

DATA MINDS



Spotify Exploratory Data Analysis (EDA) Project

TEAM MEMBERS

- **Dipanjan Halder**
- **Ekadashi Sardar**
- **Anshika Pandey**
- **Nilanjana Saren**

PRESENTED BY:
DIPANJAN HALDER



RHYTHMS IN DATA: EXPLORING SPOTIFY TRENDS

Problem Statement:

Spotify contains millions of tracks with unique features describing rhythm, mood, and sound quality. As a Music Director / Mixing Engineer, the goal is to analyze these attributes to understand what drives a song's popularity and listener engagement.

The project aims to:

- Explore patterns among key audio features such as energy, danceability, valence, and loudness
- Identify how these attributes influence a song's popularity score
- Provide data-driven recommendations to guide music production and mixing decisions



➤ PROJECT OBJECTIVES

Our Exploratory Data Analysis aims to:

- Identify key factors influencing song popularity on Spotify.
 - Analyze trends and patterns across musical features such as energy, danceability, valence, tempo, and loudness.
 - Understand correlations between audio characteristics and popularity to uncover meaningful relationships.
 - Examine changes over time, exploring how song attributes have evolved through different years.
 - Provide actionable insights for music production and mixing decisions based on data-driven findings.



DATA DESCRIPTION

Track Information

- Track ID – Unique identifier for each song on Spotify
- Track Name – Title of the song
- Artist Name – Name of the performing artist(s)
- Album Name – The album the track belongs to
- Year – Year the song was released

Acoustic Properties

- Acousticness – Likelihood of the track being acoustic
- Instrumentalness – Predicts if a track contains no vocals
- Speechiness – Detects spoken words or lyrical content
- Liveness – Presence of audience in the recording

Audio Features

- Danceability – How suitable a track is for dancing
- Energy – Perceptual measure of intensity and activity
- Valence – Musical positivity or emotional tone
- Tempo – Beats per minute (BPM)
- Loudness – Overall sound level in decibels (dB)
- Mode – Musical mode (0 = minor, 1 = major)
- Key – Musical key of the track

Engagement & Metadata

- Popularity – Popularity score ranging from 0-100
- Duration (ms) – Total track length in milliseconds
- Language – Detected language of lyrics
- Artwork URL / Track URL – Spotify media and track links

01 Univariate Analysis

Analysis of a single variable to understand its distribution, central tendency, and spread.

02 Time-series Analysis

Analysis of data over time to identify trends, patterns, and seasonality.

03 Bivariate Analysis

Analysis of the relationship between two variables.

04 Multi-variate Analysis

Analysis of more than two variables simultaneously to study complex relationships.

05 Outliers Analysis

Detecting and studying data points that deviate significantly from the rest of the dataset.

► DESCRIPTIVE STATISTICS

1. Central Tendency of 'Popularity':

Mean Popularity: 15.36

Median Popularity: 7.00

Mode Popularity: [0]

2. Statistics for 'Duration (ms)':

Mean Duration: 242603.45 ms

Median Duration: 236311.00 ms

Minimum Duration: 5000.00 ms

Maximum Duration: 4581483.00 ms

3. Distribution of 'Danceability' and 'Energy':

Danceability Description:

count 62239.000000

mean 0.596768

std 0.186262

min -1.000000

25% 0.497000

50% 0.631000

75% 0.730000

max 0.986000

Name: danceability, dtype: float64

Energy Description:

count 62239.000000

mean 0.602416

std 0.246207

min -1.000000

25% 0.440000

50% 0.639000

75% 0.803000

max 1.000000

Name: energy, dtype: float64

4. Total and Average 'Popularity' and 'Loudness':

Total Popularity (sum across tracks): 955,841.00

Average Popularity: 15.36

Average Loudness: -65.17 dB

Min Loudness: -100000.00 dB

Max Loudness: 1.23 dB

5. Skewness and Kurtosis of 'Popularity', 'Danceability', and 'Energy':

Skewness of Popularity: 1.2314

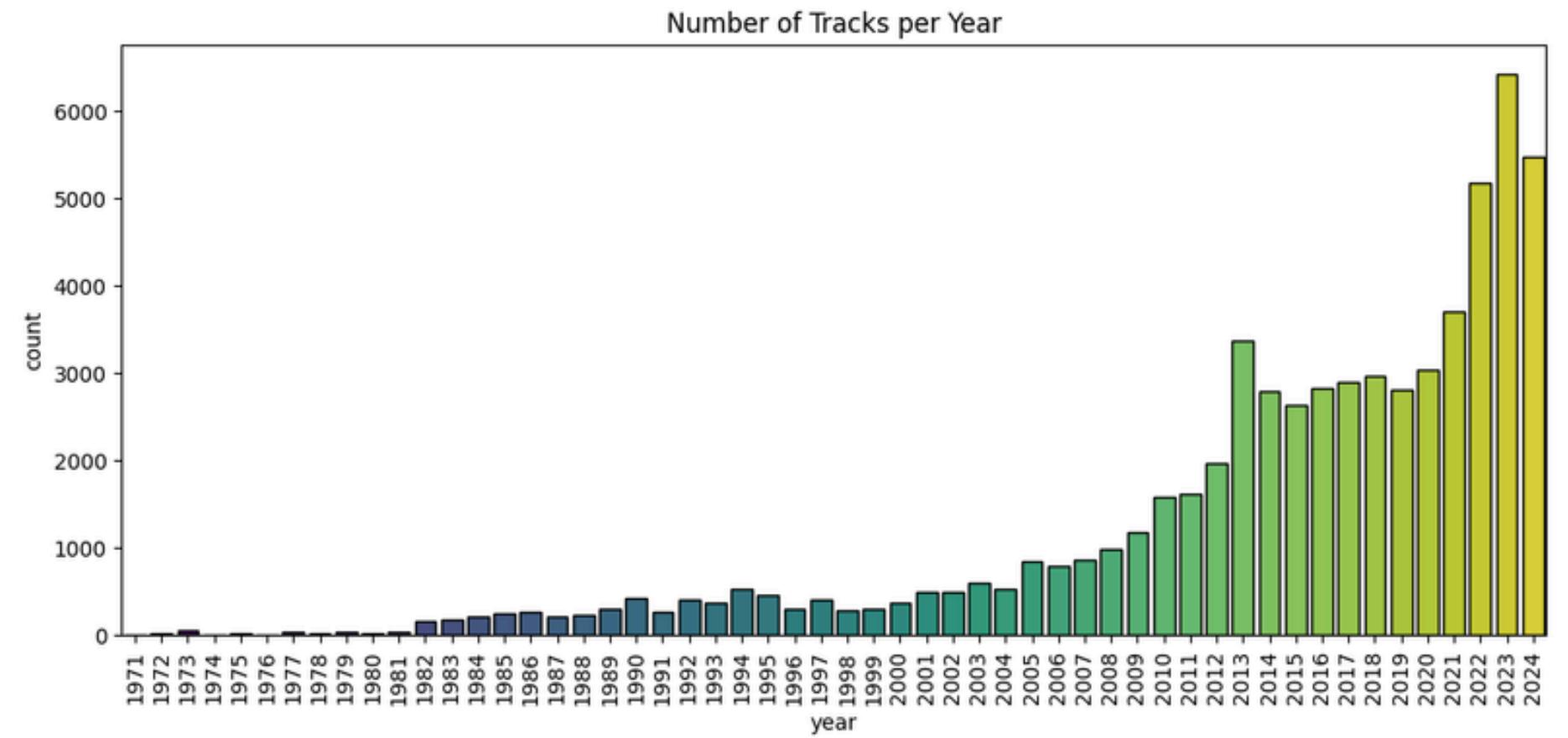
Kurtosis of Popularity: 0.6620

Skewness of Danceability: -1.0577

Kurtosis of Danceability: 2.9039

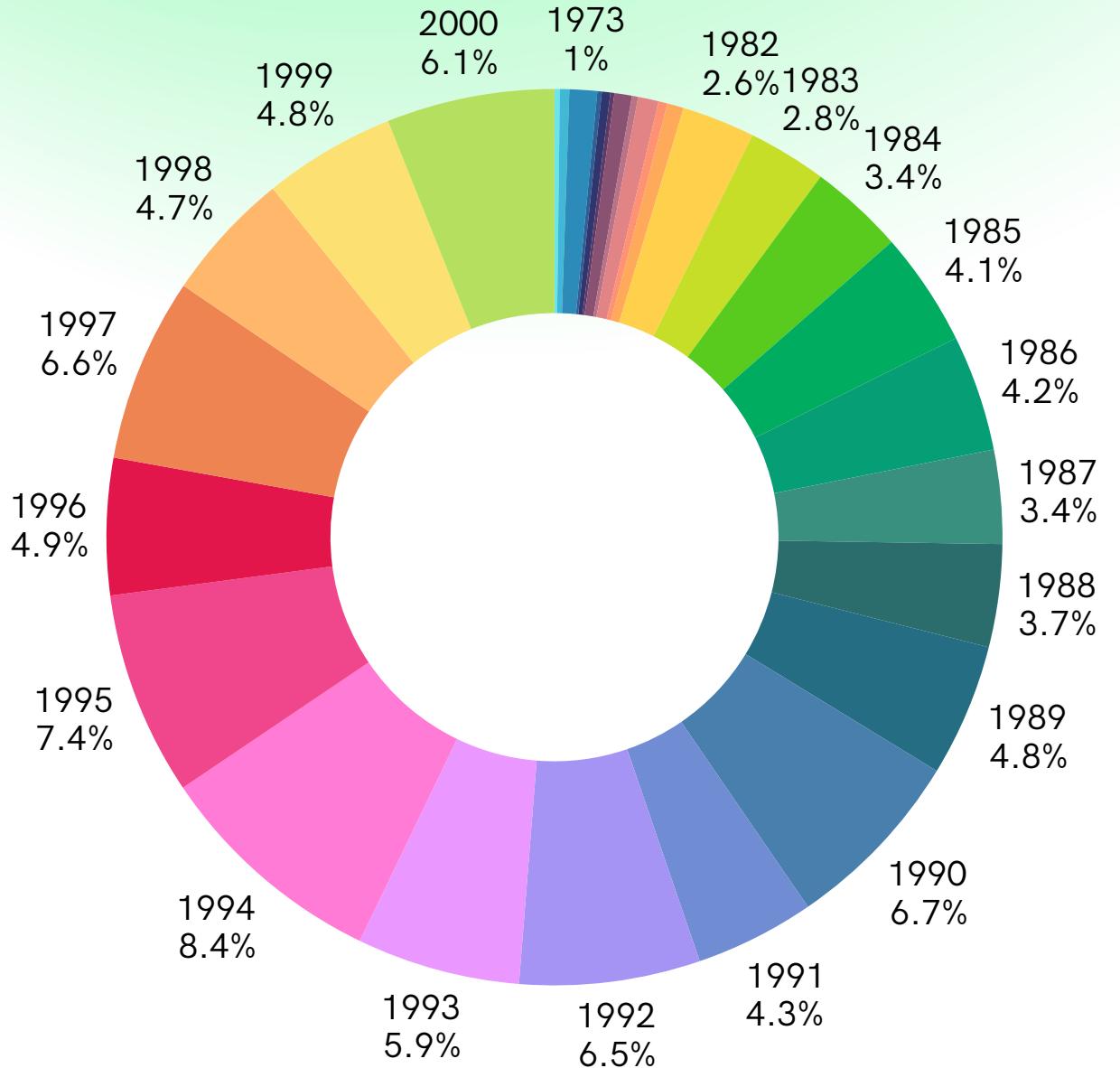
Skewness of Energy: -0.6630

Kurtosis of Energy: 0.3522

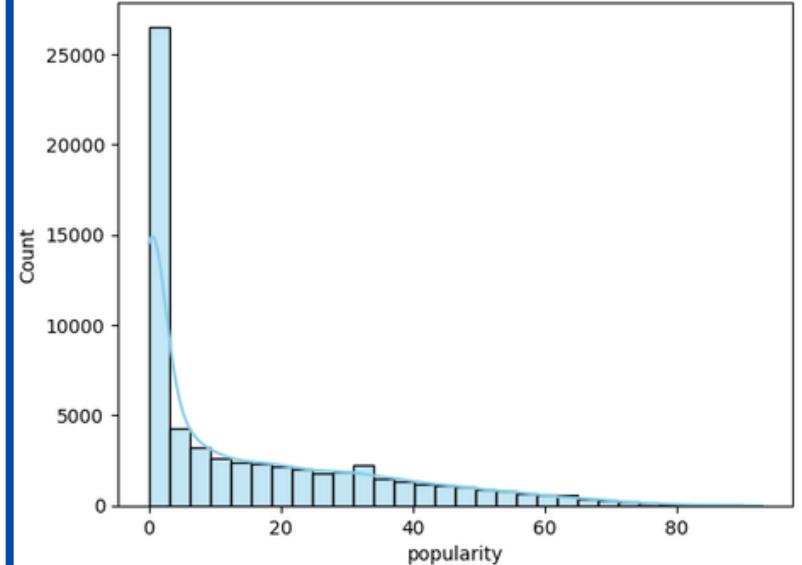


Insight :

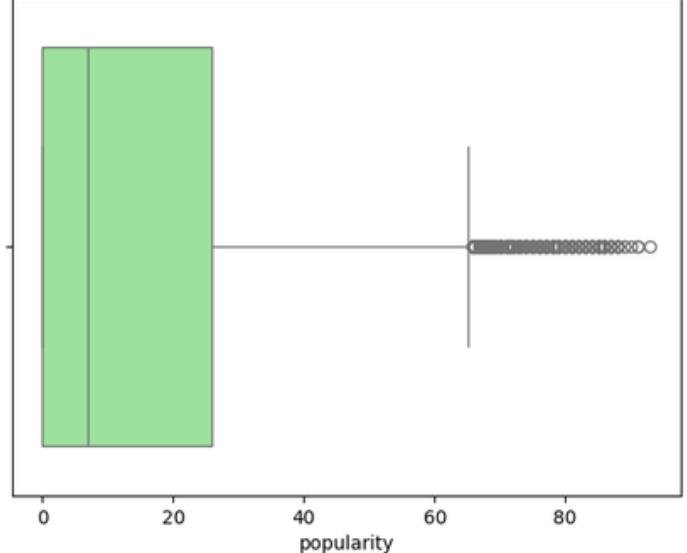
- The number of tracks released has increased dramatically over time, especially from the early 2000s onward.
- Before the 1990s, releases were relatively sparse — less than 1% of total songs came from those early years.
- The trend began to rise steadily through the late 1990s, with a notable acceleration after 2010, coinciding with the digital music and streaming boom.
- The peak occurs between 2018-2022, where yearly track counts exceed 6,000 songs, showing Spotify's rapid expansion of its music catalog.
- The donut chart emphasizes this visually: modern decades dominate, with 1994-2000 collectively contributing over 45% of the dataset.
- This distribution confirms that recent years dominate Spotify's data, aligning with how streaming platforms now host massive volumes of new music releases annually.



Distribution of Popularity

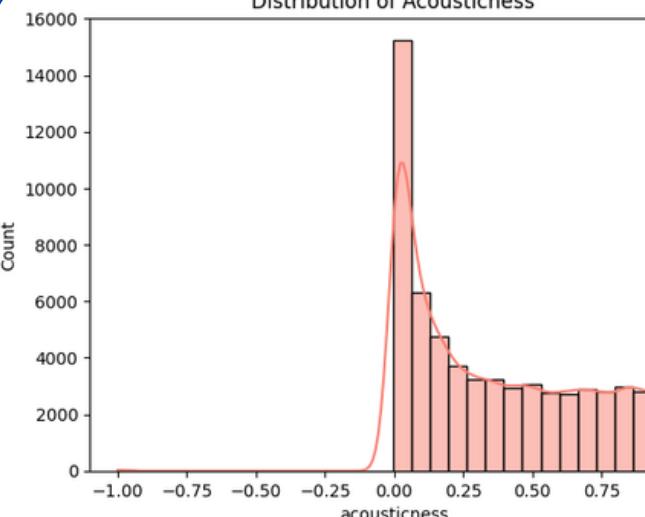


Boxplot of Popularity



popularity
count 62239.000000
mean 15.357589
std 18.630494
min 0.000000
25% 0.000000
50% 7.000000
75% 26.000000
max 93.000000

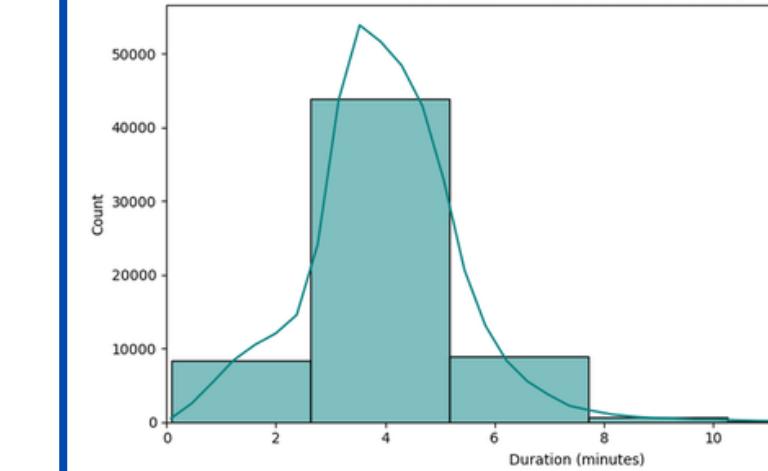
Distribution of Acousticness



acousticness

0.16400	193
0.86000	163
0.66600	159
0.11000	156
0.11500	147
0.00589	130
0.00380	129
0.01200	127
0.04900	125
0.14600	125

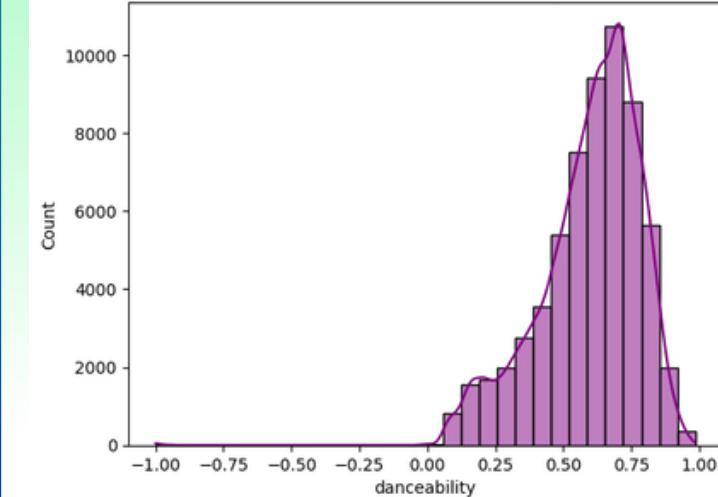
Distribution of Track Duration (minutes)



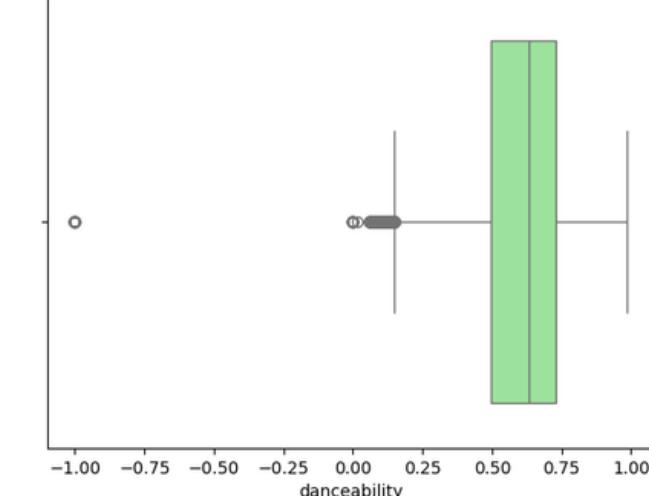
duration_min

count	62239.000000
mean	4.043391
std	1.883683
min	0.083333
25%	3.204000
50%	3.938517
75%	4.771725
max	76.358050

Distribution of Danceability



Boxplot of Danceability



danceability

count	62239.000000
mean	0.596768
std	0.186262
min	-1.000000
25%	0.497000
50%	0.631000
75%	0.730000
max	0.986000

- Popularity:

Most songs have low to moderate popularity (median = 7). Only a few tracks reach high scores — showing a right-skewed trend where viral hits are rare.

- Acousticness:

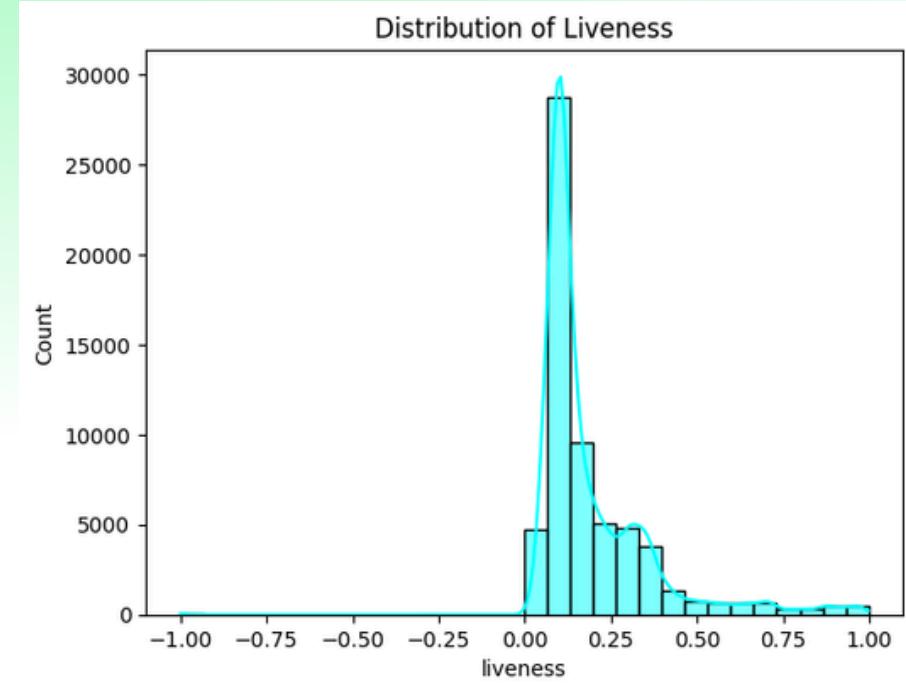
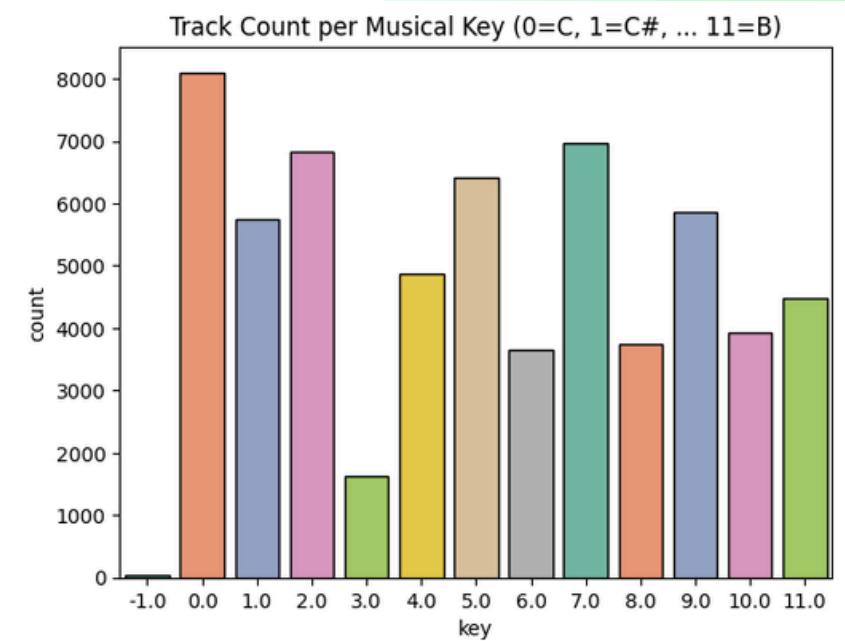
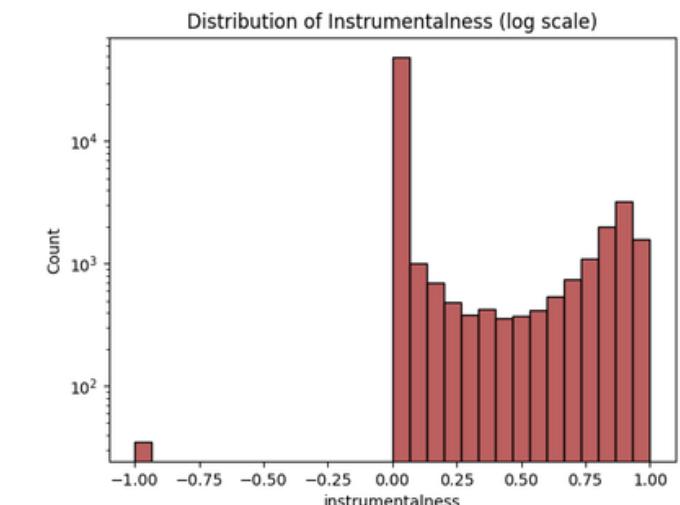
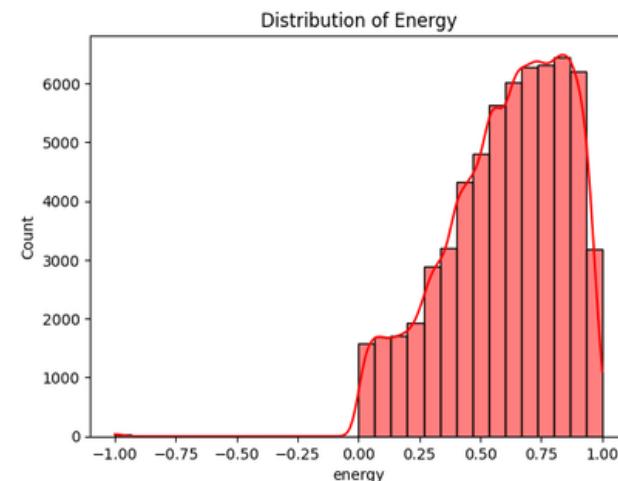
Heavily left-skewed; most songs are less acoustic and more electronic/studio-produced, with few pure acoustic tracks.

- Duration:

Average length ≈ 4 minutes; most songs lie between 3-5 min, forming a normal-like distribution with few very short or long tracks.

- Danceability:

Mostly moderate to high danceability (median ≈ 0.63). The dataset leans toward rhythmic, upbeat, mainstream music.



- **Univariate Analysis – Energy, Instrumentalness, Key & Liveness**

⚡ Energy:

- Most tracks have high energy levels (0.6-0.9), showing that the dataset is dominated by upbeat and lively songs. Only a small portion of songs are calm or low-energy.

🎸 Instrumentalness:

- Distribution is heavily concentrated near 0, meaning most tracks contain vocals, while a small group (values >0.8) are purely instrumental.

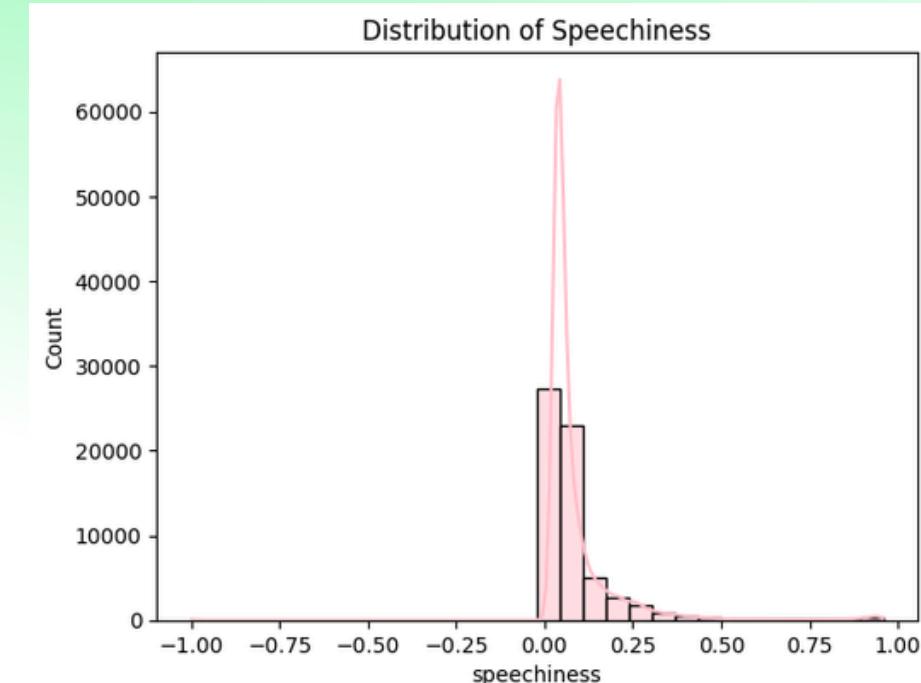
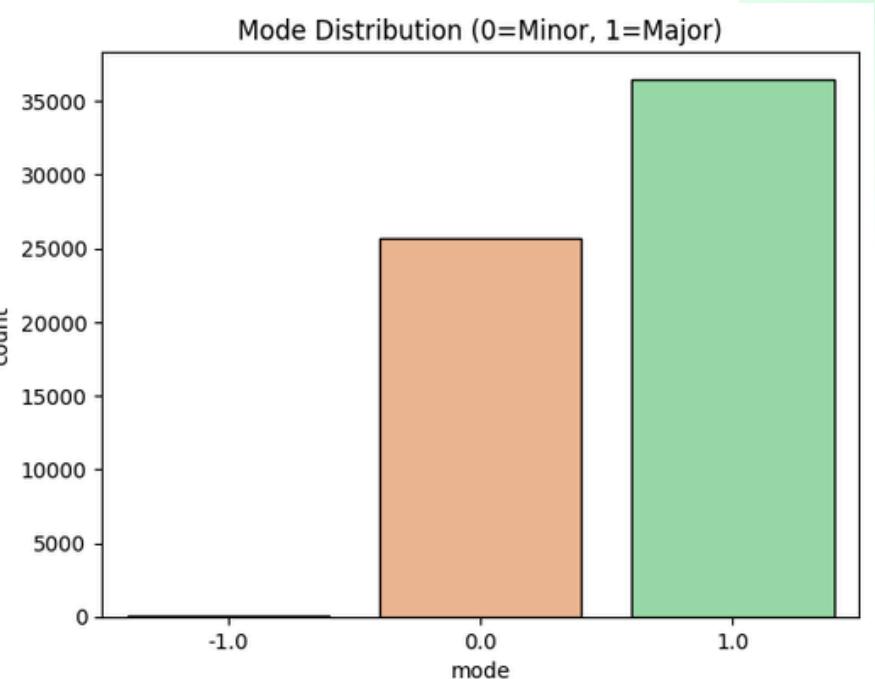
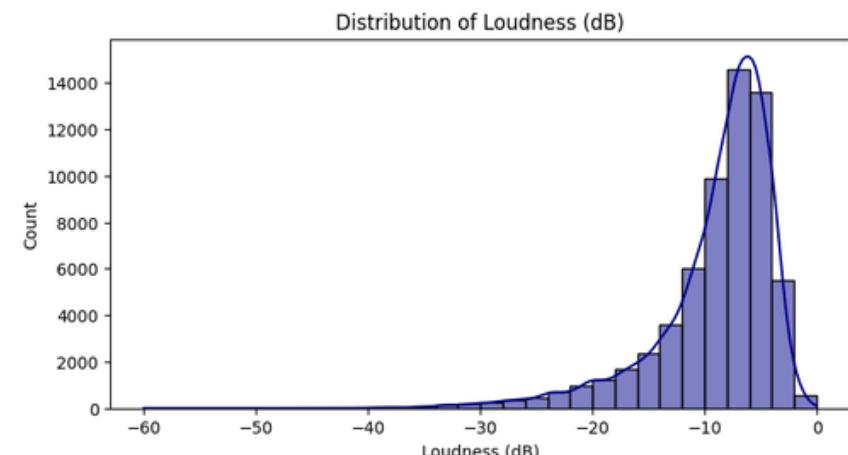
🎵 Musical Key:

- Tracks are fairly evenly spread across keys, with C, G, A, and D appearing slightly more often.
- This indicates no single key dominates, reflecting a diverse tonal variety across genres.

🎤 Liveness:

- Strong peak below 0.2, confirming that most songs are recorded in studios, not live performances.
- A few higher liveness values (>0.8) represent concert or live session recordings.





Univariate Analysis – Loudness, Mode, Speechiness & Tempo

Loudness:

- Most tracks range between -12 dB and -6 dB, typical of professionally mastered songs.
- Very quiet tracks below -40 dB are rare and likely outliers or data errors.

Mode:

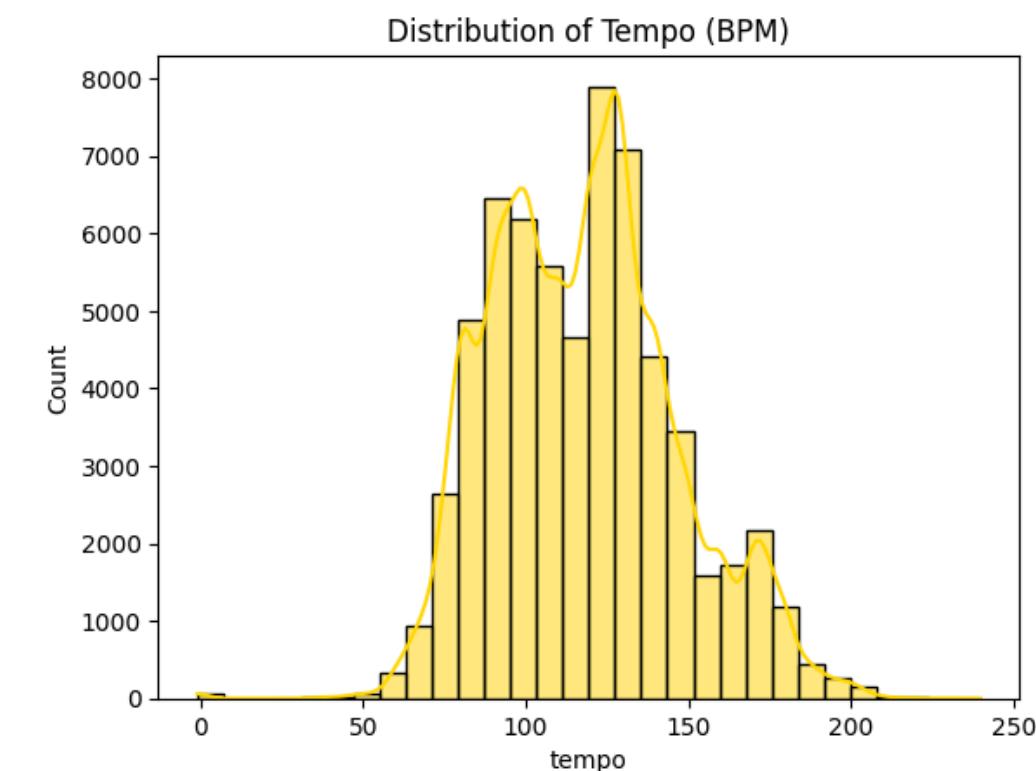
- About 60% of tracks are in Major mode (1) and 40% in Minor (0) — showing a slight preference for bright, uplifting tonalities.

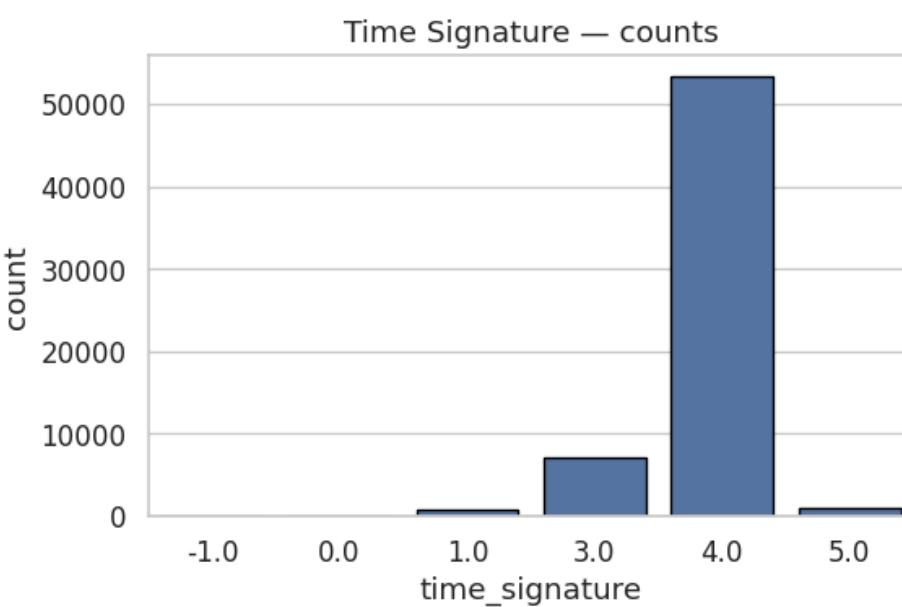
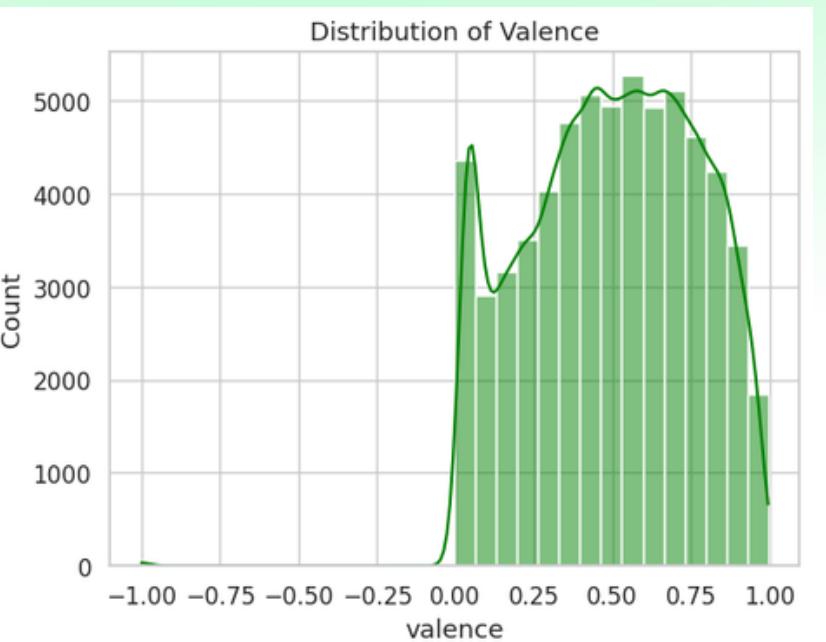
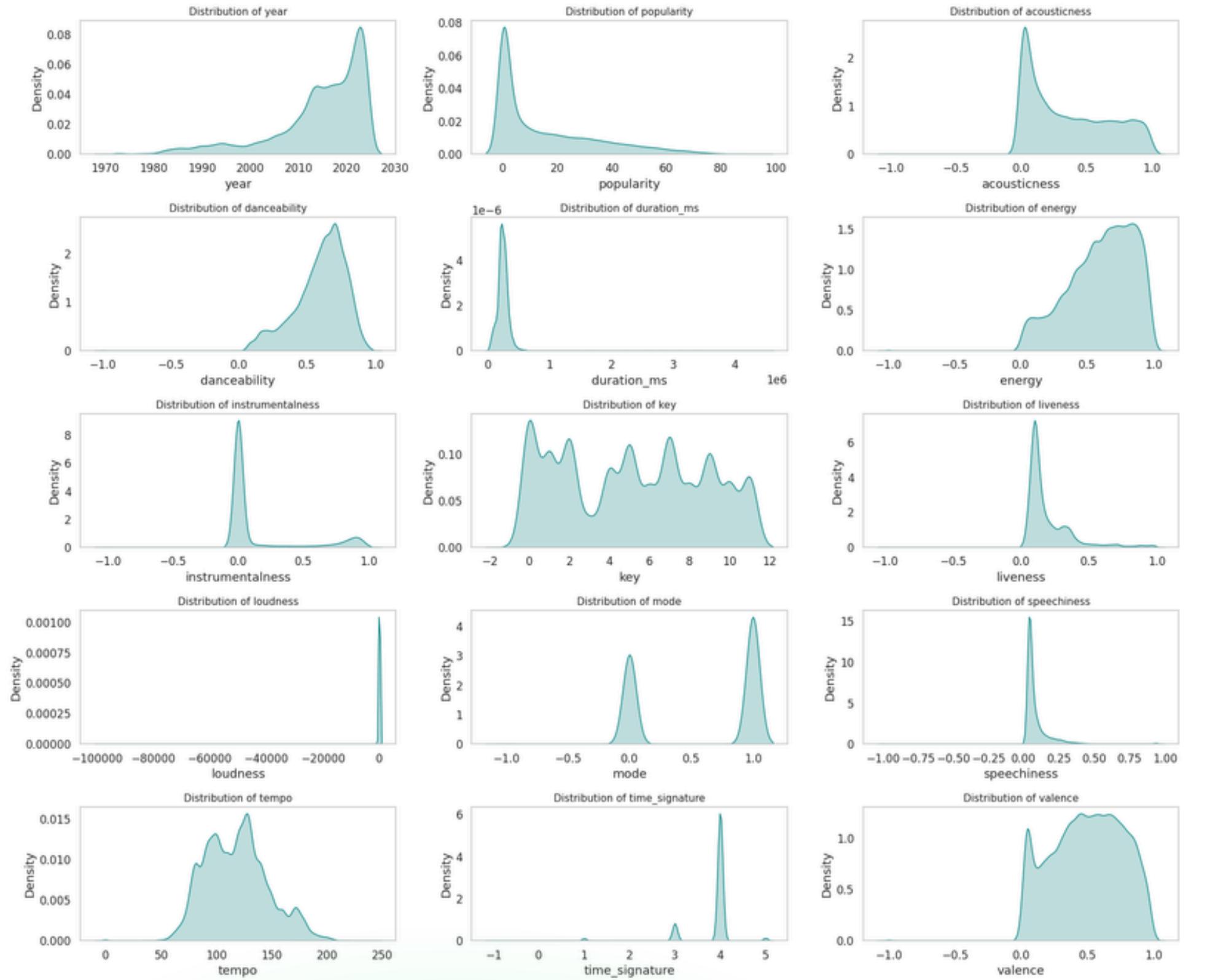
Speechiness:

- Heavily right-skewed, with most values below 0.1, meaning songs are lyrical and music-driven rather than spoken.
- Higher values correspond to rap or speech-heavy genres.

Tempo (BPM):

- Centered around 120-140 BPM, which is common for pop, dance, and upbeat genres.
- Extremely slow or fast tempos are rare, showing that mid-tempo music dominates the dataset.





Overall Insights from Univariate Analysis of Numerical Variables

Year:

- Majority of tracks are from 2000 onwards, with a sharp rise after 2010, showing that Spotify's catalog is modern and recent-heavy.

Popularity:

- The distribution is highly right-skewed — most songs have low popularity scores (<20), while only a few exceed 70, highlighting how viral hits are rare.

Acousticness:

- Most tracks are non-acoustic, with values close to 0, indicating a preference for electronic/studio production.

Danceability:

- Generally high (median ≈ 0.65), meaning most songs are rhythmic and suitable for dancing.

Duration (ms):

- Typical song lengths fall between 3-5 minutes, while a few long tracks (>20 min) act as outliers.

Energy:

- Skewed toward higher values (0.6-0.9) — songs are energetic, vibrant, and upbeat overall.

Instrumentalness:

- Most tracks are vocal-heavy (≈ 0), with only a few purely instrumental tracks (>0.8).

Key:

- Fairly uniform distribution across all 12 keys — no strong dominance of a specific key.

Liveness:

- Concentrated below 0.2, meaning most tracks were recorded in studios rather than live concerts.

Loudness (dB):

- Most tracks fall between -12 dB and -6 dB, typical of professionally mastered songs. Very quiet tracks (<-40 dB) are rare outliers.

Mode:

- About 60% major and 40% minor, suggesting a slight preference for brighter, uplifting tonalities.

Speechiness:

- Majority have low values (<0.1), indicating music-dominant tracks. Higher speechiness values correspond to rap or spoken-word styles.

Speechiness:

- Majority have low values (<0.1), indicating music-dominant tracks. Higher speechiness values correspond to rap or spoken-word styles.

Tempo (BPM):

- Centered around 120-140 BPM, common in pop and dance genres. Extremely fast or slow tempos are rare.

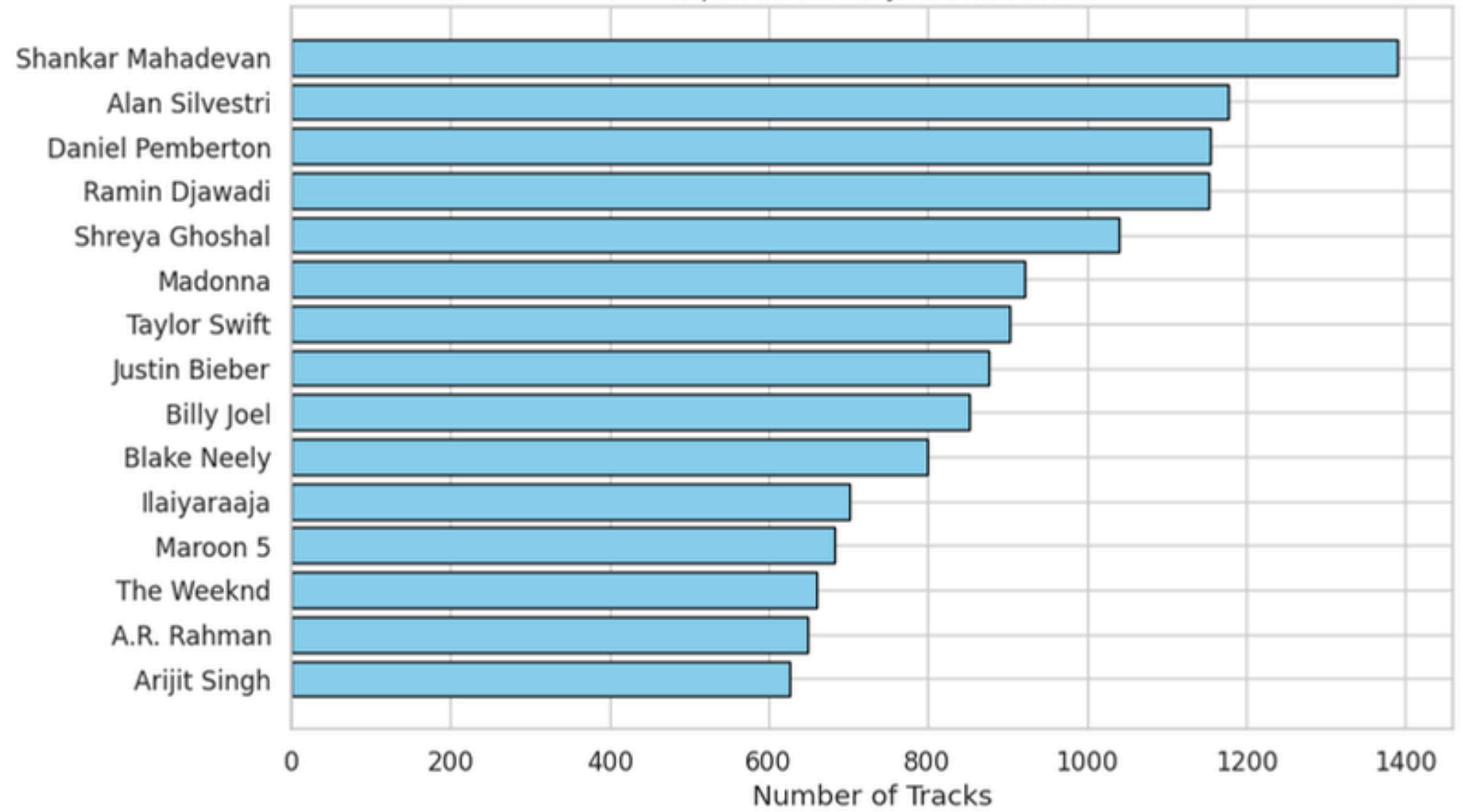
Time Signature:

- Dominated by 4/4 with small shares of 3/4 and 5/4. Invalid values (-1, 0, 1) should be cleaned.

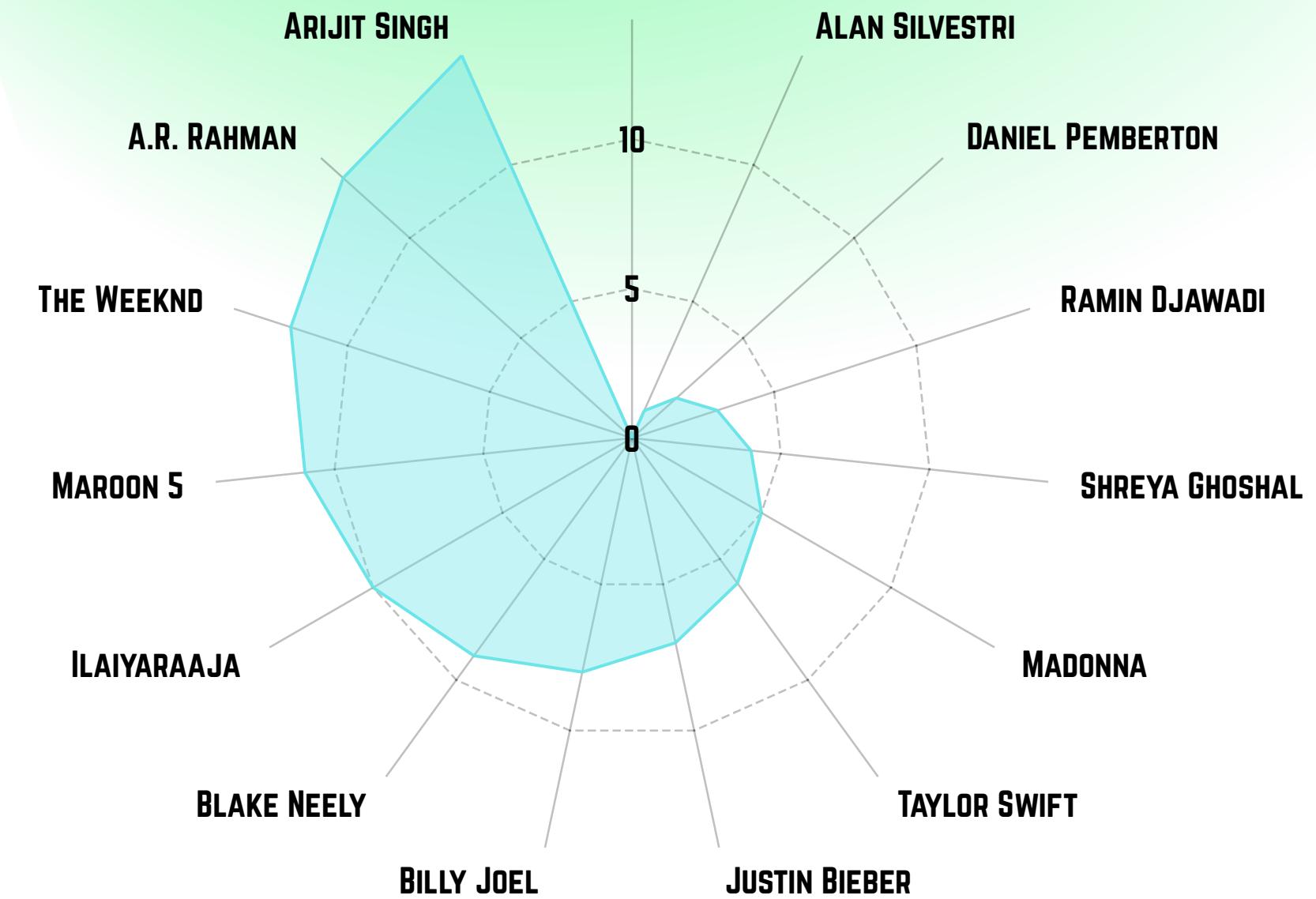
Valence:

- Most songs fall between 0.3-0.8, showing an overall positive and cheerful tone, with few dark or melancholic tracks (<0.2).

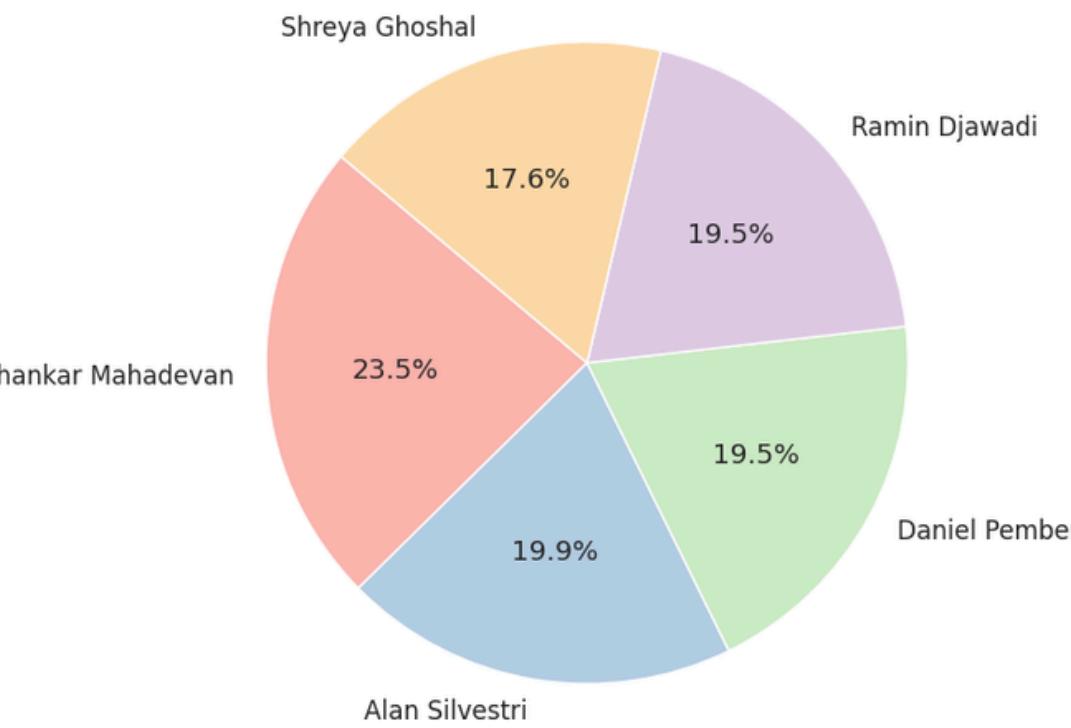
Top 15 Artists by Track Count

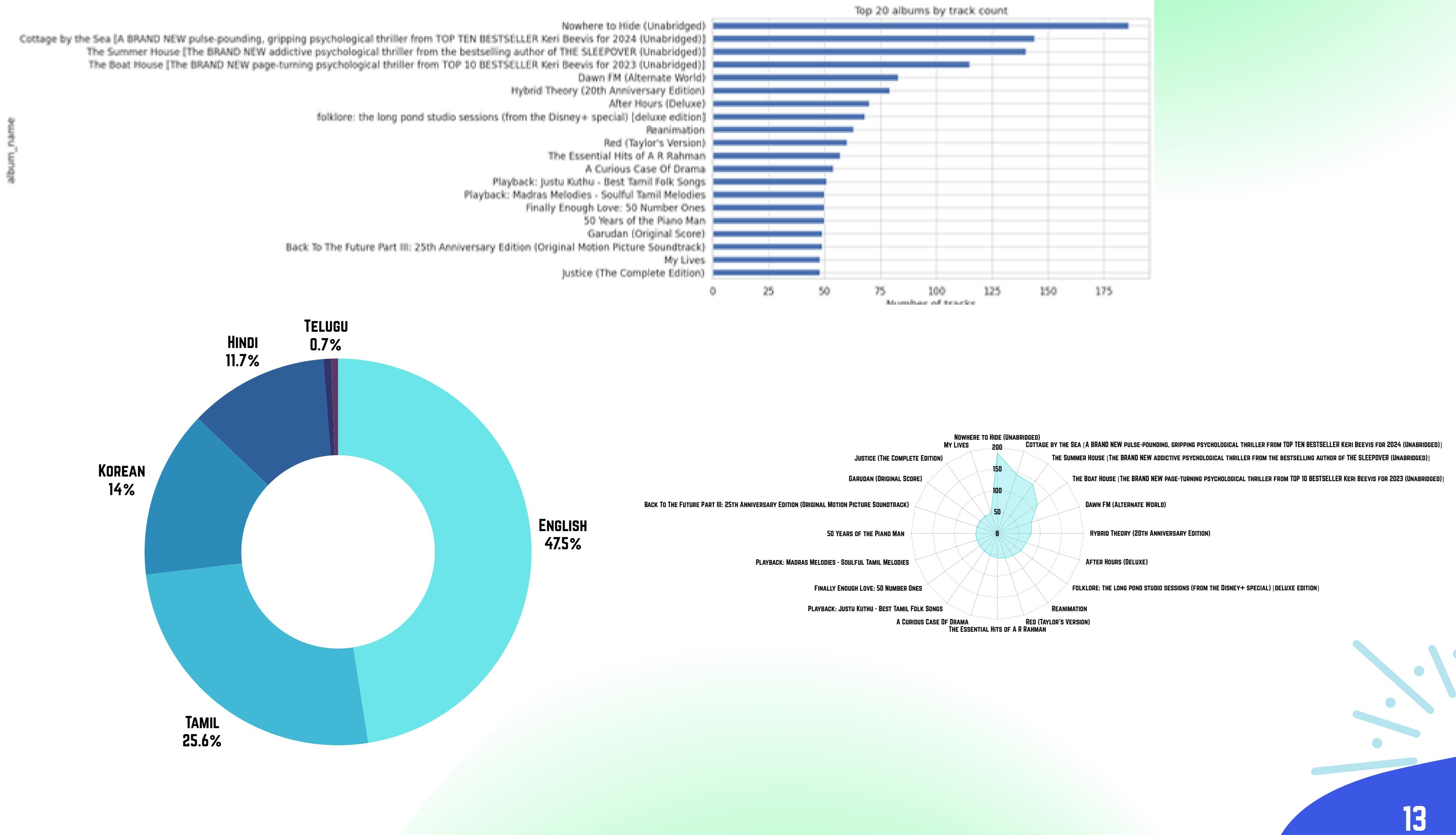


SHANKAR MAHADEVAN



Top 5 Artists by Track Count





🎯 OVERALL INSIGHTS FROM UNIVARIATE ANALYSIS OF CATEGORICAL VARIABLES

◆ Top Artists Insights

- Shankar Mahadevan leads the dataset with 1,391 tracks ($\approx 2.2\%$), followed closely by Alan Silvestri, Daniel Pemberton, and Ramin Djawadi, highlighting a strong presence of soundtrack composers.
- A diverse mix of Indian playback singers (Shreya Ghoshal, Arijit Singh, Ilaiyaraaja, A.R. Rahman) and global pop stars (Madonna, Taylor Swift, Justin Bieber, The Weeknd, Maroon 5) appear among the top contributors.
- This blend showcases a balance between film music, Indian playback, and global pop genres.
- The top 15 artists collectively contribute around 20% of the total dataset, indicating broad artist diversity.

◆ Top Albums Insights

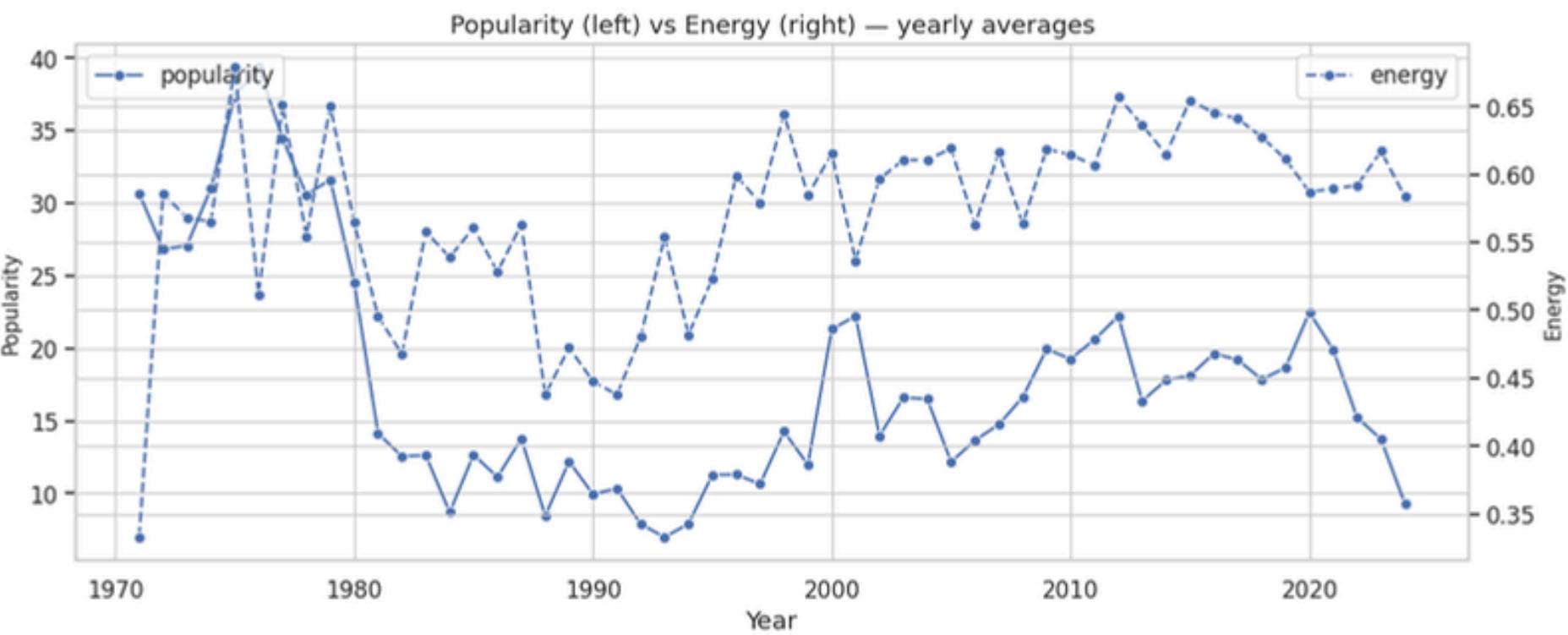
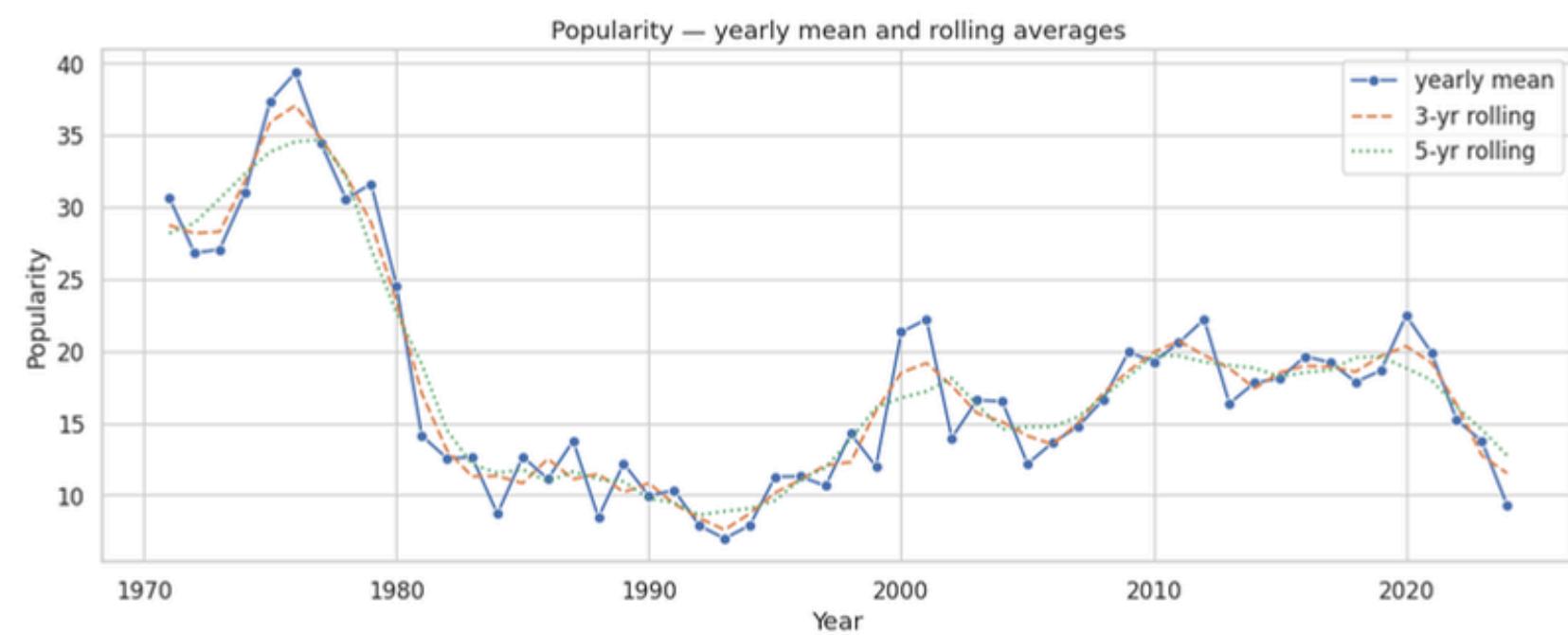
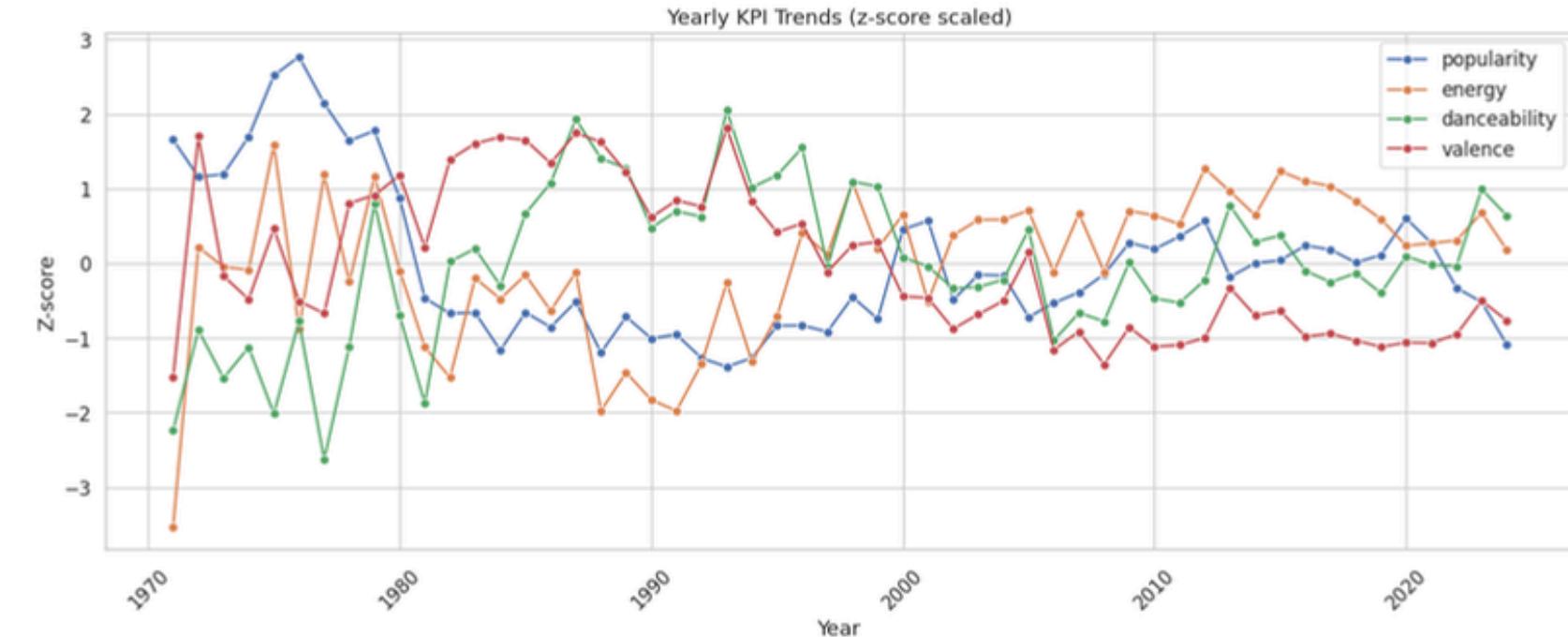
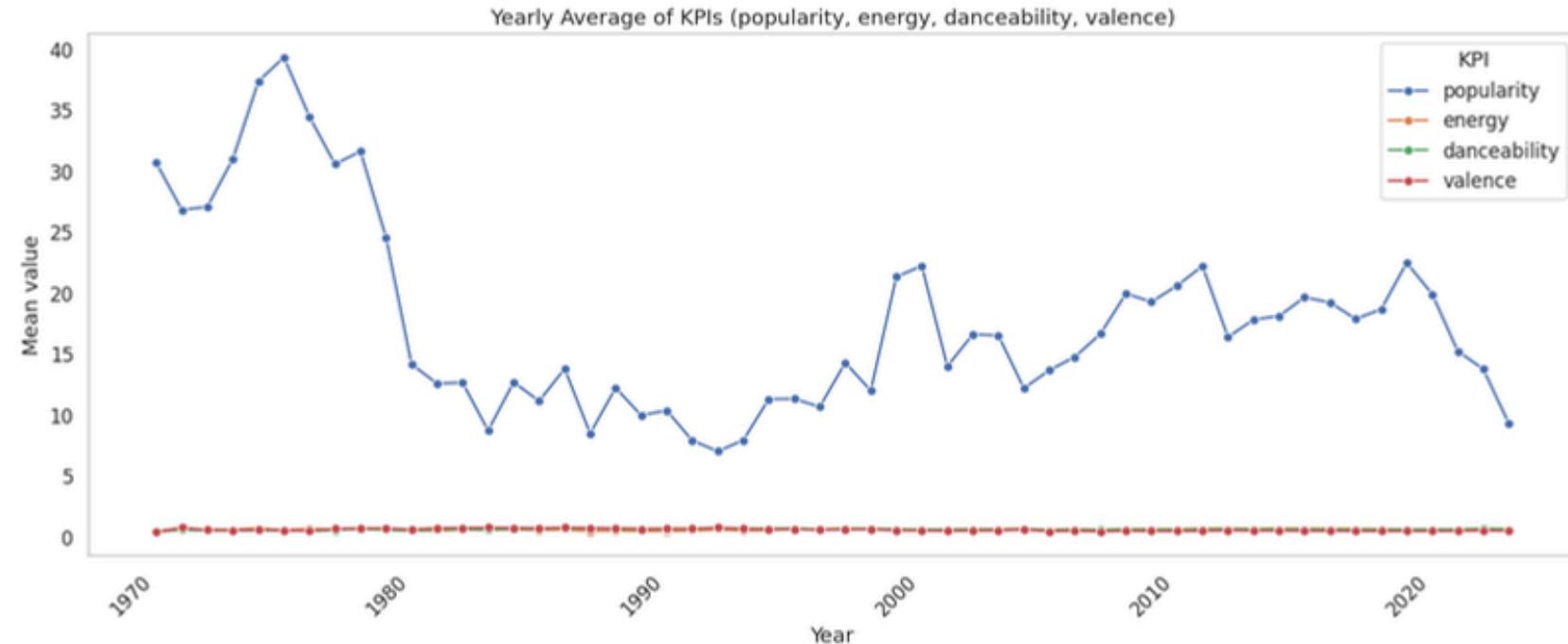
- "Nowhere to Hide (Unabridged)" is the largest album with 186 tracks, followed by audiobook-style albums like Cottage by the Sea, The Summer House, and The Boat House.
- Popular music albums such as Dawn FM, Hybrid Theory, After Hours, and Red (Taylor's Version) also feature prominently.
- Indian representation appears through albums like Essential Hits of A.R. Rahman and Tamil folk collections.
- Despite the top albums' popularity, the top 20 albums contribute less than 3% of the dataset, showing a wide distribution of songs across many albums.

◆ Language Distribution Insights

- English dominates the dataset with 47.5% (23,389 tracks).
- Tamil follows at 25.6% ($\approx 12,600$ tracks), reflecting a strong South Indian music influence.
- Korean (14%) and Hindi (11.7%) also hold notable shares, while Telugu (0.7%) and Malayalam (0.6%) are minimally represented.
- Together, English and Tamil account for over 70% of all tracks, showing a global and regional blend within Spotify's catalog.

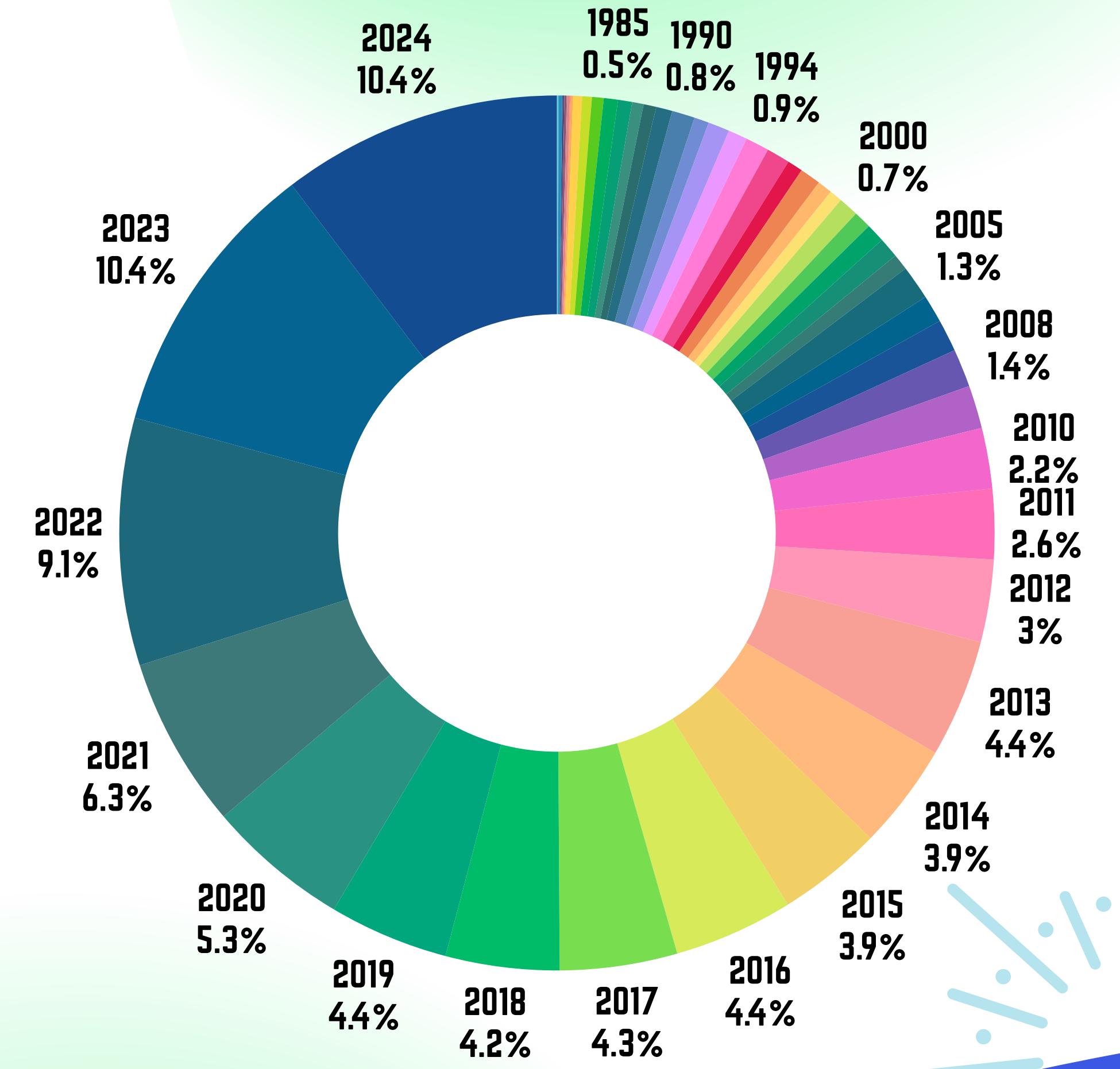
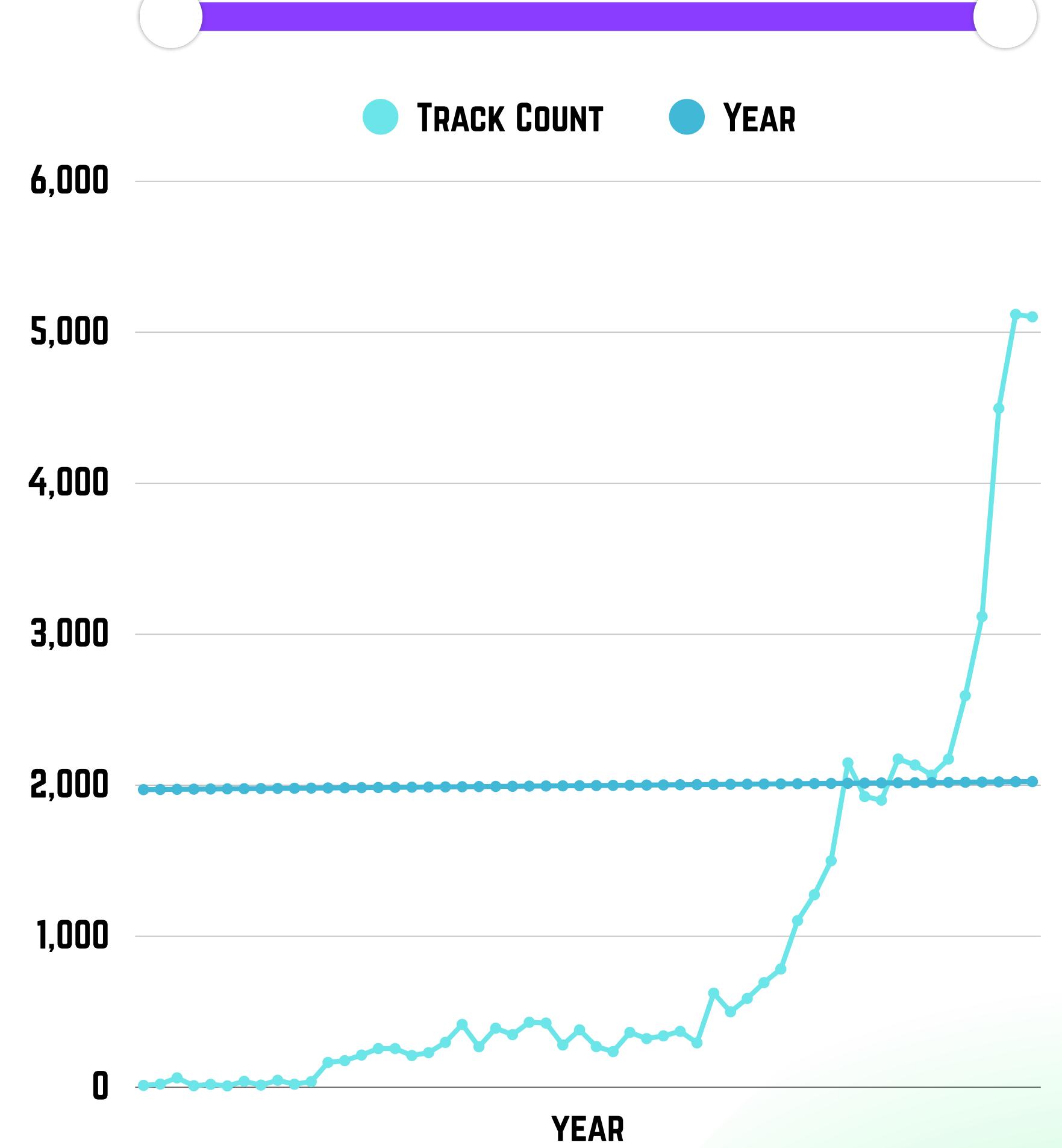


TIME SERIES ANALYSIS (ABSOLUTE VALUES OF KPIs)

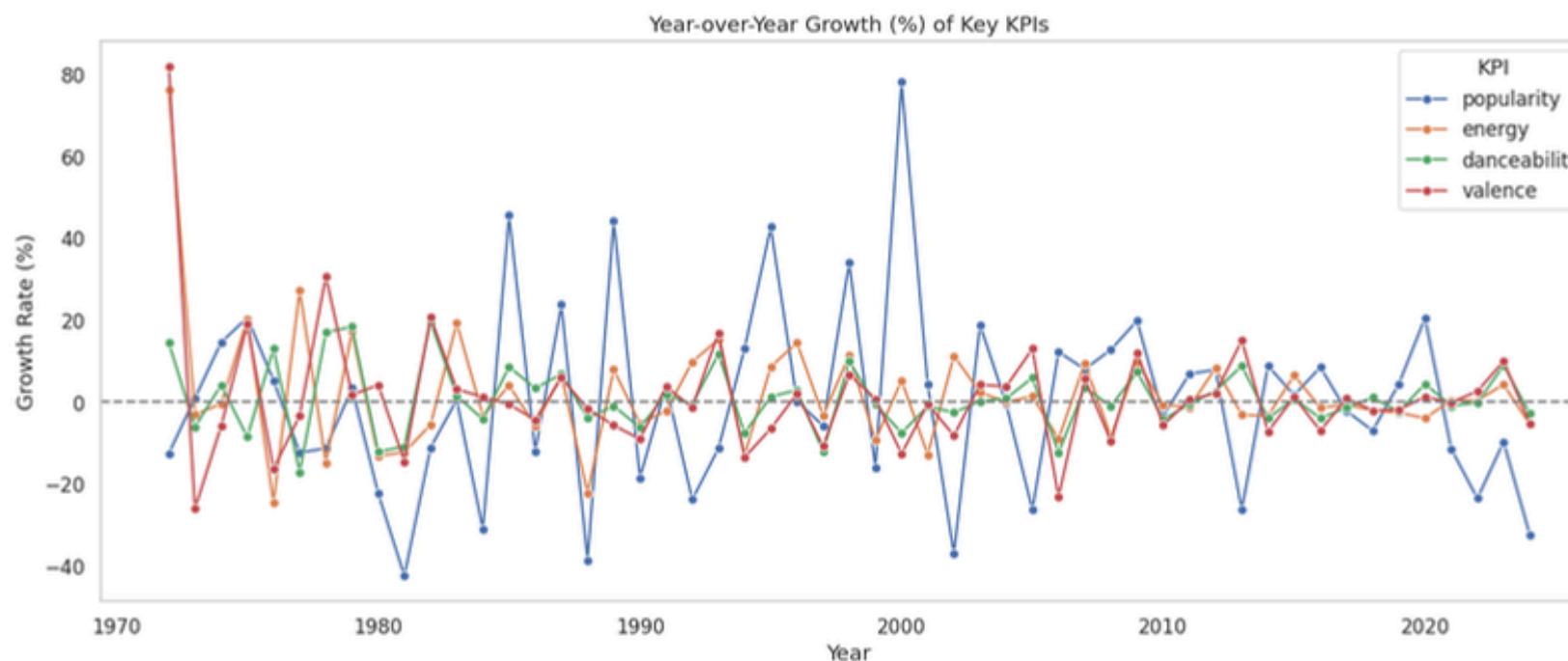


Overall KPI Trends:

- The average popularity of songs peaked during the 1970s and early 1980s, then dropped sharply through the 1990s, followed by a steady rise after 2010 – mirroring the digital streaming era boom.
- Energy, danceability, and valence show consistent upward movement from the 1990s onwards, indicating that modern songs have become more energetic, rhythmic, and emotionally positive.
- From 2015–2022, energy values stabilize around 0.6, showing a trend toward high production intensity and vibrant sonic design.



TIME SERIES ANALYSIS (GROWTH OF KPIs)

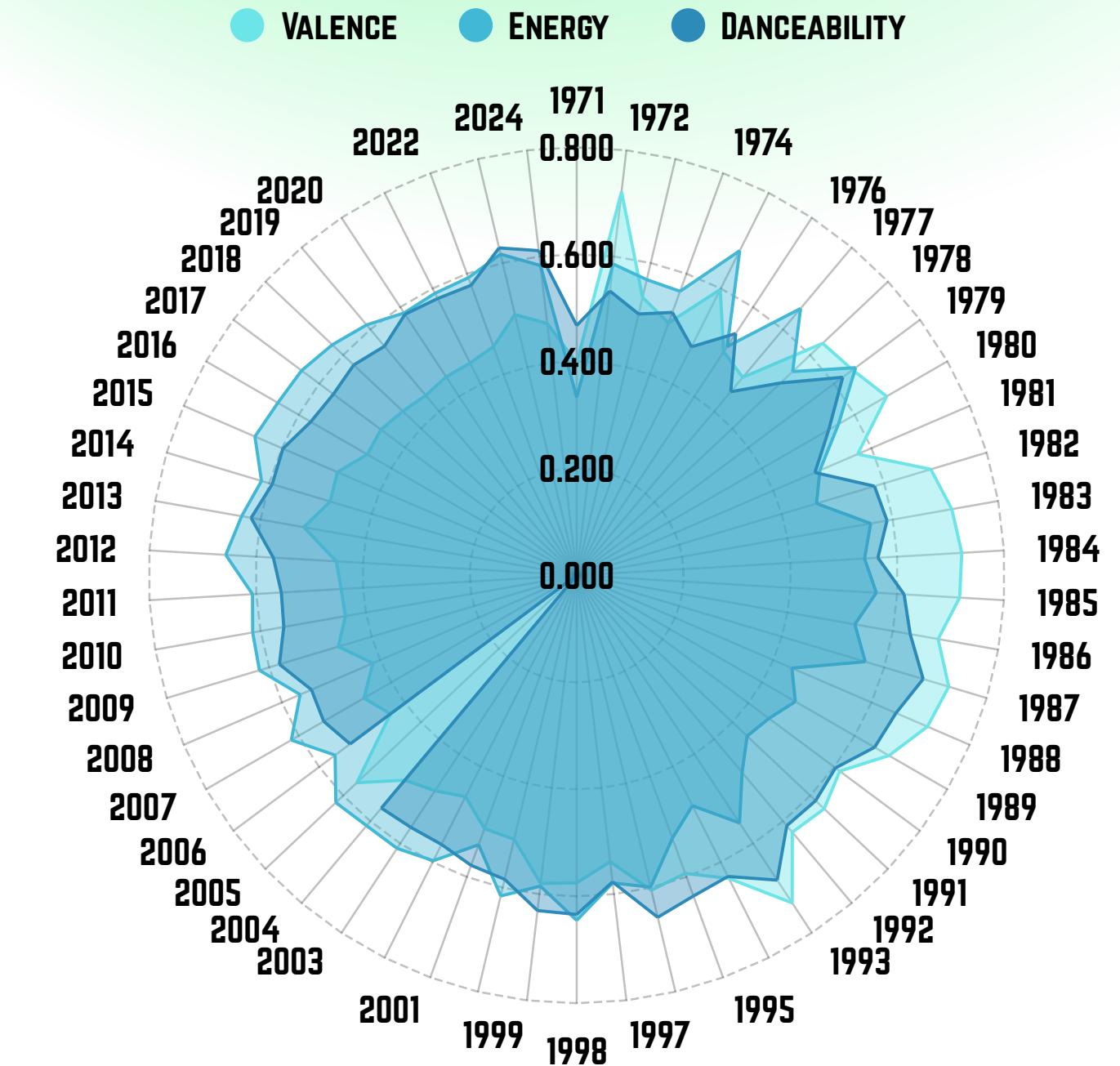


Time Series Analysis - Growth of Key Audio KPIs

- **Year-over-Year Growth Trends:**
- The growth rates of popularity, energy, danceability, and valence fluctuate significantly, reflecting shifts in listener trends and musical styles over the decades.
- Early years (1970s-1980s) show high volatility, suggesting inconsistent production volumes and recording diversity.
- From the 2000s onward, fluctuations become more stable, indicating a mature and steady music production era with consistent quality and style.
- Popularity growth occasionally spikes (notably around 2000 and 2015), likely tied to major digital platform adoption and viral streaming hits.
- Energy and danceability growth maintain positive trends post-2010, confirming modern music's consistent upbeat and rhythmic evolution.

Radar Chart Insights:

- The radar visualization shows steady and balanced growth in energy, valence, and danceability over the years.
- Periods between 1990-2010 saw gradual expansion in rhythmic intensity and emotional positivity, aligning with the rise of digital and EDM-driven music styles.



INSIGHTS – TIME SERIES GROWTH OF KPIs

Popularity Growth:

- The trend shows sharp spikes during major industry turning points – particularly the early 2000s (digital music revolution) and 2020s (streaming boom).
- This confirms that technological shifts and streaming accessibility directly drive popularity surges, as more users discover and engage with songs globally.

Energy & Danceability Relationship:

- Both energy and danceability follow similar upward growth patterns, often increasing together.
- Periods of rapid growth align with the emergence of upbeat, high-BPM genres like EDM, pop, and electronic fusion from the mid-2000s onward.
- This indicates that modern music has become progressively more rhythmic, energetic, and performance-driven.

Evolving Listener Preferences:

- Sustained positive growth across multiple KPIs since 2010 suggests that listeners increasingly favor songs that are energetic, danceable, and emotionally bright.
- The consistency of this trend across metrics like energy, valence, and danceability reveals a global shift toward modern, high-production, feel-good music that thrives on streaming platforms.

Valence (Positivity) Trends:

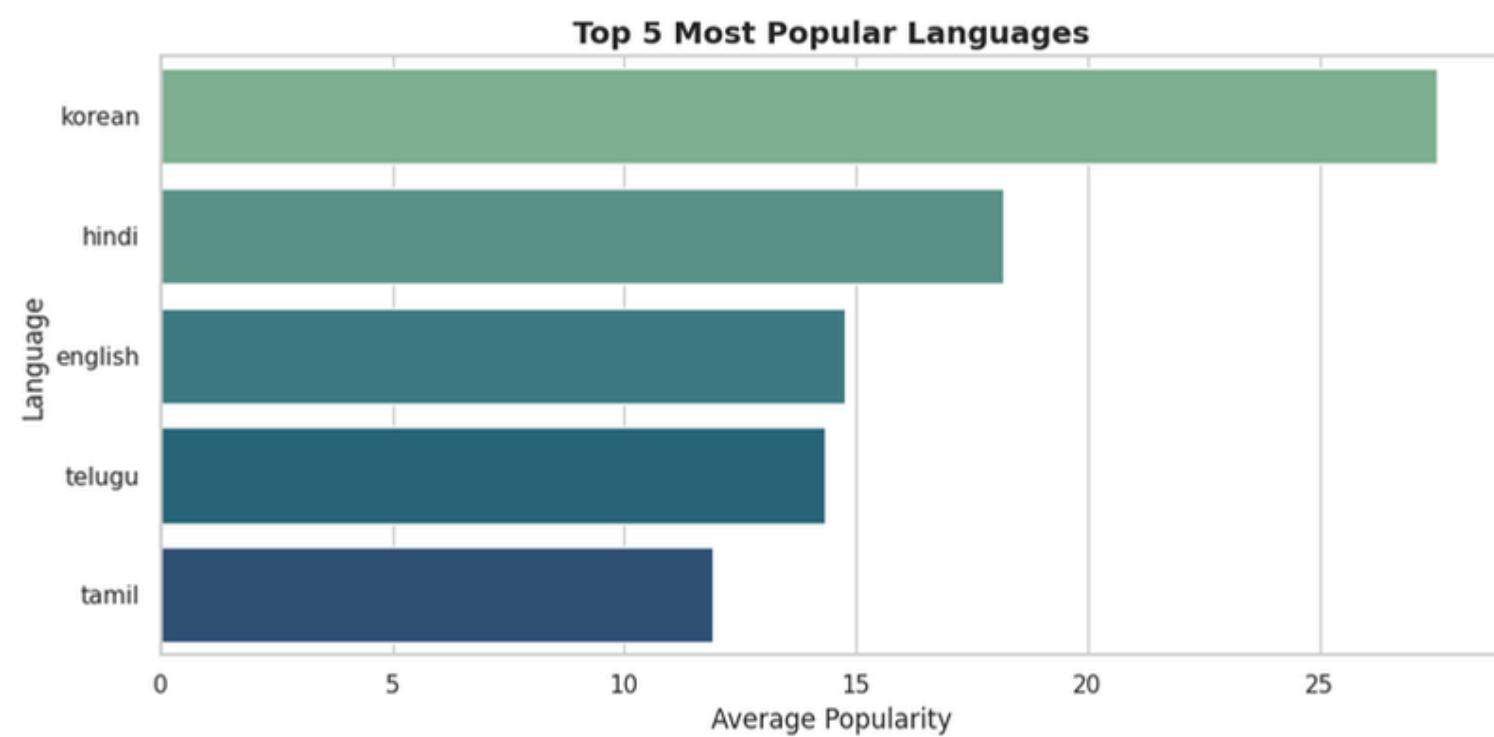
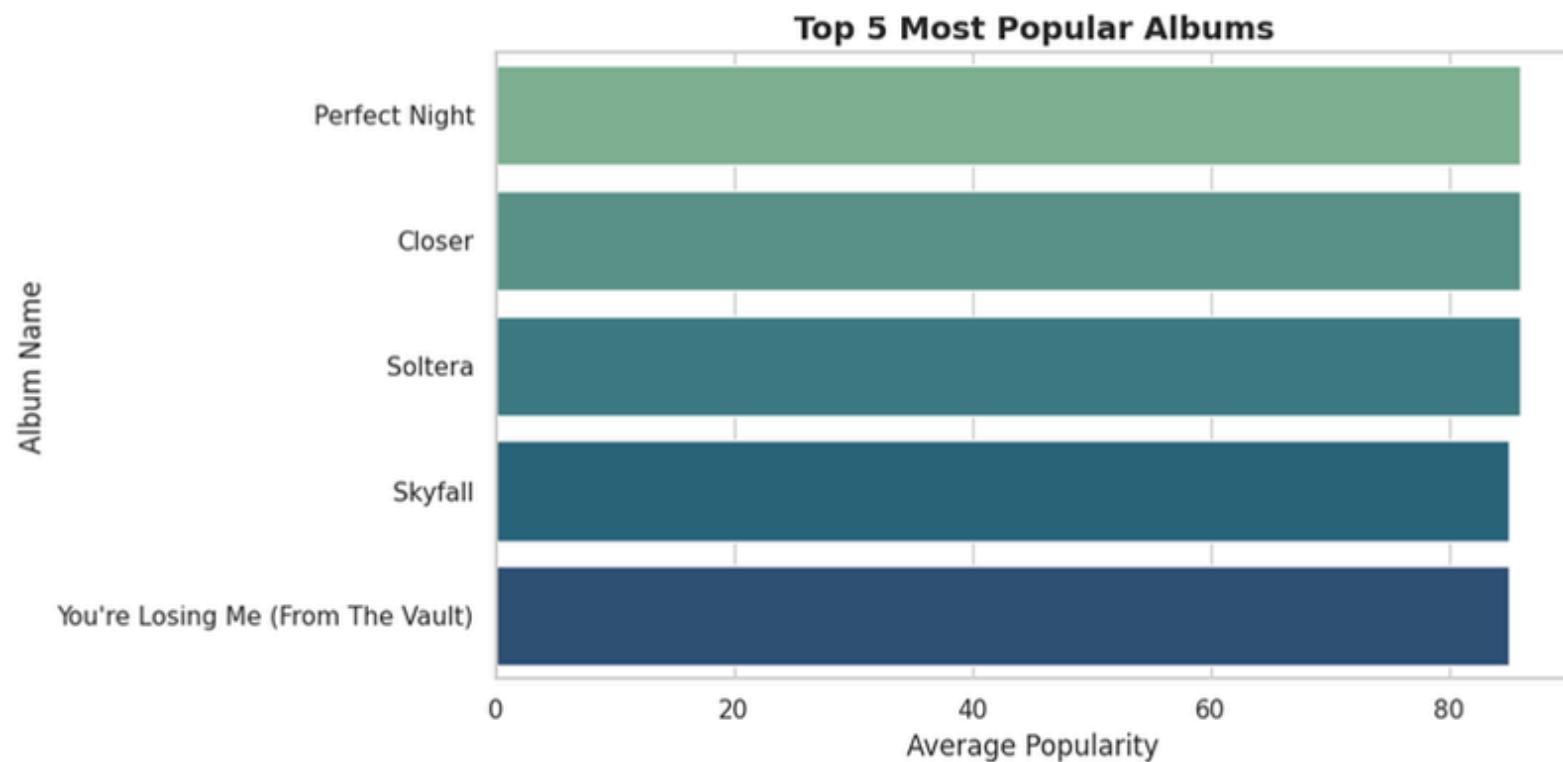
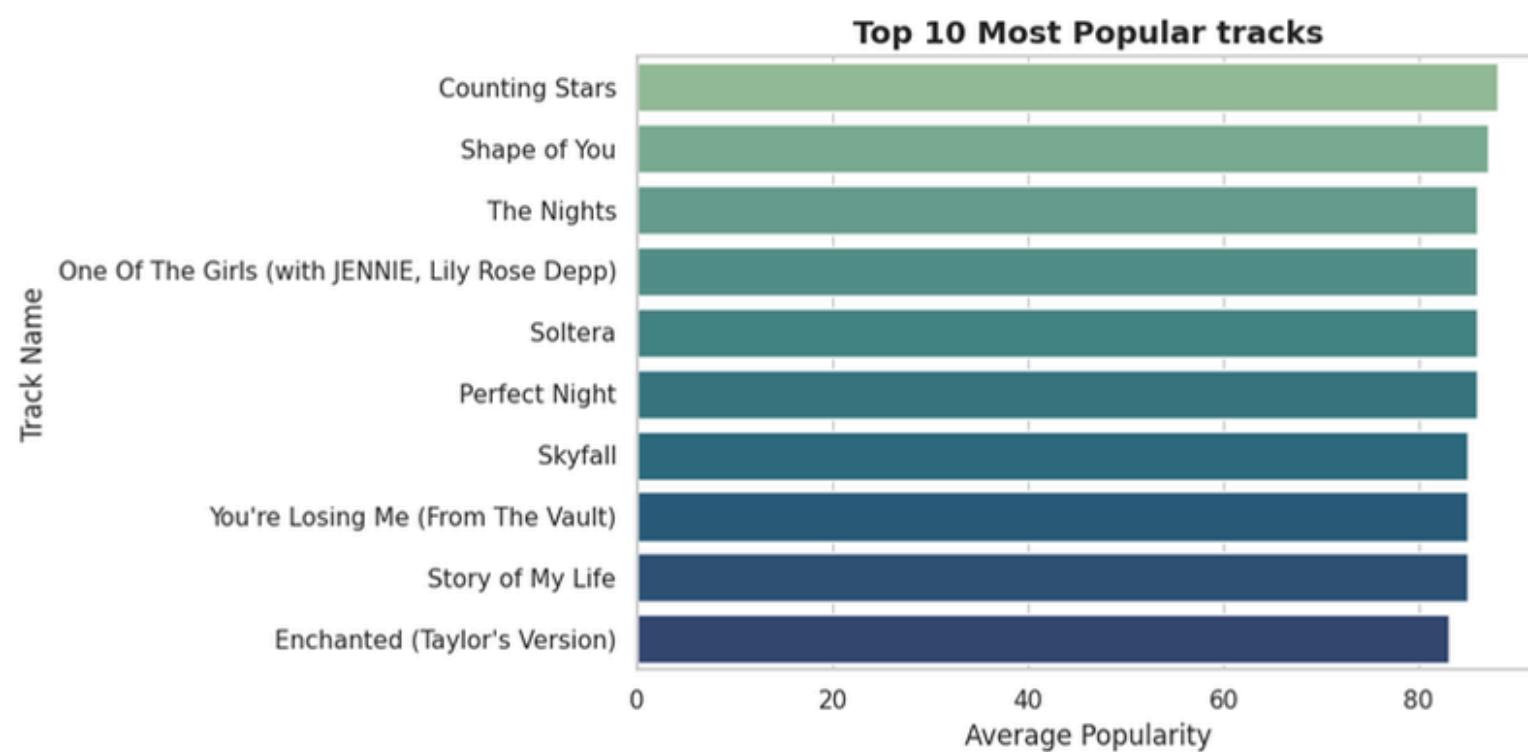
- Valence growth is fluctuating and cyclic, showing alternating periods of bright, happy tones and moodier, introspective styles.
- For example, drops in valence growth correspond to eras dominated by emotional or darker genres like grunge, alternative rock, and trap, while positive spikes reflect cheerful pop and dance eras.

Negative Growth Periods:

- Occasional sharp declines in growth – especially visible before 1980 or around 1995 – are often data-driven artifacts caused by smaller yearly track counts or incomplete metadata, not actual declines in music quality or appeal.
- Such drops should be interpreted with caution, as they may reflect sampling limitations rather than real-world listener trends.

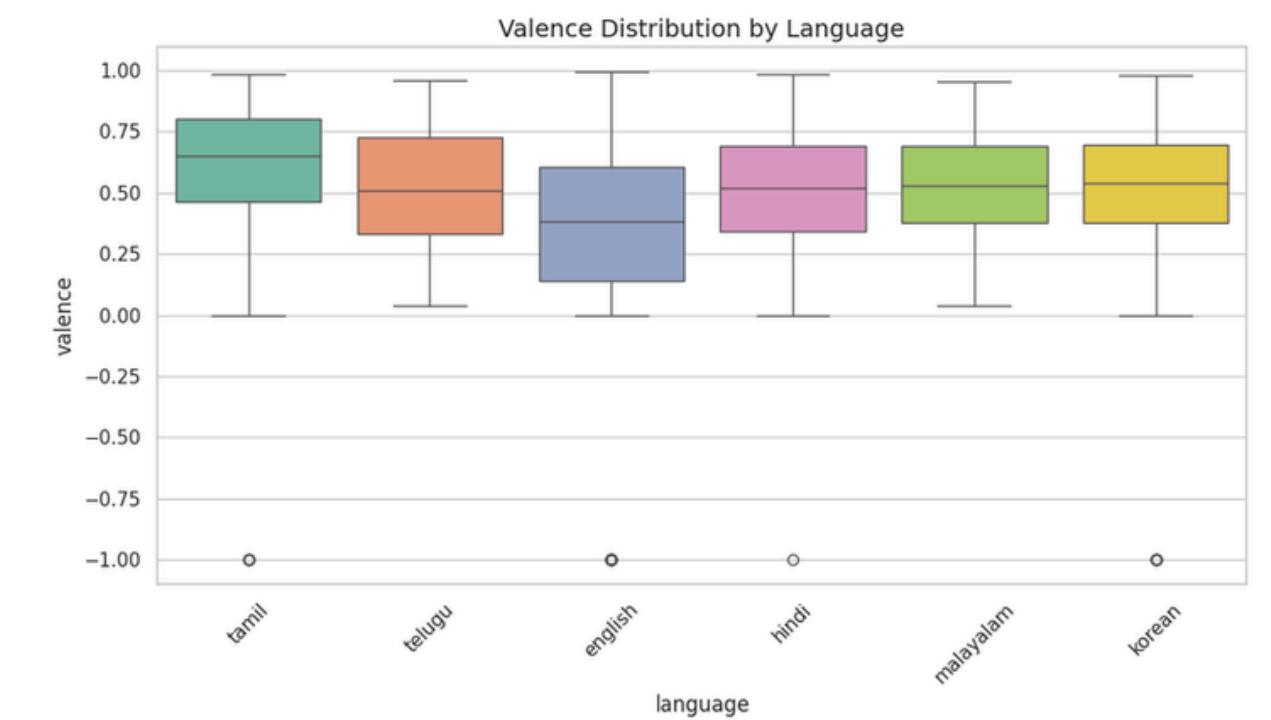
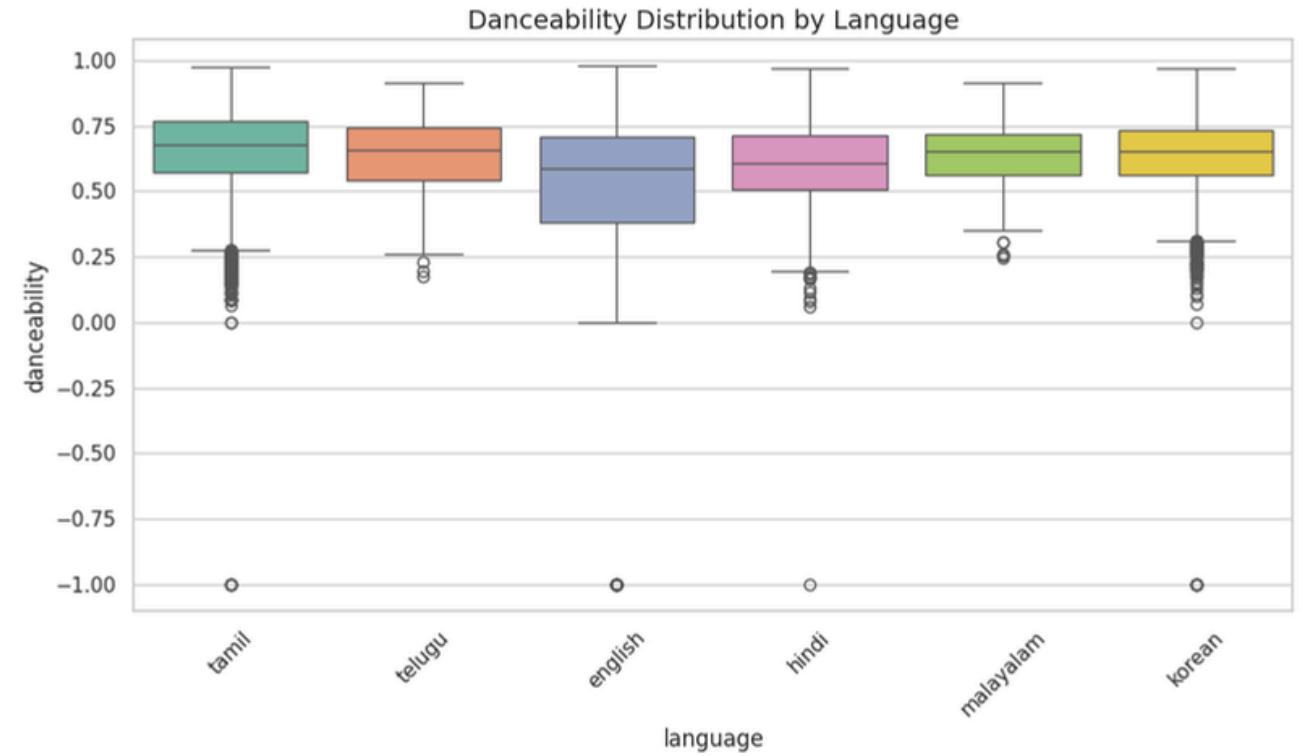
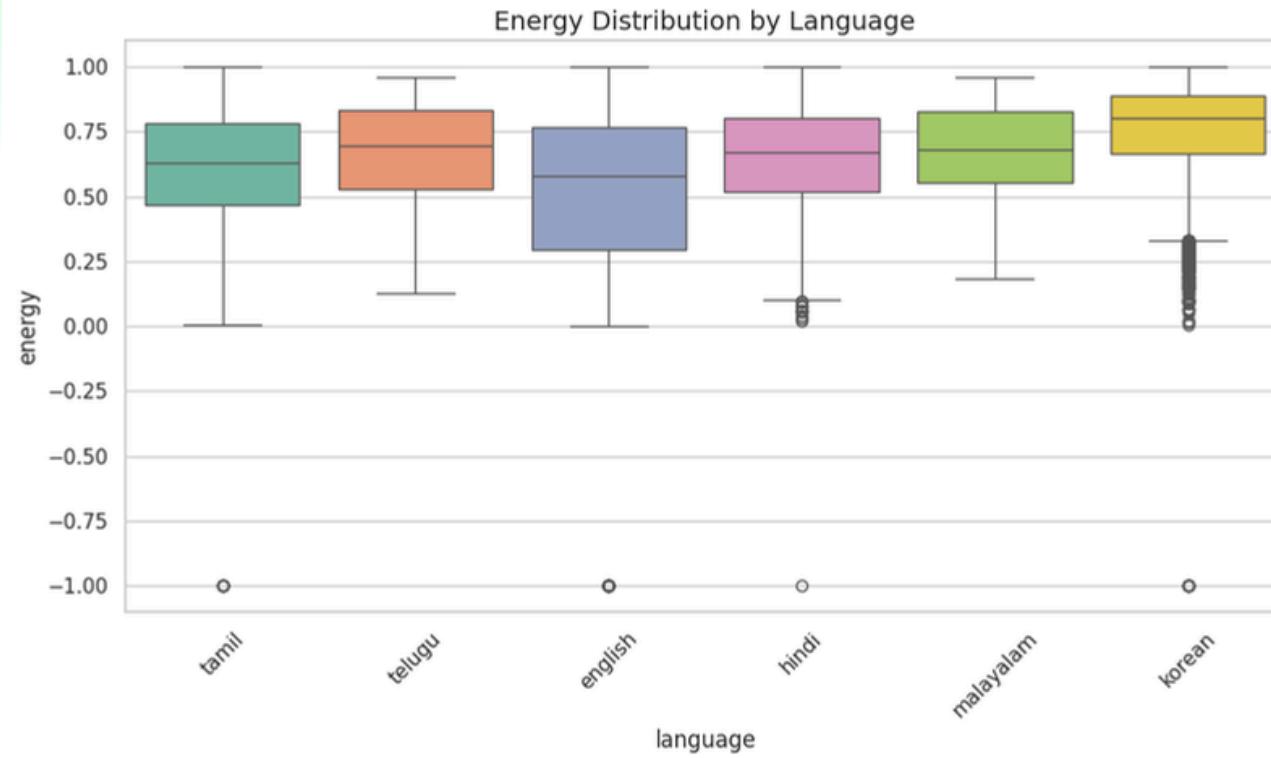
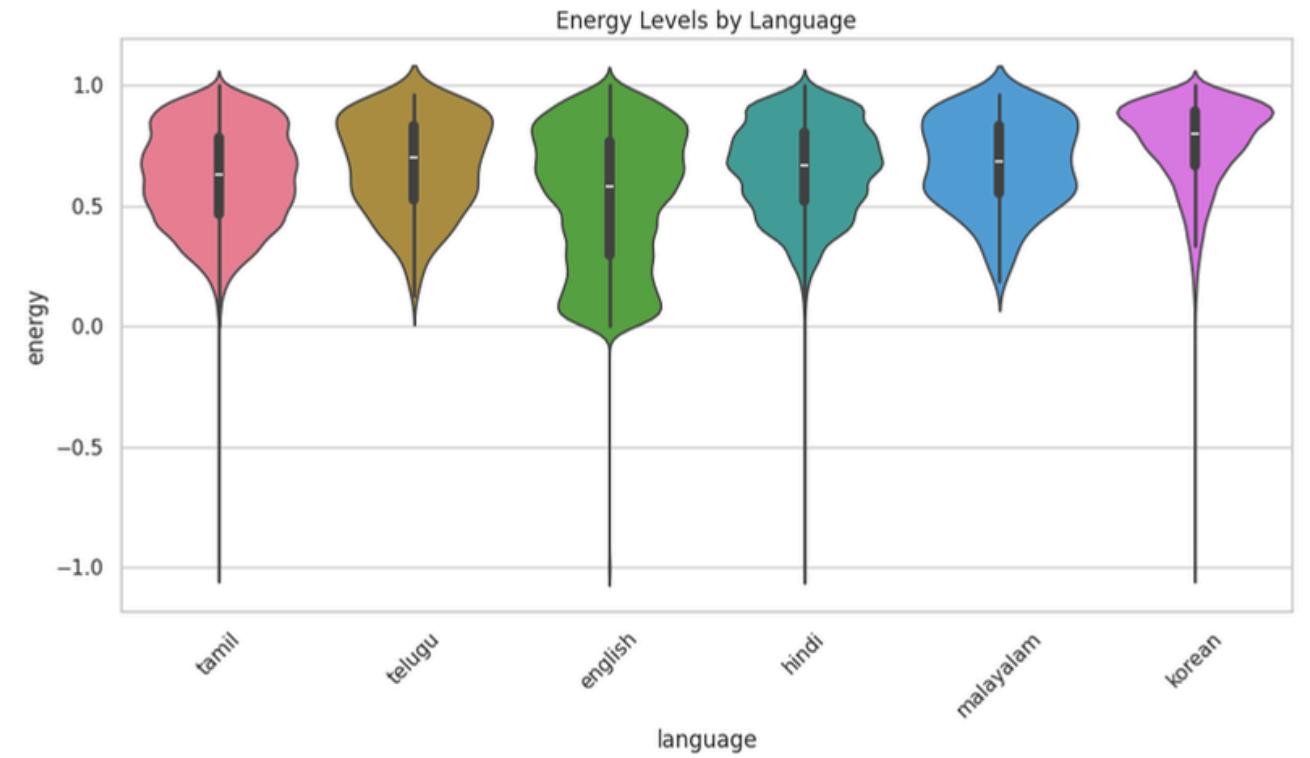
Final Takeaway:

- The Time Series Growth of KPIs shows that popularity aligns with technological and cultural milestones, while energy, danceability, and valence jointly define the modern music era.
- Over time, Spotify's catalog has transformed from low-energy, mood-diverse compositions to vibrant, high-energy, rhythm-driven tracks, perfectly matching today's listener expectations and digital consumption habits.



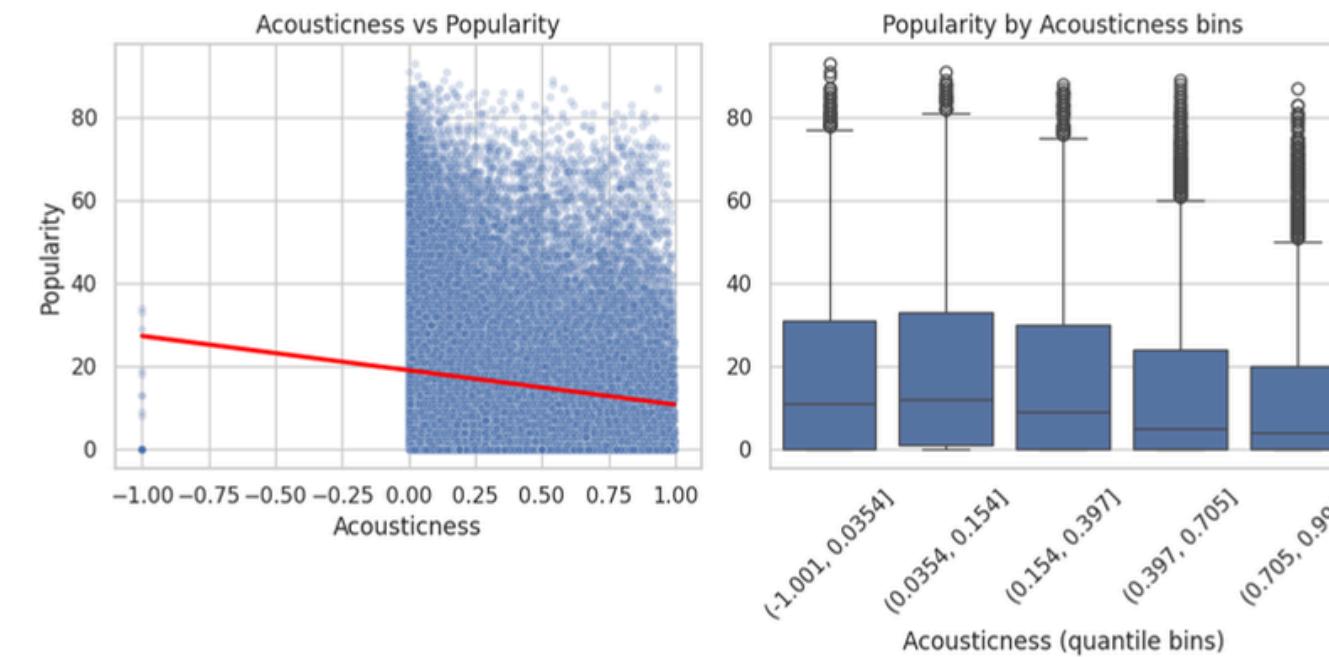
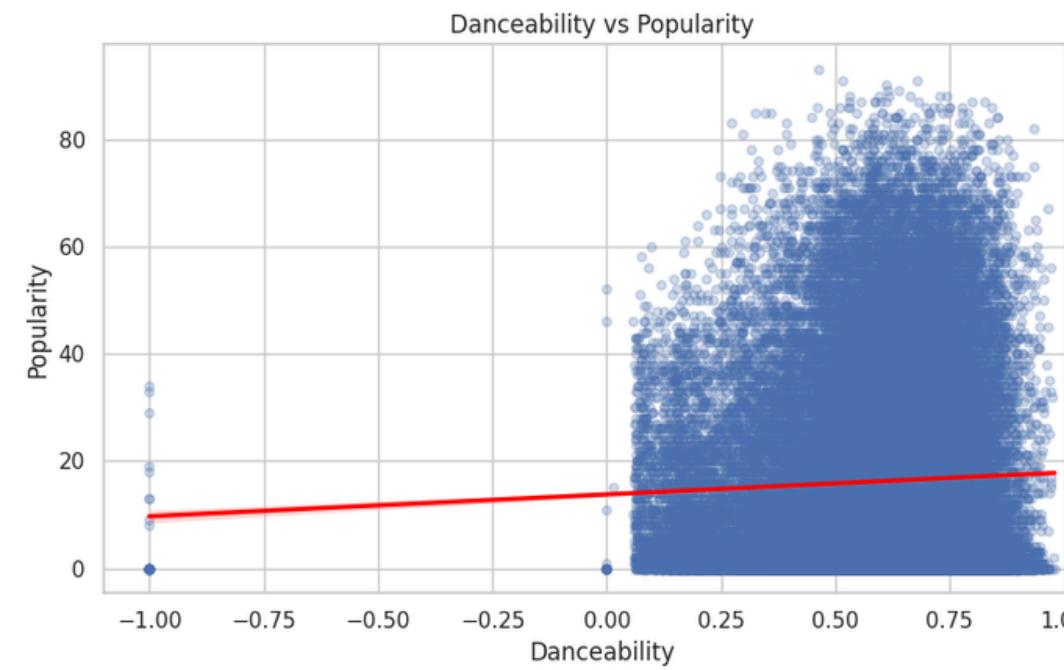
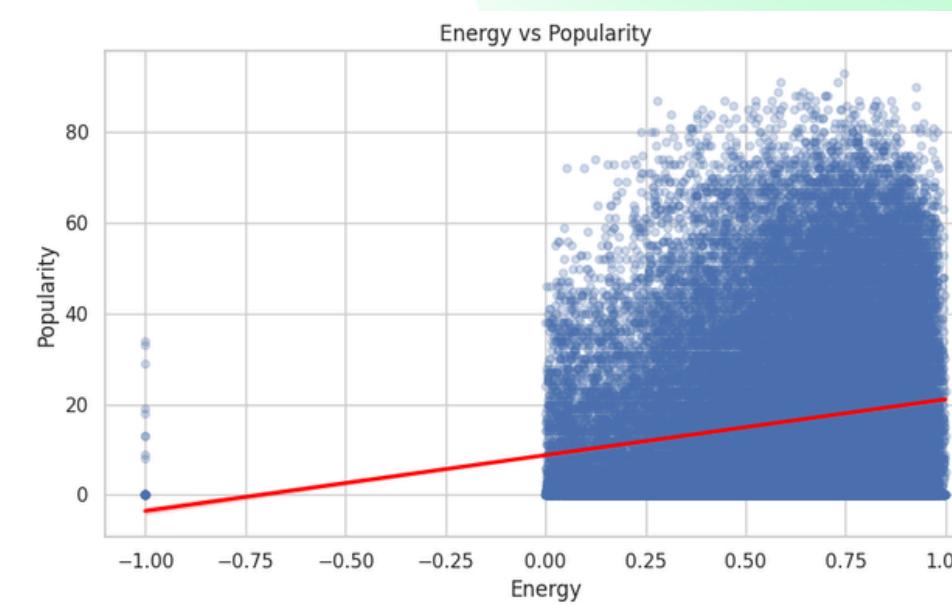
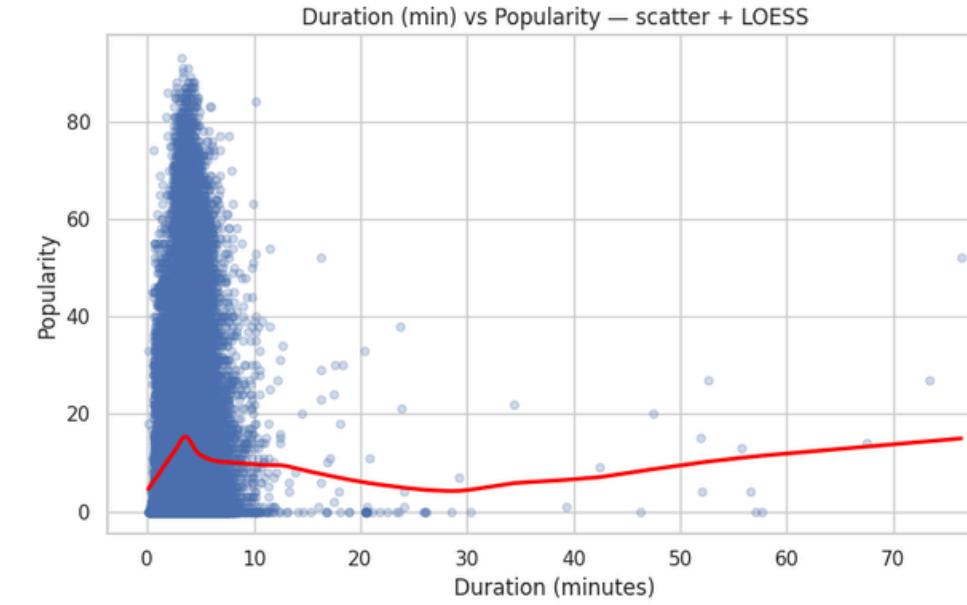
🎯 Insights – Track, Album & Language Popularity

- Top Tracks: “Counting Stars,” “Shape of You,” and “The Nights” dominate Spotify’s global charts — highlighting how upbeat, catchy, and emotionally resonant pop songs maintain long-term appeal.
- Top Albums: Perfect Night, Closer, and Soltera stand out for their consistent popularity, indicating strong listener engagement across both mainstream and niche audiences.
- Top Languages: Korean music leads with the highest average popularity — reflecting the global K-pop wave — followed by Hindi and English, showing how regional and international hits coexist in streaming trends.
- Overall, popularity trends reflect a blend of global pop dominance and diverse linguistic reach, driven by cross-cultural collaborations and streaming globalization.



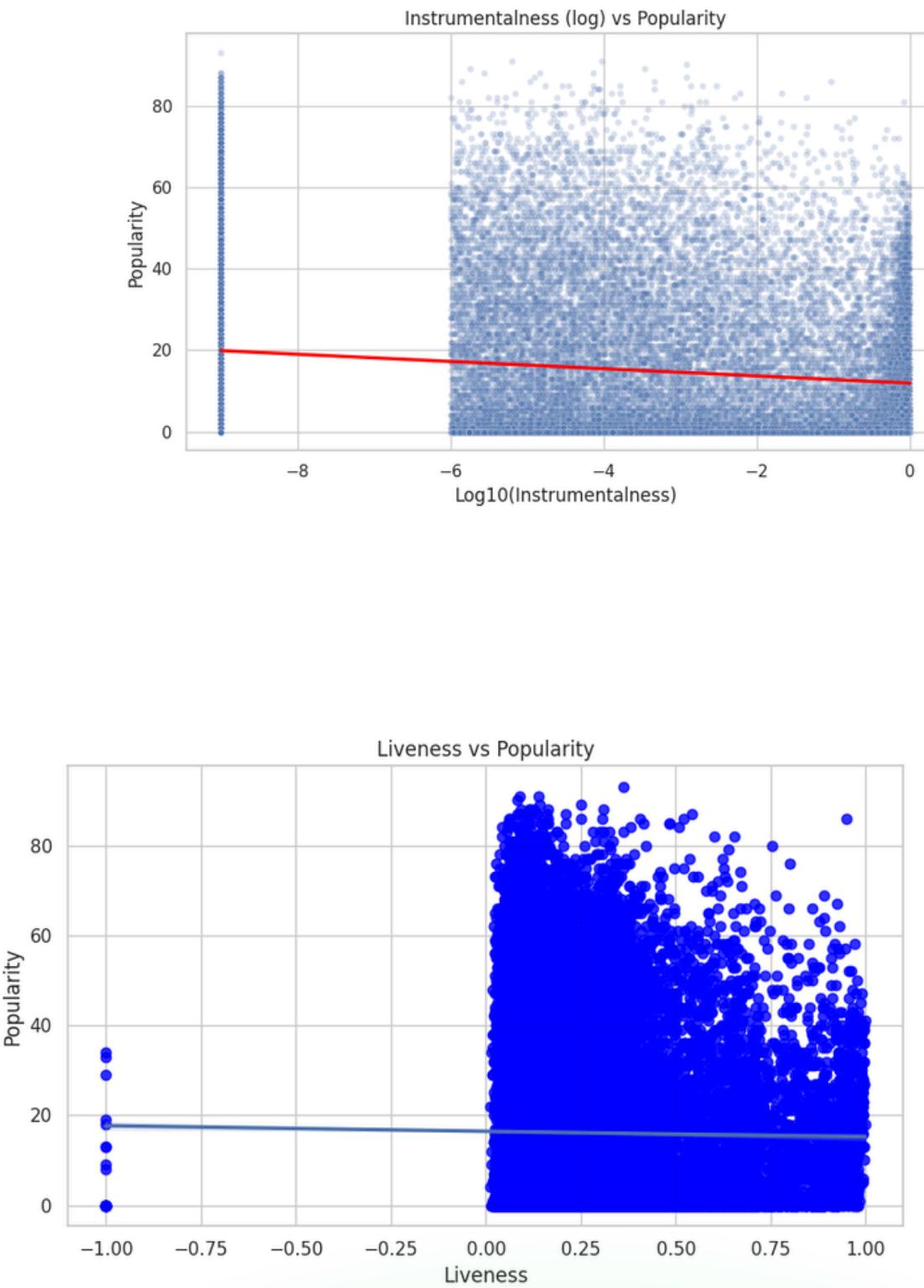
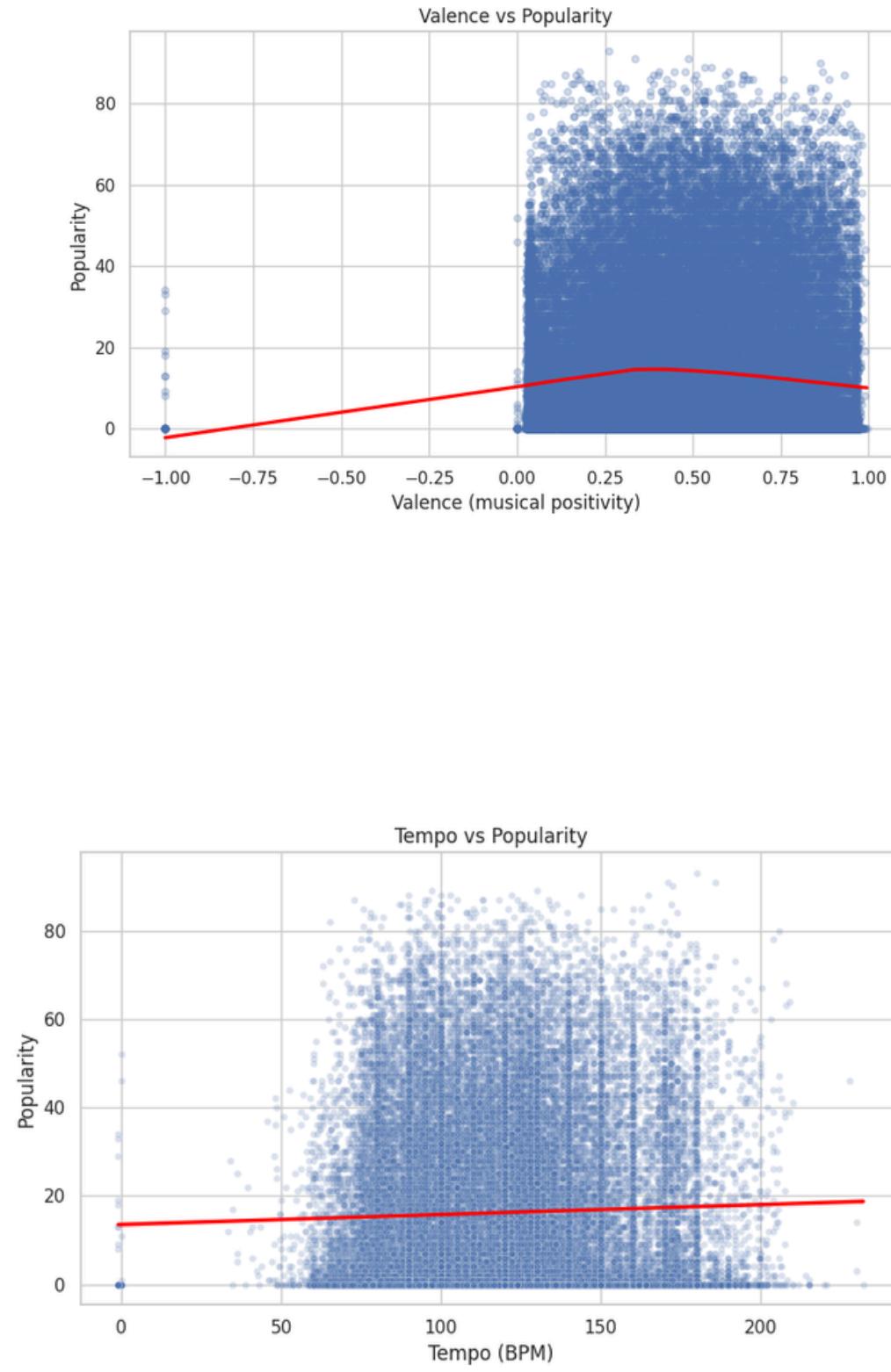
🎯 Insights – Audio Feature Trends by Language

- Korean and Tamil songs show higher energy and danceability, reflecting vibrant, performance-driven music styles.
- English tracks display wider valence variation – covering both upbeat and mellow moods.
- Hindi and Telugu songs maintain moderate levels across features, showing balanced emotional tone and rhythm.
- Overall, each language showcases a distinct audio personality, blending cultural rhythm with modern production trends.

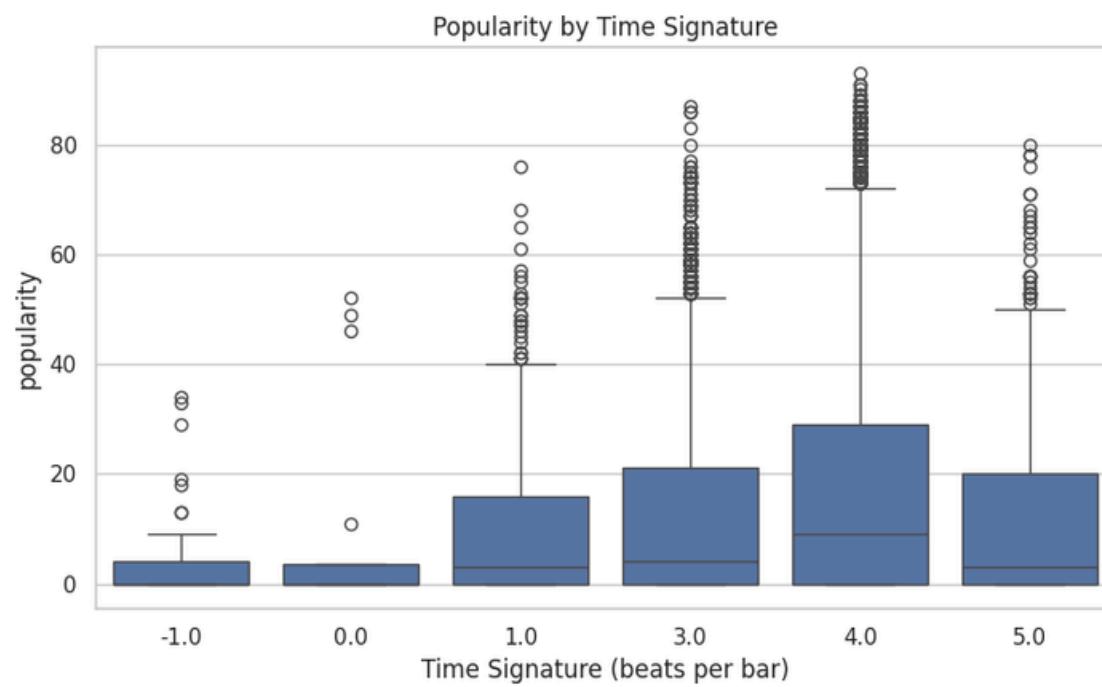
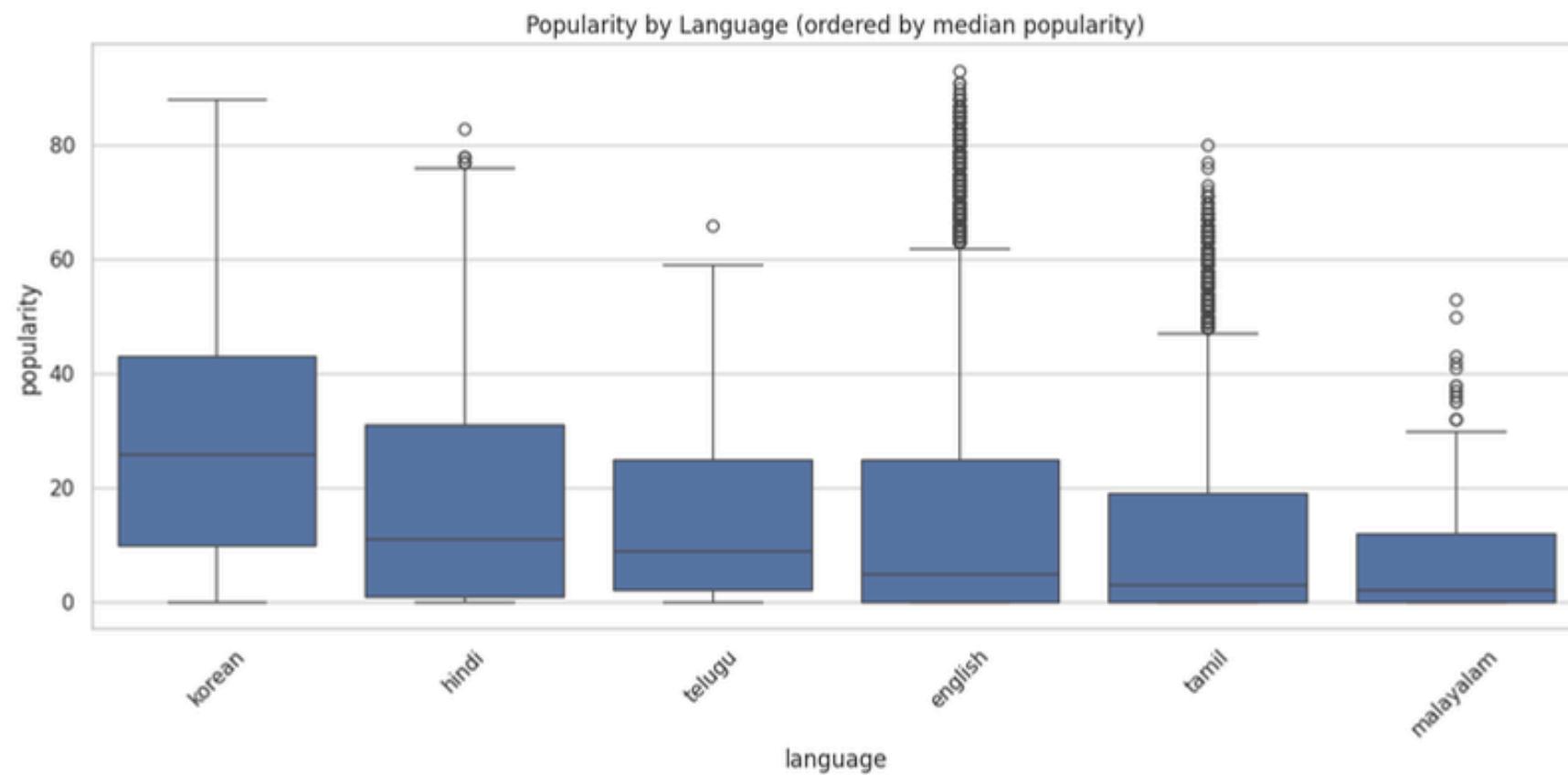


🎯 Insights — Popularity vs Audio Features

- Energy and danceability show mild positive links with popularity — energetic, rhythm-focused songs tend to attract more listeners.
- Acousticness is negatively related to popularity, showing that softer, unplugged tracks perform less commercially.
- Duration has little impact — shorter songs dominate, but extreme durations don't gain extra traction.
- Overall, catchy, energetic, and rhythm-rich tracks are the most popular across Spotify.



- 🎯 Insights — Popularity & Emotional-Tonal Features**
- Valence (musical positivity) shows a slight upward trend — moderately happy tracks perform best, while extremes drop off.
 - Instrumentalness negatively impacts popularity, confirming that vocals play a key role in mass appeal.
 - Tempo has minimal influence — both slow and fast tracks can succeed if they're catchy.
 - Liveness shows no clear pattern, suggesting live-recorded tracks are less preferred in mainstream playlists.
 - Overall, popular songs balance positive mood, vocal focus, and steady tempo for wide listener appeal.



🎯 Insights – Popularity by Musical Attributes

- Korean and Hindi tracks lead in popularity, reinforcing their strong global and regional fanbases.
- Time Signature 4/4 dominates — confirming that standard rhythmic structures appeal most to mainstream listeners.
- Key & Mode Analysis: Songs in Key 3 (D#) and major mode show the highest median popularity, suggesting brighter tonalities perform better.
- Overall, popular songs lean toward major keys, steady beats, and language-driven engagement.

INSIGHTS – BIVARIATE ANALYSIS (NUMERICAL VS CATEGORICAL)

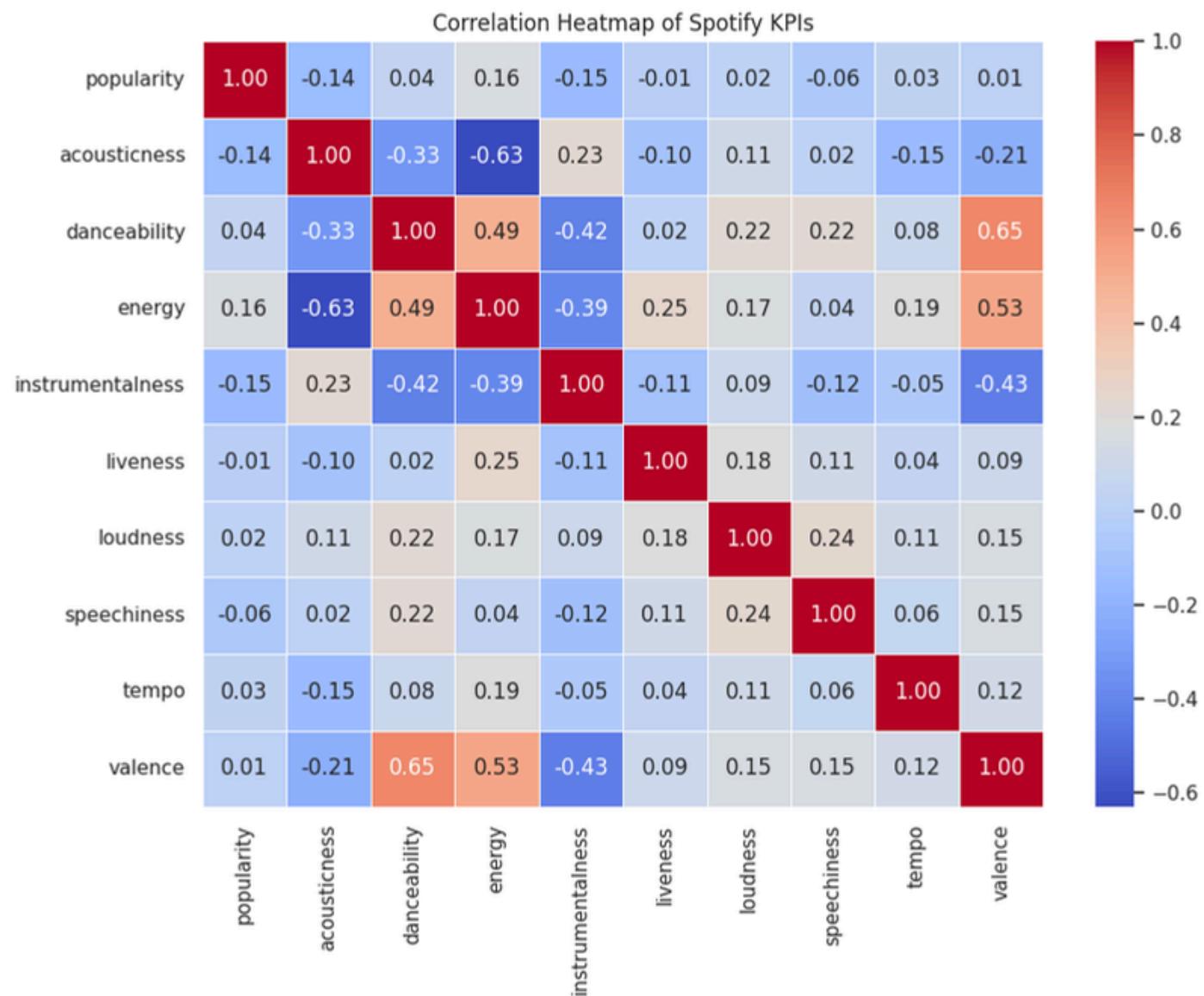
- Duration vs Popularity: Shorter songs tend to be slightly more popular – attention-grabbing, concise tracks perform better on streaming platforms.
- Danceability vs Popularity: Mild positive relationship – songs with higher danceability scores attract broader listener engagement, especially in pop and EDM genres.
- Energy vs Popularity: Higher energy correlates with greater popularity – upbeat, dynamic tracks dominate playlists.
- Loudness vs Popularity: Louder mixes are moderately more popular, reflecting modern production trends that favor intensity and clarity.
- Acousticness vs Popularity: Negatively correlated – electronic or studio-produced songs are generally more popular than purely acoustic ones.
- Valence vs Popularity: Songs with moderate positivity (not too happy or sad) are most popular – balanced emotional tone wins over extremes.
- Instrumentalness vs Popularity: Strong negative trend – tracks without vocals rarely achieve high popularity, emphasizing vocal appeal.
- Liveness vs Popularity: No clear pattern – live recordings perform less consistently compared to polished studio versions.
- Tempo vs Popularity: Minimal relationship – both fast and mid-tempo songs succeed if they're catchy and well-produced.
- Language vs Popularity: Korean and Hindi tracks dominate in average popularity, showing the strength of cultural fandoms and regional streaming bases.
- Key & Mode vs Popularity: Major keys (especially around D# and C) slightly outperform minor keys – brighter tonalities connect better with audiences.
- Time Signature vs Popularity: 4/4 remains the most popular rhythm pattern – familiar and dance-friendly beats dominate mainstream success.

Conclusion:

Popularity on Spotify is influenced most by energy, danceability, loudness, and language, while tonal and rhythmic features (key, tempo, time signature) play subtler roles. The ideal hit song blends catchiness, moderate positivity, and energetic production.



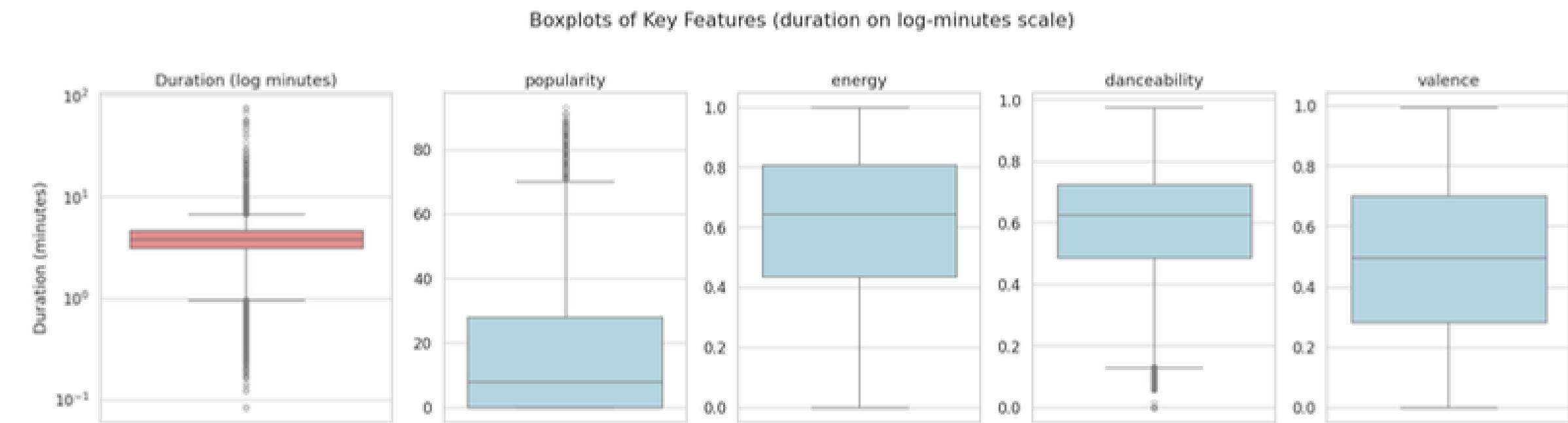
Multivariate Analysis



🎯 Insights

- Popularity shows positive correlations with energy, danceability, and loudness, suggesting that energetic, rhythm-driven, and high-volume songs are more likely to trend.
- Acousticness has a negative correlation with energy and loudness, showing that more acoustic tracks tend to be softer and calmer.
- Danceability and energy share a strong positive relationship, revealing that upbeat and energetic songs often make people want to move.
- Valence (musical positivity) connects moderately with danceability and energy, meaning happier songs often sound lively and high-energy.
- Instrumentalness shows weak or negative correlations with most KPIs — indicating that instrumental tracks generally score lower in popularity and energy.

OUTLIER ANALYSIS



🎯 Insights – Univariate Analysis (Boxplots)

- Duration (log minutes): Most tracks range between 2–6 minutes, with several outliers (very short or long songs).
- Popularity: Highly right-skewed – a few tracks achieve very high popularity, while the majority have lower scores.
- Energy: Median energy is around 0.65, indicating most songs lean toward an energetic sound.
- Danceability: Generally high across songs (median ≈ 0.6), suggesting rhythmic and listener-friendly compositions dominate Spotify.
- Valence: Fairly balanced distribution – songs span from melancholic to upbeat, showing diverse emotional tones in the dataset.



DATA-DRIVEN INSIGHTS & RECOMMENDATIONS

For Music Producers & Artists:

- Focus on high-energy, moderately danceable, and emotionally positive compositions — these traits correlate with higher popularity.
- Keep song duration around 3-4 minutes for better engagement and playlist fit.
- Consider leveraging English or bilingual (English + local) tracks to maximize reach.

For Spotify's Curation Team:

- Personalize playlists by energy and valence levels — e.g., "High Energy Workout" or "Low Valence Chill" mixes.
- Promote non-English tracks with strong engagement metrics to improve global diversity and audience discovery.

For Data Analysts & Business Teams:

- Investigate popularity prediction models using key features (energy, danceability, valence, and duration).
- Conduct deeper analysis on temporal shifts in music characteristics to understand evolving listener preferences.

For Future Research:

- Explore correlation between release year and feature trends (e.g., are modern songs shorter and more energetic?).
- Analyze genre-level insights once genre data is integrated — it can reveal cultural and regional music evolution.

💡 In essence:

Spotify's data paints a vibrant picture of modern music — energetic, emotional, and short-form — driven by both global and local influences.

These insights open doors for strategic playlisting, artist targeting, and market expansion across languages and moods.

FUTURE SCOPE / LIMITATIONS

Discuss the limitations of the dataset or analysis and suggest possible extensions for future exploration.

⚠ Limitations of the Dataset and Analysis

Incomplete Metadata:

- The dataset does not include key contextual variables like genre, artist popularity growth, or playlist inclusion — all of which could have a major influence on track success.

Popularity Bias:

- The popularity score on Spotify is dynamic and region-dependent. It changes with user activity, meaning the current values may not reflect long-term trends.

Missing Listener Demographics:

- Without information on age, region, or listener behavior, we can't link musical attributes to audience preferences directly.

Static Snapshot of Data:

- The dataset likely represents a single time frame, not continuous updates. Therefore, temporal evolution of music trends (e.g., how energy levels changed across years) remains partially unexplored.

Data Imbalance:

- Some languages or styles have very few entries (like Malayalam, Telugu), which can distort statistical conclusions and visualizations.

Correlation, Not Causation:

- Observing that high energy tracks tend to be more popular does not imply causation — external factors like promotion, social media trends, or artist fame also play key roles.





CONCLUSION

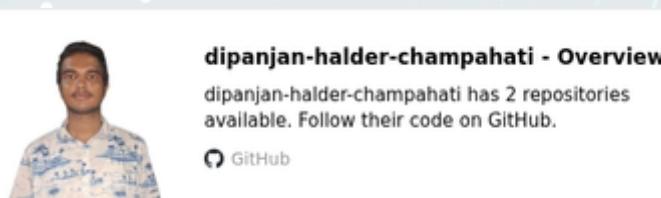
The Spotify dataset reveals how data mirrors modern music culture – upbeat, emotionally diverse, and digitally driven. By understanding these patterns, artists and data professionals can both shape the future of sound and optimize listener engagement.



 "DATA DOESN'T JUST DESCRIBE MUSIC – IT HELPS US HEAR IT DIFFERENTLY."

THANK YOU

GITHUB



EMAIL

mr.dipanjanhalder@gmail.com

PRESENTED BY:
DIPANJAN HALDER