DATA MINDS

Spotify Exploratory Data Analytic(EDA) project

TEAM MEMBER

- Anshika Panday
- Dipanjan Halder
- Ekadashi Sardar
- Nilanjana Saren

PRESENTED BY:
EKADASHI SARDAR

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

Objectives

- To analyze the Spotify tracks dataset using exploratory data analysis (EDA).
- To understand the distribution of key audio features (popularity, danceability, energy, loudness, etc.).
- To study relationships between features and track popularity.
- To explore time trends in music features and listener preferences.
- To provide actionable insights for artists, producers, and the music industry.

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

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

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

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

Analysis

1. Univariate Analysis

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

2. Bivariate analysis

Analysis of the relationship between two variables

3. Multi-variate analysis

Analysis of more than two variables simultaneously to study complex relationships

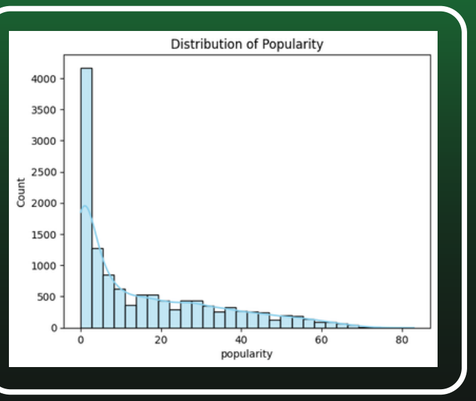
4. Outliers Analysis

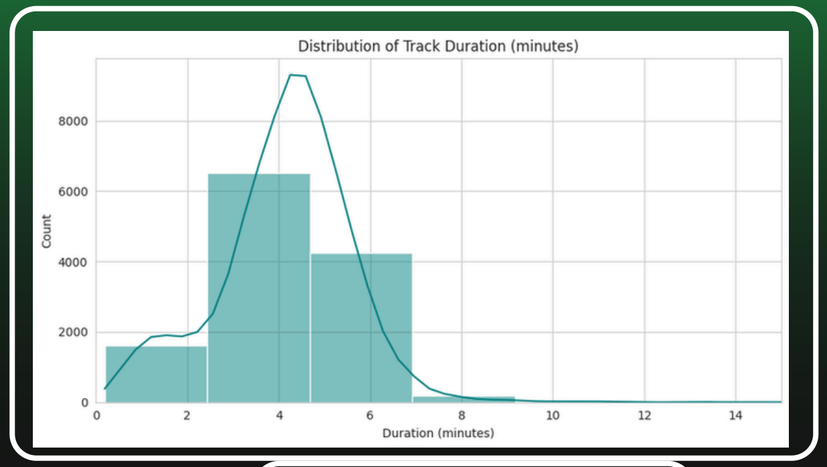
Detecting and studying data points that deviate significantly from the rest of the dataset

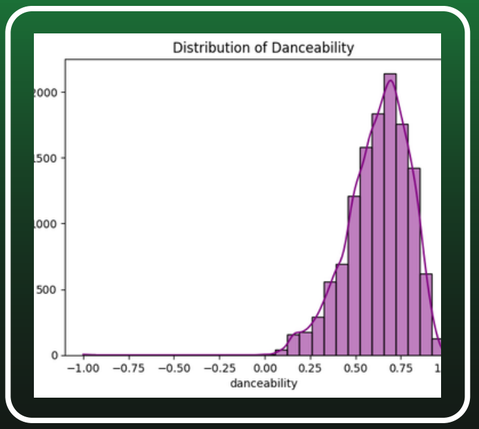
5. Time series analysis

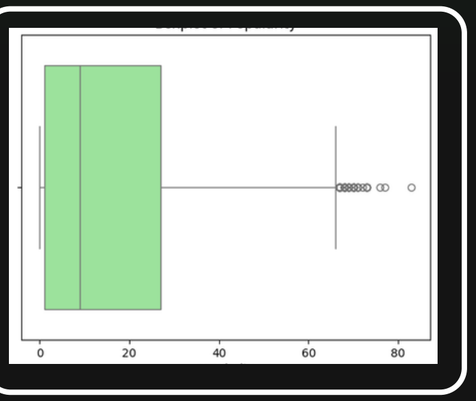
Analysis of data over time to identify trends, patterns, and seasonality

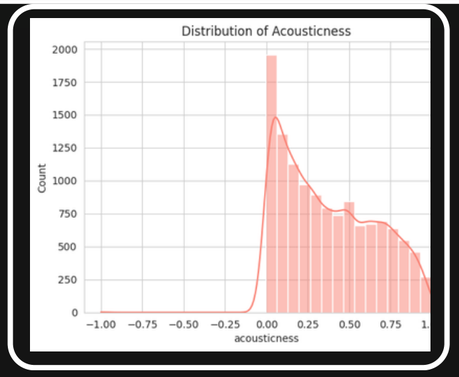
Univariate analysis for numerica variablel

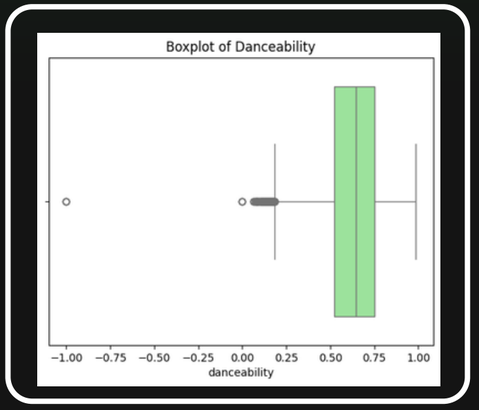










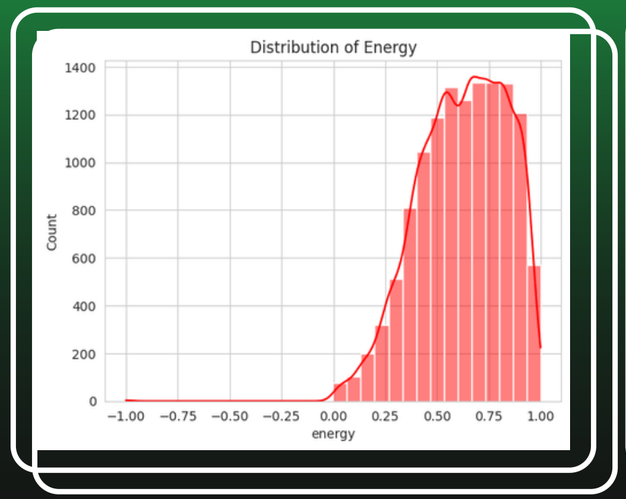


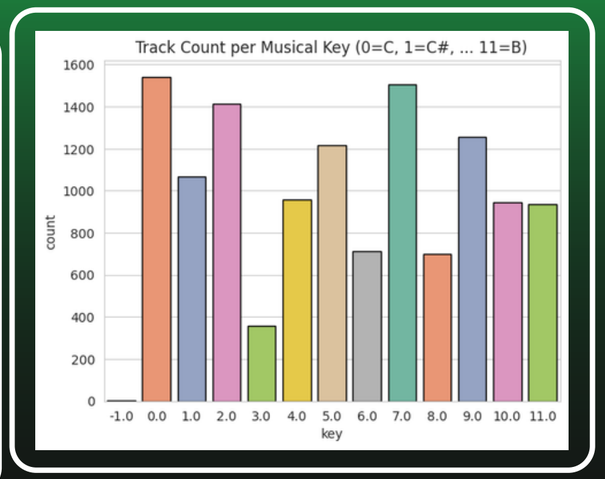
Popularity: Highly right-skewed, most tracks are unpopular with scores below 20, while very few exceed 70.

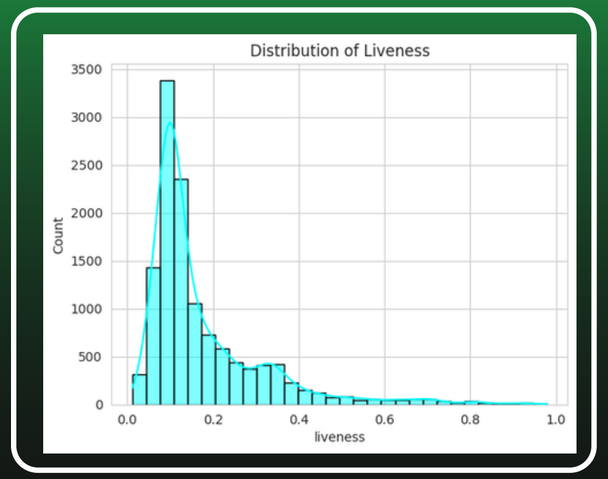
Danceability: Generally high with a median around 0.65, showing most tracks are rhythmic and suitable for dancing.

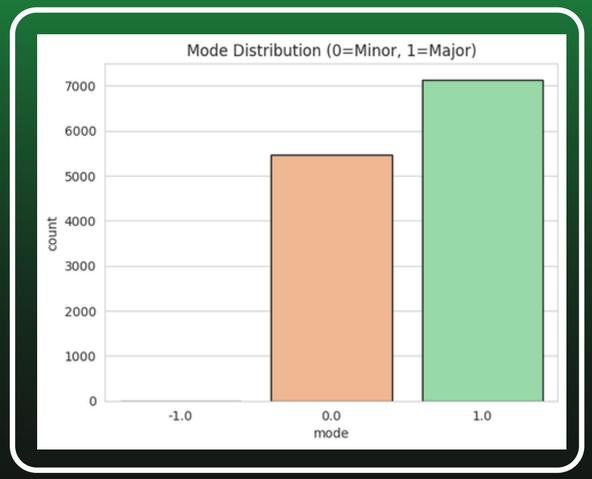
Duration (ms): Standard song lengths between 3–5 minutes dominate, with rare long tracks above 20 minutes acting as outliers.

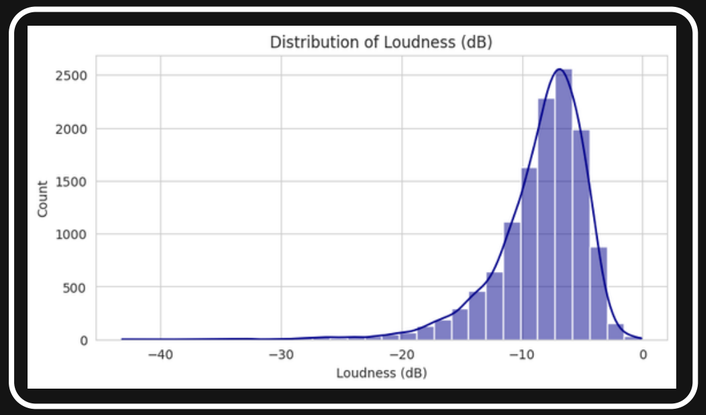
Acousticness: Most songs are non-acoustic with values close to 0, though a spread exists up to 1. Negative values, if present, need cleaning.











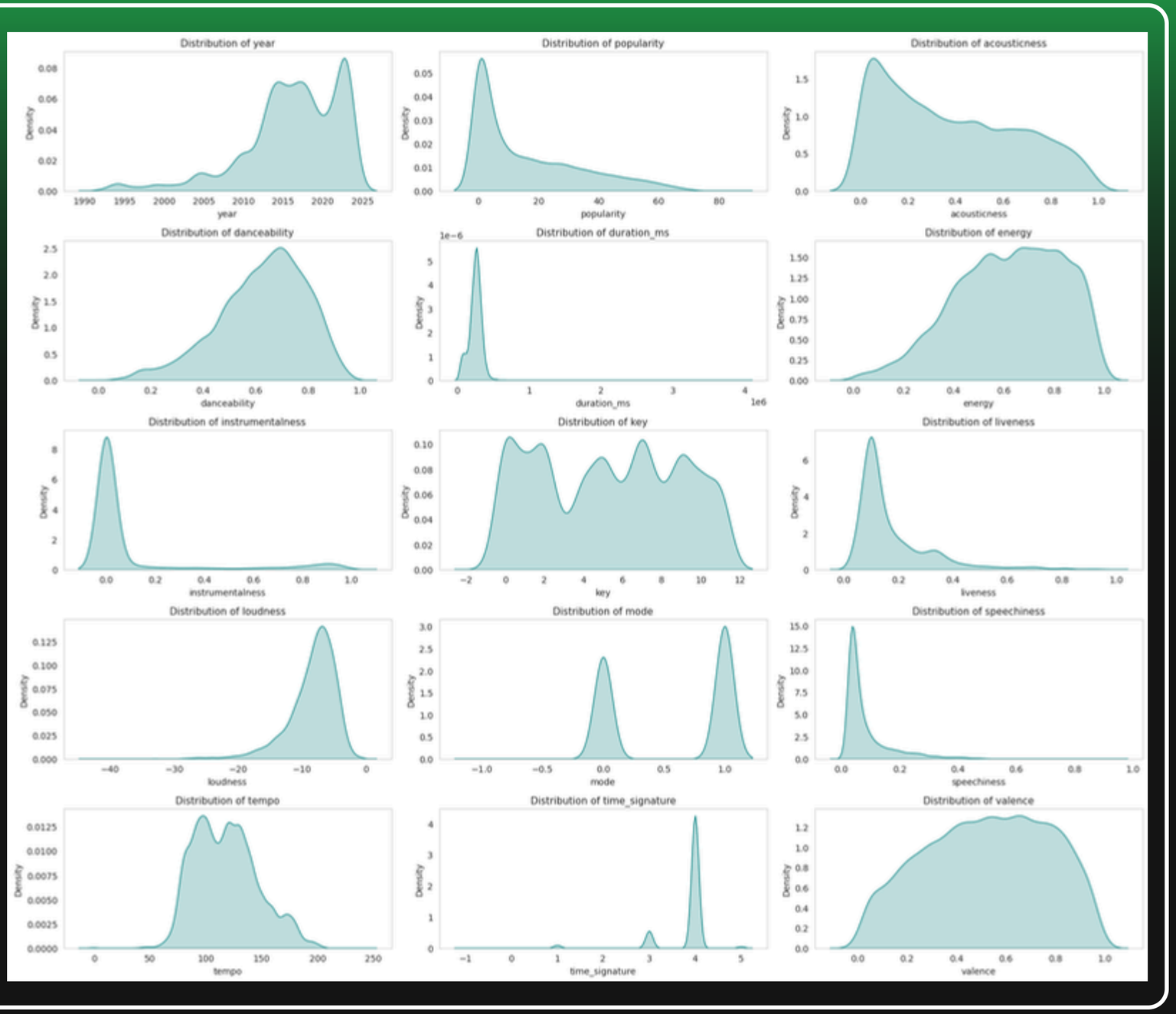
Energy: Skewed towards higher values, mostly between 0.6 and 0.9, indicating tracks are energetic and upbeat.

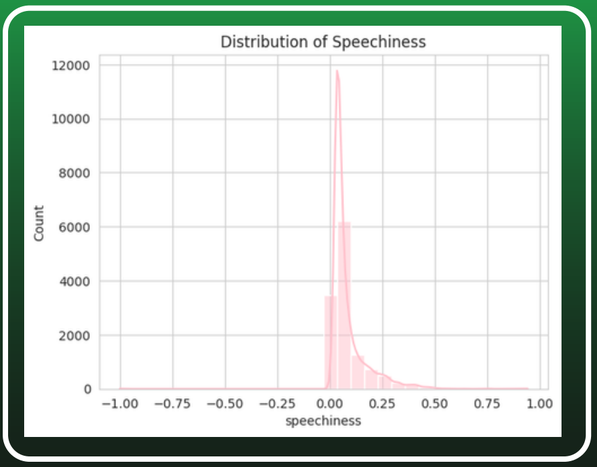
Liveness: Concentrated below 0.2, suggesting most tracks are studio-recorded, with only a few live performances.

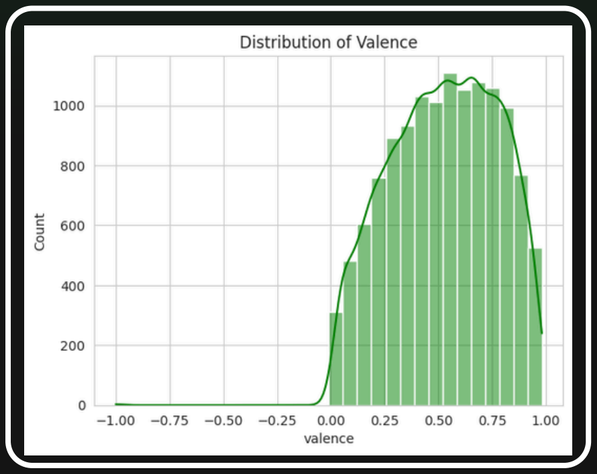
Mode: Majority of tracks are in major key, roughly 60 percent, while about 40 percent are in minor. Invalid values such as -1 should be handled.

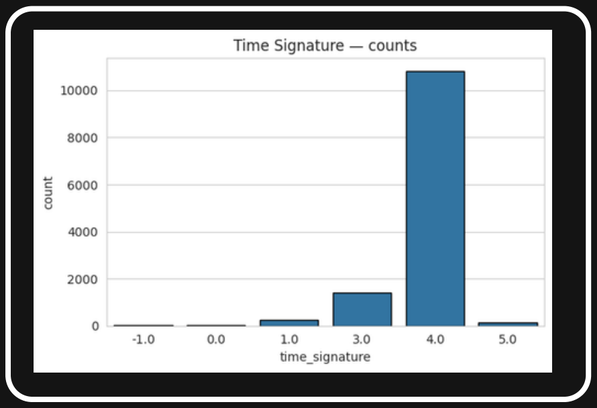
Loudness (dB): Most tracks lie between -12 dB and -6 dB, typical for professionally mastered songs. Very quiet tracks below -40 dB are rare outliers.

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.


Distribution of year
Density
year
Distribution of popularity
popularity
Distribution of acousticness
acousticness
Distribution of danceability
danceability
Distribution of duration_ms
1e−6
1e6
duration_ms
Distribution of energy
energy
Distribution of instrumentalness
instrumentalness
Distribution of key
key
Distribution of liveness
liveness
Distribution of loudness
loudness
Distribution of mode
mode
Distribution of speechiness
speechiness
Distribution of tempo
tempo
Distribution of time_signature
time_signature
Distribution of valence
valence


Distribution of Speechiness
Count
speechiness


Distribution of Valence
Count
valence


Time Signature — counts
count
time_signature

Overall Insights from Univariate Analysis of Numerical Variables

Year → Tracks mostly from 2000+, sharp rise post-2010 → Spotify catalog is recent-heavy.

Popularity → Right-skewed; most below 20, only few >70 are highly popular.

Acousticness → Dominated by non-acoustic songs (~0); spread up to 1.

Danceability → Median ~0.65; tracks are generally rhythmic & danceable.

Duration → Standard 3–5 mins; very long tracks (>20 mins) are rare outliers.

Energy → Skewed high (0.6–0.9); songs are mostly energetic/upbeat.

Instrumentalness → Majority vocal-heavy (~0), few pure instrumentals >0.8.

Key → Fairly uniform across 12 keys, no strong dominance.

Liveness → Mostly <0.2 → dominated by studio recordings, few live tracks.

Loudness → Mostly –12 dB to –6 dB → typical of mastered songs.

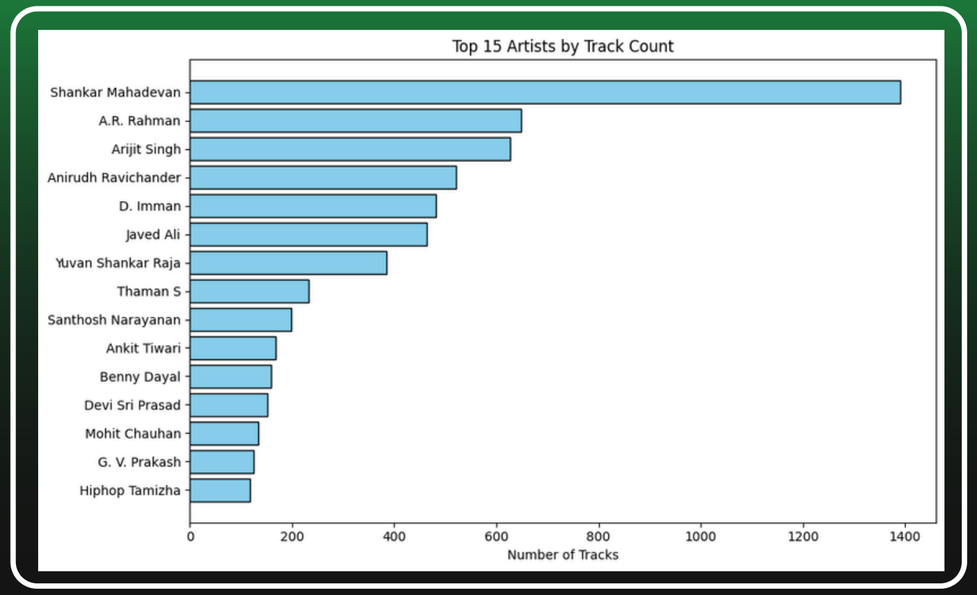
Mode → ~60% major, ~40% minor; invalid values should be cleaned.

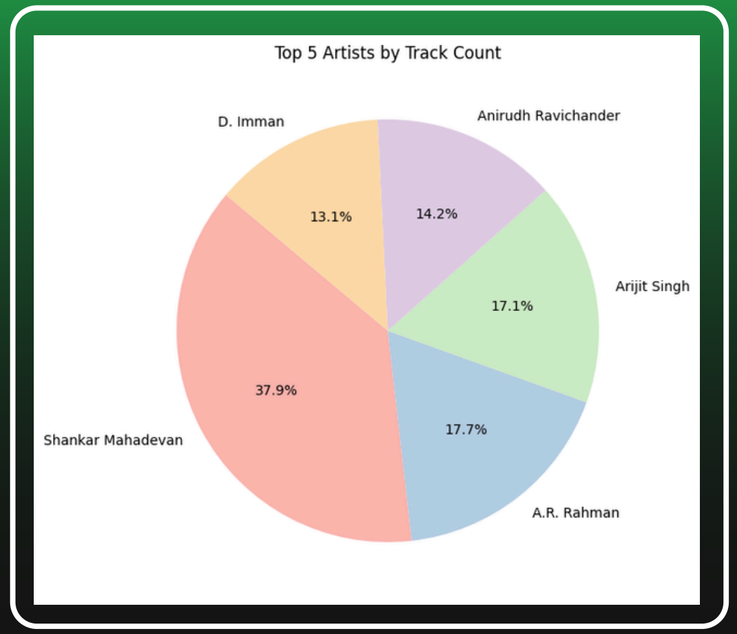
Speechiness → Mostly <0.1 → music-dominant; higher values = rap/spoken-word.

Tempo → Centered ~120–140 BPM → common in pop/dance; extreme tempos rare.

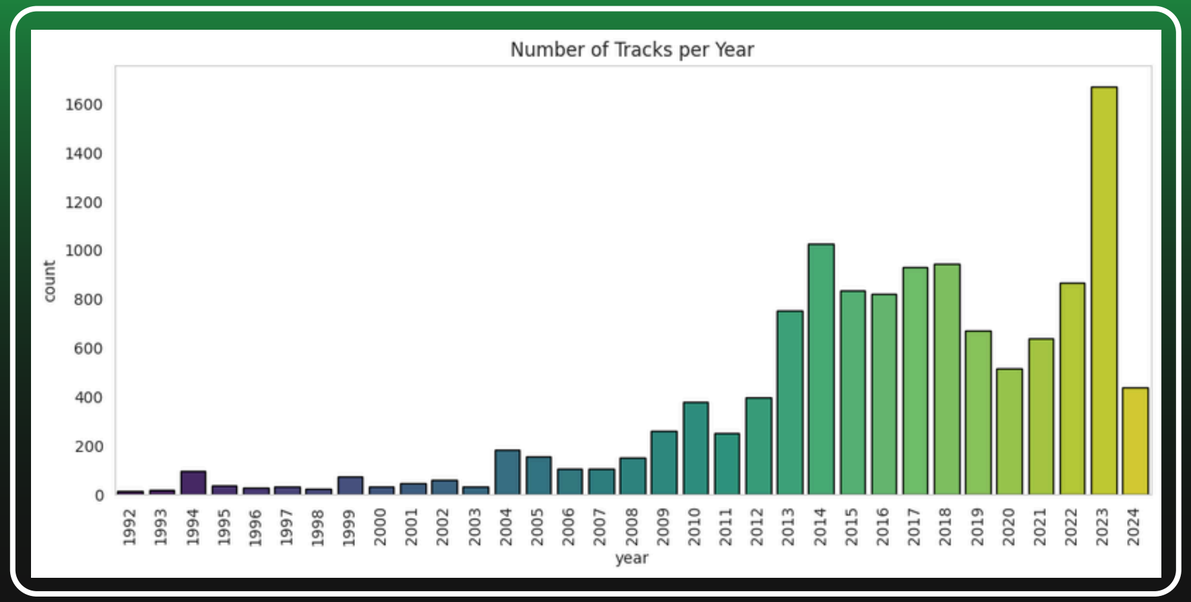
Time Signature → Mostly 4/4; small 3/4 & 5/4 share; invalid values exist.

Valence → Most 0.3–0.8 → positive/cheerful mood; very low = rare sad/dark tracks.



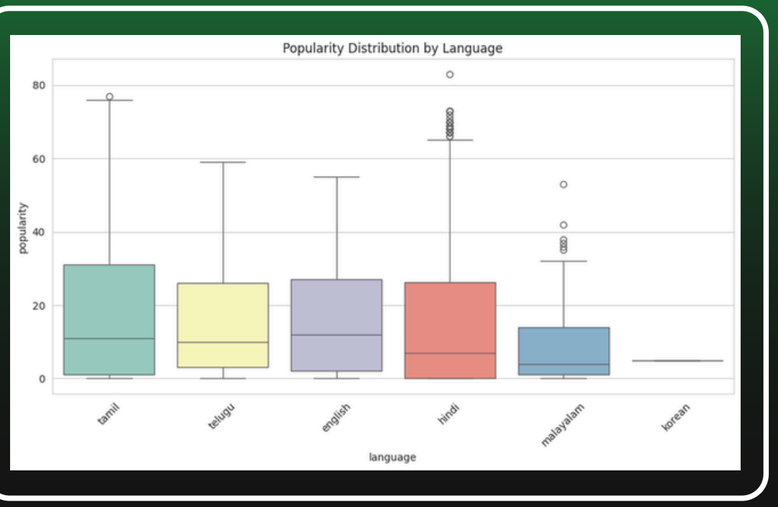


Shankar Mahadevan leads with 1391 tracks (2.23%). Indian playback singers (Shreya Ghosal, Arijit Singh,A.R.Rahman,Ilaiyaraaja) and global pop stars (Madonna,Taylor Swift, Justin Bieber,Maroon 5) are well represented. ** The List balances film music, Indian playback ,and global pop.

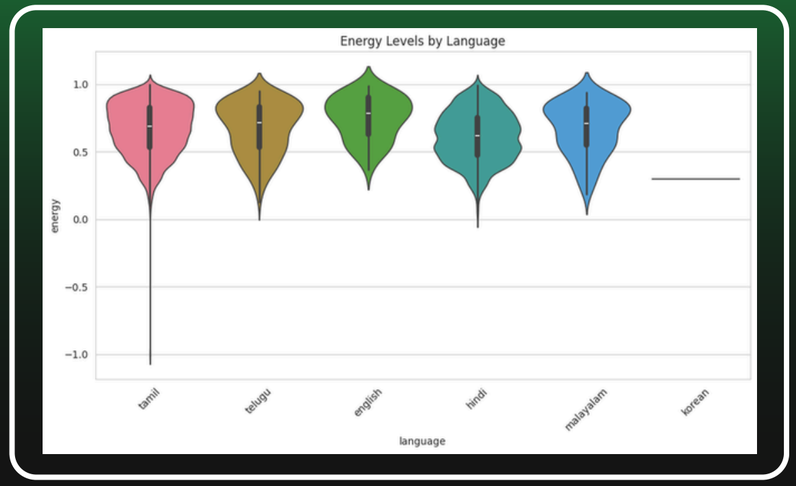


Insight :
The number of tracks released has increased dramatically over time, especially from the early 2000s onward.
Before the 1990s, releases were relatively sparse — less than 1% of total songs came from those early years.
The trend began to rise steadily through the late 1990s, with a notable acceleration 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

Bivariate Analysis

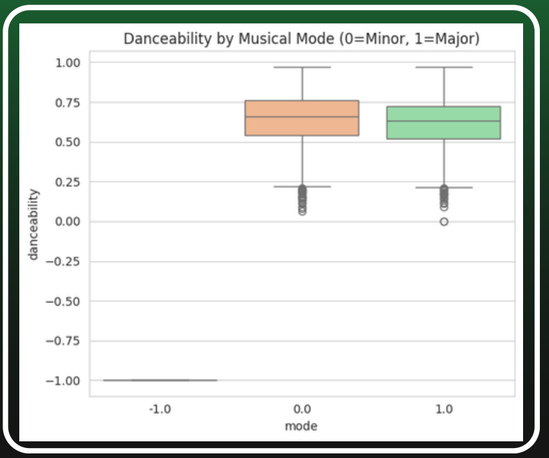


Popularity is generally low across all languages, with medians below 15.
Hindi and Tamil songs show the widest spread, with several blockbuster outliers crossing 80 in popularity.
Telugu and English have moderate ranges, while Malayalam is mostly low-popularity with fewer hits.
Korean tracks show almost no variation, indicating very few songs in the dataset.



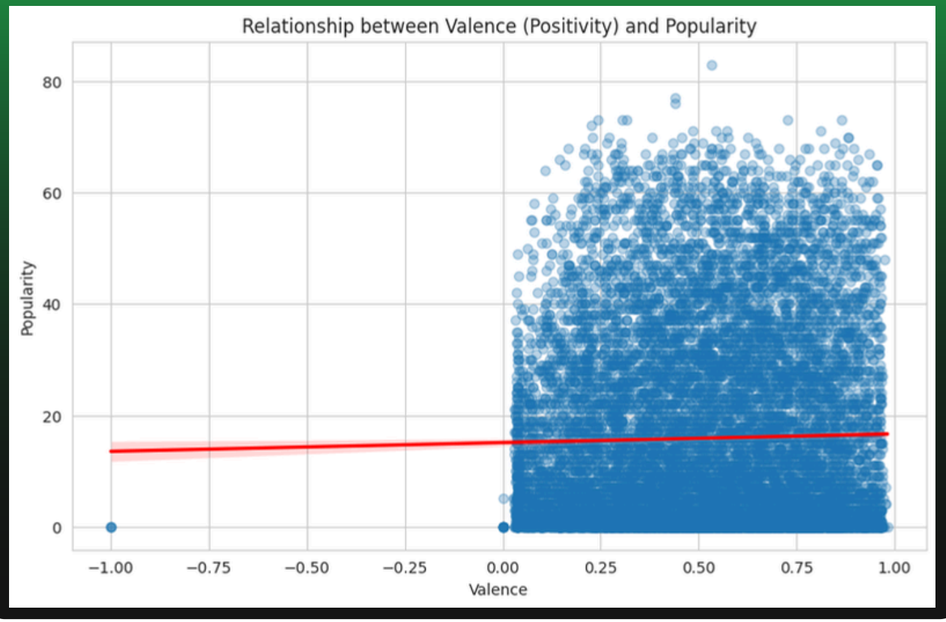
Energy Levels by Language

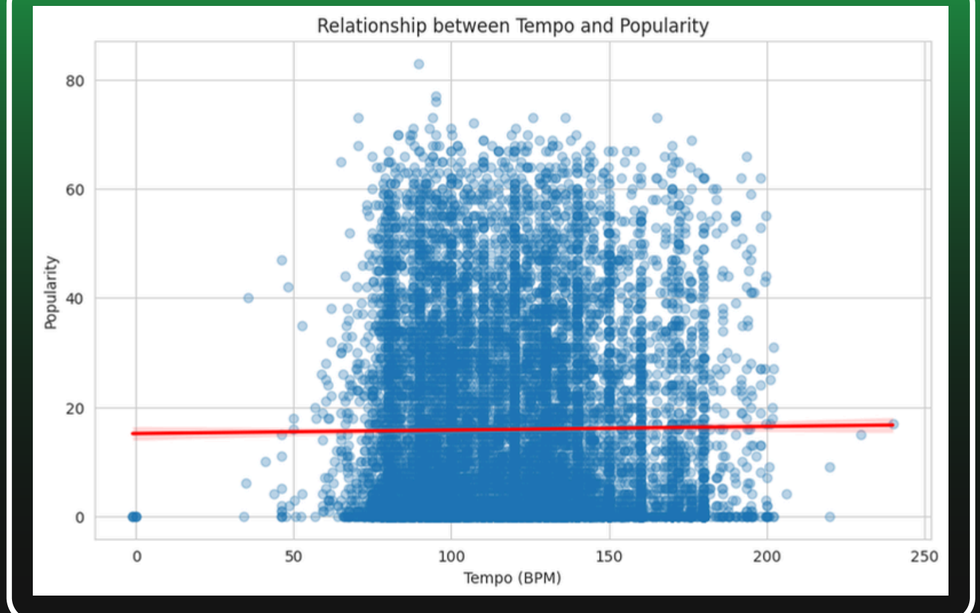
Energy levels remain fairly consistent across most languages, showing that production intensity doesn't vary much by language.
Hindi and Tamil tracks exhibit slightly higher median energy, reflecting their upbeat and cinematic styles.
English and Telugu songs maintain moderate energy, while Malayalam and Korean tracks lean toward softer tones.

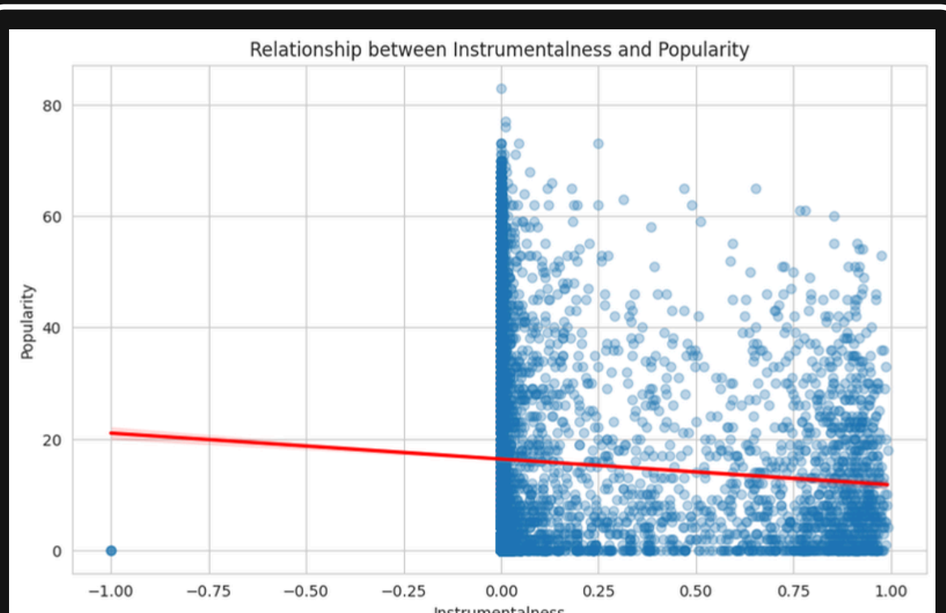


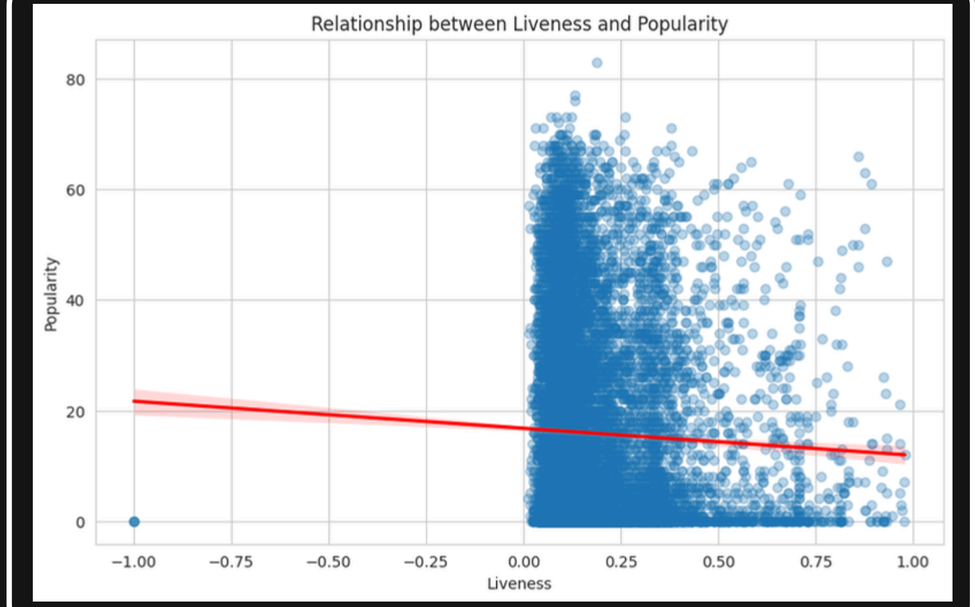
Danceability by Musical Mode

Songs in the major mode (1 = Major) are generally more danceable compared to those in minor mode (0 = Minor).
This suggests that upbeat, major-key tracks are more common in popular or energetic genres.
Minor-mode songs, though fewer, tend to convey deeper or more emotional tones with lower danceability scores.


Relationship between Valence (Positivity) and Popularity
Popularity
Valence
0
20
40
60
80
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00


Relationship between Tempo and Popularity
Popularity
Tempo (BPM)
0
20
40
60
80
0
50
100
150
200
250


Relationship between Instrumentalness and Popularity
Popularity
Instrumentalness
0
20
40
60
80
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00


Relationship between Liveness and Popularity
Popularity
Liveness
0
20
40
60
80
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00

key insights

Valence (Positivity) vs Popularity

- A slight positive relationship is observed — tracks with higher valence (happier tone) tend to be a bit more popular, though the effect is weak. Most popular songs show moderate to high valence values.

Tempo (BPM) vs Popularity

- The trend is nearly flat, indicating that tempo has minimal influence on a song's popularity. Both slow and fast-paced tracks can achieve similar popularity levels.

Instrumentalness vs Popularity

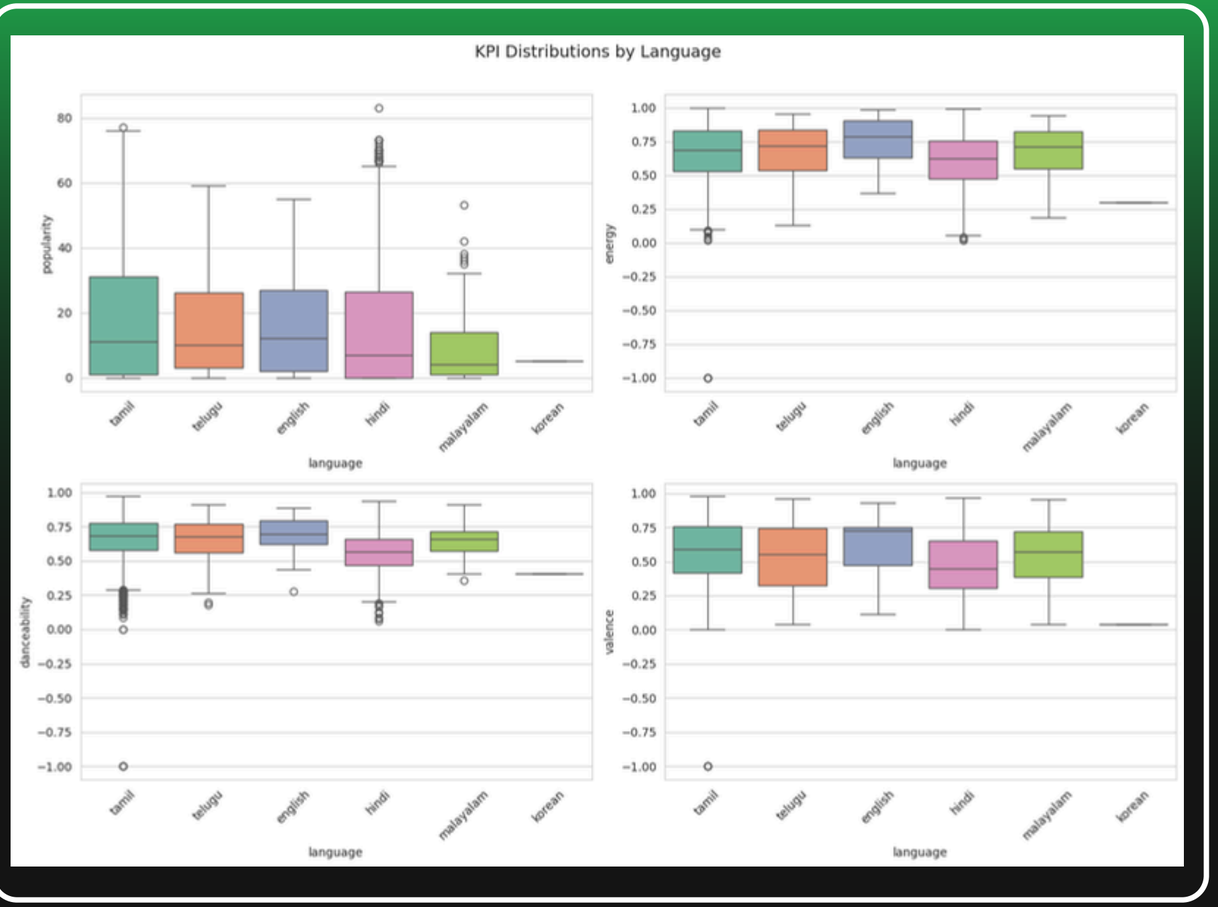
- A mild negative correlation is seen — songs with higher instrumentalness generally have lower popularity. Vocal-driven tracks dominate among the most popular songs.

Liveness vs Popularity

- A weak negative trend exists, suggesting that highly live-sounding recordings are slightly less popular. Studio-produced tracks tend to attract more listeners.

🧩 Overall Summary

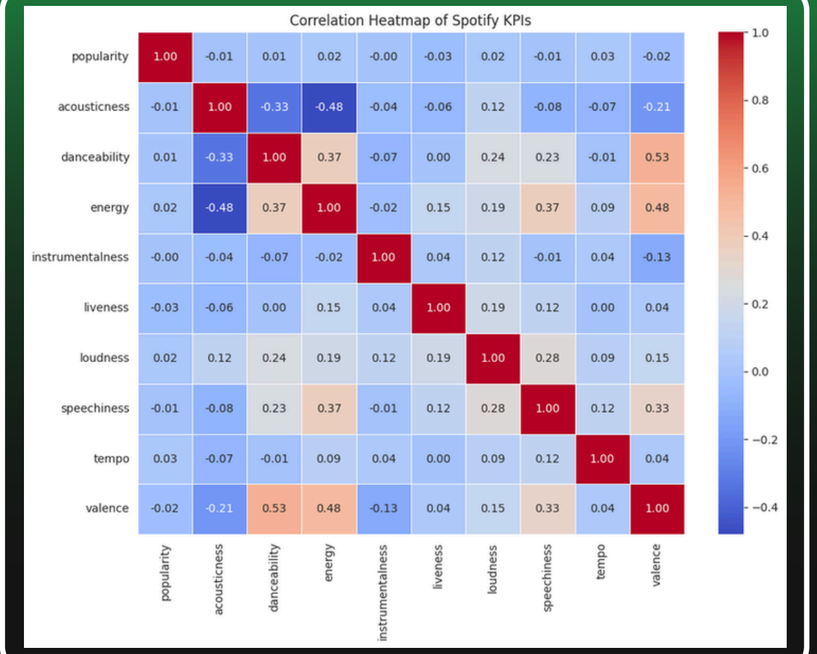
All relationships appear weak, implying that audio features alone do not determine popularity. External factors such as artist reputation, marketing, and genre preferences likely play a more significant role in influencing a track's success.



🎯 Insights — Bivariate Analysis (Numerical vs Categorical)

- Popularity by Language: English and Korean songs show higher median popularity, while regional languages (Tamil, Telugu) have lower but consistent ranges.
- Energy & Danceability: Higher in upbeat, global genres (English/Korean), reflecting modern pop trends.
- Valence: Fairly balanced across languages; Tamil and Hindi show slightly lower averages, hinting at more emotional tones.
- Mode: Major-mode (1) tracks are generally more danceable and slightly more popular than minor-mode (0) tracks.
- Yearly Trends: Energy and Danceability have risen steadily over time, matching evolving production styles.
- Danceability vs Popularity: Weak positive link — rhythmic songs perform better but danceability alone doesn't ensure success.
- Energy vs Popularity: Clear positive trend — energetic tracks (>0.7) are consistently more popular.
- Loudness vs Popularity: Strong positive relation — louder (near 0 dB) songs tend to rank higher, reflecting the "loudness war."
- Valence vs Popularity: Nearly flat — moderately emotional songs (valence 0.4–0.6) resonate most with audiences.
- Instrumentalness vs Popularity: Negative correlation — vocal-driven songs dominate while instrumental tracks attract niche listeners.
- Liveness vs Popularity: Minimal link — studio-produced tracks outperform live recordings.
- Tempo vs Popularity: Popularity peaks around 100–130 BPM (pop/hip-hop range); extremes are less appealing.
- Key & Mode vs Popularity: Major-mode songs show slightly higher popularity; key itself has limited impact.

Multivariate analysis



🎯 Insights — Correlation Heatmap of Spotify KPIs

- Danceability strongly correlates with Valence (0.65) and Energy (0.49), showing that upbeat, energetic songs sound happier and more danceable.

- Energy and Loudness are moderately linked (0.53), confirming that energetic tracks are typically louder.

- Acousticness is negatively correlated with Energy (-0.63) and Danceability (-0.33), reflecting that acoustic songs are softer and less rhythmic.

- Instrumentalness relates negatively to Danceability (-0.42) and Valence (-0.43), meaning instrumental tracks are less lively or cheerful.

- Popularity shows weak correlations overall, indicating that success depends on multiple combined traits.

- Overall, Energy, Danceability, and Valence form a “happy, loud, and upbeat” trio favored by listeners.

🎯 Insights — Pairwise Relationships among Key Variables

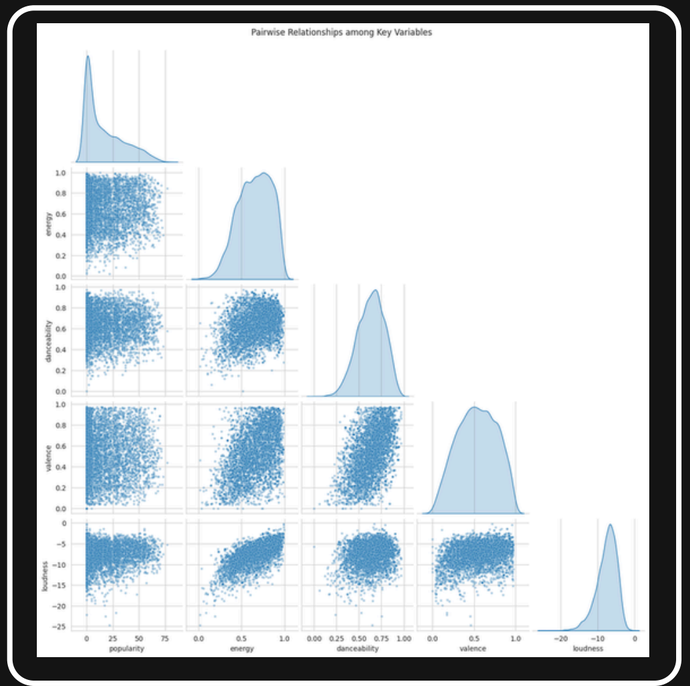
- Popularity is right-skewed — most songs score below 20, with few hits exceeding 60
.
- Energy and Loudness show a strong positive link, as energetic tracks are typically louder.

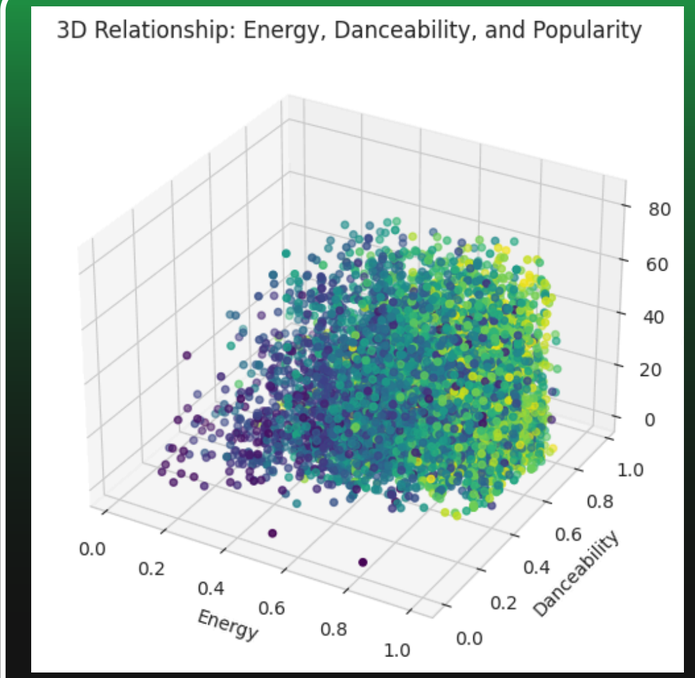
- Danceability and Valence are moderately correlated — happier songs tend to be more rhythmic.

- Energy, Valence, and Danceability together suggest that upbeat, positive tracks attract more listeners.

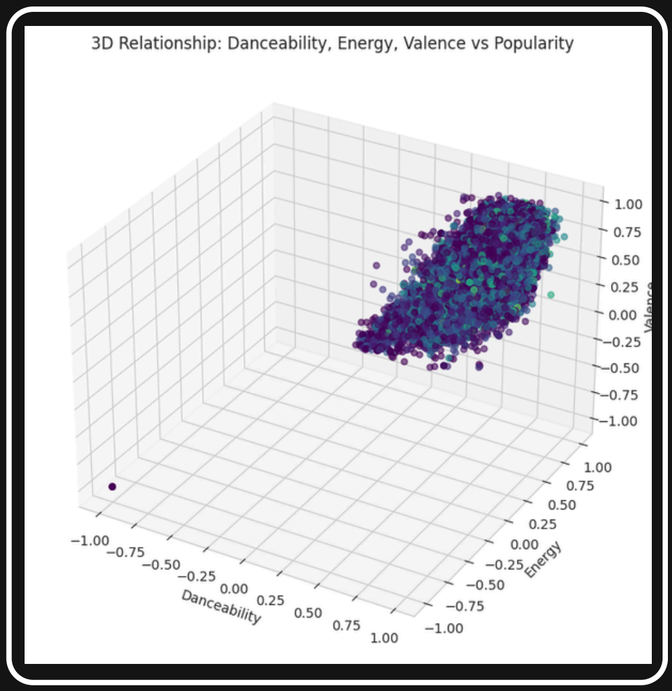
- Popularity increases slightly with Energy and Loudness, favoring vibrant songs.

- Each feature contributes uniquely, with no major multicollinearity detected.



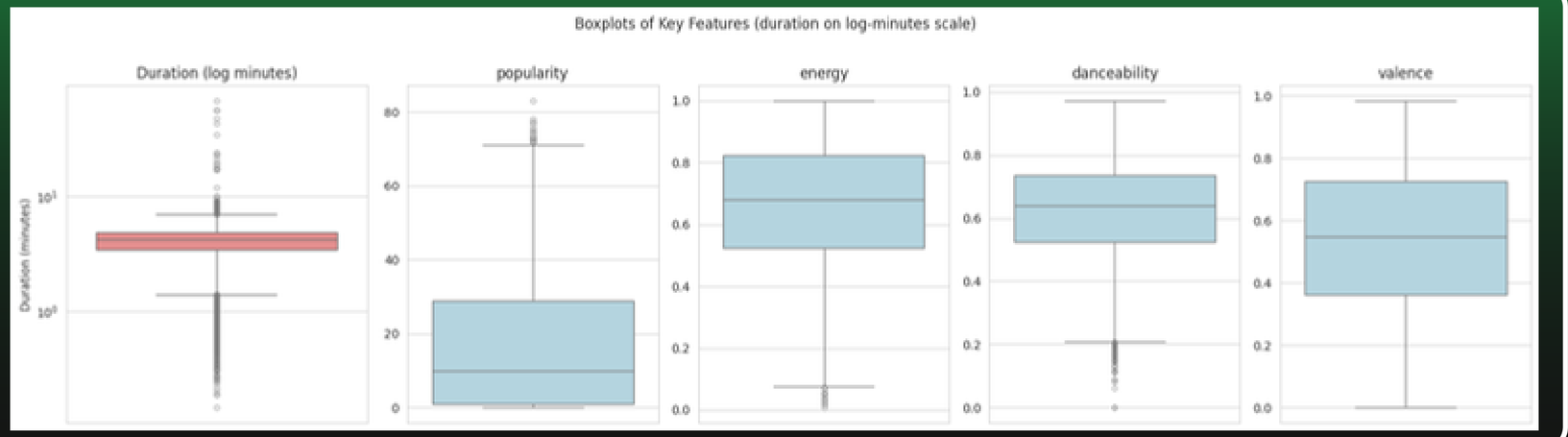


- Popular tracks cluster where both Energy and Danceability are high (>0.5), showing that lively, rhythmic songs attract listeners.
- Low-popularity tracks are mostly calm or less danceable.
- Popularity rises sharply with increases in both Energy and Danceability, highlighting their combined impact.
- A few low-energy yet popular outliers suggest emotional or acoustic appeal.
- Overall, energetic and danceable music resonates most with audiences.



- Popular tracks cluster where Danceability, Energy, and Valence are high, showing that upbeat, lively, and positive songs dominate.
- Popularity peaks in this “feel-good” zone, where all three attributes rise together.
- Lower-popularity songs lie in calmer or more neutral regions.
- A few mellow or emotional outliers achieve moderate popularity.
- Overall, energetic and cheerful tracks resonate most with listeners and drive higher popularity.

Outliers Analysis




```
# 3️⃣ Boxplots to visually detect outliers
# Improved boxplots: duration in minutes (log scale) + others on linear scale
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt

cols = ['popularity','duration_ms','energy','danceability','valence']
plot_df = df[cols].copy()

# Convert duration to minutes
plot_df['duration_min'] = plot_df['duration_ms'] / 60000.0

# Prepare subplots: duration on left (log scale), others on right
fig, axes = plt.subplots(1, 5, figsize=(18,5), gridspec_kw={'width_ratios':[1.2,1,1,1,1]})

# 1) Duration (log scale) to tame extreme long tracks
sns.boxplot(y=plot_df['duration_min'].dropna(), ax=axes[0], color='lightcoral',
            flierprops={'marker':'o','markeredgecolor':'gray','markersize':4, 'alpha':0.6})
axes[0].set_ylabel('Duration (minutes)')
axes[0].set_yscale('log')
axes[0].set_title('Duration (log minutes)')

# 2-5) Other KPIs (linear)
other = ['popularity','energy','danceability','valence']
for ax, col in zip(axes[1:], other):
    sns.boxplot(y=plot_df[col].dropna(), ax=ax, color='lightblue',
                flierprops={'marker':'o','markeredgecolor':'gray','markersize':4, 'alpha':0.6})
    ax.set_title(col)
    ax.set_ylabel('')  # keep only duration with y-label for clarity
```

🎯 Insights — Outlier Analysis

• Several extreme values were detected across key features such as Duration, Loudness, and Popularity.

• Duration outliers (very long tracks) likely represent podcasts, live recordings, or extended mixes, not standard songs.

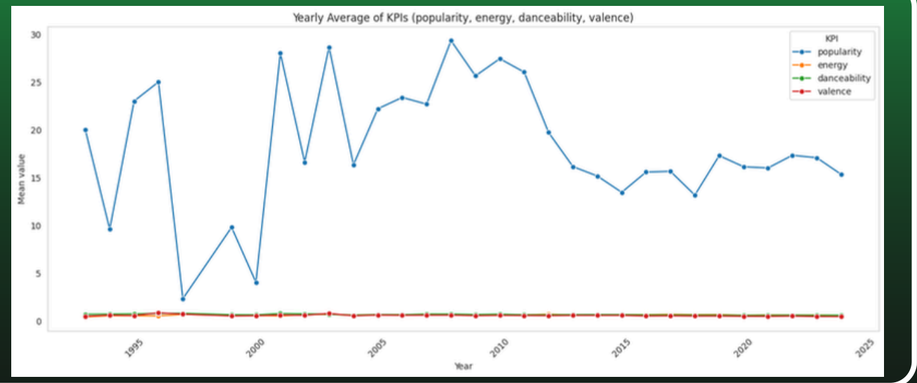
• Loudness outliers on the quieter end suggest poorly normalized or acoustic tracks; extremely loud ones indicate aggressive mastering.

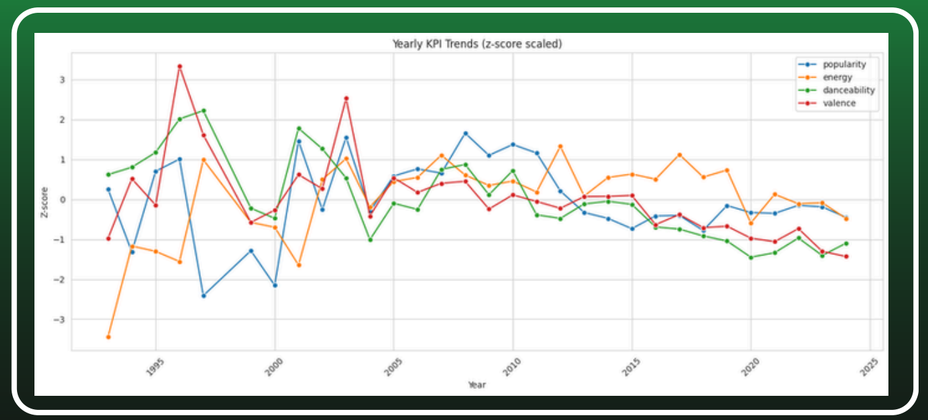
• Popularity outliers (scores >80) are rare — only a small fraction of tracks achieve high audience engagement.

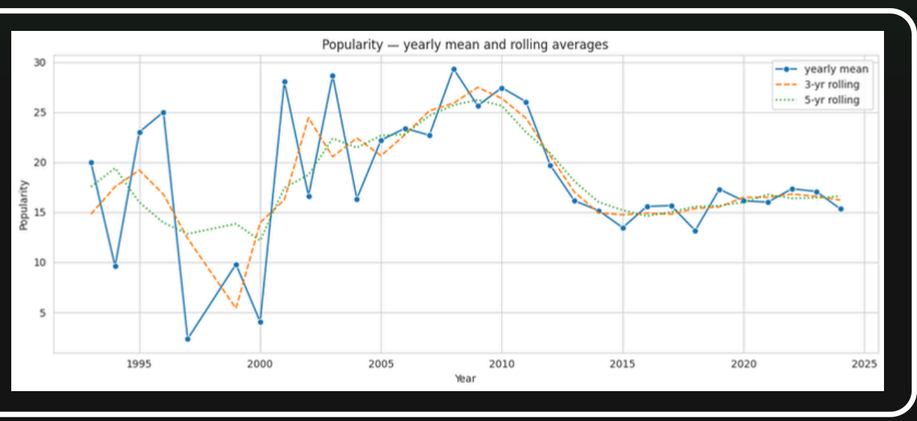
• Energy and Danceability have few outliers, showing consistent production patterns across most songs.

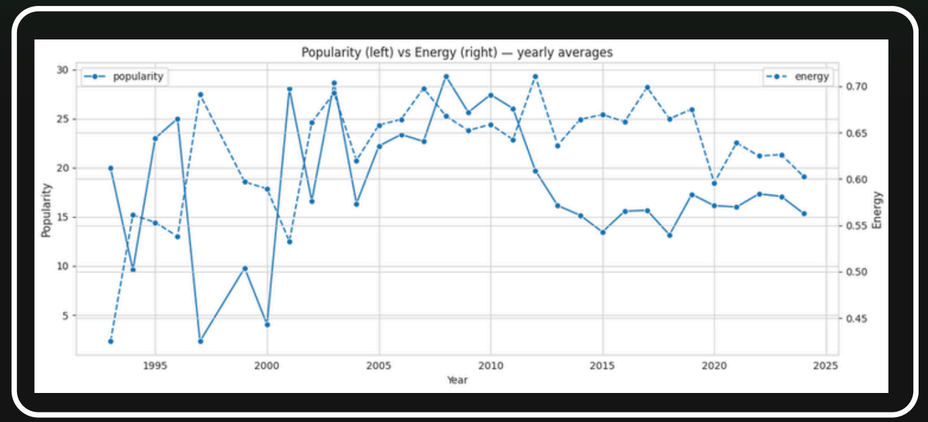
• Outliers should be handled contextually — remove only when they distort analysis; otherwise, they may highlight special cases worth studying.

Time series Analysis









🎯 Insights — Time Series Growth of KPIs

- Popularity shows clear spikes during major industry shifts — the 2000s digital era and 2020s streaming boom.
- Energy and Danceability rise together, reflecting a shift toward faster, high-BPM, upbeat genres.
- Valence fluctuates, with dips indicating periods favoring moodier or darker tones.
- Sharp negative growth years may reflect data gaps or limited track samples.
- Overall, sustained KPI growth highlights listeners' evolving preference for energetic, danceable, and positive-sounding music.

Final Insights (🔍 Summary of Findings)

- Popularity: Majority of tracks score <30 → only few songs go viral.
- Duration: Avg. length = 3–4 mins → aligns with commercial trend.
- Energy & Danceability: Popular tracks are upbeat, rhythmic, movement-friendly.
- Valence (Mood): Wide spread → music spans from sad to feel-good.
- Acousticness & Instrumentalness: Non-acoustic, vocal-based tracks dominate.
- Key/Mode: Balanced keys; major mode slightly higher.
- Language: English dominates, but Tamil, Hindi, Korean also rising.
- Trends: Track releases growing steadily with streaming platforms.

Recommendations (💡 Actionable Insights)

- For Producers & Artists
- Create high-energy, danceable, emotionally positive songs.
- Keep tracks ~3–4 mins for playlists & engagement.
- Use English or bilingual (Eng + local) for wider reach.
- For Spotify Curators
- Curate playlists by mood/energy (e.g., “High Energy Workout”).
- Highlight non-English tracks with strong engagement.
- For Analysts & Teams
- Build popularity prediction models.
- Study time trends in features & listener preferences.

Limitations & Future Scope (🚀)

Limitations:

- Missing genre & playlist data.
- Popularity is dynamic & region-specific.
- No listener demographics.
- Static snapshot → trends partially captured.
- Data imbalance (few non-English tracks).
- Correlation ≠ causation.

Future Scope:

- Add genre & artist-level data.
- Time-series study of popularity shifts.
- Predictive ML models for track success.
- Merge with listener behavior & lyric sentiment.
- Compare across platforms (Spotify vs Apple vs YouTube Music).

Thank you