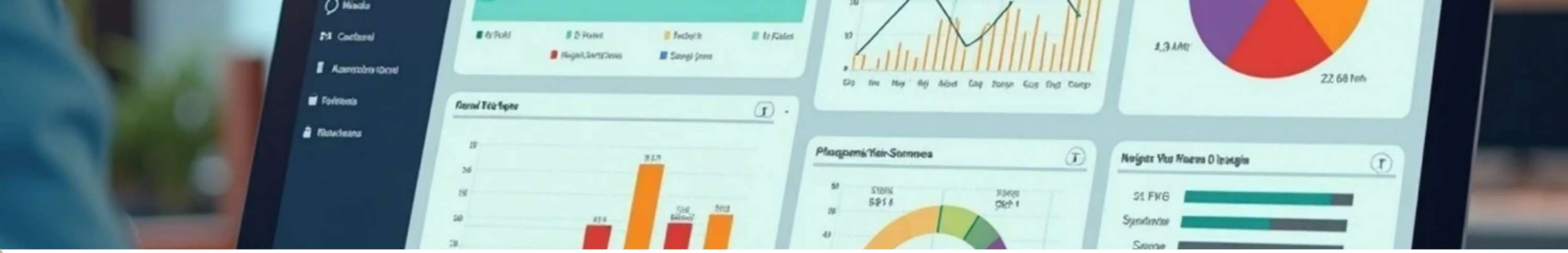
Analysis of Spotify Tracks Dataset

Presented by : DS Explorers

EDA Project

This project delves into a thorough analysis of the Spotify Tracks Dataset, aiming to uncover trends, relationships, and patterns within music features. By examining these insights, we strive to provide valuable information that can guide mixing engineers in their craft, ultimately enhancing the quality and appeal of music production.



Overview

This project analyzes the Spotify Tracks Dataset to :

- Understand patterns in song features
- Explore relationships between variables
- Identify trends across time
- Provide insights for mixing engineers

Project Roadmap Overview

1

Data Collection & Preprocessing

Initial steps to gather and clean the dataset.

2

Exploratory Data Analysis (EDA)

Analyzing data distributions and relationships for Univariate Bivariate & Multivariate analysis

3

Time Series Trends

Observing changes over the years.

4

Elite Performances

Identifying the leading artists and tracks.

5

Recommendations & Insights

Providing actionable suggestions.

Dataset & Preprocessing

The dataset is comprised of various track features such as duration, popularity, and musical attributes. A thorough data cleaning process was executed to eliminate missing values and duplicates, alongside necessary data type conversions. Additionally, new features were generated to enhance analysis capabilities.

Removing Unknowns

```
Number of rows in the dataset at initial stage: 62317

Duplicate entries: 78
```

```
Removed 78 duplicate entries
Number of rows in the dataset after dropping duplicates: 62239
```

Dataset size :

Number of rows : 49234
Number of columns : 22

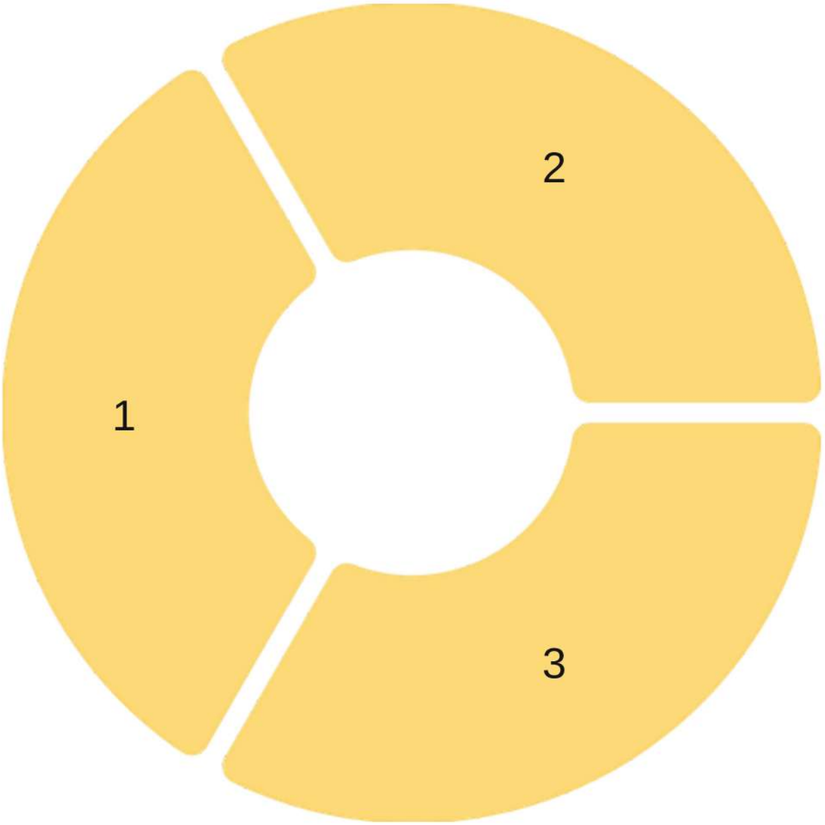
Deriving New Features

Starting year: 1971
End year: 2024

Univariate Analysis (Numerical)

Distributions of Key Variables

Examined distributions of numerical variables such as popularity, danceability, energy, acousticness, tempo, and duration.



Popularity Insights

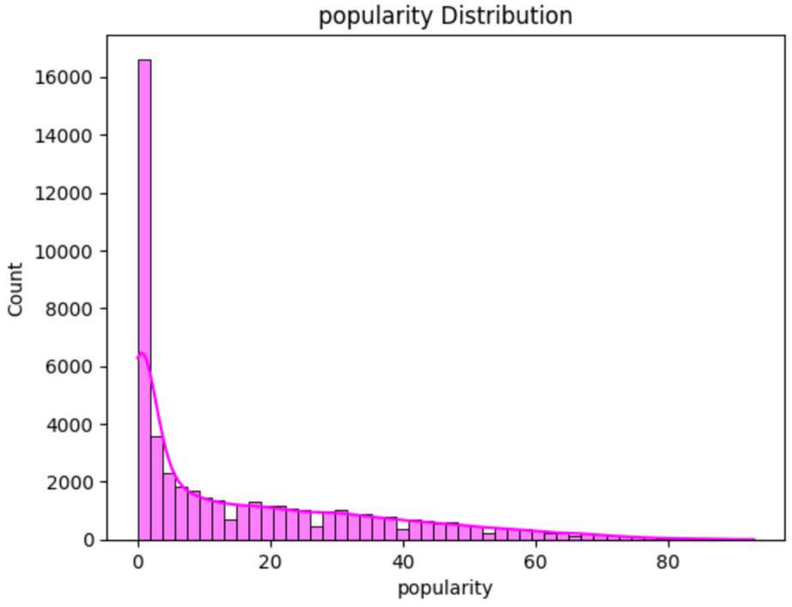
Insights reveal that most tracks exhibit moderate popularity.

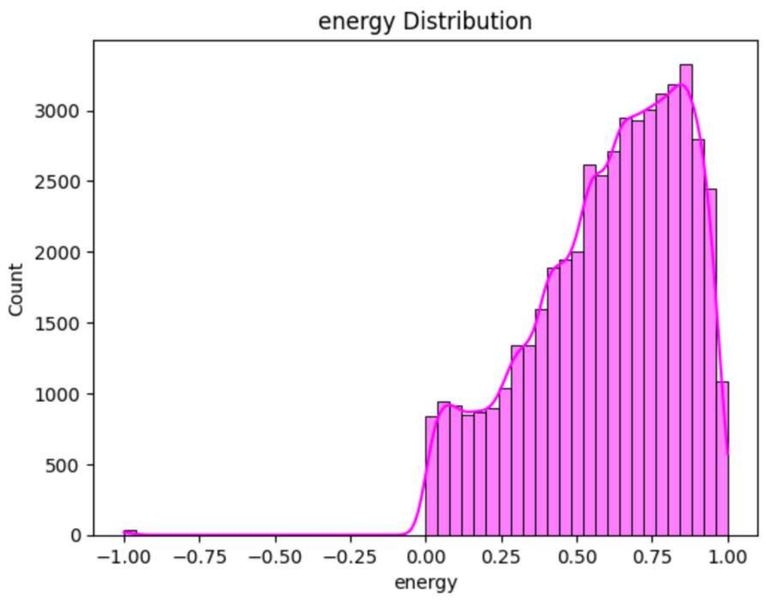
Danceability and Energy Metrics

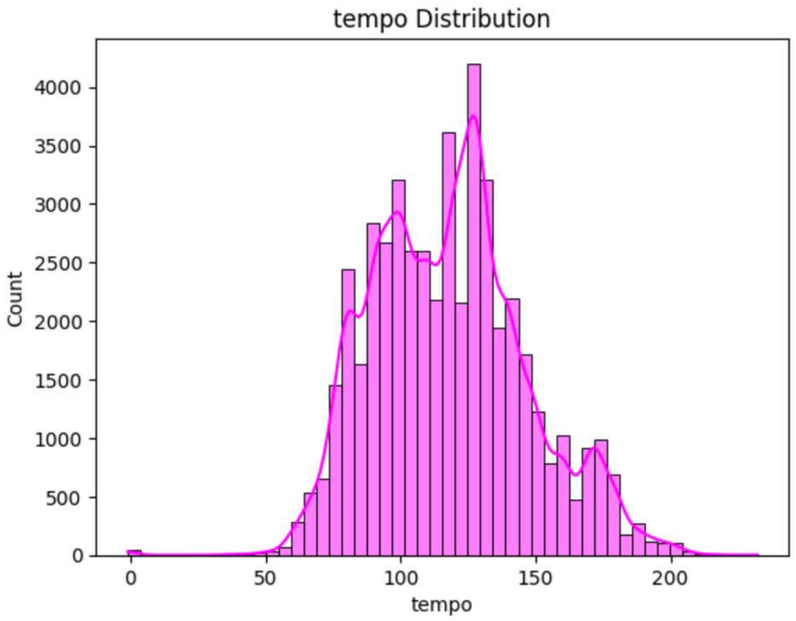
Metrics are positively skewed, with a greater prevalence of energetic tracks compared to acoustic ones.

EDA- Univariate Analysis of Numerical Variables

An initial check of distributions across basic track features.







Key Insights :
Most tracks have moderate popularity
Energy is positively skewed

EDA- Univariate Analysis of Numerical Variables

```
1. Analysing popularity
Central Tendency of popularity:
   Mean popularity: 16.18
   Median popularity: 8.00
   Mode popularity: [0]

   popularity Description:
count    49234.000000
mean        16.175651
std         19.077595
min          0.000000
25%          0.000000
50%          8.000000
75%         28.000000
max         93.000000
Name: popularity, dtype: float64

   Kurtosis of popularity: 0.5117
   Skewness of popularity: 1.1741
```

```
5. Analysing energy
Central Tendency of energy:
   Mean energy: 0.60
   Median energy: 0.65
   Mode energy: [0.65]

   energy Description:
count    49234.000000
mean         0.601245
std          0.252940
min         -1.000000
25%          0.434000
50%          0.645000
75%          0.807000
max          1.000000
Name: energy, dtype: float64

   Kurtosis of energy: 0.4354
   Skewness of energy: -0.7100
```

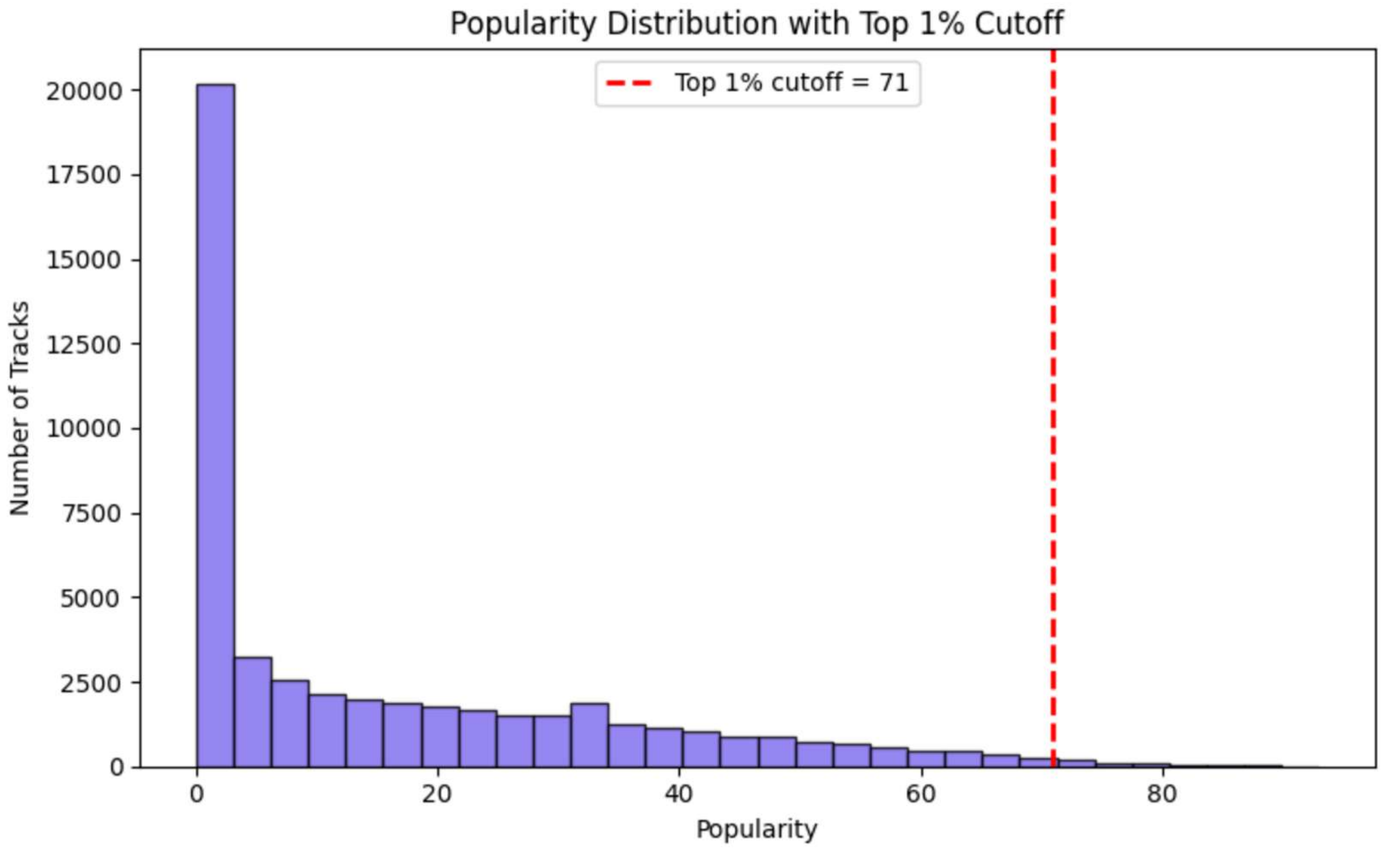
```
9. Analysing tempo
Central Tendency of tempo:
   Mean tempo: 117.99
   Median tempo: 118.17
   Mode tempo: [90.015]

   tempo Description:
count    49234.000000
mean       117.986141
std         28.596865
min         -1.000000
25%         95.999000
50%        118.167000
75%        135.026750
max        232.198000
Name: tempo, dtype: float64

   Kurtosis of tempo: 0.0038
   Skewness of tempo: 0.3107
```

EDA- Univariate Analysis of Numerical Variables

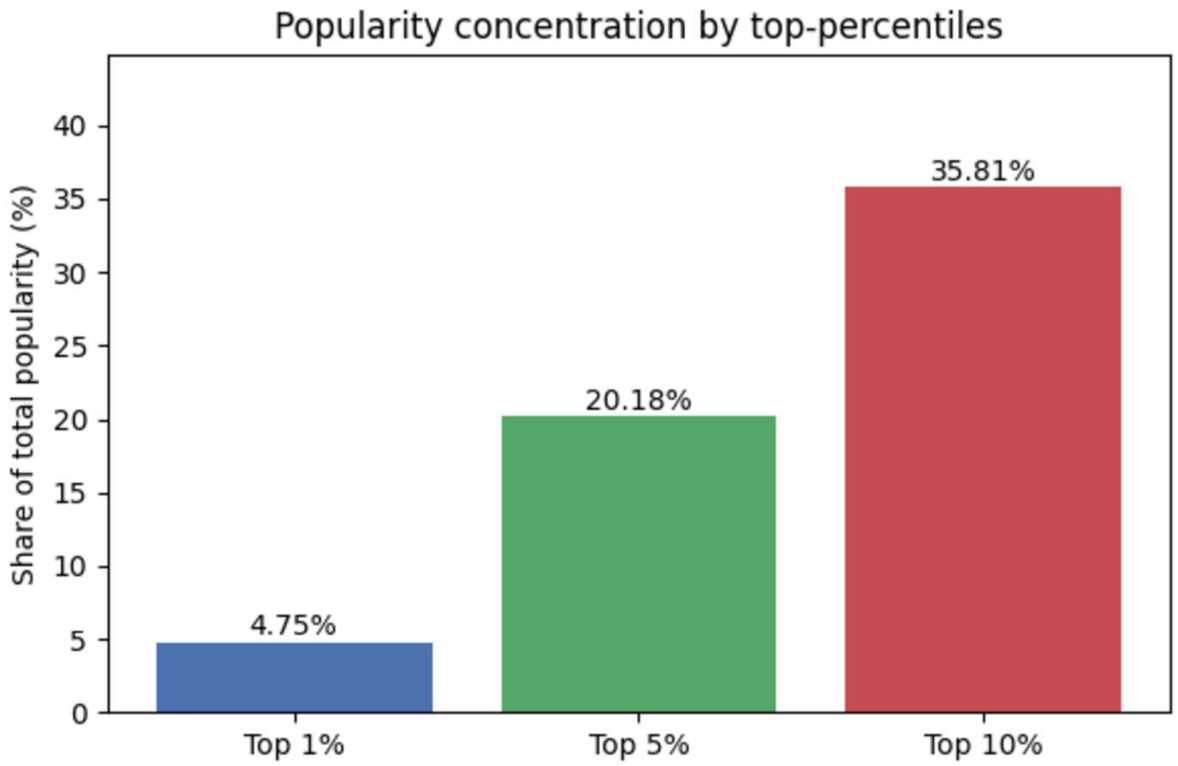
Popularity Insights



- **Most tracks have very low popularity scores**
- **Popularity distribution is highly right-skewed**
- **Top 1% of tracks start at a popularity score of 71**

EDA- Univariate Analysis of Numerical Variables

Popularity Insights



- **The top 1% holds only 4.75% of total popularity, showing popularity is not extremely concentrated.**
- **Popularity jumps dramatically from top 1% to top 5%, increasing from 4.75% to 20.18%.**
  The top 10% controls over one-third (35.81%) of all popularity.
- **Most popularity (64.19%) still belongs to the bottom 90% of the population.**

Univariate Analysis of Categorical Data

This presentation focuses on the univariate analysis of categorical variables, specifically examining the distributions of genres, keys, and modes in the dataset. Our observations reveal insightful trends about musical characteristics and their effects on listener perception.

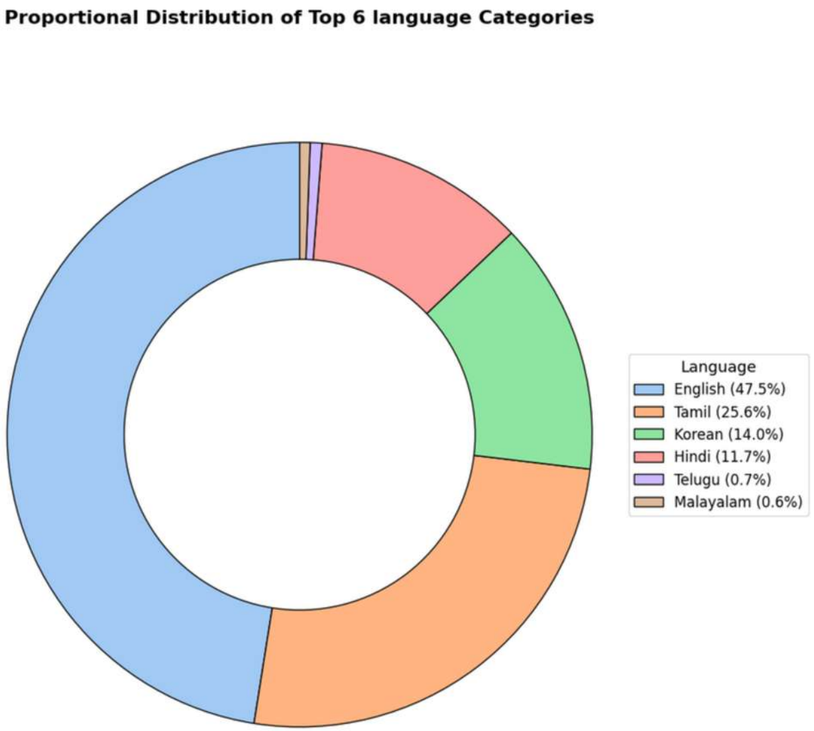
Visualization with Donut Chart
Proportional Distribution of Languages

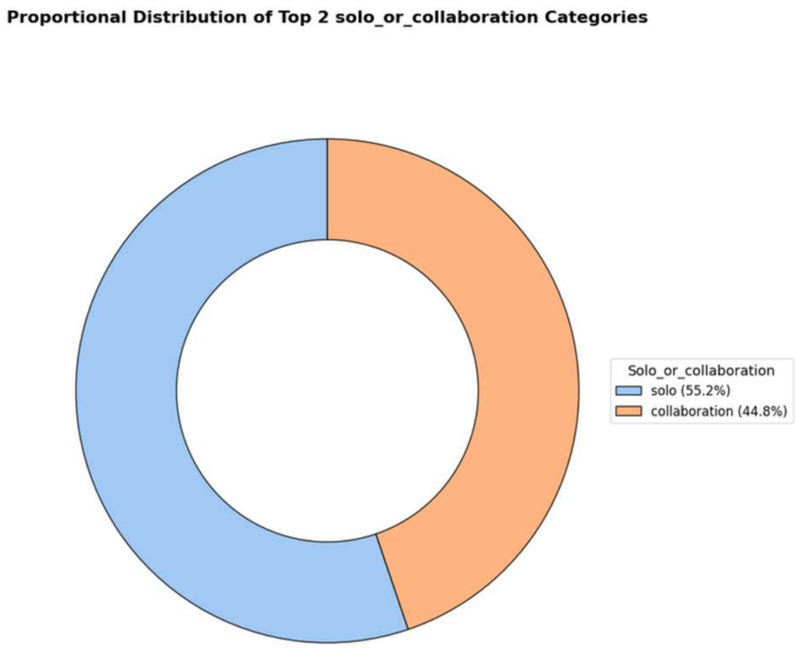
Visualization with Donut Chart
Proportional distribution of solos and collaborations

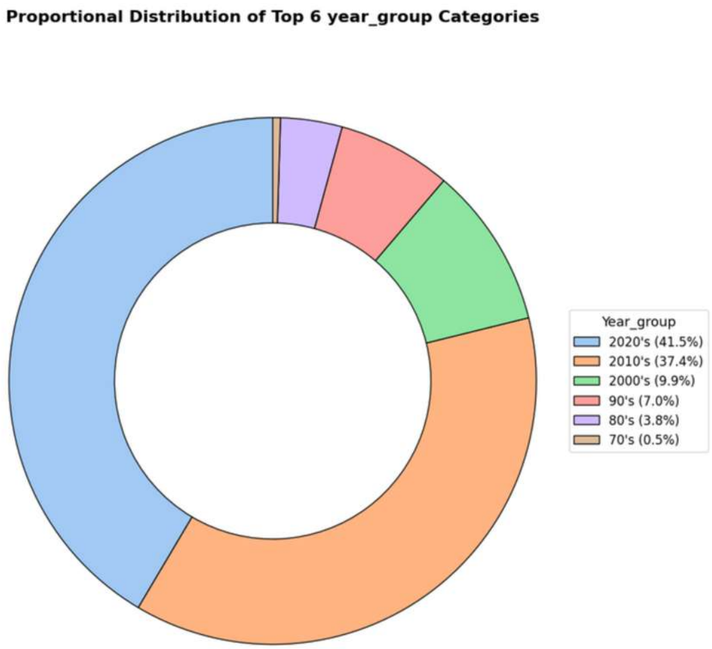
Visualization with Donut Chart
Proportional Distribution of Year Groups

EDA- Univariate Analysis of Categorical Data

Visualization with Donut Chart







Key Insight :
English dominates at 47.5% while Tamil takes second place at 25.6%, with the remaining four languages (Korean, Hindi, Telugu, Malayalam) collectively accounting for less than 27%

Key Insight :
Solo work slightly dominates at 55.2% while collaborations make up 44.8%, showing a relatively balanced split between individual and collaborative efforts.

Key Insight :
The 2020's and 2010's dominate with nearly 80% of the distribution, while older decades from the 70's-90's represent less than 12% combined.

Bivariate Analysis (Num vs Num)

This analysis explores the relationships between numerical variables to uncover key insights. A strong correlation is found between danceability and energy, indicating that energetic tracks are more likely to gain popularity. Additionally, tracks that are either too long or too short tend to have lower popularity ratings.

Analysis with Joint Plots

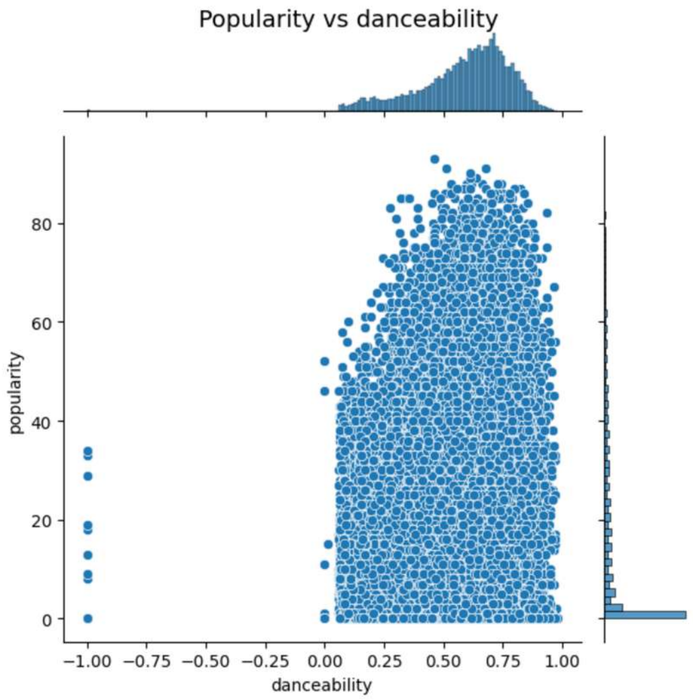
Popularity vs Duration , Danceability & Energy

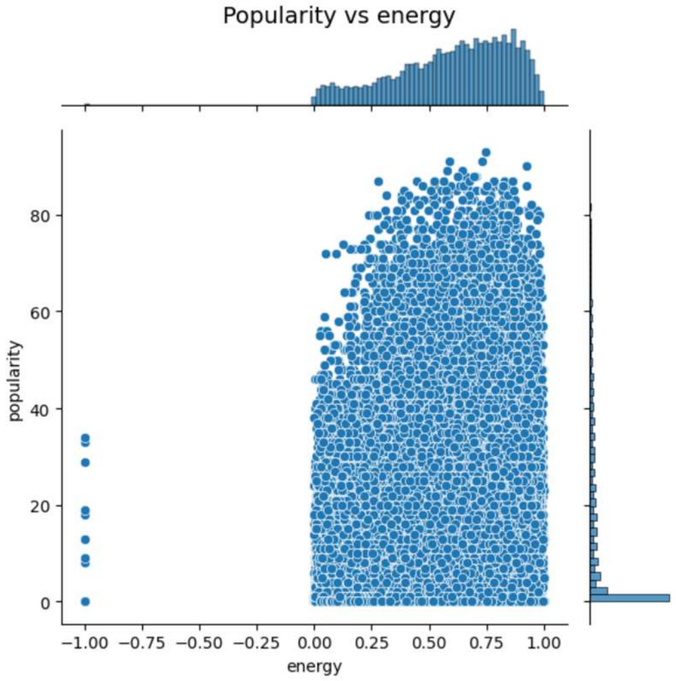
Visualisation with Heatmap

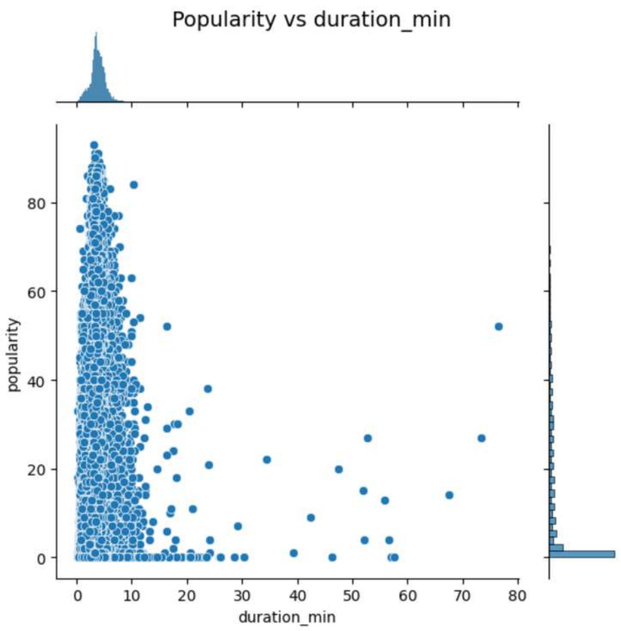
With all numerical features

Bivariate Analysis (Num vs Num)

Analysis with Joint Plots







Key Insights :

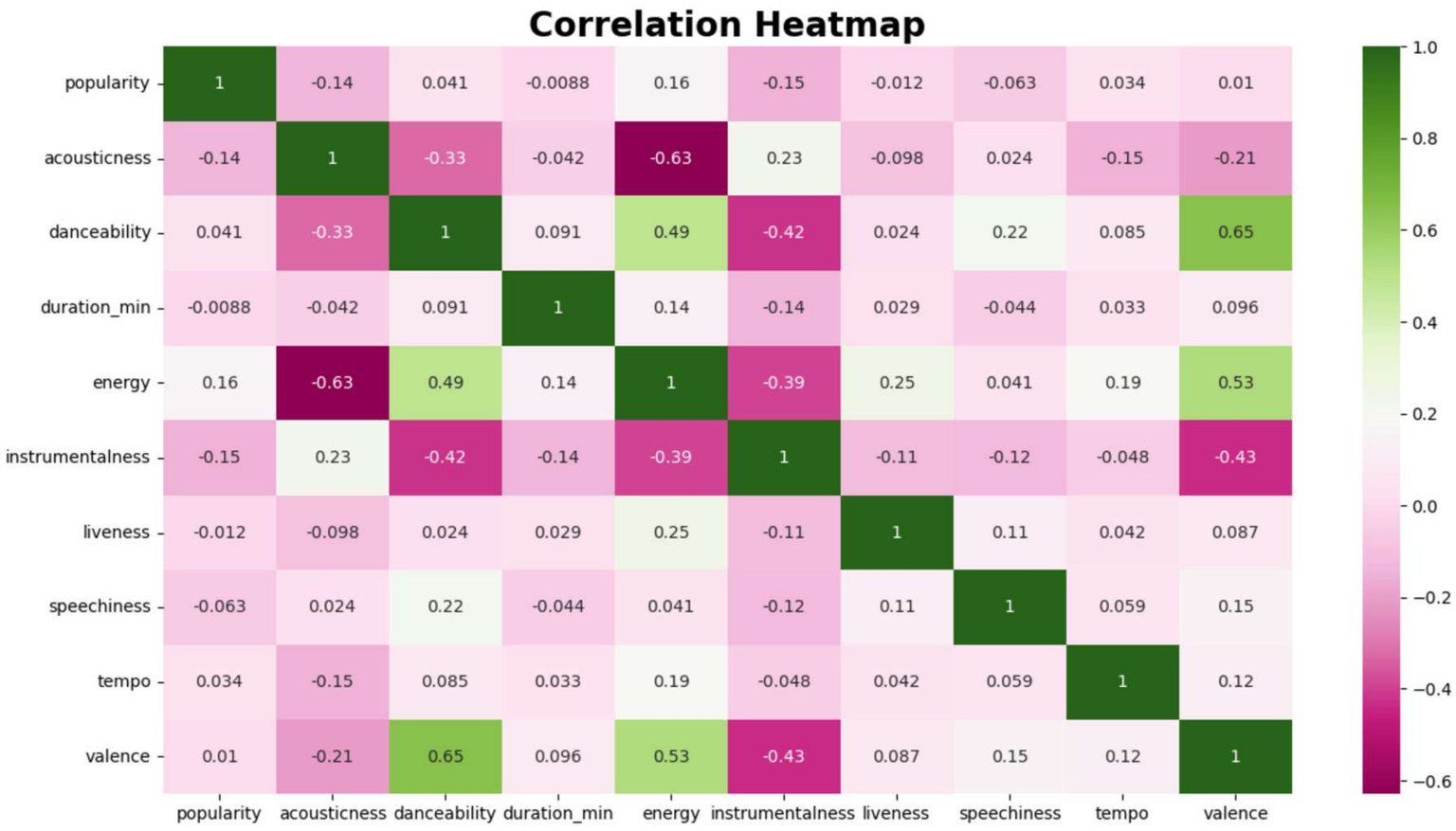
Duration vs Popularity: Strong negative correlation - shorter songs dominate popularity, with virtually no popular songs over 20 minutes.

Danceability vs Popularity: Weak positive correlation - most popular songs cluster in mid-to-high danceability range (0.3-0.8).

Energy vs Popularity: Weak positive correlation - popular songs require moderate-to-high energy levels with few low-energy hits.

Bivariate Analysis (Num vs Num)

Visualisation with Heatmap

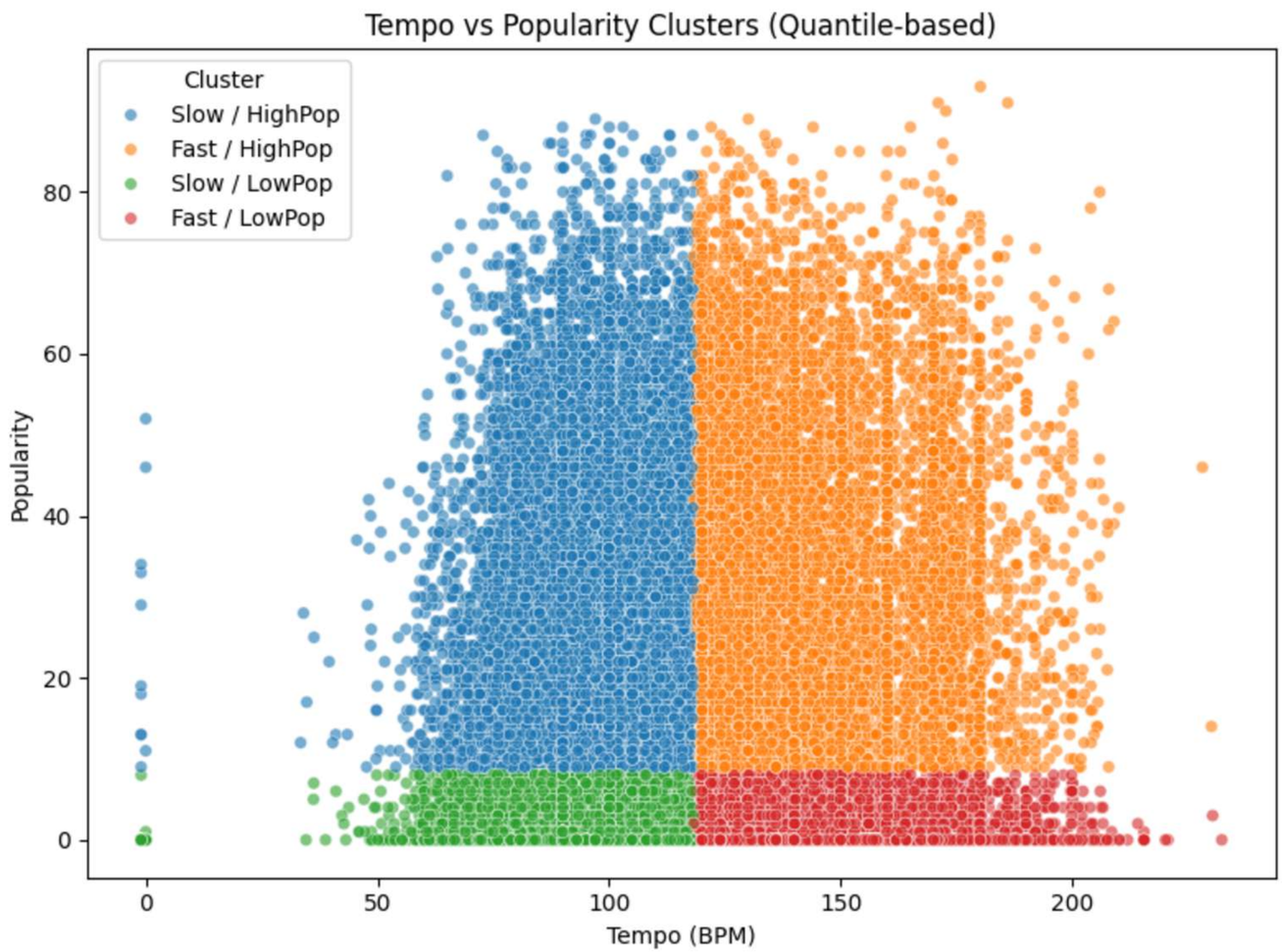


Key Insights :

- **Strong positive musical trait clusters:** Danceability, energy, and valence are strongly correlated (0.49-0.65).
- **Acoustic songs are typically calm:** Acousticness negatively correlates with energy (-0.63) and danceability (-0.33).
- **Popularity is weakly correlated:** All features show weak correlations with popularity (under ±0.2).

Bivariate Analysis (Num vs Num)

Scatter plot with categorical clustering

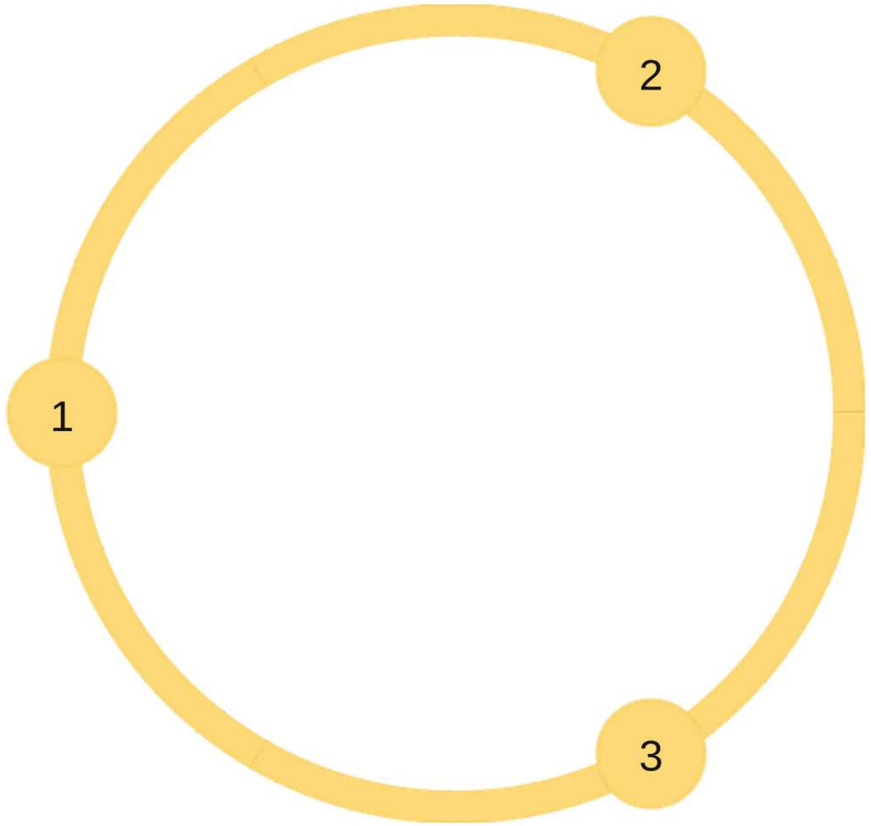


Key Insights :

- **Fast songs dominate high popularity:** Orange cluster (Fast/HighPop) shows the highest concentration of popular songs above 60 popularity score.
- Tempo range varies by popularity: High popularity songs span 80-200 BPM, while low popularity songs cluster in narrower tempo ranges.
- **Clear quadrant separation:** Four distinct clusters emerge based on tempo (slow vs fast) and popularity (high vs low) thresholds.

Bivariate Analysis: Numerical vs Categorical Variables

The bivariate analysis reveals interesting insights into how popularity varies with categorical variables . Our analysis shows significant differences in key features among music genres.



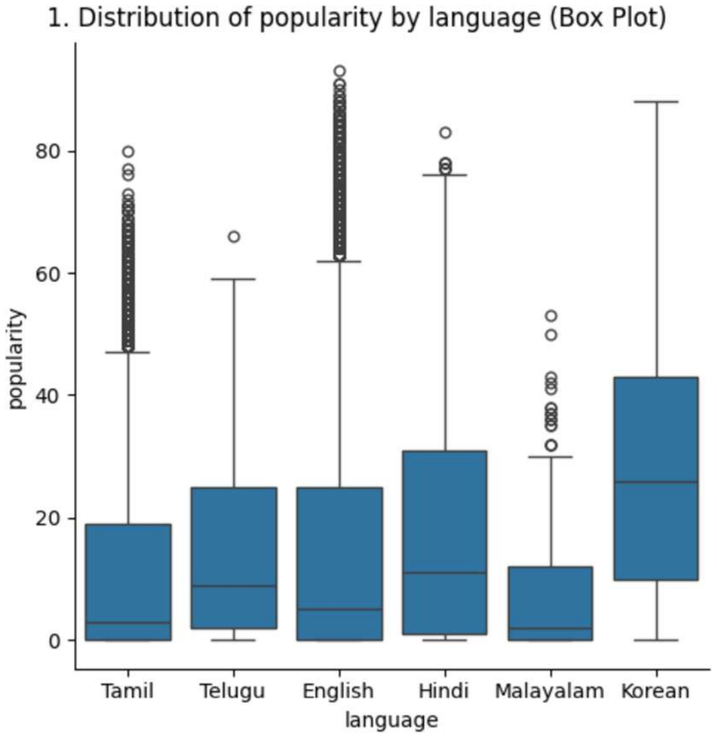
Visualisation with Box plots

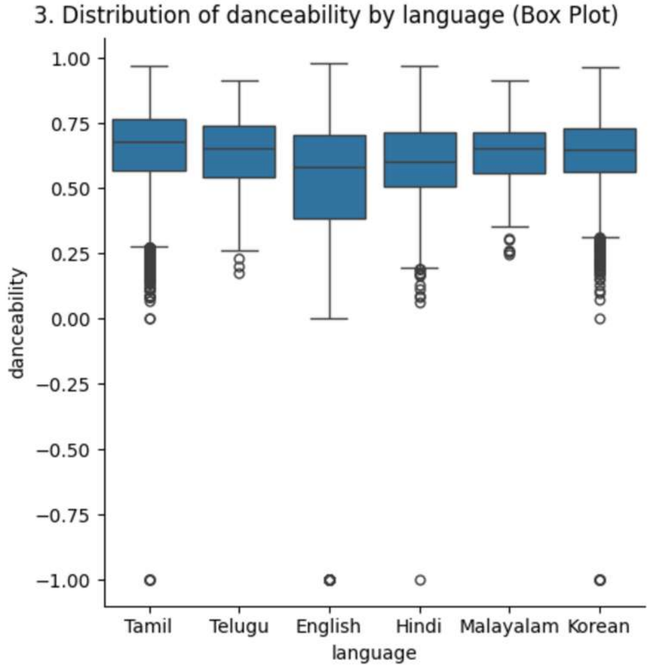
How Popularity Changes with Mode & Collaboration

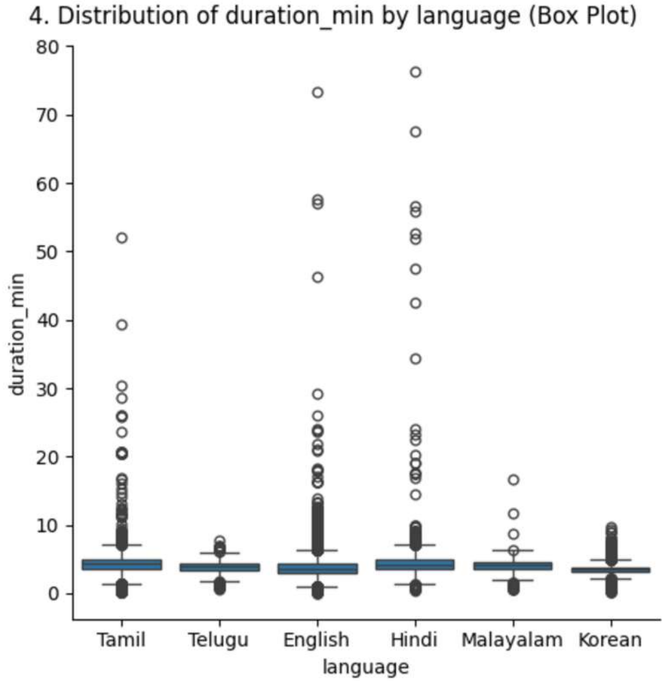
Average Popularity Distribution Between Sad And Happy Songs

Bivariate Analysis: Numerical vs Categorical Variables

Visualisation with Box plots







Key Insight:

1. Popularity by Language: **English shows highest median popularity (~62) with many outliers, while other languages have lower medians (8-30).**

Key Insight:

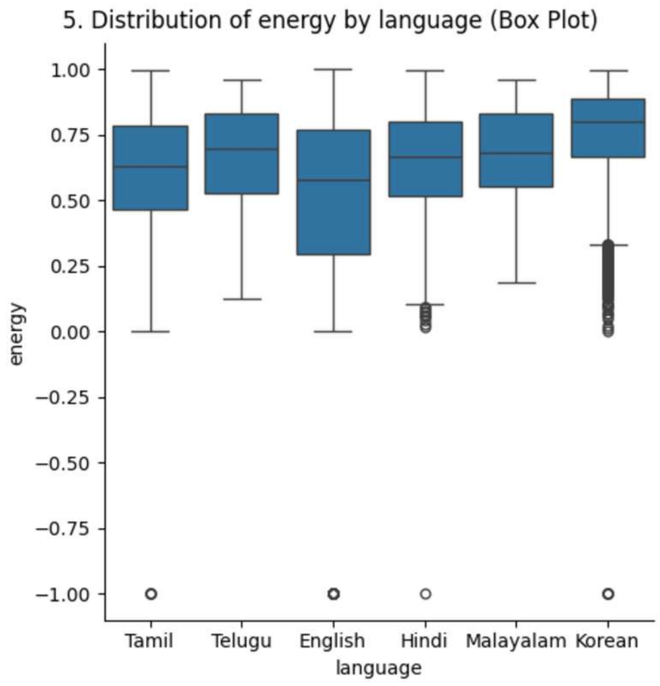
1. Danceability by Language: **All languages show similar danceability distributions with medians around 0.6-0.7, indicating consistent musical rhythm patterns.**

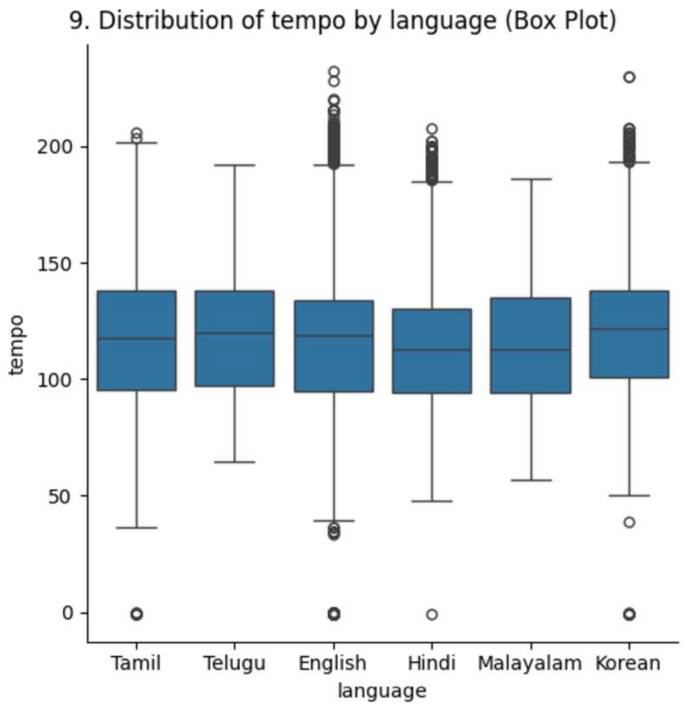
Key Insight:

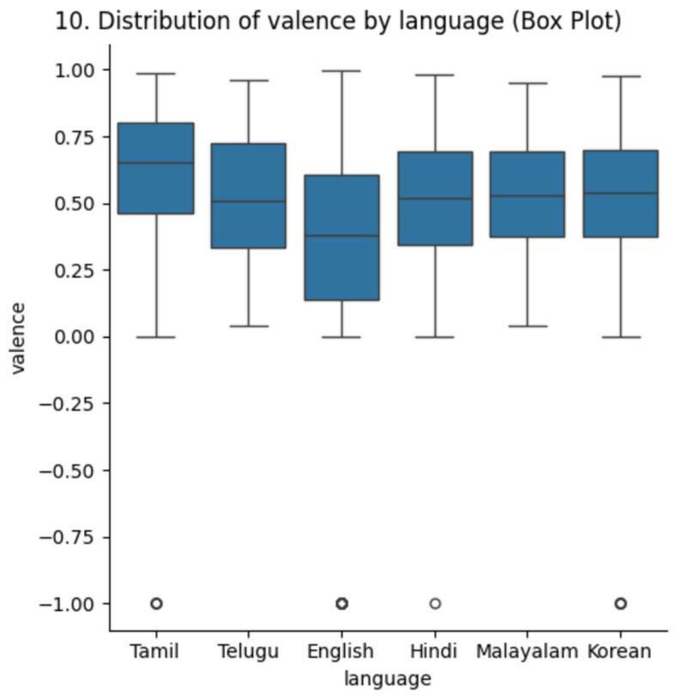
1. Duration by Language: **Most languages have similar short song durations (~3-5 minutes median), with English and Hindi showing more variation in longer songs.**

Bivariate Analysis: Numerical vs Categorical Variables

Visualisation with Box plots







Key Insight:

1. Energy by Language: **Korean shows highest median energy (~0.8), while English has widest variation with lower median (~0.5)**.

Key Insight:

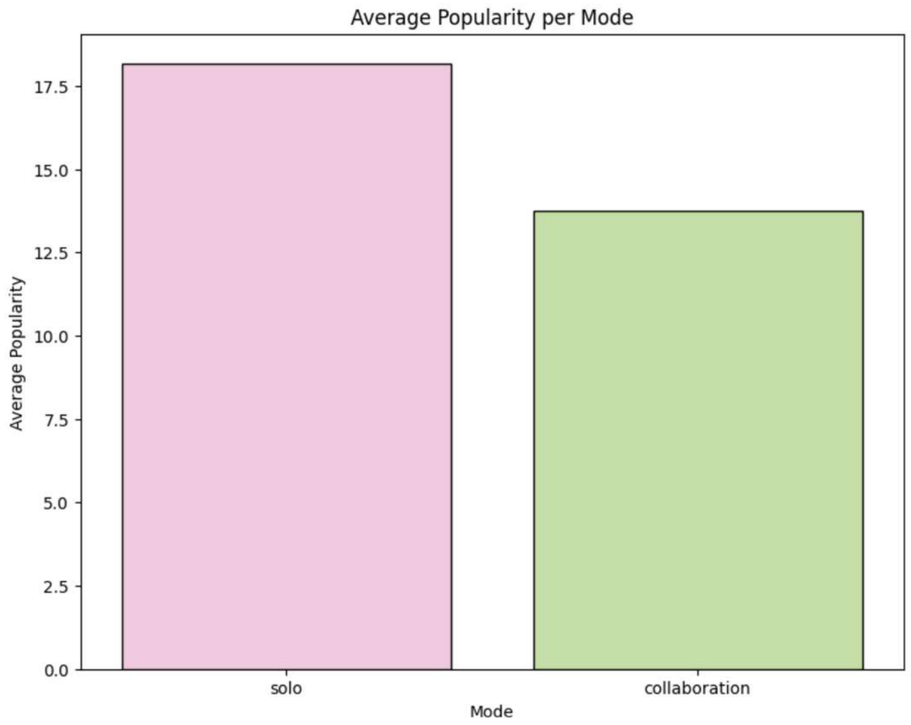
1. Tempo by Language: **All languages show similar tempo distributions (100-140 BPM median) with consistent cross-cultural rhythm preferences.**

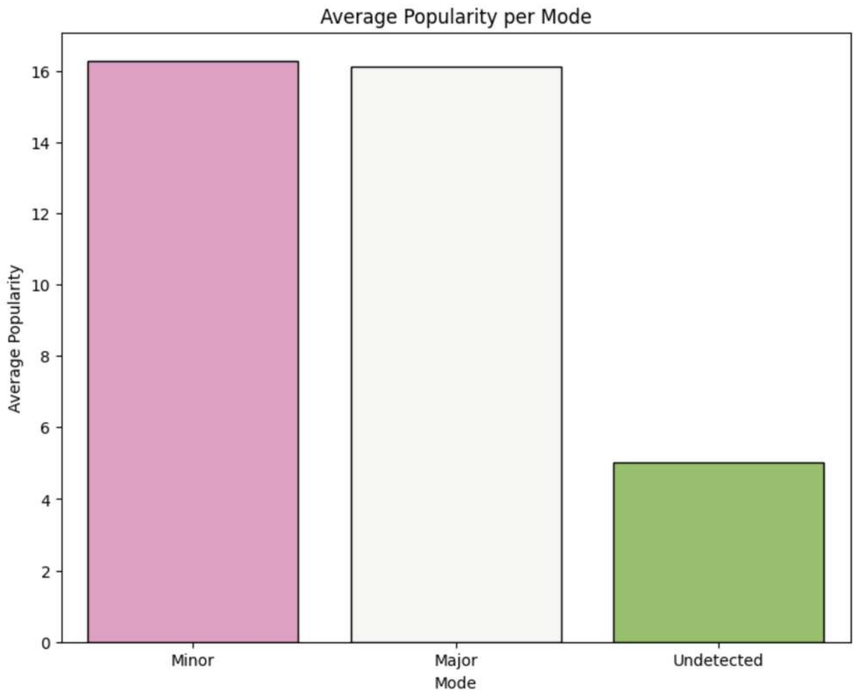
Key Insight:

1. Valence by Language: **Tamil shows highest positivity (median ~0.7), while English has lowest median valence (~0.4) indicating more negative emotional tone.**

Bivariate Analysis: Numerical vs Categorical Variables

How Popularity Changes with Mode & Collaboration



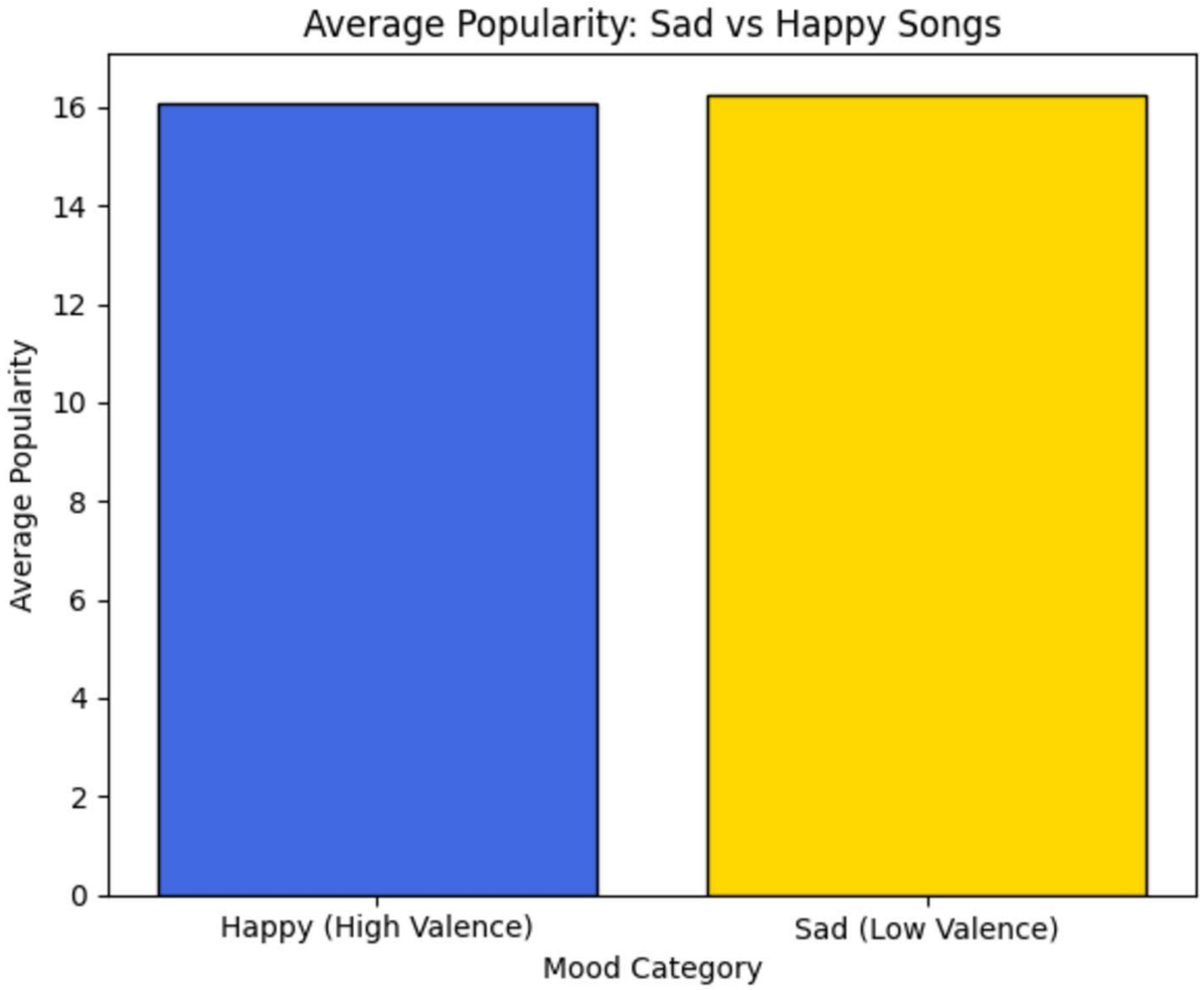


Key Insights:

- **Musical Mode Popularity: Minor and Major modes show nearly equal high popularity (~16), while Undetected mode has significantly lower popularity (~5).**
- **Solo vs Collaboration: Solo tracks achieve higher average popularity (~18) compared to collaboration tracks (~14), suggesting individual artists perform better commercially.**

Bivariate Analysis: Numerical vs Categorical Variables

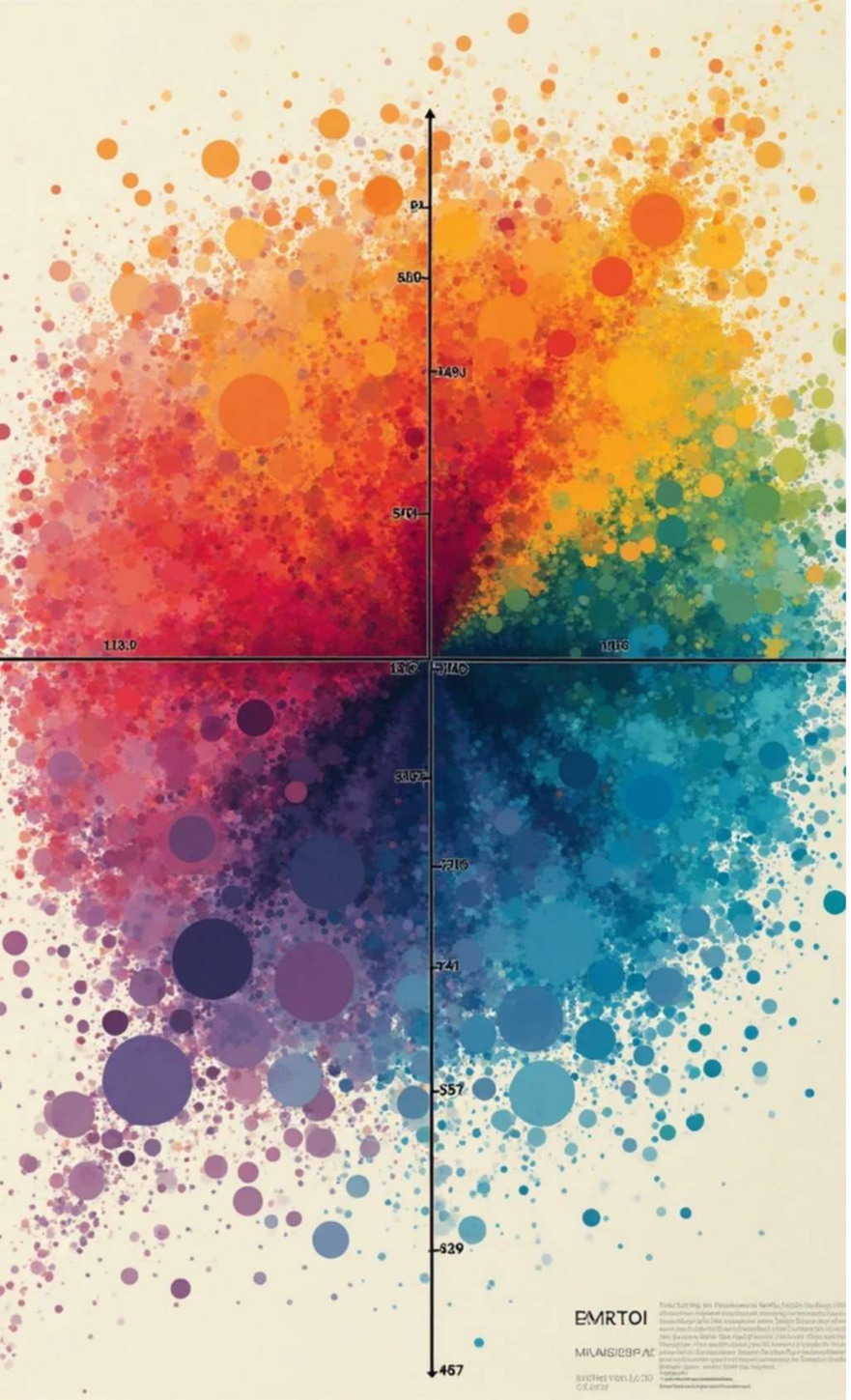
Average Popularity Distribution Between Sad And Happy Songs



Key Insight:

Emotional Valence vs Popularity: **Happy songs (high valence) and sad songs (low valence) show nearly identical average popularity (~16), indicating that emotional tone alone doesn't determine commercial success.**

	mood	popularity
0	Happy (High Valence)	16.087280
1	Sad (Low Valence)	16.262831



Multivariate Analysis Overview

Data Exploration

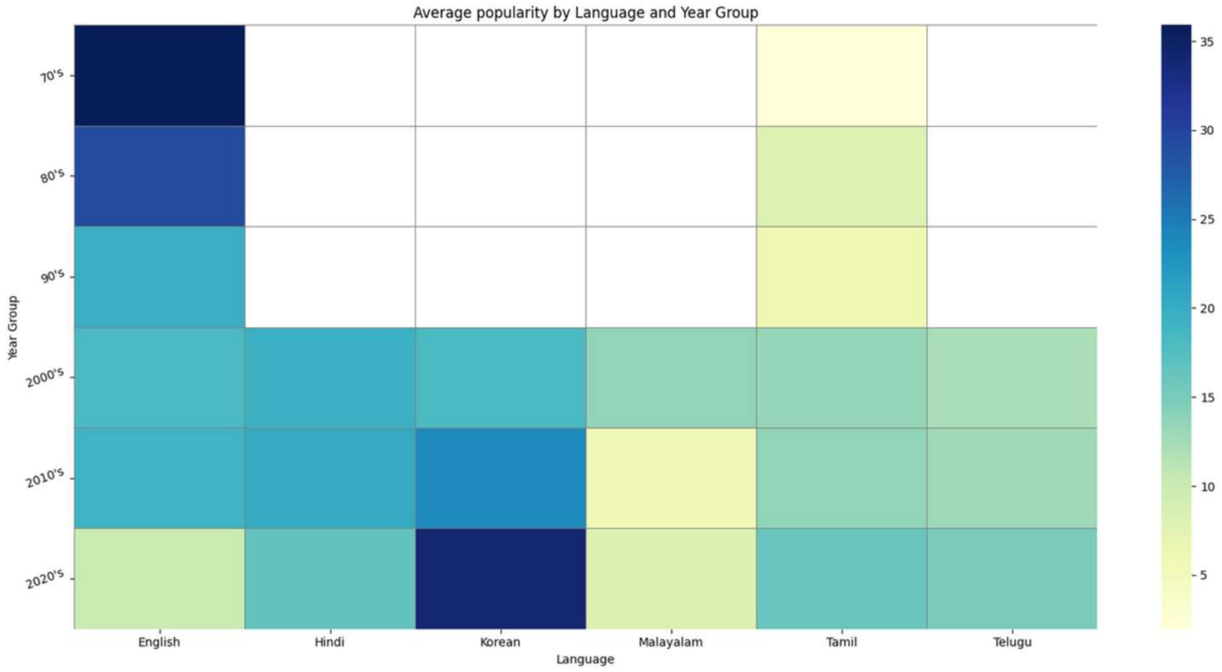
Language wise Analysis of Various Numerical Variables over Year Groups

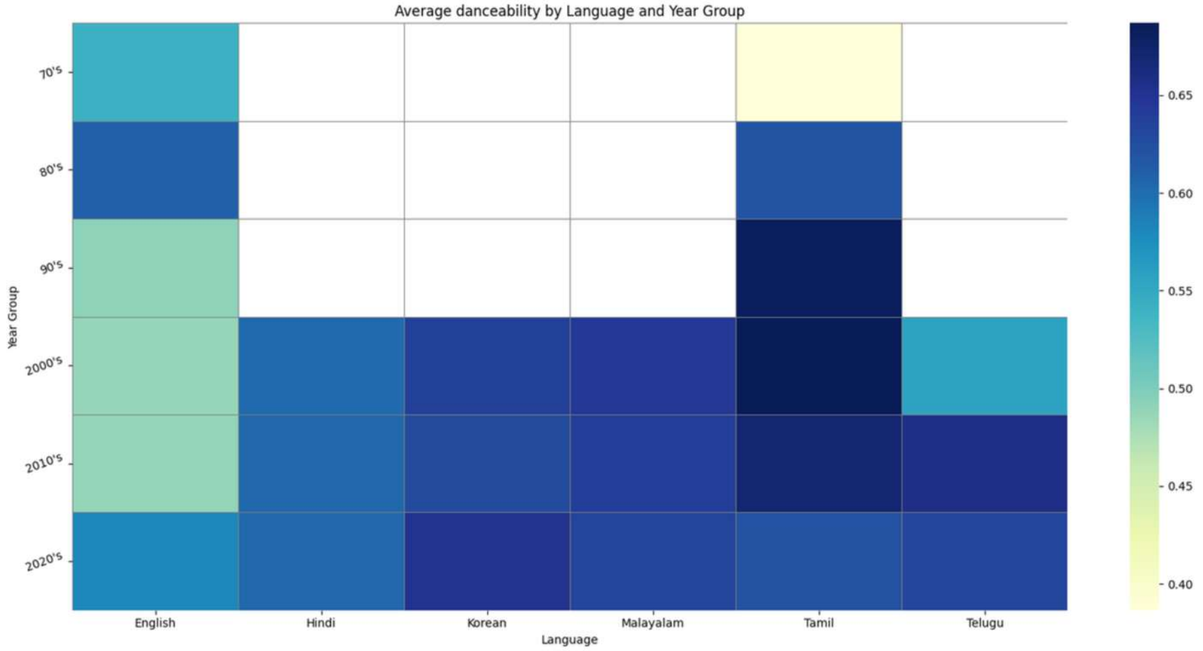
Key Insights

Visual Representation

Multivariate Analysis Overview

Language wise Analysis of Various Numerical Variables over Year Groups



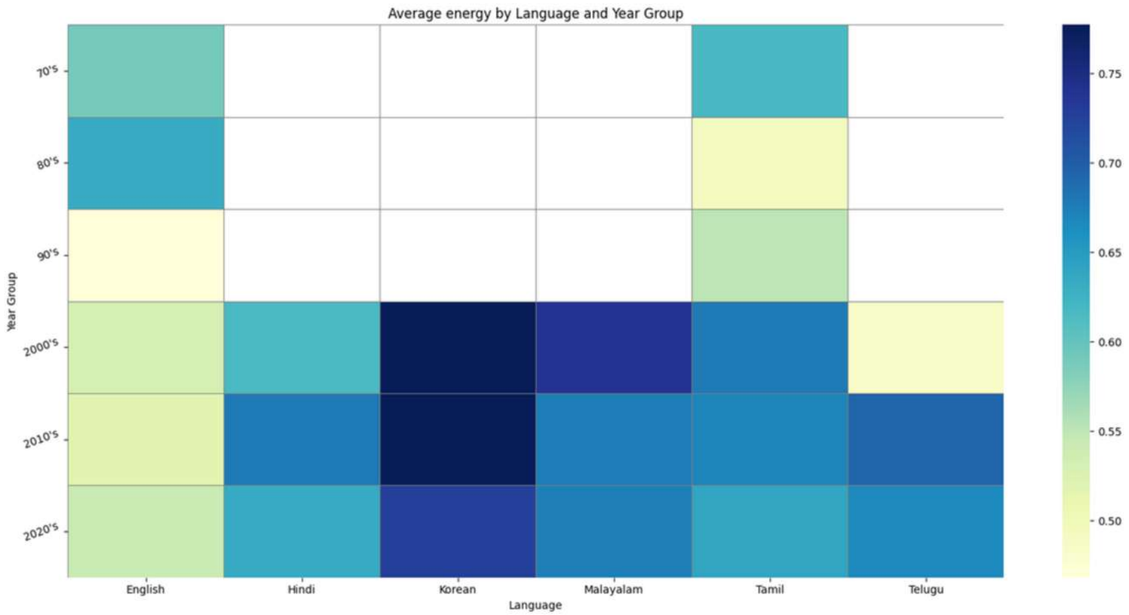


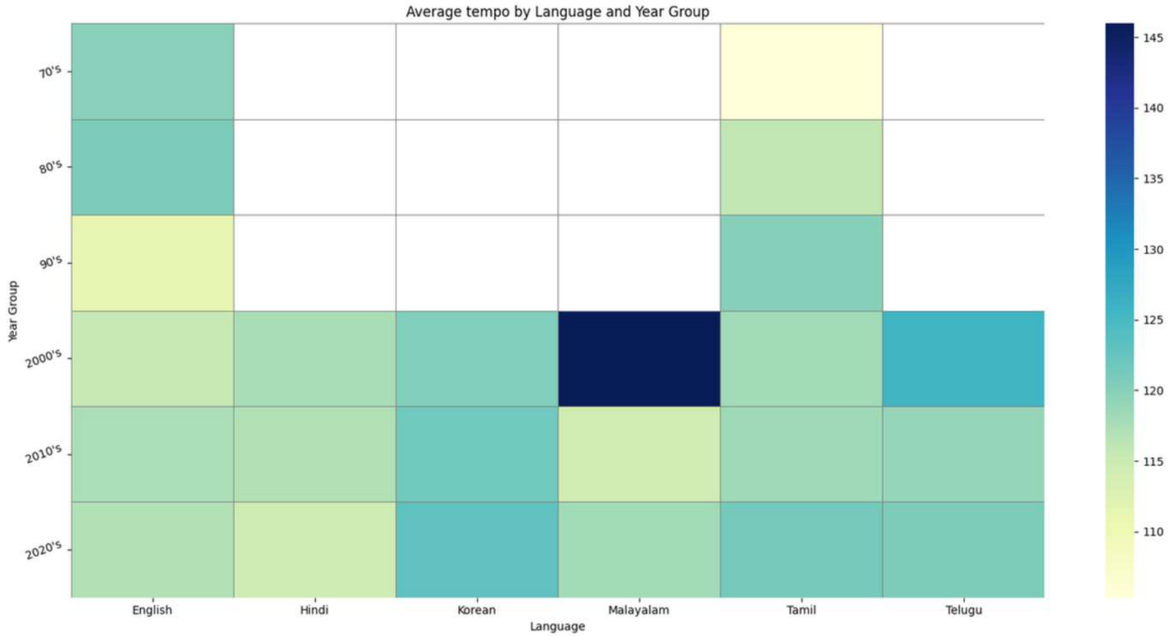
Key Insights:

1. **Popularity by Language-Year: English shows highest popularity in older decades (70s-90s ~35), while Korean peaks in recent 2020s; most languages show declining popularity over time.**
2. **Danceability by Language-Year: Tamil shows highest danceability in 90s-2000s (~0.67), while most languages maintain consistent moderate danceability (0.5-0.6) across decades with minimal temporal variation.**

Multivariate Analysis Overview

Language wise Analysis of Various Numerical Variables over Year Groups





Key Insights:

1. **Tempo by Language-Year: Malayalam shows highest tempo in 2000s (~145 BPM), while most languages maintain consistent moderate tempo (115-125 BPM) across decades with minimal variation.**
2. **Energy by Language-Year: Korean shows highest energy in 2000s-2010s (~0.77), while English shows consistently lower energy across all decades; most languages cluster around 0.6-0.7 energy levels.**

Time Series Analysis of Music Trends

Track Releases Over the Years

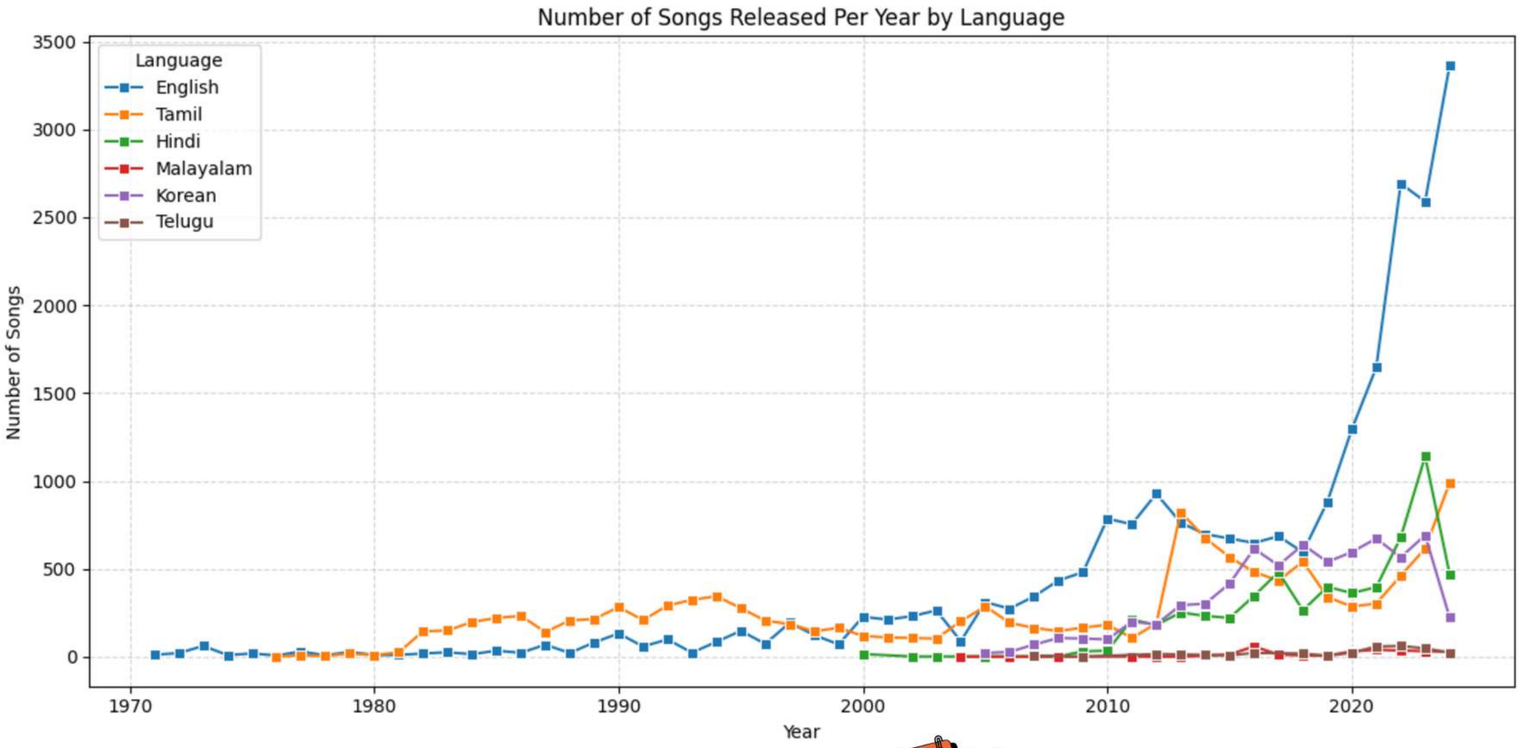
Music Feature Trends Over Years

Change in Solo vs Collaboration over Time per each Language

Time Series Analysis of Music Trends

Track Releases Over the Years

Analyzed the trend of track releases alongside yearly popularity metrics.



Key Insights :

English: **Explosive growth from ~900 to 3,300+ songs between 2019-2025, dominating all other languages by a massive margin.**
Hindi: **Remained minimal for decades before experiencing a dramatic late surge starting in 2023, reaching over 1,100 songs**.
Korean: **Rose from near-zero in 2010 to 600-700 songs by 2025, directly reflecting the global K-pop phenomenon's timeline.**

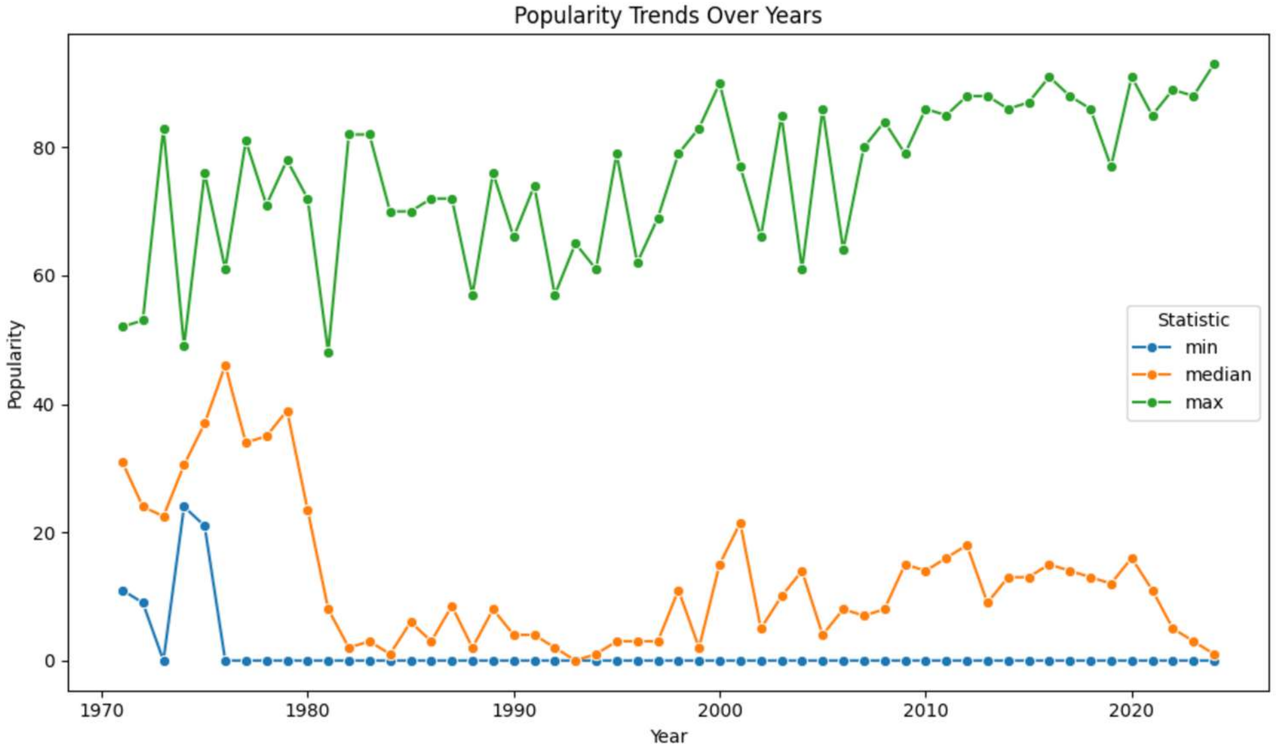
Overall Pattern:

- 2019-2020: Major inflection point for all languages
- Digital platforms democratized music production

Time Series Analysis of Music Trends

Popularity Trends Over Years

Analyzed the trend of track releases alongside yearly popularity metrics.



Key Insights :

Maximum Popularity:

- **High volatility throughout, ranging 50-90 with frequent sharp fluctuations, but shows an overall upward trend from ~52 in 1970 to ~93 by 2025.**

Median Popularity:

- **Started around 30 in 1970, peaked at ~46 in mid-1970s, then dropped dramatically to under 10 by early 1980s where it has remained relatively stable (5-20 range) ever since.**

Minimum Popularity:

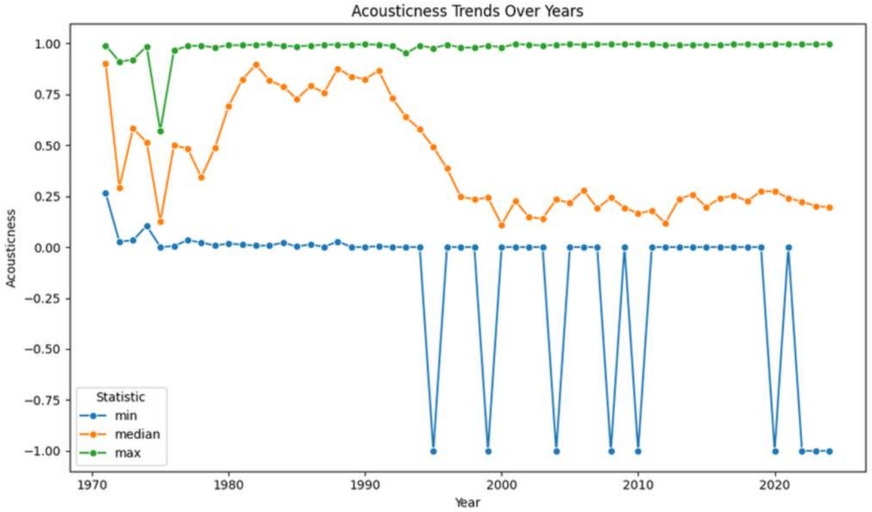
- **Highly volatile in the 1970s (ranging 0-24), then collapsed to near-zero by 1976 and has remained essentially at zero (0-1) for the past 50 years.**

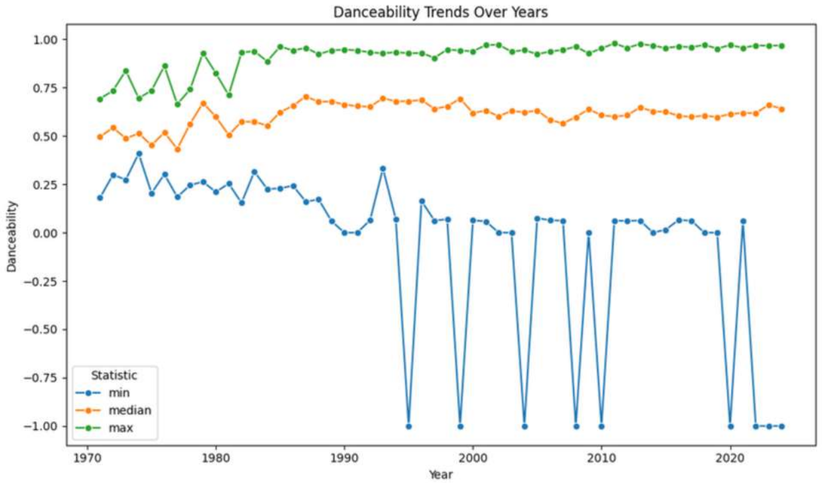
Overall Pattern:

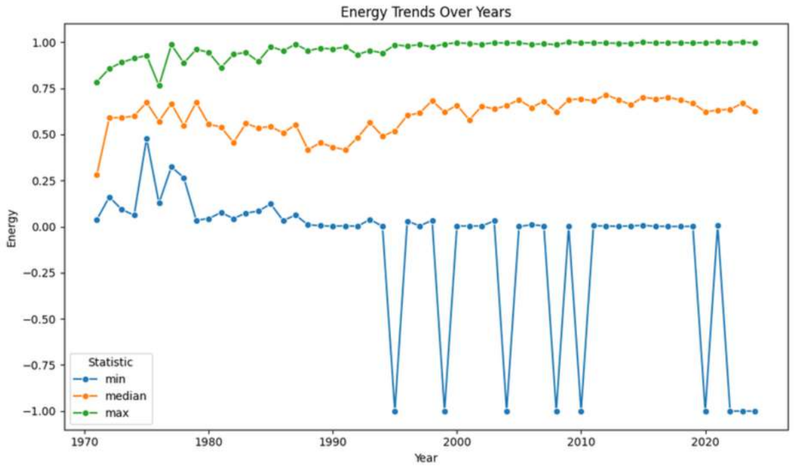
- **Clear divergence after 1980: maximum popularity trending upward while median and minimum remain suppressed, indicating increasing inequality in song popularity with a few mega-hits dominating while most songs get minimal attention.**

Time Series Analysis of Music Trends

Music Feature Trends Over Years





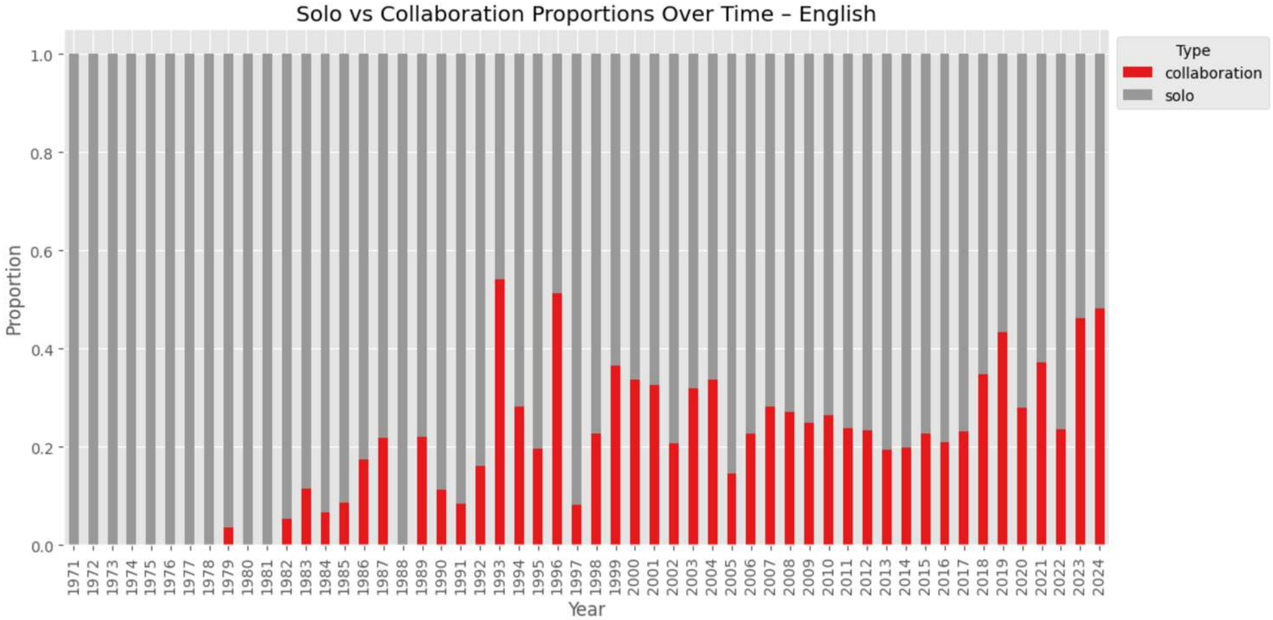


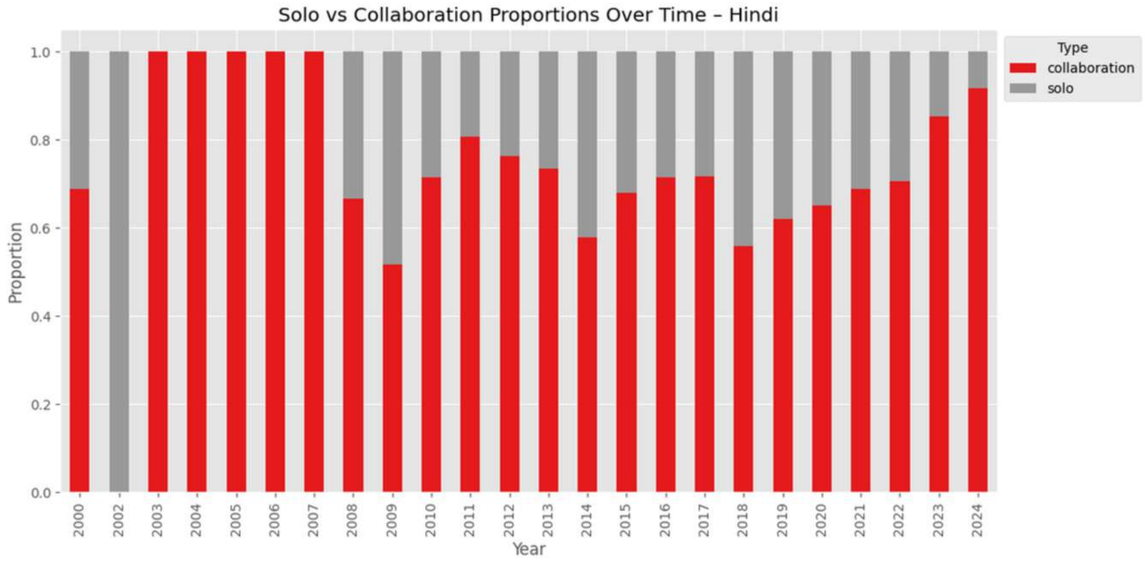
Overall Pattern:

- Acousticness: **Songs went from mostly acoustic instruments (1970s) to mostly electronic/digital sounds (2000s onwards).**
- Danceability: **Songs have stayed equally danceable over the past 50 years, no major change.**
- Energy: **Songs have become more energetic and intense over time, especially since 2000.**

Time Series Analysis of Music Trends

Change in Solo vs Collaboration over Time per each Language



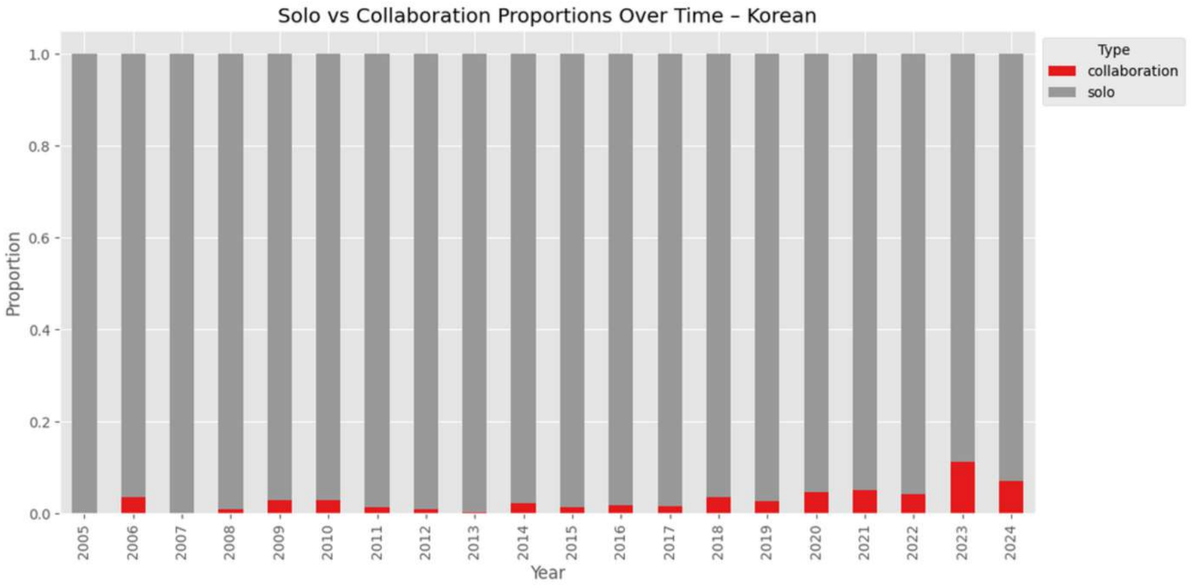


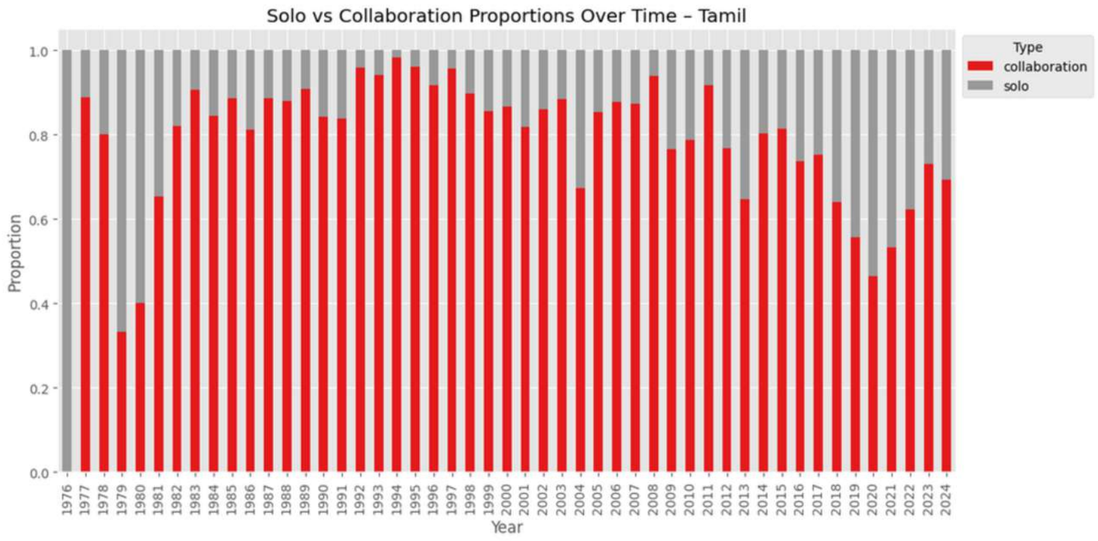
Overall Pattern:

- Hindi: **Collaborations dominated (100%) from 2003-2007, then solo artists took over, with recent years (2023-2024) showing collaborations increasing again to 85-92%.**
- English: **Almost entirely solo artists (95-100%) until the 1990s, then collaborations gradually increased, reaching 40-50% in recent years (2019-2024).**

Time Series Analysis of Music Trends

Change in Solo vs Collaboration over Time per each Language



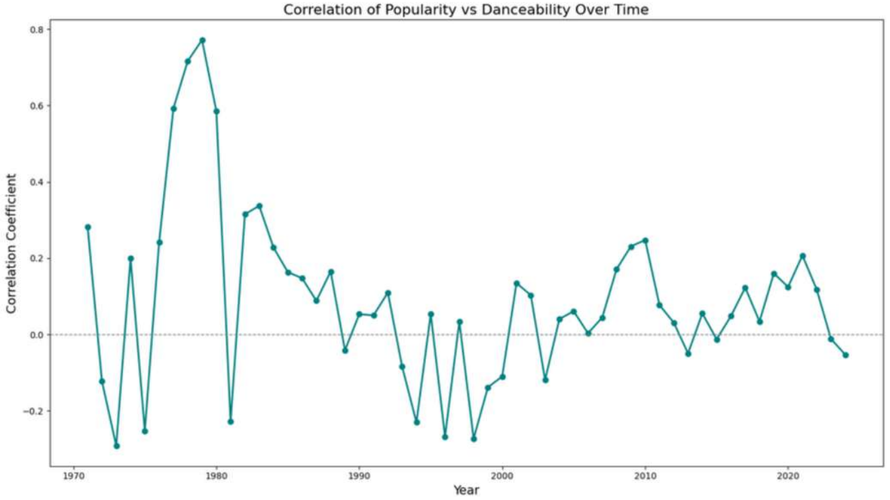


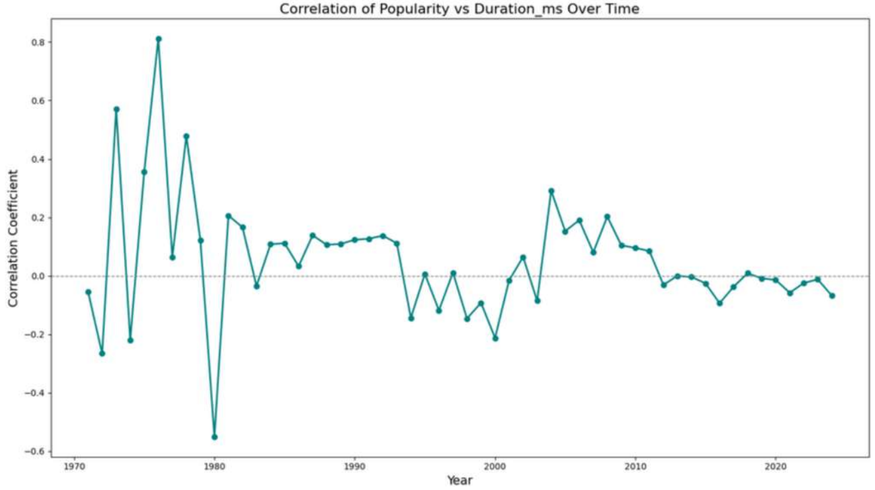
Overall Pattern:

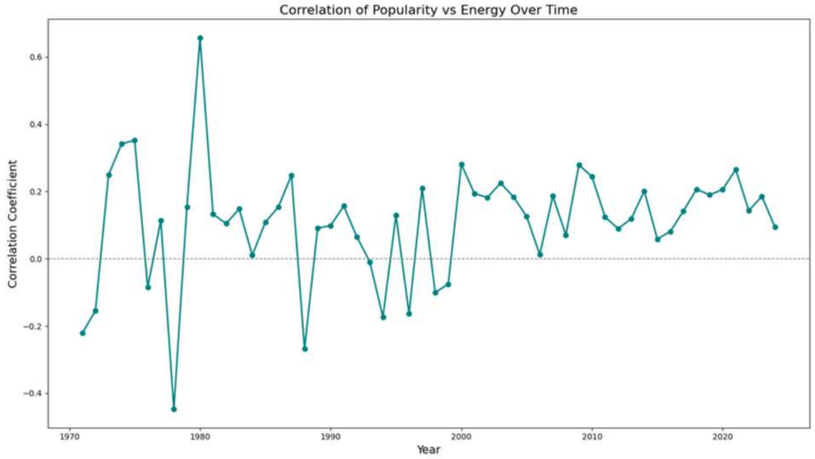
- Korean: **Overwhelmingly solo artists (95-100%) throughout entire period, with collaborations barely visible until recent years (2023-2024) reaching only ~10-12%.**
- Tamil: **Heavily collaboration-focused (65-100%) from 1976-2010, then shifted dramatically to solo-dominated (70-100%) from 2015 onwards, showing complete reversal in industry structure.**

Time Series Analysis of Music Trends

Change over Time of the Correlations of Numerical Variables with Popularity







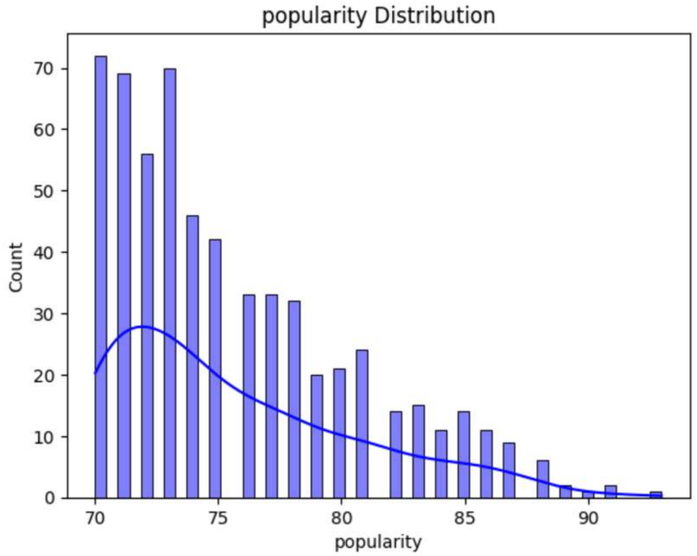
Overall Pattern:

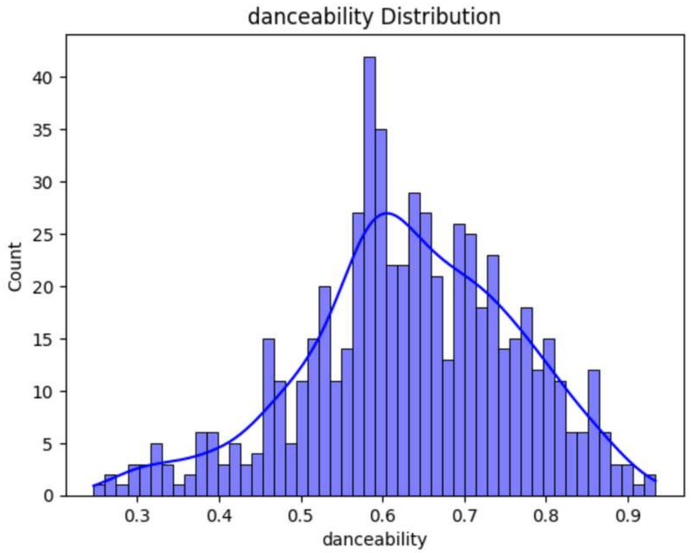
- Popularity vs Danceability: **Strong link in 1970s (0.77 peak), now near zero—danceability doesn't drive popularity anymore.**
- Popularity vs Duration: **Strong link in 1970s (0.81 peak), now near zero—song length doesn't matter for hits anymore.**
- Popularity vs Energy: **Moderate link in 1970s (0.67 peak), weakened but still slightly positive—energy helps a bit, but much less than before.**

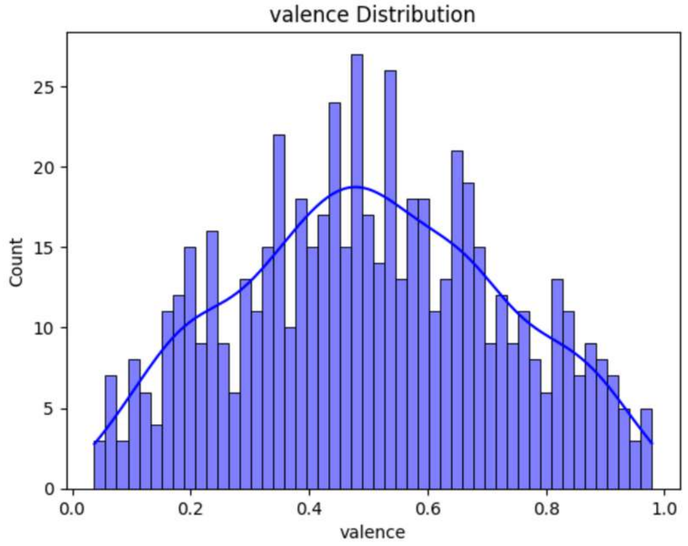
Time Series Analysis of Music Trends

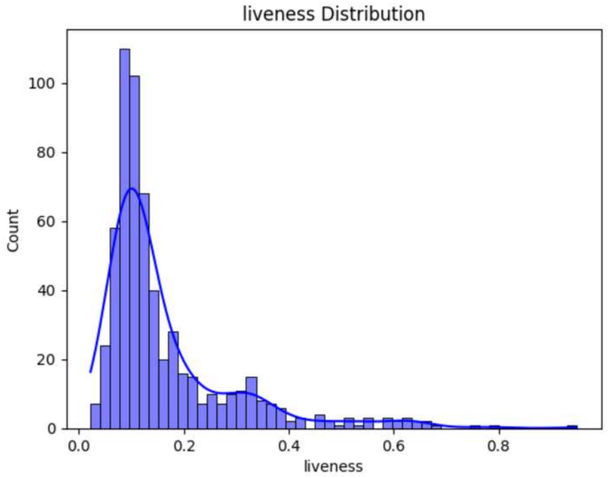
Year_group wise Outlier Distribution

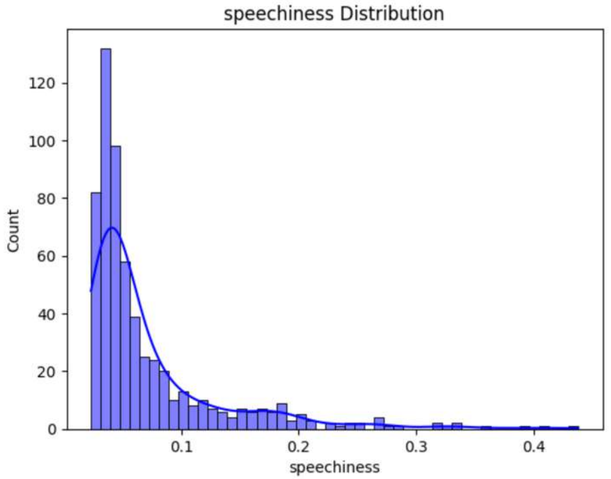
A Brief Analysis of Outliers with Different Numerical variables









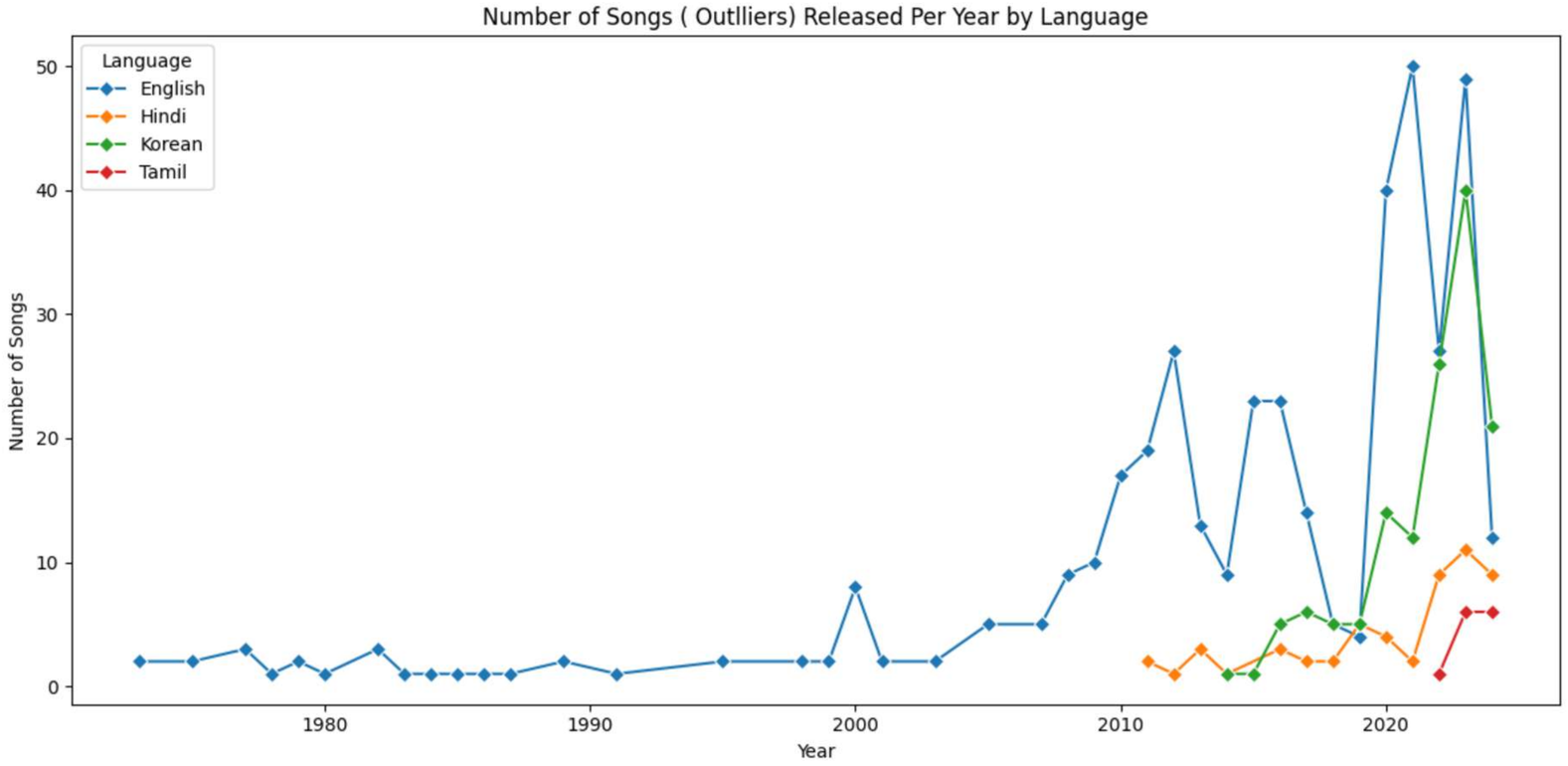


Overall Pattern:

- Danceability: Bell curve centered at 0.6—most songs moderately danceable.
- Valence: Normal distribution centered at 0.45—songs slightly lean negative/sad.
- Liveness: Heavy left skew at 0.1—most songs are studio recordings, not live.
- Speechiness: Heavy left skew near 0—most songs are music, not spoken word.
- Popularity: Right-skewed at 70-73—most songs have low popularity, few become hits.

Time Series Analysis of Music Trends

Number of Songs (Outlliers) Released Per Year by Language



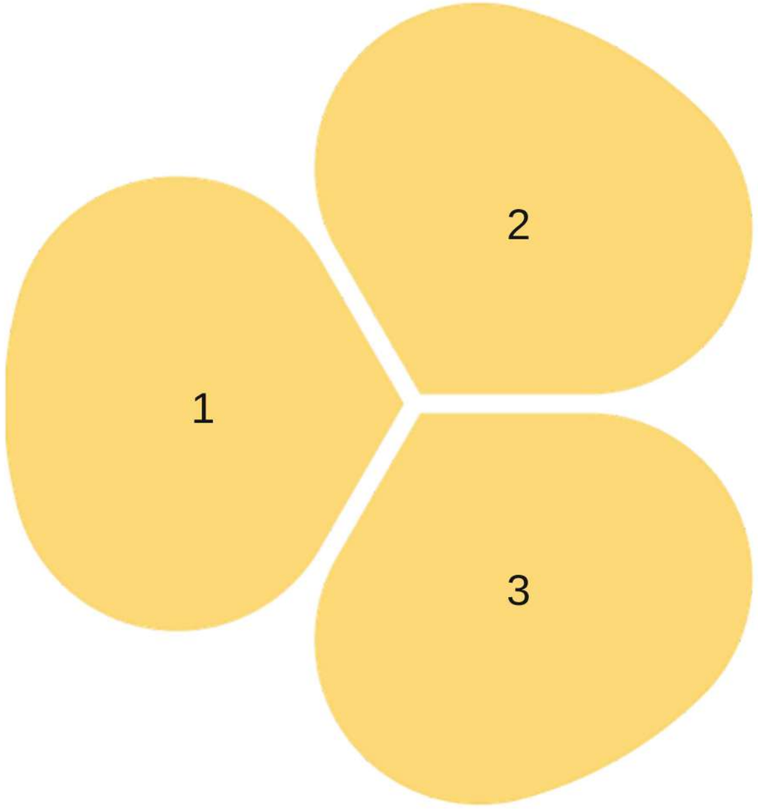
English: **Steady at 1-3 outlier songs until 2008, then explosive growth reaching 50 outliers by 2021.**
Korean: **Minimal until 2019, then sudden spike to 40 outliers by 2023—coincides with K-pop's global explosion.**
Hindi & Tamil: **Both remain low (under 12 outliers) throughout entire period, with slight increases only after 2020**.

Elite Performances

The analysis presents the top 10 artists by track numbers, popular tracks, and genre distributions. The insights reveal artist dominance and genre trends that reflect global Spotify listening patterns, which can inform strategic market decisions.

Top 10 Artists

Identified artists with the highest number of tracks.



Popular Tracks

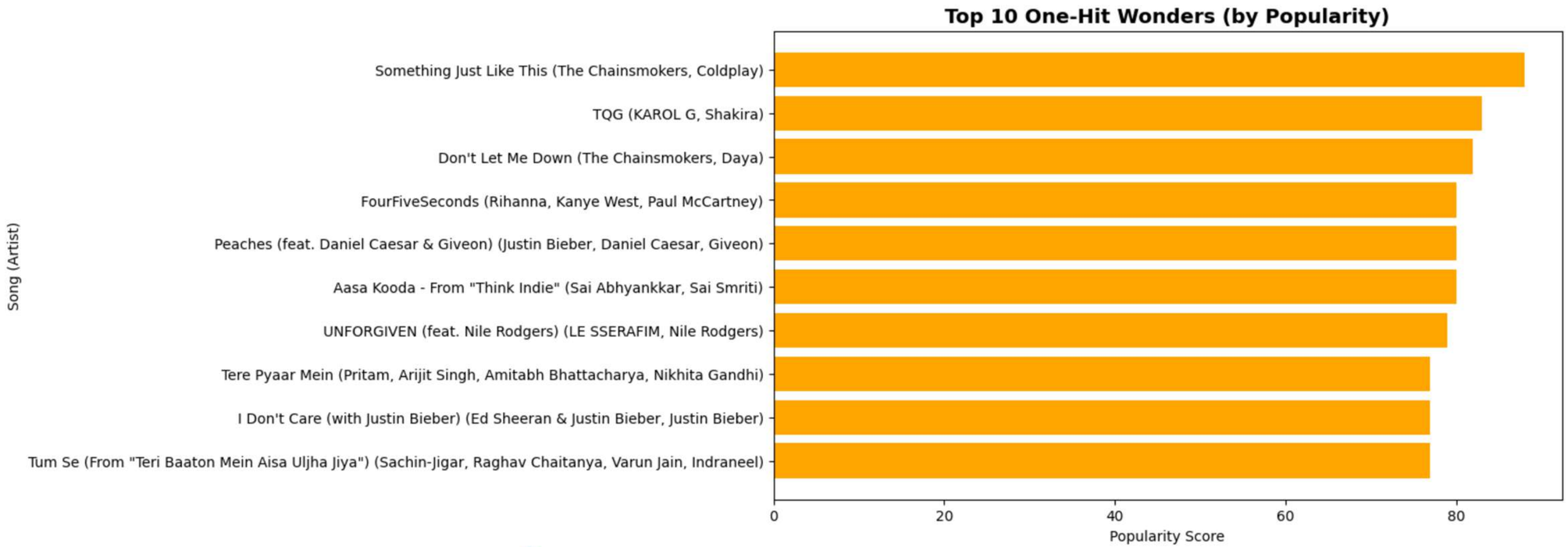
Highlights of the most streamed tracks in the dataset.

Prevalent Genres

Analysis of genres that are dominant in the dataset.

Elite Performances

Top 10 One Hit Wonders (by Popularity)



Insights

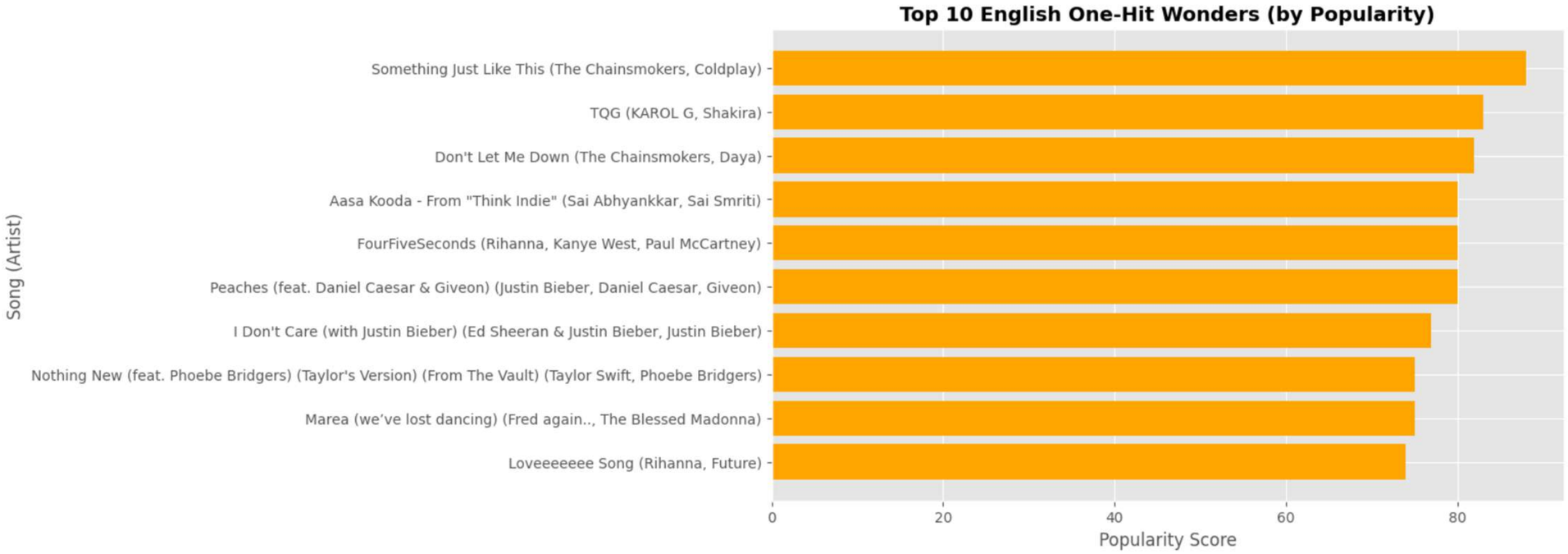
Chainsmokers Dominance: **Two Chainsmokers collaborations in top 3—"Something Just Like This" (#1) and "Don't Let Me Down" (#3).**

International Collaborations Win: **Most top one-hit wonders are collaborations between major artists from different countries (US, UK, India).**

Bollywood Presence: **Three Indian songs made the top 10, showing strong crossover appeal despite being one-hit wonders globally.**

Elite Performances

Top 10 English One Hit Wonders (By Popularity)



Insights

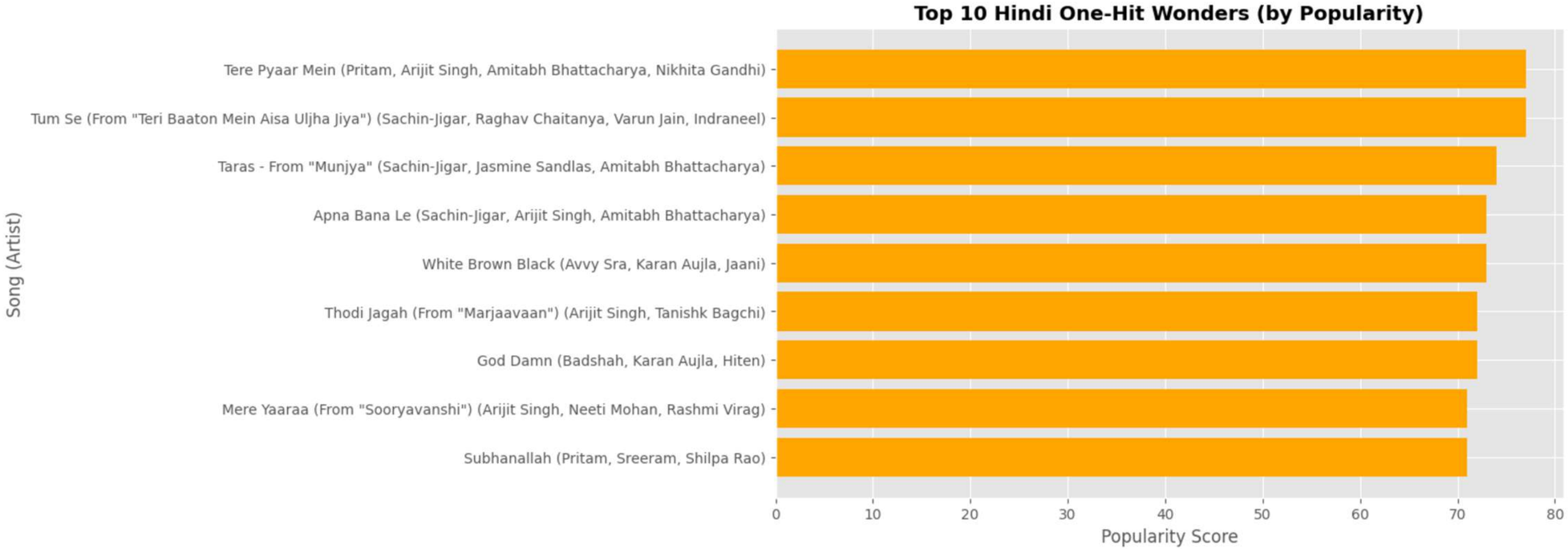
Chainsmokers + Collaborations: **Chainsmokers take #1 and #3 spots with collaborations (Coldplay and Daya).**

Justin Bieber Features: **Justin Bieber appears in 3 songs (with Ed Sheeran, Daniel Caesar/Giveon, and Taylor Swift).**

Mix of Genres: **Top 10 spans EDM, pop, R&B, and indie—showing diverse paths to one-hit wonder status.**

Elite Performances

Top 10 Hindi One Hit Wonders (By Popularity)



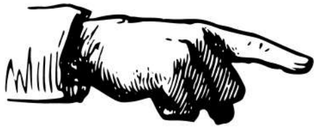
Insights

Film Songs Dominate: **9 out of 10 are Bollywood movie soundtracks—film music drives Hindi one-hit wonders.**

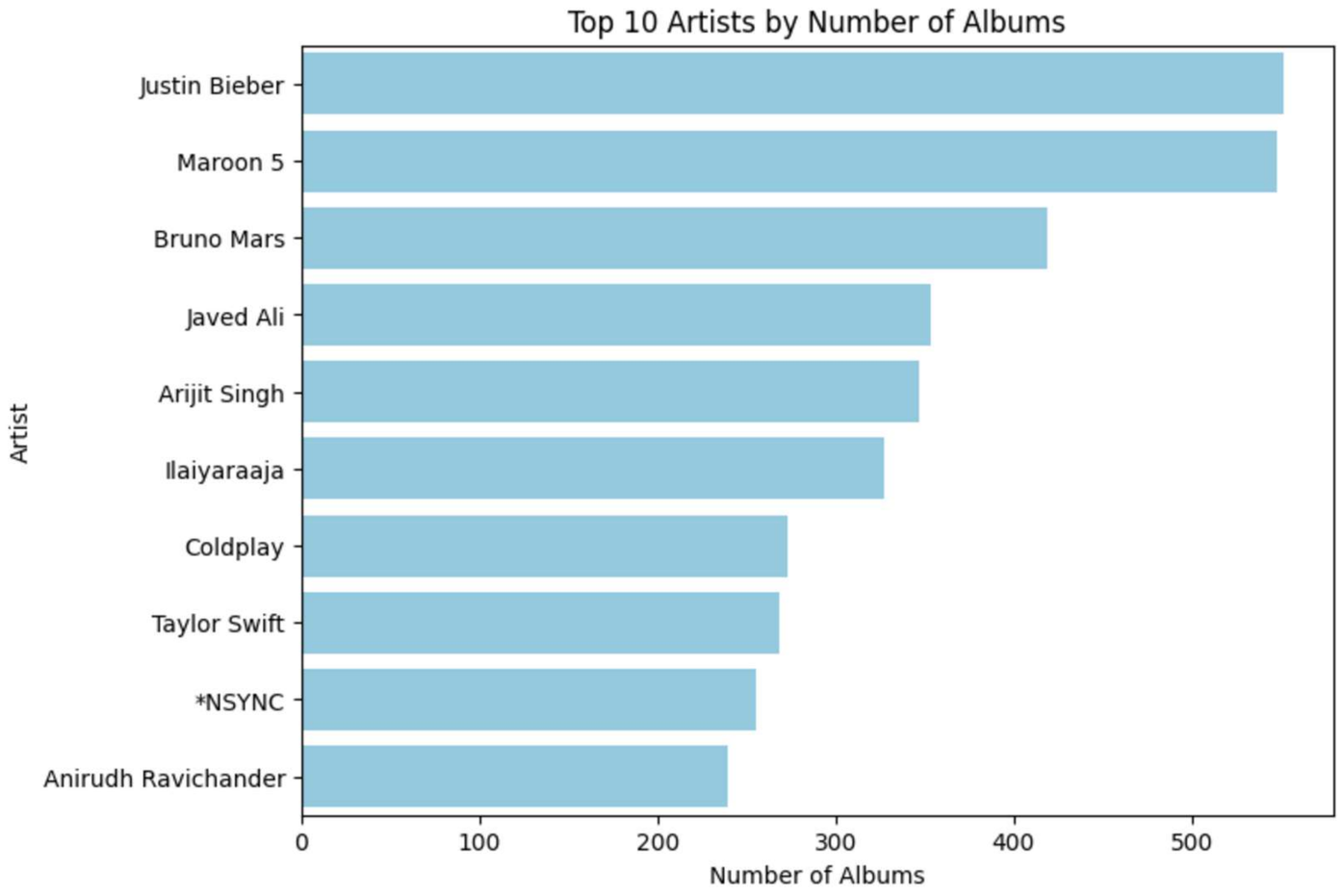
Sachin-Jigar's Success: **Composer duo Sachin-Jigar appears in 3 of top 4 songs, showing their hitmaking consistency.**

Arijit Singh Presence: **Arijit Singh (popular playback singer) features in 4 songs, making him the most frequent voice.**

Elite Performances



Top 10 Artists by Number of Albums



Insights :

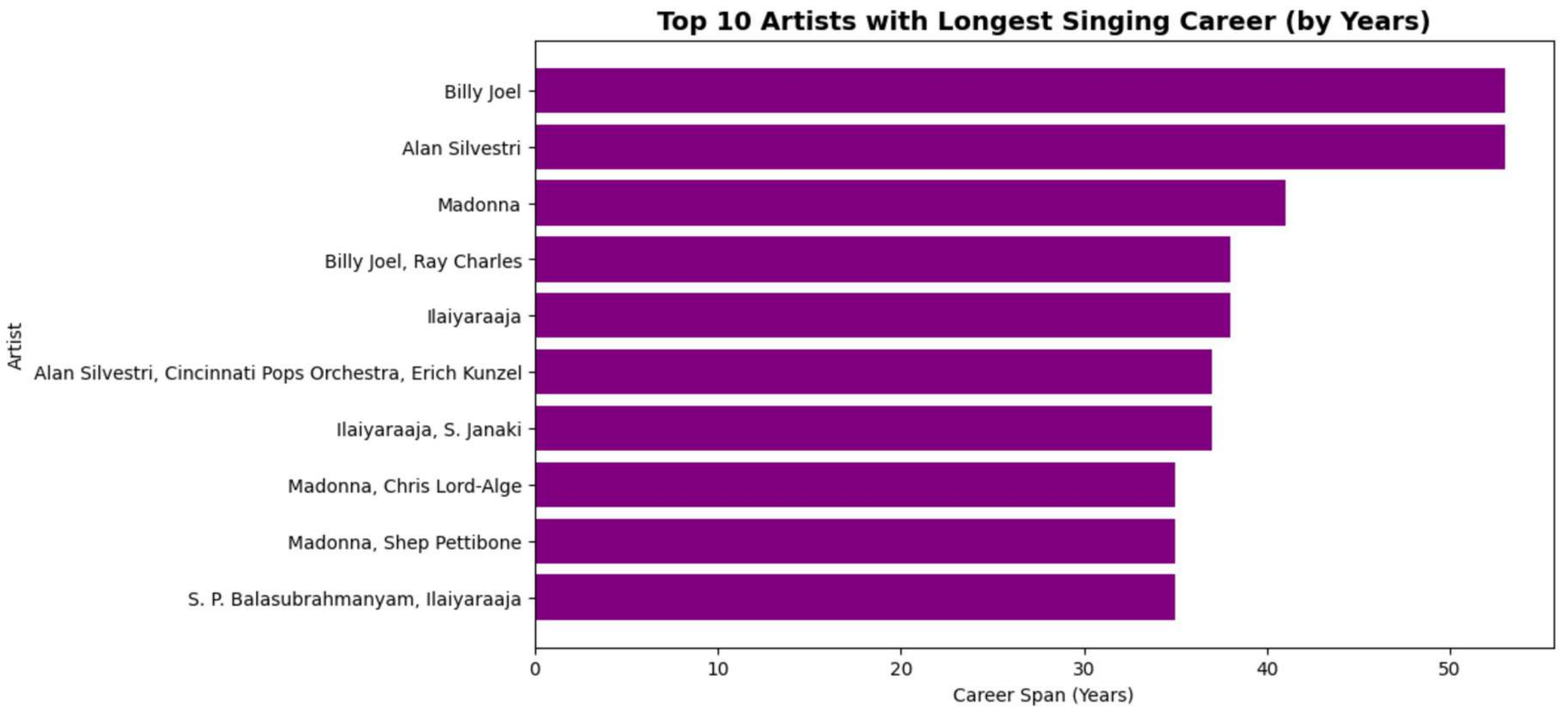
Justin Bieber & Maroon 5 Lead: **Both have released 550+ albums, far ahead of others—showing prolific output over long careers.**

Bruno Mars in Third: **With 420 albums, he's the closest competitor but still significantly behind the top two.**

Indian Artists Strong: **4 Indian artists (Javed Ali, Arijit Singh, Ilaiyaraaja, Anirudh Ravichander) in top 10—reflecting Bollywood's high volume production**.

Elite Performances

Top 10 Artists with Longest Singing Career



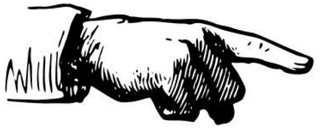
Insights :

Billy Joel & Alan Silvestri Tie: **Both have 53+ year careers, the longest in the dataset—remarkable longevity in music industry.**

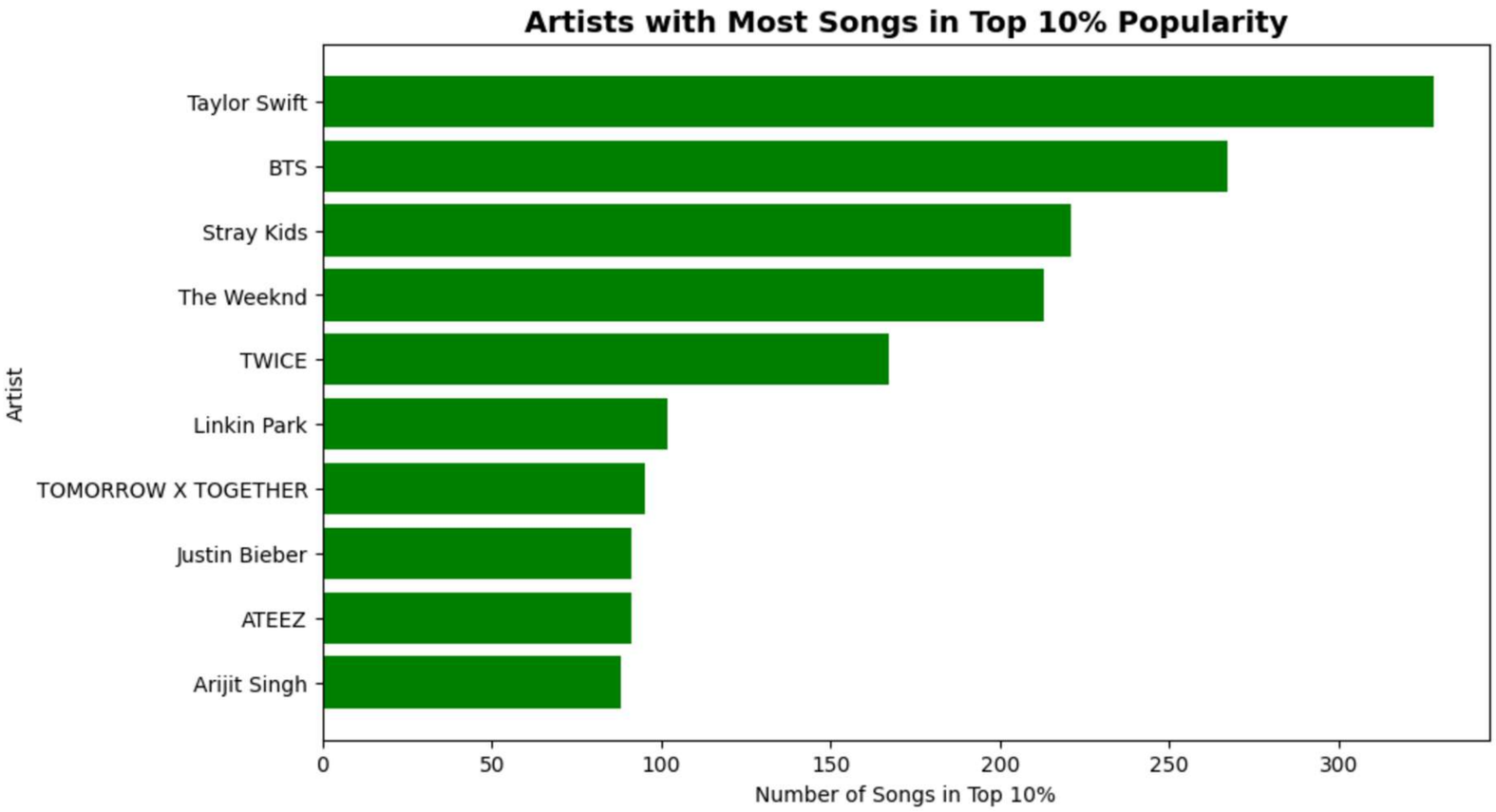
Madonna's Endurance: **41-year career places her third, showing sustained success across multiple decades and genre shifts.**

Indian Classical Dominance: **Ilaiyaraaja appears 3 times (solo and collaborations), highlighting Indian classical musicians' exceptionally long careers.**

Elite Performances



Artists with Most Songs in Top 10% Popularity



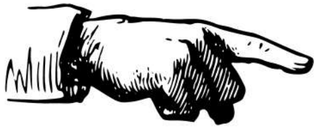
Insights :

Taylor Swift Dominates: **330+ songs in top 10% popularity—far ahead of all others, showing unmatched commercial success**.

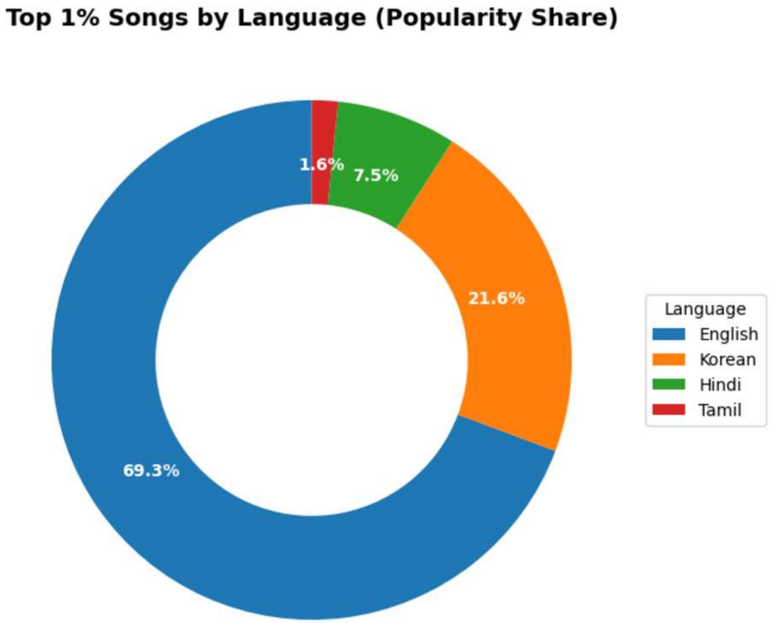
K-pop Strong Presence: **5 K-pop groups (BTS, Stray Kids, TWICE, TXT, ATEEZ) in top 10—demonstrating genre's global popularity power.**

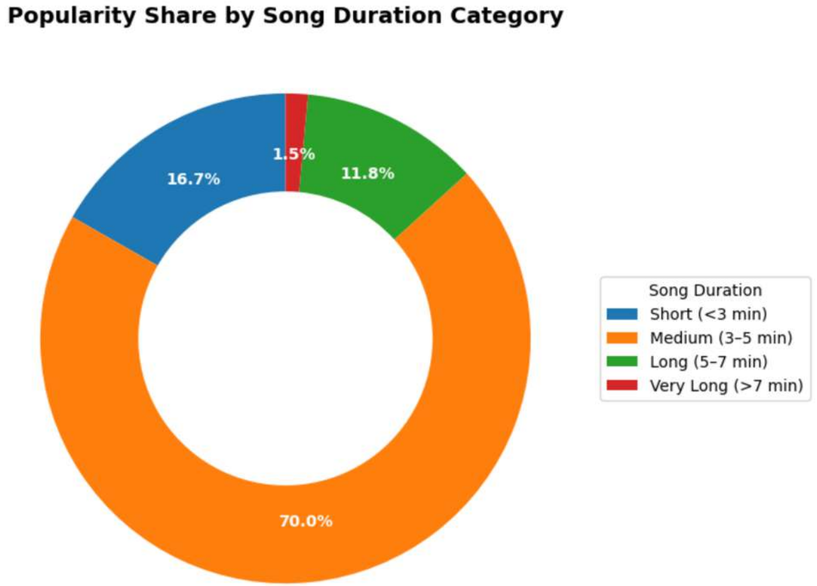
Mix of Generations: **List spans veteran (Linkin Park), current pop (The Weeknd, Justin Bieber), and newer acts (Stray Kids)—showing diverse paths to sustained popularity.**

Elite Performances



Top 1% Songs by Language (Popularity Share) & Popularity Share by Song Duration Category





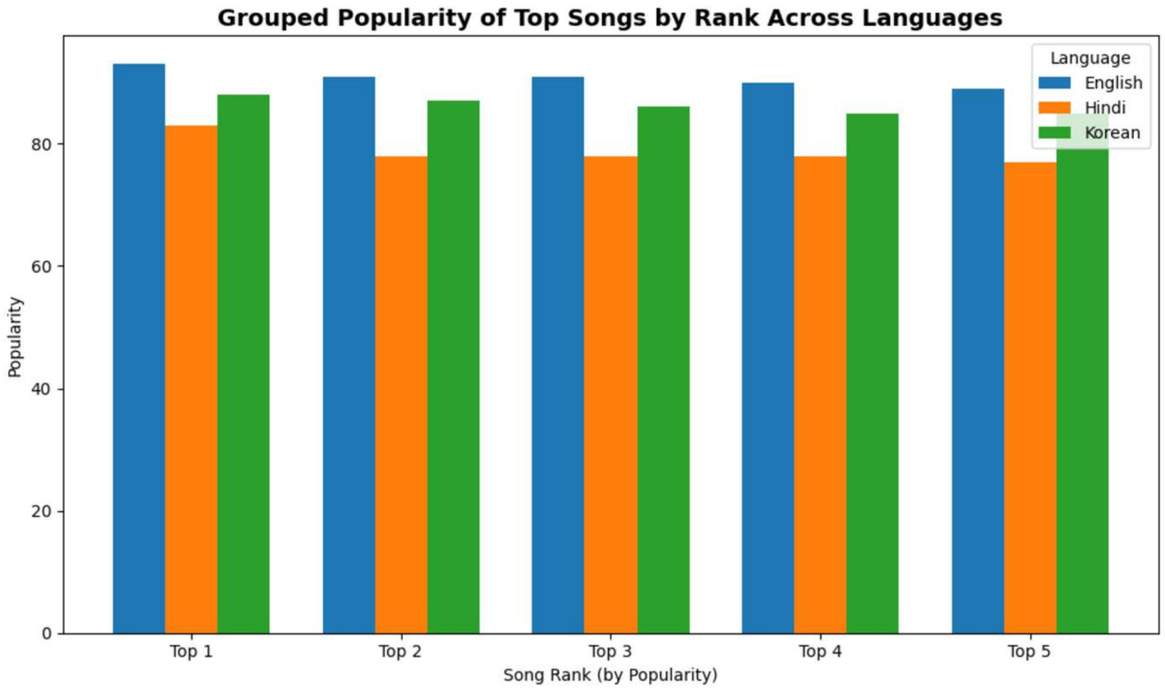
Insights :

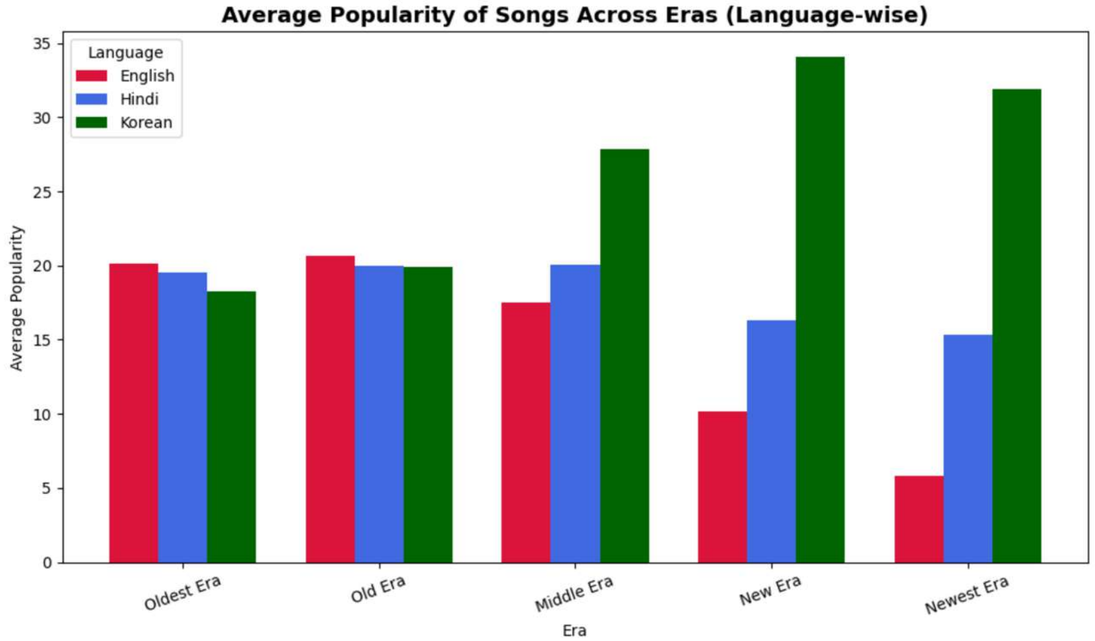
Top 1% Songs by Language: **English dominates with 69.3%, Korean strong at 21.6%, Hindi and Tamil minimal (7.5% and 1.6%)—English is the global hit language.**

Popularity by Song Duration: **Medium length (3-5 min) wins massively at 70%, short songs at 16.7%, long songs combined under 13%—listeners prefer standard radio-friendly lengths.**

Elite Performances

Grouped Popularity of Top Songs by Rank Across Languages & Average Popularity of Songs Across Eras (Language-wise)





Insights :

Grouped Popularity by Rank: **All languages maintain consistent high popularity (77-92 range) across top 5 ranks—top songs are universally popular regardless of language.**

Average Popularity Across Eras: **Korean rises sharply in recent eras (28→34), English declines (20→6), Hindi stays stable (15-20)—Korean music's modern surge is clear.**

Recommendations for Mixing Engineers

1. **Core Track Structure**

- Keep duration in the 3–5 min sweet spot; avoid extremes.
- Target tempo range: 80–200 BPM, balancing radio-friendly and energetic pacing.

2. **Sound & Energy Profile**

- Mix for moderate-to-high energy and danceability (0.5–0.8).
- Favor digital/electronic textures; minimize excessive acousticness.
- Maintain loudness clarity with dynamic control—energy should feel strong but not distorted.

Recommendations for Mixing Engineers

3. **Emotional & Creative Factors**

- Both happy (high valence) and sad (low valence) songs can succeed; focus on mix quality over mood.

- Experiment with genre fusions—cross-style collaborations often yield standout tracks.

-

4. **Market & Audience Orientation**

- English remains dominant, but K-pop and Hindi show rapid growth—adapt mixing aesthetics accordingly.

- Solo tracks perform slightly better on average, but international collaborations create global hits.

Analysis of Spotify Data Insights

The analysis of Spotify data reveals strong relationships among track features, illustrating how energy, danceability, and tempo influence popularity. Additionally, understanding time-related trends is crucial for effective music production strategies.

Market & Language Insights :

- Top 1% Songs by Language: English (69.3%) dominates, Korean (21.6%) rising, Hindi (7.5%) and Tamil (1.6%) smaller shares.
- Popularity by Language: English highest median (~62), others lower (8–30).
- Energy by Language: Korean highest (~0.8), English lowest median (~0.5).
- Tempo by Language: Stable medians around 115–125 BPM, Malayalam peaked at ~145 BPM (2000s).
- Danceability by Language: All languages stable (~0.6–0.7).
- Valence by Language: Tamil highest (~0.7), English lowest (~0.4).

Analysis of Spotify Data Insights

The analysis of Spotify data reveals strong relationships among track features, illustrating how energy, danceability, and tempo influence popularity. Additionally, understanding time-related trends is crucial for effective music production strategies.

Elite Performances (Artists & Tracks) :

- Chainsmokers: 2 collabs in global top 3 one-hit wonders cross-collaboration success.
- Justin Bieber: Featured in 3 top one-hit wonders - frequent collaborator pattern.
- Hindi One-Hit Wonders: 9 of top 10 from films; Sachin-Jigar appear 3 times, Arijit Singh in 4 songs.
- Top Artists by Albums: Justin Bieber & Maroon 5 lead with 550+ albums; Bruno Mars follows with 420.
- Longest Careers: Billy Joel & Alan Silvestri ~53 years; Madonna 41 years; Ilaiyaraaja multiple appearances.
- Most Songs in Top 10% Popularity: Taylor Swift dominates with 330+ songs; 5 K-pop groups in top 10.
- Song Duration Popularity (Top 1%): Medium (3–5 min) = 70%, Short = 16.7%, Long = ~13%.

Analysis of Spotify Data Insights

Time Series Insights for Mixing Engineers

1. Track Releases & Growth

- English: Explosive rise from ~900 (2019) to 3,300+ songs (2025) - dominant production volume.
- Hindi: Minimal for decades, but sharp surge post-2023, reaching 1,100+ songs.
- Korean: Grew from near zero (2010) to 600–700 songs (2025) - reflects K-pop's global boom.
- Overall: 2019–2020 = inflection point across languages due to digital platforms.

2. Popularity Trends Over Time

- Maximum Popularity: Climbed from ~52 (1970) to ~93 (2025), but with high volatility (50–90).
- Median Popularity: Peaked ~46 (mid-1970s), collapsed below 10 in 1980s, stayed low (5–20) since.
- Minimum Popularity: Fell near-zero after 1976, stayed at 0–1 for 50+ years.
- Pattern: Rising inequality - few mega-hits dominate, while most songs remain obscure.

Analysis of Spotify Data Insights

Time Series Insights for Mixing Engineers

3. Music Feature Evolution

- Acousticness: Shift from mostly acoustic (1970s) to mostly electronic/digital (2000s onwards).
- Danceability: Stable over 50 years (~0.6).
- Energy: Significant rise post-2000 → songs became more intense and energetic.

4. Collaboration Trends by Language

- Hindi: 100% collaborations (2003–2007), then solos dominated; recently (2023–2024) collabs surged again to 85–92%.
- English: Predominantly solos until 1990s (95–100%), then collabs rose to 40–50% in 2019–2024.
- Korean: Consistently solo (95–100%), tiny rise to ~10–12% collabs in 2023–2024.
- Tamil: Opposite trend—65–100% collabs (1976–2010), then flipped to 70–100% solos post-2015.

Analysis of Spotify Data Insights

Time Series Insights for Mixing Engineers

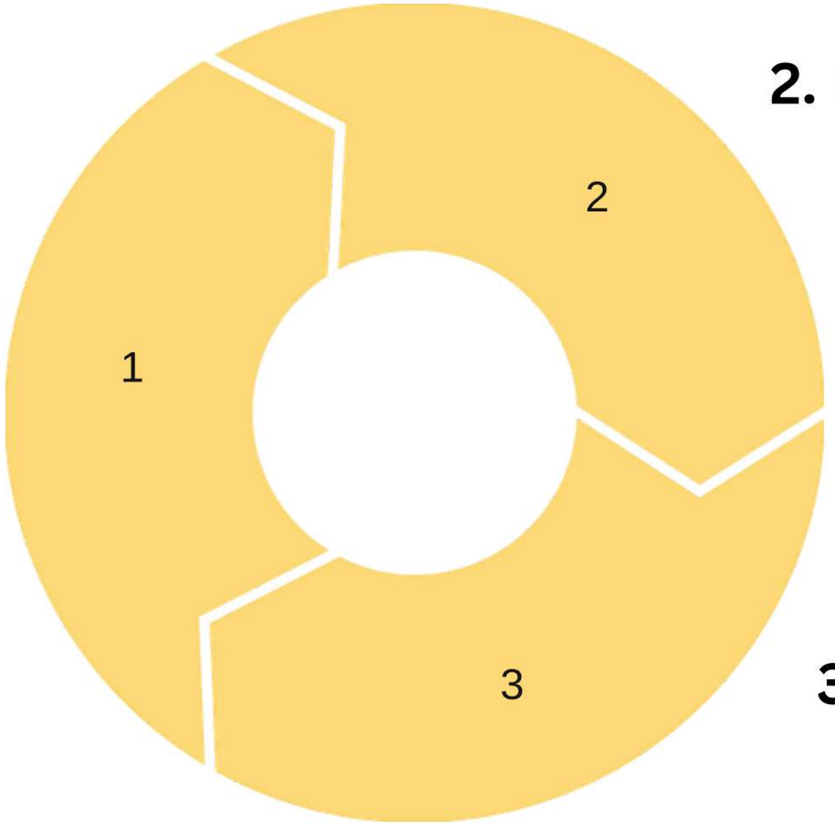
5. Changing Correlations with Popularity

- Danceability vs Popularity: Strong in 1970s (0.77), now near zero.
- Duration vs Popularity: Strong in 1970s (0.81), now near zero.
- Energy vs Popularity: Moderate in 1970s (0.67), still slightly positive today.

6. Outliers & Distributions

- Danceability: Bell curve centered at ~0.6.
- Valence: Normal distribution around ~0.45 (slightly negative lean).
- Liveness & Speechiness: Heavily skewed low → most songs are studio-recorded, not speech-heavy.
- Popularity: Right-skewed, top cluster ~70–73 → only a few tracks achieve high success.

Future Work Directions



1. User Listening Behaviors

Study how listeners engage with tracks over time.

2. Recommendation System Development

Build smarter systems using feature relationships.

3. Exploring Lyrics Aspects

Extend analysis beyond audio to lyrical content.

Thank You

Appreciation

Thank you for your attention.