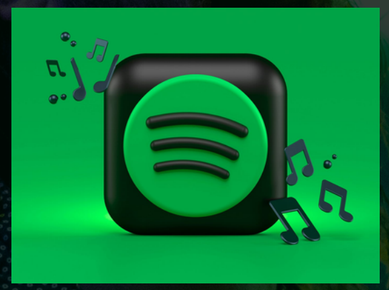
THE MANHATTAN PROJECT

P R E S E N T S



Exploratory Data Analysis

on

Spotify Tracks

01

02

03

~ By Ritav Paul

Problem Statement

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.

Objectives

- Initial Data Inspection
- Handling Missing Values
- Distribution Analysis
- Relationship Analysis
- Categorical Feature Analysis
- Trend Analysis
- Uncovering Insights

01

02

03

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 1.0.
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.

Data Description

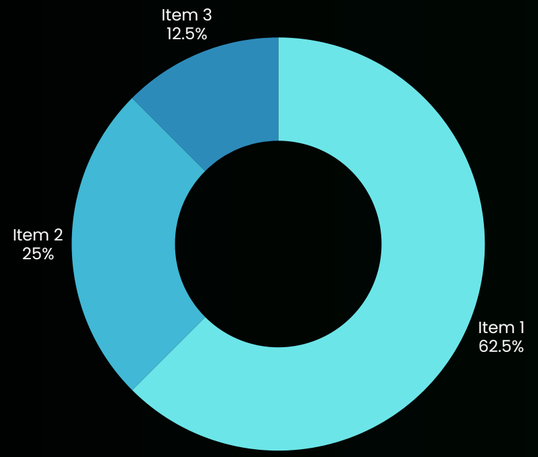
01

02

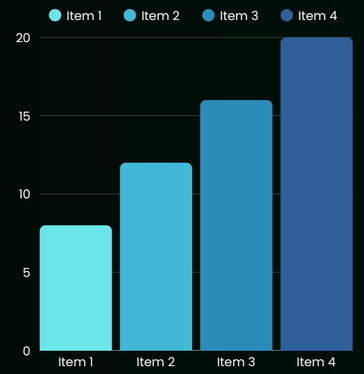
03

Workflow

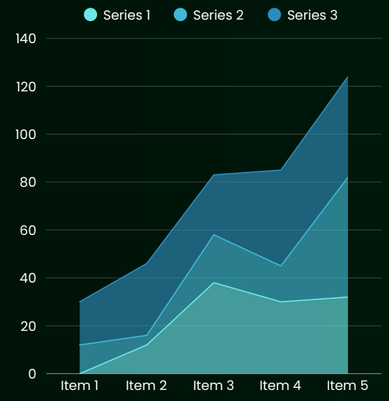
Univariate Analysis



Bivariate Analysis



Multivariate Analysis

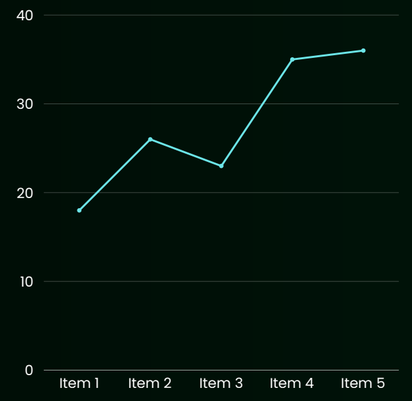


01

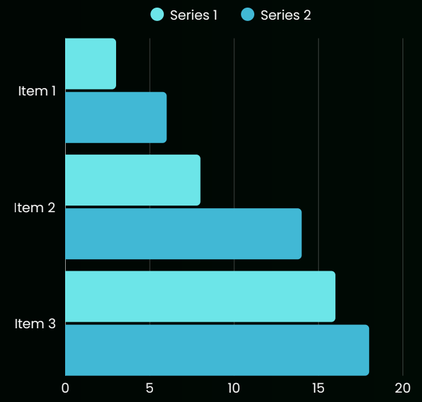
02

03

Time Series Analysis



Outlier Analysis



Univariate Analysis

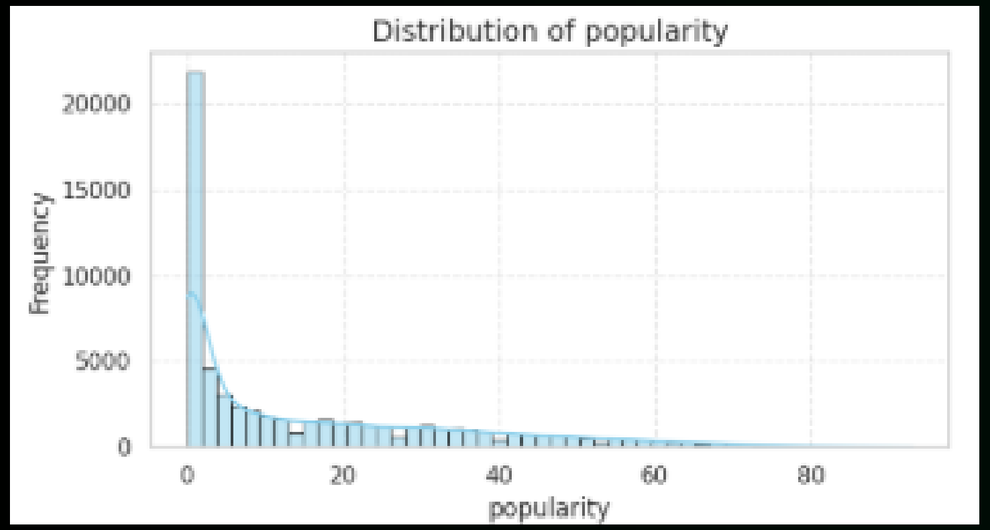
Learn More

01

02

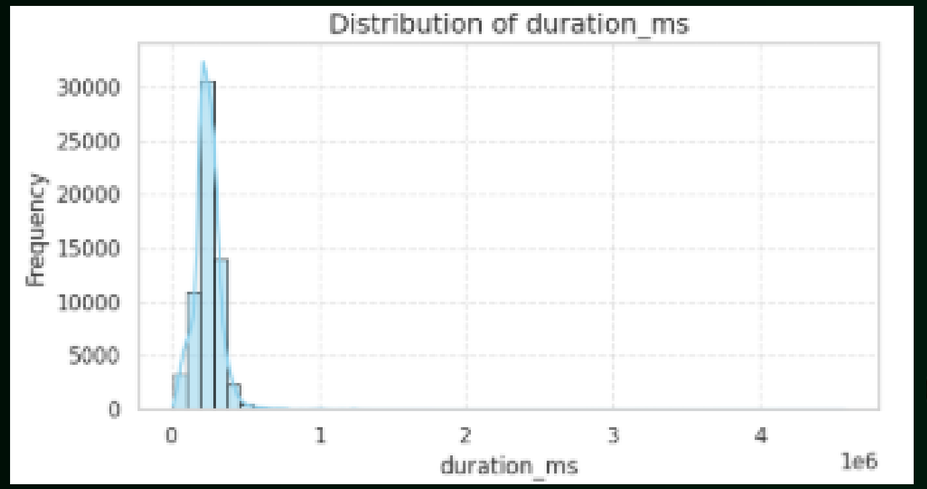
03

Distribution of Popularity Scores



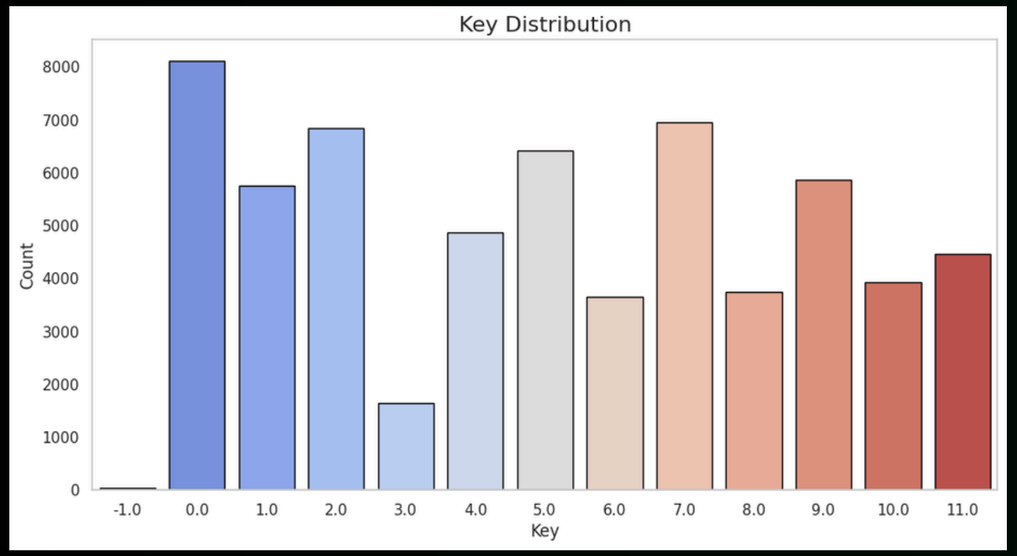
Popularity shows a slightly right-skewed distribution, with most songs clustered around low to moderate popularity values and fewer songs reaching very high popularity scores.

Average and typical song duration



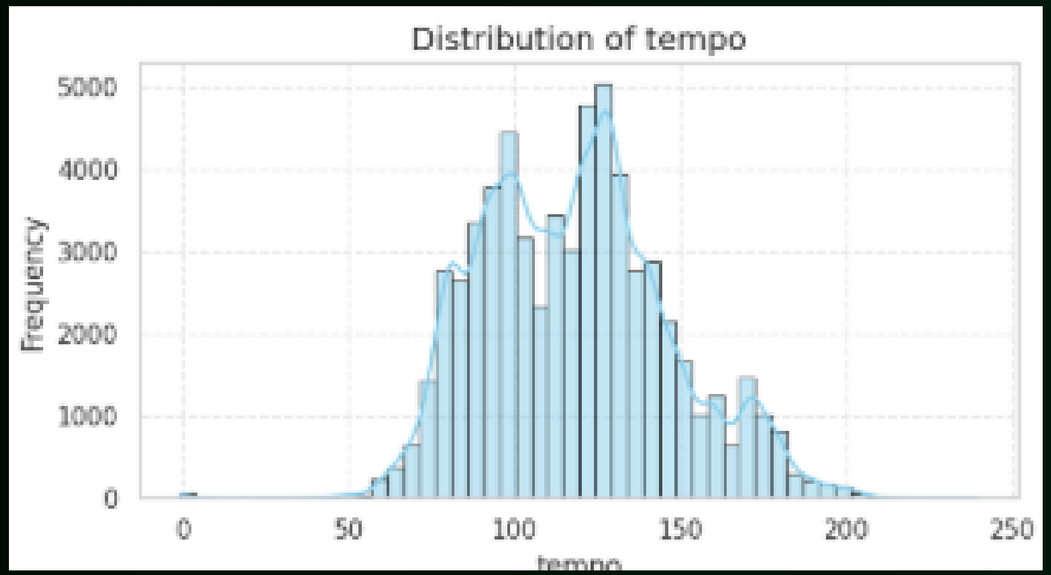
Song durations are right-skewed, with most songs between 2–5 minutes and a few extremely long tracks creating a long tail.

Most Frequent Keys



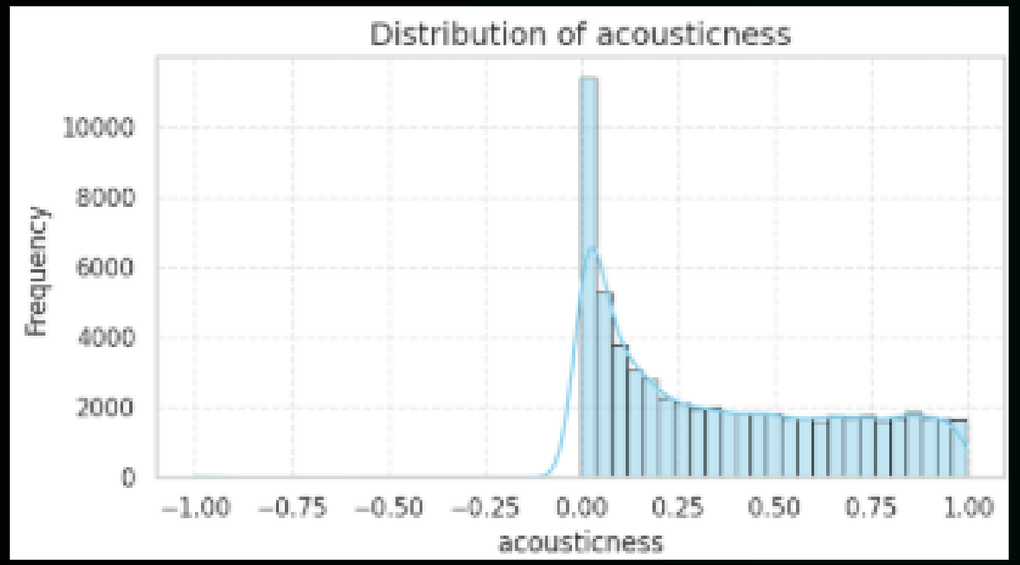
Most songs are concentrated in certain keys, such as C, G, and D, which are the most common in Western music. Less frequently used keys appear as lower bars, showing that some keys are rarely chosen by artists.

Distribution of Tempo



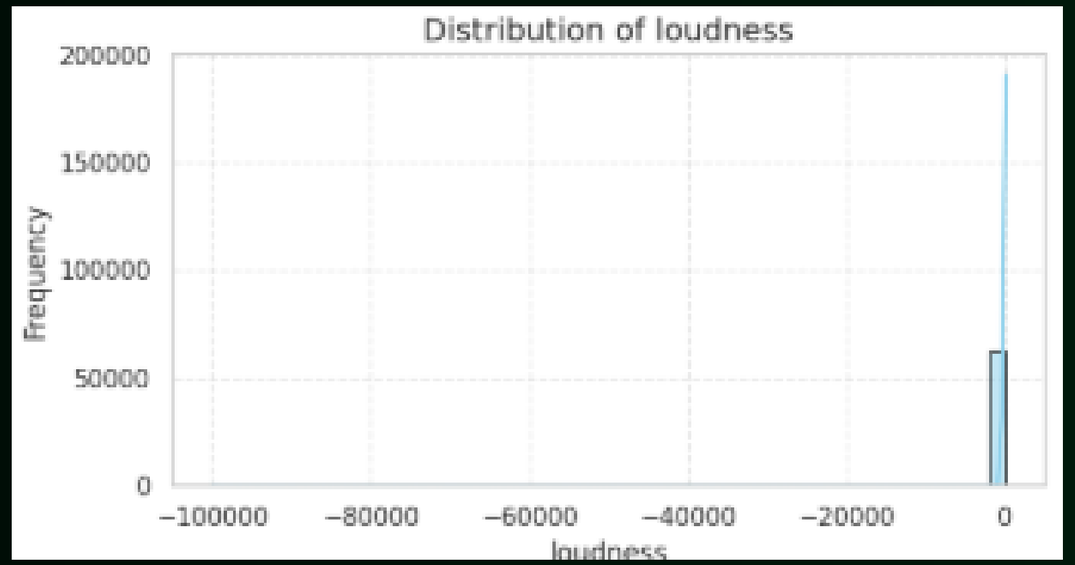
Tempo shows a moderately normal distribution, but with small peaks at common tempo values reflecting popular beats per minute.

Distribution of Acousticness



Acousticness displays a bimodal pattern, with prominent peaks near 0 and 1. This suggests that the dataset includes a mix of highly acoustic and minimally acoustic tracks, reflecting both traditional and electronically produced music.

Typical Loudness Levels (dB)



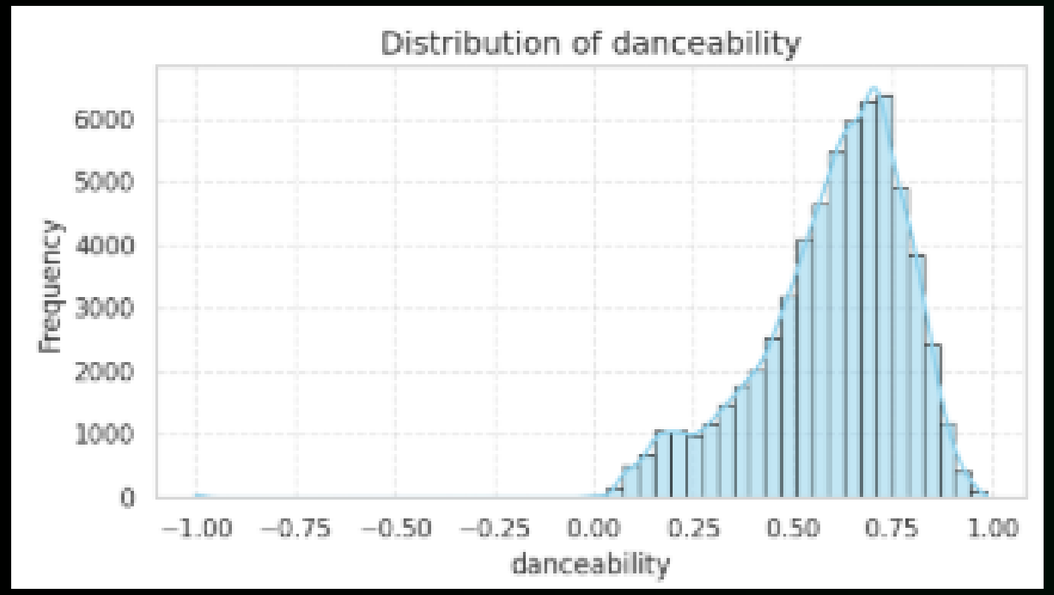
Loudness distribution is approximately normal, but with slight skew due to extremely quiet or extremely loud tracks.

01

02

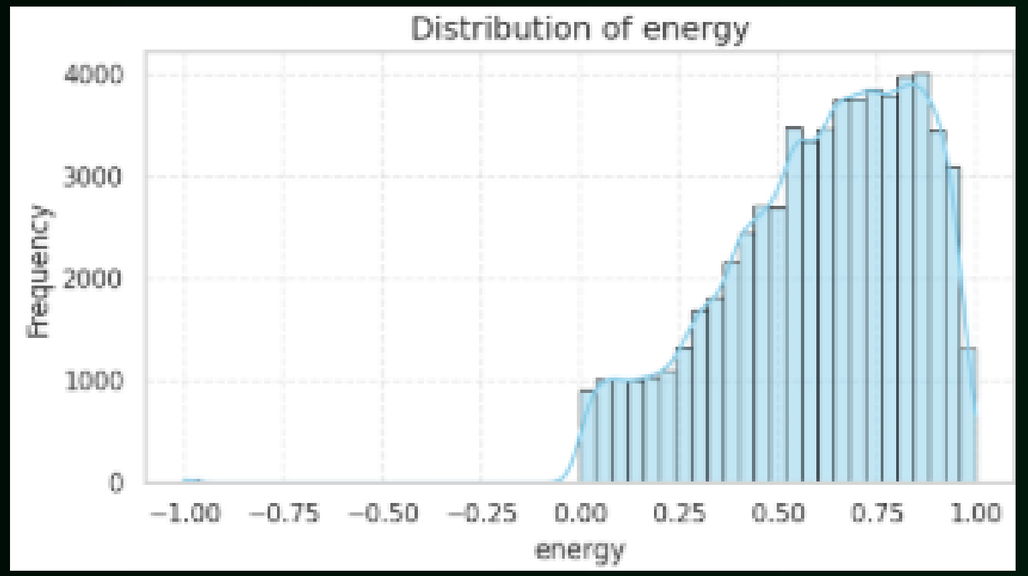
03

Distribution of Danceability



Danceability exhibits a slight left skew, with a concentration of songs in the moderate to high range. This indicates that the dataset primarily consists of rhythmically engaging, dance-oriented tracks.

Distribution of Energy Levels



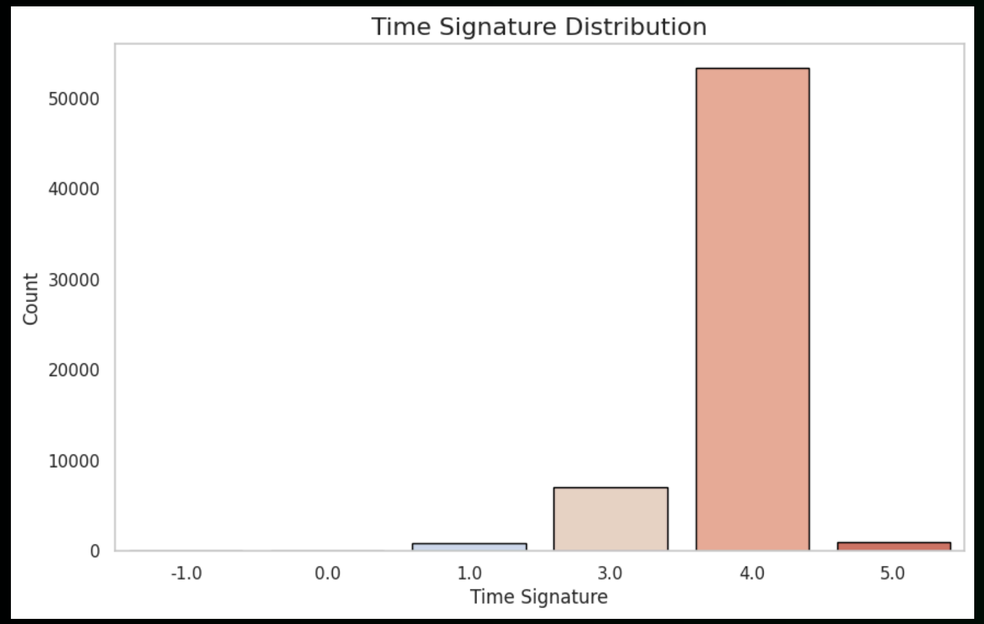
Energy exhibits a slightly left-skewed distribution, suggesting many songs have moderate to high energy levels.

01

02

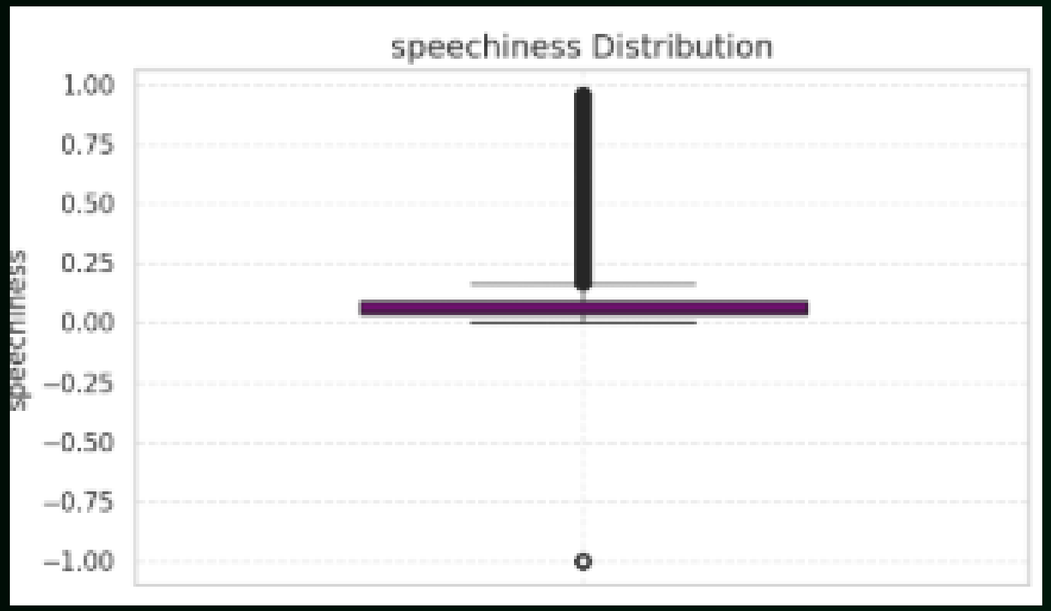
03

Most Common Time Signatures



Most songs use the standard 4/4 time signature, which is the dominant rhythm in modern music. Other time signatures like 3/4 or 6/8 occur rarely, appearing as shorter bars in the plot, indicating their less frequent use in popular tracks.

Distribution of Speechiness



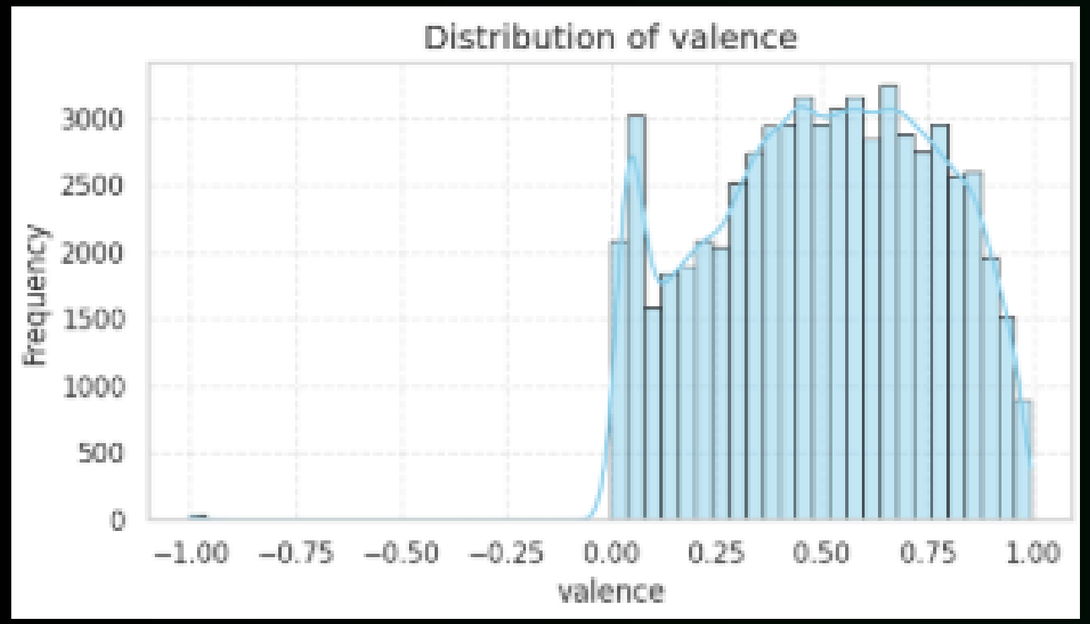
Speechiness is highly right-skewed. Most songs have low speech content, with rap or spoken-word tracks as high-value outliers.

01

02

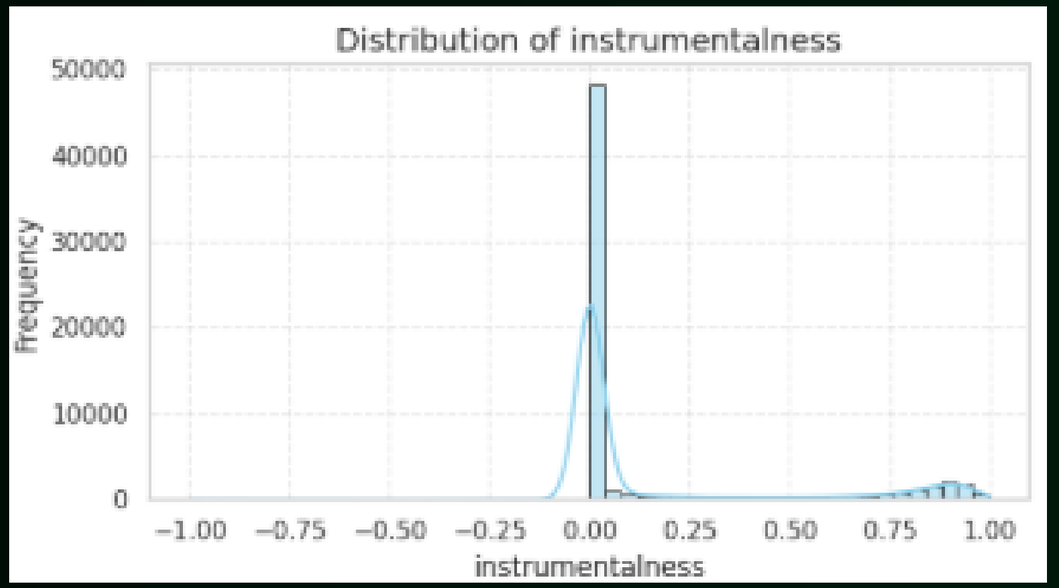
03

Distribution of Valence (Musical Positivity)



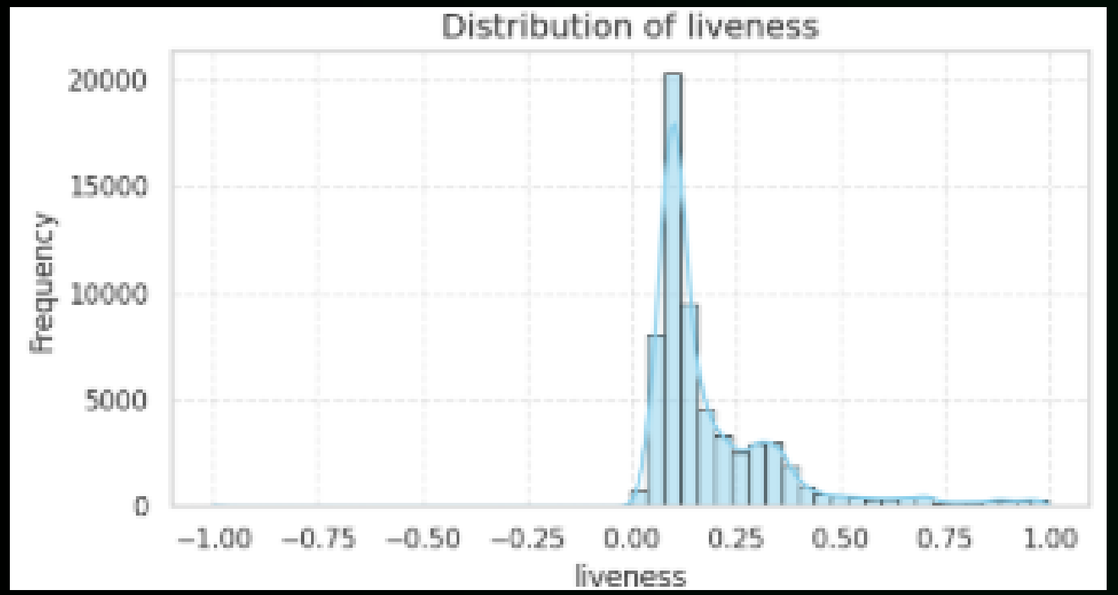
Valence is slightly bimodal, with peaks near low and high values, indicating songs tend to be either very sad/negative or very happy/positive in mood.

Distribution of Instrumentalness



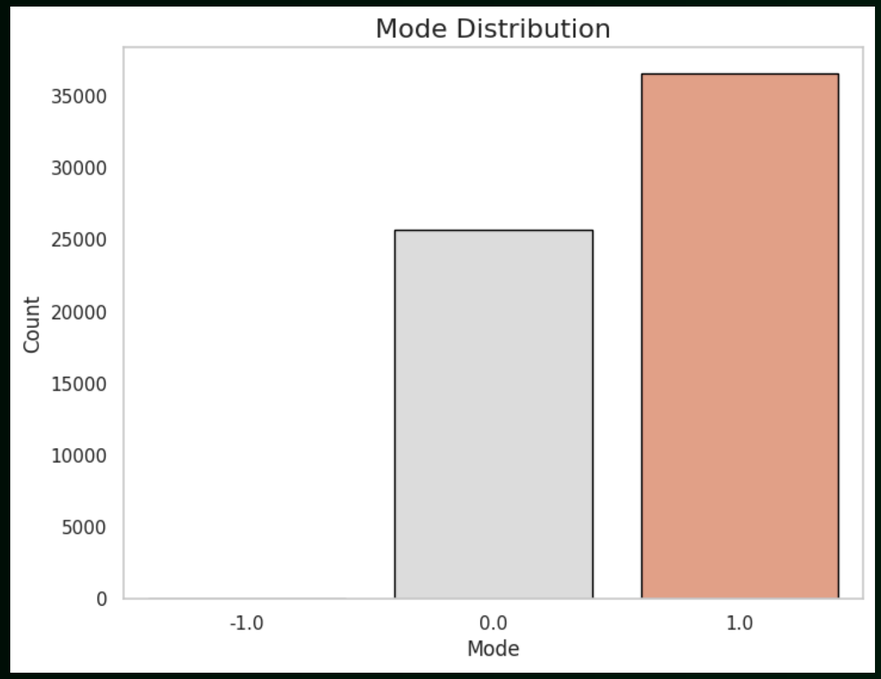
Instrumentalness is highly right-skewed, as most tracks have low instrumentalness, but a small number are fully instrumental, creating a long tail.

Distribution of Liveness



Liveness is right-skewed, with most songs having low liveness scores, indicating studio recordings, and a few songs with high liveness representing live recordings.

Common Values for Mode (Major/Minor)



The majority of songs are in major mode, with a smaller proportion in minor mode. This reflects the general preference for brighter, major-key tracks in popular music.

01

02

03

Median & Quartiles for Popularity and Duration

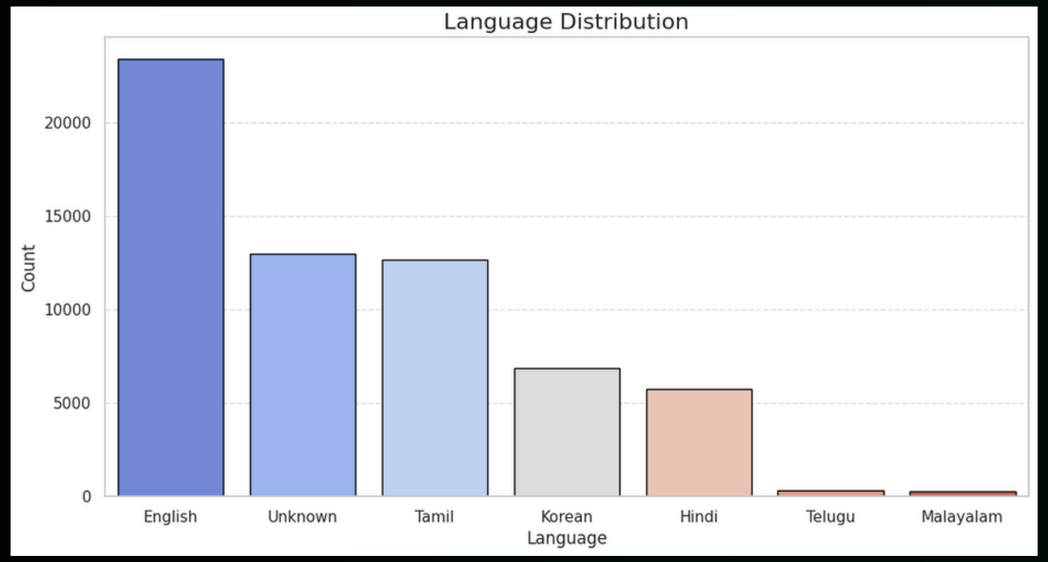
Popularity

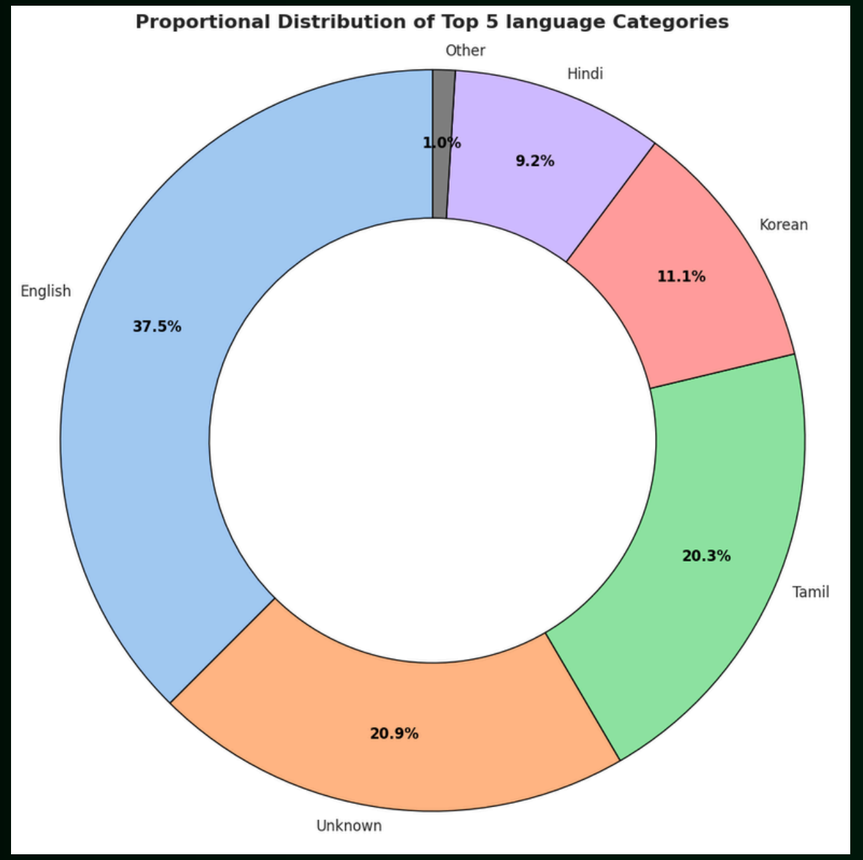
```
Central Tendency of 'popularity':
Mean: 15.3584
Median: 7.0000
Mode: [0]

Spread of 'popularity':
Minimum: 0.0000
Maximum: 93.0000
Standard Deviation: 18.6269
25th Percentile (Q1): 0.0000
75th Percentile (Q3): 26.0000
```

Duration

```
Central Tendency of 'duration_ms':
Mean: 242527.0400
Median: 236267.0000
Mode: [232627.0]

Spread of 'duration_ms':
Minimum: 5000.0000
Maximum: 4581483.0000
Standard Deviation: 112999.9326
25th Percentile (Q1): 192160.0000
75th Percentile (Q3): 286240.0000
```

Distribution Across Language Categories





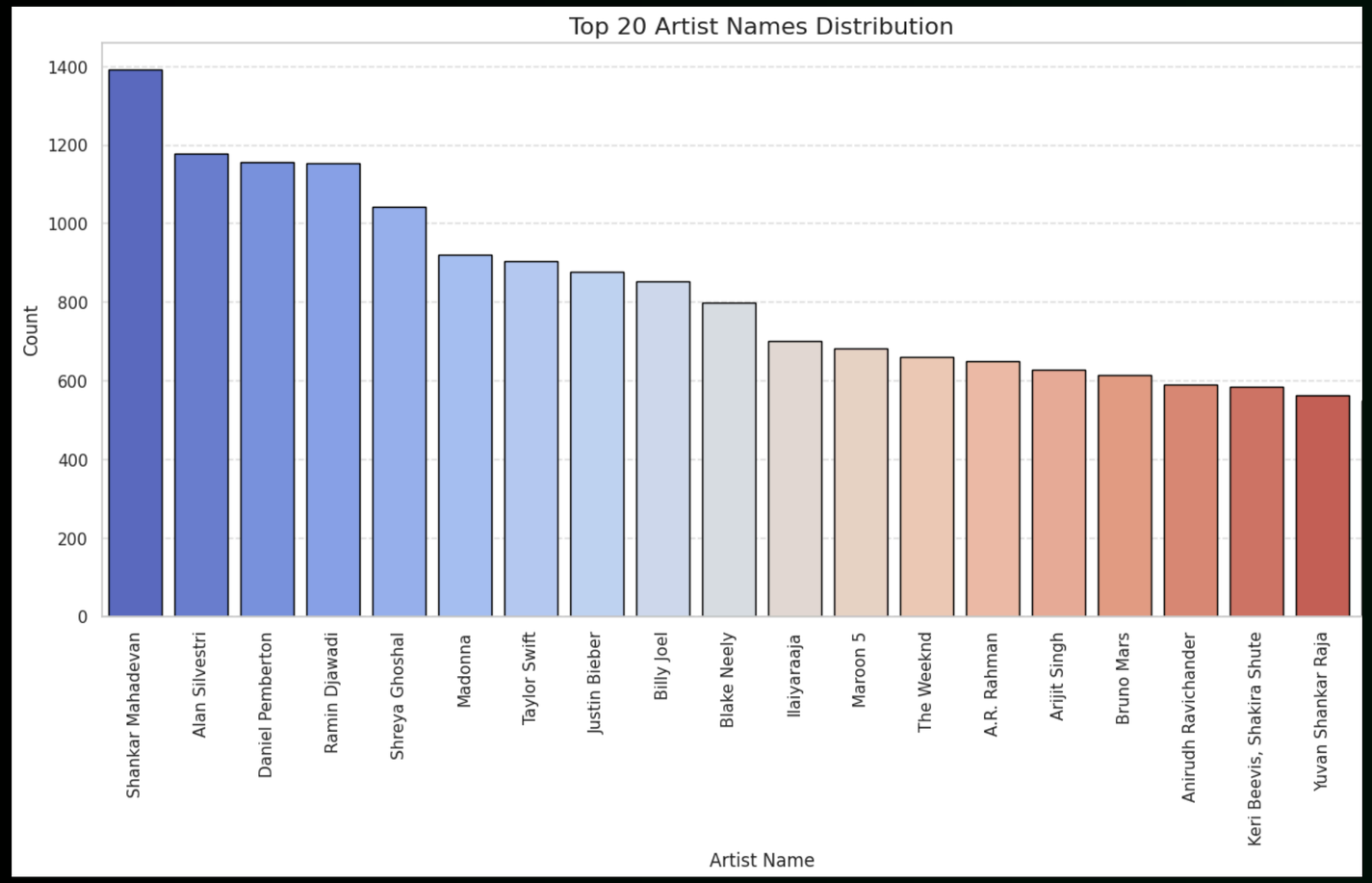
The dataset is dominated by English songs, followed by Unknown and Tamil tracks. Korean and Hindi are moderately represented, while Telugu and Malayalam have very few songs, indicating a strong bias toward English music in the dataset

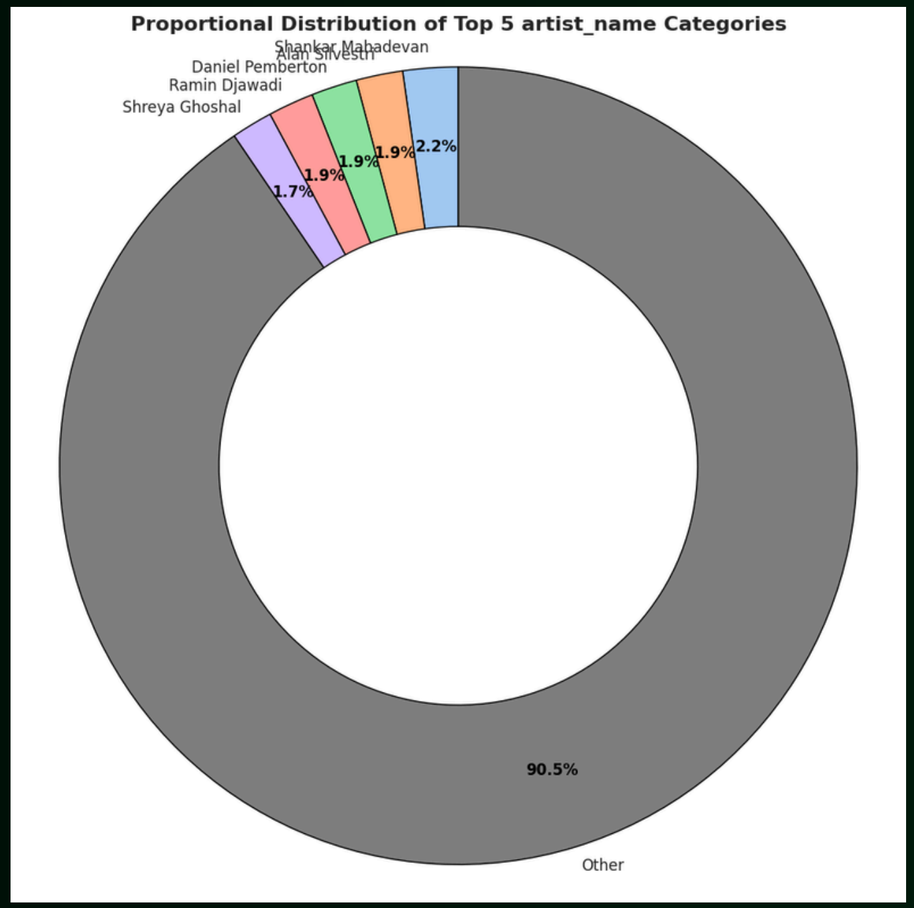
01

02

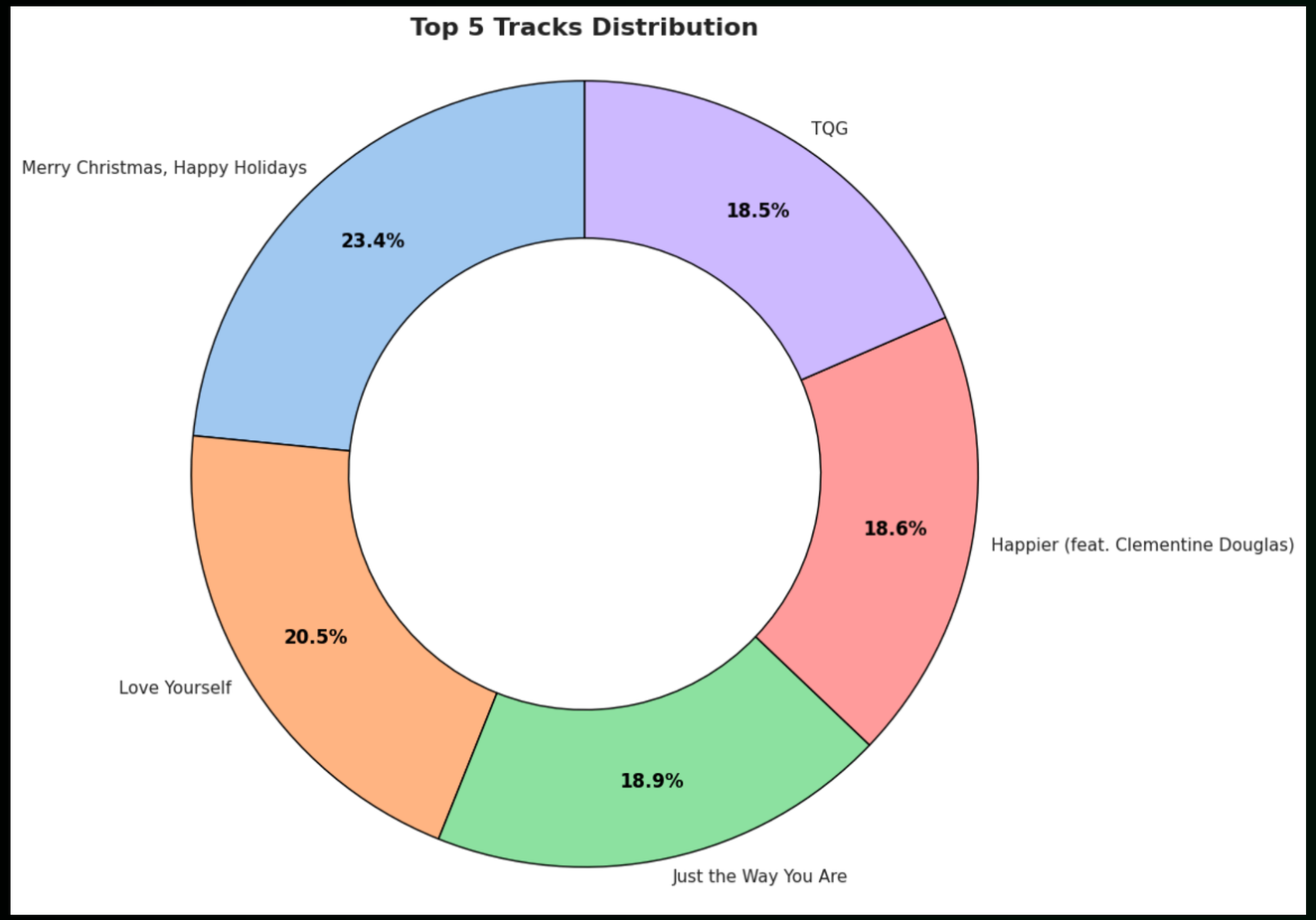
03

Meet The Top 20 Artists

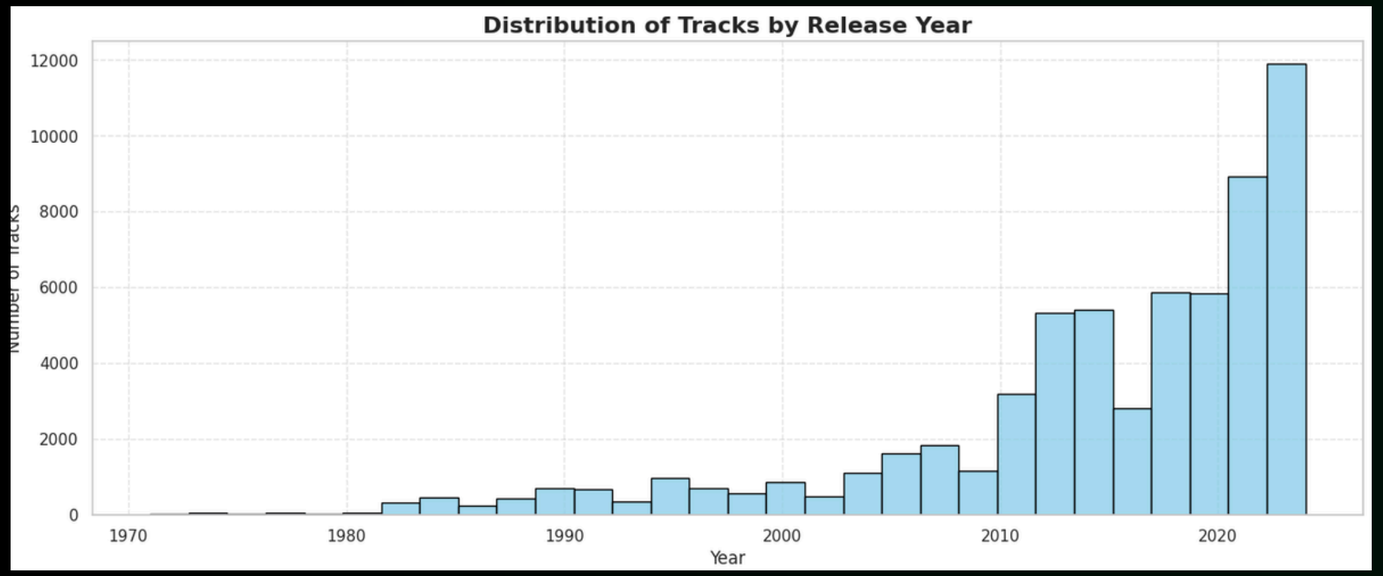




Know The Top 5 Tracks



Distribution of Tracks by Release Year



This graph shows a dramatic exponential surge in track releases, with the volume remaining low and stable until the late 1990s.

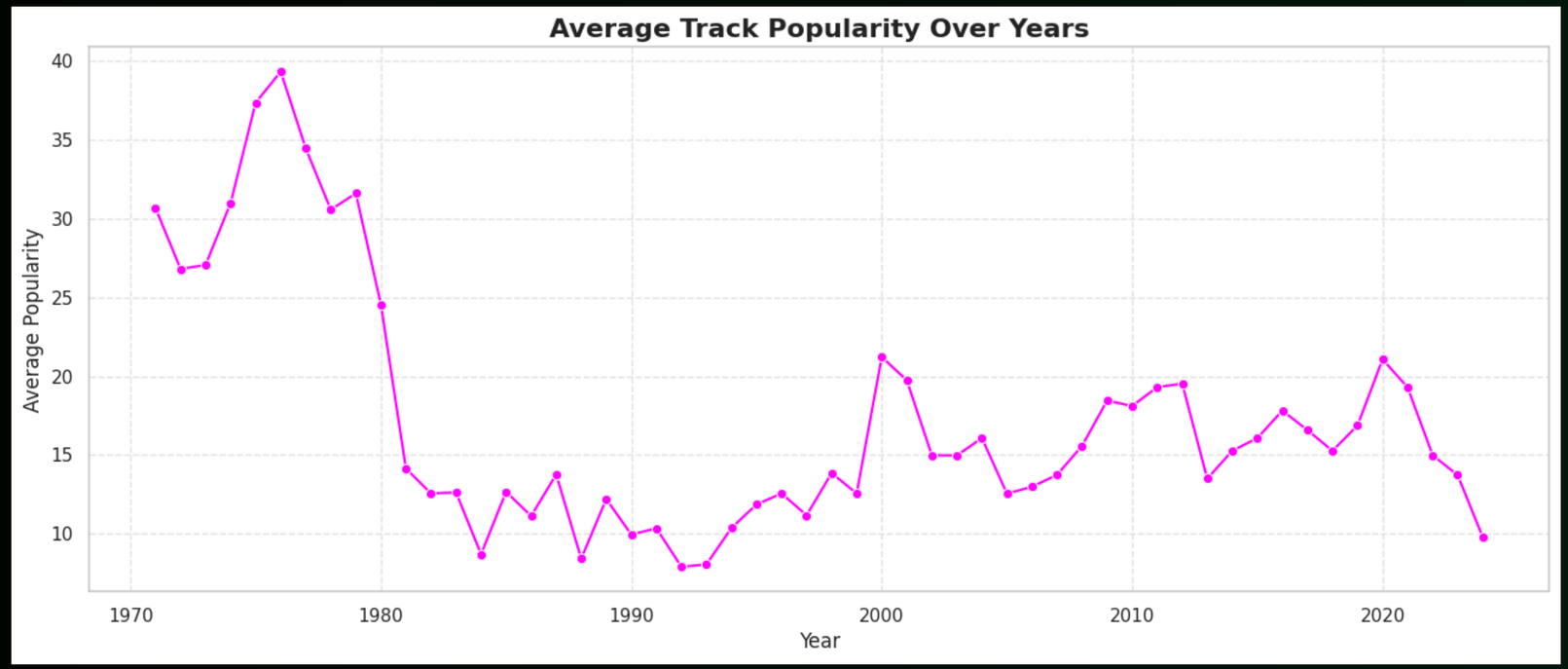
Pre-2000: Minimal and steady output.

2000s: Gradual growth begins.

Post-2010: Explosive growth, culminating in the highest peak (over 12,000 tracks) in the final year shown.

Takeaway: Music production has been democratized and hyper-accelerated by digital platforms.

Average Track Popularity over years



The graph shows average track popularity has declined sharply from a high peak in the mid-1970s.
1970s: High popularity, peaking near 40.
Post-1980: Dramatic drop and sustained low popularity, generally below 20.
Modern Era: Stability at a much lower level, with a recent final year showing a low point.
Conclusion: The average track faces more competition and is significantly less popular now than in the 1970s.

Bivariate Analysis

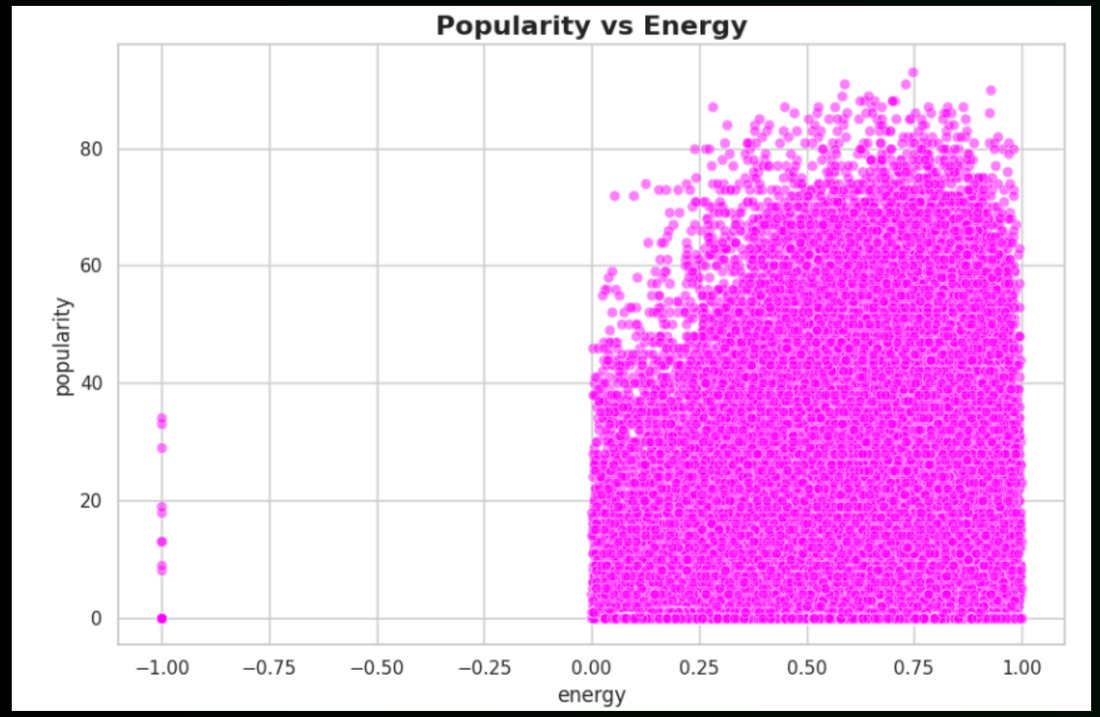
Learn More

01

02

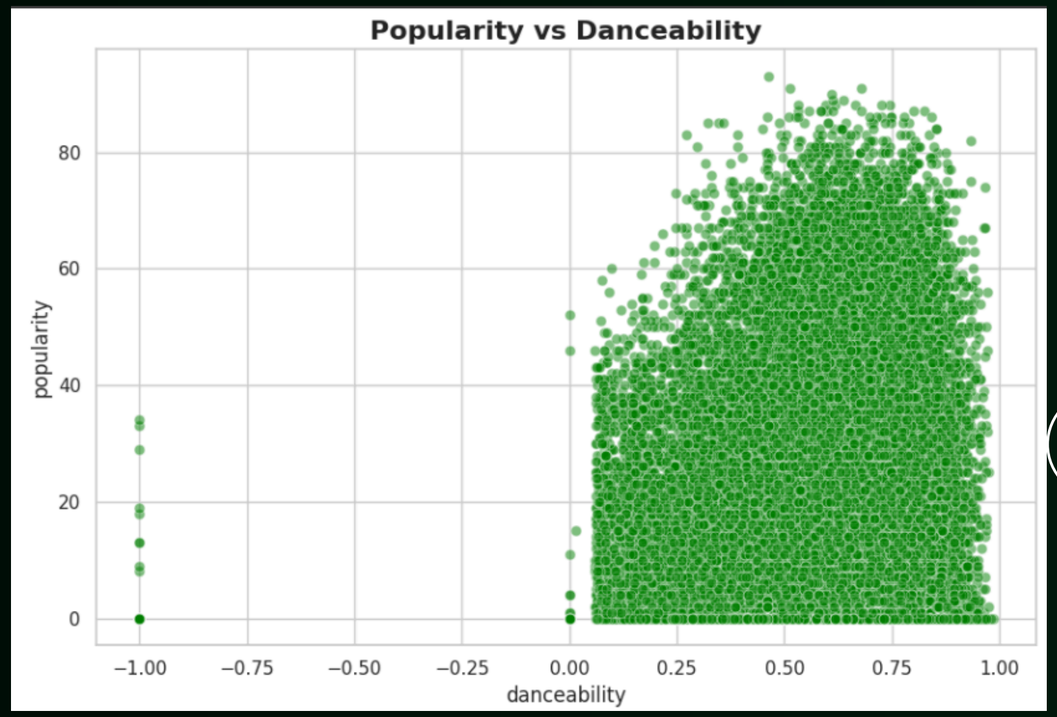
03

Popularity vs Energy



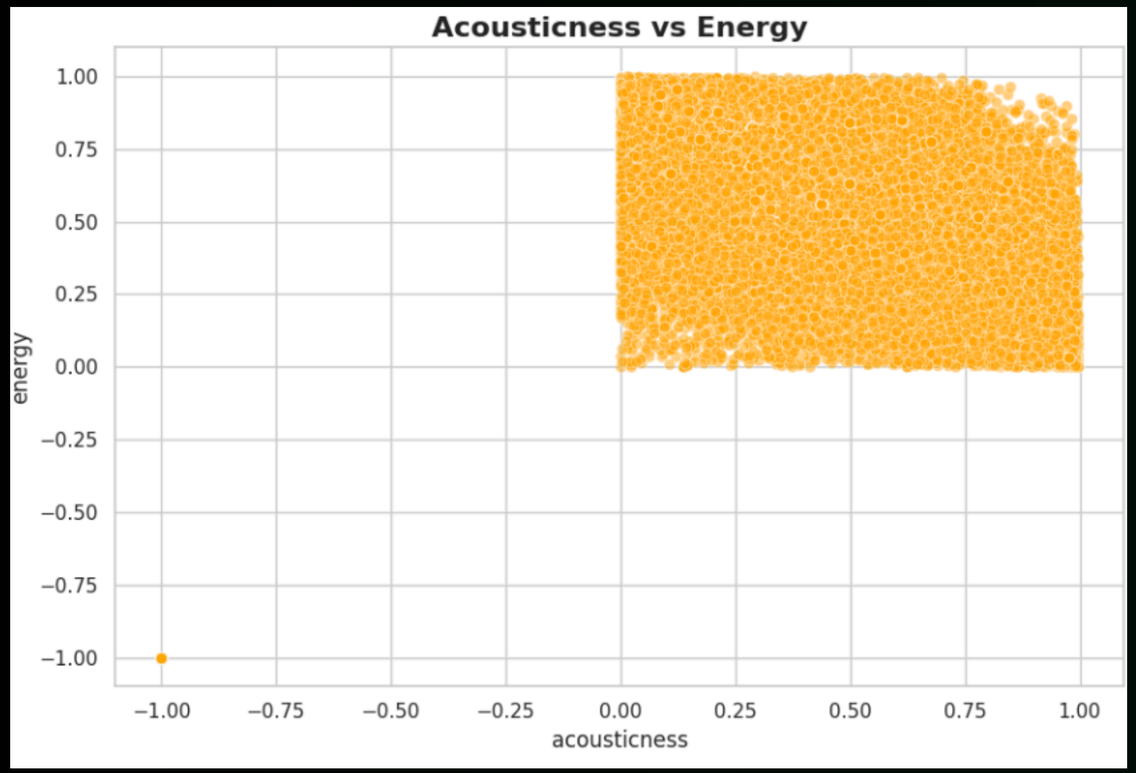
Most high-energy tracks tend to have higher popularity.
Spread shows variability; some moderate-energy tracks are also popular.

Popularity vs Danceability



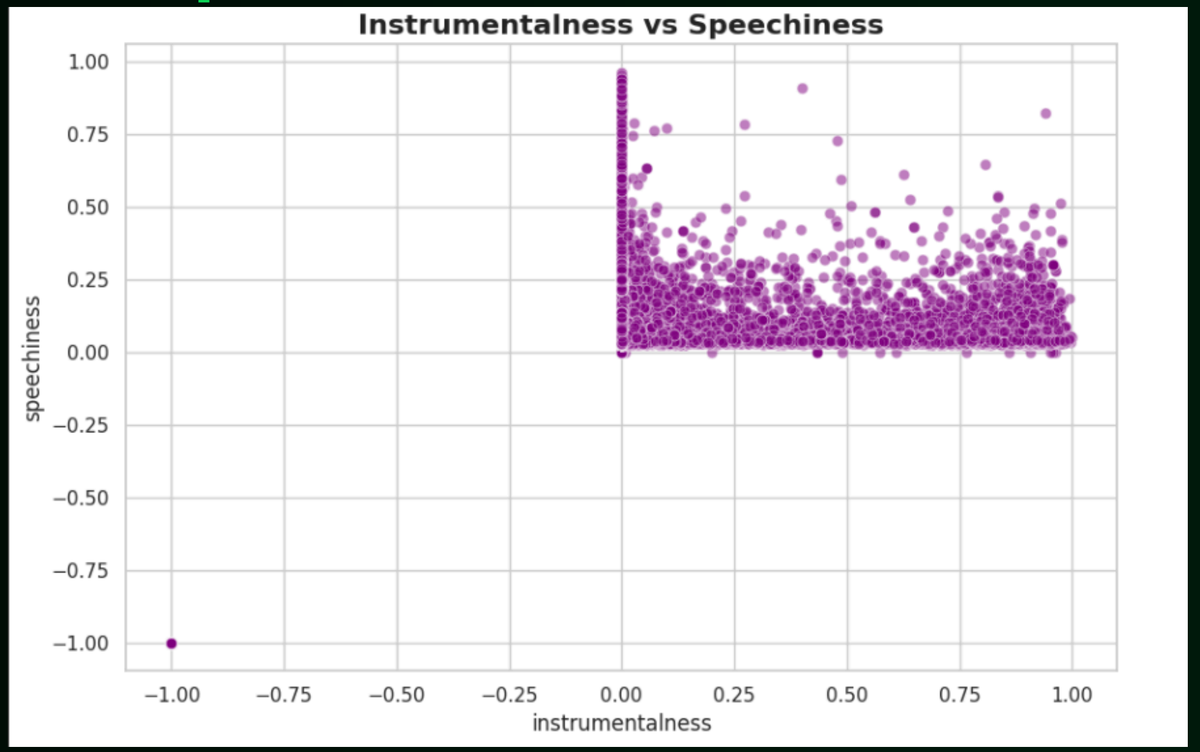
Positive trend; tracks that are more danceable generally achieve higher popularity.
Outliers exist with very popular tracks having low danceability.

Acousticness vs Energy



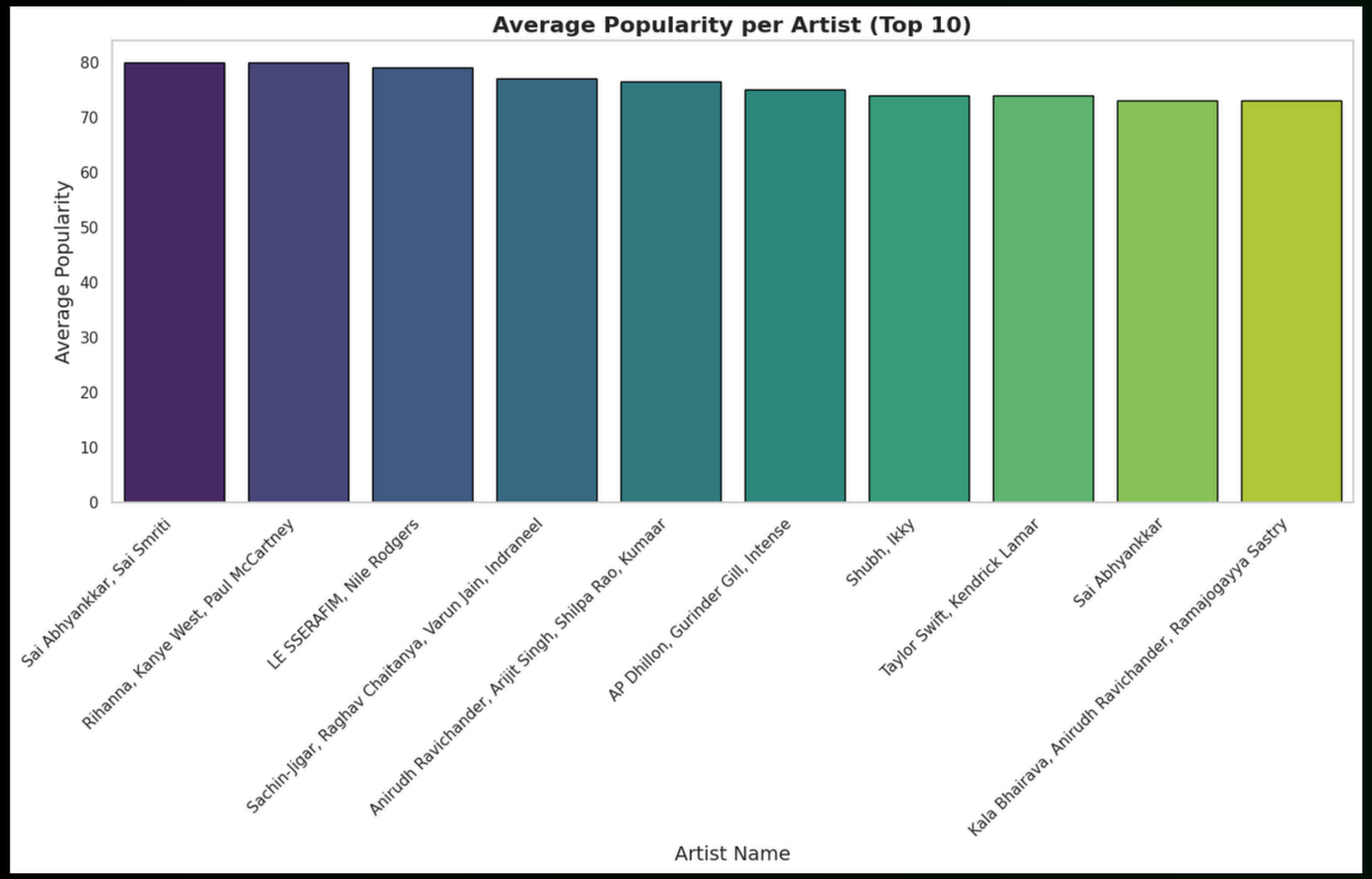
Negative correlation; highly acoustic tracks tend to have lower energy.
Few acoustic tracks appear with high energy (outliers).

Instrumentalness vs Speechiness



Strong negative correlation; instrumental tracks have low speechiness, as expected.
Tracks with vocals have high speechiness and low instrumentalness.

Average Popularity per Artist (Top 10)

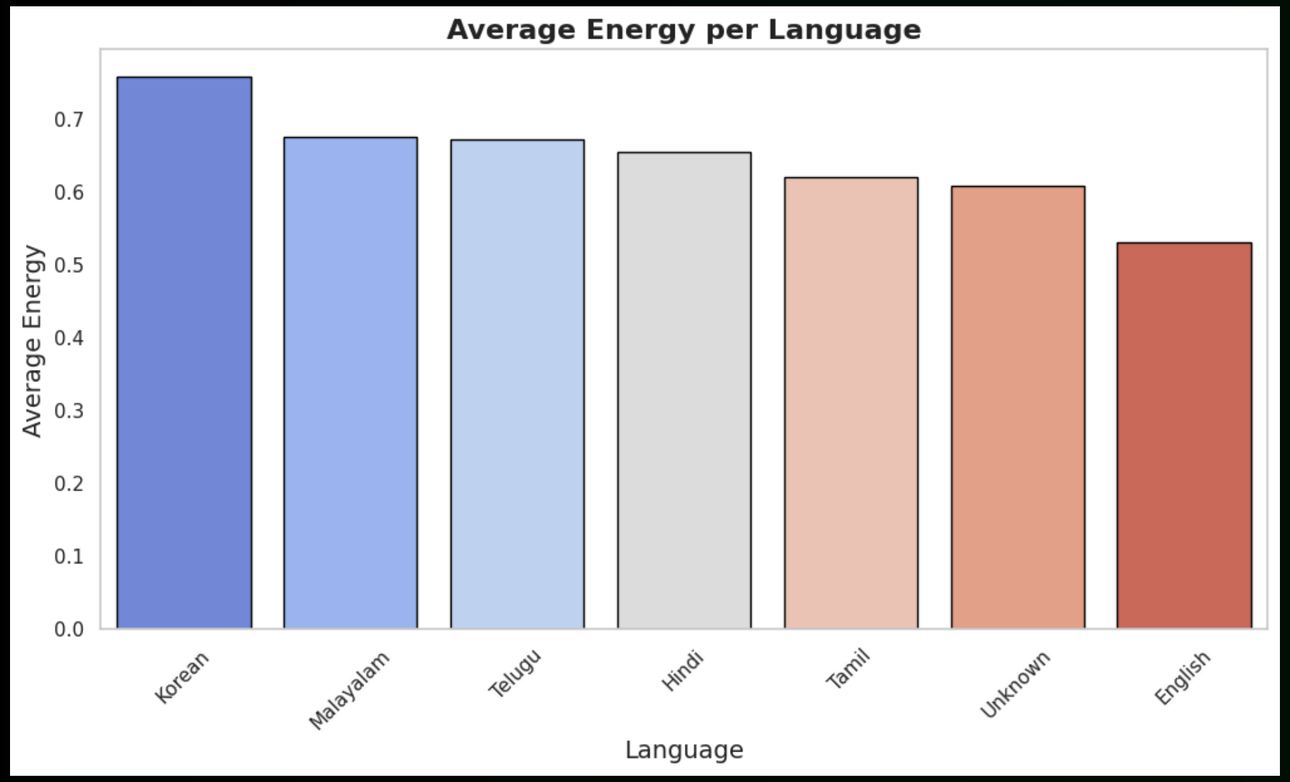


High and Even Popularity: All top artists exhibit a very high and tightly clustered average popularity, with scores generally between 74 and 80.

The Top Tier: Sai Abhyankkar, Rihanna/Kanye West/Paul McCartney, and LE SSERAFIM/Nile Rodgers form the highest tier, all averaging near 80.

Key Insight: There is minimal differentiation among the top artists in terms of average popularity, suggesting that the very highest tier of music, regardless of genre or region, maintains a consistently high level of appeal.

Average Popularity per Language



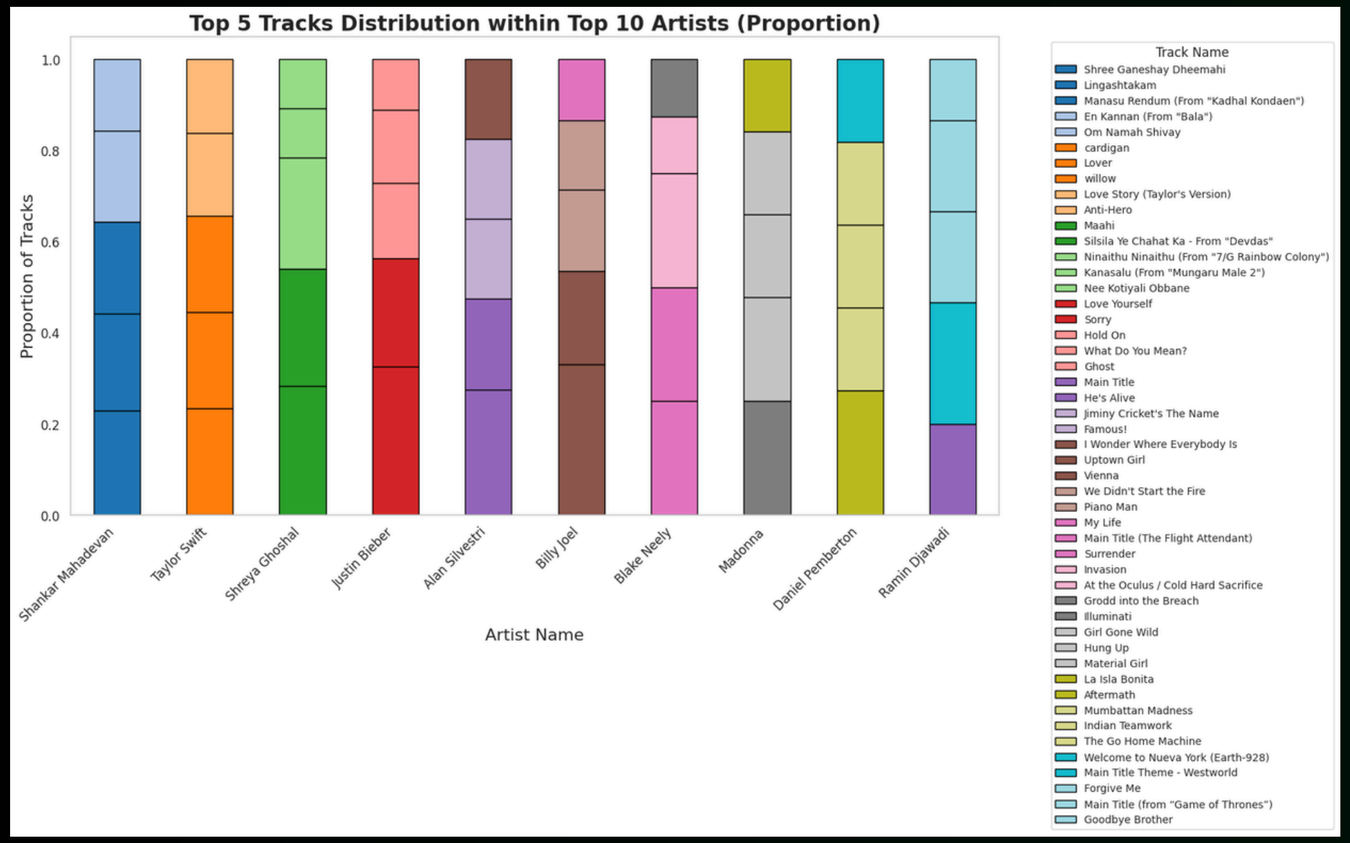
Korean Dominance: Korean music (K-Pop) has the highest average energy score (above 0.7), suggesting a predominantly high-tempo and active musical style in the dataset.

Clustered High Energy: Malayalam, Telugu, Hindi, and Tamil music are also clustered with relatively high average energy scores (around 0.62 to 0.68).

Lowest Energy: English music, followed by a category labeled "Unknown," has the lowest average energy score (around 0.53), indicating a broader inclusion of lower-tempo or more subdued tracks.

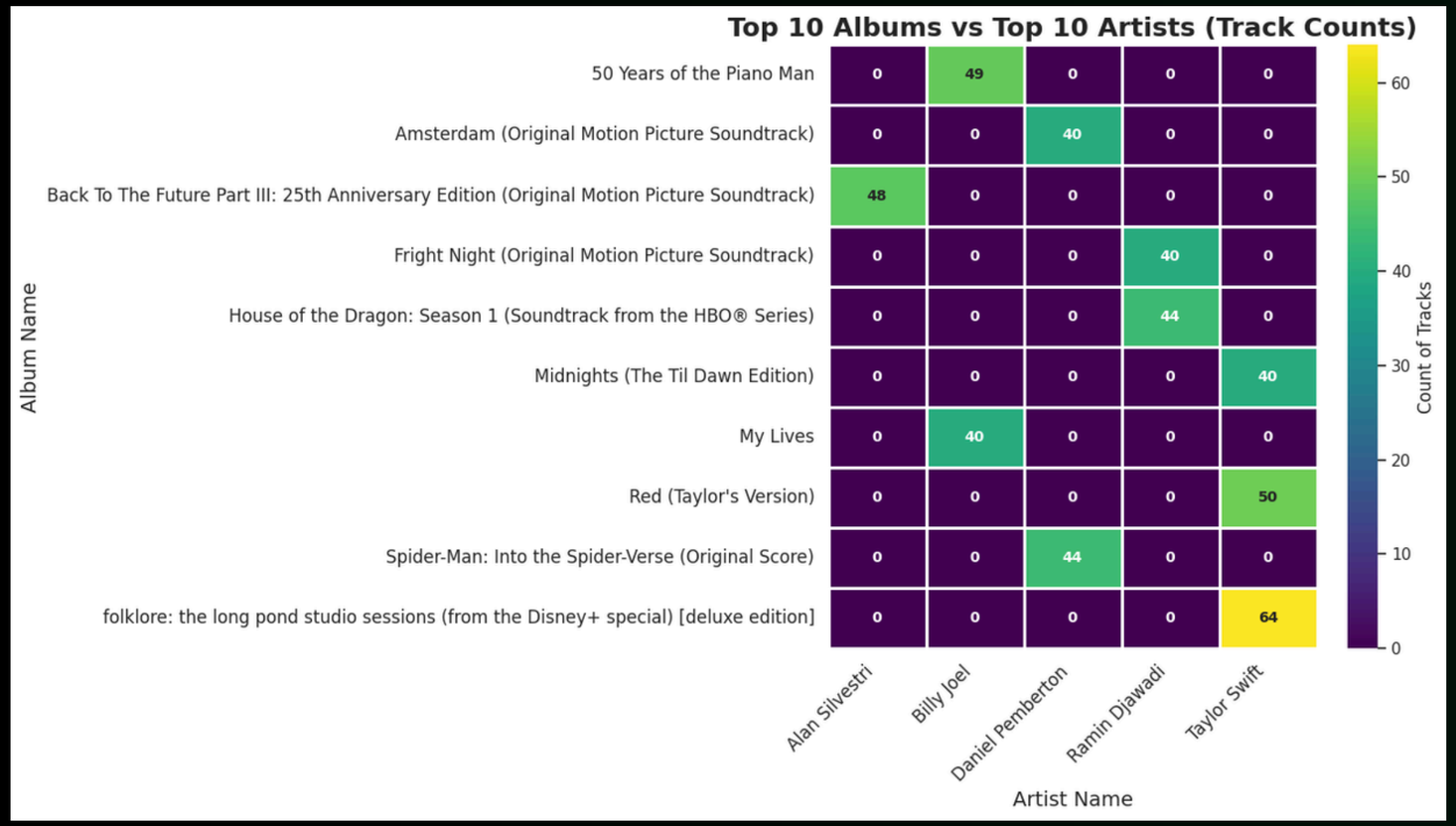
Overall: Non-English language tracks in this dataset tend to exhibit a higher average musical energy compared to their English counterparts.

Distribution of the Top 5 Tracks as a proportion of total tracks for the Top 10 Artists



- Track Concentration Varies: Some artists (e.g., Shreya Ghoshal, Justin Bieber, Billy Joel) show a very uneven distribution, with a few songs (large colored blocks) making up a high proportion of their total tracks in the dataset.
- Track Diversity: Other artists (e.g., Shankar Mahadevan, Taylor Swift, Blake Neely) show a more even distribution across their top tracks (smaller, more equal-sized colored blocks), indicating a broader variety of popular songs driving their total track count.
- Key Insight: The chart highlights which artists' popularity is heavily reliant on a few dominant tracks versus those whose popularity is driven by a larger catalog of well-performing songs.

Top 10 Artists vs Top 10 Albums



- Strong Artist-Album Dominance: The data shows clear, concentrated relationships. Artists generally dominate a single album in this list, with most cells showing zero interaction.
- Taylor Swift's High Concentration: Taylor Swift has the highest single concentration of tracks, with 64 tracks from the folklore: the long pond studio sessions (deluxe edition) and 50 tracks from Red (Taylor's Version).
- Billy Joel's Catalog: Billy Joel's contribution is concentrated on the compilation album, 50 Years of the Piano Man, accounting for 49 tracks.
- Soundtrack Specialists: Artists like Daniel Pemberton (Spider-Man: Into the Spider-Verse) and Ramin Djawadi (House of the Dragon) show clear specialization in their respective soundtracks, with 44 tracks each.
- Key Takeaway: The top artists' track counts are not broadly distributed across many albums; instead, they are heavily concentrated in specific, successful, or comprehensive release packages (deluxe editions, compilations, or soundtracks).

Multivariate Analysis

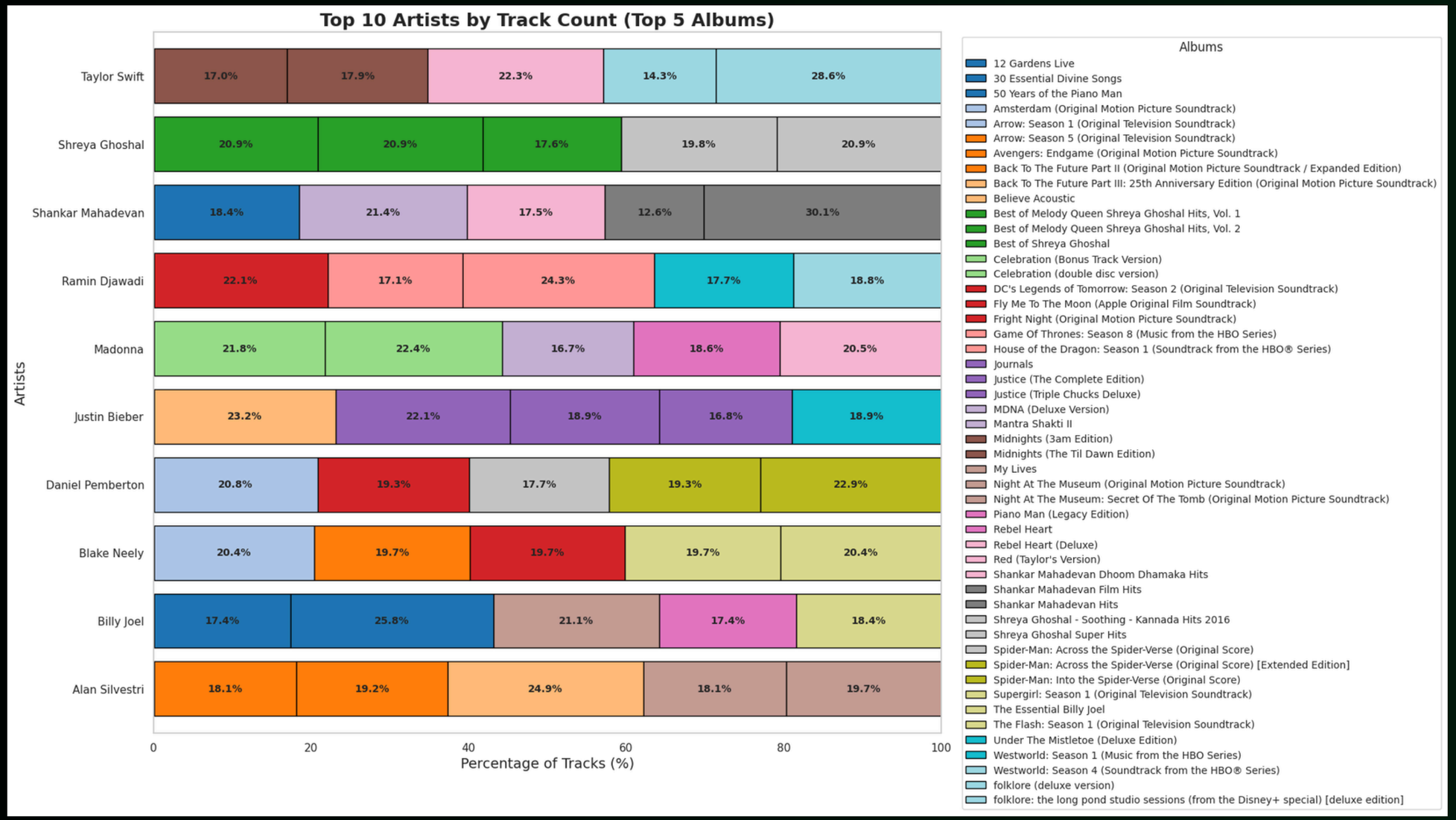
Learn More

01

02

03

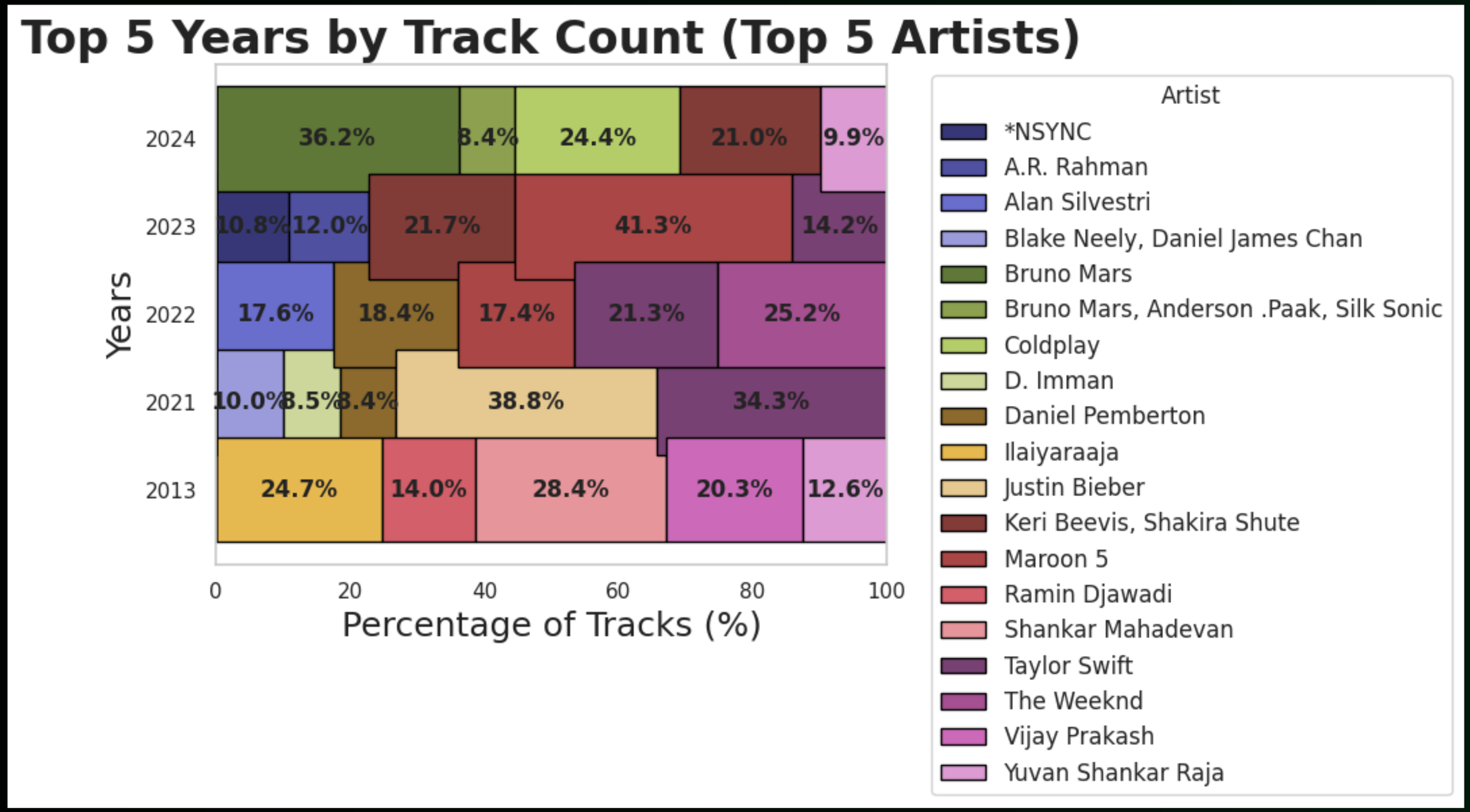
Top 10 artists by track count across their top 5 albums



Interpretation

- **Diversification**: Artists like Taylor Swift and Shreya Ghoshal exhibit high diversification, with no single album dominating (all top 5 albums contribute between 17% and 30%). This indicates a pattern of releasing multiple substantial albums.
- **Concentration**: Other artists show slight concentration, where one album accounts for a larger share of the total tracks from the top five. For instance, Alan Silvestri's and Billy Joel's largest individual albums contribute nearly 25% each.
- **Genre Influence**: Soundtrack composers (e.g., Ramin Djawadi, Alan Silvestri) frequently top the list due to the nature of their work—scores often have high track counts across multiple seasons or extended editions, driving their overall total.
- **Top Album Variance**: The contribution of the top individual album varies widely among artists, highlighting significant differences in the track quantity and packaging strategy (deluxe vs. standard) within the highest-ranking releases.

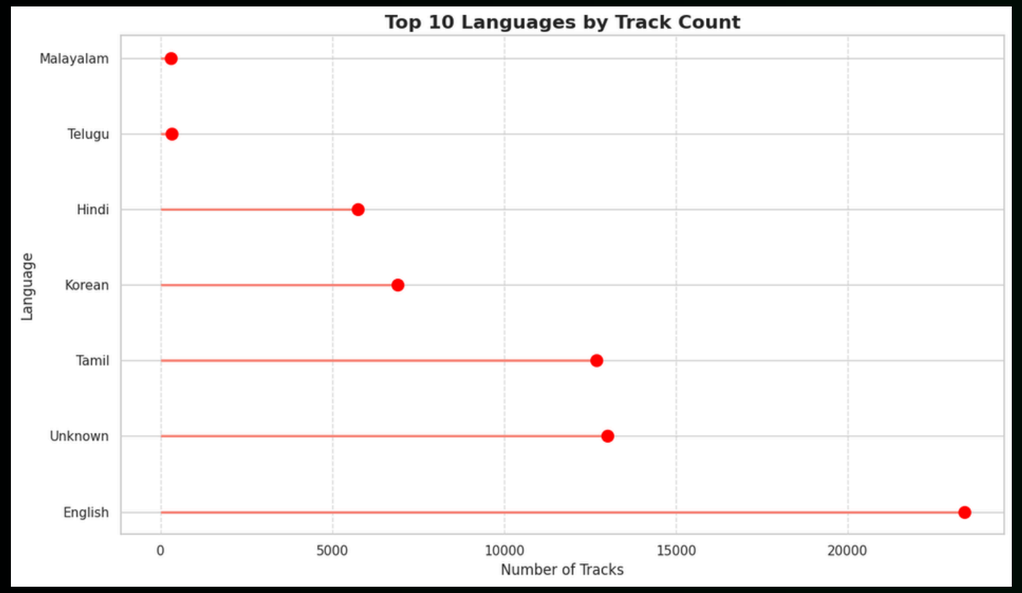
Top 5 years by track count for top 5 artists

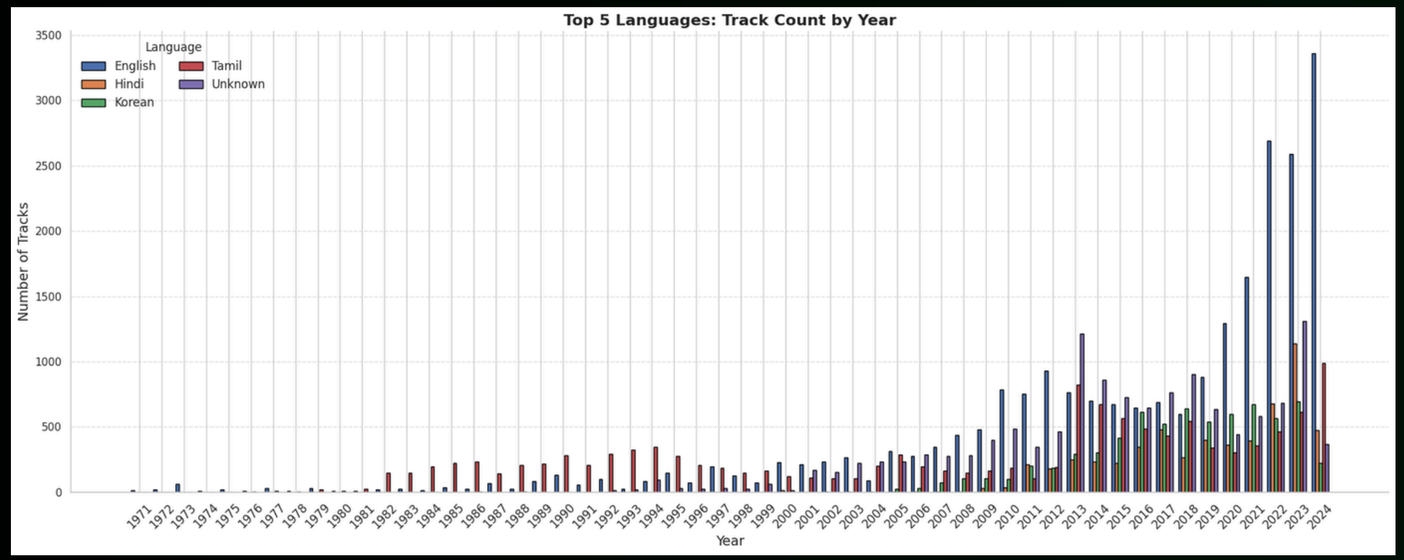


Interpretation

- Yearly Dominance: The chart reveals which artists drove the track count in each of the top five years. For example, 2021 was heavily dominated by Justin Bieber (38.8%) and The Weeknd (34.3%), accounting for over 73% of that year's total tracks from the featured artists.
- Artist Concentration: 2023 shows the highest concentration, with Ramin Djawadi (41.3%) contributing over two-fifths of the tracks. 2024 is more diversified, led by Bruno Mars (36.2%) and Coldplay (24.4%).
- Historical Context: 2013, the earliest year shown, features a strong split between Justin Bieber (24.7%), Ramin Djawadi (28.4%), and Vijay Prakash (20.3%).
- Recurring Impact: Ramin Djawadi (Red) and Justin Bieber (Yellow/Tan) appear as major contributors in multiple years, suggesting consistent release of high-track-count albums across the examined period.

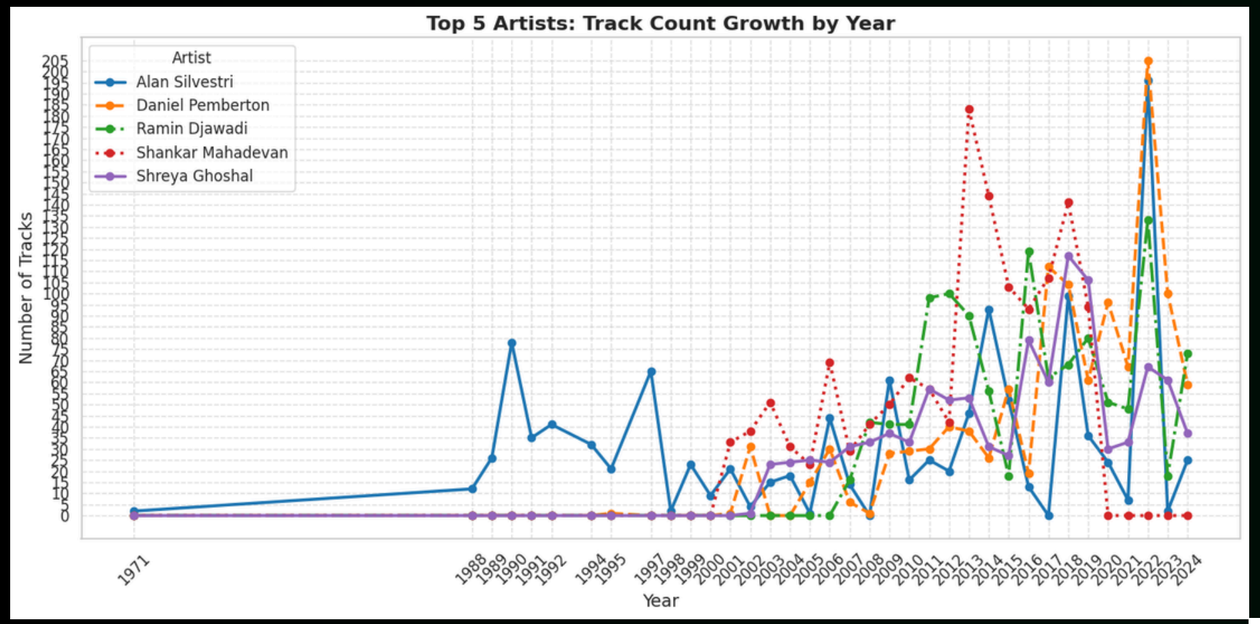
Top Languages by Track Count

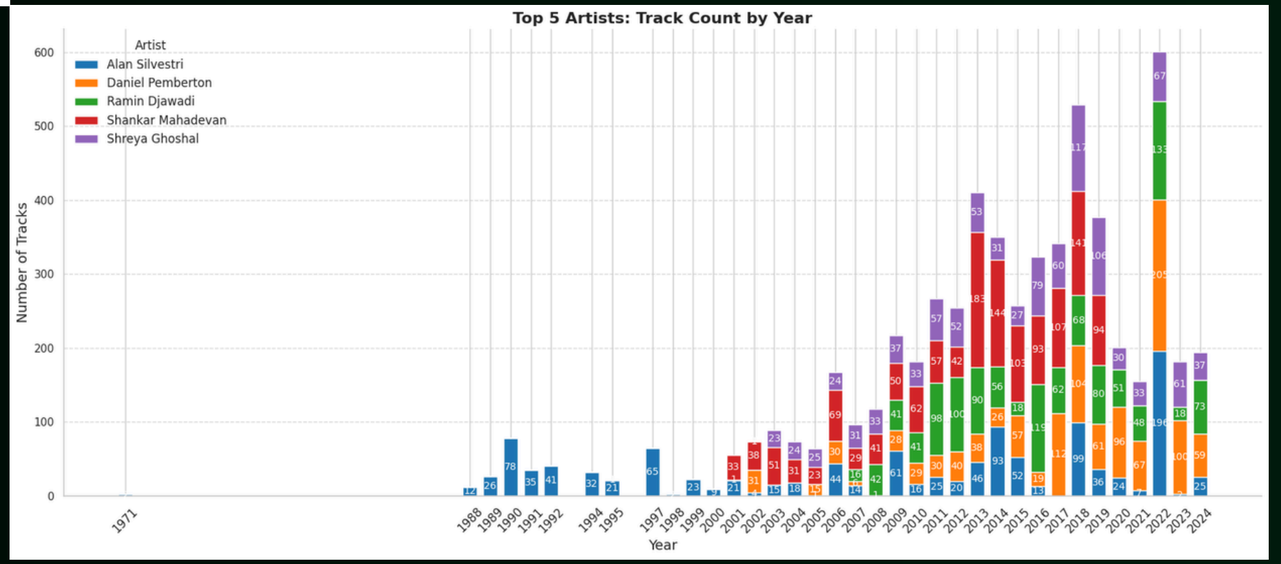




The chart clearly indicates English is the dominant language by track count, vastly exceeding all others (over 22,500 tracks). Tamil and Unknown languages rank second and third, respectively, followed by Hindi and Korean. The remaining languages (Malayalam, Telugu, etc.) have significantly lower track counts, suggesting the dataset is heavily weighted toward English-language content, with a strong presence from South Indian languages and Korean.

Top 5 Artists Track Count by Year





Interpretation

- Volatility and Peaks: Track releases are highly volatile for most artists, marked by significant annual spikes followed by sharp drops.
- Daniel Pemberton shows the most extreme peak, releasing over 200 tracks in 2023, far exceeding any other single-year count.
- Shankar Mahadevan also exhibits high volatility, with notable peaks around 2014 and 2016-2017.
- Composer Activity: Soundtrack composers like Alan Silvestri, Daniel Pemberton, and Ramin Djawadi show intermittent bursts of high activity, often corresponding to major film or TV releases.
- Ramin Djawadi shows a sustained increase in high-track years from around 2010 to 2017.
- Shreya Ghoshal's Consistency: Shreya Ghoshal displays the most relatively consistent track output in the later years (post-2010), maintaining a steady, moderate level of releases compared to the dramatic spikes seen by others.
- Early Career: Alan Silvestri has the longest recorded history, showing intermittent track counts dating back to 1988-1990, while Shreya Ghoshal's output begins later, showing consistent low-level activity from the early 2000s.

Time Series Analysis

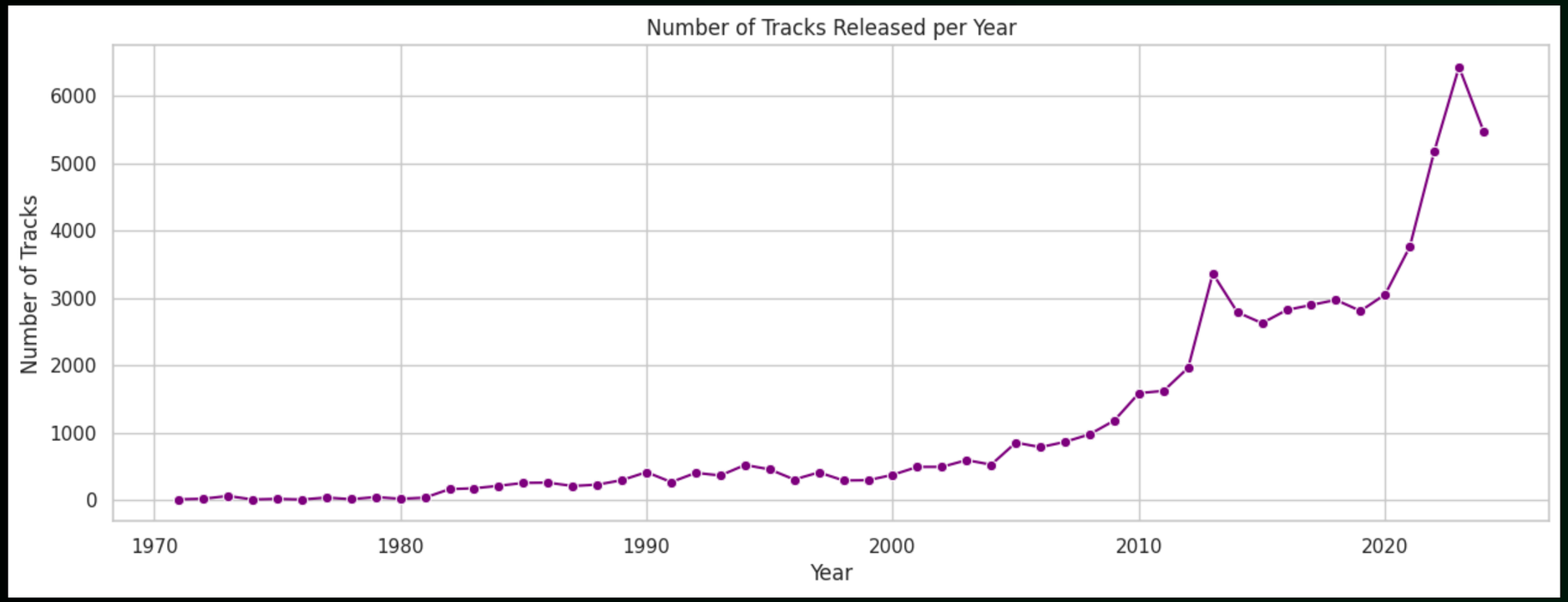
Learn More

01

02

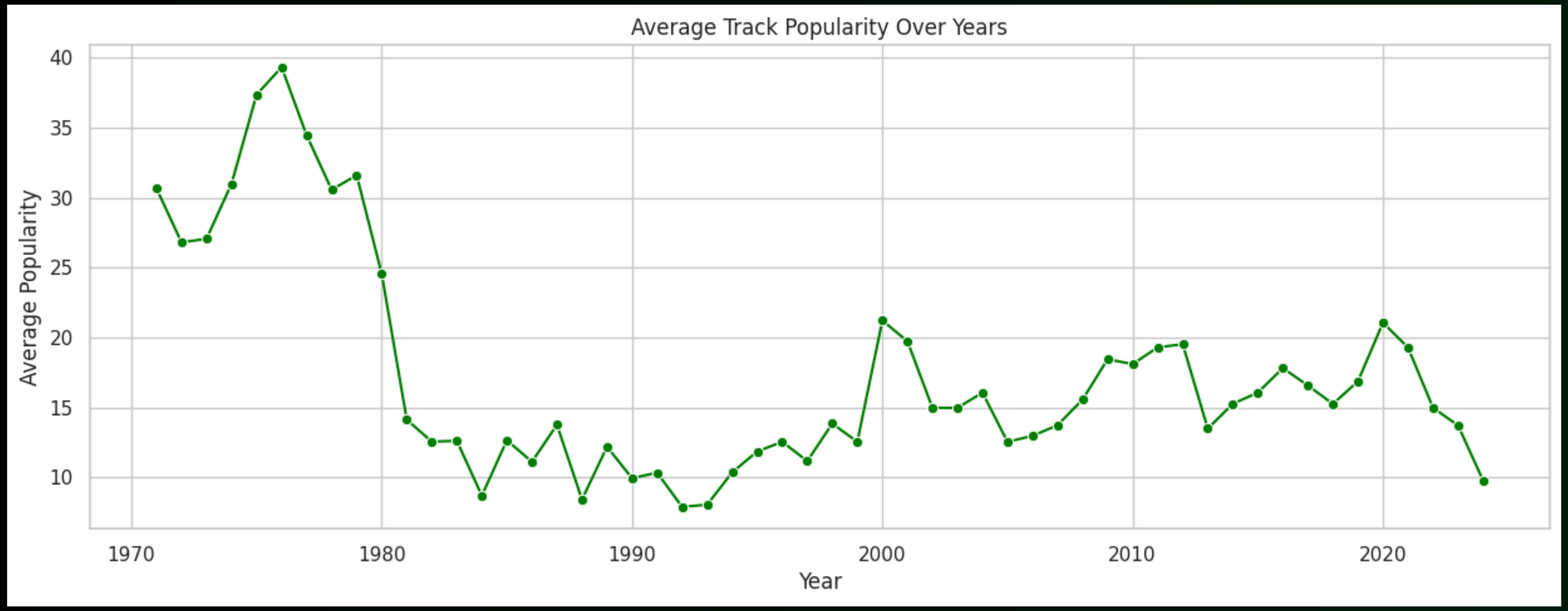
03

Total Number of Tracks Released per Year



- **The data shows a trajectory of explosive growth in the number of tracks released. Track counts remained relatively flat and low (under 1,000) from the 1970s until around 2007. A sharp increase began around 2010, and while there were minor fluctuations, the number of tracks released annually surged dramatically, peaking at over 6,000 tracks in the most recent years (around 2023). This highlights a significant and recent escalation in overall track production or recording/cataloging activity within the dataset.**

Average Track Popularity over the years



The chart shows extreme volatility in average track popularity. It peaked around 1978 (near 40) before undergoing a dramatic sharp decline in the early 1980s. Popularity then remained low and stable (mostly 8–15) until the late 1990s. From 2000 onward, there's been a moderate, fluctuating recovery (mostly 15–20), with a noticeable recent decline after a small peak around 2020. Overall, the average popularity of tracks in the dataset is significantly lower in the modern era than in the late 1970s.

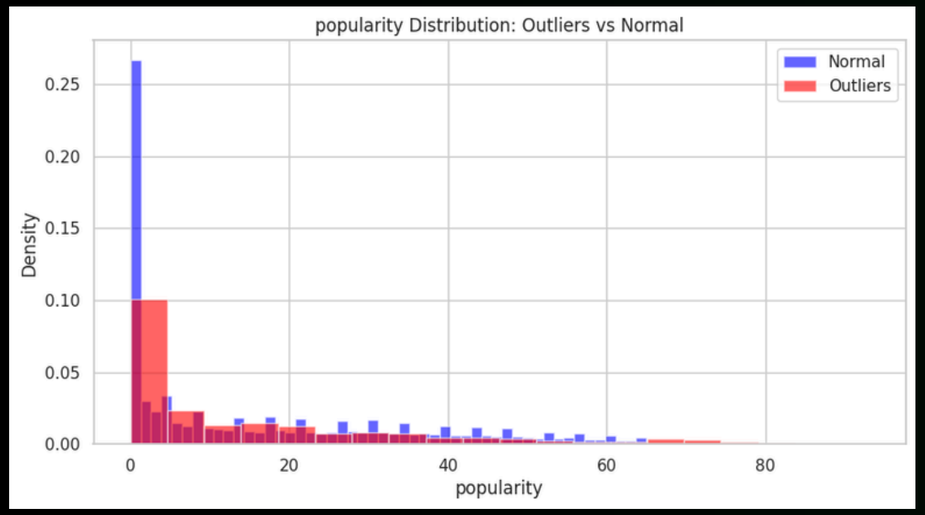
Outlier Analysis

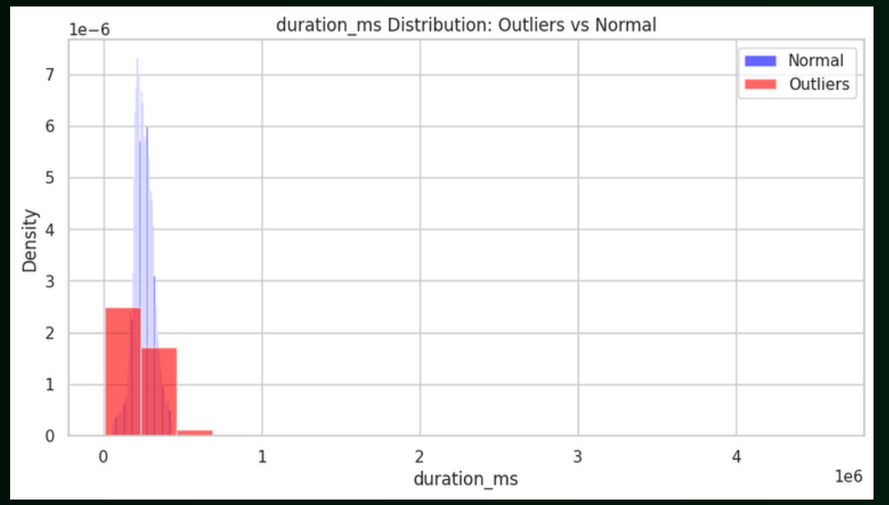
Learn More

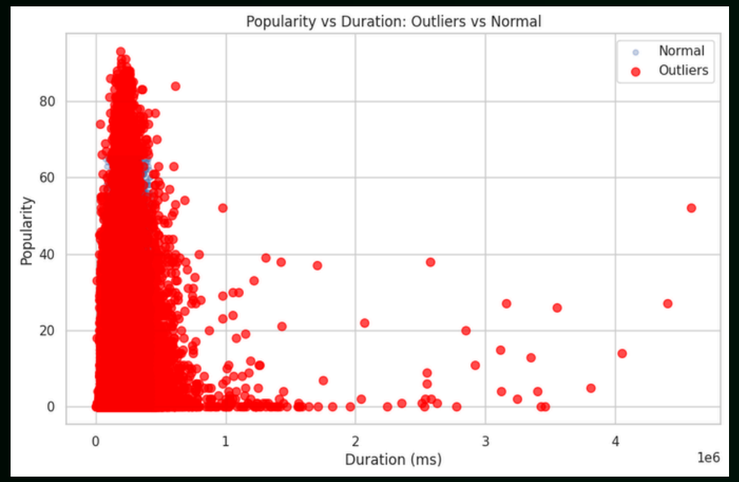
01

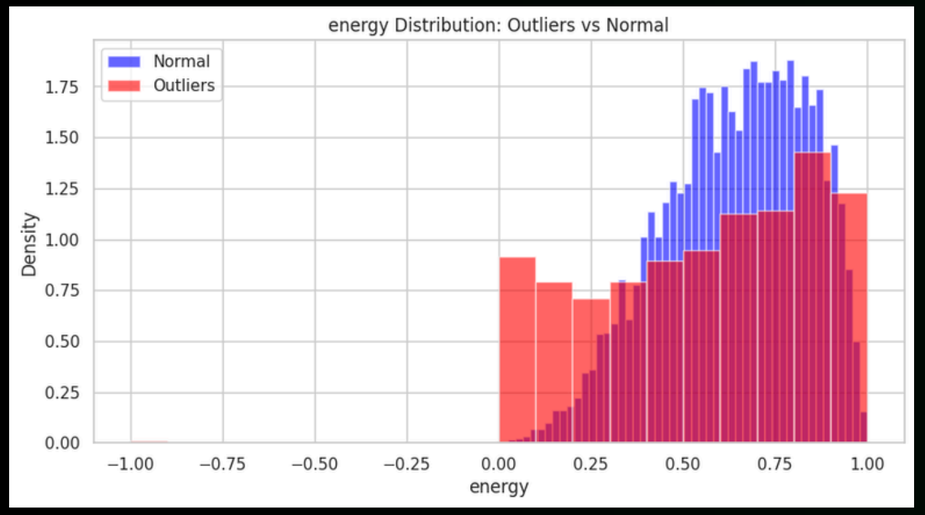
02

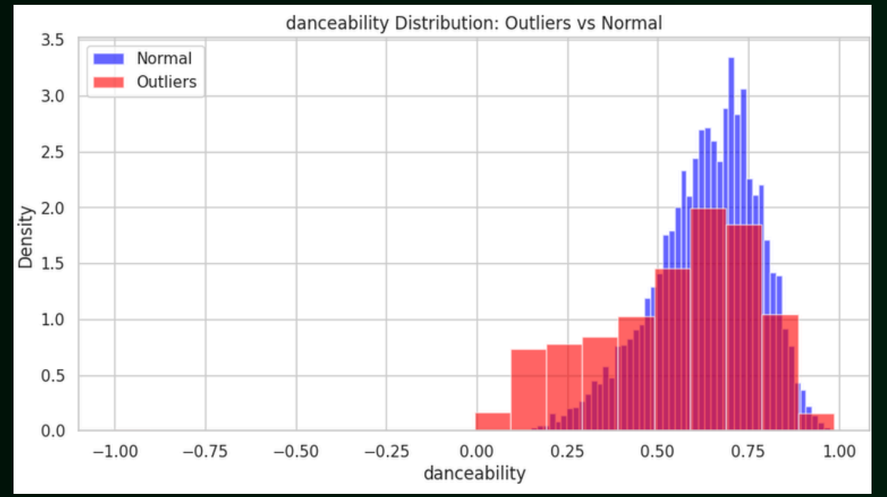
03


popularity Distribution: Outliers vs Normal
Normal
Outliers
Density
popularity


duration_ms Distribution: Outliers vs Normal
1e−6
Normal
Outliers
Density
duration_ms
1e6


Popularity vs Duration: Outliers vs Normal
Normal
Outliers
Popularity
Duration (ms)
1e6

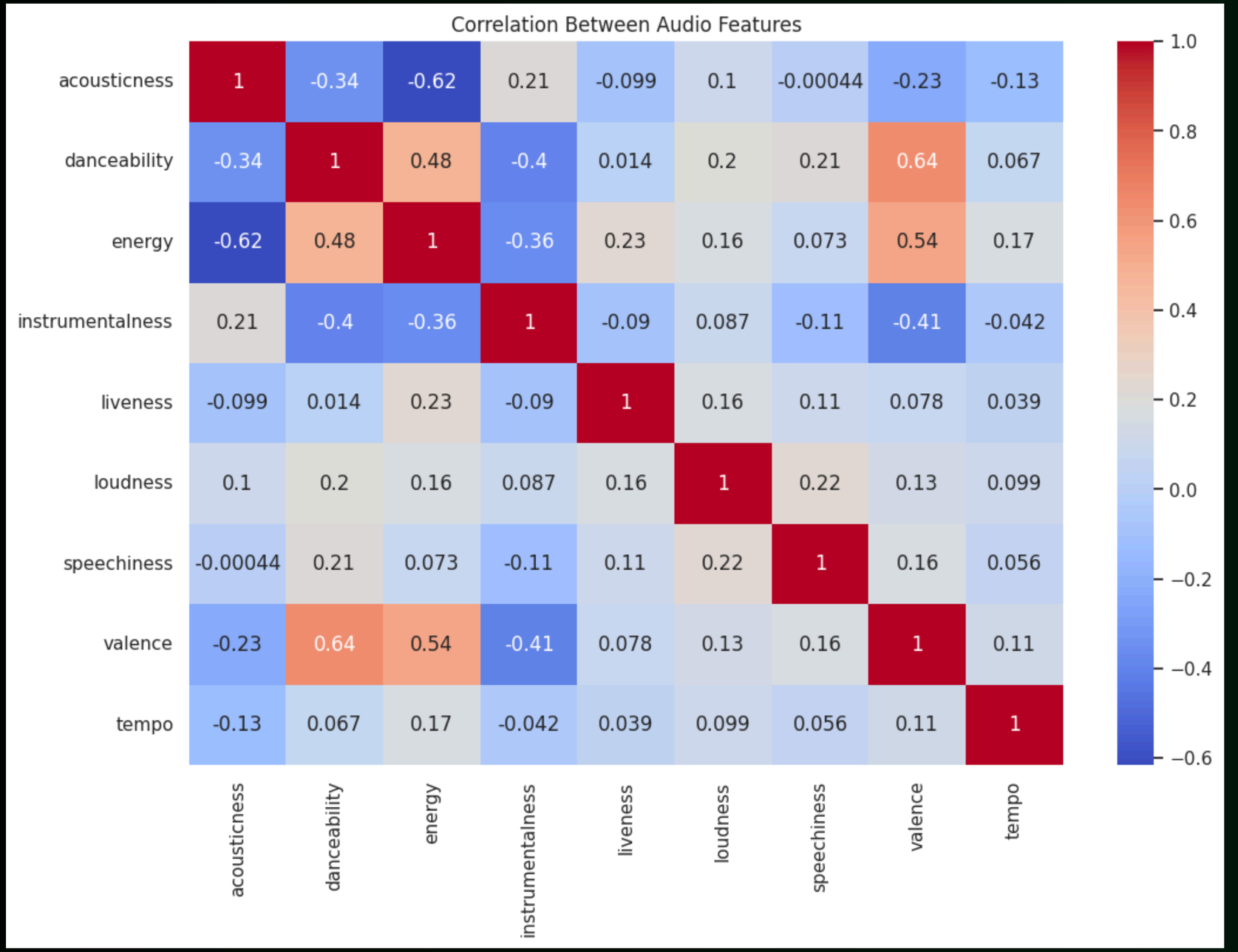

energy Distribution: Outliers vs Normal
Normal
Outliers
Density
energy


danceability Distribution: Outliers vs Normal
Normal
Outliers
Density
danceability

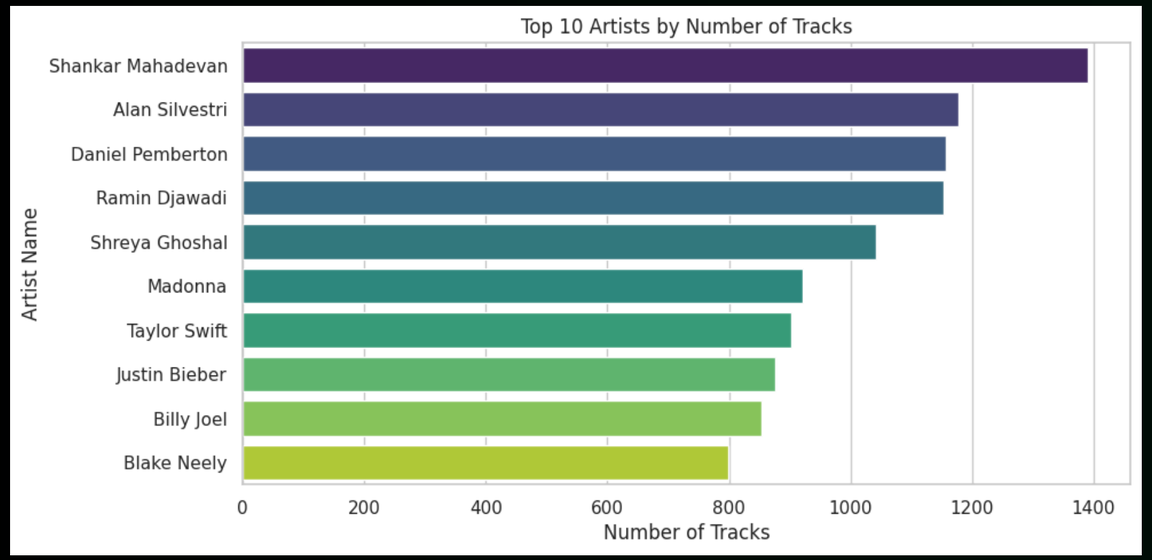
Interpretation

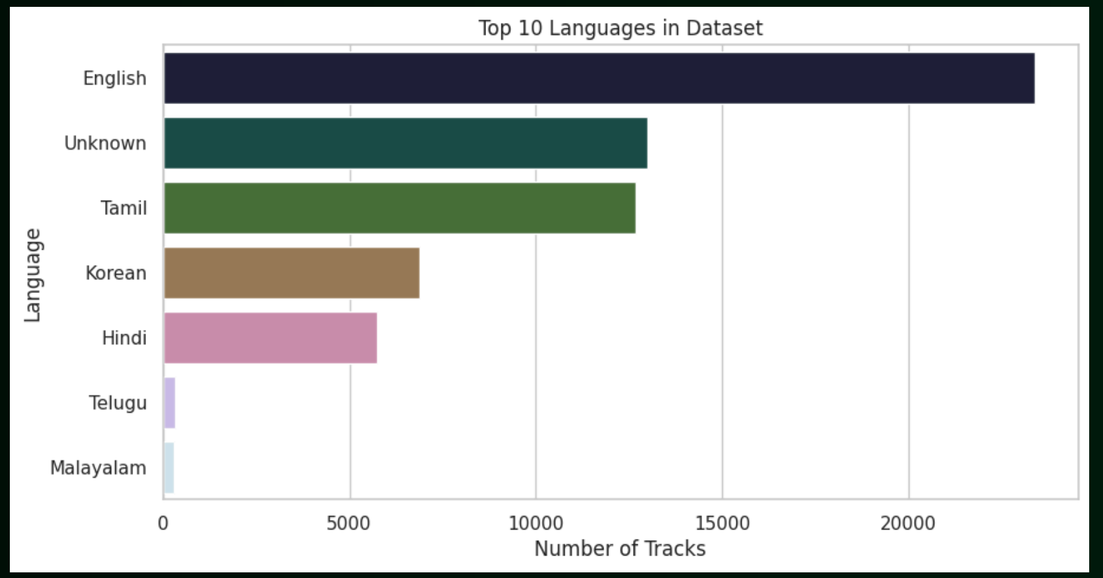
The Outlier classification effectively isolates tracks that significantly deviate from this popular but undistinguished core in three key dimensions:

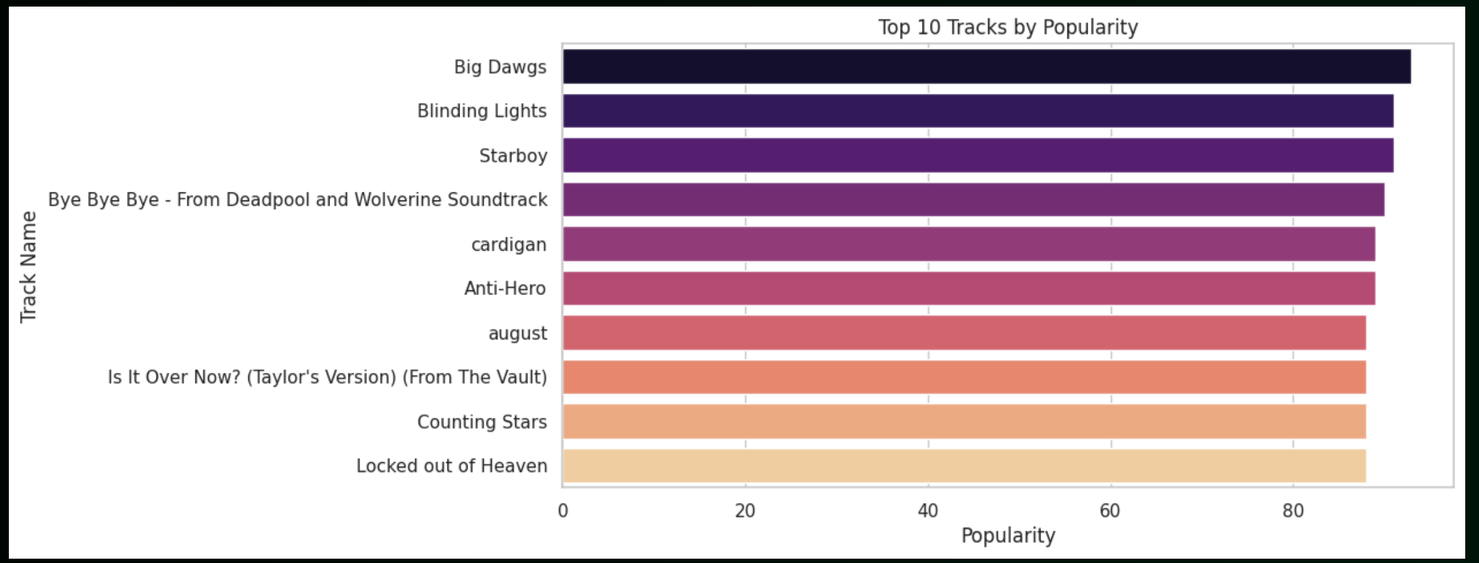
- Commercial Extremes (Popularity): Tracks with uncaracteristically high popularity (scores up to ≈90) are classified as outliers, regardless of their length or feature scores.
- Structural Extremes (Duration): Tracks with uncaracteristically long durations (up to 75+ minutes) are classified as outliers, representing a clear structural deviation from the norm.
- Feature Extremes (Energy/Mood): Tracks with uncaracteristically low danceability or low energy scores are flagged as outliers, suggesting the model isolates songs that are exceptionally subdued or difficult to dance to, departing from the dataset's high-energy standard.


Correlation Between Audio Features
acousticness 1 -0.34 -0.62 0.21 -0.099 0.1 -0.00044 -0.23 -0.13
danceability -0.34 1 0.48 -0.4 0.014 0.2 0.21 0.64 0.067
energy -0.62 0.48 1 -0.36 0.23 0.16 0.073 0.54 0.17
instrumentalness 0.21 -0.4 -0.36 1 -0.09 0.087 -0.11 -0.41 -0.042
liveness -0.099 0.014 0.23 -0.09 1 0.16 0.11 0.078 0.039
loudness 0.1 0.2 0.16 0.087 0.16 1 0.22 0.13 0.099
speechiness -0.00044 0.21 0.073 -0.11 0.11 0.22 1 0.16 0.056
valence -0.23 0.64 0.54 -0.41 0.078 0.13 0.16 1 0.11
tempo -0.13 0.067 0.17 -0.042 0.039 0.099 0.056 0.11 1
acousticness
danceability
energy
instrumentalness
liveness
loudness
speechiness
valence
tempo
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

Presenting The Top 10







Strategic Playbook: Data-Driven Music Production

Pillar 1: The Sonic Sweet Spot (Maximizing Engagement)

Insight

Popular tracks consistently exhibit high Danceability (>0.7) and Energy (>0.8).

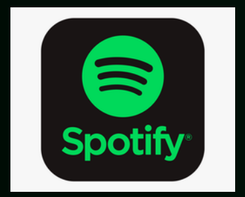
Recommendation

Focus on an "Energetic & Danceable" Profile. Engineer tracks to maximize both features, aligning your sound with the statistical profile of commercial success.

Pillar 2: The Modern Mix (Efficiency & Impact)

Insight

Average loudness has steadily increased while track duration has decreased. Modern music is louder and more concise.



Recommendation

Master for a Modern, Impactful Mix. Target a competitively loud mix. Use a shorter, efficient song structure to boost listener engagement and streaming completion rates.

Pillar 3: Global Reach (Market Expansion)

Insight

Non-English languages, particularly Spanish and Korean, show a significant, rapid increase in popularity over the last decade.

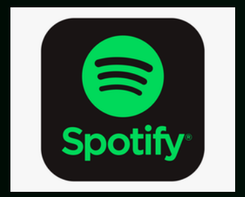
Recommendation

Leverage Shifting Language Trends. Explore collaborations with artists in these ascending markets or develop multilingual versions of key tracks.

Pillar 4: Niche Clarity (Feature Specialization)

Insight

Speechiness is highly bimodal (tracks are either highly lyrical or purely instrumental).



Recommendation

Target a Clear Niche: Lyrical vs. Instrumental. For instrumental tracks, focus production efforts on maximizing secondary features like Energy and Valence to drive popularity within that segment.

Thank You

For Watching this Presentation

~ By Ritav Paul

01

02

03