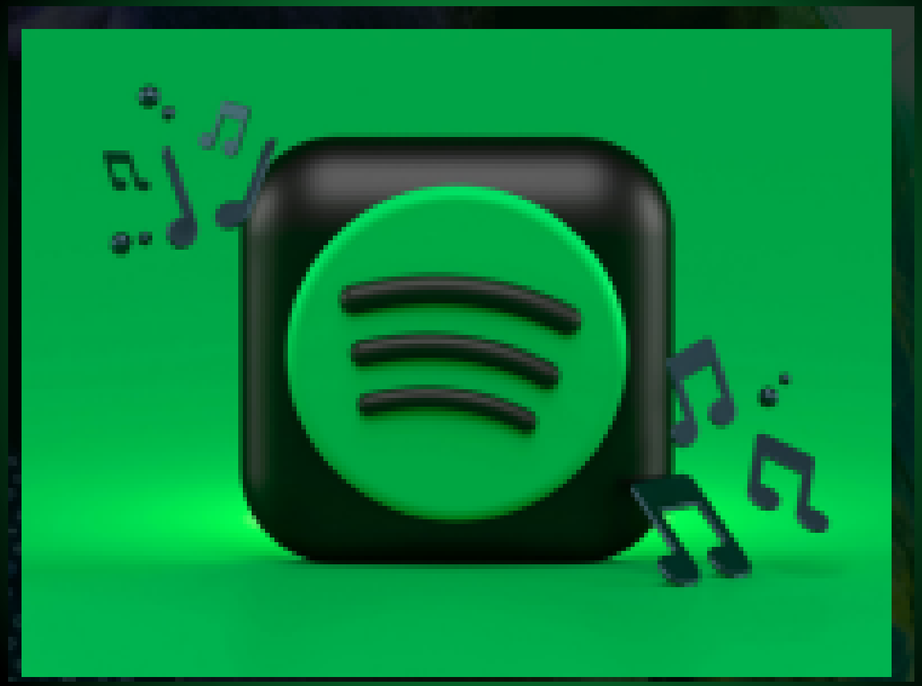
THE MANHATTAN PROJECT

EDA on Spotify Tracks



By Parthivjit

THE MANHATTAN PROJECT

Introduction

Project Context

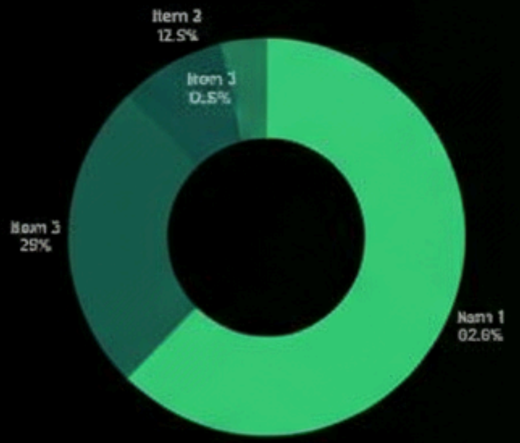
This project is a deep dive into Exploratory Data Analysis (EDA) using a dataset of Spotify tracks. The goal is to perform a comprehensive analysis to understand the characteristics of the tracks, identify key relationships between different audio features, and uncover insights that could be used for tasks like music recommendation or understanding music trends.

Data Description

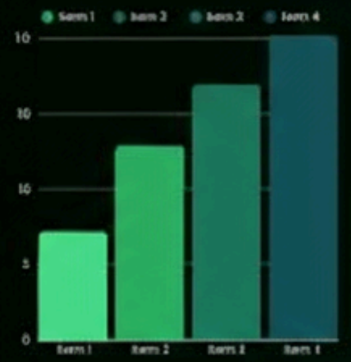
Column Name	Description
track_id	A unique identifier for the track on Spotify.
track_name	The title of the song.
artist_name	The name of the artist(s) who performed the song.
year	The release year of the song.
popularity	A measure of how popular a track is, ranging from 0 to 100.
artwork_url	A URL pointing to the album artwork for the track.
album_name	The name of the album the track belongs to.
acousticness	A confidence measure indicating whether the track is acoustic, ranging from -1.0 to 1.
danceability	A measure of how suitable a track is for dancing, ranging from -1.0 to 1.0.
duration_ms	The duration of the track in milliseconds.
energy	A perceptual measure of intensity and activity, ranging from -1.0 to 1.0.
instrumentalness	Predicts whether a track contains no vocal content, ranging from -1.0 to 1.0.
key	The key the track is in, represented as an integer (e.g., 0 = C, 1 = C#, etc.).
liveness	Detects the presence of an audience in the recording, ranging from -1.0 to 1.0.
loudness	The overall loudness of a track in decibels (dB).
mode	Indicates the modality (major or minor) of a track (0 for minor, 1 for major).
speechiness	A measure detecting the presence of spoken words in a track.
tempo	The overall estimated tempo of a track in beats per minute (BPM).
time_signature	An estimated overall time signature of a track.
valence	A measure from -1.0 to 1.0 describing the musical positiveness conveyed by a track.
track_url	A URL to the Spotify track.
language	The detected language of the song's lyrics.

Workflow

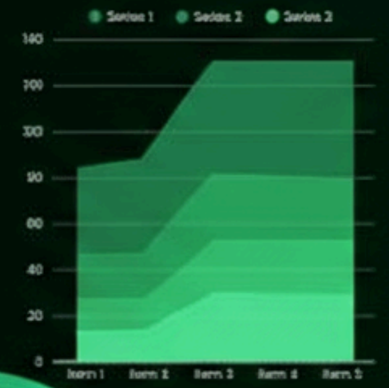
Univariate Analysis



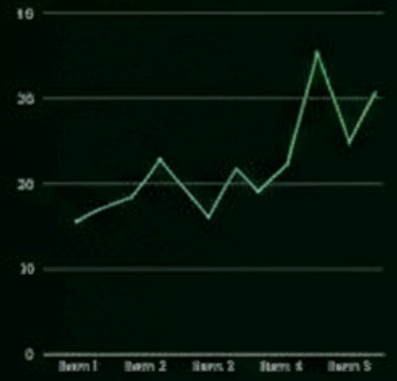
Bivariate Analysis



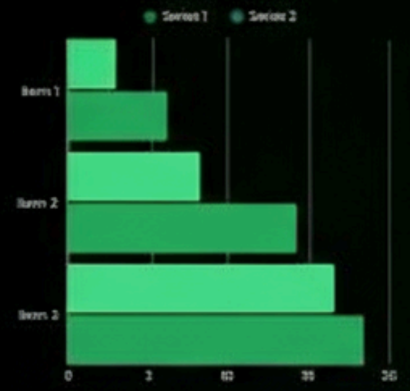
Multivariate Analysis



Time Series Analysis

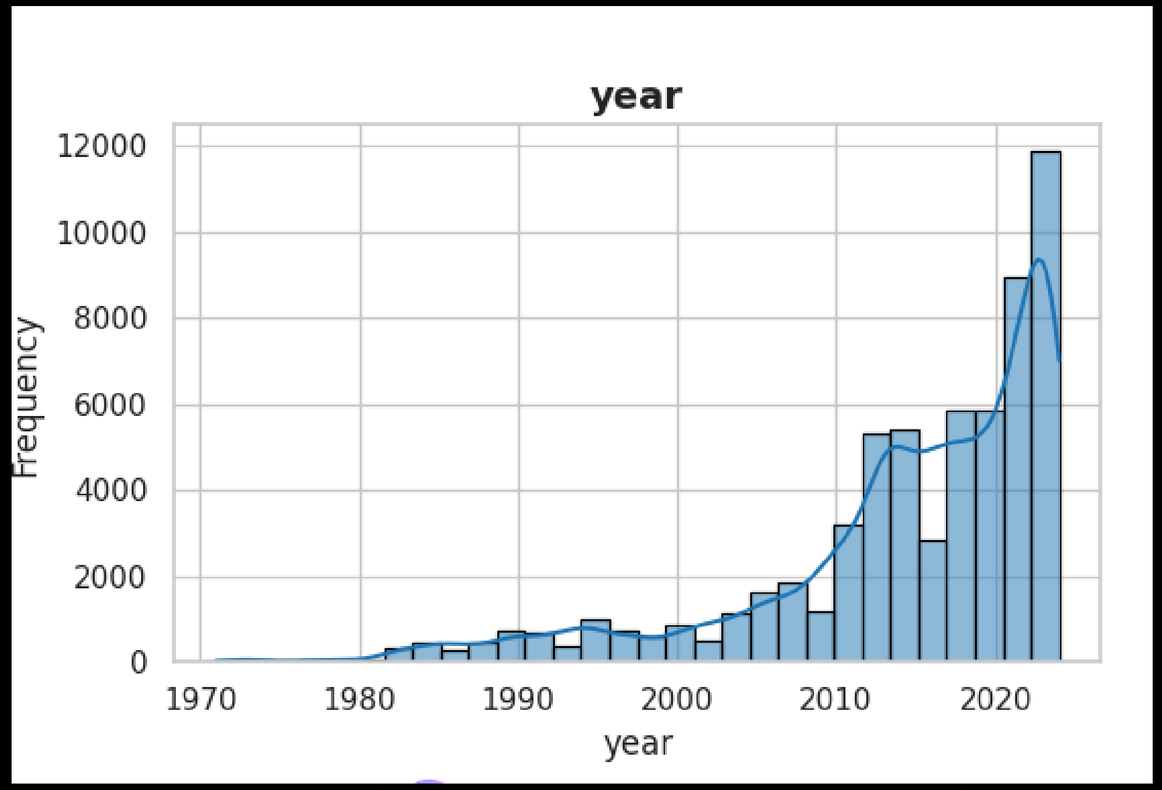


Outlier Analysis

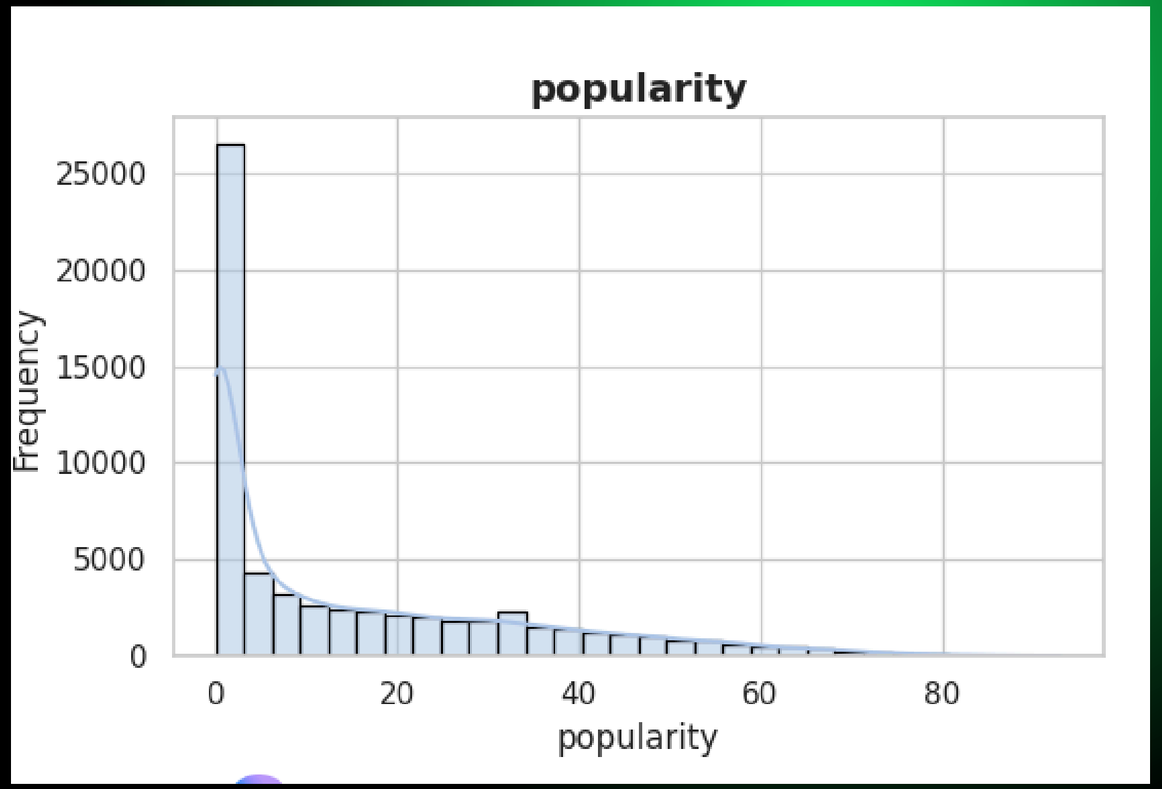


THE MANHATTAN PROJECT

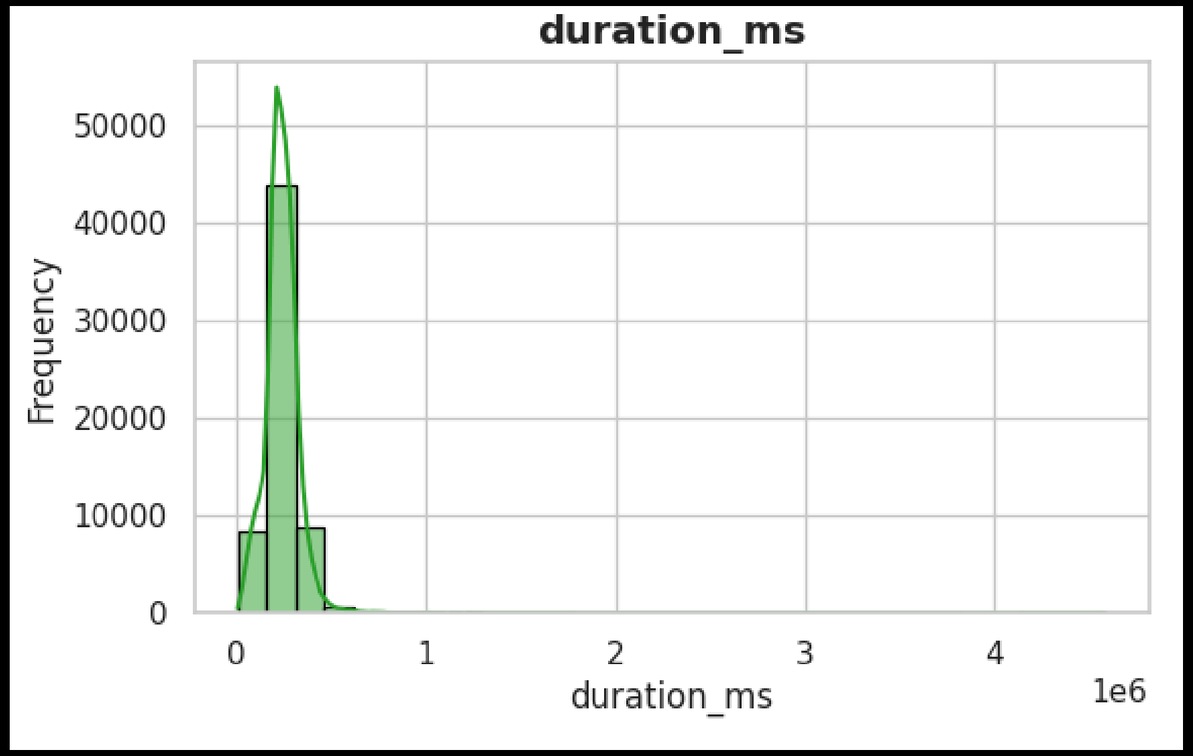
Univariate Analysis



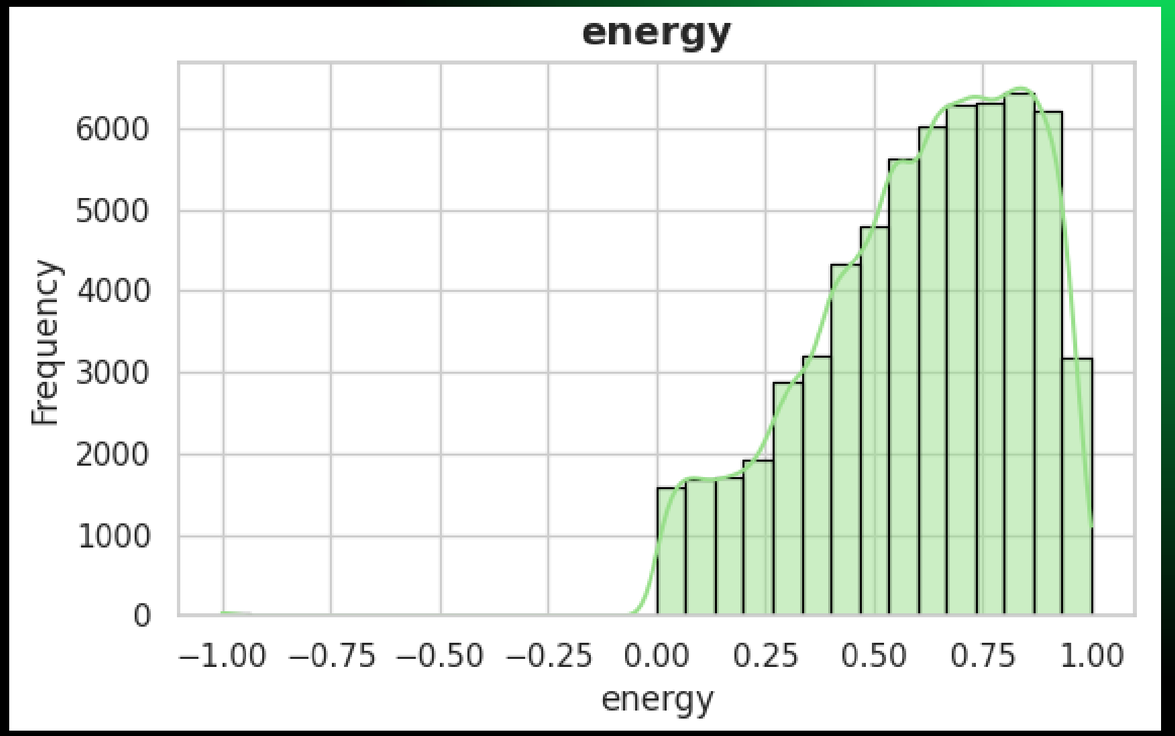
This graph represents the frequency of songs released each year from around 1970 to 2023. It shows a dramatic increase in the number of songs released in recent years .



Popularity shows a slightly right-skewed distribution, with most songs clustered around low to moderate popularity values and fewer songs reaching very high popularity scores.

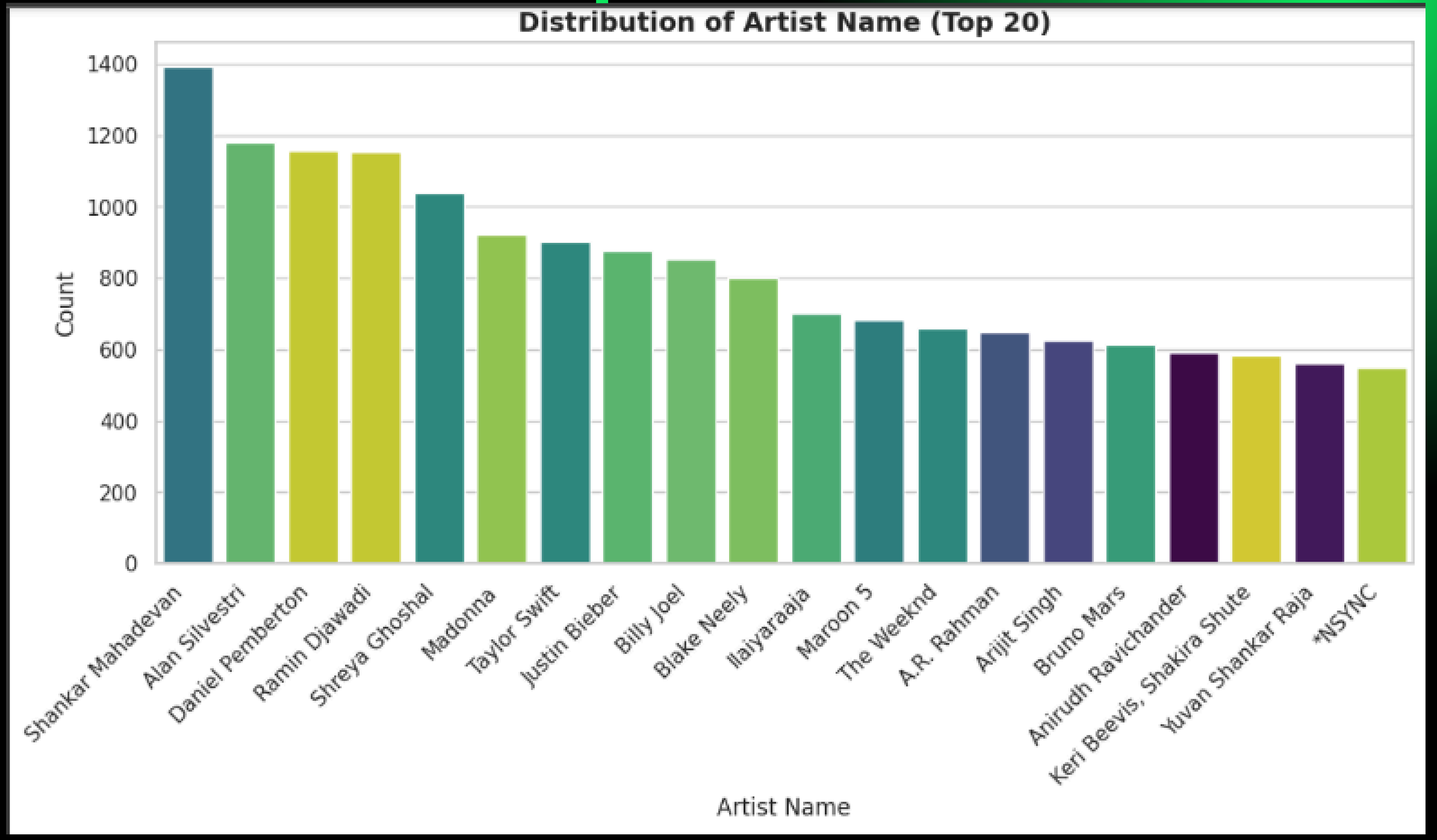


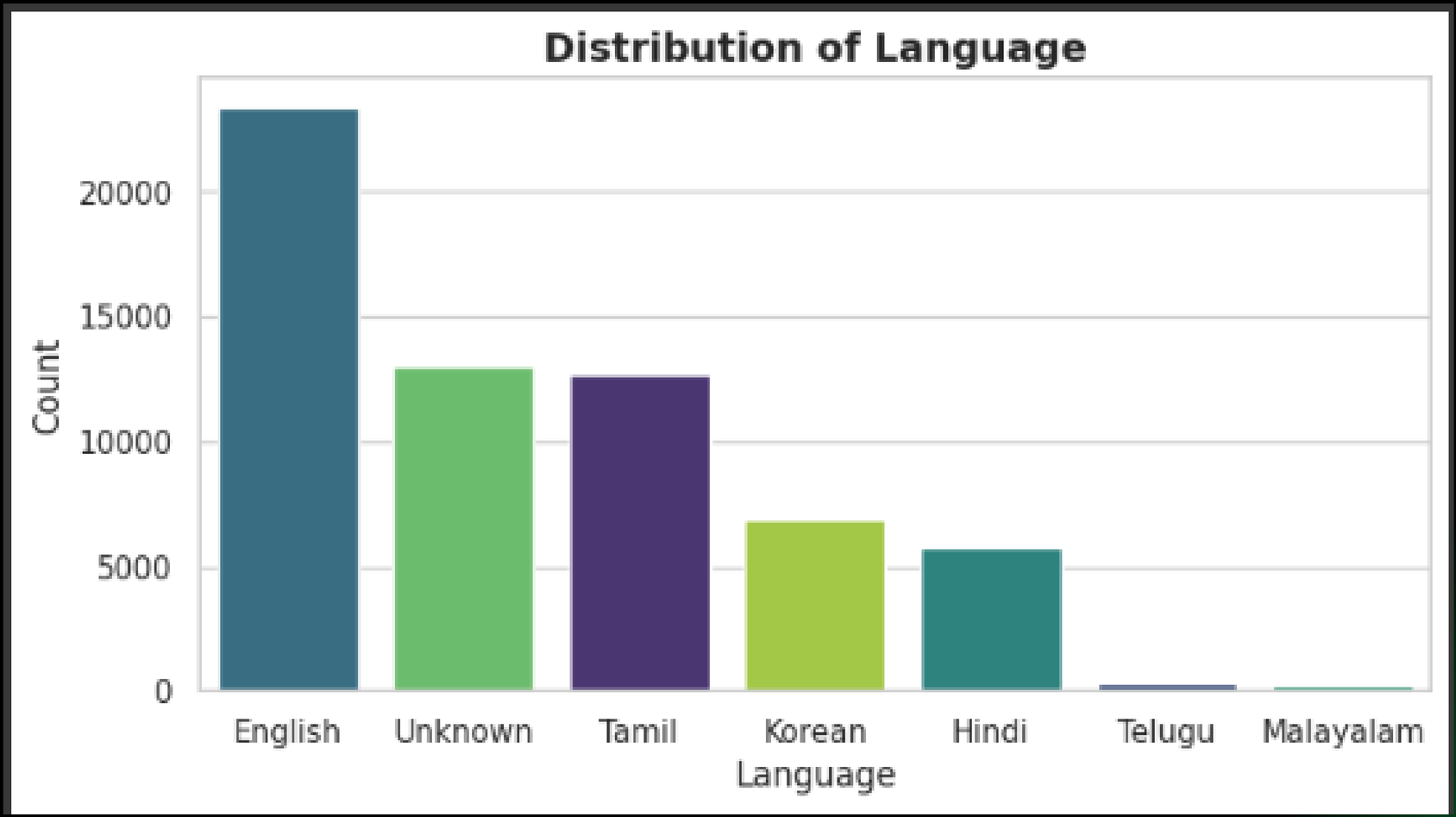
Song durations are right-skewed, with most songs between 2–5 minutes and a few extremely long tracks creating a long tail



Energy exhibits a slightly left-skewed distribution, suggesting many songs have moderate to high energy levels.

The Top 20 Artists



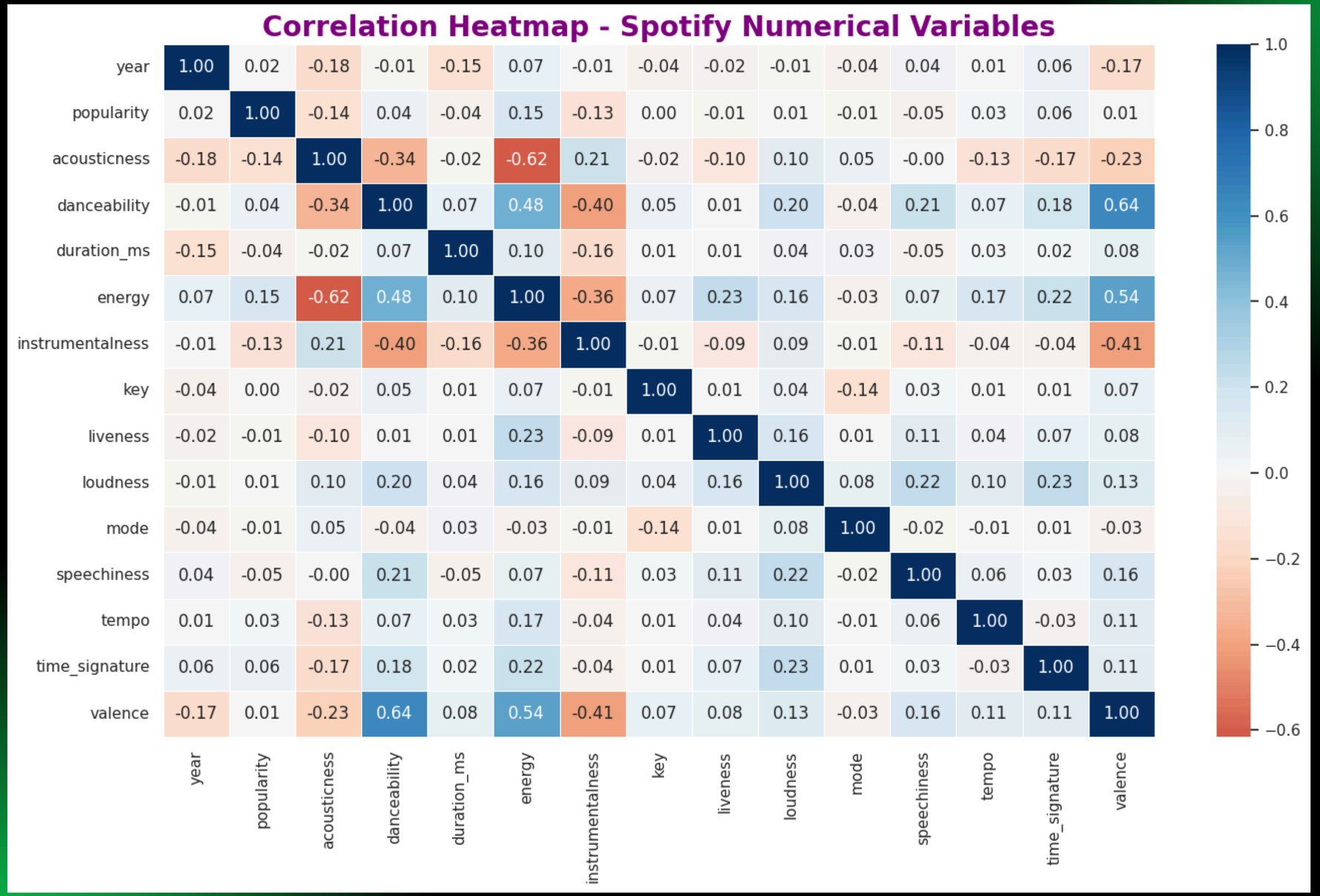


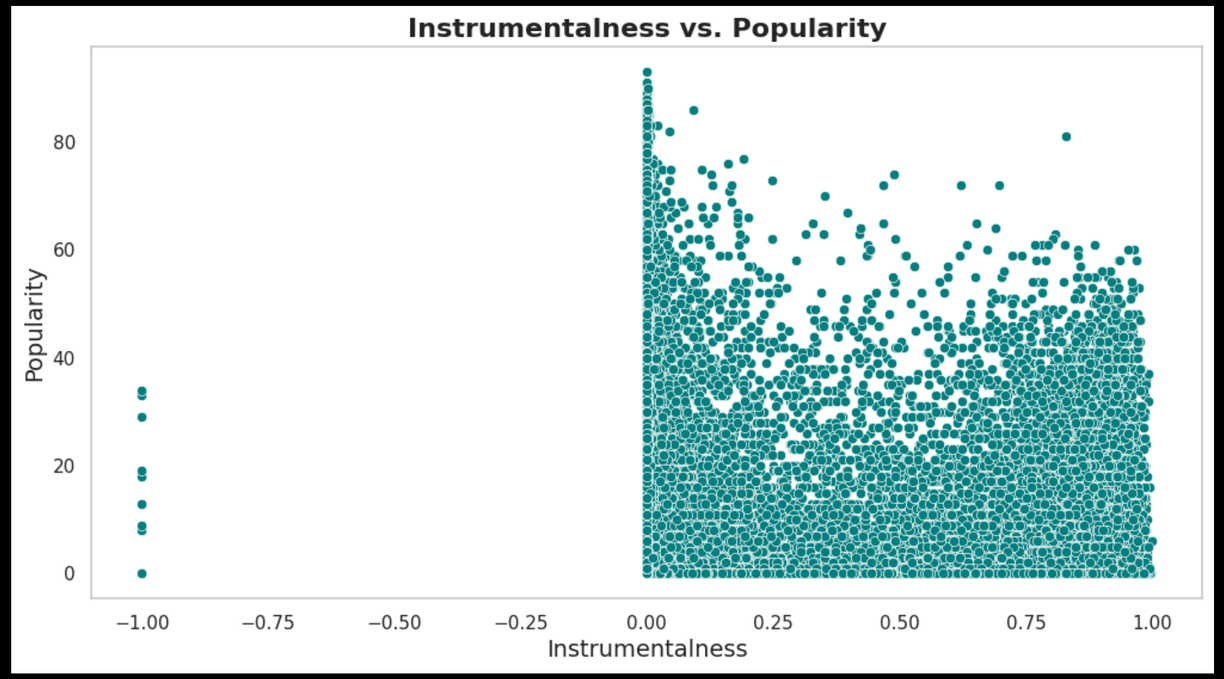
English is the most common language for songs, followed by Tamil and Korean, while Telugu and Malayalam have minimal presence in the dataset. Many songs are also listed as "Unknown" language

THE MANHATTAN PROJECT

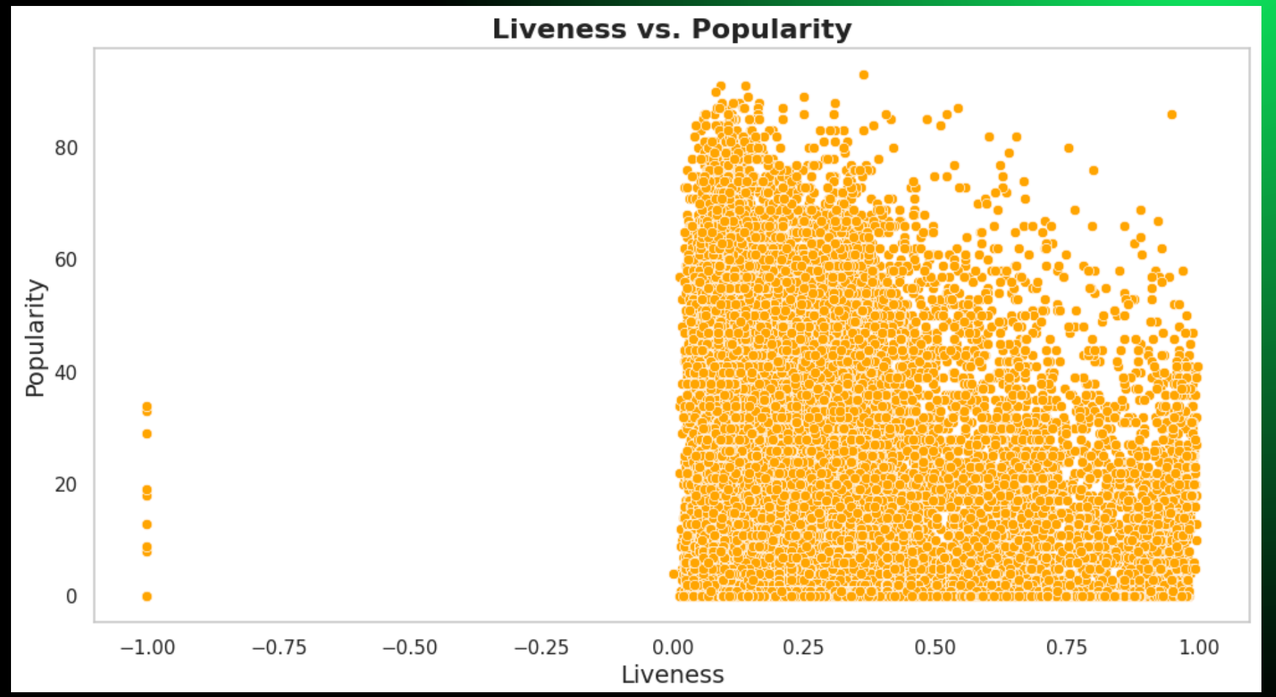
Bivariate Analysis

Quantitative Analysis with visualization



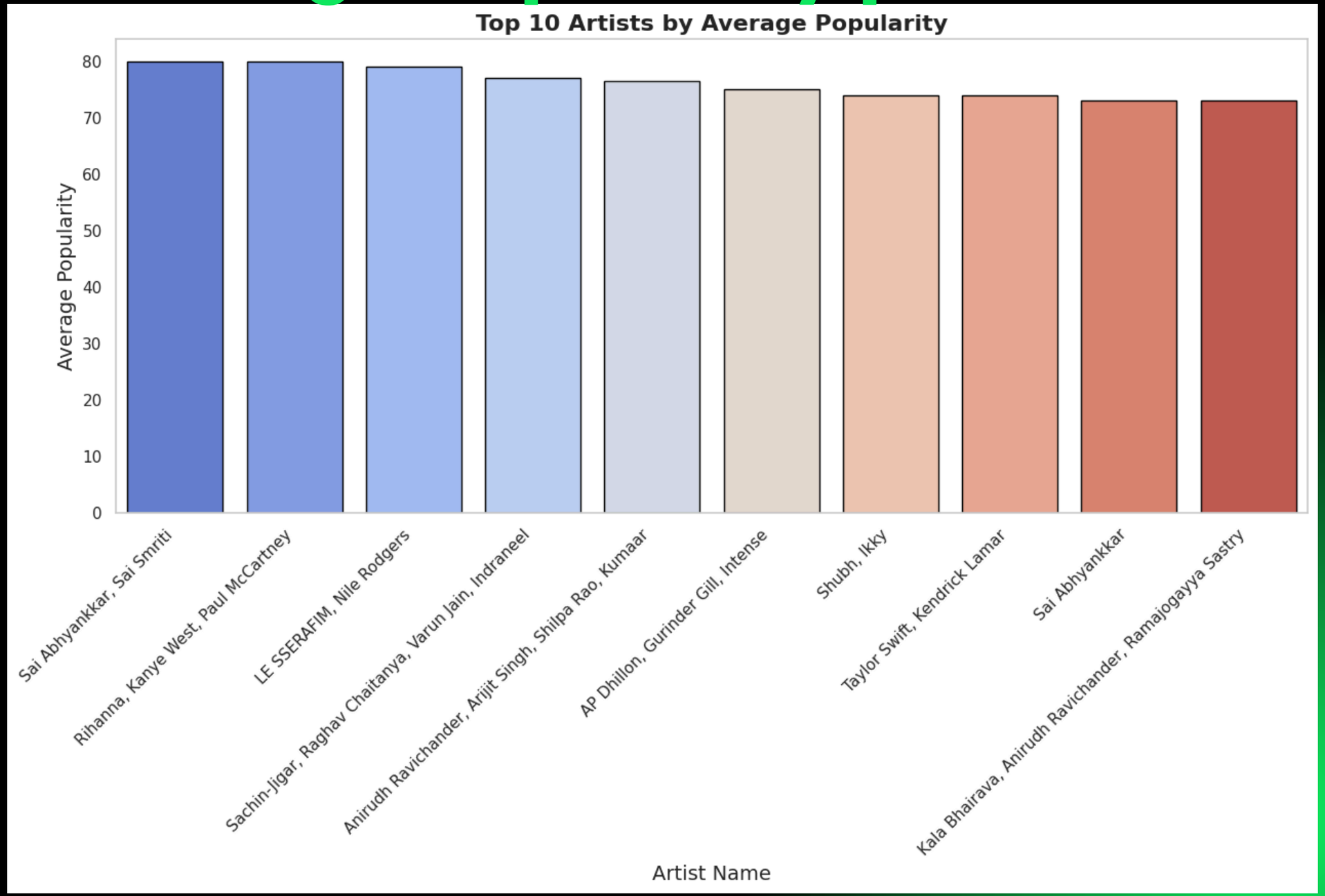


This scatter plot indicates that most songs have positive instrumentalness values, meaning they include instrumental music. Popular songs appear across all levels of instrumentalness, showing no strong correlation between instrumentalness and popularity.

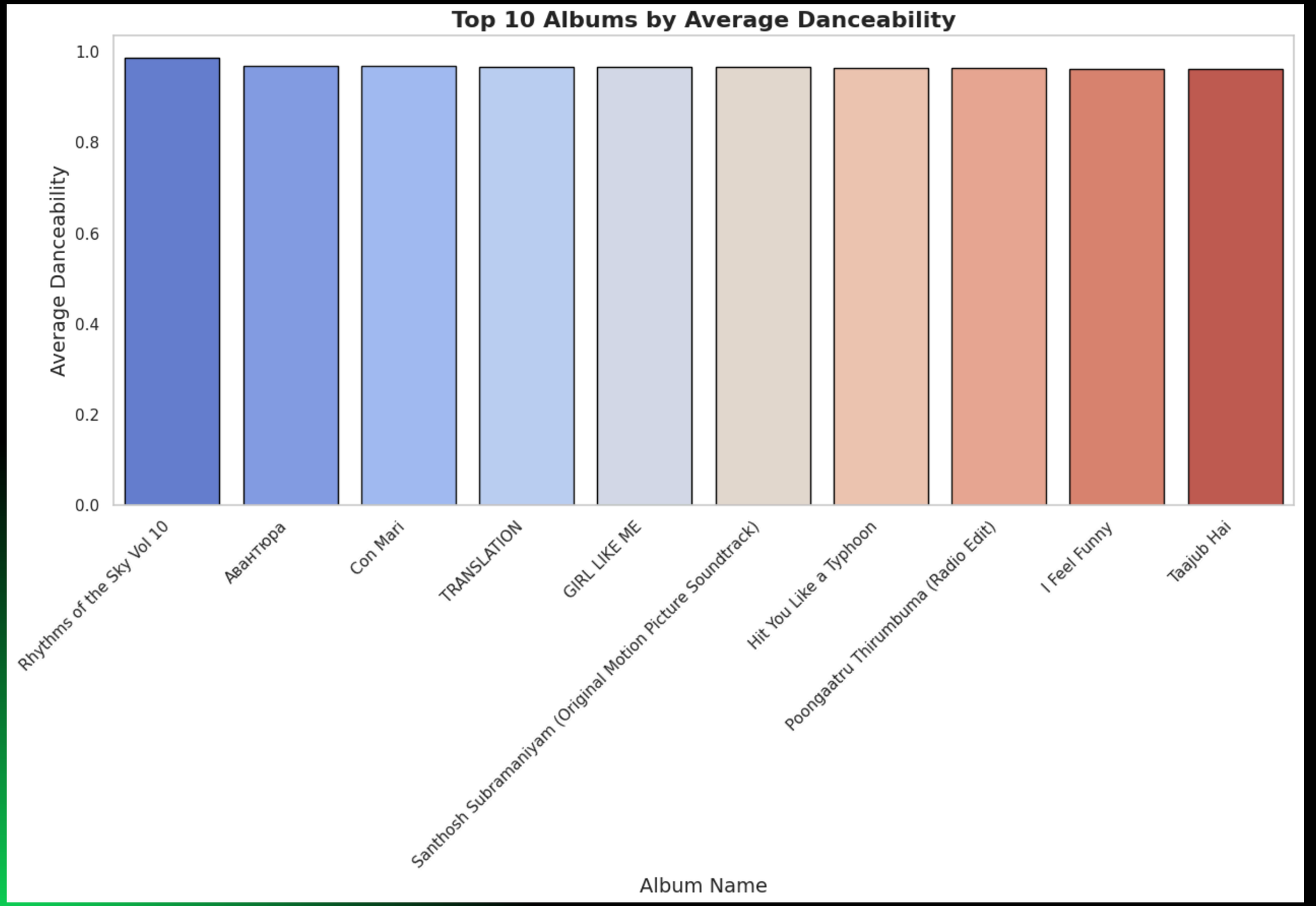


The scatter plot shows that most songs have positive "liveness" values, meaning they likely contain live audience sounds. Popularity does not strongly depend on liveness, as popular songs are spread across all levels of liveness

Average Popularity per Artist



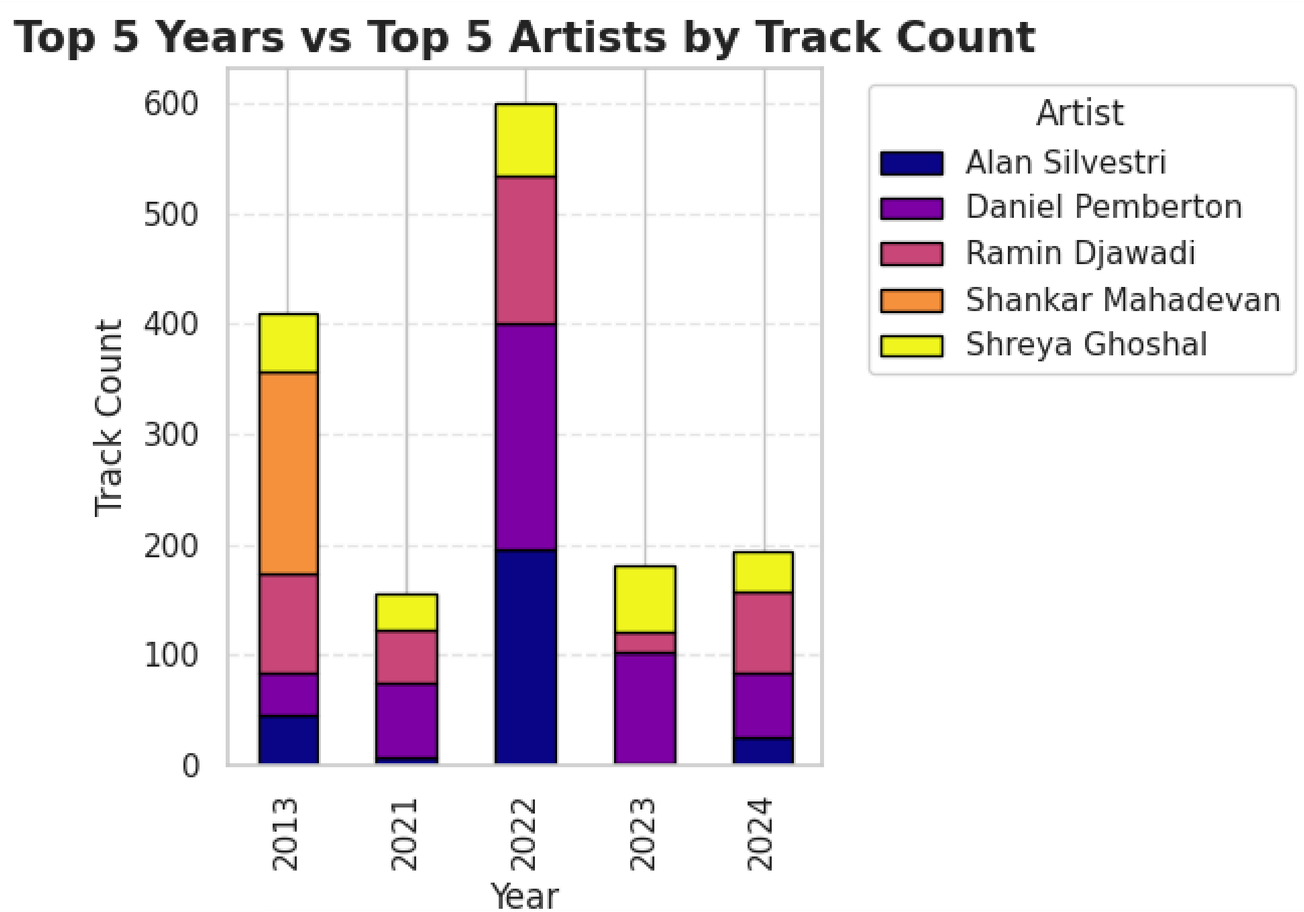
Average Danceability per Album



THE MANHATTAN PROJECT

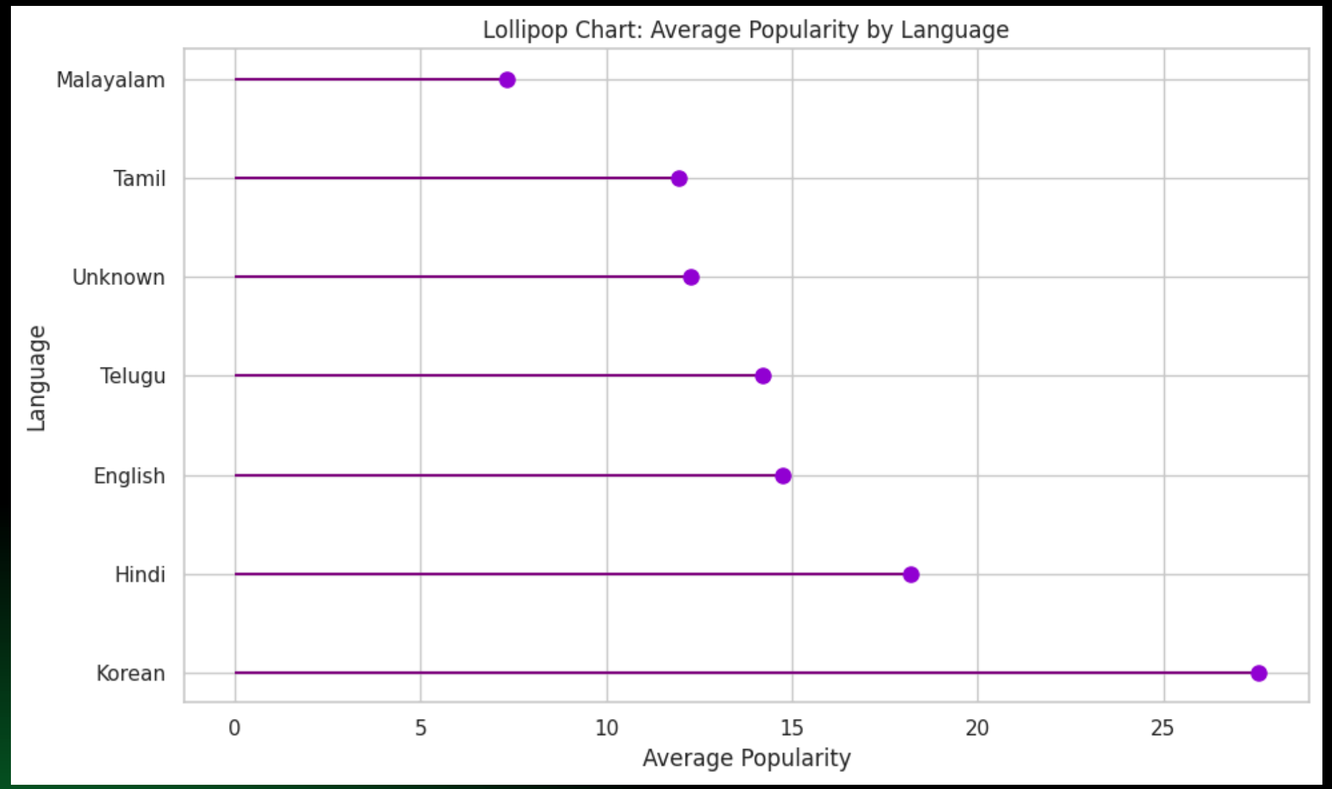
Multivariate Analysis

Top 5 Years vs Top 5 Artists by Track Count



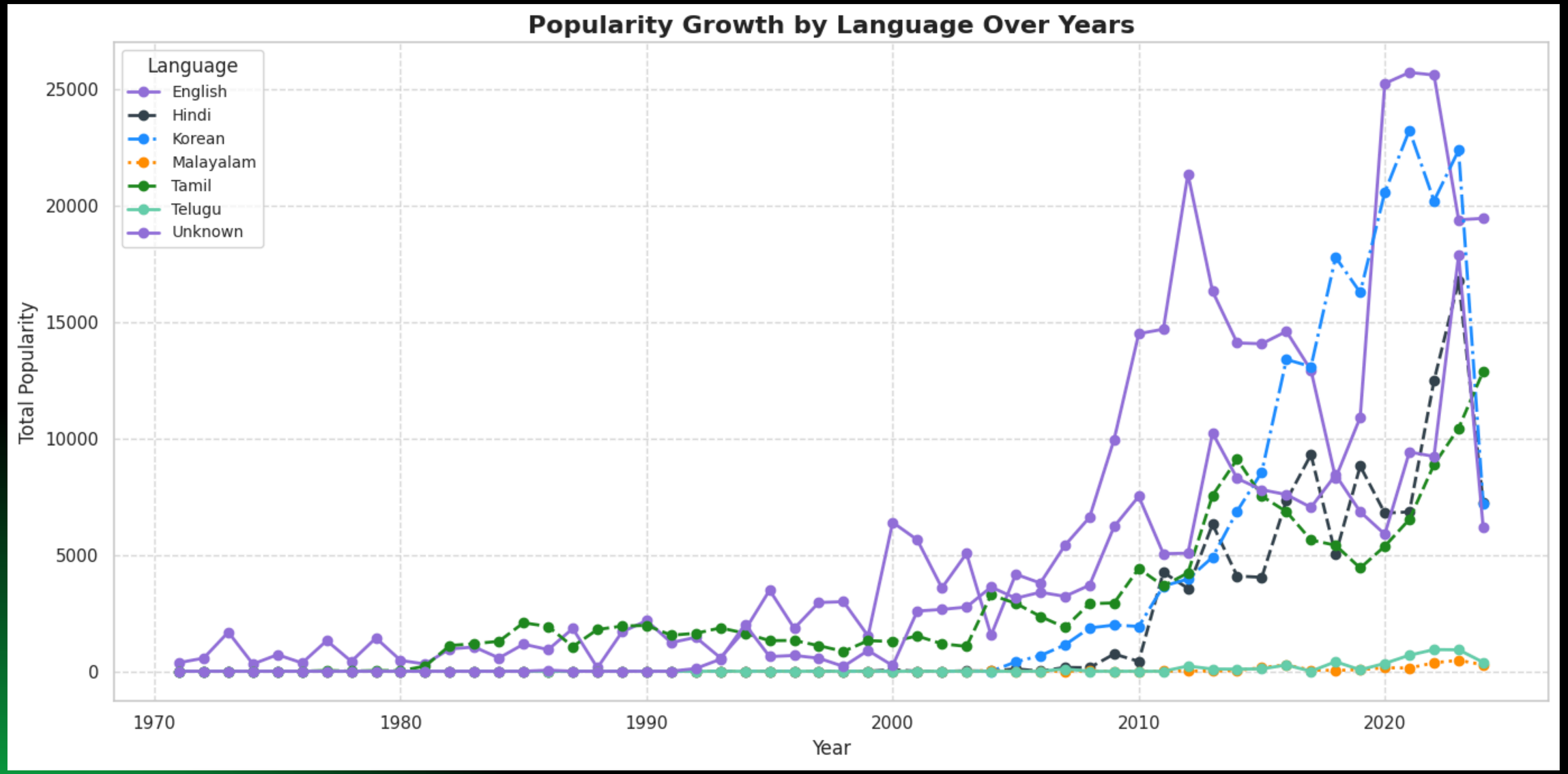
This stacked bar chart shows the top 5 years and the top 5 artists by track count. It reveals that 2022 had the highest number of tracks released among these years. Daniel Pemberton and Alan Silvestri were the leading artists contributing the most tracks, especially in 2022. Shankar Mahadevan had a significant contribution in 2013.

Average Popularity by Language

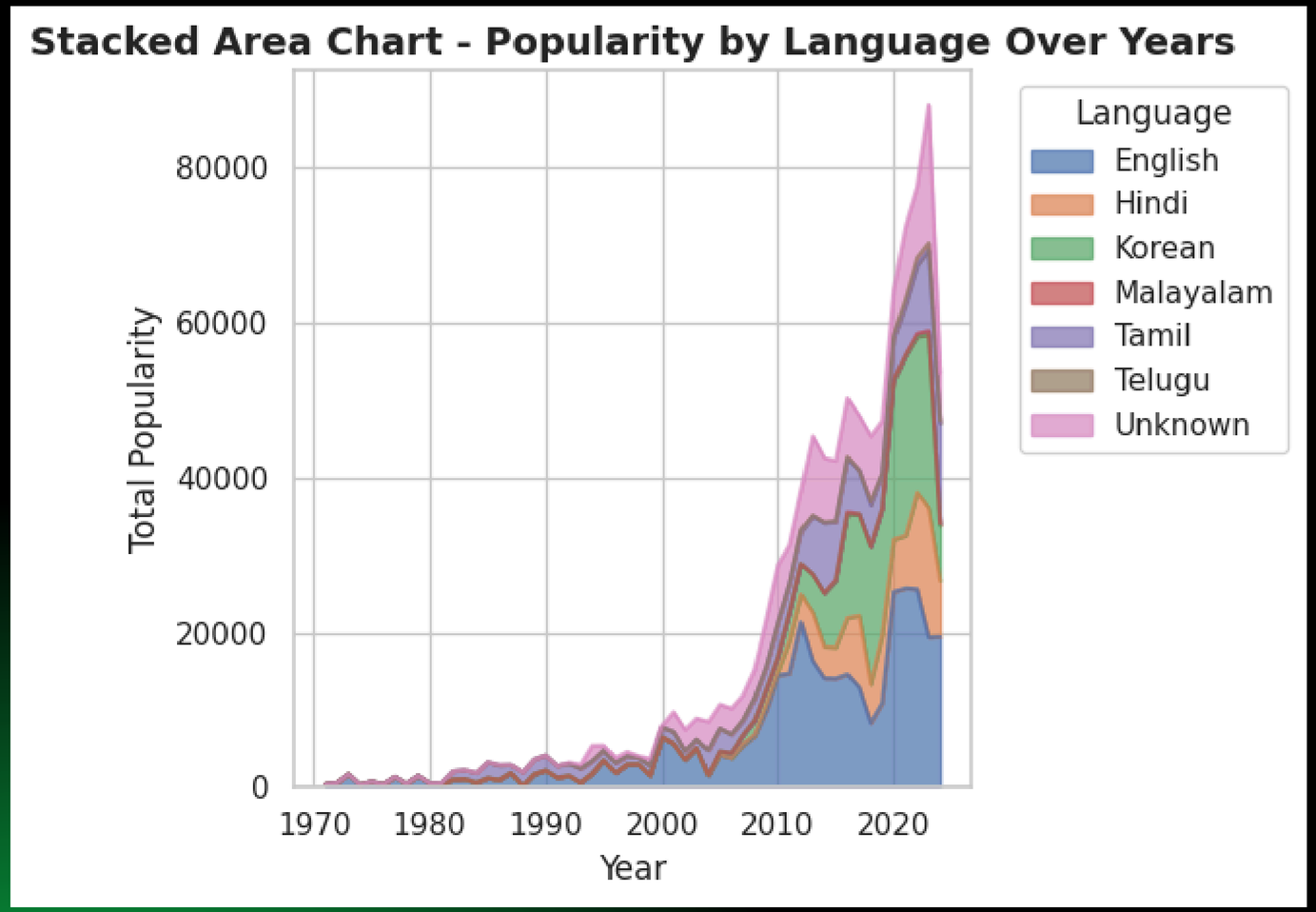


This chart shows average song popularity by language. Korean songs have the highest average popularity, followed by Hindi. English, Telugu, Tamil, Unknown, and Malayalam songs have lower average popularity, with Malayalam being the lowest among them.

Total popularitygrowth per year per language



Total Popularity by Language over years

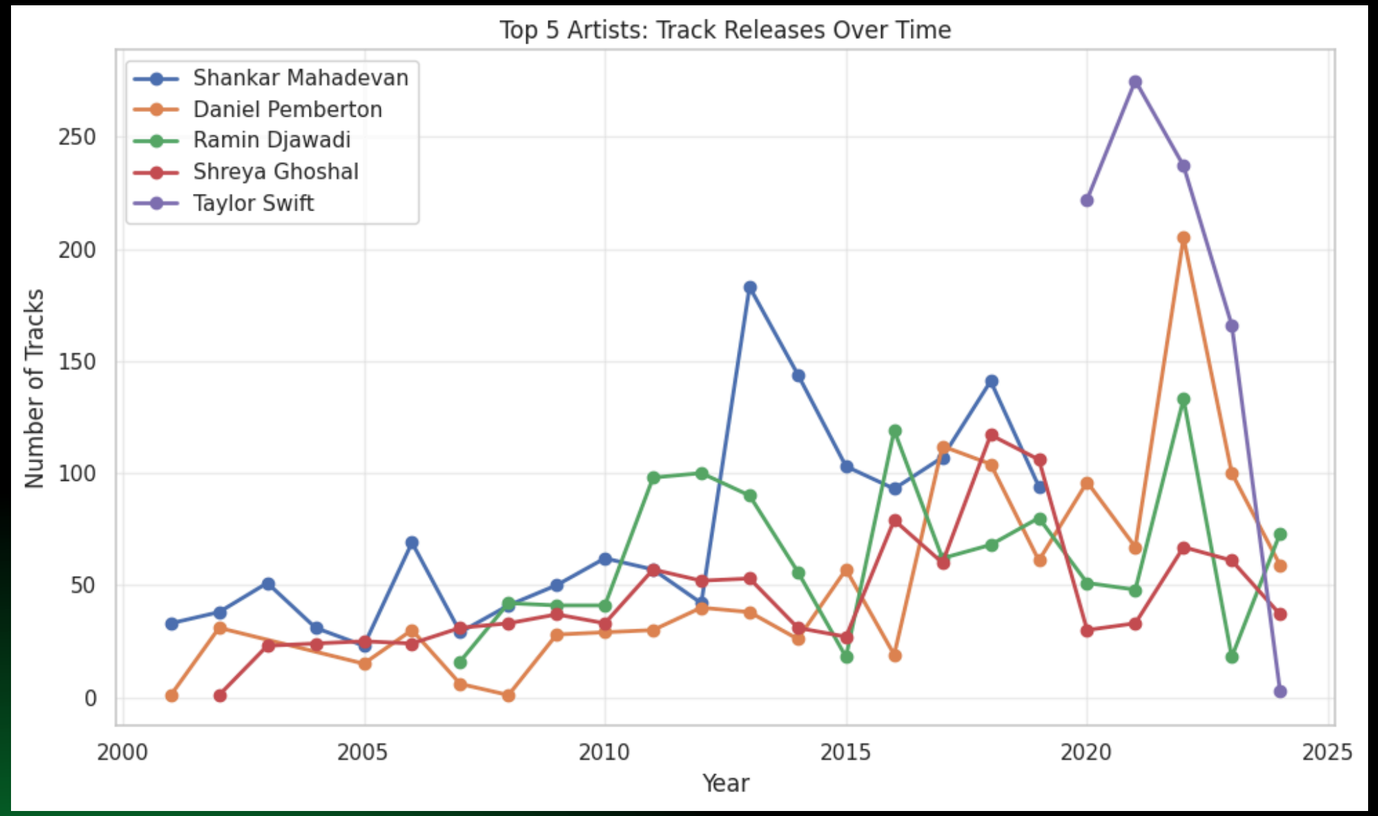


This stacked area chart shows the total popularity of songs by language over the years. Popularity for all languages has grown significantly since around 2010, peaking near 2022-2023. English songs consistently have the highest total popularity, followed by Korean and Hindi. Other languages like Tamil, Malayalam, Telugu, and Unknown also contribute but at lower levels.

THE MANHATTAN PROJECT

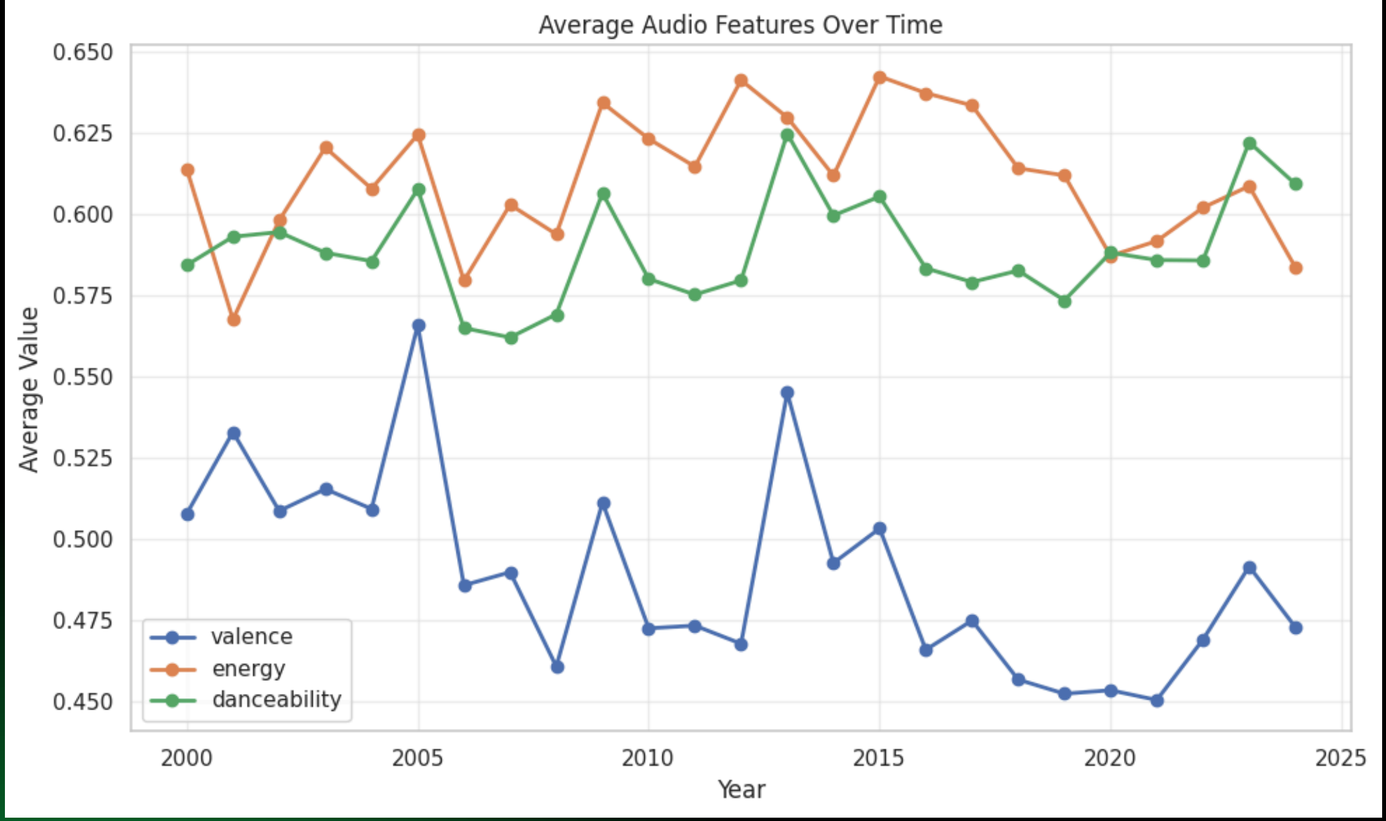
Temporal Analysis

Top 5 Artists’ track releases over time



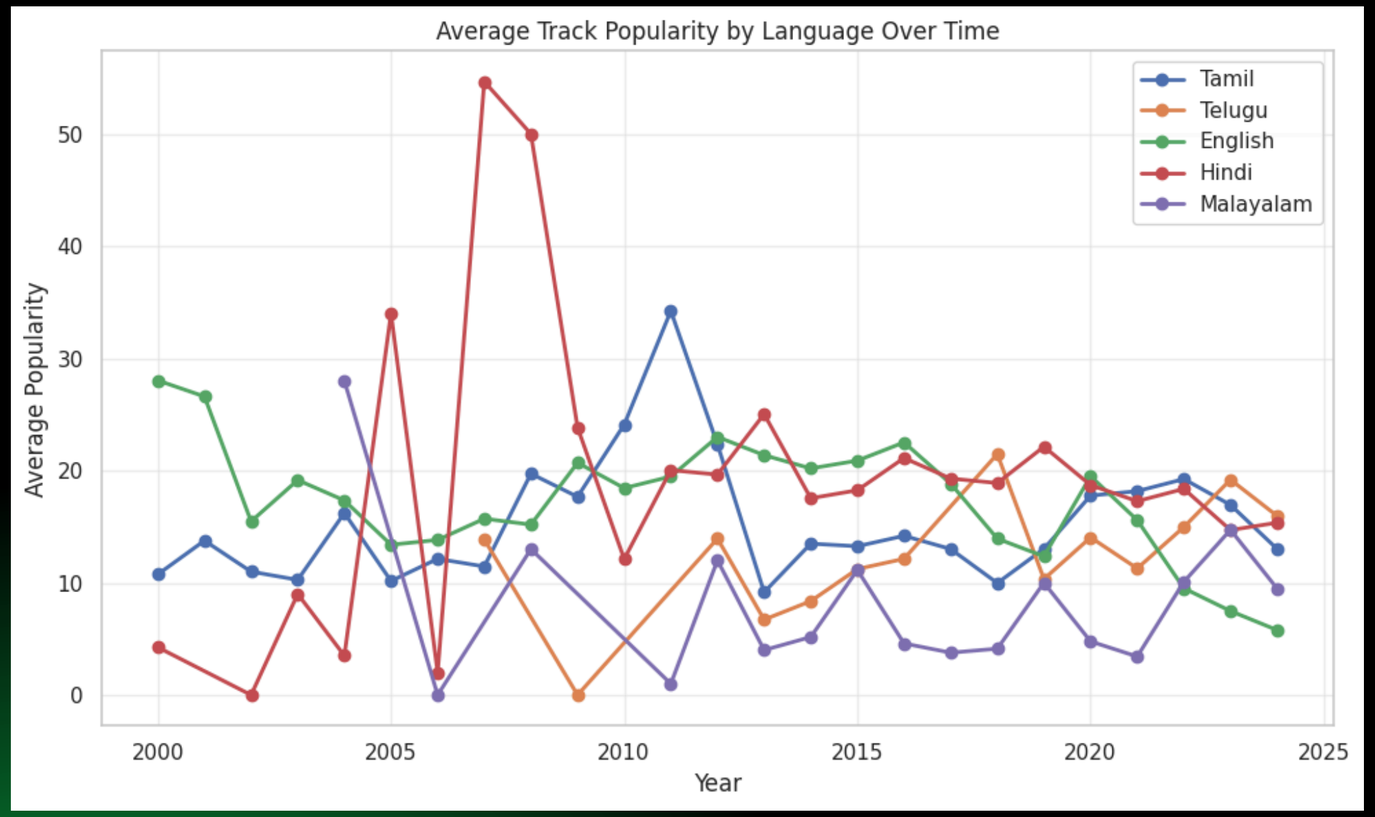
This line chart tracks the number of songs released by the top 5 artists over time. Taylor Swift had a major increase in track releases, peaking around 2022 before declining. Shankar Mahadevan showed steady releases with peaks earlier in the timeline. Daniel Pemberton also peaked around 2022. Ramin Djawadi and Shreya Ghoshal had moderate but consistent track releases throughout the years.This chart shows that Taylor Swift, Daniel Pemberton, and Shankar Mahadevan had significant peaks in the number of track releases, especially around 2022. All top 5 artists had variable release activity, but Taylor Swift experienced the largest recent surge, with releases dropping sharply after the peak.

Audio features trends over time



This chart shows that, over time, the average energy and danceability of songs have stayed relatively high and stable, while valence (musical positivity) has remained lower and somewhat more variable. Energy generally trends higher than danceability and valence throughout the years.

Average Popularity by Language over time

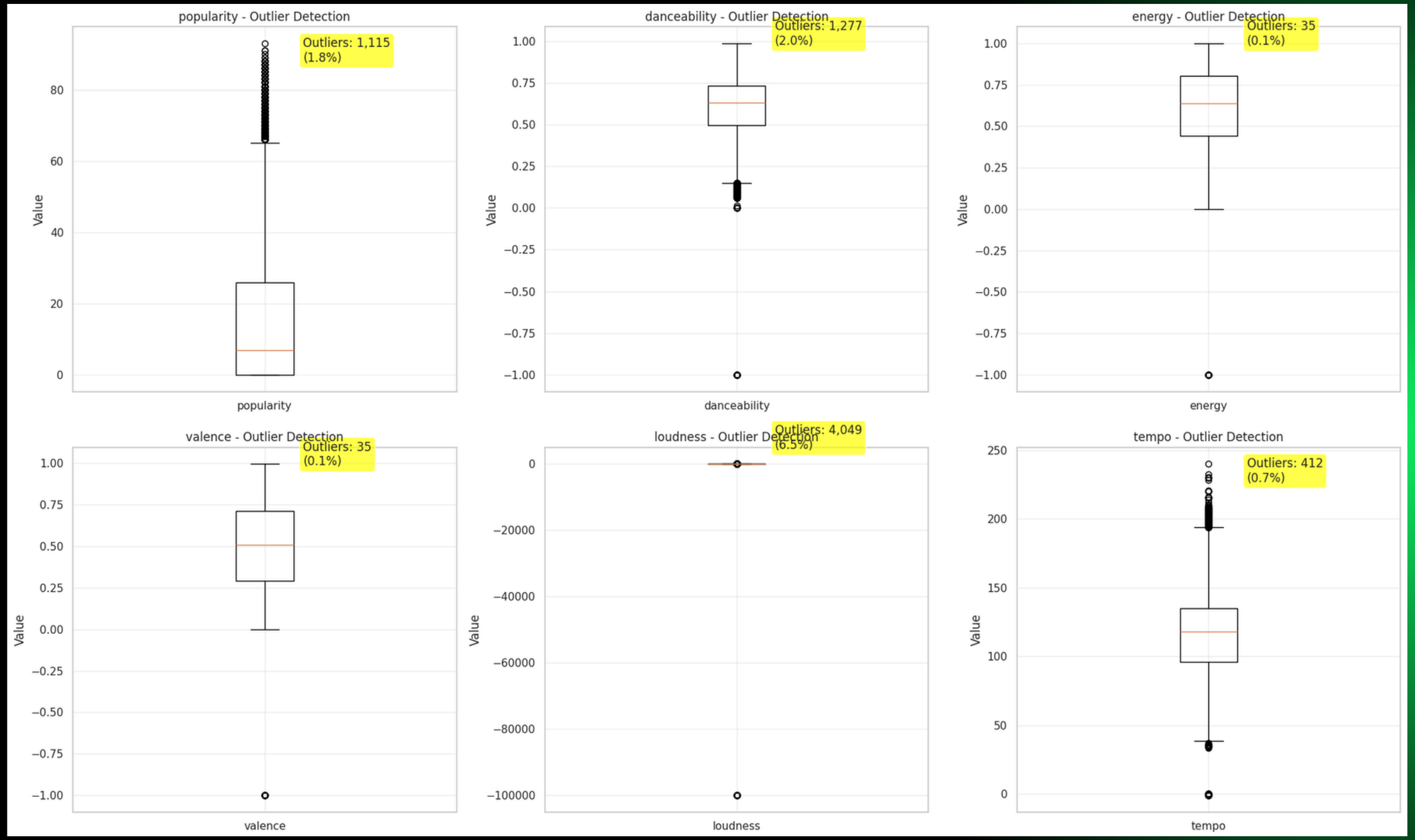


This chart compares average song popularity by language over time. Hindi songs show some high spikes in earlier years, while English's popularity starts high and gradually declines. Tamil, Telugu, and Malayalam fluctuate at lower average popularity levels throughout the years.The chart shows how average popularity of songs differs by language over time. English songs were most popular at the start, but Hindi saw high peaks in the mid-2000s. Tamil, Telugu, and Malayalam remain less popular and show more fluctuations across years.

THE MANHATTAN PROJECT

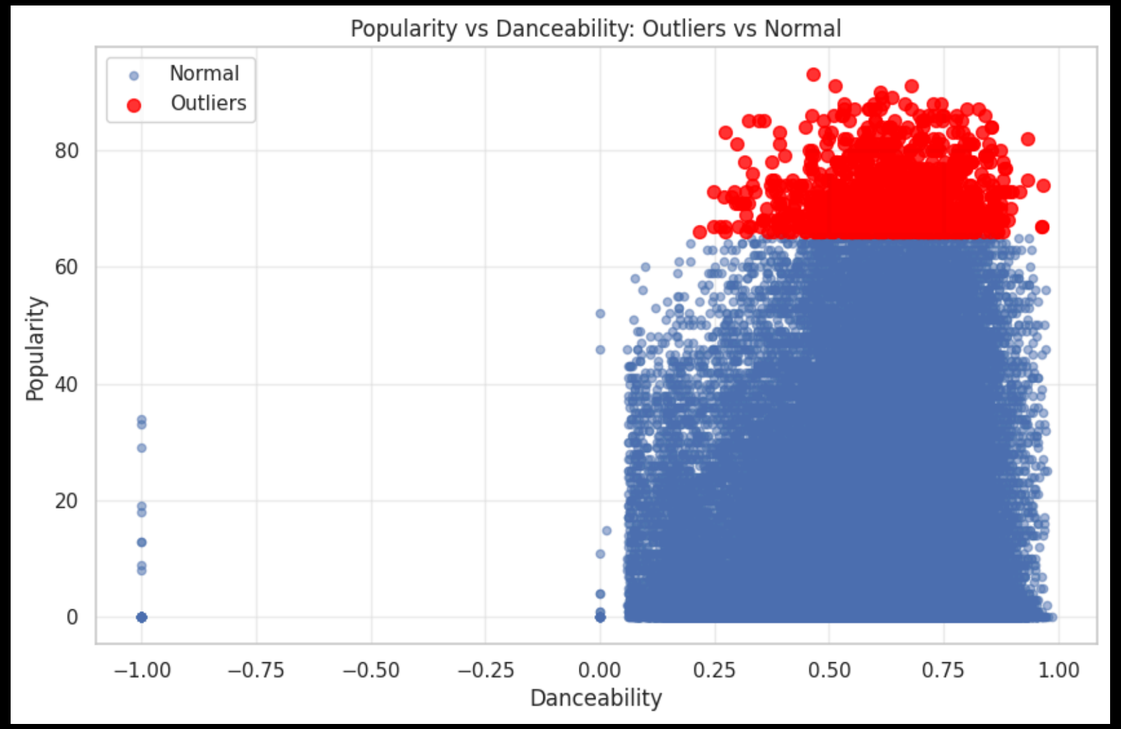
Outlier Analysis

Outlier Analysis Using IQR Method

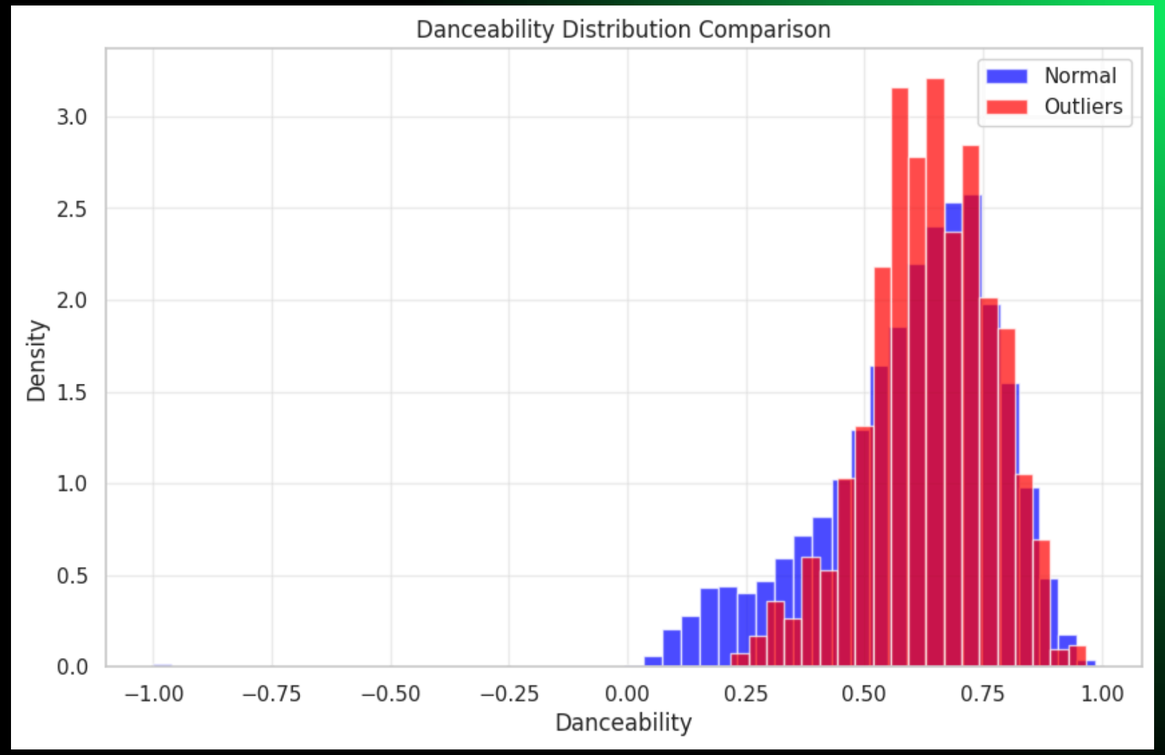


Interpretation

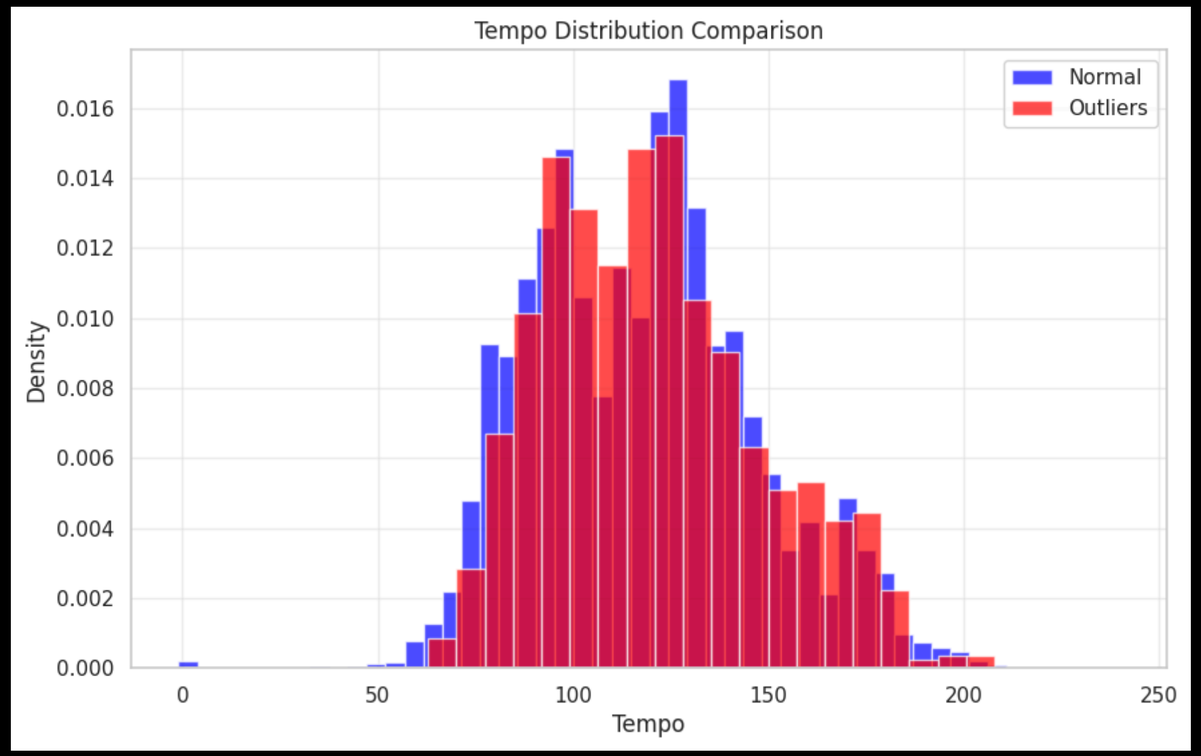
These boxplots show outlier detection for audio features: popularity, danceability, energy, valence, loudness, and tempo. Most features have very few outliers (0.1–2%), except loudness, which has many outliers (6.5%), indicating unusual values. Popularity, danceability, and tempo also show more outliers compared to other features.This figure displays boxplots with outlier counts for various song features. Most features, like energy and valence, have very few outliers (under 1%), while loudness has many (6.5%), and popularity, danceability, and tempo show moderate numbers of outliers (around 1-2%). Outliers in these features indicate data points with unusual or extreme values compared to the rest.



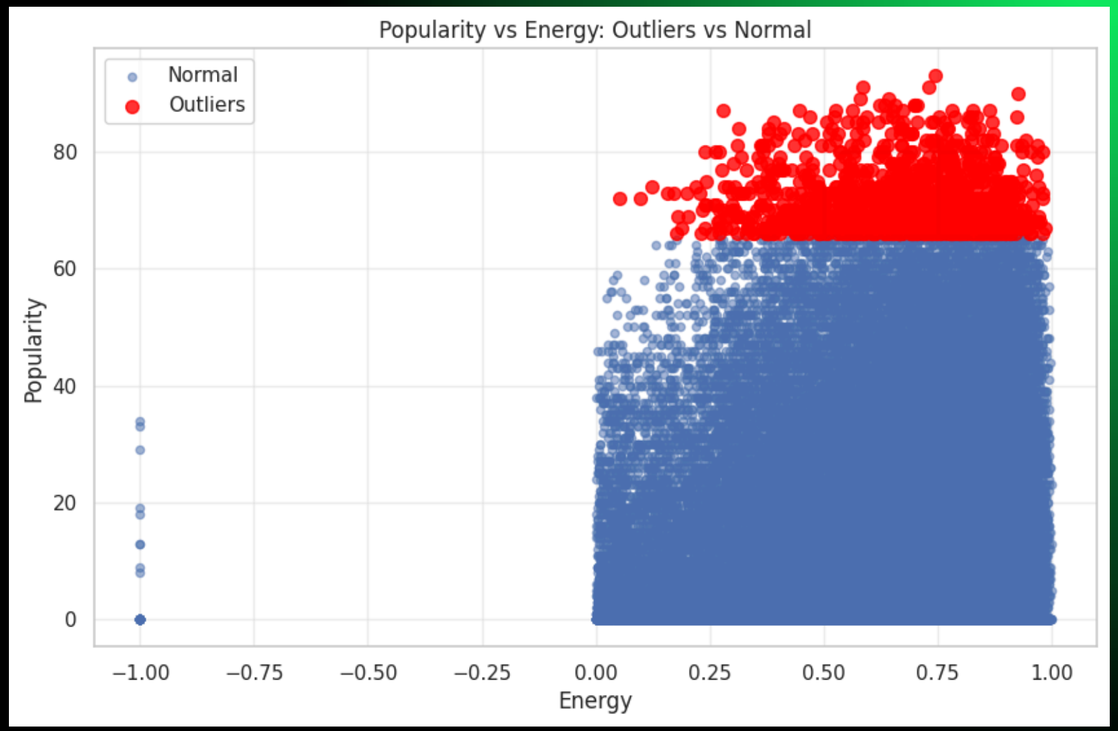
This scatter plot shows the relationship between danceability and popularity. Red points represent outliers, which have very high popularity at moderate-to-high danceability scores. The majority of songs (blue) have lower to mid popularity and are spread across all danceability values. Outliers tend to cluster at the top, showing the most popular songs are often more danceable.



This histogram compares the distribution of danceability scores between normal songs (blue) and outlier songs (red). Most outliers have danceability scores centered around 0.6 to 0.75, indicating they tend to be more danceable than the normal songs, which have a wider danceability range and a lower peak density.

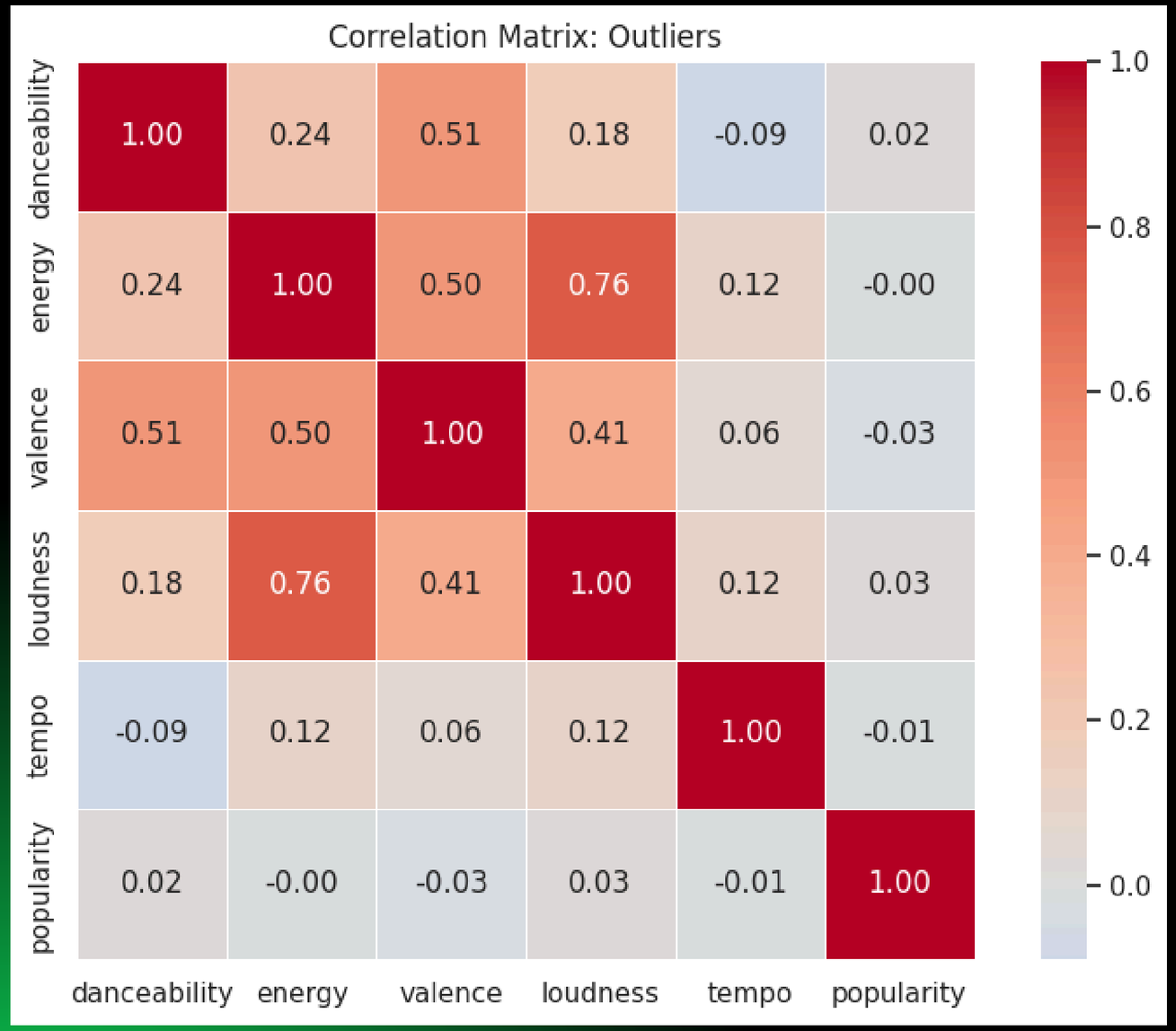


This chart compares tempo values for normal songs and outliers. Both have a similar distribution, centered around 100–130 BPM, but outliers are distributed a bit more widely, with slightly more presence at extreme tempo values compared to normal songs.



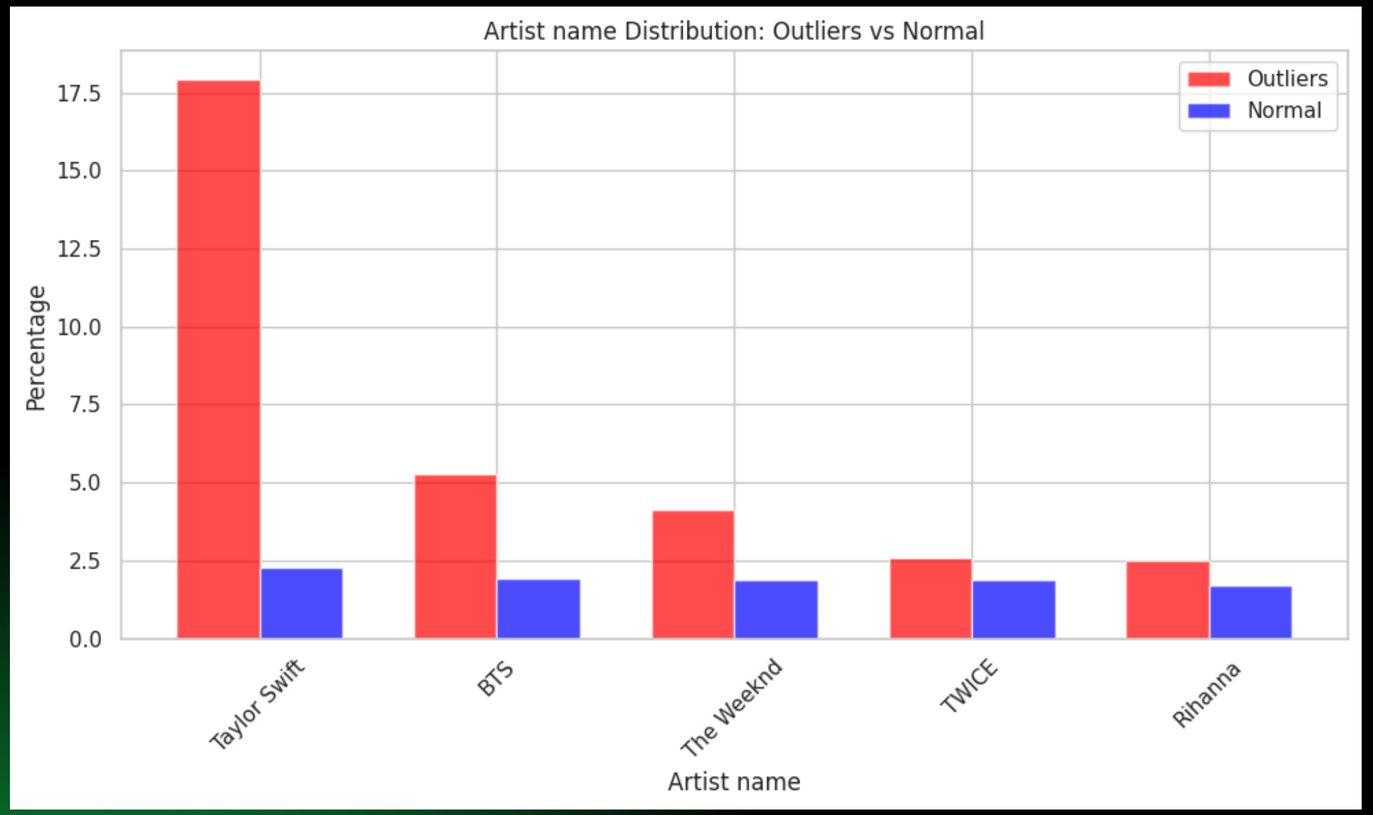
This scatter plot shows that songs with higher energy levels are more likely to be outliers with high popularity. Most normal songs (blue) have moderate popularity, while the most popular outliers (red) cluster at higher energy values. High energy is a common feature in the most popular songs.

Correlation analysis among outliers



This correlation matrix shows that for outlier songs, danceability, energy, valence, and loudness are moderately correlated with one another, especially energy with loudness (0.76) and valence. However, popularity is not strongly correlated with any of these features in the outlier group.

Top Artists in outliers vs normal



This bar chart compares the percentage of songs by top artists in outlier and normal song groups. Taylor Swift dominates the outlier group, far exceeding her presence in the normal group. Other artists like BTS, The Weeknd, TWICE, and Rihanna also have a higher share among outliers compared to normal songs, but the effect is strongest for Taylor Swift.

Summary and Actionable Insights — Spotify Dataset

🎯Recommendation 1: Focus on an “Energetic & Danceable” Sound Profile

🧩 Insight: Analysis shows that the top 25% of most popular tracks consistently exhibit high:

Danceability (median ≈ 0.72+)

Energy (median ≈ 0.80+)

These tracks balance rhythmic intensity and upbeat dynamics — forming the “sweet spot” of modern mainstream music.

🎵 Recommendation: Producers and engineers should prioritize energetic, rhythm-driven compositions.

For new projects, target the 0.7–0.9 range for both danceability and energy.

Use percussive layers, strong groove patterns, and compressed dynamic range to achieve the desired “dancefloor” sound profile.

Summary and Actionable Insights — Spotify Dataset

🎯 Recommendation 2: Master for a Modern, Impactful Mix

🧩 Insight: Longitudinal trend analysis revealed:
Loudness levels in popular songs have steadily increased over the years.
Duration (in ms) has decreased slightly, reflecting a preference for concise, high-impact tracks.
Modern hits are louder, tighter, and more dynamic — consistent with streaming-era listener behavior.
🎵 Recommendation: During mixing and mastering:
Aim for competitive loudness levels aligned with chart-topping standards (approx. -5 to -6 LUFS range).
Keep song length efficient (around 2.5–3.5 minutes) to improve completion and replay rates.
Use automated mastering tools or reference matching for consistency across streaming platforms.

Summary and Actionable Insights — Spotify Dataset

🎯 Recommendation 3: Leverage Shifting Language Trends

🧩 Insight: Language-based trend analysis reveals:
English songs remain dominant, but
Spanish and Korean songs have seen significant rises in average popularity within the last decade.
Cross-cultural hits are expanding listener markets, showing that language is becoming less of a barrier to global success.
🎵 Recommendation:
Explore multilingual collaborations or bilingual versions of key tracks.
Consider genre fusions (e.g., Latin Pop, K-Pop influences) to reach global audiences.
Utilize data-driven playlist strategies to position multilingual songs within high-engagement regions.

Summary and Actionable Insights — Spotify Dataset

🎯 Recommendation 4: Optimize Temporal Sound Evolution

🧩 Insight: Temporal feature analysis shows:
Valence (positivity) has fluctuated — happier songs dominate certain decades, while moodier tones have trended recently.
Tempo has seen mild declines, possibly reflecting the rise of chill/pop subgenres.
🎵 Recommendation:
Adjust production style according to evolving decade trends.
For current market alignment, slightly slower tempos (100–115 BPM) and moderately positive valence may yield stronger listener resonance.

THE MANHATTAN PROJECT

Conclusion

This phase concludes that popularity in modern music is driven by a synergy of energy, rhythm, and accessibility. Tracks that optimize both danceability and energy, maintain modern mastering levels, and adapt to shifting linguistic and emotional trends are most likely to succeed.

Thank You

For Watching this Presentation

By Parthivjit Basak