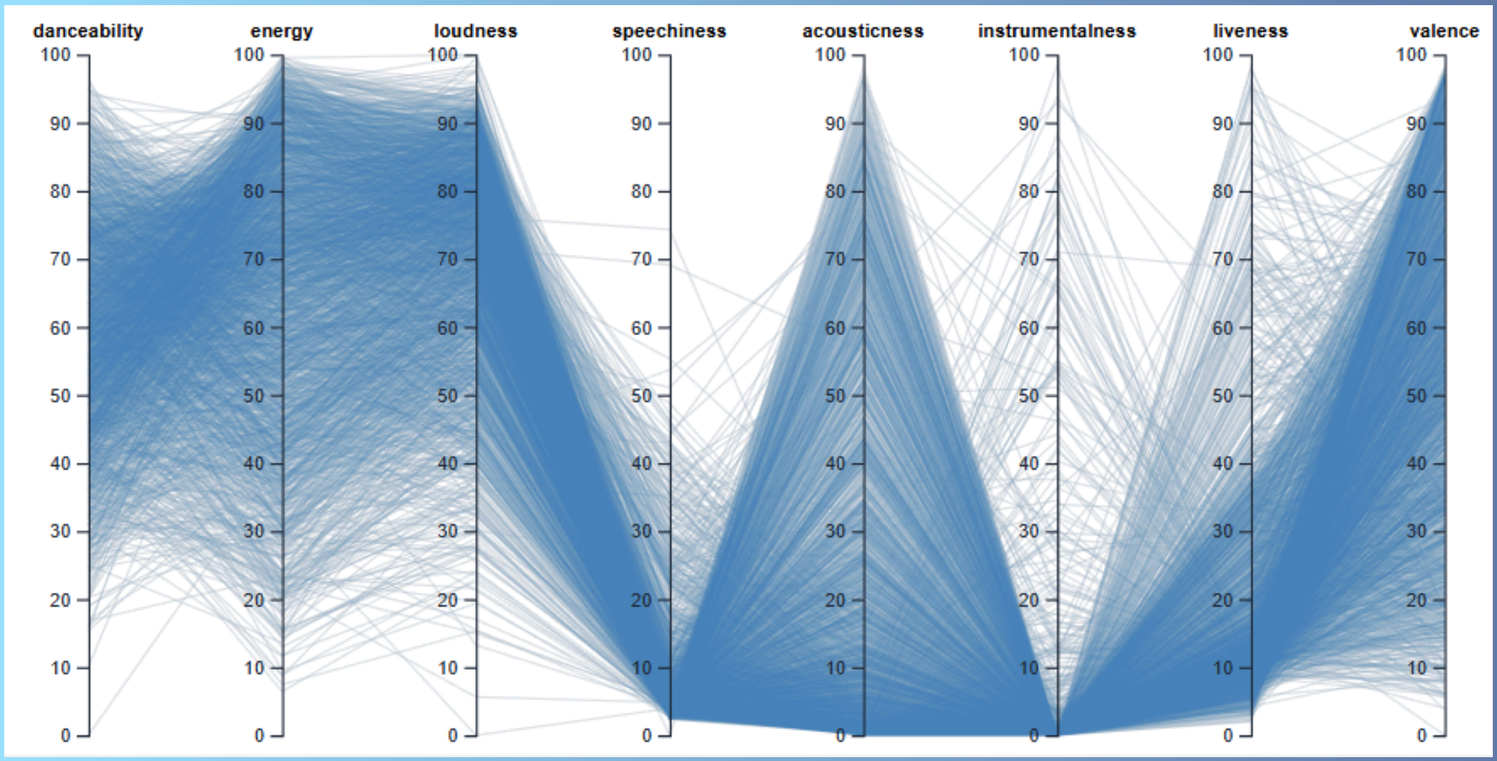
Bump chart

- Show the trend across the years of the top 5 genres from 1958 to 2021.
- Possibility of add or remove genre.
- Possibility to filter by year or range of years.
- Automatically computes the top 5 genres for each range of time we insert.



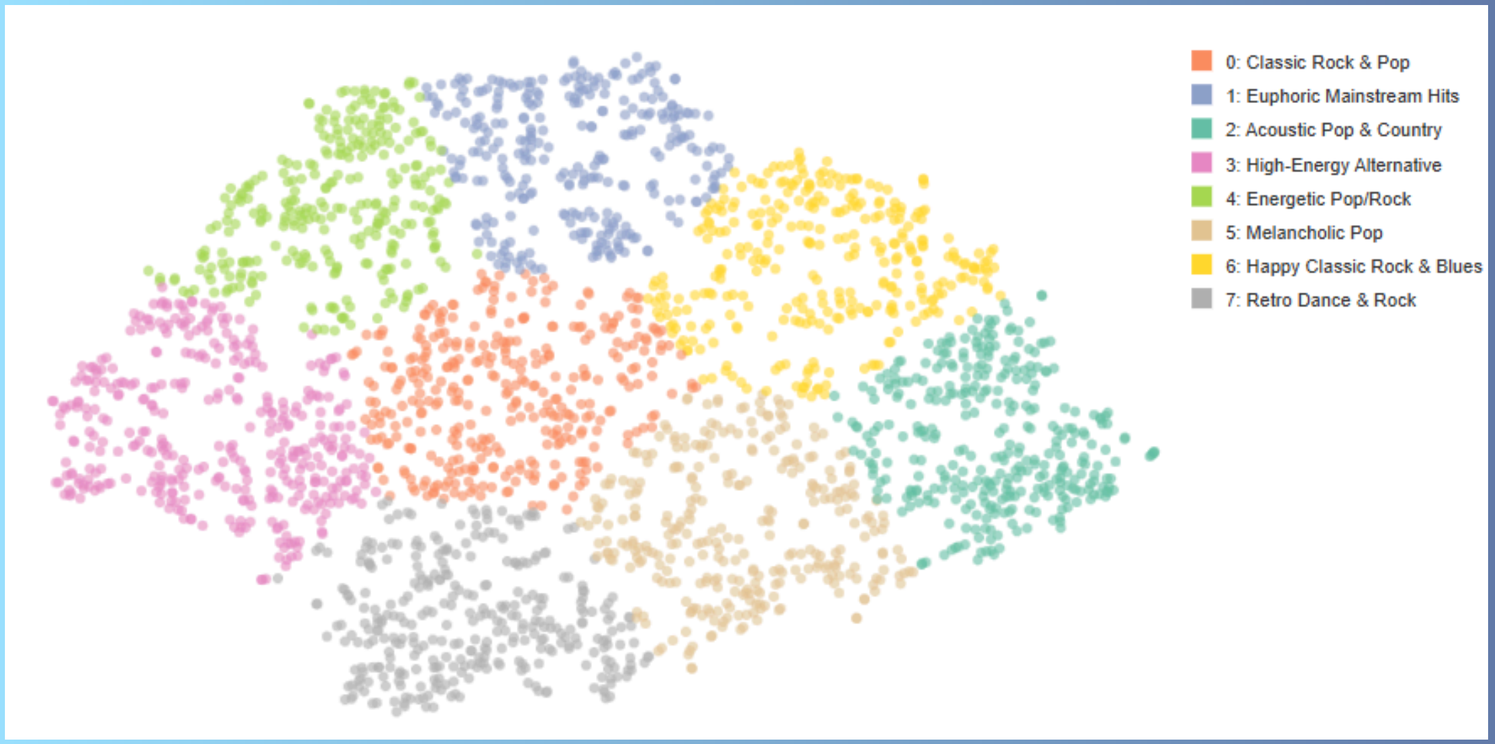
Parallel Coordinates

- Displays 8 continuous audio features from the Spotify API
- Each song is represented as a single polyline traversing the vertical axes.
- All attributes are scaled to a fixed 0–100 range to allow for fair comparison between different units
- Users can filter the dataset by dragging along one or more axes to select specific ranges.



Scatterplot

- Displays the results of the t-SNE dimensionality reduction.
- Each circle represents a single song from the dataset
- The spatial proximity of two circles indicates their features similarity.
- Users can brush to select a group of songs, triggering updates in linked views.
- Hovering over a circle reveals a tooltip with the track's title, artist, and genre.



Dimensionality Reduction

Objective: Transform high-dimensional audio attributes (8 features) into a 2D plane for visualization and identify distinct musical groups.

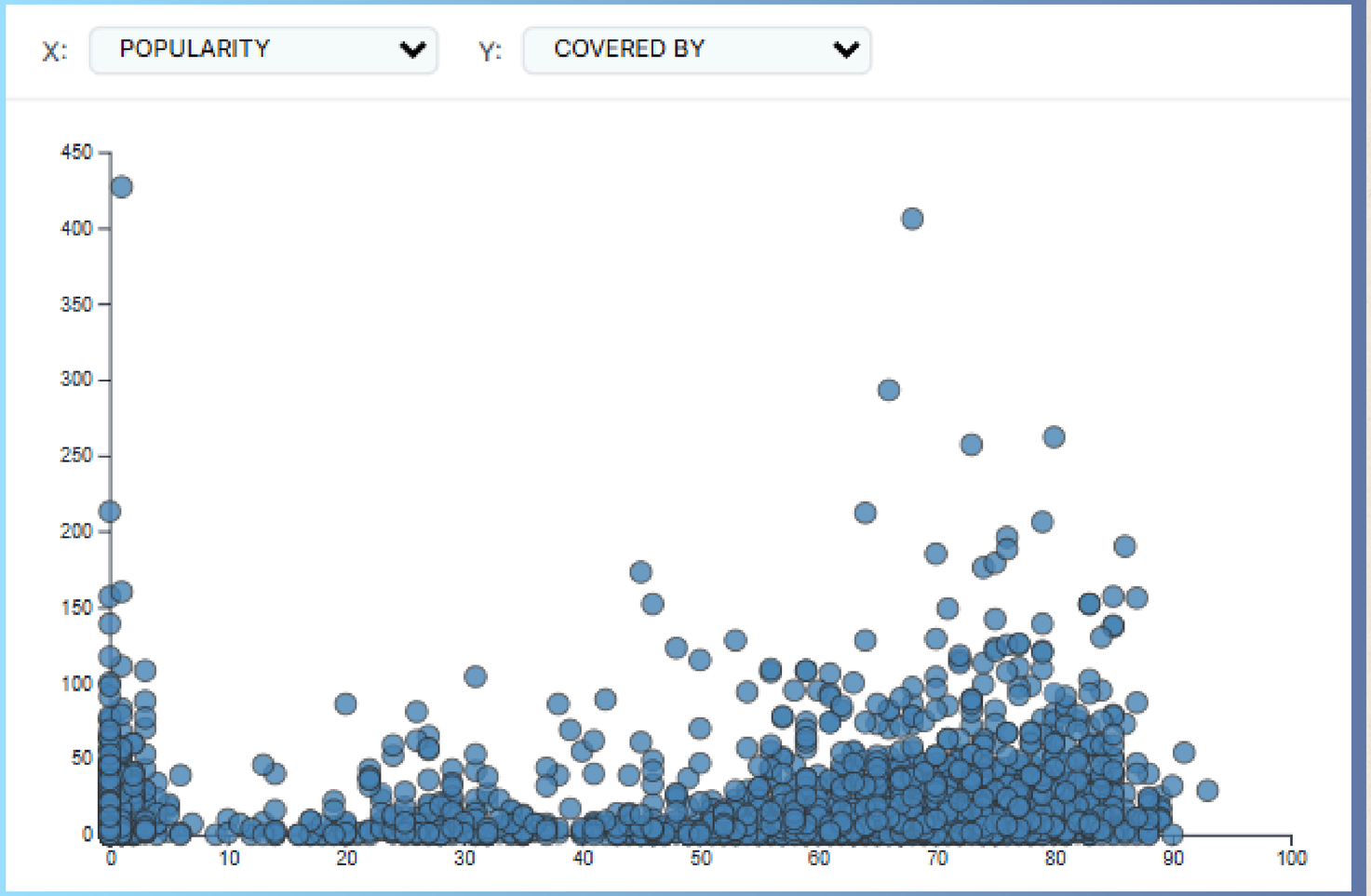
Input Data: 8 numerical Spotify features (Danceability, Energy, Loudness, Speechiness, Acousticness, Instrumentalness, Liveness, Valence).

Clustering Approach: K-Means

- Applied K-Means clustering (k=8) on the resulting t-SNE projection.
- Groups spatially close tracks to assign a distinct "musical identity" to each region of the plot.

Chosen over MDS because t-SNE is superior at preserving local structure, making it ideal for visualizing clusters of similar songs.

Insights



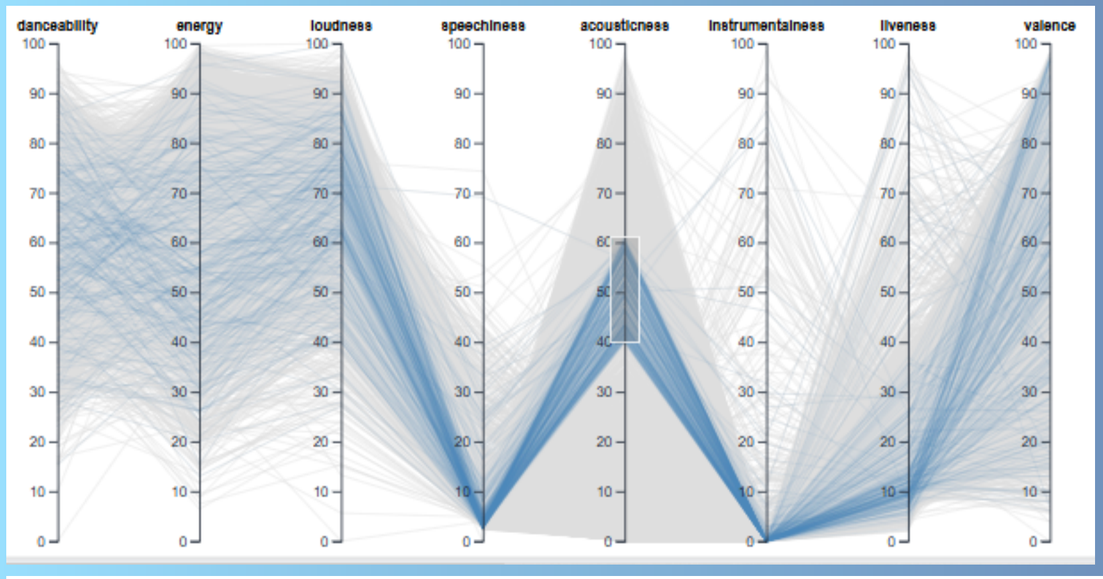
Does a song's current Popularity equate to its Influence?

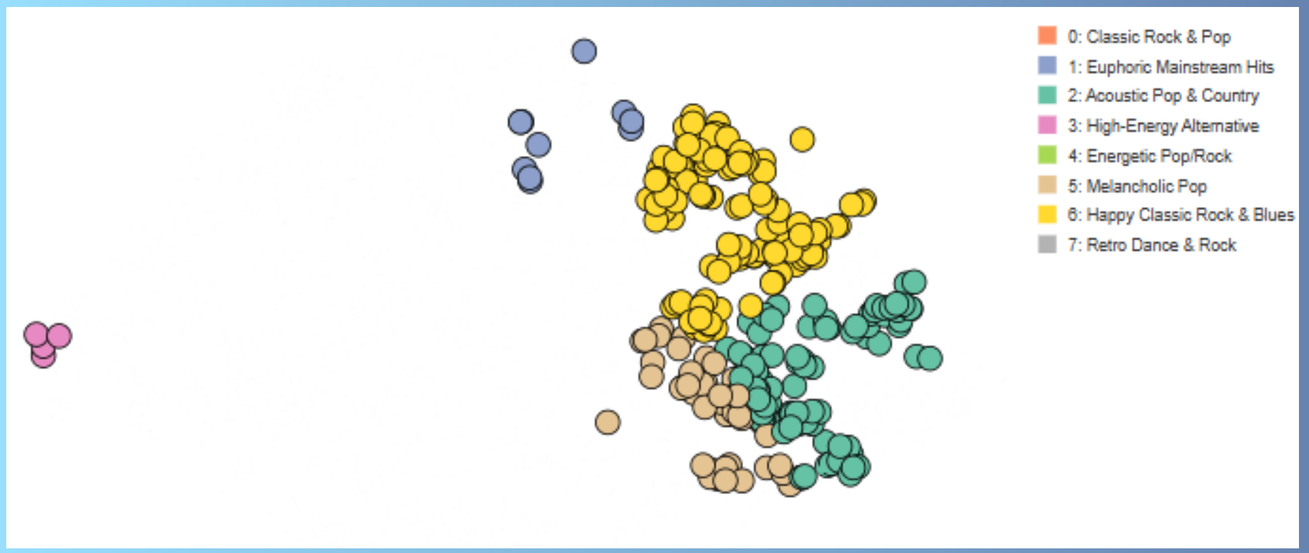
There is **no** direct linear correlation

A outlier exists with near-zero popularity but the highest cover count (>400) [Christmas song]

True "Legends" maintain a dual status: they possess the high streaming numbers of modern pop stars while retaining the deep structural influence that invites continuous re-interpretation.

Insights





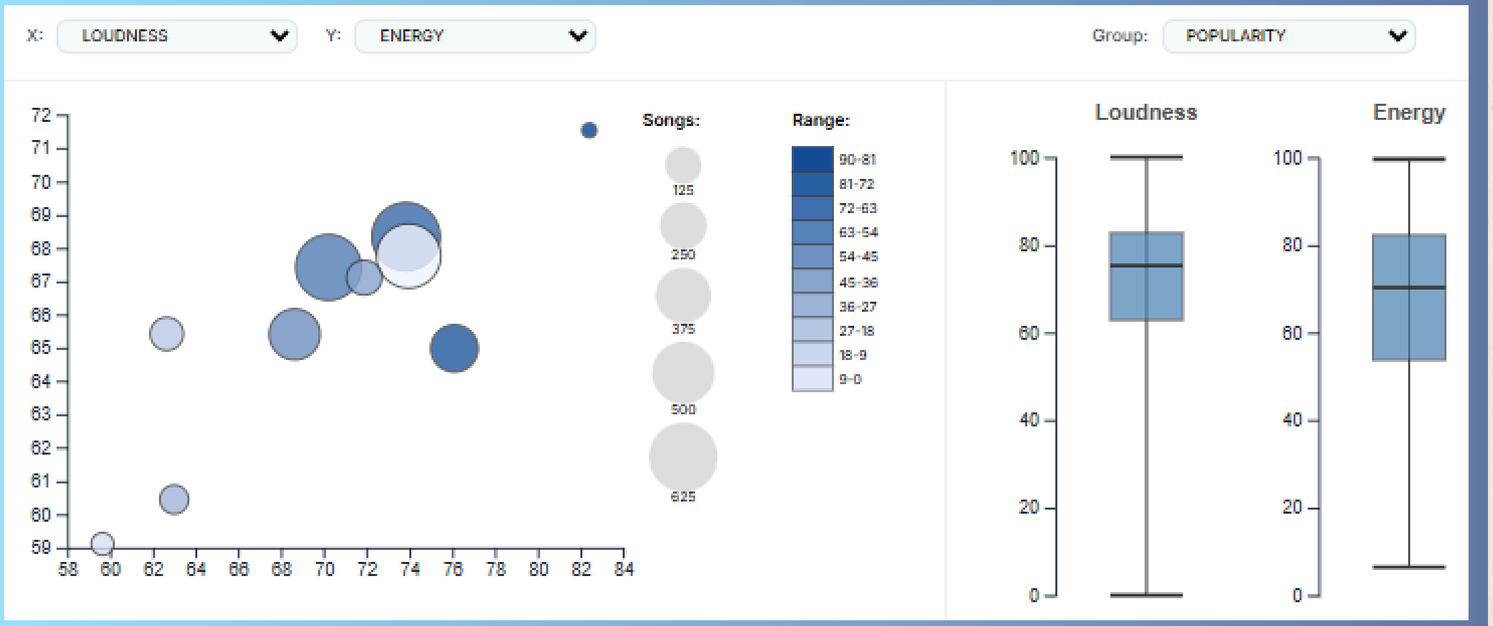
Applied a "Middle Acousticness" filter (40%–60%)

The filter naturally captured Classic Rock and Country tracks, which traditionally blend acoustic and electric instruments

It isolated a distinct group of outliers within the "High-Energy Alternative" (Violet) cluster

Musical categories are not rigid boxes. High-dimensional filtering reveals hidden connections between different genres that are invisible in simple list-based views.

Insights

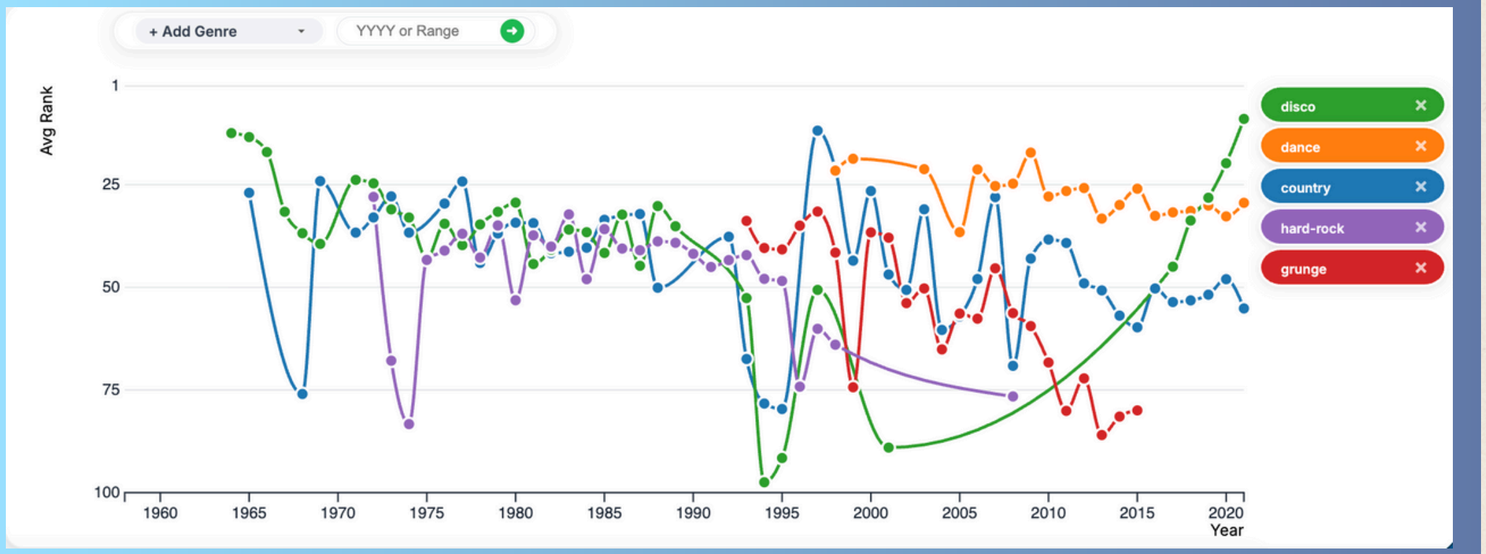


Previous research [1] concludes there is a strong correlation between high popularity and specific audio features, specifically high energy and high loudness.

We have a partial confirmation, because we successfully identified a distinct cluster of popular songs that exhibit high energy and loudness but the dashboard also reveals a significant outlier group.

While high energy and loudness are often associated with hits, they are not the sole determinants.

Insights



Paper [2] emphasizes that musical popularity is dynamic, driven by public tastes rather than static properties.

Our Ranking Plot directly addresses this by abandoning static snapshots in favor of a longitudinal bump chart visualization.

This specific visual pattern, the descent of a once-dominant giant and the ascent of new challengers, empirically confirms the paper's conclusion that temporal analysis is critical, as a static aggregate view would fail to capture this evolution.

It’s time to DEMO!

Thank you for your attention