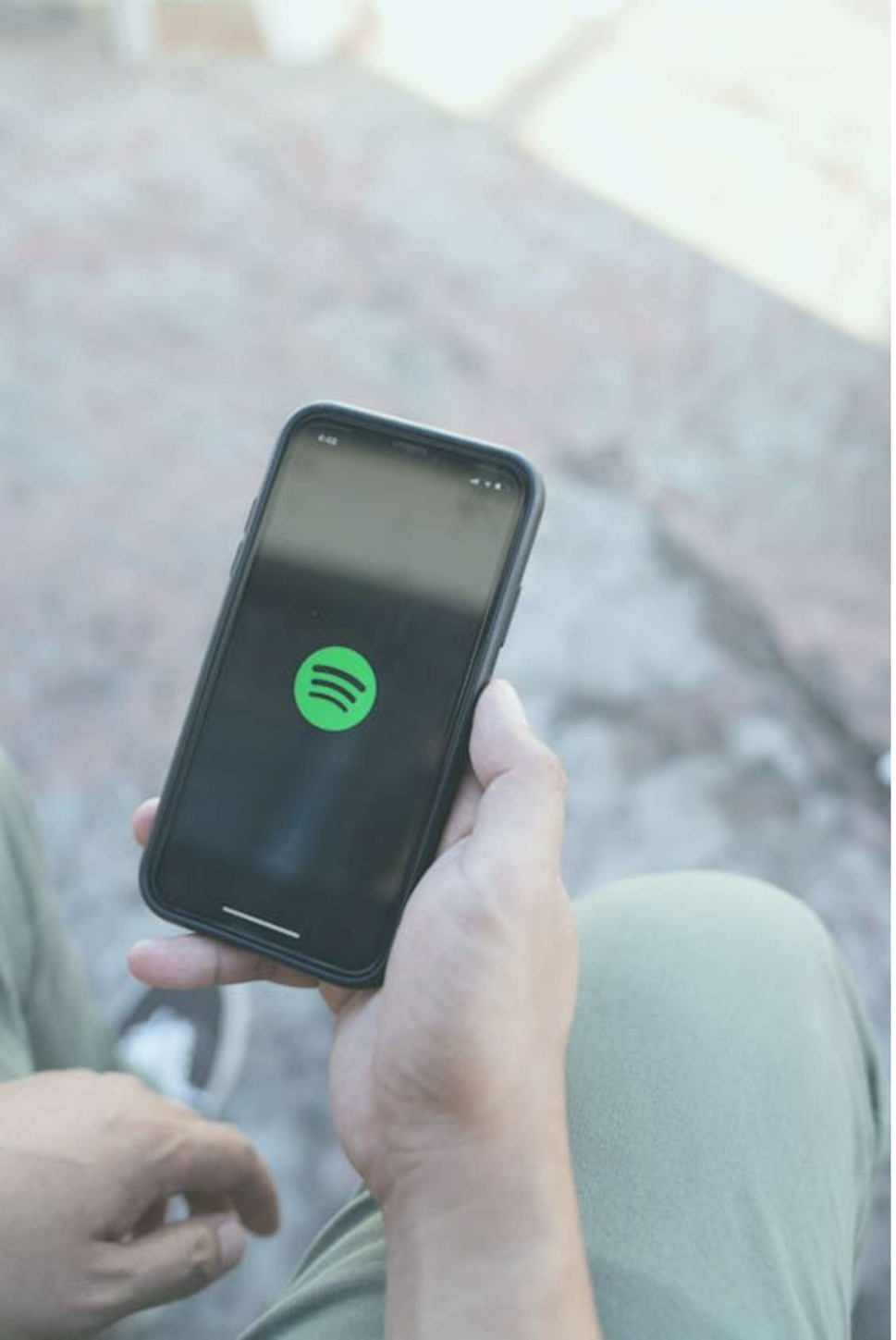
# EXPLORATORY DATA ANALYSIS OF Spotify TRACKS

PRESENTED BY RITTIK DAS



# Spotify

## 🎼Problem Statement

Consider you are Music Director/Mixing Engineer aiming to optimize new songs for popularity, leveraging insights from this dataset is key.

## 🔑 Important Points

- Popular tracks are 3–4 minutes, 100–130 BPM, and high in danceability, energy, and loudness.
- Vocal-driven songs (low instrumentalness, moderate speechiness) outperform purely instrumental tracks.
- Danceability, Energy, and Valence show strong correlation with popularity — upbeat tracks trend higher.

## 🎯 Expectations

- Mix for clarity + loudness, keep songs engaging and upbeat, and ensure strong vocal presence.
- Balance emotional tone (valence) — neither too dark nor overly bright.

## ⚡ Key Challenges

- Avoid making tracks too formulaic while following trends.
- Popularity depends on genre context and external factors (algorithms, culture) beyond audio features.

Exploring Musical Analysis

# Data Description

**Track Information**

- track_id : A unique identifier for the track on Spotify.
- track_name : The title of the song.
- artist_name : The name of the artist(s) who performed the song.
- album_name : The name of the album the track belongs to.
- year : The release year of the song.
- language : The detected language of the song's lyrics.
- artwork_url : A URL pointing to the album artwork for the track.
- track_url : A URL to the Spotify track.

**Popularity / Engagement**

- popularity : A measure of how popular a track is, ranging from 0 to 100.

# Data Description

**Audio Features (Musical Attributes)**

- acousticness : A confidence measure indicating whether the track is acoustic, ranging from -1.0 to 1.0.
- danceability : A measure of how suitable a track is for dancing, ranging from -1.0 to 1.0.
- energy : A perceptual measure of intensity and activity, ranging from -1.0 to 1.0.
- instrumentalness : Predicts whether a track contains no vocalcontent, ranging from -1.0 to 1.0.
- liveness : Detects the presence of an audience in the recording, ranging from -1.0 to 1.0.
- loudness : The overall loudness of a track in decibels (dB).
- speechiness : A measure detecting the presence of spoken words in a track.
- valence : A measure from -1.0 to 1.0 describing the musical positiveness conveyed by a track

**Technical Features (Musical Composition)**

- duration_ms : The duration of the track in milliseconds.
- tempo : The overall estimated tempo of a track in beats per minute (BPM).
- key : The key the track is in, represented as an integer (e.g., 0 = C, 1 = C#, etc.).
- mode : Indicates the modality (major or minor) of a track (0 for minor, 1 for major).
- time_signature : An estimated overall time signature of a track.

# Final Table Structure

**The dataset combines continuous audio features, discrete metrics, categorical metadata, and temporal markers. This diverse mix not only captures how songs are structured, but also reflects their cultural context and listener reception. By analyzing these variables, we can uncover patterns that guide music production, mixing decisions, and audience targeting, ultimately shaping what makes a track popular across time and genres.**

## Numerical Variables

- acousticness
- danceability
- duration_ms
- duration_sec
- energy
- instrumentalness
- liveness
- loudness
- energy_dance_ratio
- speechiness
- tempo
- valence
- popularity
- key
- mode
- time_signature
- decade

## Categorical Variables

- track_id
- track_name
- artist_name
- album_name
- artwork_url
- track_url
- language
- mood
- popularity_segment

## Temporal Variables

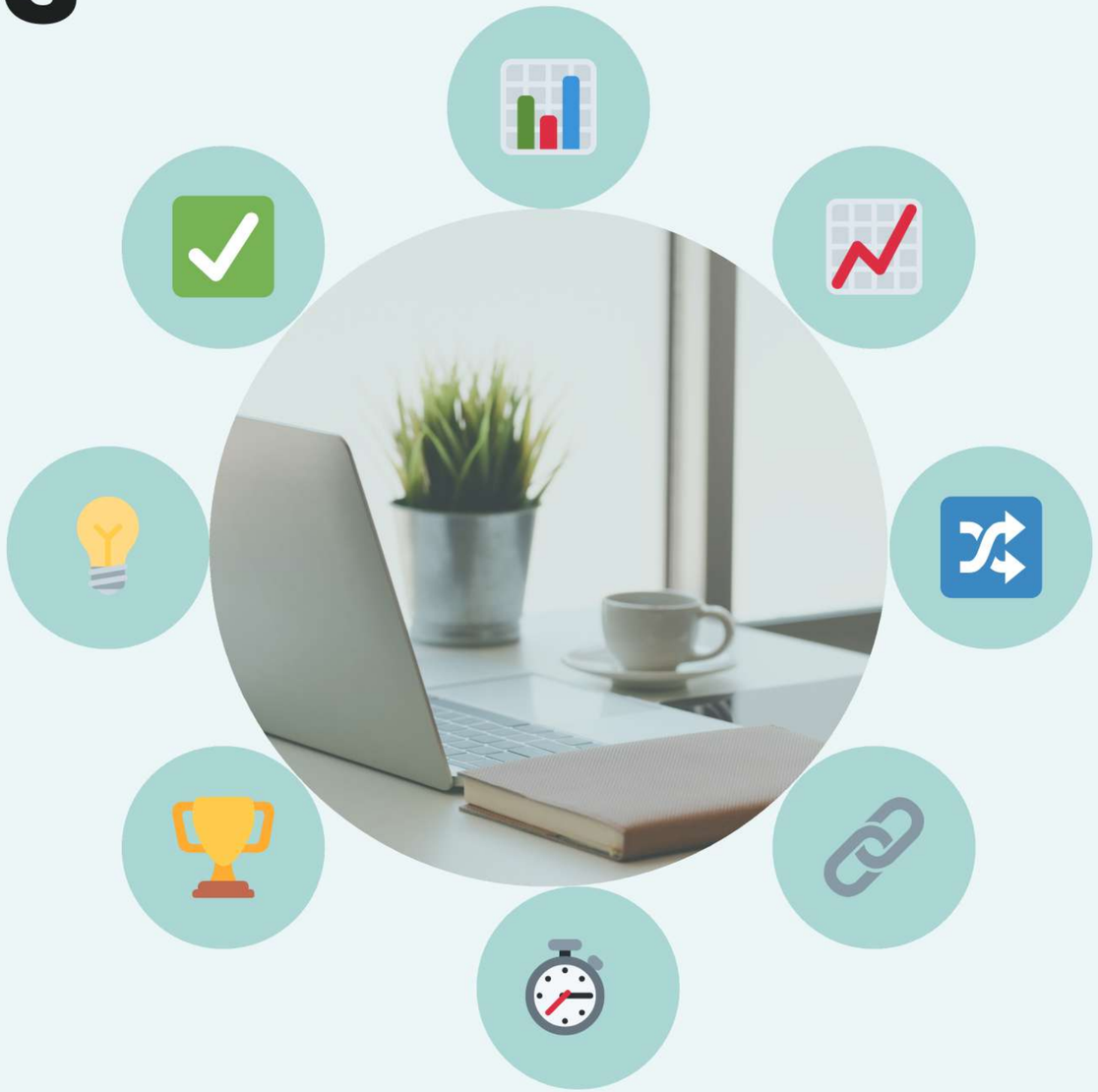
- year

Spotify

# EDA ANALYSIS

# Analysis Contains

- 📊 Statistical Analysis
- 📈 Univariate Analysis
- 🔀 Bivariate Analysis
- 🔗 Multivariate Analysis
- ⏱️ Time Series Analysis
- 🏆 Top N & Bottom N Analysis
- 💡 Some Hypothetical Analysis
- ✅ Final Recommendation



**Clear Steps for Comprehensive Insights**
Break down complex data into understandable parts for effective decision-making.

**Actionable Outcomes Driving Better Decisions**
Translate analysis into practical business recommendations and strategies.

Spotify

# *Statistical Analysis*

# Statistical Summary of Numerical Variables

```
Distribution of Numerical Variables:

1. Distribution of POPULARITY

Mean      : 16.18
Median    : 8.00
Mode      : [0]
Std Dev   : 19.08
Min       : 0.0000
25% (Q1)  : 0.00
50% (Q2)  : 8.00
75% (Q3)  : 28.00
Max       : 93.0000
Skewness  : 1.17
Kurtosis  : 0.51
============================================================

2. Distribution of DURATION_SEC

Mean      : 233.52
Median    : 228.84
Std Dev   : 102.41
Min       : 10.0270
25% (Q1)  : 188.89
50% (Q2)  : 228.84
75% (Q3)  : 276.71
Max       : 4581.4830
Skewness  : 10.32
Kurtosis  : 314.23
============================================================

3. Distribution of KEY

Mean      : 5.10
Median    : 5.00
Mode      : [0.0]
Std Dev   : 3.56
Min       : 0.0000
25% (Q1)  : 2.00
50% (Q2)  : 5.00
75% (Q3)  : 8.00
Max       : 11.0000
Skewness  : 0.06
Kurtosis  : -1.27
============================================================
```

**Popularity**

- Mean : 16, Median: 8, Mode: 0 → **Most tracks are at the low end of popularity.**
- Skewness = 1.17 → long right tail (a few highly popular tracks raise the mean).
- The track set is **dominated by unpopular or niche songs,** only a small fraction achieve very high popularity.

**Duration**

- Mean : 233 sec (~3.9 min), Median: 229 sec.
- Typical range (Q1–Q3): 189 – 277 sec (~3.1–4.6 min).
- **some extreme tracks up to 76 minutes** (4581 sec) → skewness 10.3.
- The **typical song length is ~3–4 minutes,** aligning with standard commercial music. Outliers may represent live recordings or experimental tracks.

**Keys**

- Mode : Key = 0 (C major).
- Spread is relatively uniform (Q1=2, Median=5, Q3=8).
- Songs are distributed across keys, but C major is slightly more common. No strong bias toward unusual keys.

# Statistical Summary of Numerical Variables

```
4. Distribution of TEMPO

Mean      : 118.08
Median    : 118.34
Std Dev   : 28.44
Min       : 0.0000
25% (Q1)  : 96.01
50% (Q2)  : 118.34
75% (Q3)  : 135.03
Max       : 232.1980
Skewness  : 0.36
Kurtosis  : -0.15
============================================================

5. Distribution of ACOUSTICNESS

Mean      : 0.35
Median    : 0.26
Std Dev   : 0.32
Min       : 0.0000
25% (Q1)  : 0.06
50% (Q2)  : 0.26
75% (Q3)  : 0.62
Max       : 0.9960
Skewness  : 0.54
Kurtosis  : -1.12
============================================================

6. Distribution of LOUDNESS

Mean      : -8.98
Median    : -7.31
Std Dev   : 5.64
Min       : -45.9200
25% (Q1)  : -10.81
50% (Q2)  : -7.31
75% (Q3)  : -5.29
Max       : 0.8940
Skewness  : -1.78
Kurtosis  : 3.80
============================================================
```

**Tempo**

- Mean/Median : ~118 BPM.
- Typical range : 96–135 BPM (Q1–Q3).
- Balanced distribution, no strong skew.
- Most songs fall into mid-tempo range, typical for pop and mainstream music (danceable but not extremely fast/slow).

**Acousticness**

- Median : 0.26 (fairly low).
- Distribution spread across full 0–1 range, but right-skewed.
- The track set leans toward electronic/produced tracks but still includes many acoustic songs.

**Loudness**

- Median : -7.3 dB, Typical range: -10.8 to -5.3 dB.
- **Most tracks are loud and compressed,** consistent with modern production standards.

# Statistical Summary of Numerical Variables

```
7. Distribution of DANCEABILITY

Mean      : 0.59
Median    : 0.63
Std Dev   : 0.19
Min       : 0.0000
25% (Q1)  : 0.49
50% (Q2)  : 0.63
75% (Q3)  : 0.73
Max       : 0.9790
Skewness  : -0.76
Kurtosis  : 0.06
============================================================

8. Distribution of ENERGY

Mean      : 0.60
Median    : 0.65
Mode      : [0.65]
Std Dev   : 0.25
Min       : 0.0002
25% (Q1)  : 0.43
50% (Q2)  : 0.65
75% (Q3)  : 0.81
Max       : 1.0000
Skewness  : -0.57
Kurtosis  : -0.56
============================================================

9. Distribution of TIME_SIGNATURE

Mean      : 3.87
Median    : 4.00
Mode      : [4.0]
Std Dev   : 0.47
Min       : 0.0000
25% (Q1)  : 4.00
50% (Q2)  : 4.00
75% (Q3)  : 4.00
Max       : 5.0000
Skewness  : -3.25
Kurtosis  : 16.43
============================================================
```

**Danceability**

- Median : 0.63, Typical range: 0.49–0.73
- Slight skew toward higher values.
- **Songs are generally danceable, which fits with mainstream modern track sets**.

**Energy**

- Median : 0.65, Mode: 0.65 → clustering around medium-high energy.
- Balanced distribution, slight left skew.
- **The track set leans toward energetic tracks rather than calm/ambient music.**

**Time Signature**

- Mode = 4, overwhelmingly dominant.
- Very few tracks in 3 or 5.
- **4/4 is the standard time signature, as expected in pop and electronic music.**

# Statistical Summary of Numerical Variables

```
10. Distribution of SPEECHINESS

Mean      : 0.09
Median    : 0.05
Std Dev   : 0.12
Min       : 0.0000
25% (Q1)  : 0.04
50% (Q2)  : 0.05
75% (Q3)  : 0.09
Max       : 0.9590
Skewness  : 4.47
Kurtosis  : 24.80
============================================================

11. Distribution of VALENCE

Mean      : 0.49
Median    : 0.50
Std Dev   : 0.26
Min       : 0.0000
25% (Q1)  : 0.28
50% (Q2)  : 0.50
75% (Q3)  : 0.70
Max       : 0.9950
Skewness  : -0.10
Kurtosis  : -1.04
============================================================

12. Distribution of INSTRUMENTALNESS

Mean      : 0.16
Median    : 0.00
Std Dev   : 0.32
Min       : 0.0000
25% (Q1)  : 0.00
50% (Q2)  : 0.00
75% (Q3)  : 0.04
Max       : 0.9990
Skewness  : 1.63
Kurtosis  : 0.87
============================================================
```

**Speechiness**

- Median : 0.05 → most songs are low speechiness (sung rather than spoken).
- Long tail up to 0.95 → rap/spoken-word tracks.
- The track set is mostly lyrical, but includes some rap songs.

**Valence**

- Median : 0.50 → balanced mood.
- Spread : wide across 0–1.
- **The track set contains both happy and sad tracks, with no strong lean.**

**Instrumentalness**

- Median : 0.00, Q3=0.04 → most tracks are non-instrumental (vocal).
- **Few high values near 1 → fully instrumental tracks.**
- **The track set leans heavily toward vocal tracks, but still represents instrumental music.**

# Statistical Summary of Numerical Variables

```
13. Distribution of LIVENESS

Mean      : 0.20
Median    : 0.13
Std Dev   : 0.18
Min       : 0.0119
25% (Q1)  : 0.09
50% (Q2)  : 0.13
75% (Q3)  : 0.25
Max       : 0.9980
Skewness  : 2.25
Kurtosis  : 5.41
============================================================

14. Distribution of MODE

Mean      : 0.58
Median    : 1.00
Mode      : [1.0]
Std Dev   : 0.49
Min       : 0.0000
25% (Q1)  : 0.00
50% (Q2)  : 1.00
75% (Q3)  : 1.00
Max       : 1.0000
Skewness  : -0.32
Kurtosis  : -1.90
============================================================

15. Distribution of DECADE

Mean      : 2010.45
Median    : 2010.00
Mode      : [2020]
Std Dev   : 10.98
Min       : 1970.0000
25% (Q1)  : 2010.00
50% (Q2)  : 2010.00
75% (Q3)  : 2020.00
Max       : 2020.0000
Skewness  : -1.30
Kurtosis  : 1.25
============================================================
```

**Liveness**

- Median : 0.13, typical range 0.09–0.25.
- Some very high values (~1.0) → live recordings.
- **Mostly studio recordings,** with a smaller presence of live performances.

**Mode**

- Mode = 1 (major), median also 1 → majority of songs are in major keys.
- The track set leans major (happier), but minor-key songs are also present.

**Decade**

- Median : 2010, Mode: 2020.
- Majority of songs from 2010s and 2020s.
- The track set is modern-heavy, less historical coverage before 2000.

# Statistical Summary of Numerical Variables

```
16. Distribution of ENERGY_DANCE_RATIO

Mean      : 85.84
Median    : 1.03
Std Dev   : 6180.76
Min       : 0.0003
25% (Q1)  : 0.76
50% (Q2)  : 1.03
75% (Q3)  : 1.31
Max       : 810000.0000
Skewness  : 91.28
Kurtosis  : 9513.81
============================================================
```

**Energy-Dance Ratio**

- Median : ~1.0, IQR: 0.76–1.31 → balanced energy vs danceability.
- **Ignoring outliers, most songs maintain a natural balance between energy and danceability.**

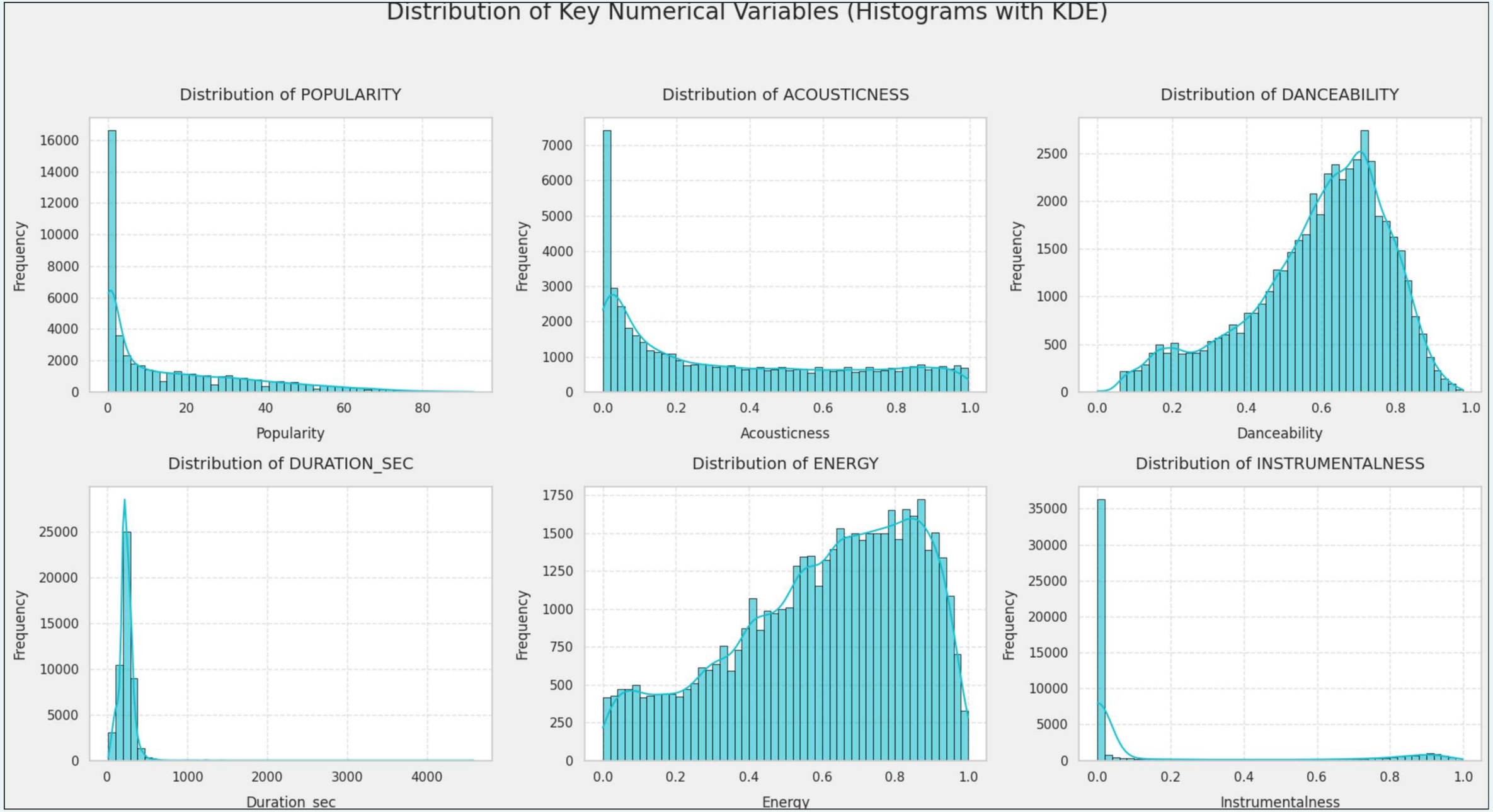
**General Overview**

- **Typical Song Profile : A ~3.5 min, mid-tempo (118 BPM),** danceable, energetic, studio-recorded pop track in a major key.
- **Bias : Skews toward recent, mainstream, high-energy, major-key songs.**

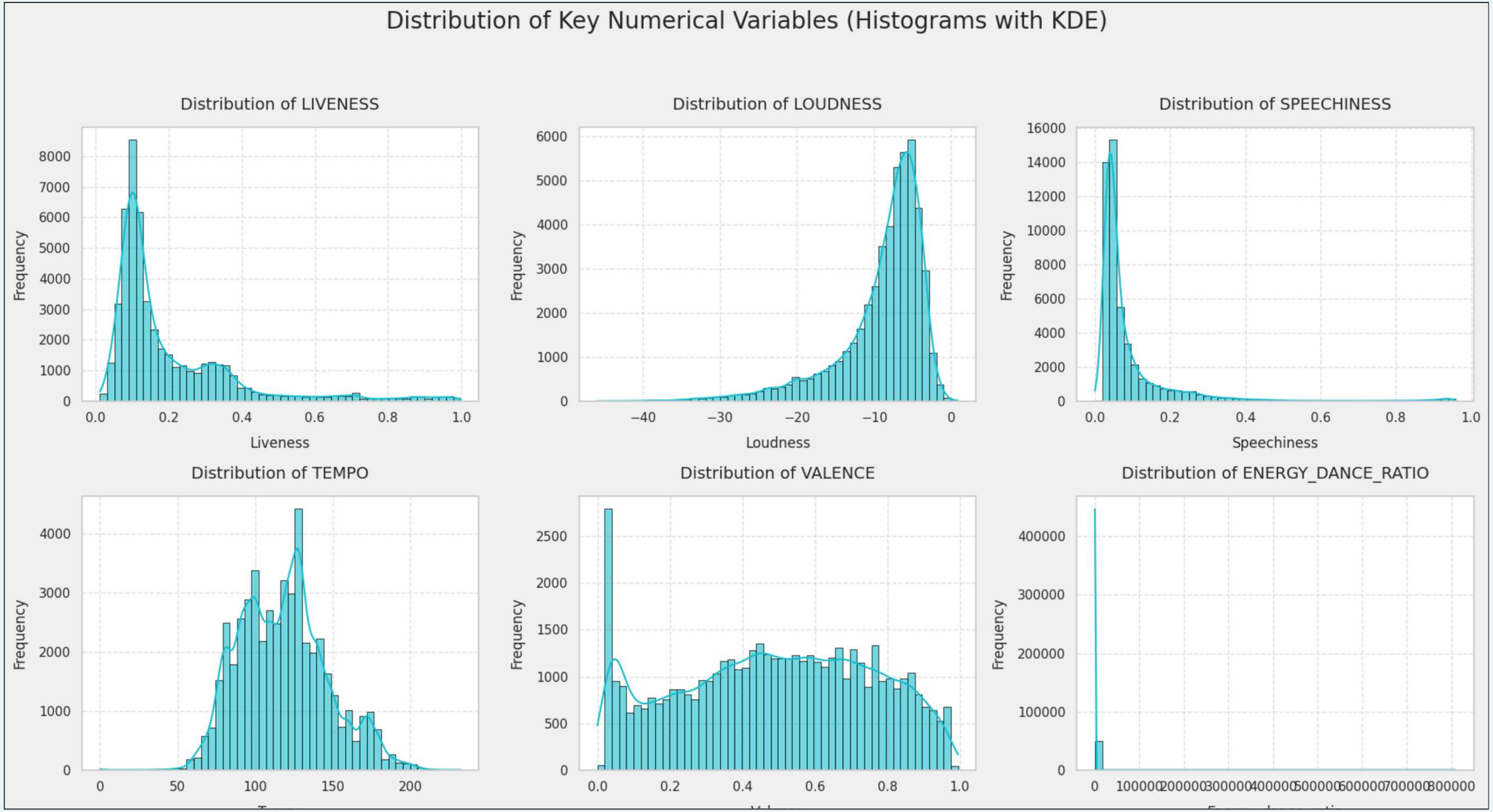
Spotify

# *Univariate Analysis*

# Visual Analysis Of Numerical Variables



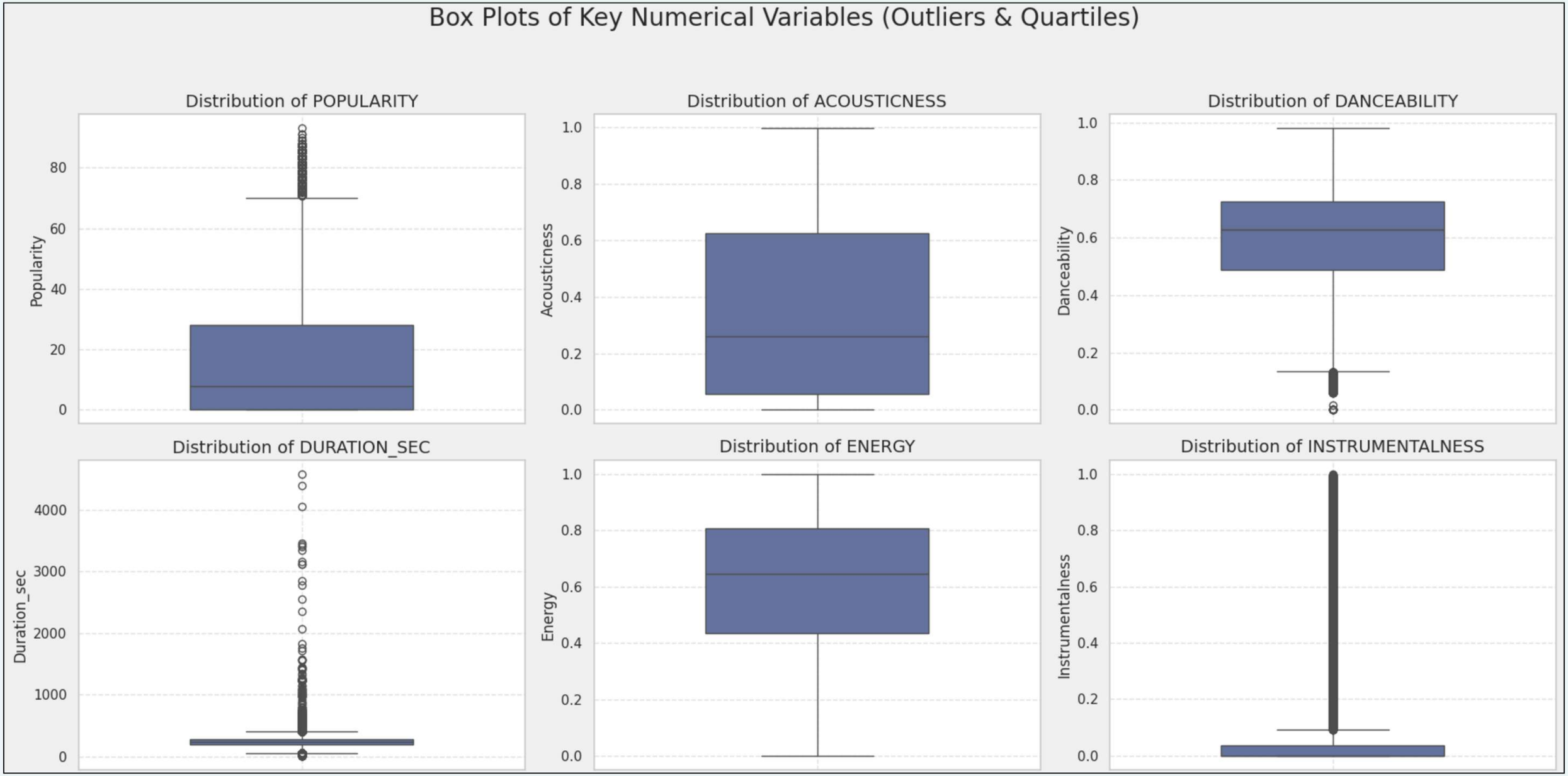
# Visual Analysis Of Numerical Variables



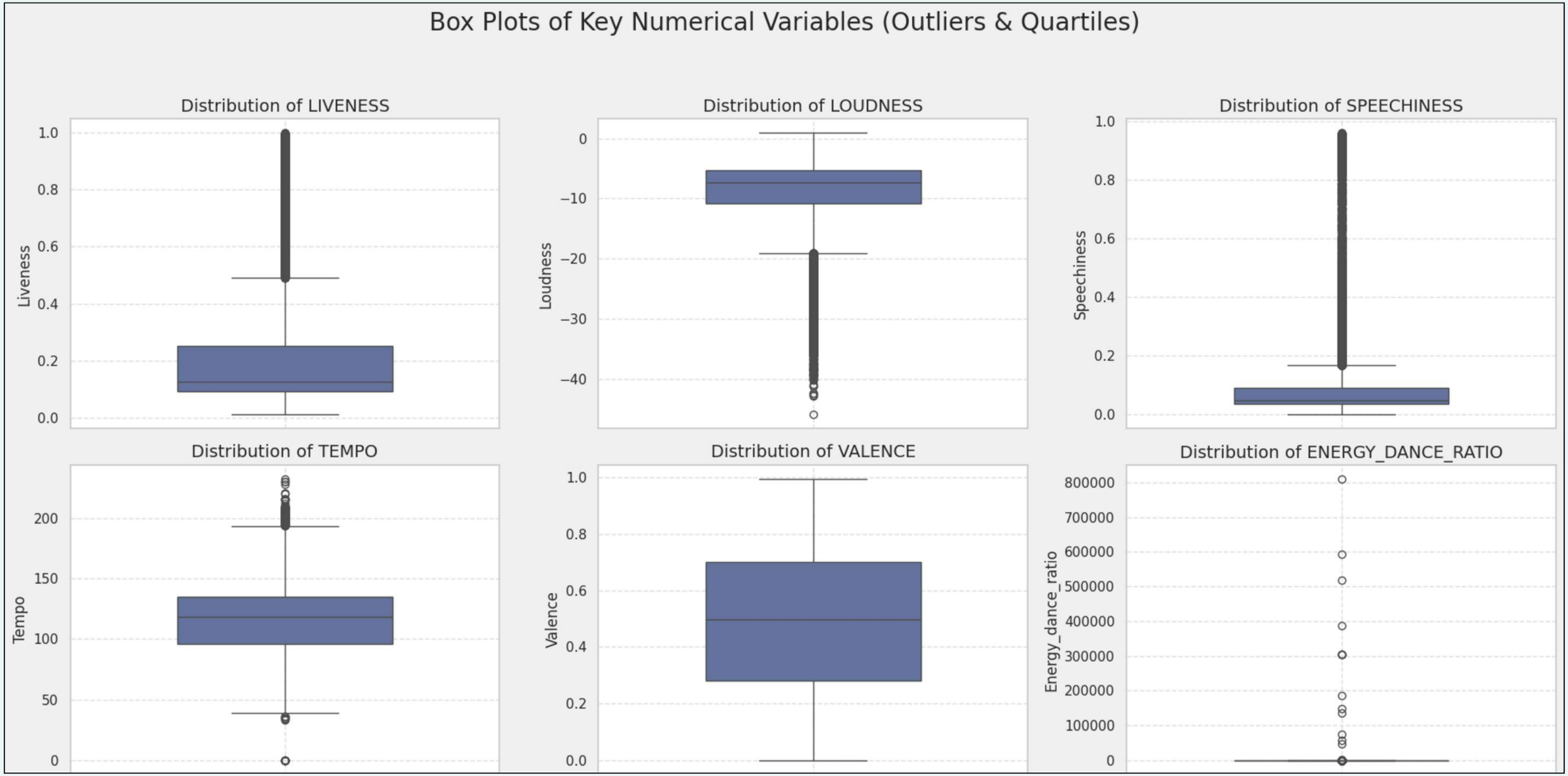
# Overview

- **Popularity :** **Highly right-skewed with most tracks clustered at very low popularity,** only a small tail reaches moderate-to-high popularity, indicating a long-tail distribution where hits are rare.
- **Acousticness :** Broad spread with many tracks at low acousticness and a steady presence across mid-to-high values; catalog leans more electric/produced than purely acoustic.
- **Danceability :** Unimodal concentration around 0.6–0.8, implying the dataset **is generally oriented toward moderately to highly danceable** tracks.
- **Duration :** Strong right skew centered roughly around 180–240 seconds, with a small tail of very long tracks; typical song lengths dominate.
- **Energy :** Distribution weighted toward 0.6–0.9 with fewer low-energy tracks, **showing a preference for energetic music.**
- **Instrumentalness :** Mass piled near zero with a minor high-end bump, indicating most tracks contain vocals with a small subset of true instrumentals.
- **Liveness :** Predominantly low liveness around 0.1–0.2 with a long right tail, suggesting most tracks are studio recordings with few live-performance cues.
- **Loudness :** Concentration between roughly −12 to −5 dB with very few near 0 dB, indicating generally well-compressed but not clipping mixes typical of commercial releases.
- **Speechiness :** Heavily skewed near 0 with sparse high values, meaning most tracks are music-dominant with limited spoken-word or rap-like segments.
- **Tempo :** Bimodal-to-multimodal clusters around ~95–110 BPM and ~120–135 BPM, **aligning with common pop/hip-hop and dance tempos.**
- **Valence :** Fairly even spread with a tilt toward mid-to-high valence (0.4–0.8), **implying the catalog leans more positive/cheerful than dark.**
- **Energy Dance Ratio :** **Mass near small values with extreme outliers, signaling the ratio is unstable for some records** and dominated by a few anomalously large cases; consider winsorizing or using log/robust scaling before modeling.

# Visual Analysis Of Numerical Variables



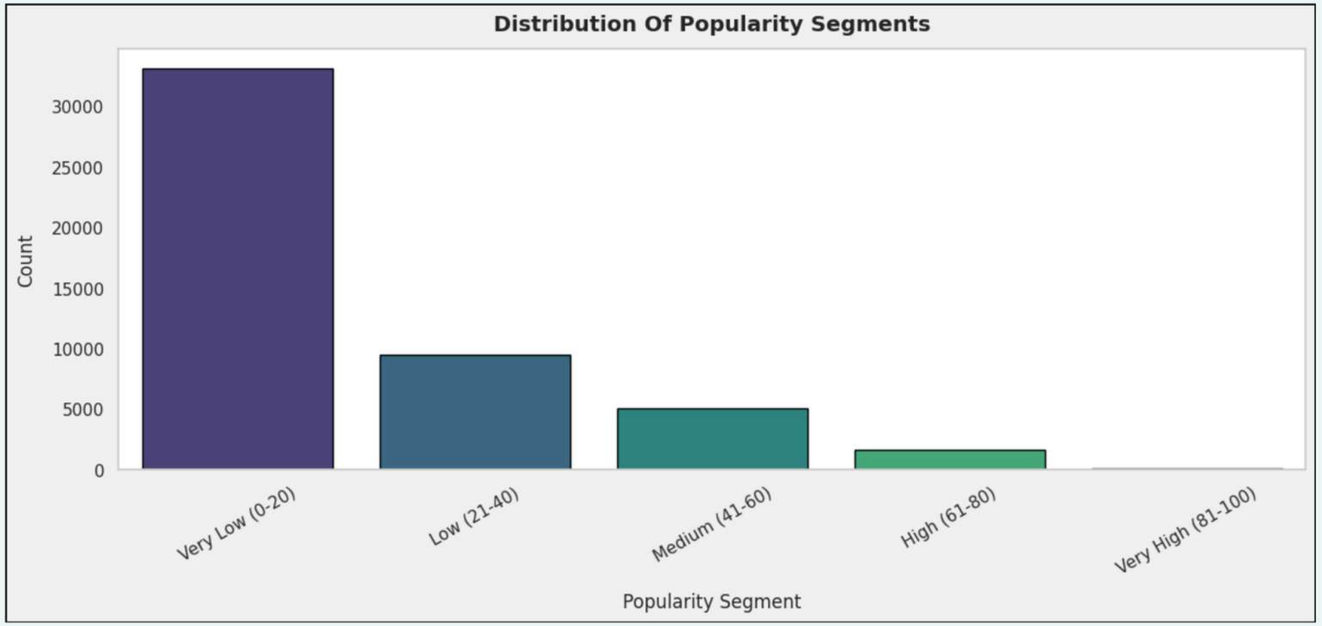
# Visual Analysis Of Numerical Variables

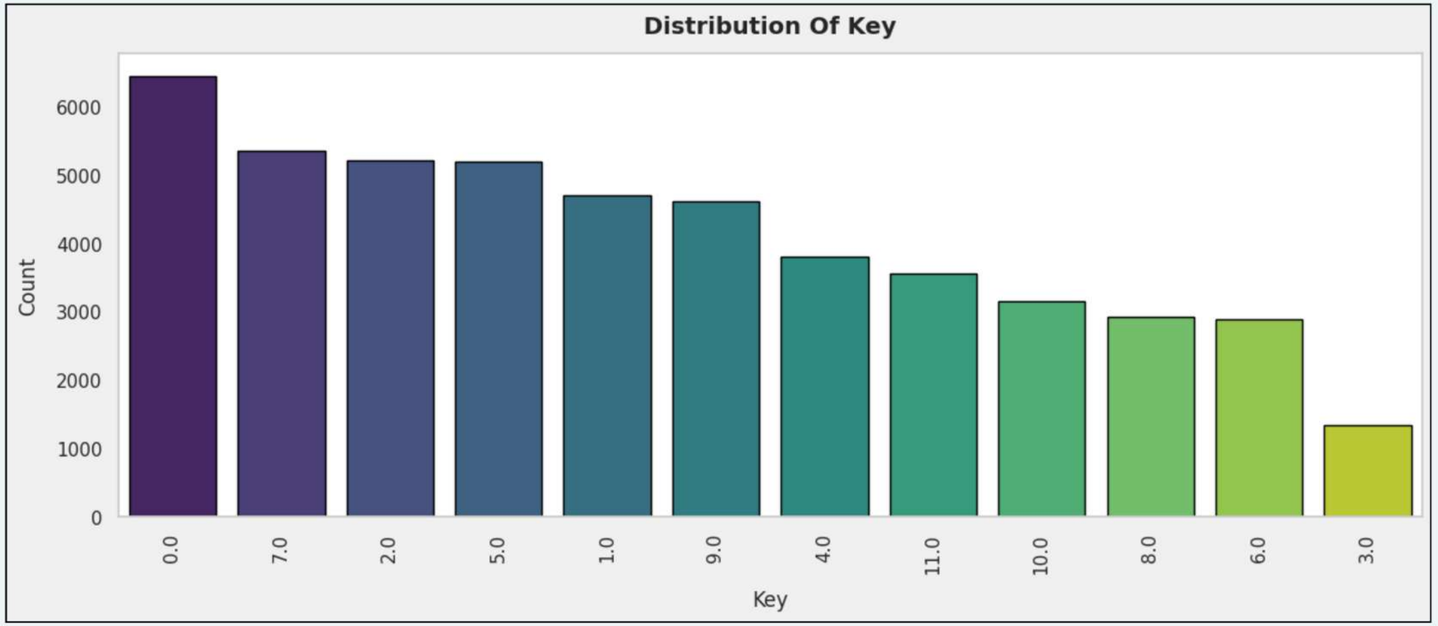


# Overview

- **Popularity :** Median near the low single digits with IQR roughly 1–25, and a long upper whisker with many high-end outliers up to ~90, **confirming a heavy right skew where a small fraction of tracks are disproportionately popular.**
- **Acousticness :** Median around ~0.25 with IQR ~0.07–0.60, showing broad variability; numerous low-end and high-end points beyond whiskers indicate both very electric and very acoustic extremes present as outliers.
- **Danceability :** Median near ~0.62 with a compact IQR ~0.50–0.72, **suggesting consistency toward moderate–high danceability**, sparse low-end outliers exist but few extreme highs beyond ~0.95.
- **Duration :** Median around standard song length (~200–240s) with a tight IQR centered near typical pop durations, but extreme upper outliers extend well past 1000s, **highlighting a long tail of unusually long tracks and a handful of short anomalies.**
- **Energy :** Median near ~0.65 with IQR ~0.45–0.80, **indicating a tilt toward energetic content**, both low-energy and near-1.0 values appear as outliers, reflecting niche ambient and peak-intensity cuts.
- **Instrumentalness :** Median near zero with a very low IQR close to 0–0.05, and a dense cluster of high-end outliers up to 1.0, **confirming that vocals dominate while a minority are strongly instrumental.**
- **Liveness :** Median ~0.15 with IQR ~0.08–0.25; numerous high-end outliers beyond ~0.45 up to 1.0 indicate most tracks are studio-like while a minority show strong live characteristics.
- **Loudness :** Median around −8 to −9 dB with IQR roughly −11 to −6 dB; substantial lower outliers extend below −20 dB suggesting a tail of very quiet/mastered-low tracks.
- **Speechiness :** Median near ~0.05 with tight IQR ~0.03–0.08; a dense spike of outliers above ~0.15 up to ~1.0 reflects a small subset with spoken-word/rap features.
- **Tempo :** **Median ~115–120 BPM with IQR ~100–130 BPM**; outliers include very slow (<40 BPM) and very fast (>190 BPM) pieces, **revealing edge-case tempos.**
- **Valence :** Median ~0.5 with IQR ~0.3–0.7; **symmetric spread with moderate tails and few extremes indicates a balanced mix of moods from sad to cheerful.**
- **Energy Dance Ratio :** Median near zero with extremely compressed IQR and massive positive outliers reaching ~800k, **indicating a highly unstable ratio dominated by anomalies and requiring clipping/log-transform before modeling**

# Visual Analysis Of Discrete Numerical Variables

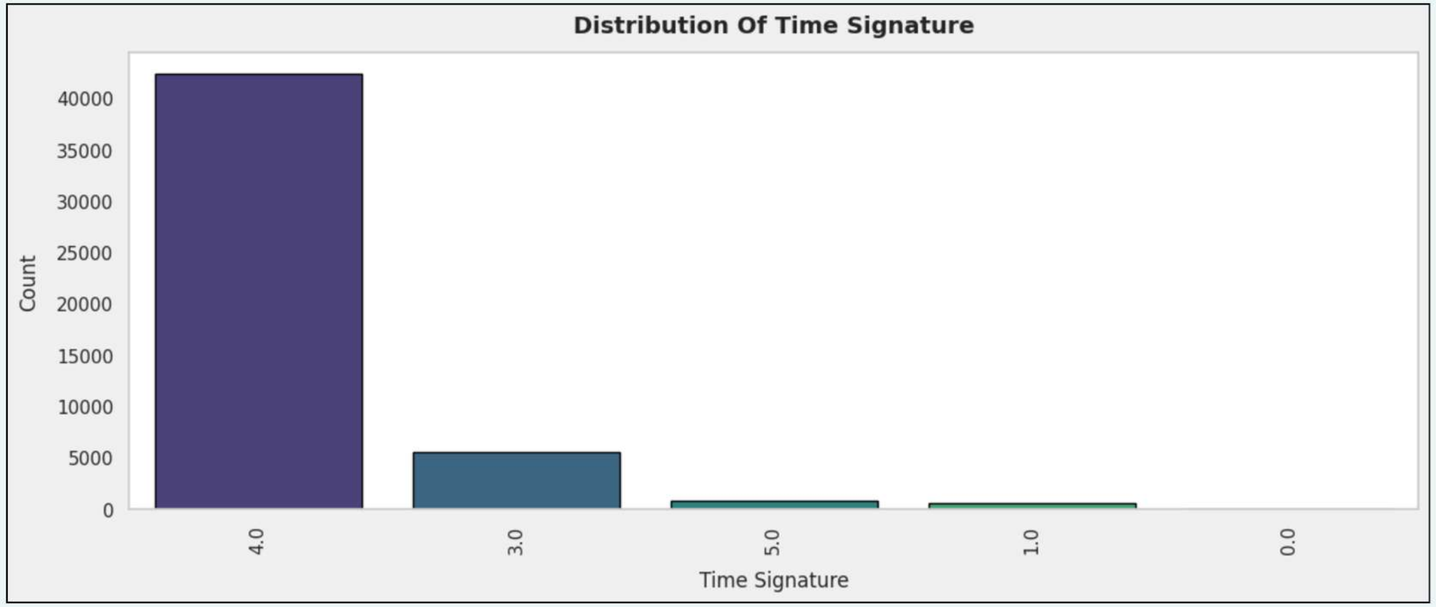


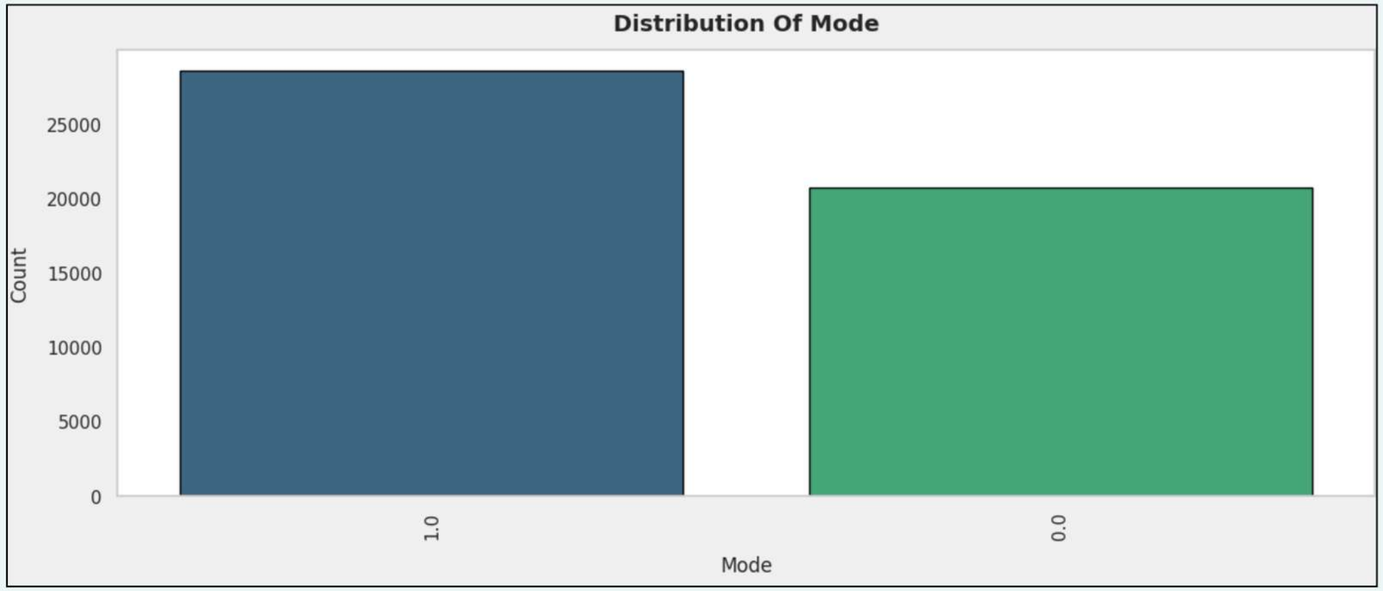

- **Popularity Segments : The catalog is dominated by the "Very Low" bucket, with a steep drop as popularity rises**— prioritize a two-tier strategy: growth hacks (collabs, playlist pitching, TikTok snippets) for the long tail, and concentrated ad spend on the small "High" segment to push proven tracks into "Very High" for outsized ROI.

- **Key : Key "0" i.e. "C" leads by a wide margin, with keys 7("G"), 2("D"), and 5("F") also strong,** suggesting playlists and releases skew toward a few musical keys— target these keys for covers, remixes, or catalog expansion to align with listener comfort zones and maximize completion rates.

# Visual Analysis Of Discrete Numerical Variables

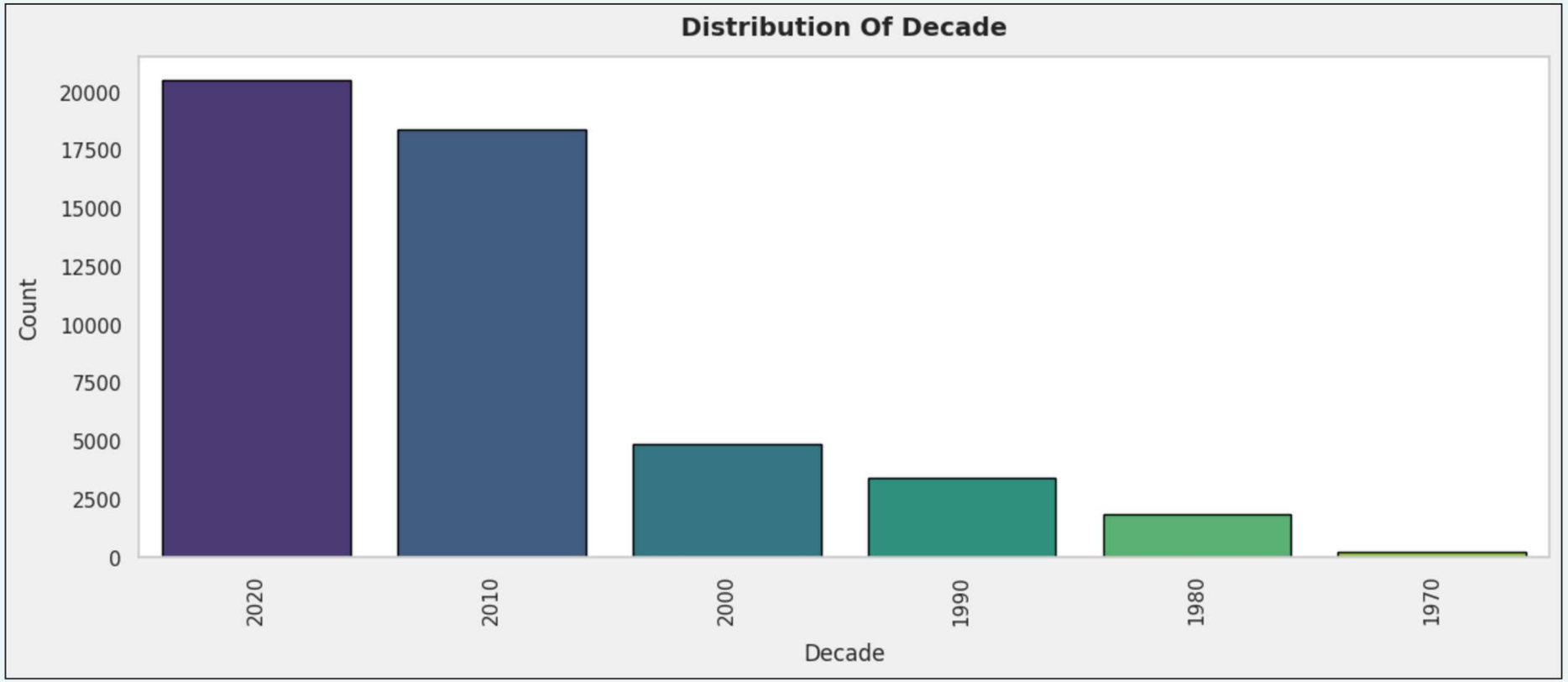



- **Time Signature : 4/4 overwhelmingly dominates,** with **small niches in 3/4 and 5/4**—focus production and A&R on 4/4 for scale while using 3/4 selectively to differentiate ballads and sync pitches without sacrificing mass appeal.



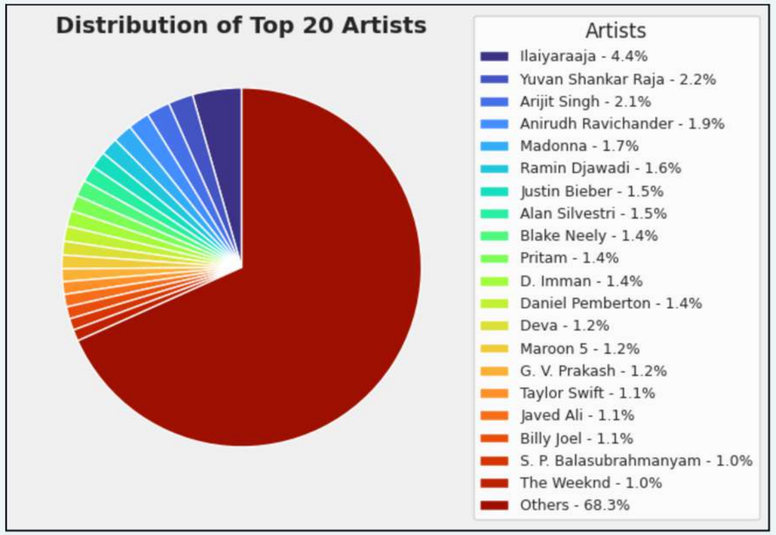

- **Mode (Major vs Minor) :** Major mode slightly outweighs minor—maintain **a core release mix that leans major for radio and playlist friendliness,** while reserving minor-mode tracks for mood/genre verticals (e.g., alt, trap, cinematic) where engagement is deeper even if audience is narrower.

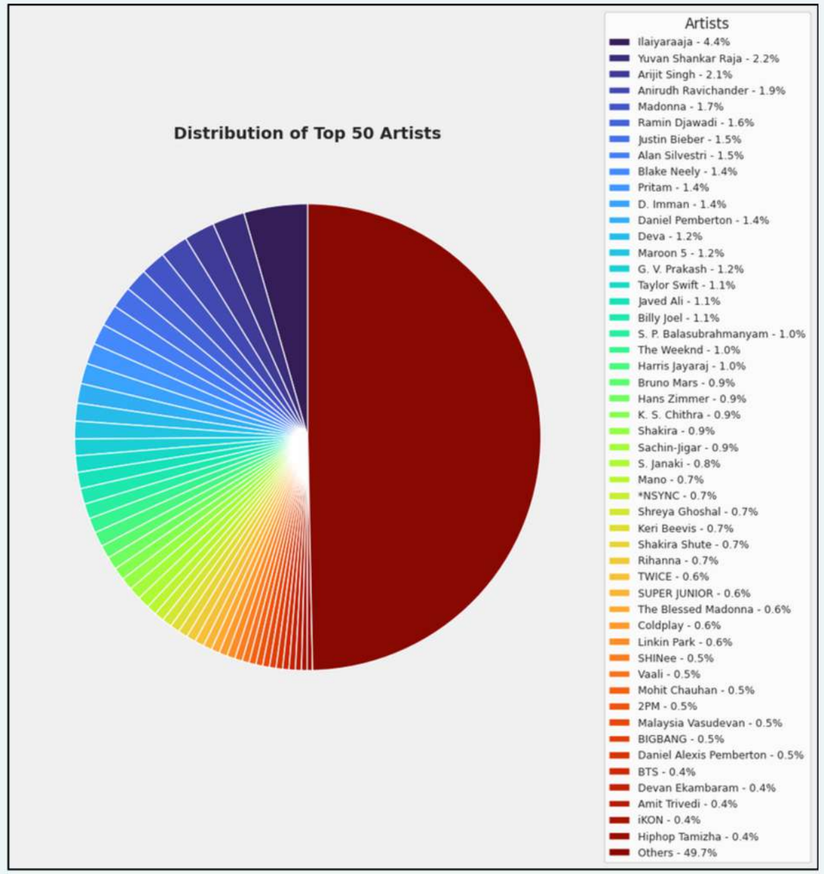
# Visual Analysis Of Discrete Numerical Variables

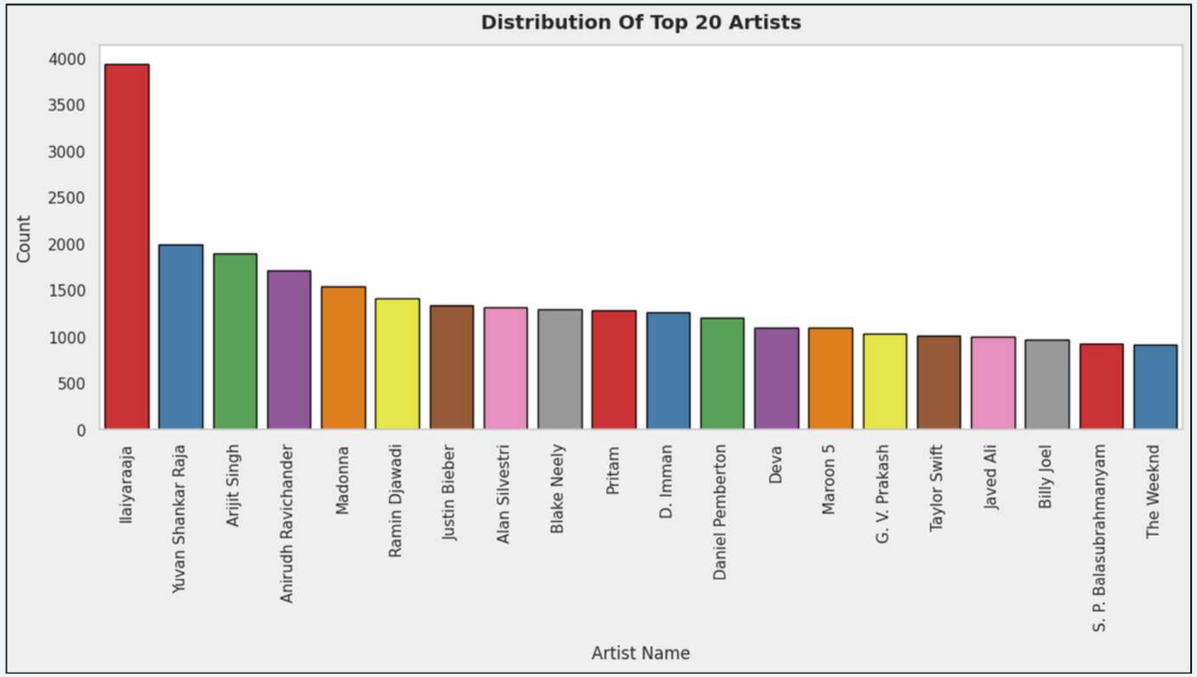



- **Decade :** **Recent music dominates the catalog, with the 2020s and 2010s contributing the bulk of tracks** and **sharp falloff before 2000**—focus marketing, playlisting, and licensing on modern catalogs for scale, while packaging pre-2000 content as niche “throwback” products to capture nostalgia without over-investing inventory.

# Visual Analysis Of Categorical Variable - Artist Name



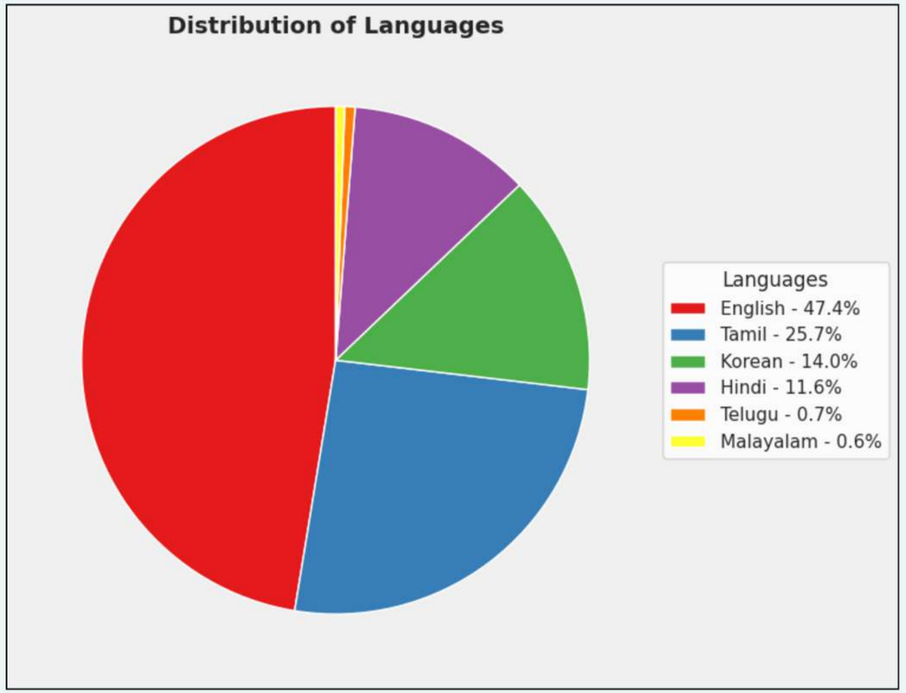



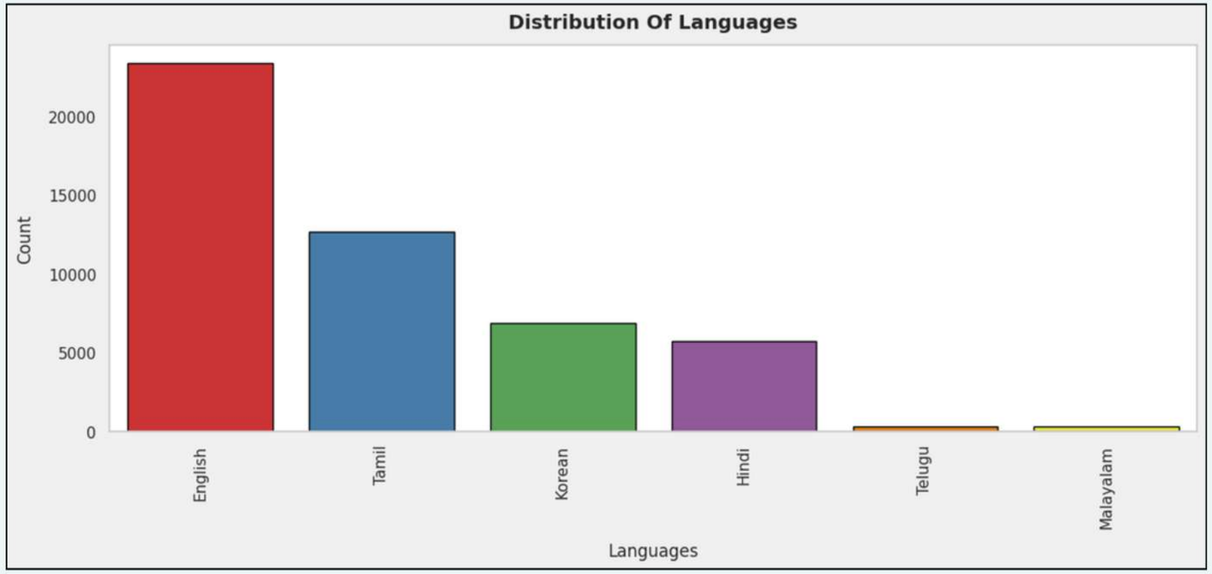



# Overview

- **Heavy long tail :** **"Others" make up ~68%, showing a highly fragmented catalog**—use algorithmic discovery playlists and UGC to surface niche artists efficiently. **50% of all songs come from just the top 50 artists, meaning only ~0.78% of singers account for half the catalog**—build discovery products for the long tail while concentrating marketing on this elite cohort for efficient scale.
- **Clear tier-1 leaders :** **Ilaiyaraaja (around 4.4%) is the single largest draw, followed by Yuvan Shankar Raja (around 2.2%), Arijit Singh (around 2.1%), and Anirudh Ravichander (around 1.9%)**—anchor marquee campaigns and catalog remasters around these names for maximum uplift.
- **Mixed global-regional appeal :** **Presence of Madonna, Justin Bieber, Taylor Swift, The Weeknd alongside South Asian leaders suggests dual-track programming**—global pop for scale, regional icons for depth and loyalty.
- **Catalog leverage :** **Top-20 artists each contribute ~1–4%—bundle greatest-hits**, mood-based compilations, and era boxes to drive replays with minimal new production cost.
- **Growth play in mid-tiers :** Artists ranked 5–20 have similar counts—target them with incremental promo (playlist top slots, short ads) where marginal spend can push tracks into breakout territory.
- **Creator partnerships :** The long-tail volume is ideal for influencer packs and editorial "discovery" lanes, converting small fan pockets into steady stream growth.

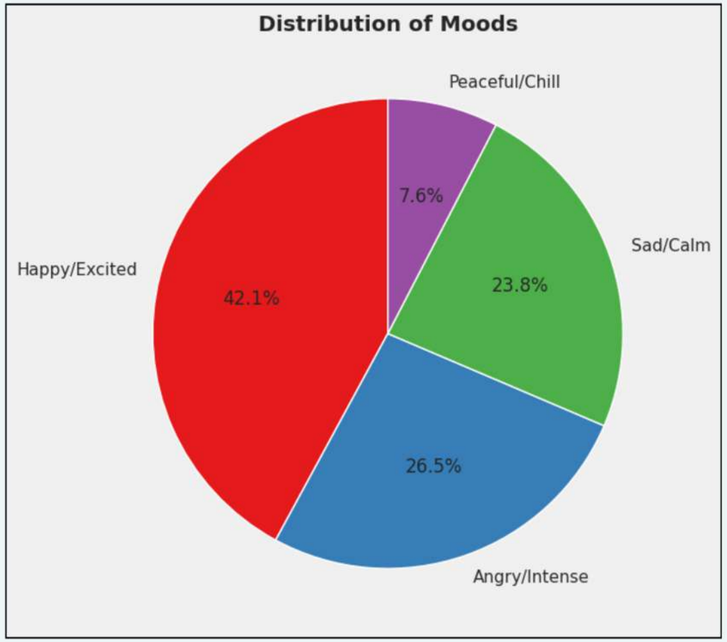
# Visual Analysis Of Categorical Variable - Language

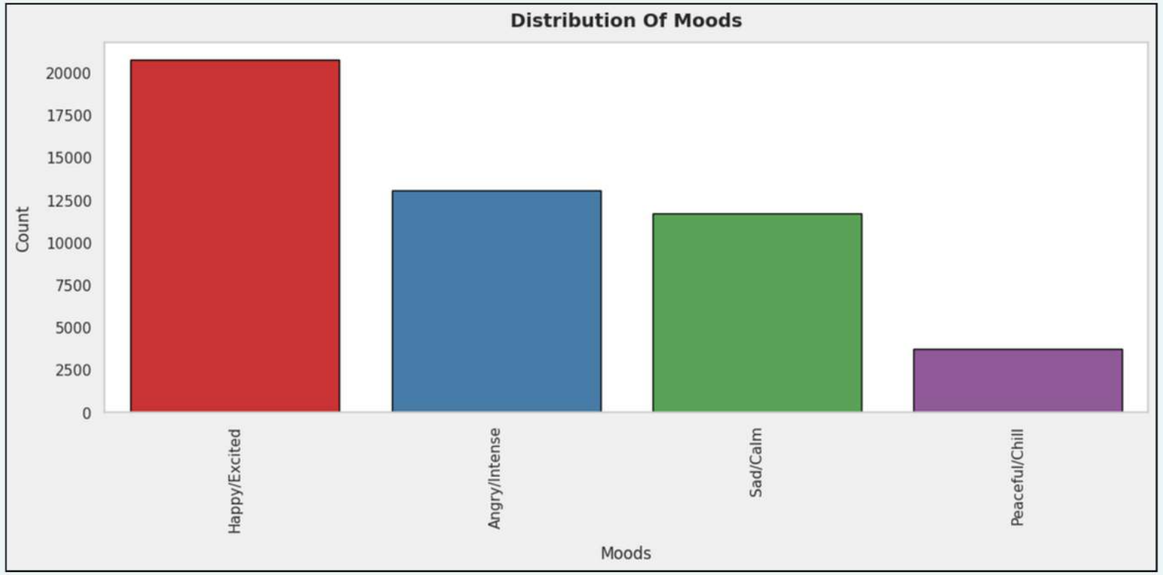



- **English is the anchor market :** **Nearly half of the catalog is in English (47.4%), making it the natural priority for mainstream playlisting**, global ad targeting, and partnerships; optimize SEO, metadata, and release calendars with English-first campaigns to drive scale.
- **Strong South Indian footprint :** **Tamil accounts for a quarter of the library (25.7%), indicating a sizable regional opportunity**—curate Tamil-first editorial shelves, creator collabs, and festival-timed drops to deepen engagement and subscription conversion in TN diaspora clusters.
- **Rising K-wave potential :** **Korean content at ~14.0% is substantial enough to justify dedicated K-pop/K-indie lanes**—bundle fan-centric features (fancams, lyrics, photocards) and sync with comeback calendars to boost repeat listens and merch tie-ins.
- **Hindi is underweighted vs market size :** At 11.6%, **Hindi share is smaller than general audience demand in India**—expand acquisition of Hindi catalog and label deals, and repurpose high-performing English/Tamil formats into Hindi to lift reach.
- **Niche tail (Telugu, Malayalam) is small but strategic :** Each is below 1%, ideal for high-ROI micro-verticals—launch city-level playlists, hyperlocal influencer series, and vernacular podcast crossovers to grow cheaply while improving regional retention.
- **Portfolio balance for growth :** **Current mix supports a dual strategy—use English and Tamil for scale, while Korean and Hindi drive growth spikes through event-led campaigns;** allocate experimental budget to Telugu/Malayalam to discover breakout regional hits.

# Visual Analysis Of Categorical Variable - Mood



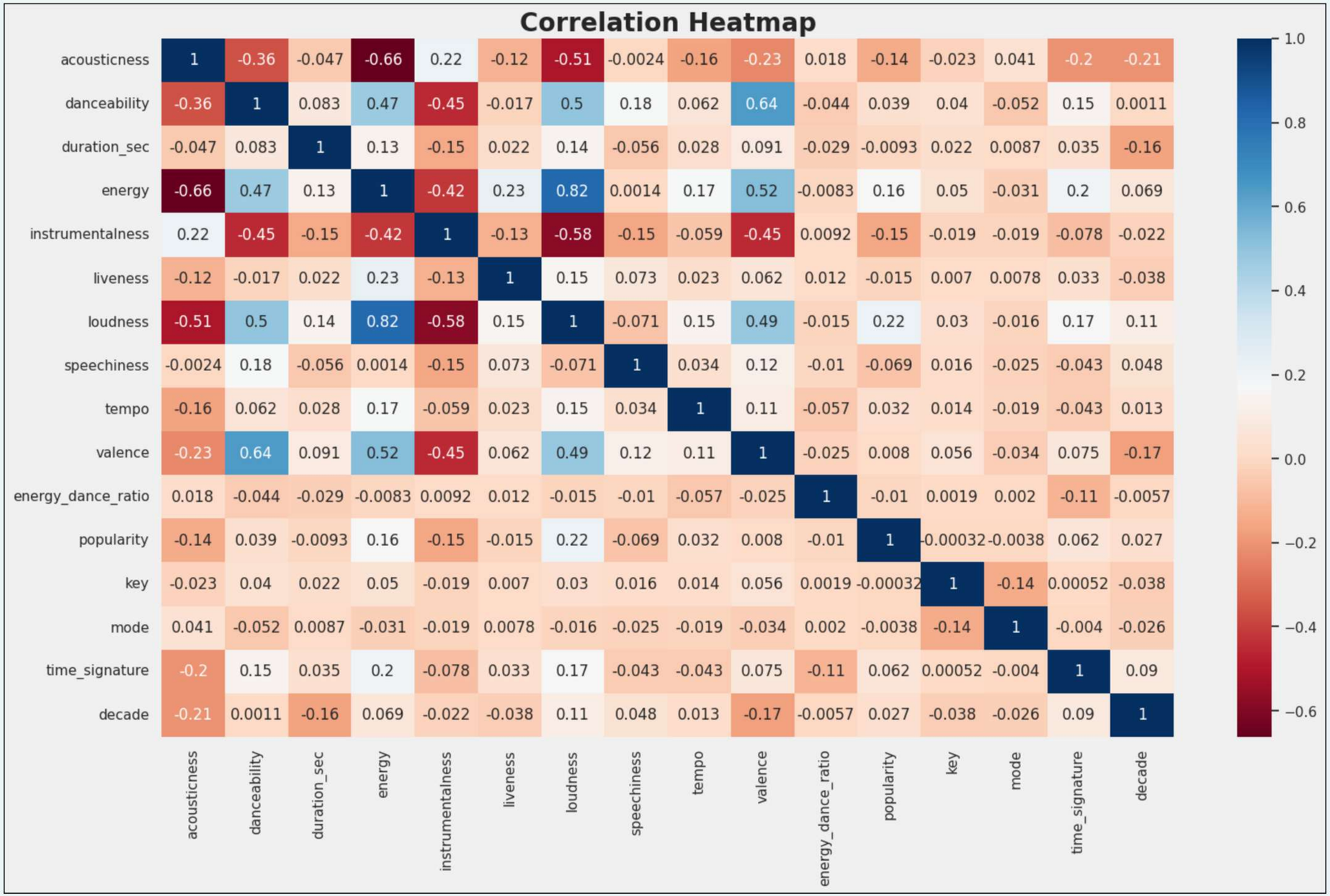




- **Happy/Excited (42.1%) :** **Core demand** center—prioritize upbeat release calendars, **workout/party playlists, and short-form promo** to maximize streams and ad yield.
- **Angry/Intense (26.5%) :** **Strong secondary lane—lean into gym/gaming niches and brand collabs needing high-energy soundtracks for reliable engagement.**
- **Sad/Calm (23.8%) :** Stable long-play category—optimize for late-night, study, and focus playlists to drive session length and retention.
- **Peaceful/Chill (7.6%):** **Small** but high-stickiness—build lo-fi/ambient products and wellness bundles where time spent per user is high despite lower volume.
- **Portfolio strategy :** Keep majority investment in Happy/Excited; use Angry/Intense for growth spikes; expand evergreen Sad/Calm and Peaceful/Chill to lift average listening time.

Spotify

# *Bivariate Analysis*

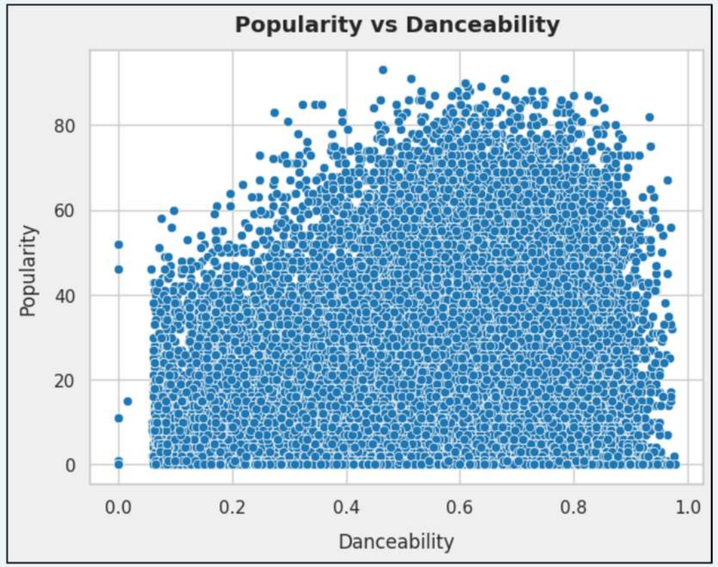
# Quantitative Analysis Using Correlation Heatmap

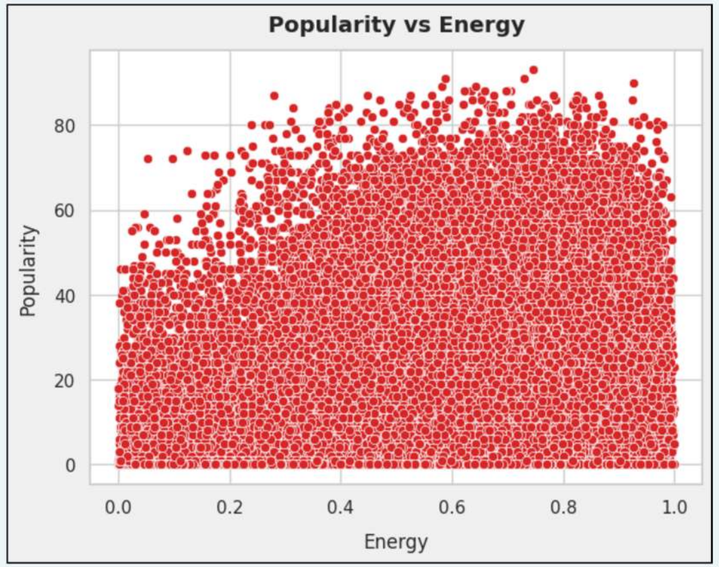


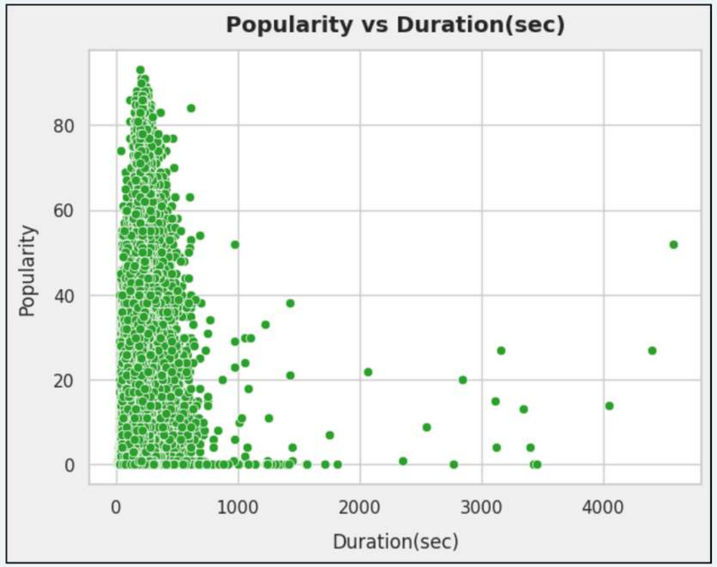
# Overview

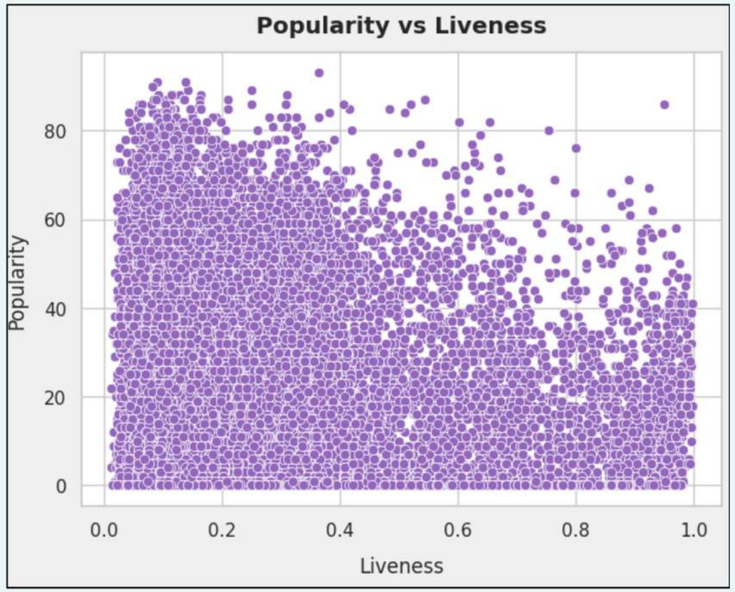
- **Energy vs loudness is strongly positive (≈0.82), so boosting production loudness reliably correlates with perceived energy; high-energy playlist candidates will also test loud in mixes.**
- The correlation between duration and popularity is essentially zero, so neither shorter nor longer songs are inherently more popular in this dataset.
- **Acousticness is strongly negative with energy (≈−0.66) and loudness (≈−0.51), indicating acoustic tracks skew softer/less energetic—position them for calm/focus products, not hype lists.**
- Danceability and popularity are nearly uncorrelated, indicating higher danceability does not systematically translate to higher popularity here.
- **Danceability links positively with energy (≈0.47) and valence (≈0.64), suggesting upbeat, danceable tracks also feel happier; use this triad to brand "feel-good" playlists.**
- Energy shows a very weak positive correlation with popularity, meaning high-energy tracks are only marginally more popular on average.
- Instrumentalness is negatively correlated with energy (≈−0.42), danceability (≈−0.45), loudness (≈−0.58), and valence (≈−0.45), implying instrumentals are quieter, calmer, and less "happy"; prioritize them for study, ambient, and sync beds.
- Loudness has a weak positive correlation with popularity, suggesting slightly louder mixes tend to perform a bit better, though the effect is small.
- **Valence has moderate positive ties with loudness (≈0.49) and energy (≈0.52), reinforcing that happier songs sound louder and more energetic; applicable for party/workout curation.**
- Acousticness has a weak negative correlation with popularity, implying more "organic/acoustic" tracks are modestly less popular than less acoustic/electronic ones on average.
- **Popularity shows weak correlations across features (highest ~0.22 with loudness), meaning hits are not driven by any single audio trait; promotion, timing, and artist brand likely dominate outcomes.**
- Tempo relationships are mild overall, with small positives to energy (≈0.17) and valence (≈0.11); tempo alone is a weak predictor—pair with danceability/energy for targeting.
- **Energy_dance_ratio correlations are near zero, signaling instability; treat this derived metric cautiously in modeling and consider winsorizing or alternative normalizations.**
- Structural tags (key, mode, time_signature, decade) have minimal correlation with popularity or mood proxies, so they are better for A&R guidance than hit prediction.
- Feature collinearity to note for modeling: energy with loudness (high), acousticness with energy/loudness (high negatives), danceability with valence (moderate-high); avoid redundant inputs or apply regularization.

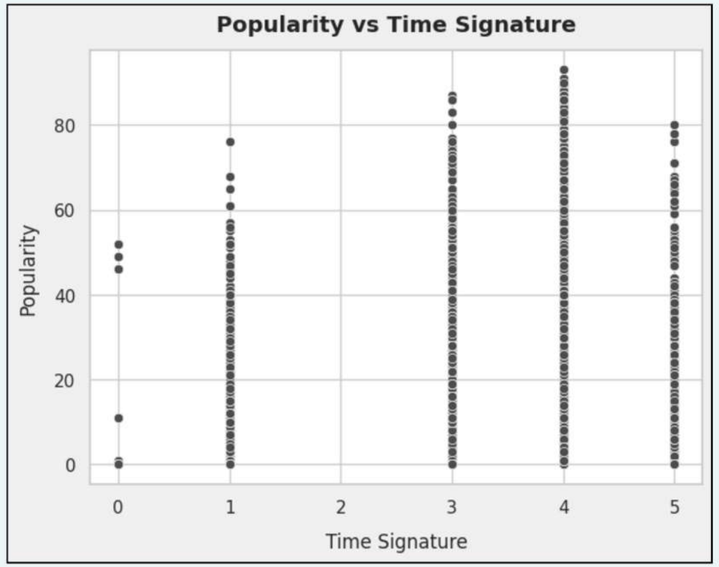
# Analysis Of Popularity

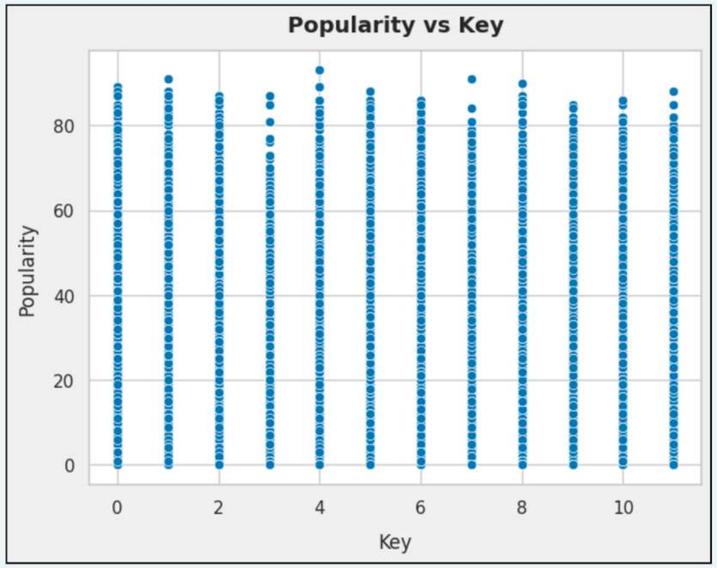


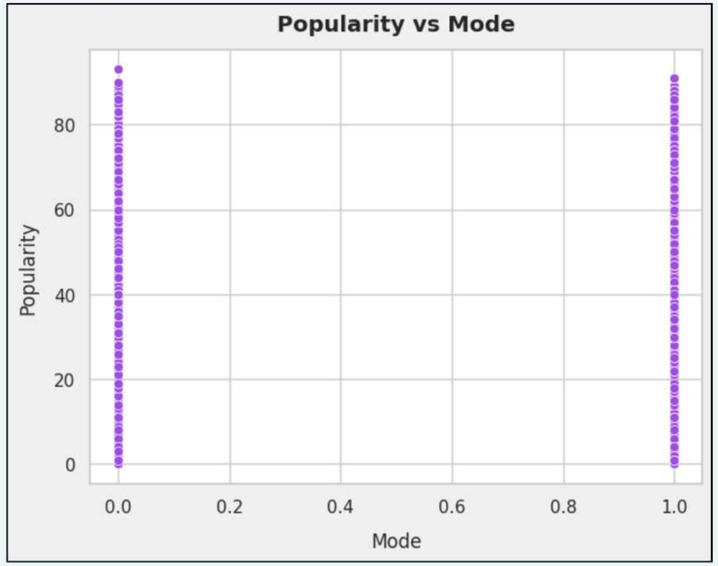



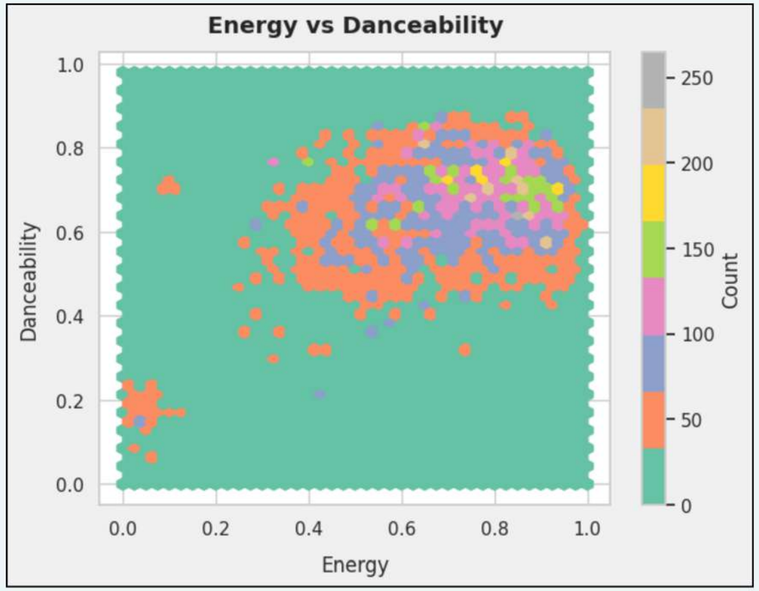
# Analysis Of Popularity

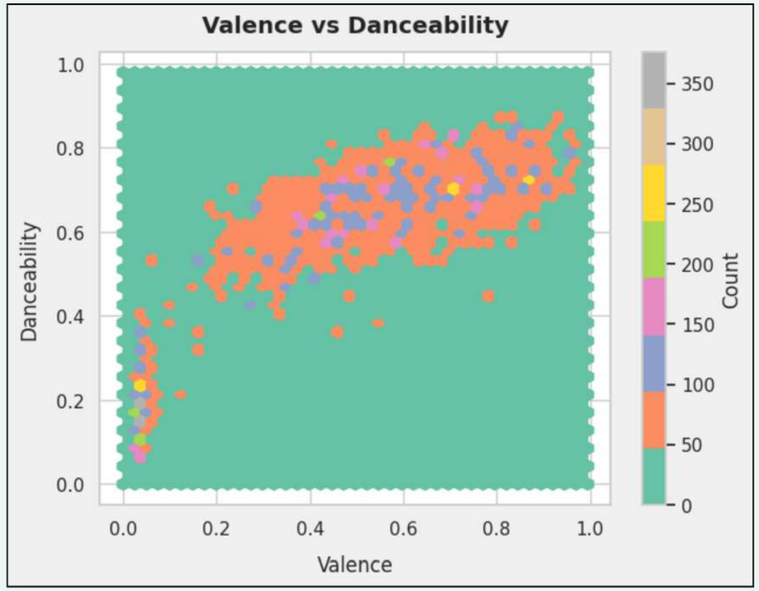


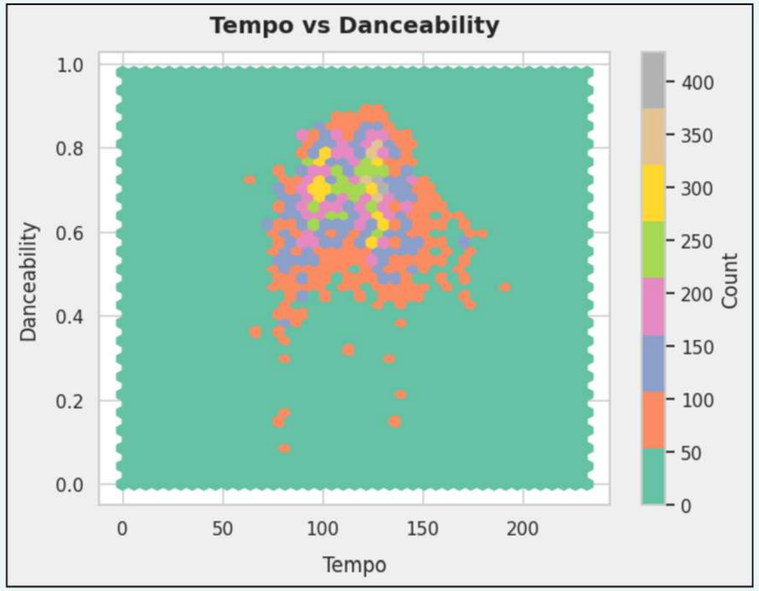

- **Popularity vs Energy :** The upper envelope slopes upward—**higher energy levels host more of the top popularity points—yet the spread is wide at all energies, confirming only a weak positive relationship; energy helps but doesn't determine hits.**
- **Popularity vs Duration :** High-popularity songs cluster around standard lengths (~2–4 minutes), while very long tracks rarely reach high popularity, indicating practical listening limits; duration beyond ~6–7 minutes correlates with lower hit likelihood.
- **Popularity vs Danceability : The densest high-popularity points sit in the mid-to-high danceability band (~0.4–0.8), but strong variance persists across the range, implying danceability is a supportive—but not decisive—factor.**
- **Liveness vs popularity :** Popular songs cluster at low-to-moderate liveness, and the density thins out as liveness increases—**live-leaning tracks tend to be less popular than studio-style recordings on average.**

- **Time signature vs popularity : 4/4 contains the largest band of higher-popularity points,** while other signatures (3, 5, etc.) also host successes but with thinner density—mainstream gains are most common in 4/4, though non-4/4 can still produce hits.
- **Key vs popularity :** Popularity is broadly spread across all keys with no clear winner, indicating **musical key does not materially drive hit potential** in this dataset.
- **Mode vs popularity :** Both minor (0) and major (1) show similar popularity spreads, suggesting mode has negligible effect on average popularity.

# Analysis Of Danceability

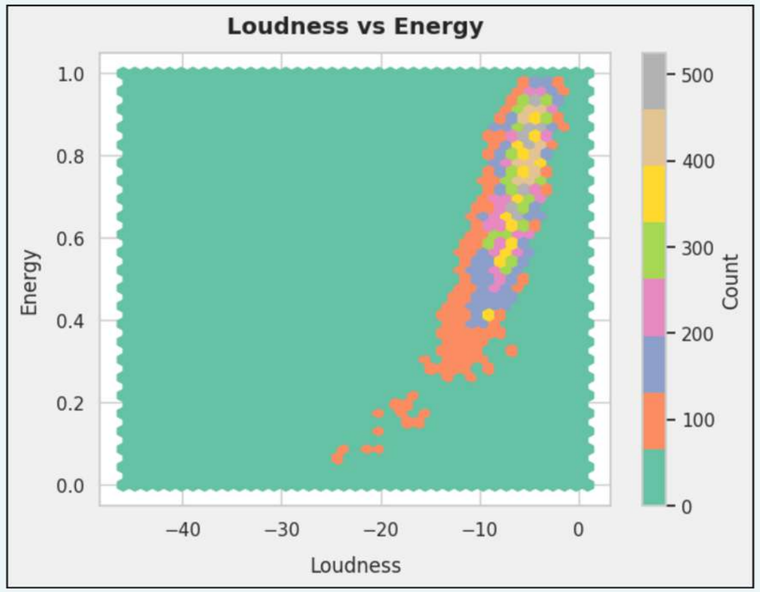


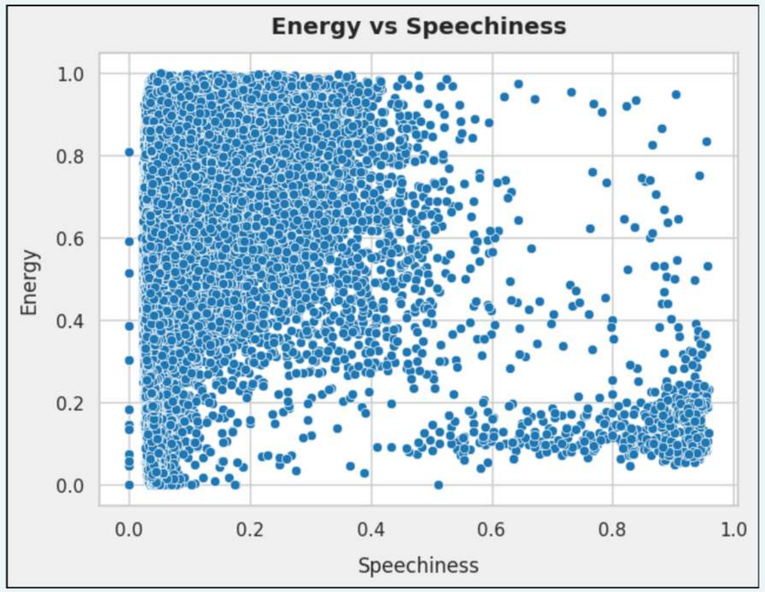



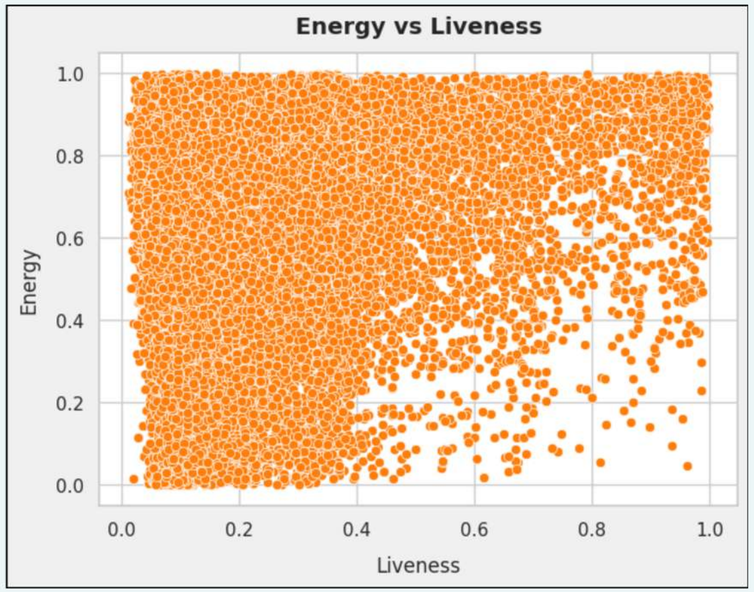



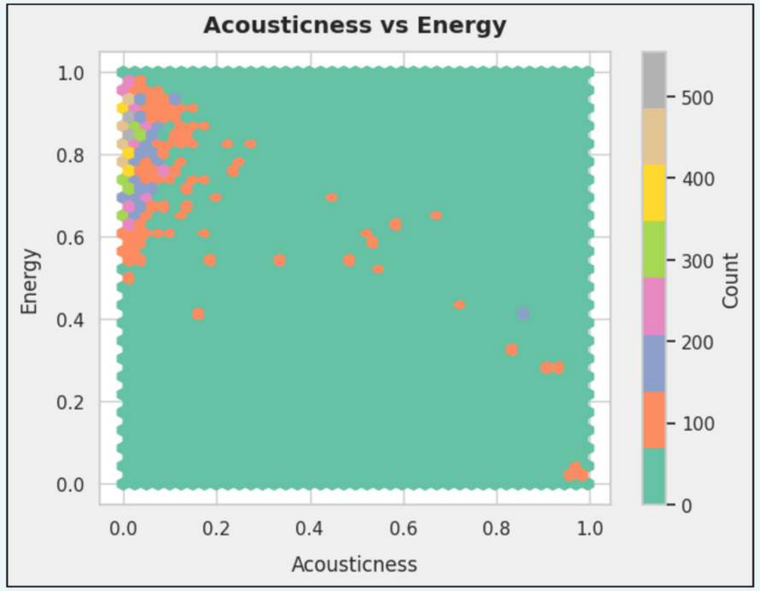

- **Tempo vs danceability :** The densest cluster of **highly danceable tracks sits around 100–130 BPM**, with the hottest spot near roughly 110–125 BPM; outside this band, danceability is less concentrated, implying mid-tempo grooves are the sweet spot for movement.
- **Valence vs danceability :** Danceability rises with valence, forming an upward trend from low valence/low danceability to high valence/high danceability; **happier songs tend to be easier to dance to, which is ideal for feel-good playlist branding.**
- **Energy vs danceability :** The tightest concentration appears at mid-to-high energy (≈0.6–0.85) paired with mid-to-high danceability (≈0.55–0.8); **extremely low energy underperforms on danceability, and ultra-high energy doesn't guarantee higher danceability,** suggesting a balanced "energetic but not maximal" profile works best.

# Analysis Of Energy



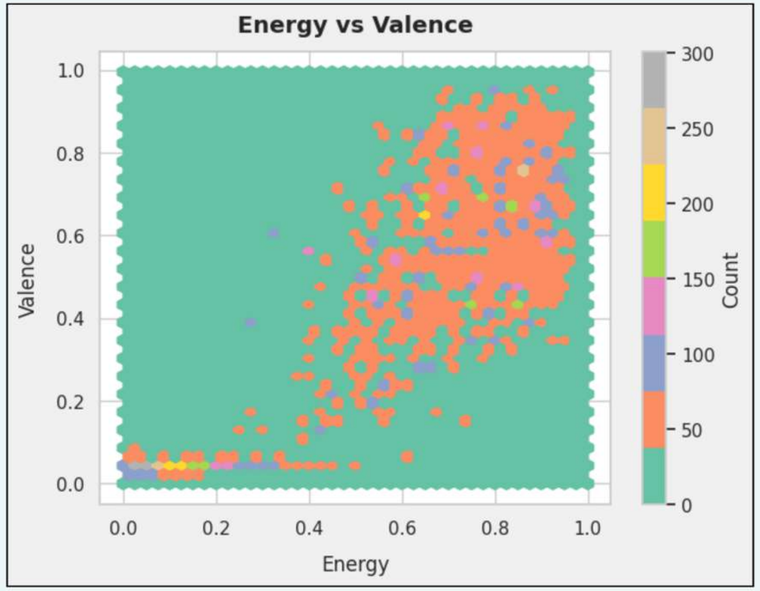



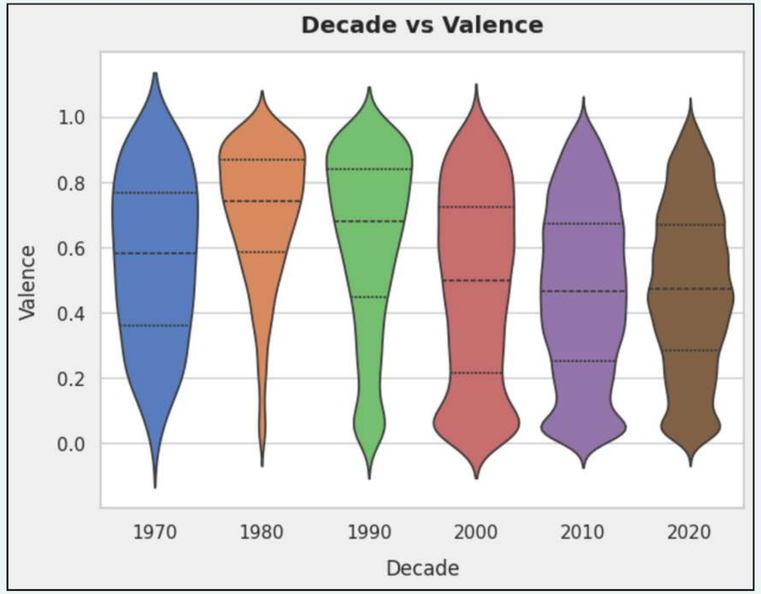



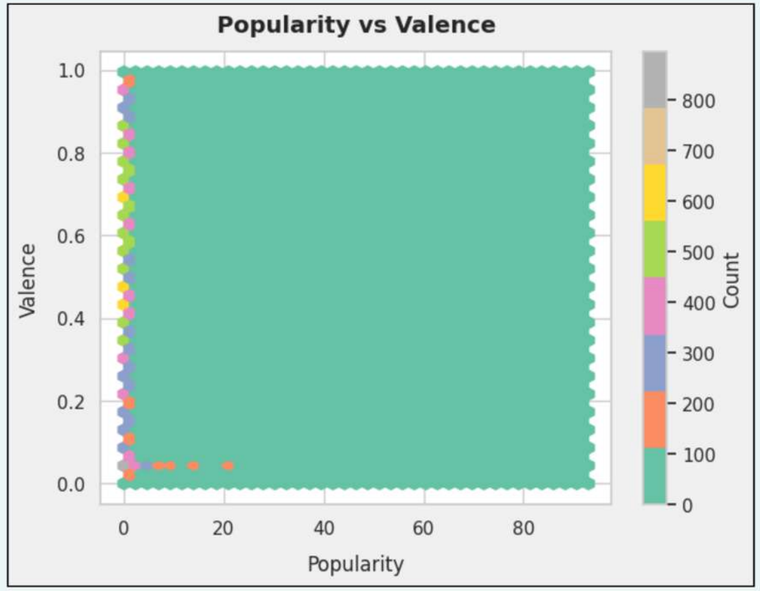




- **Energy vs liveness :** **High-energy tracks cluster at low-to-mid liveness,** and as liveness rises the density of energetic songs thins, indicating studio-style production better sustains perceived energy than live-leaning recordings.
- **Energy vs speechiness :** Most high-energy cuts sit at low speechiness, while higher speechiness bands are dominated by mid-to-low energy, suggesting spoken-word/rap-like segments generally reduce perceived energy except for niche outliers.
- **Loudness vs energy :** **There is a tight, upward diagonal—louder mixes map to higher energy almost linearly—reinforcing that mastering levels are a strong proxy for energy perception.**
- **Acousticness vs energy :** **As acousticness increases, energy drops sharply,** with the highest acoustic tracks clustering at very low energy; acoustic/organic recordings trend calm rather than intense.

# Analysis Of Valence

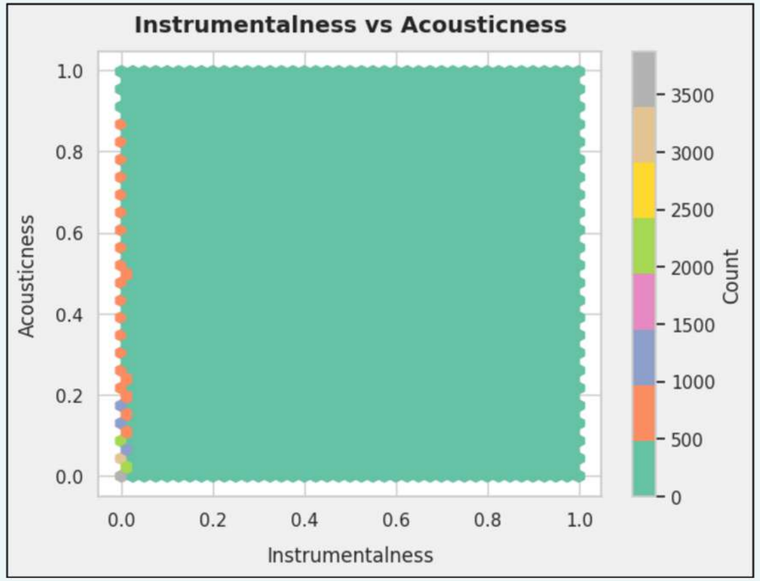


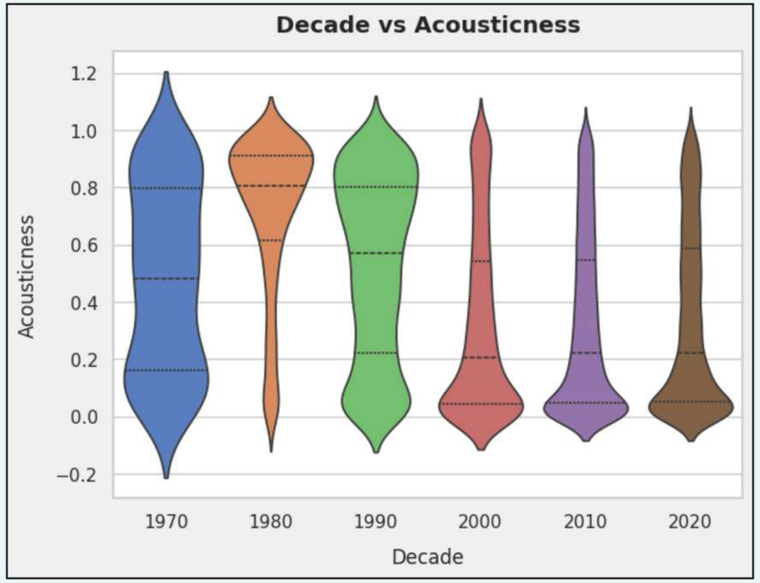



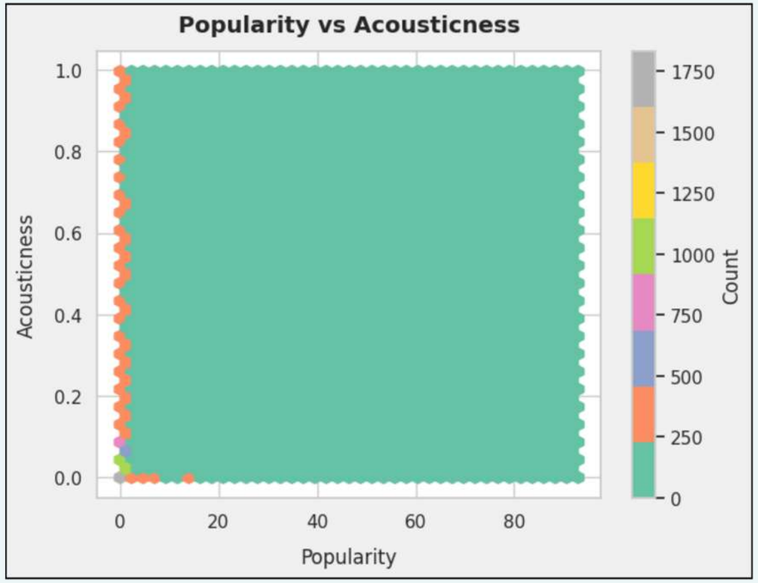




- **Energy vs valence :** There is a clear upward trend—**higher energy aligns with higher valence**—so **energetic tracks tend to feel more positive,** reinforcing the "feel-good equals energetic" programming logic.
- **Decade vs valence :** Earlier decades show wider mood spread, while recent decades center around mid valence with fewer extremes, indicating modern releases cluster toward balanced, **neither-too-happy-nor-too-somber palettes for broader appeal.**
- **Popularity vs valence :** Popularity points are broadly distributed across the full valence range with no strong tilt toward very happy or very somber; at most, there is a slight concentration around mid-to-high valence, meaning cheerful songs may have a small edge, but mood alone does not determine hits.

# Analysis Of Acousticness

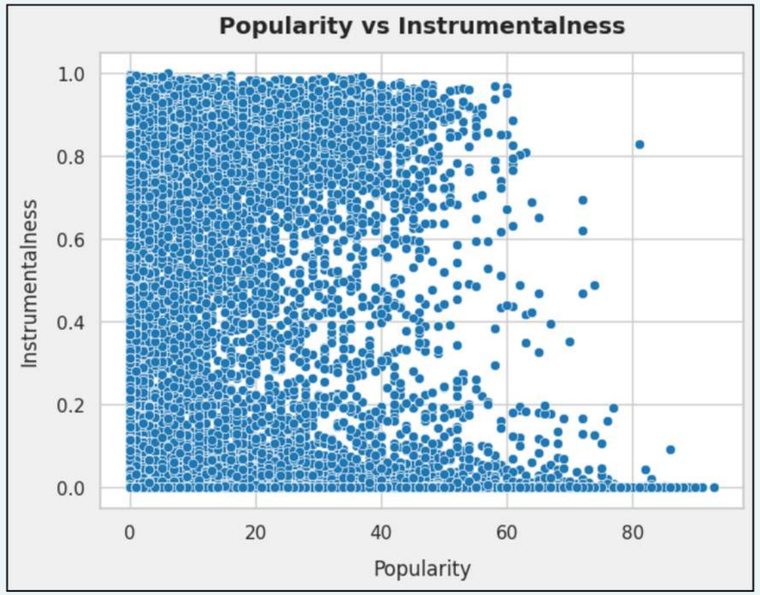


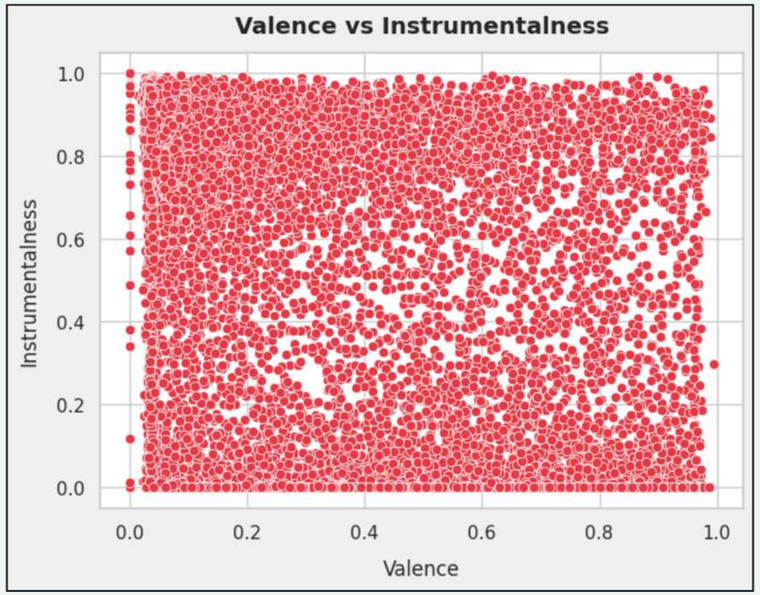



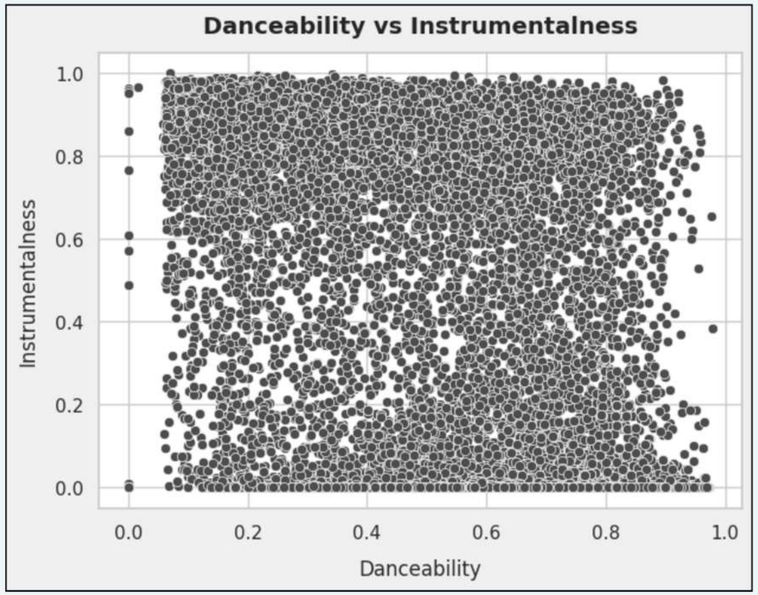




- **Instrumentalness vs acousticness :** As acousticness increases, points concentrate more toward higher values while instrumentalness remains predominantly low, **indicating most acoustic-leaning tracks still contain vocals and true instrumentals are comparatively rare.**
- **Decade vs acousticness :** Older decades show broader and higher acousticness distributions, while 2000s onward cluster lower, confirming a shift toward more produced, less acoustic sound profiles in modern releases.
- **Popularity vs acousticness :** **Popularity is spread across the full acousticness range,** but high counts sit in mid-to-low acousticness, supporting a slight **preference for produced/electronic textures over purely acoustic ones.**

# Analysis Of Instrumentalness

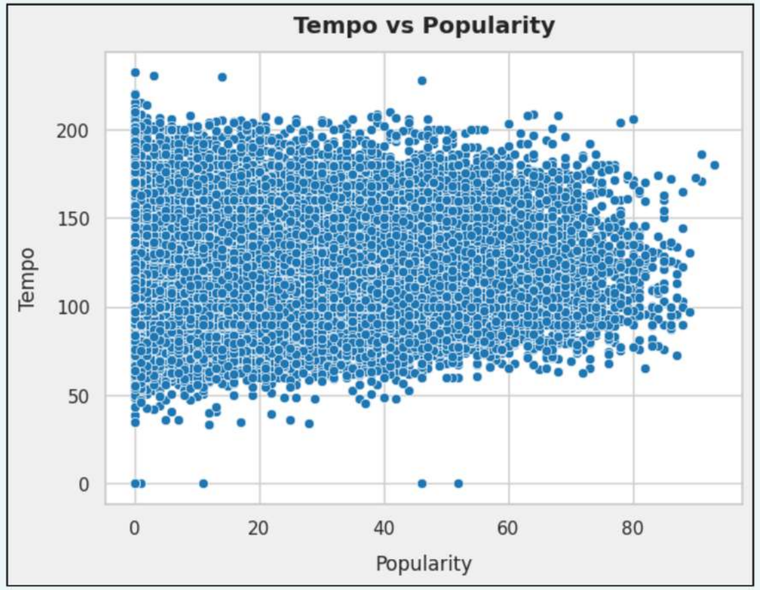


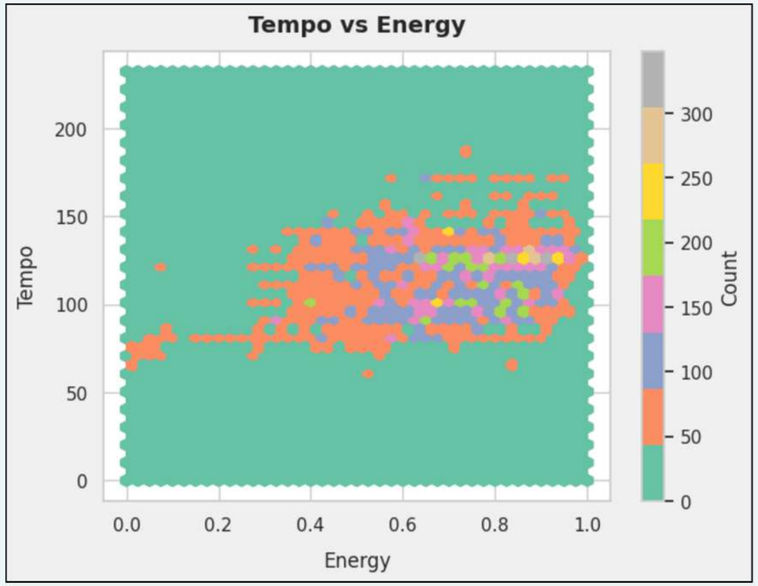




- **Popularity vs instrumentalness :** Higher popularity clusters appear toward lower instrumentalness, and the density of very popular points thins as instrumentalness rises—**tracks without vocals are generally less likely to be highly popular.**
- **Valence vs instrumentalness : No strong directional pattern;** instrumental pieces span the full mood spectrum, but **mood alone doesn't offset the popularity gap created by the absence of vocals.**
- **Danceability vs instrumentalness :** Danceability is spread across the instrumentalness range but with stronger concentration at lower instrumentalness, implying **vocals/cohesive toplines aid movement cues and mainstream dance appeal.**

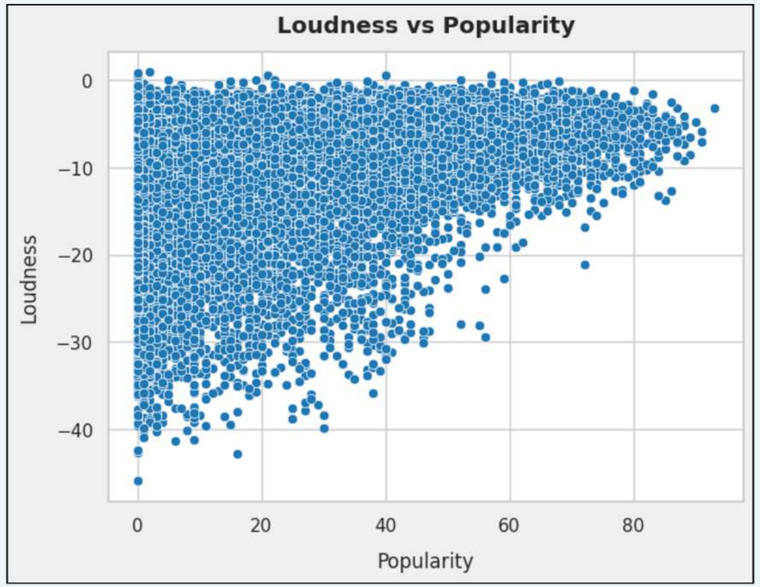
# Analysis Of Tempo

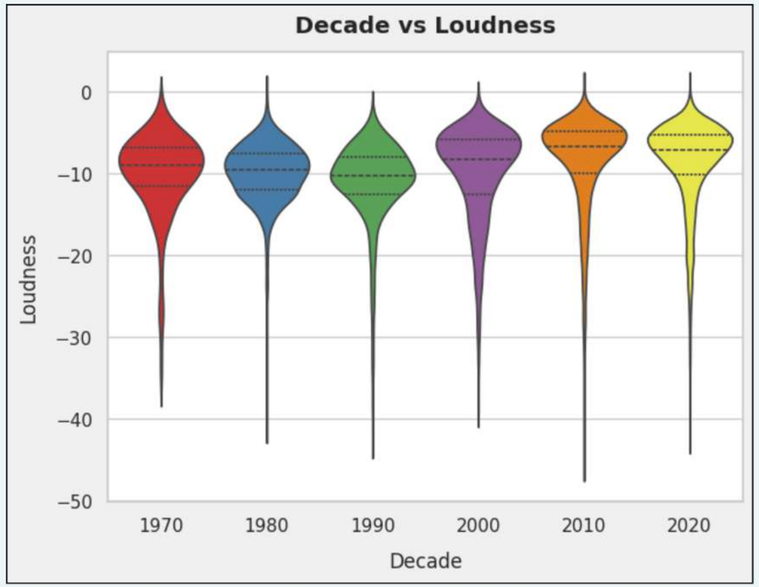



- **Tempo vs Popularity :** **Popular tracks occur across the full tempo range** with no clear tilt toward faster or slower songs; popularity clusters where the catalog density is highest (roughly 90–140 BPM), **indicating tempo alone doesn't drive hits.**
- **Tempo vs Energy :** Energy rises with tempo up to a broad mid band; the densest, most energetic region sits around 90–140 BPM paired with mid-to-high energy, showing that pairing a moderate–**fast tempo with solid intensity is the most common high-energy recipe.**

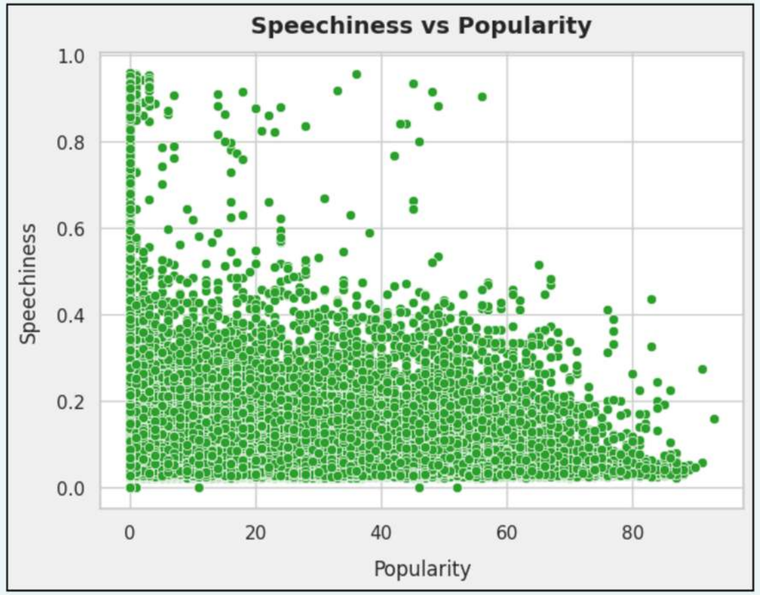
# Analysis Of Loudness

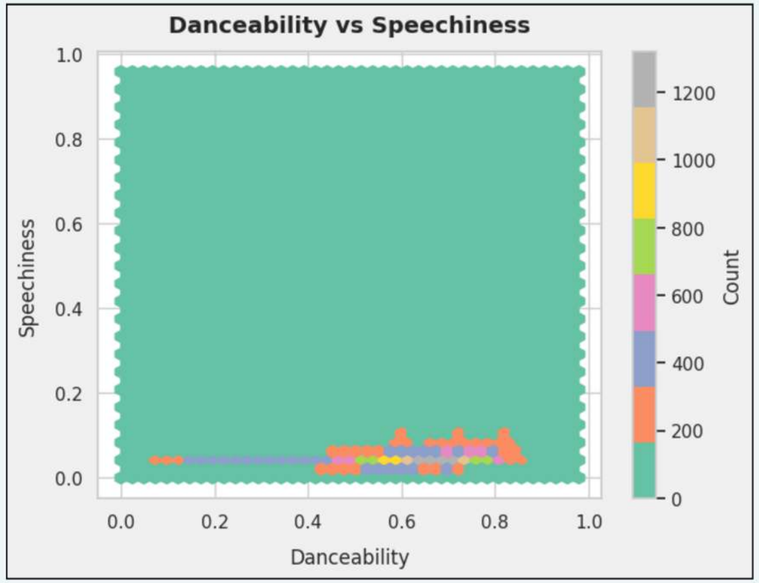



- **Loudness vs popularity :** The upper boundary trends toward louder mixes as popularity increases—hits skew slightly less negative in LUFS—but the scatter remains broad, so **loudness helps marginally without guaranteeing success.**
- **Decade vs loudness :** **Median loudness steadily rises from earlier decades to 2020s** (less negative dBFS), confirming the modern “loudness war” trend where contemporary masters are consistently louder than historical releases.

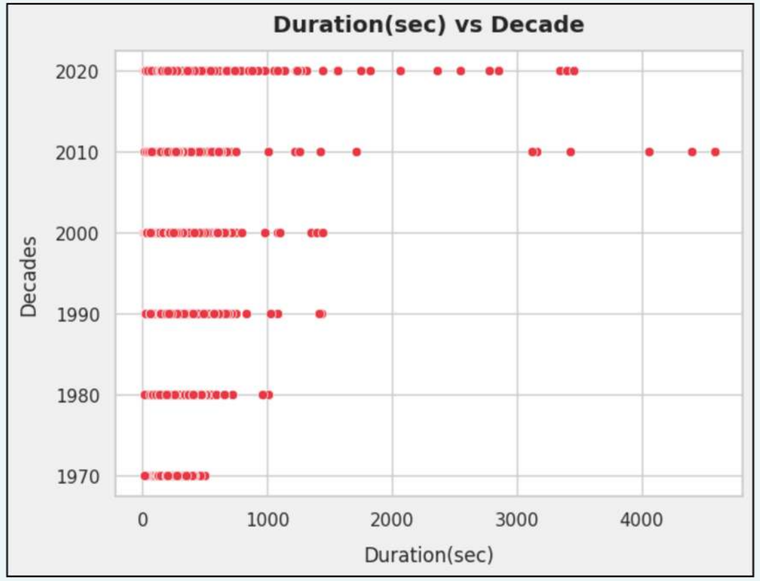
# Analysis Of Speechiness

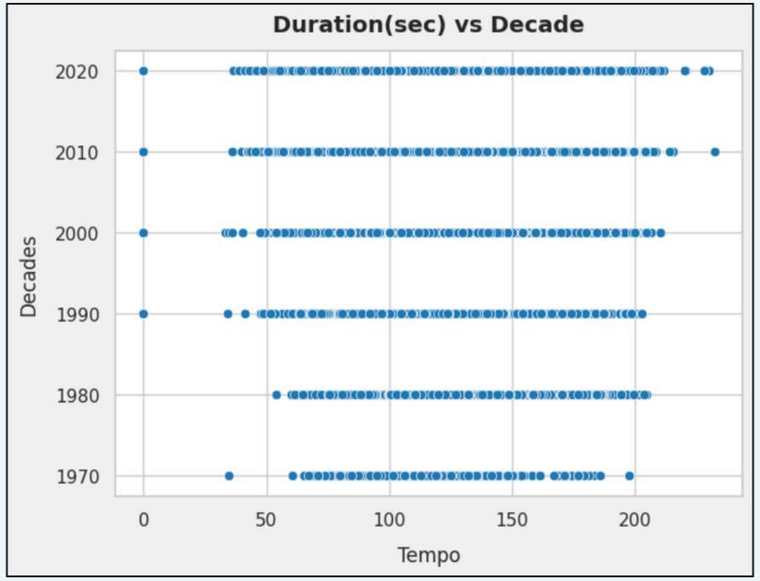


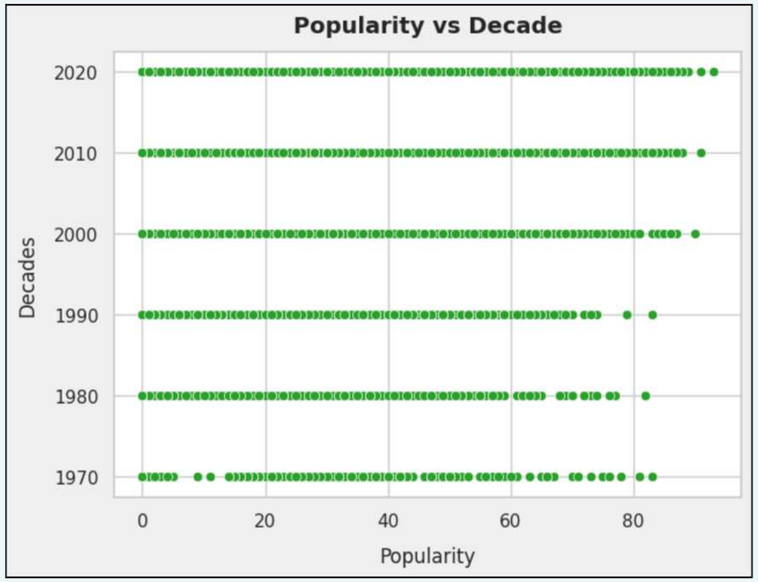




- **Speechiness vs popularity :** As popularity increases, speechiness values compress toward very low levels, indicating mainstream hits tend to contain minimal spoken-word content; **high-speechiness outliers exist but are uncommon among top tracks.**
- **Danceability vs speechiness :** The densest region of danceable songs aligns with very low speechiness, suggesting vocals structured for melody—**not spoken delivery—better support movement and radio/club appeal.**

# Analysis Of Decade

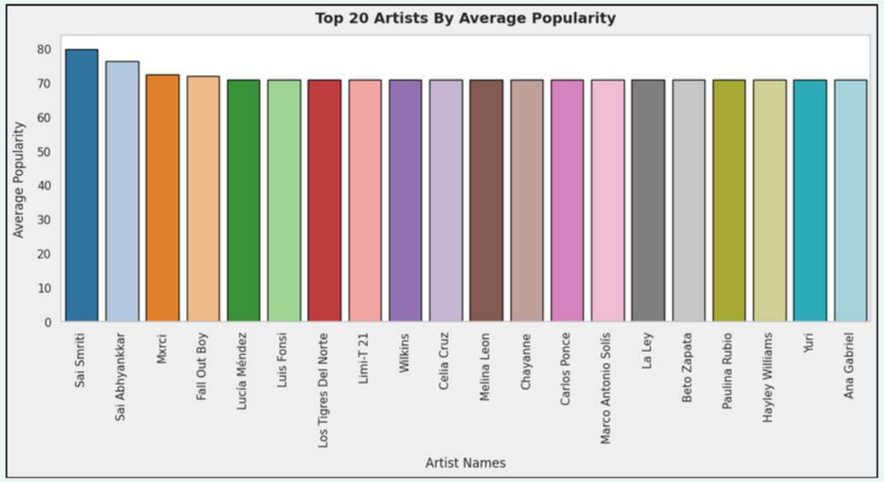


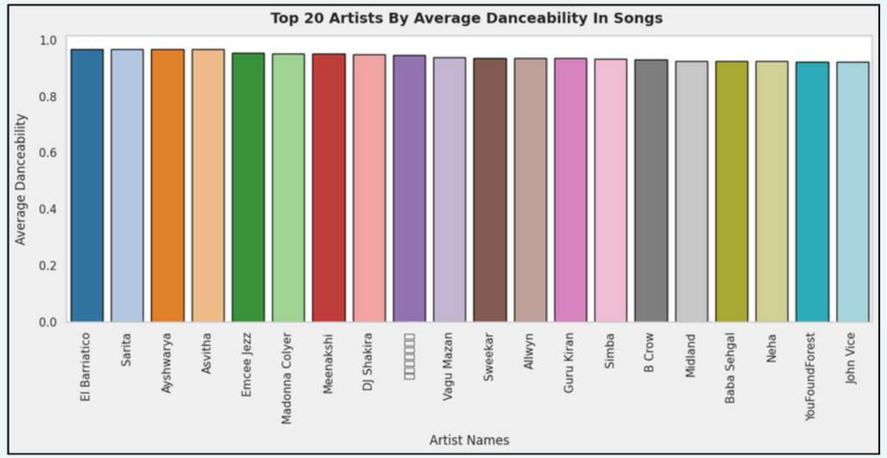



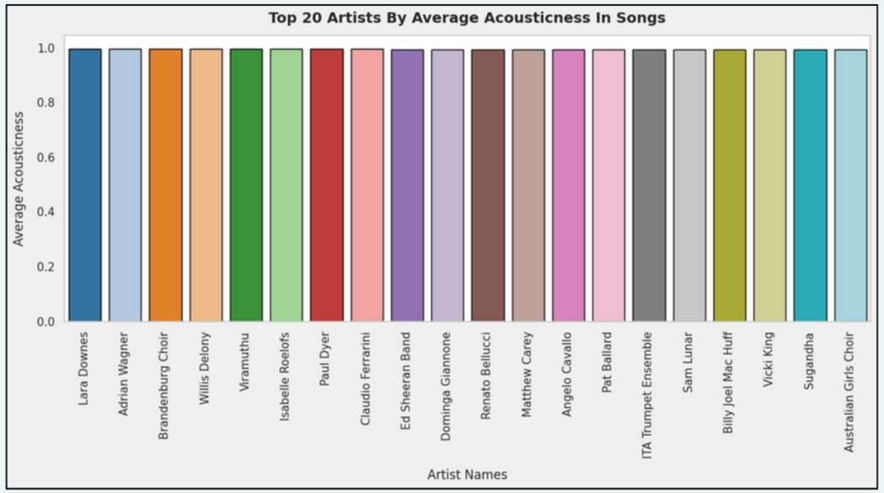



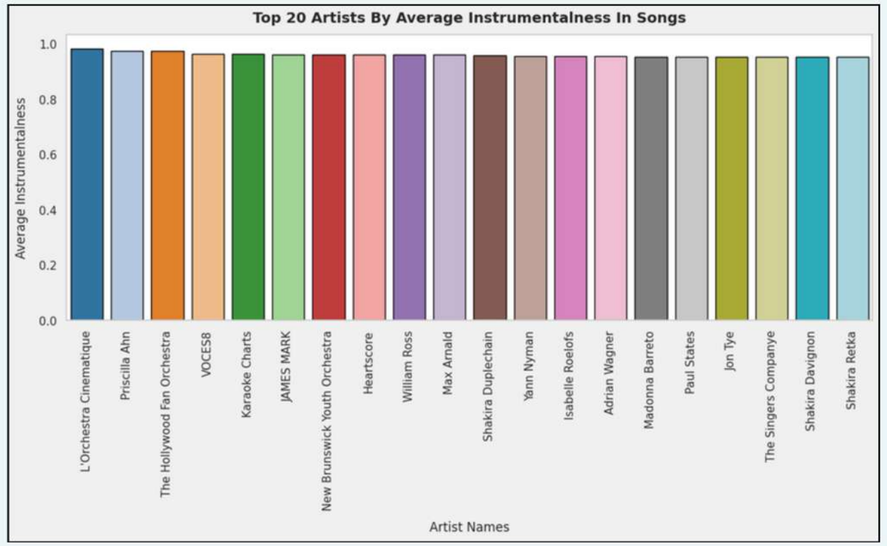

- **Duration vs decade :** Typical song lengths cluster around 2–4 minutes in every decade, with occasional long-form outliers more visible in recent years, **indicating the standard radio-friendly duration persists over time.**
- **Tempo vs decade :** Common tempos sit in the ~90–140 BPM band across decades, with only sparse extremes, showing rhythmic norms are remarkably stable historically.
- **Popularity vs decade :** High-popularity points appear in every decade, but density is greatest in recent decades due to larger catalog volume, **implying discovery advantage rather than an inherent recency popularity boost.**

# Artist Level Analysis

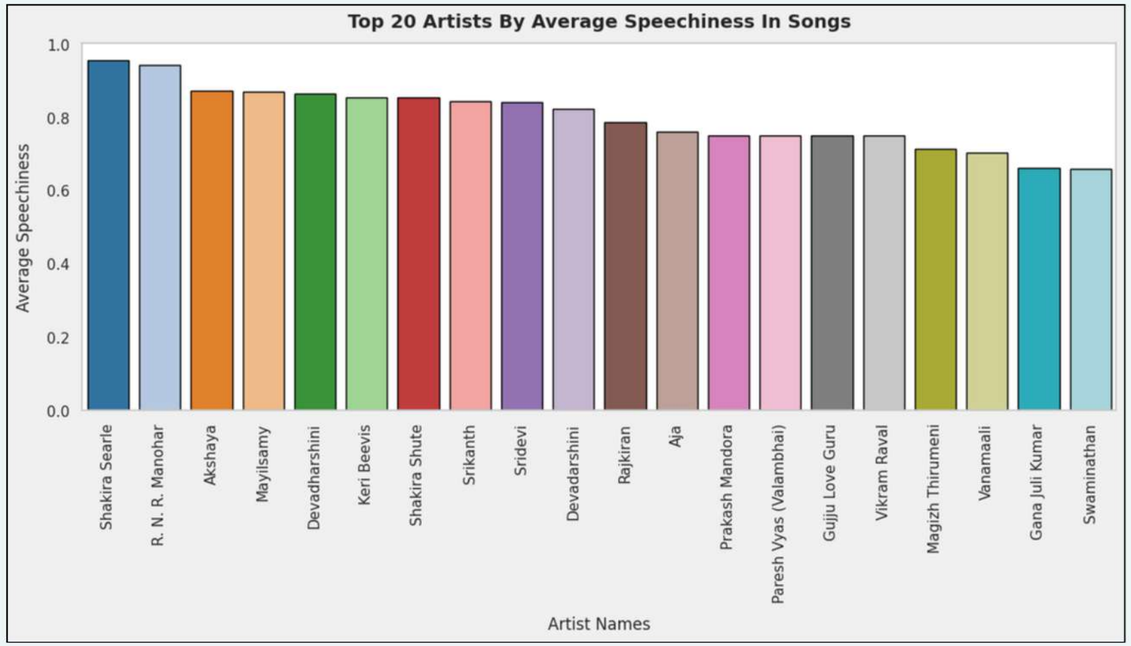


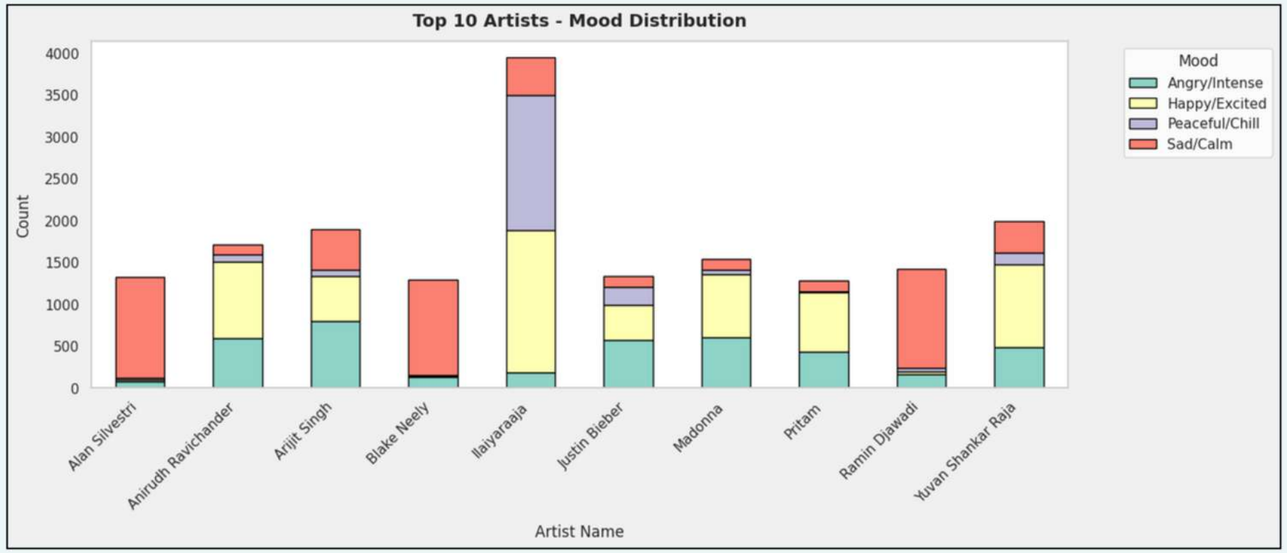
# Artist Level Analysis

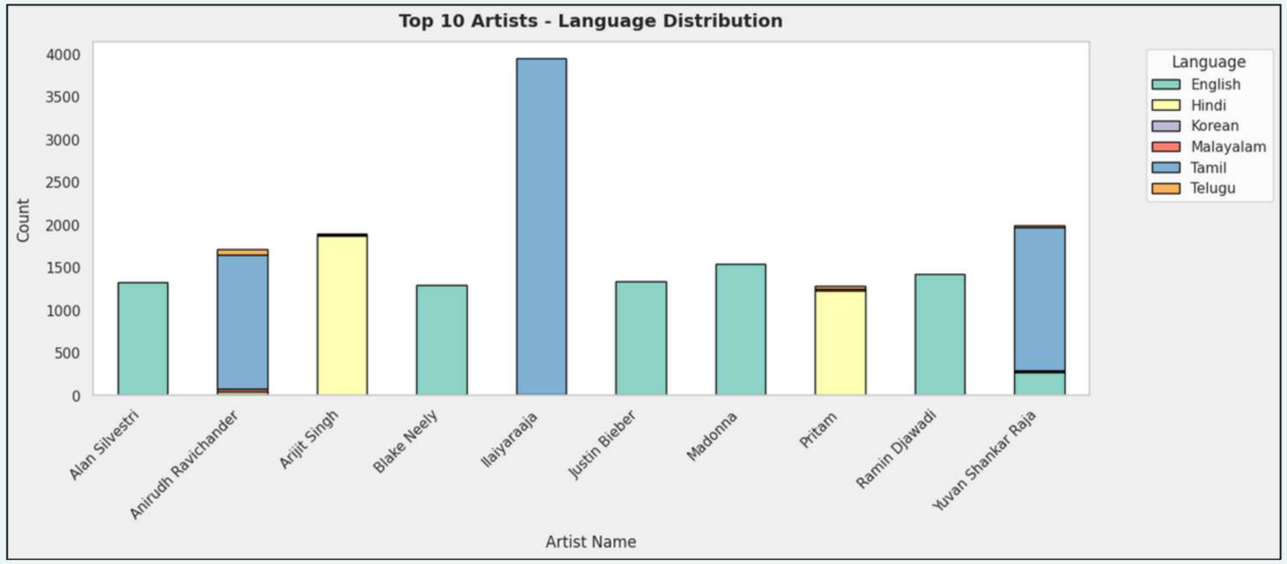



- **Top 20 by average popularity :** The leaders cluster tightly in the low-70s to ~80 range, indicating consistent audience pull rather than one-off spikes; prioritize these artists for headline placements, re-releases, and cross-promos to compound dependable engagement.
- **Top 20 by average danceability :** Nearly all sit above 0.92, making them plug-and-play for party/workout/club playlists; pair with mid-to-high energy tracks and mid-tempo grooves to create reliable high-movement sets and ad-friendly formats.
- **Top 20 by average acousticness :** Values ≈0.98–1.00 show strongly organic recordings; package for unplugged/coffeehouse/editorial storytelling lanes and live-session content where authenticity sells over loudness.

- **Top 20 by average instrumentalness :** Scores near 0.95–0.99 confirm these catalogs are overwhelmingly instrumental; position them for focus, ambient, and sync products where vocals are not required, and bundle into long-play experiences to maximize time-spent.
- **The top 20 by average speechiness :** Here artists have extremely high speechiness (≈0.7–1.0), indicating rap/spoken-word–heavy catalogs that are ideal for hip-hop and lyrical storytelling playlists, brand cyphers, and rhythm-driven campaigns.

# Artist Level Analysis

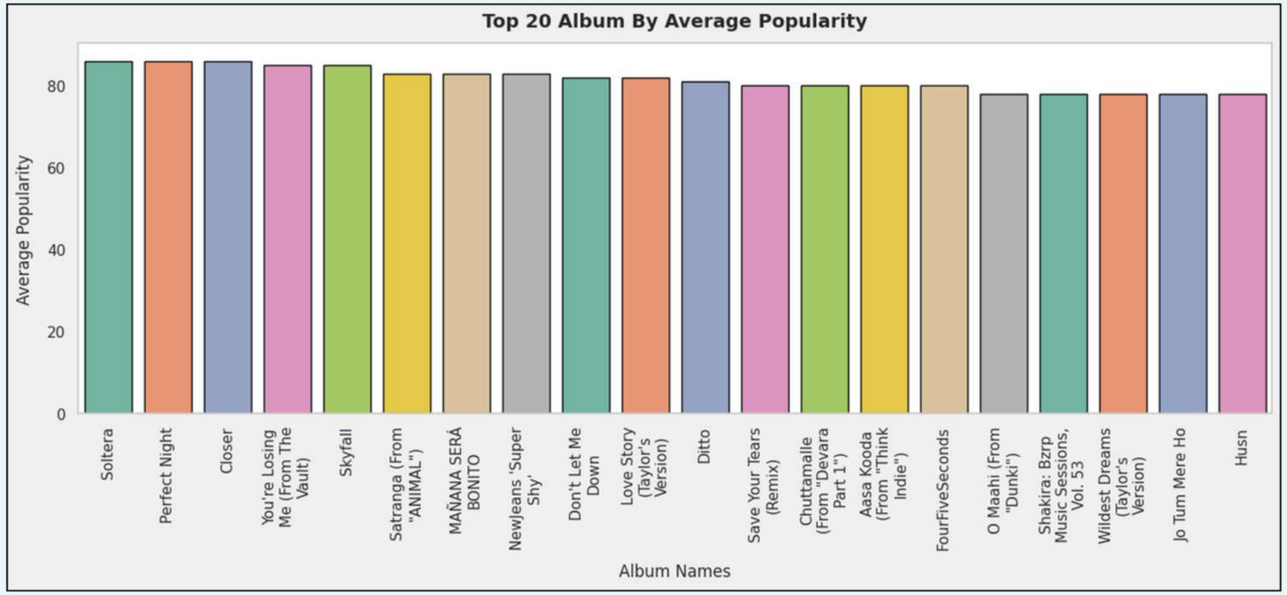


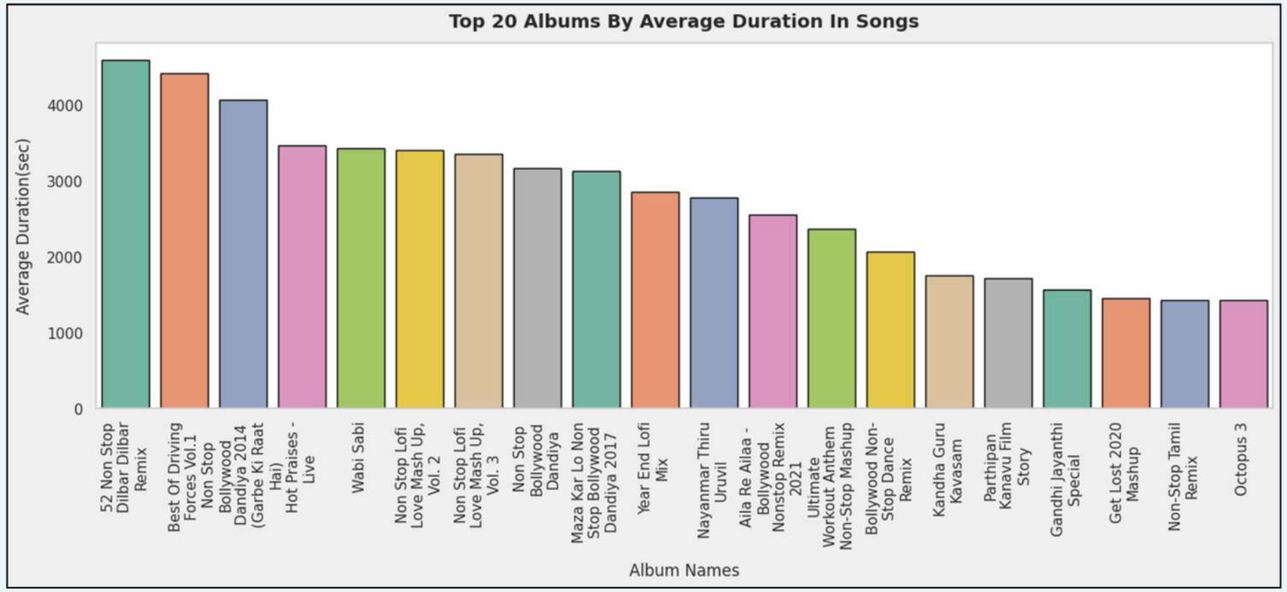




- **Mood mix across top artists :** Film composers like Ilaiyaraaja and Ramin Djawadi show large Peaceful/Chill and Sad/Calm shares alongside Happy/Excited, reflecting score diversity, while pop acts (Justin Bieber, Madonna) lean Happy/Excited and Angry/Intense; Arijit and Pritam balance Angry/Intense with Happy/Excited for mainstream Bollywood curation.

- **Language mix across top artists :** The distribution is highly segmented by market—**Ilaiyaraaja and Yuvan Shankar Raja skew heavily Tamil, Arijit Singh and Pritam are predominantly Hindi, and Alan Silvestri, Blake Neely, Justin Bieber, Madonna, and Ramin Djawadi are English**-led—so editorial should localize by artist rather than assuming cross-language breadth.

# Album Level Analysis

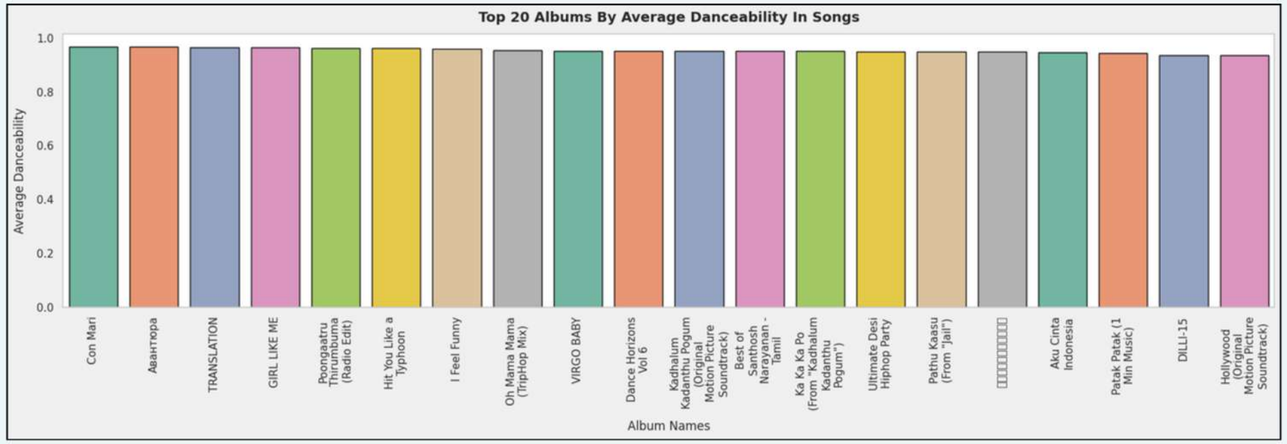


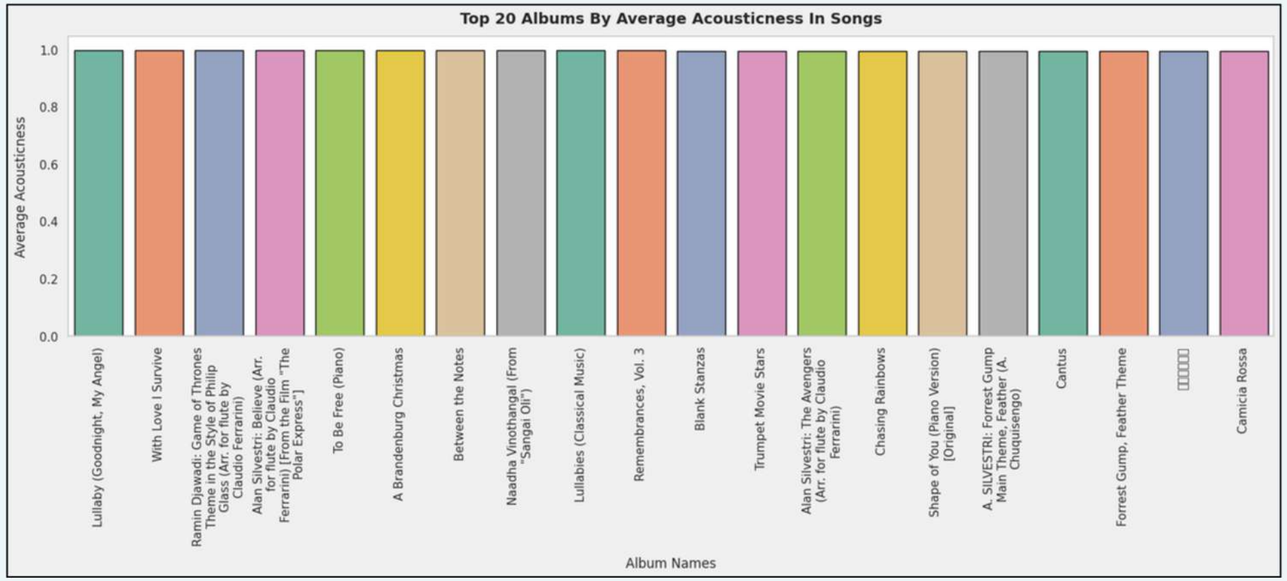

- **Top albums by average popularity :** Scores concentrate in the low-80s to high-80s with a blend of English, Hindi, and K-pop titles—strong cross-market appeal for featured carousels and new-user onboarding playlists.

- **Top albums by duration :** The leaders are long, non-stop mixes and mashups (≈2,400–4,600 seconds), **ideal for continuous-play formats** that maximize time-spent and ad inventory.

# Album Level Analysis

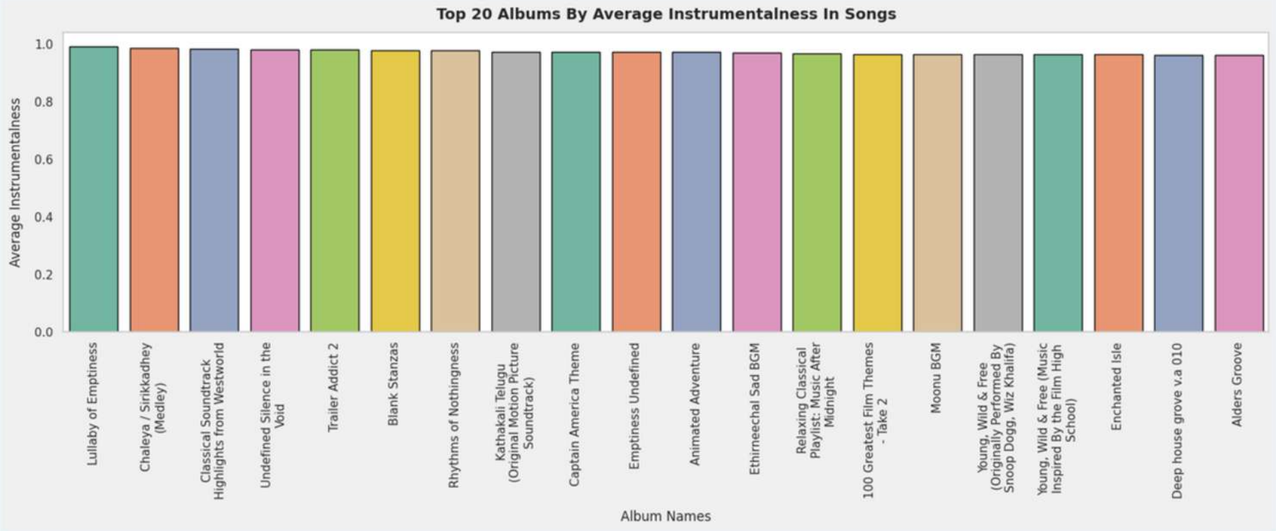


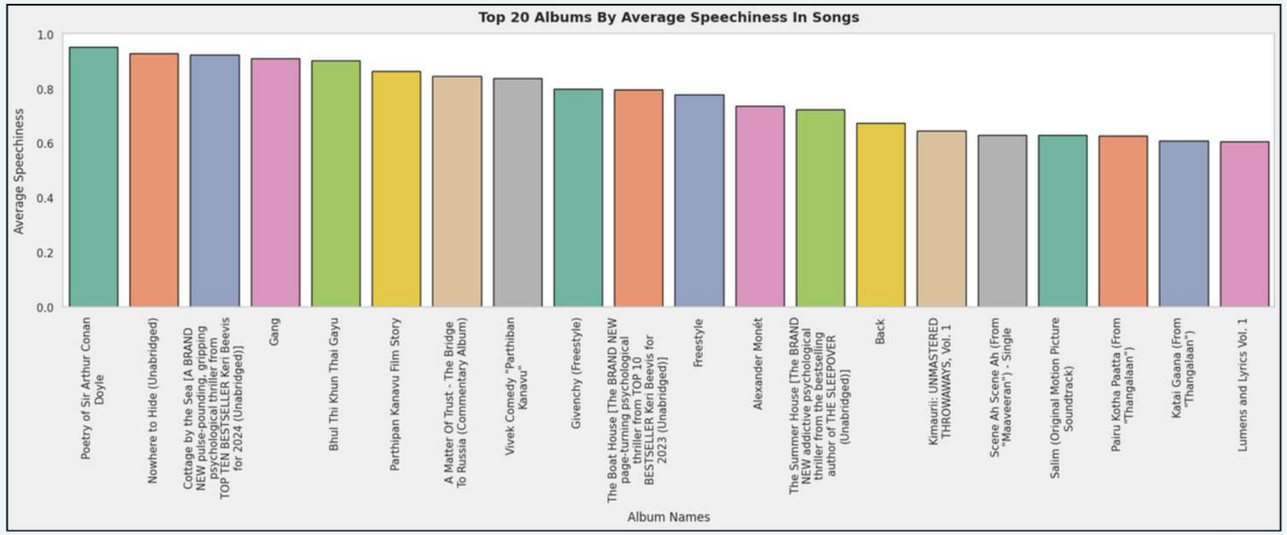

- **Top albums by danceability :** Values cluster around 0.94–0.97, mixing global pop, hip-hop, and Tamil film tracks—**ready-made sources for party/workout playlists** and short-form-friendly edits.




- **Top albums by acousticness :** Nearly all entries sit at ~0.98–1.00 acousticness, dominated by piano, orchestral, lullaby, and film-score arrangements—**prime candidates for focus, study,** and “unplugged” editorial shelves.

# Album Level Analysis

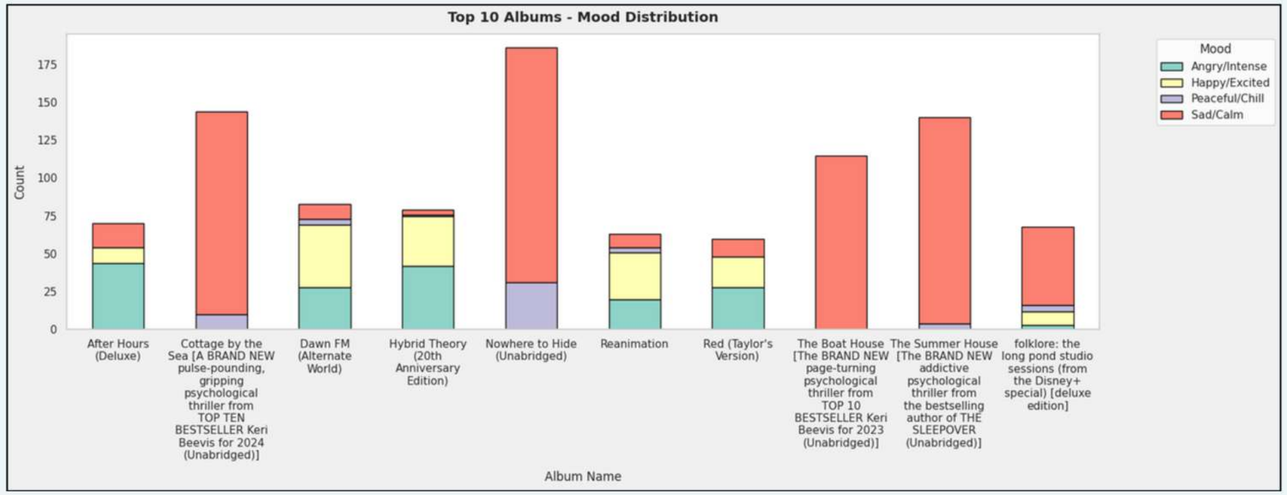


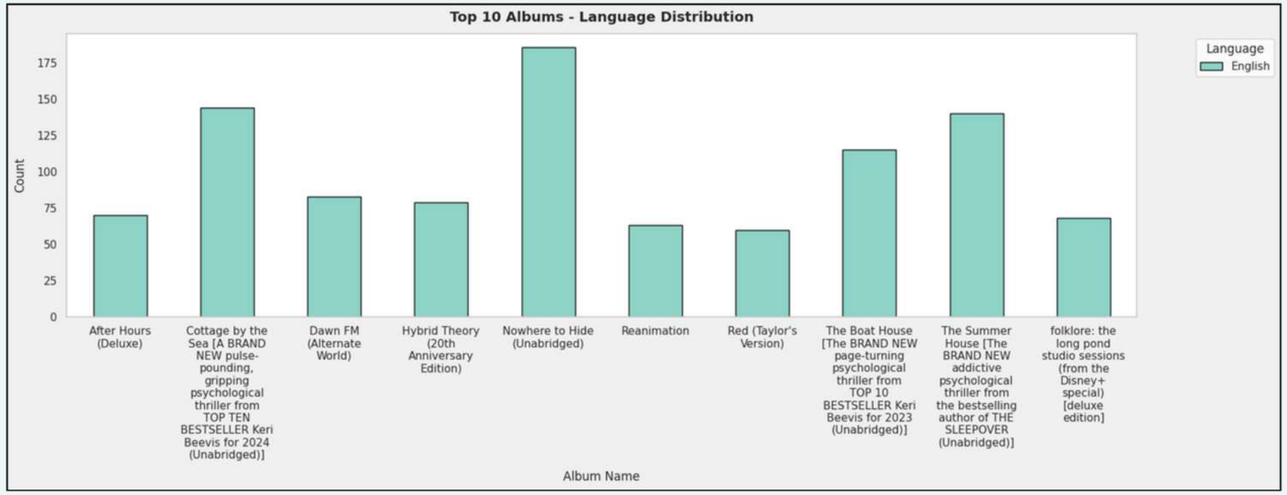

- **Top albums by instrumentalness :** Values cluster near 0.95–0.99 across film scores, classical sets, and ambient/electronic instrumentals—**perfect for focus, cinematic, and sync-friendly background playlists.**

- **Top albums by speechiness :** This list skews to **audiobooks, commentary, and freestyle/rap** projects with speechiness ~0.7–0.97, ideal for spoken-word and lyrical storytelling shelves rather than music-only playlists.

# Album Level Analysis

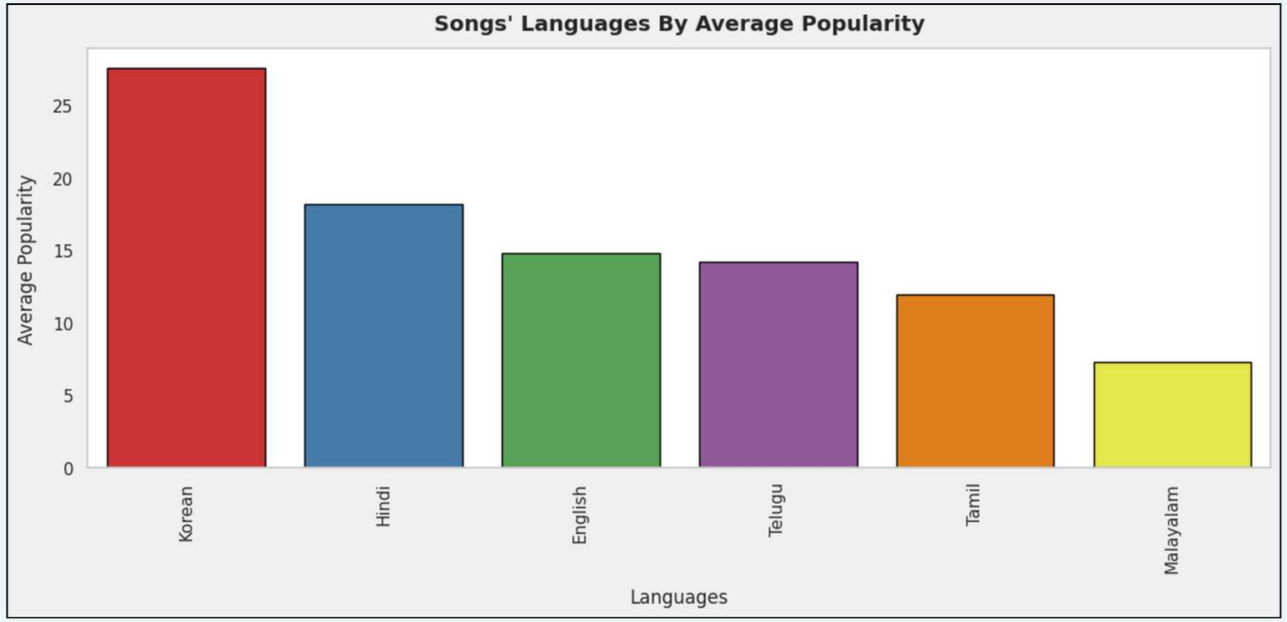



- **Mood distribution across top albums :** Several titles are dominated by Sad/Calm cuts, especially audiobook-style entries and reflective sessions, while pop/alt albums mix Happy/Excited with some Angry/Intense; only a few show meaningful Peaceful/Chill segments, **so mood targeting should differ by album intent.**

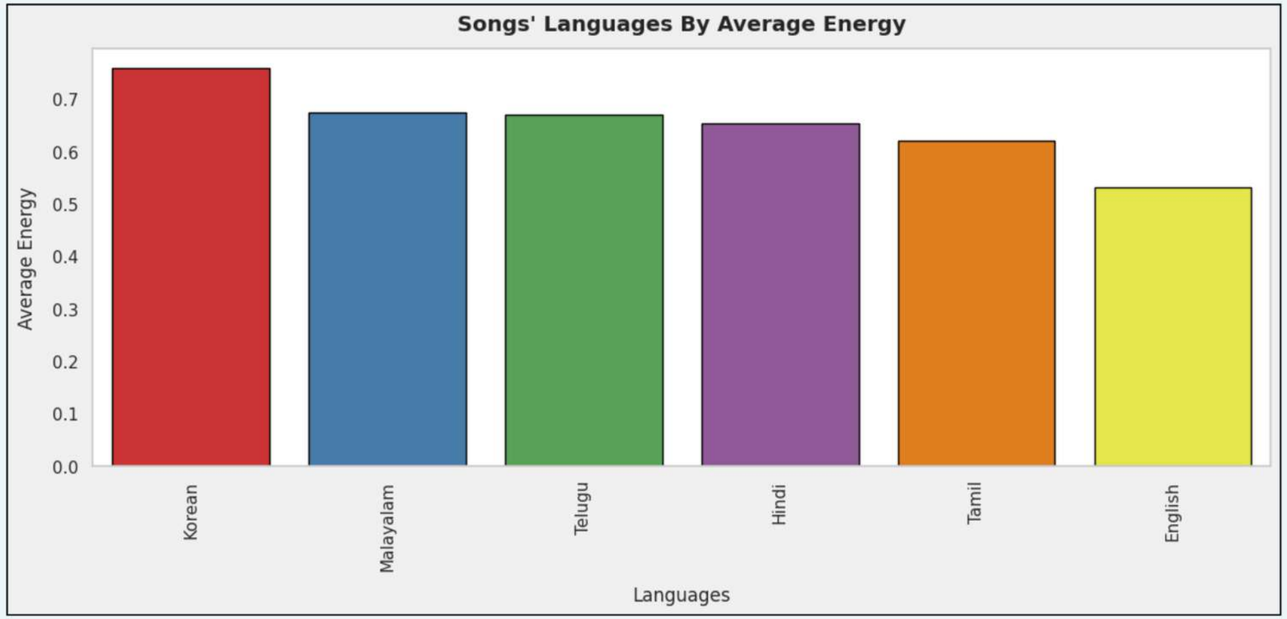



- **Language distribution across top albums :** The set is entirely English, indicating that this **"top albums" slice is English-led**; localization needs will center on mood/context rather than language switching for these specific albums.

# Language Wise Analysis

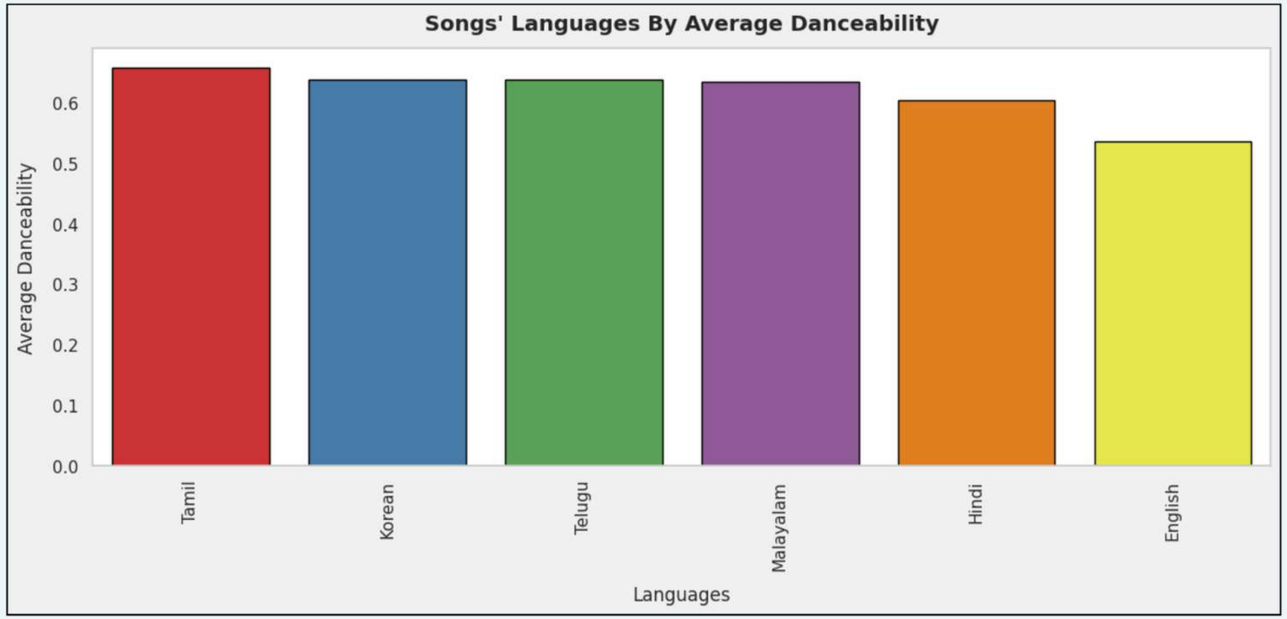


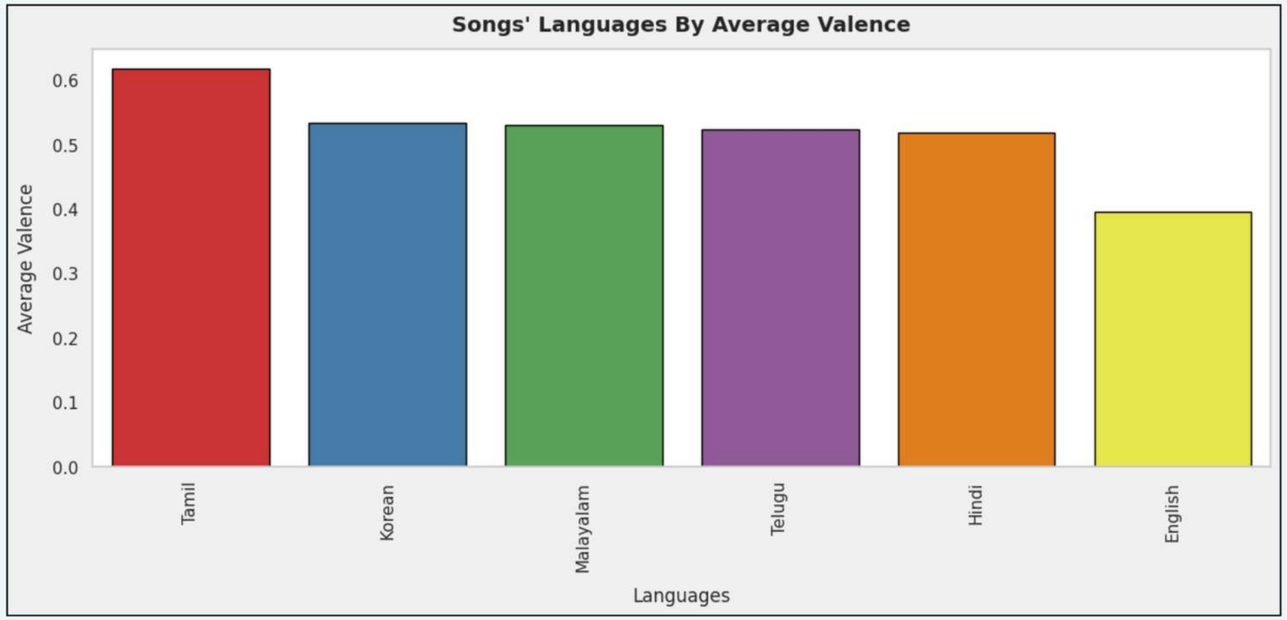

- **Popularity by language :** Korean leads by a clear margin, followed by Hindi and English, with Telugu, Tamil, and Malayalam trailing—indicating current audience heat is strongest for Korean releases.




- **Energy by language :** Korean tracks have the highest average energy, followed by Malayalam and Telugu, then Hindi and Tamil, with English the lowest—so high-octane, hype-ready selections are most abundant in Korean catalogs, while English skews comparatively mellow.

# Language Wise Analysis

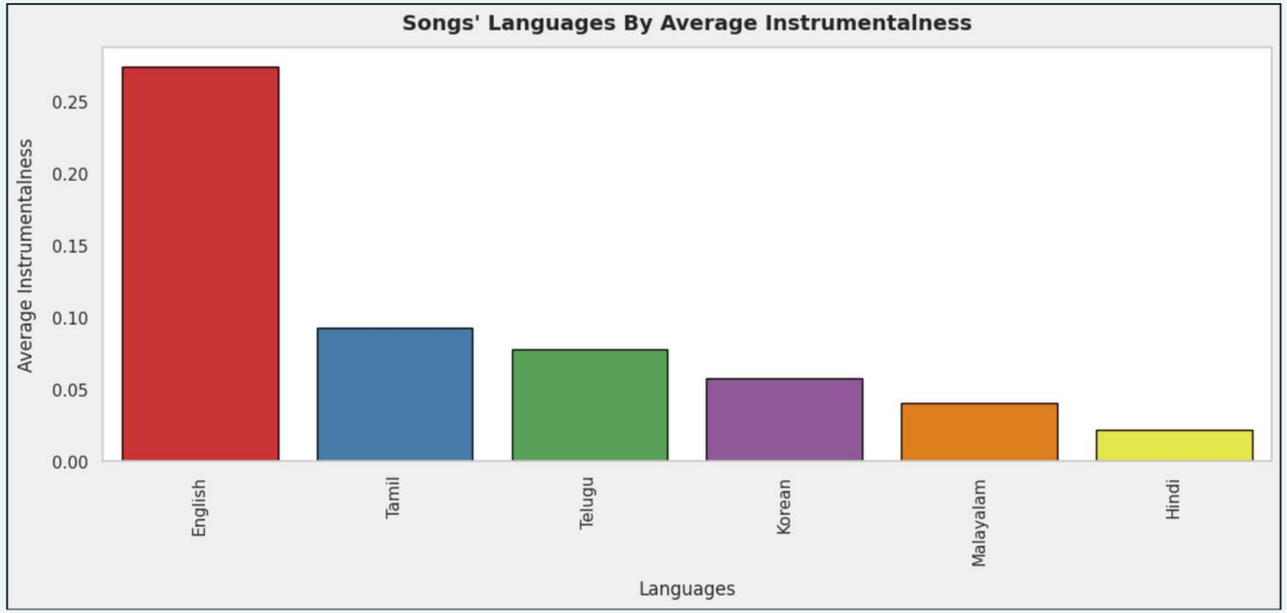


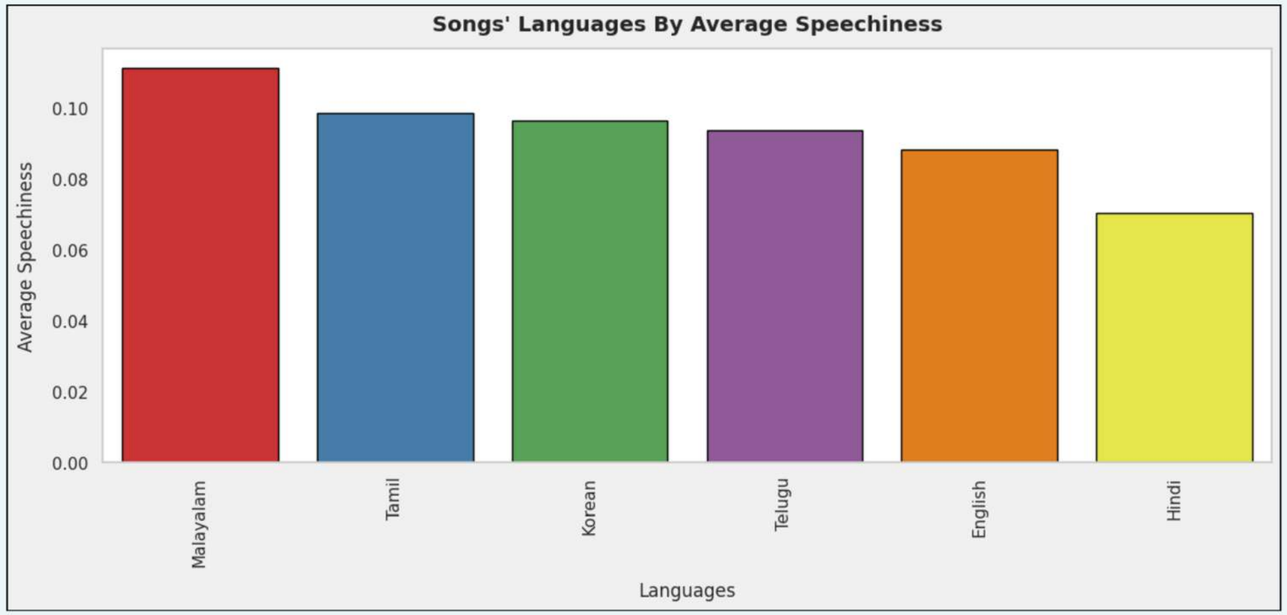




- **Danceability by language :** High danceability maps to “party” songs; Tamil ranks highest, followed closely by Korean, Telugu, and Malayalam, while English is lowest.

- **Valence by language :** Tamil shows the brightest mood profile, while Korean, Malayalam, Telugu, and Hindi sit mid; English trends lower, suggesting Tamil aligns best with “feel-good” positioning.

# Language Wise Analysis

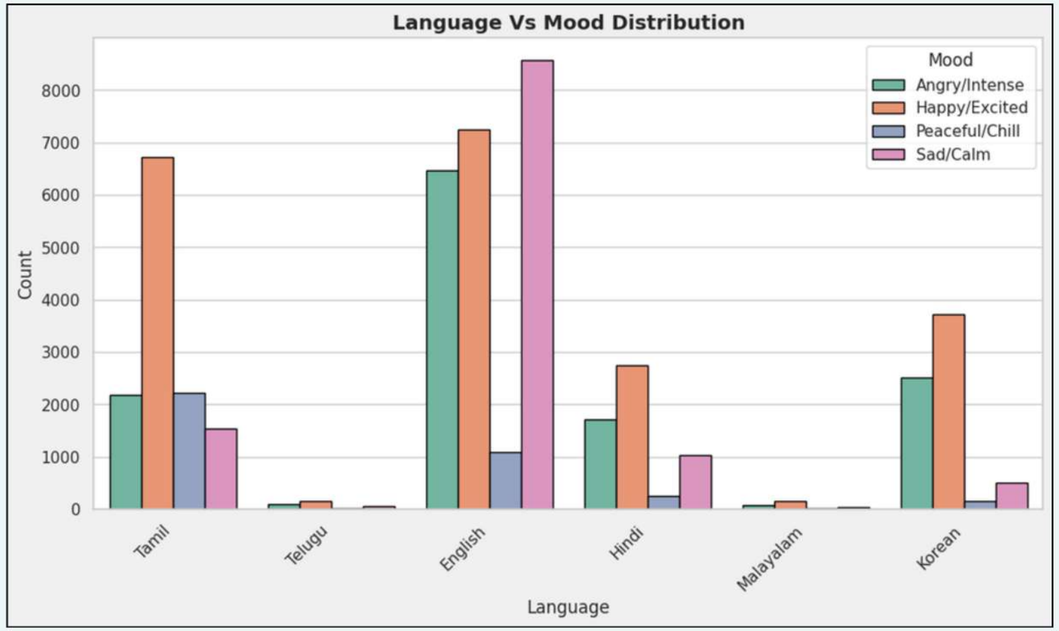


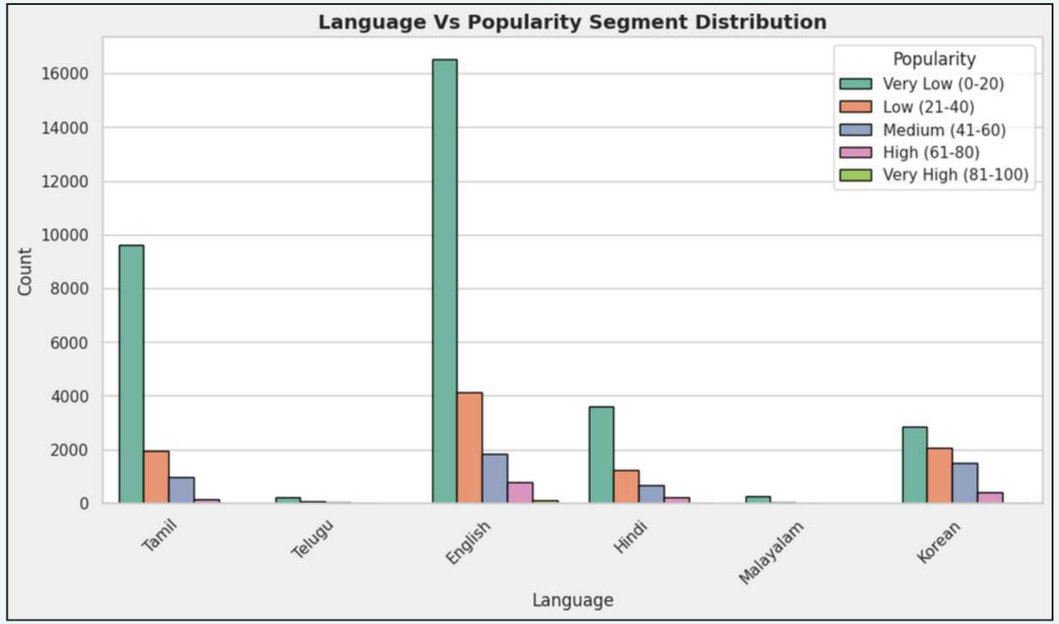




- **Instrumentalness by language :** English shows the highest average instrumentalness by a wide margin, indicating more instrumental or vocal-light tracks in English catalogs; **Hindi is the lowest,** with Tamil/Telugu/Korean/Malayalam clustered at modest levels.

- **Speechiness by language :** Malayalam tops average speechiness, with Tamil and Korean close behind, then Telugu and English, while Hindi is lowest—**suggesting rap/spoken-word prevalence is highest in Malayalam and Tamil, and least common in Hindi within this dataset.**

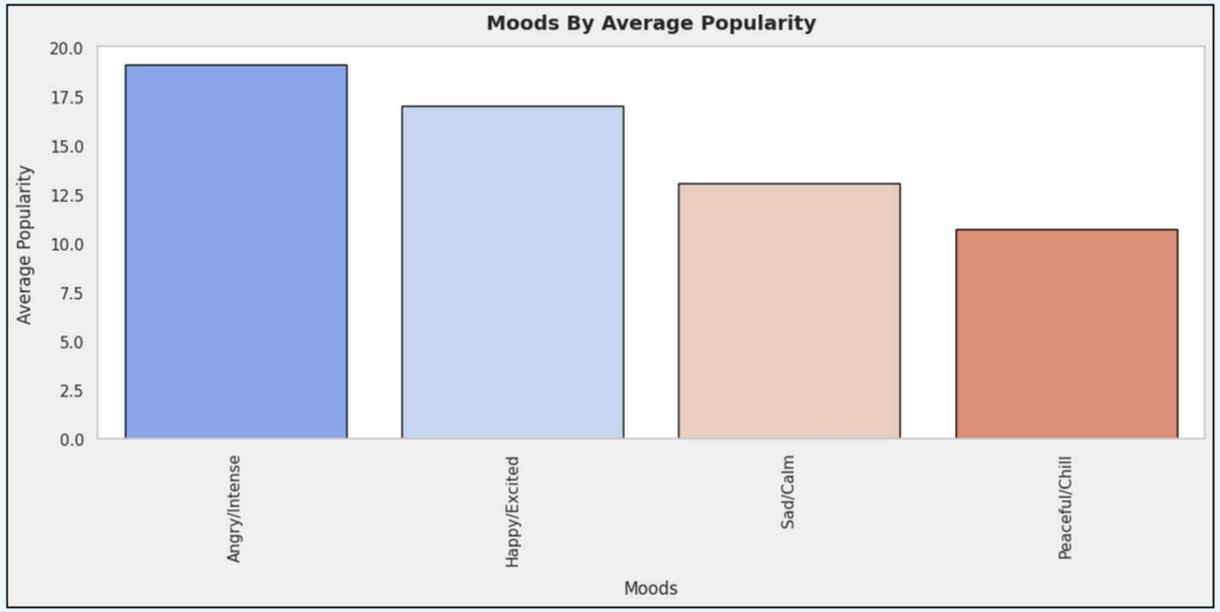
# Language Wise Analysis



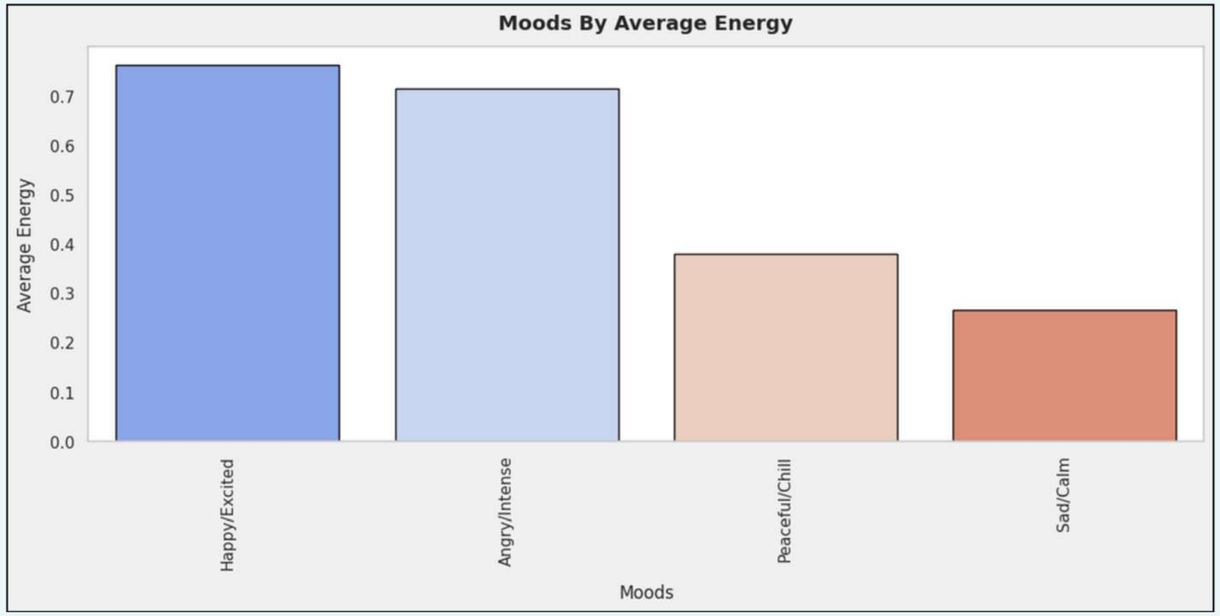




- Across languages and moods, **English dominates overall** volume with especially large Happy/Excited and Sad/Calm catalogs, while Tamil and Korean are the strongest regional drivers for upbeat, high-energy programming—Tamil excels in party-leaning Happy/Excited, Korean balances big Angry/Intense and Happy/Excited segments, **Hindi mixes energetic and romantic moods**, and Telugu/Malayalam contribute smaller but focused pockets—so for scale use English, for party target Tamil, for hype target Korean, and for mainstream film/romance lean on Hindi.
- **English and Korean have the broadest share of higher-popularity segments,** with English leading total volume and showing the largest Medium–High bins, while Korean over-indexes in Medium and retains a meaningful High slice; **Tamil and Hindi are sizable but skew toward Very Low/Low,** Telugu and Malayalam are smallest overall.

# Mood Wise Analysis

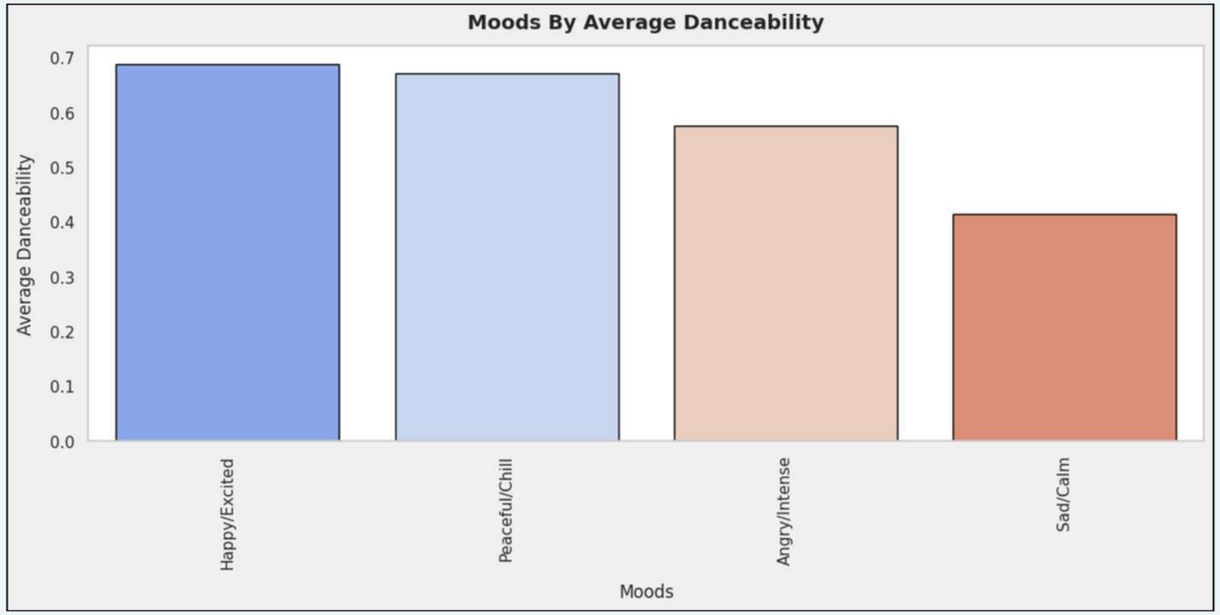



- **Popularity by mood :** Angry/Intense leads, then Happy/Excited, with Sad/Calm and Peaceful/Chill trailing—**high-arousal moods attract more average popularity** in this dataset.

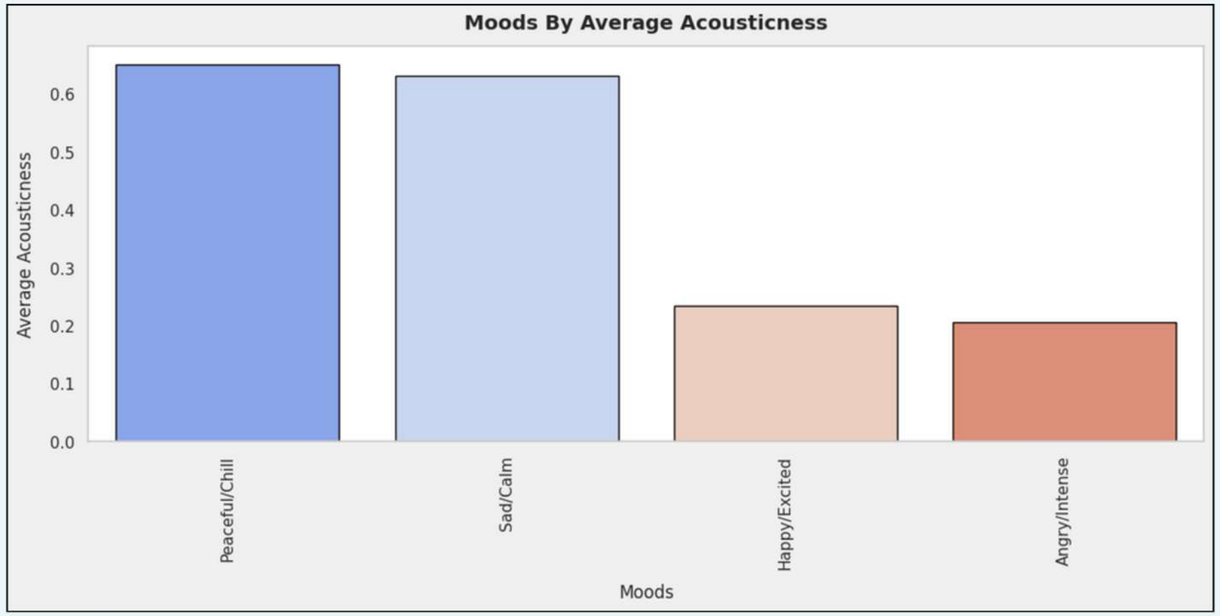



- **Energy by mood :** Happy/Excited has the highest energy, Angry/Intense is slightly lower but still high, Peaceful/Chill is mid-low, and Sad/Calm is lowest—confirming energy is a primary separator across moods.

# Mood Wise Analysis

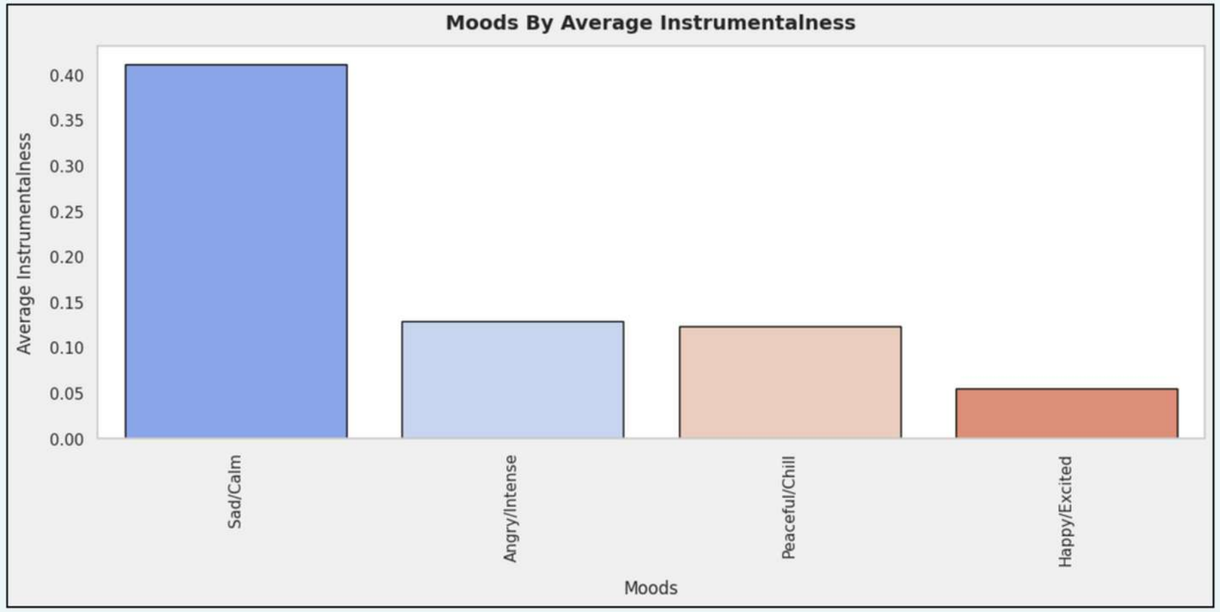



- **Danceability by mood :** Happy/Excited is most danceable and Peaceful/Chill is close behind, while Angry/Intense is moderate and Sad/Calm is lowest—party programming aligns best with Happy/Excited.

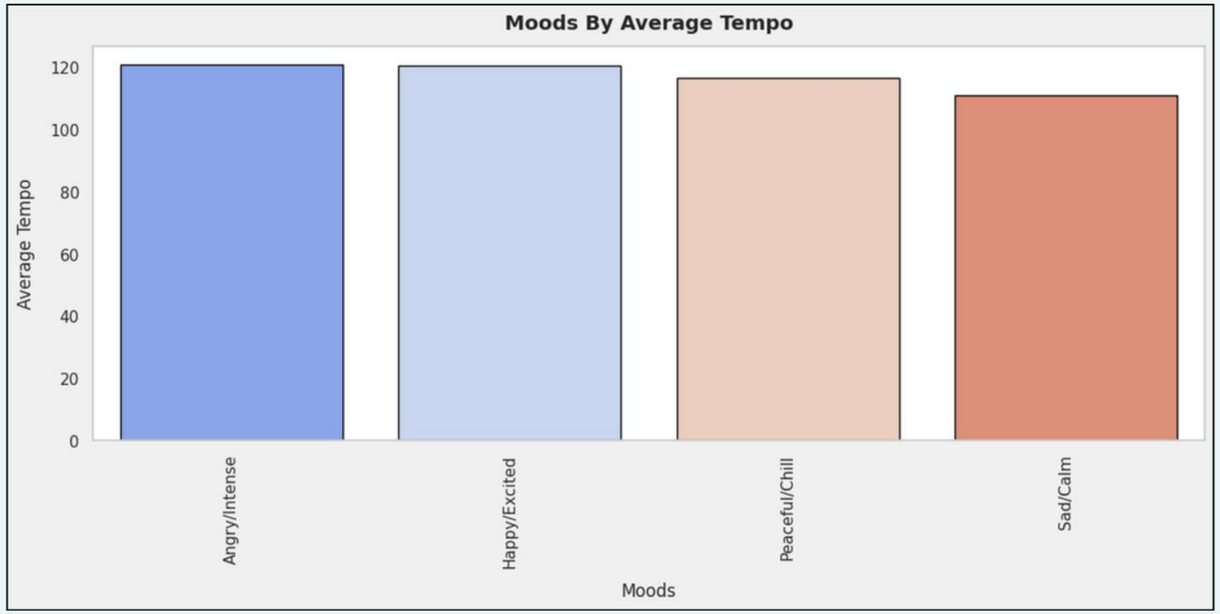



- **Acousticness by mood :** Peaceful/Chill and Sad/Calm are highly acoustic, whereas Angry/Intense and Happy/Excited are low on acousticness—organic textures map to calm moods, processed/electric textures to energetic moods.

# Mood Wise Analysis




- **Instrumentalness by mood :** Sad/Calm is markedly more instrumental than other moods, while Happy/Excited is the least—vocals drive upbeat tracks, whereas calmer pieces lean instrumental.

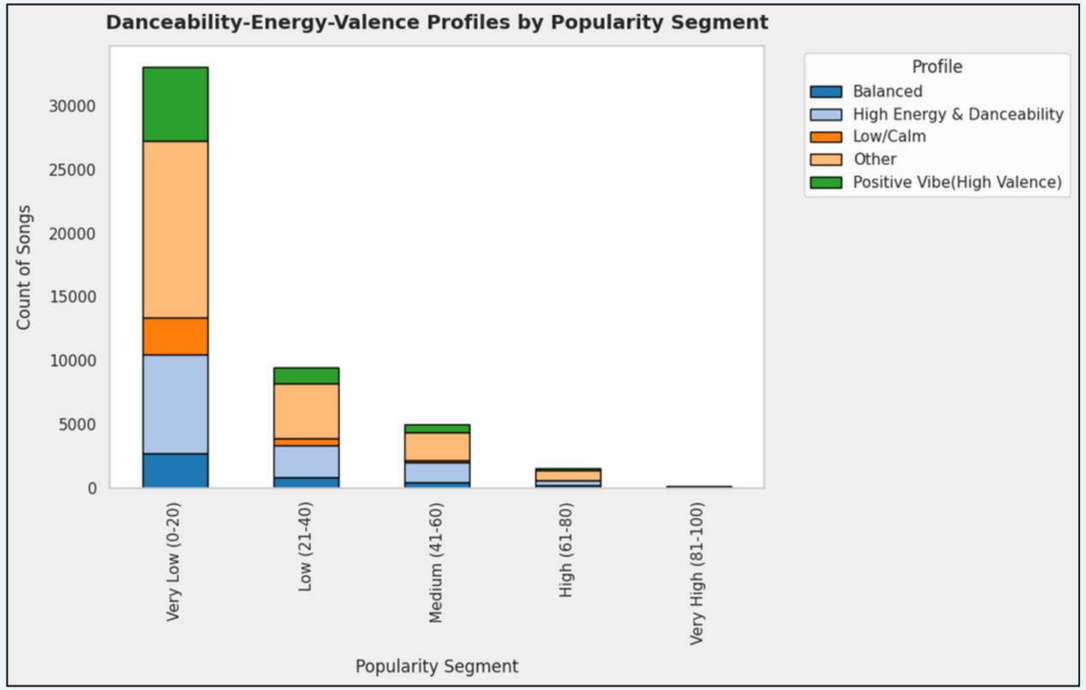


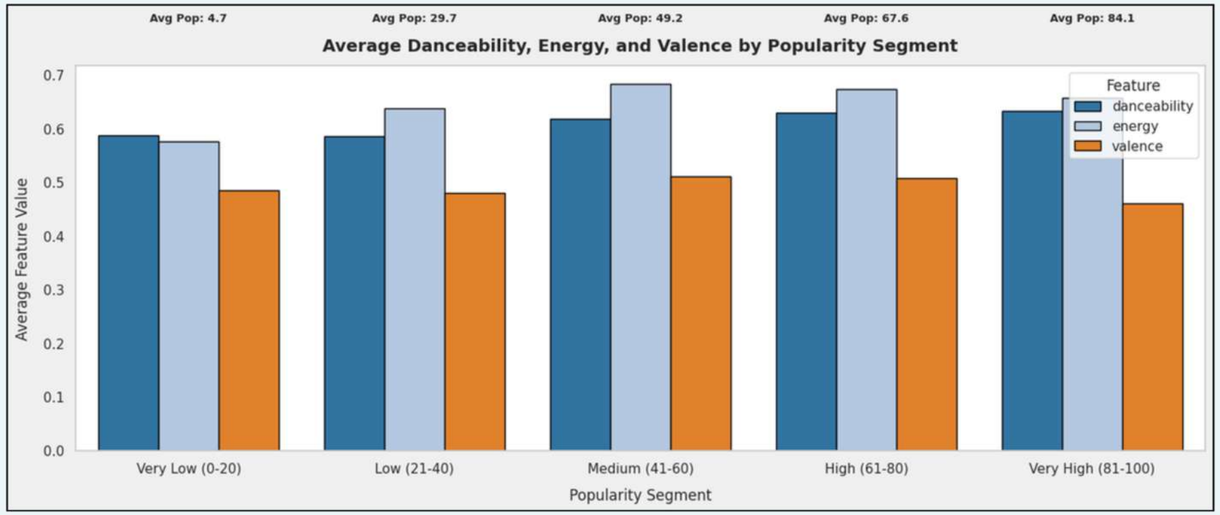

- **Tempo by mood :** Angry/Intense and Happy/Excited sit around ~120 BPM, Peaceful/Chill slightly lower, and Sad/Calm the slowest—tempo aligns with arousal but differences are modest versus energy gaps

Spotify

# *Multivariate Analysis*

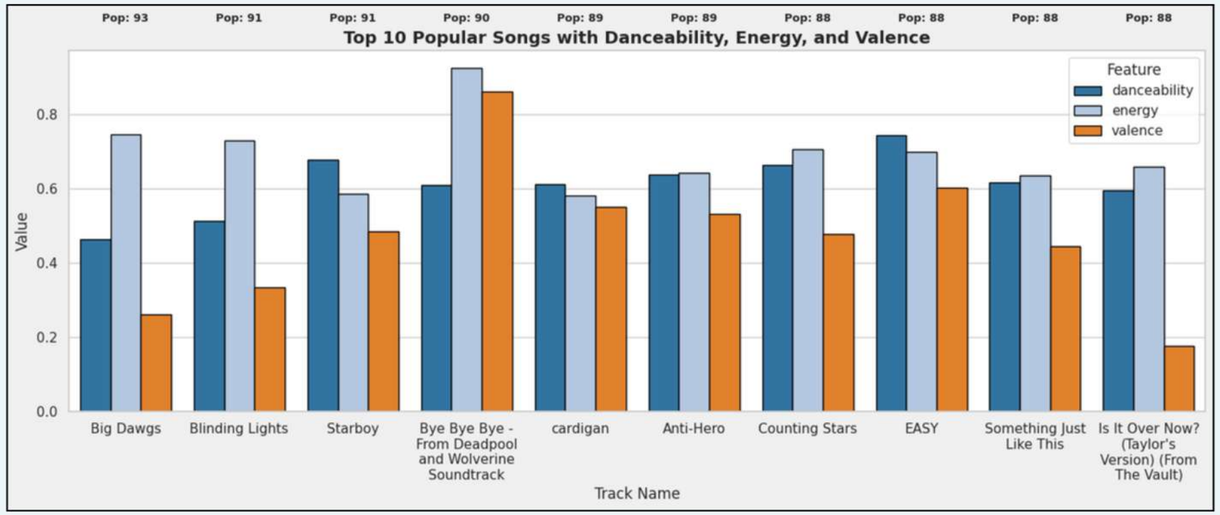
# Popularity Wise Analysis

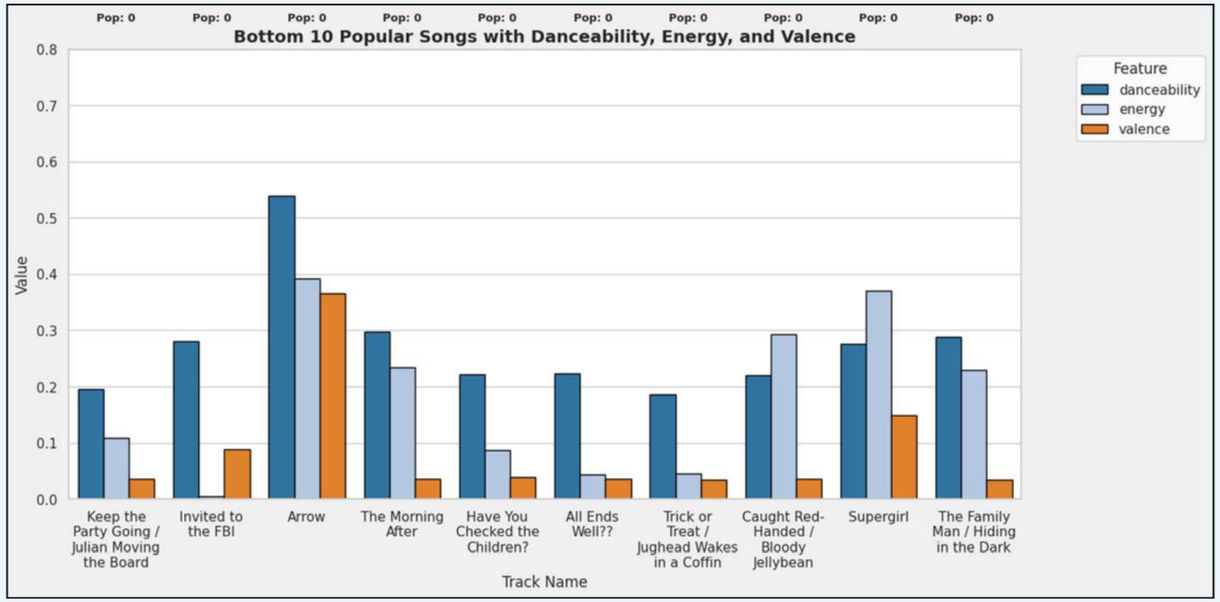



- **Segment profiles :** As popularity rises from very low to very high, average energy and danceability steadily increase, but valence peaks around the middle and softens at the very top—**hits are energetic and movable, not necessarily the "happiest."**
- **Mix archetypes by segment :** High and very-high segments are enriched for "High Energy & Danceability" profiles and have fewer "Low/Calm" cuts, suggesting curation should bias toward drive and groove for hit-seeking playlists.

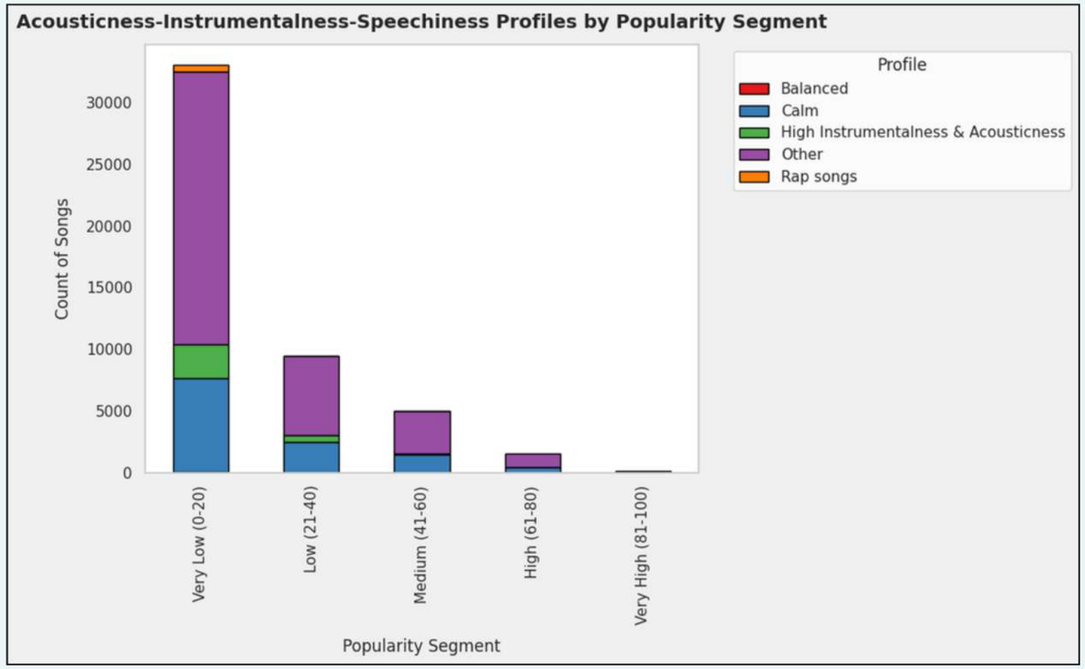
# Popularity Wise Analysis

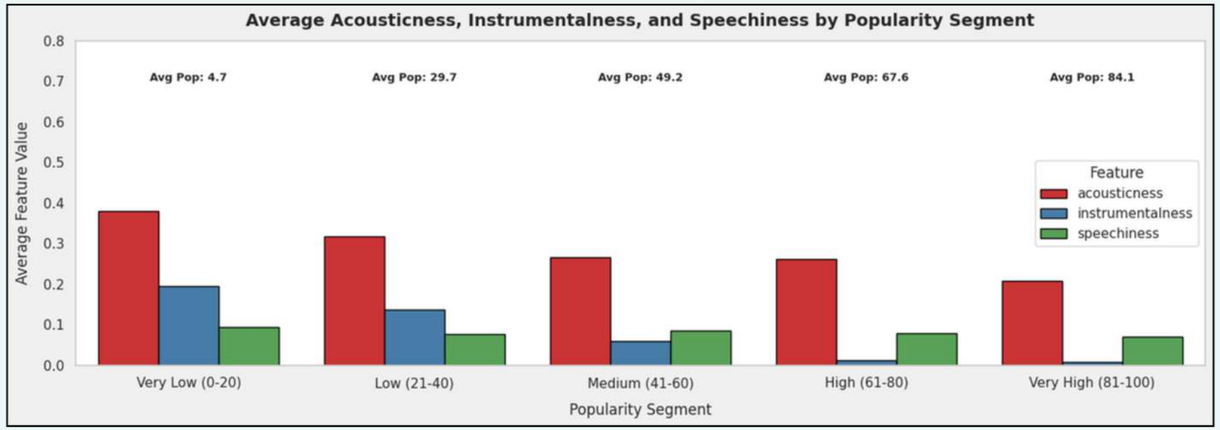



- **Top vs bottom songs :** The most popular tracks pair mid-high danceability with **high energy and moderate valence,** while the least popular skew **low on all three or lack consistent energy/danceability,** underscoring that punch and movement matter more than pure "happiness."

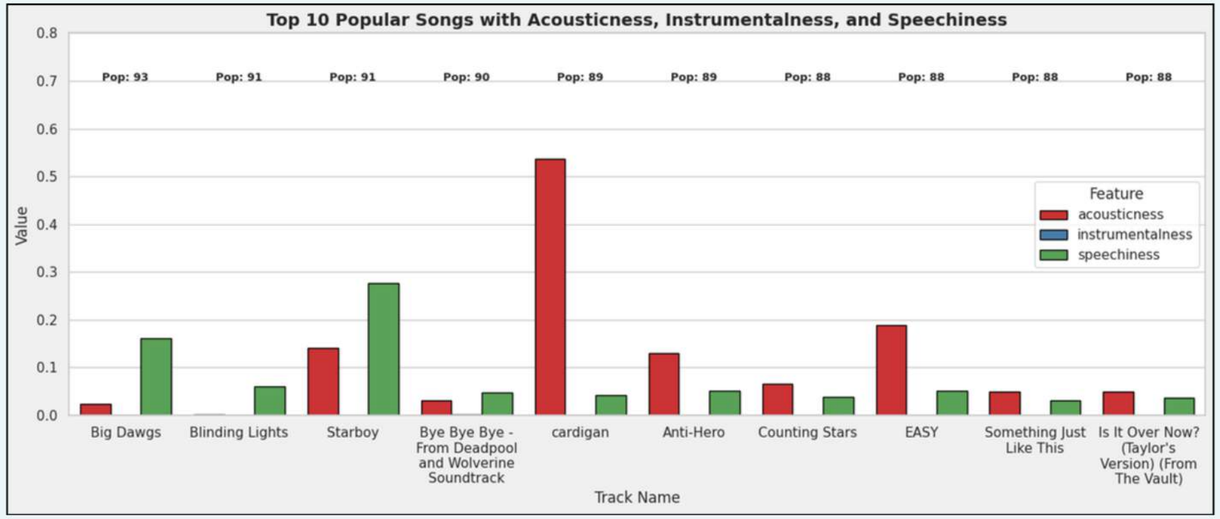
# Popularity Wise Analysis

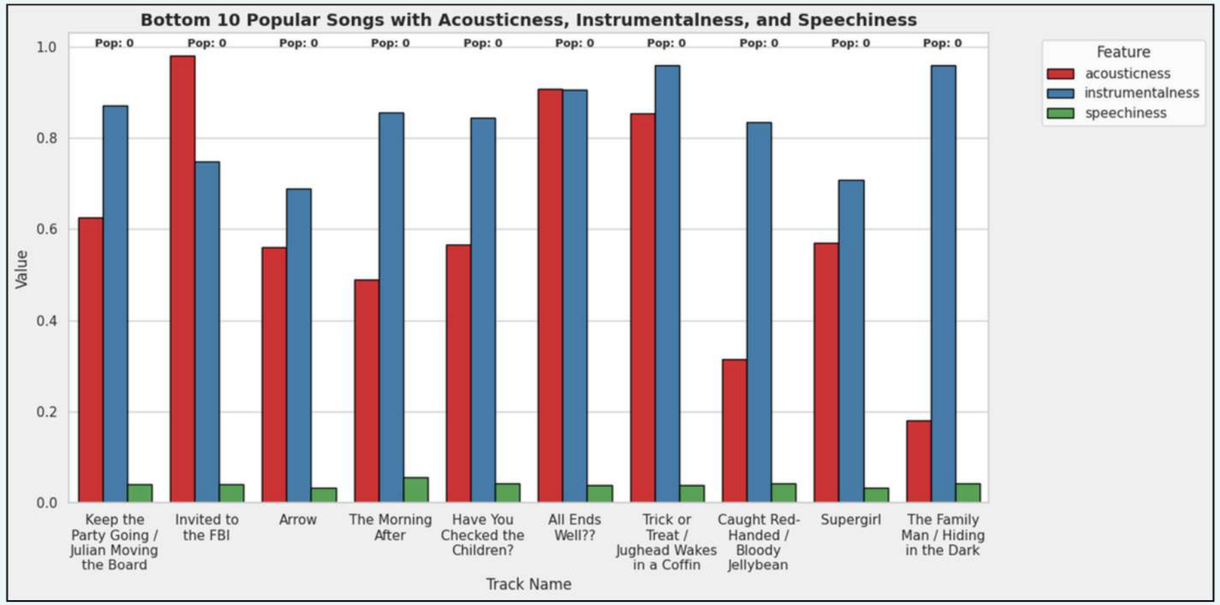



- **Profile shift with popularity :** As popularity rises, average instrumentalness drops sharply and acousticness declines, while speechiness edges down—not many rap-high or instrumental-only tracks sit in the top tier.
- **Composition archetypes by segment :** "Calm" and "High Instrumentalness & Acousticness" dominate the very-low and low segments, **whereas higher segments thin out these profiles in favor of more balanced, vocal-centric productions.**

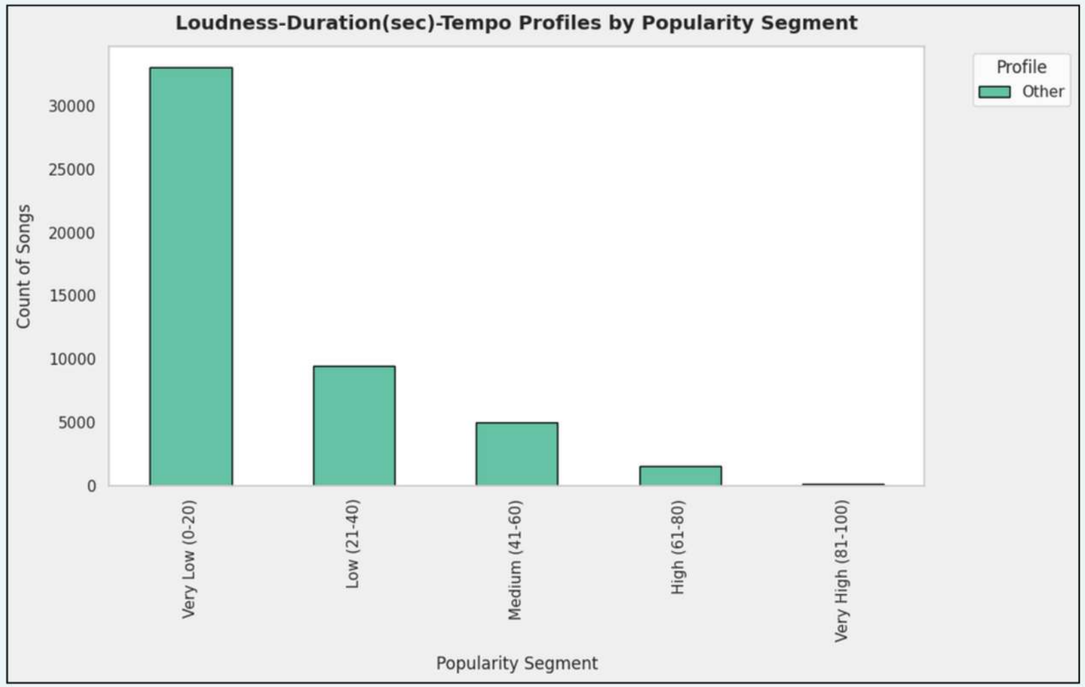
# Popularity Wise Analysis

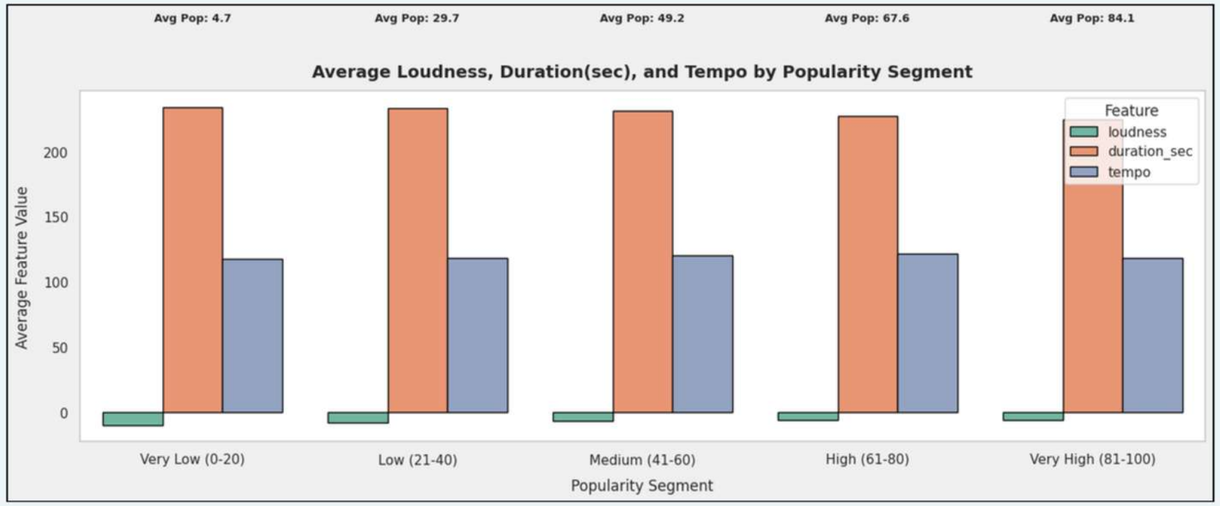



- **Bottom vs top songs :** The least popular tracks are very high in instrumentalness and acousticness with low speechiness, while the most popular are low on instrumentalness, moderately acoustic at most, and keep speechiness modest—confirming that **vocal-forward, less-acoustic mixes fare better.**

# Popularity Wise Analysis

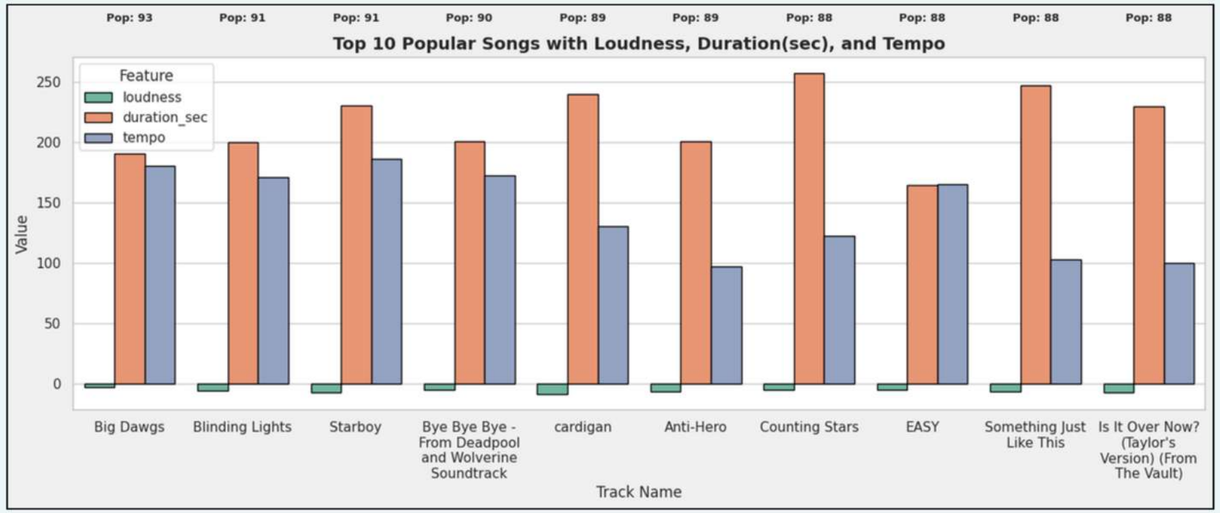


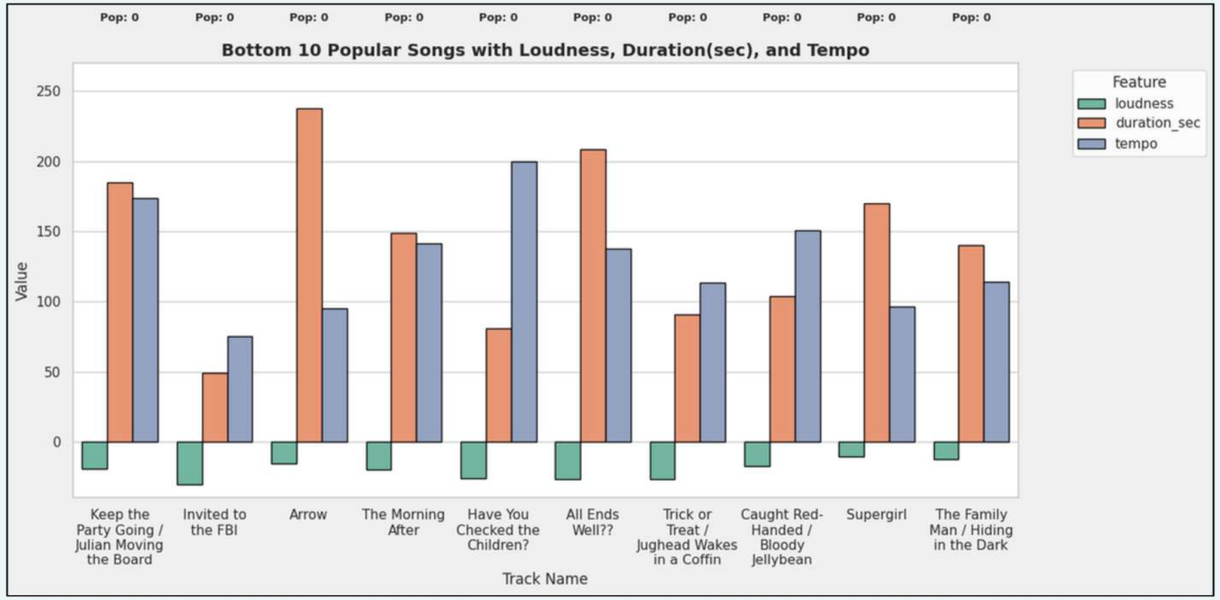




- **Segment trends :** Across popularity segments, tempo stays in a tight mid-tempo band, duration shortens slightly at the top, and average loudness becomes less negative—suggesting mastering level and tight format fit matter more than raw speed.

- **Segment mix profile :** The distribution is dominated by a single “other” profile across segments, implying no single L-D-T archetype guarantees success; instead, staying within tight professional bounds is necessary but not sufficient.

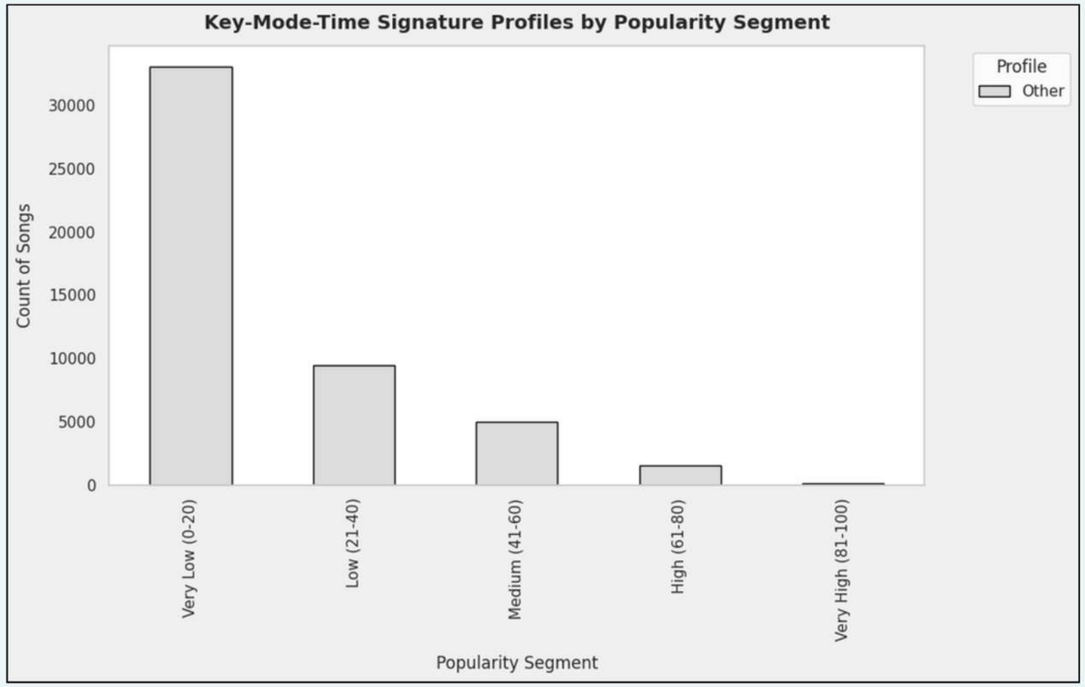
# Popularity Wise Analysis

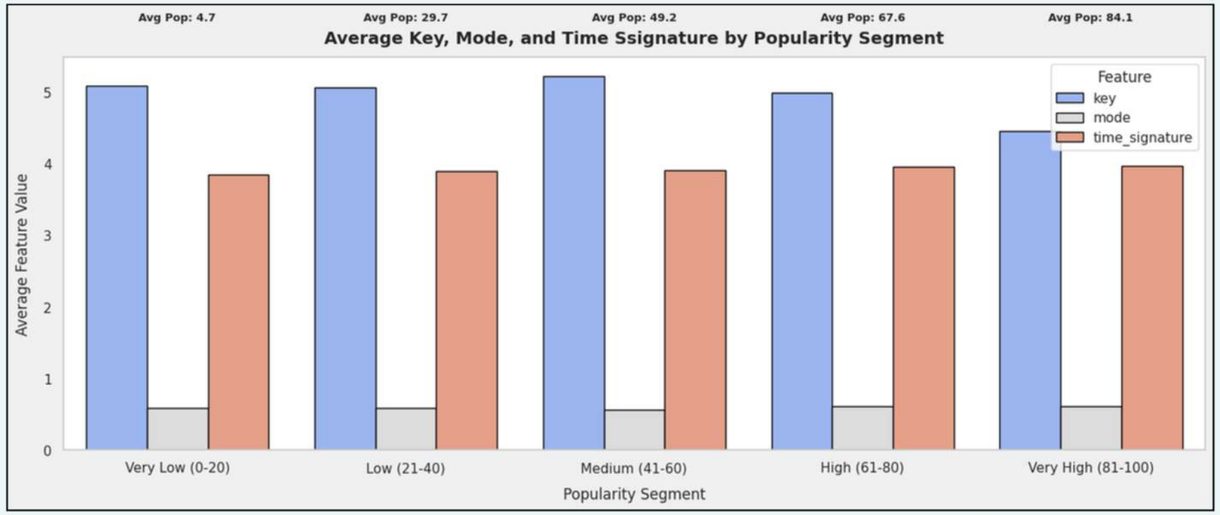



- **Top-hit anatomy :** The most popular tracks cluster around radio-friendly lengths (2.7–4.3 min), moderately fast tempos (~100–185 BPM), and consistently loud masters (around −6 to −4 dBFS), reinforcing **"short, punchy, and loud" as the prevailing hit recipe.**
- **Bottom-tier contrast : Least-popular tracks vary widely in tempo and duration**, but are generally quieter (more negative loudness), showing that under-powered masters correlate with poor traction even when tempo is similar.
- **Per-track view (top 10) :** Individual hits mix mid-to-high tempos with assertive loudness and concise runtimes; outliers with slower tempos compensate via higher loudness and strong hooks.

# Popularity Wise Analysis

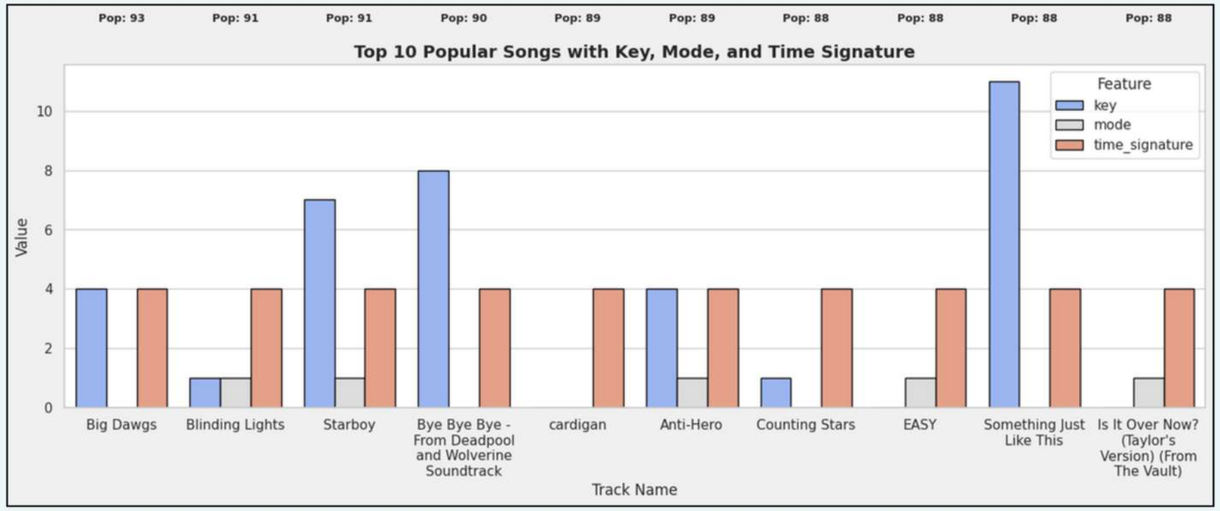


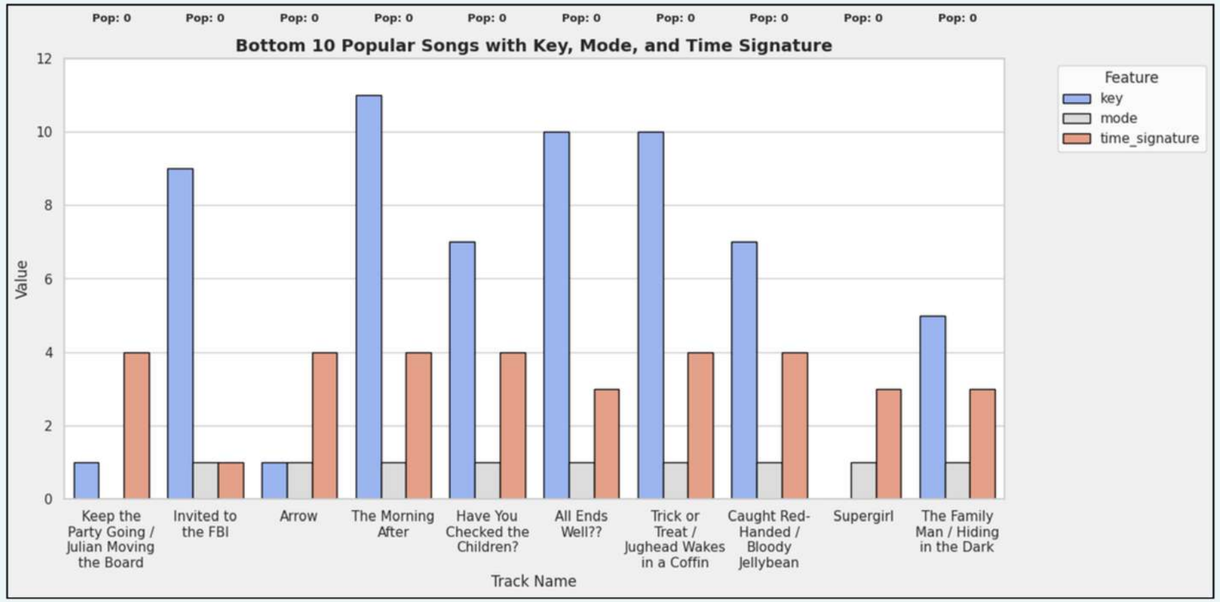




- **Key and time signature aren't decisive for hits :** Across popularity bands, average key index hovers mid-scale and time signature centers on 4/4 with tiny variation, indicating these traits are broadly standardized and weak predictors of success.

- **Segment distribution :** The stacked profile by popularity shows no distinctive key-mode-meter archetype emerging at higher tiers, so **optimizing melody, energy, and hook design should take precedence** over theoretical key/meter selection.

# Popularity Wise Analysis

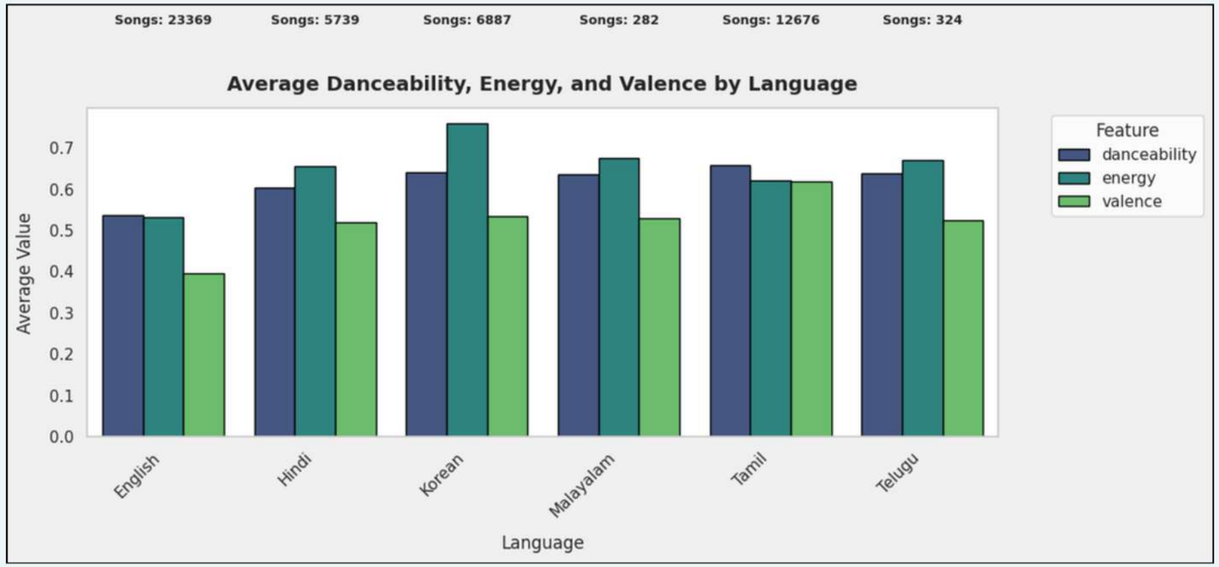


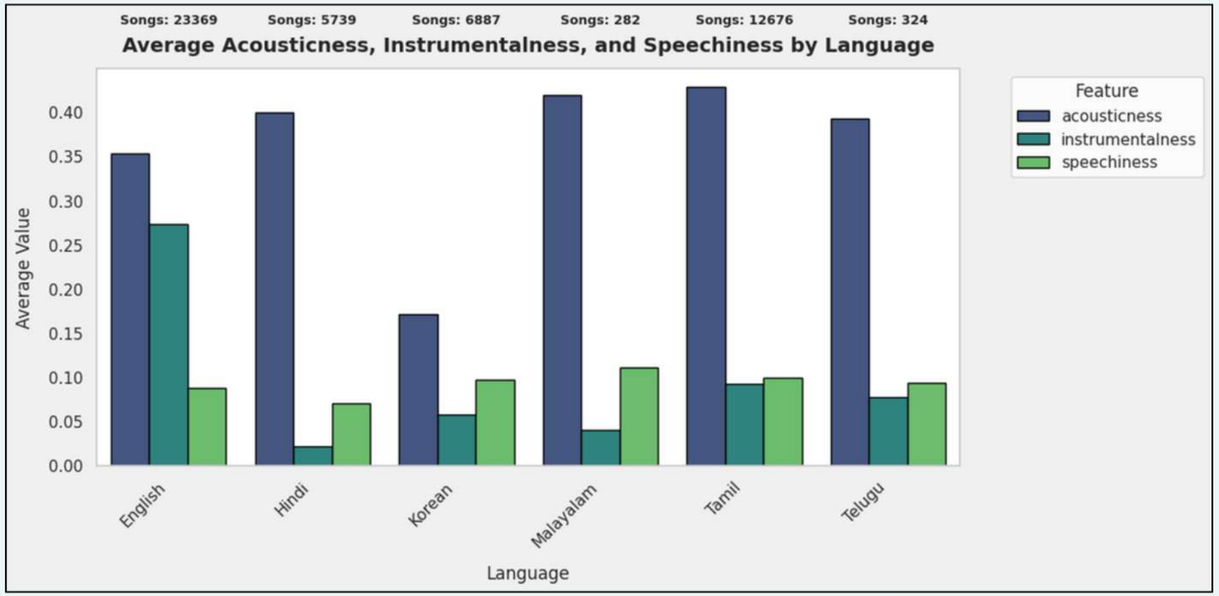



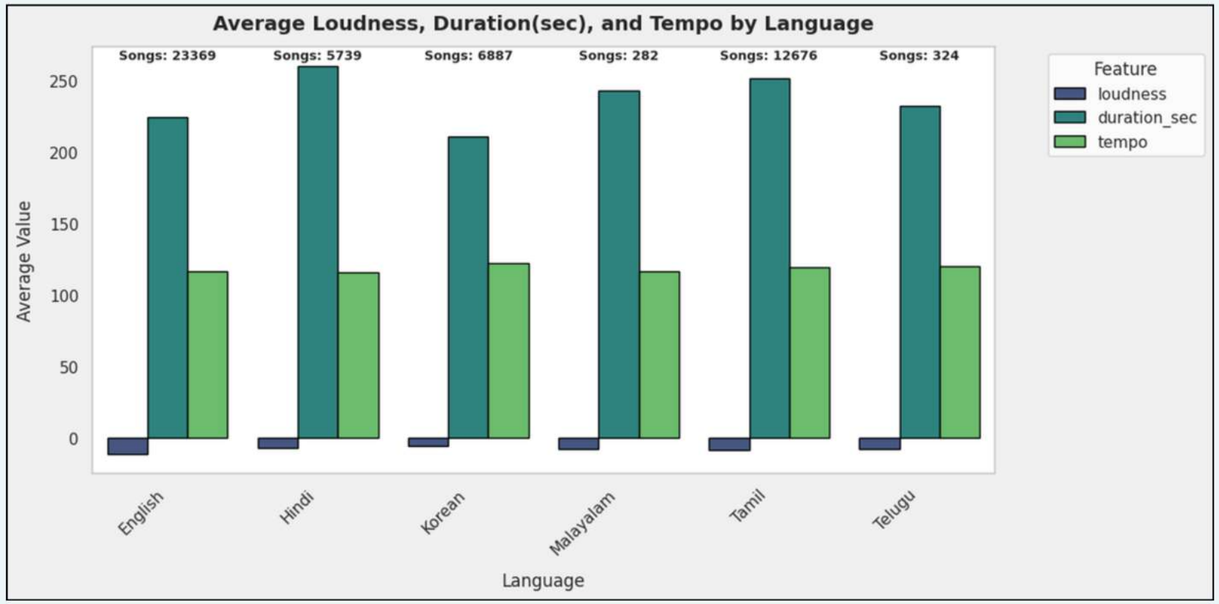

- **Mode skews major for everyone :** The average mode is consistently near 1.0 in all segments, showing a strong bias toward major-key tracks regardless of popularity—minor isn't penalized, just less common.

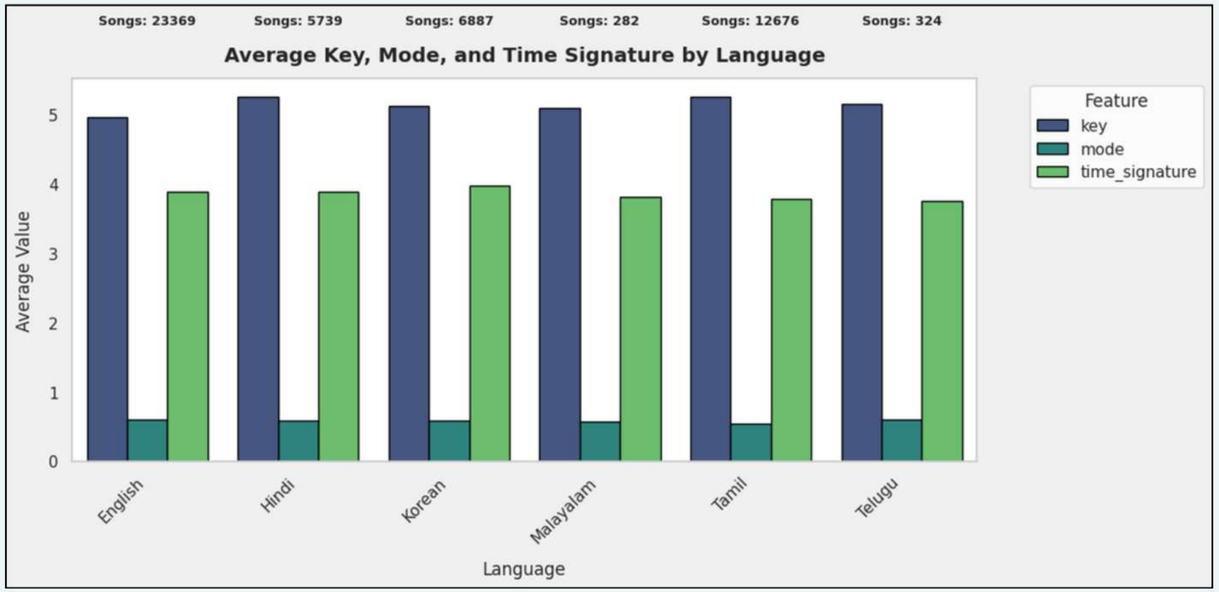
- **Top vs bottom tracks :** Both top-10 and bottom-10 examples overwhelmingly use 4/4 and a mix of keys in major mode, reinforcing that songwriting and production factors matter more than key/meter choices.

# Language Wise Analysis



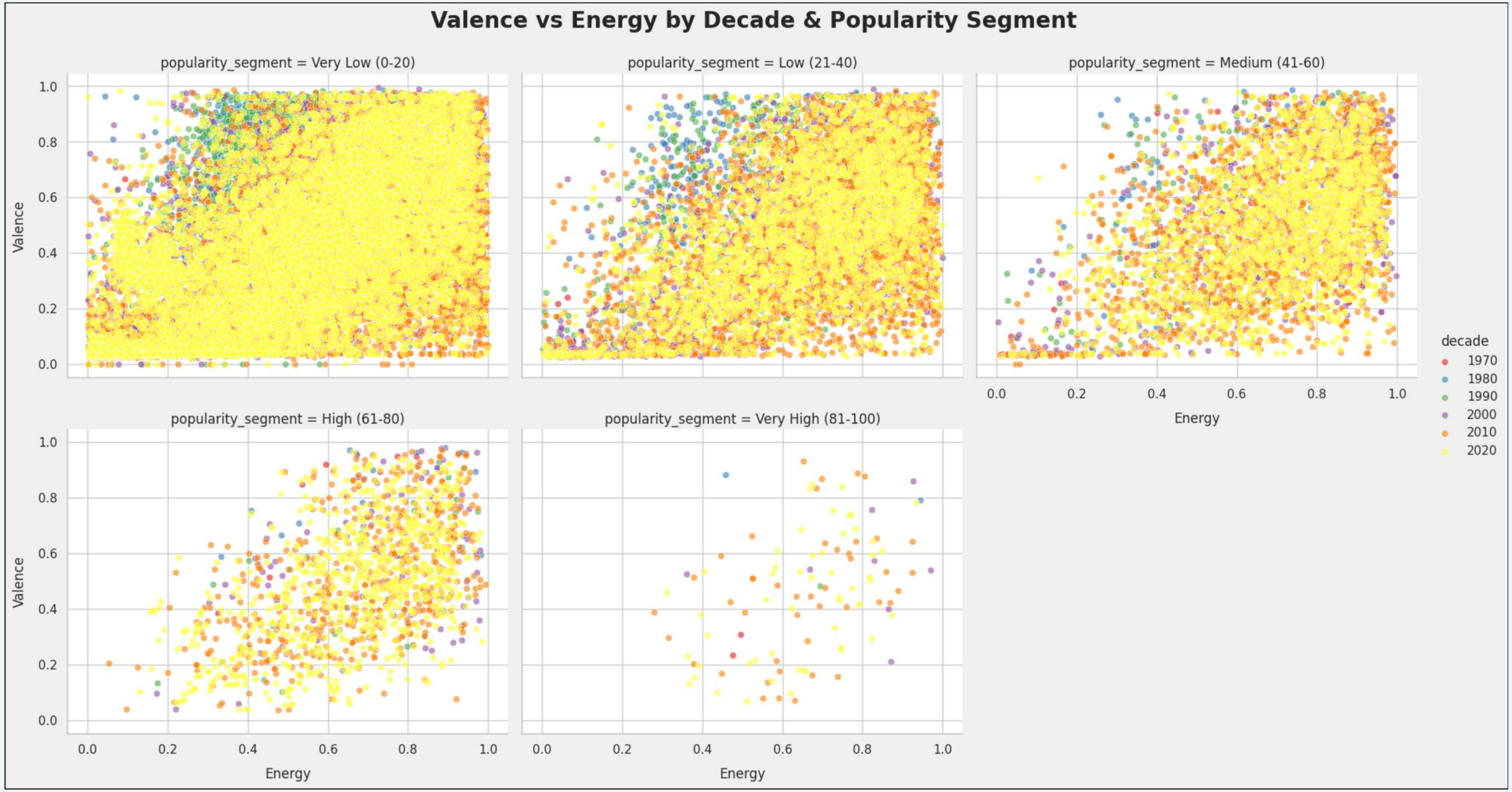



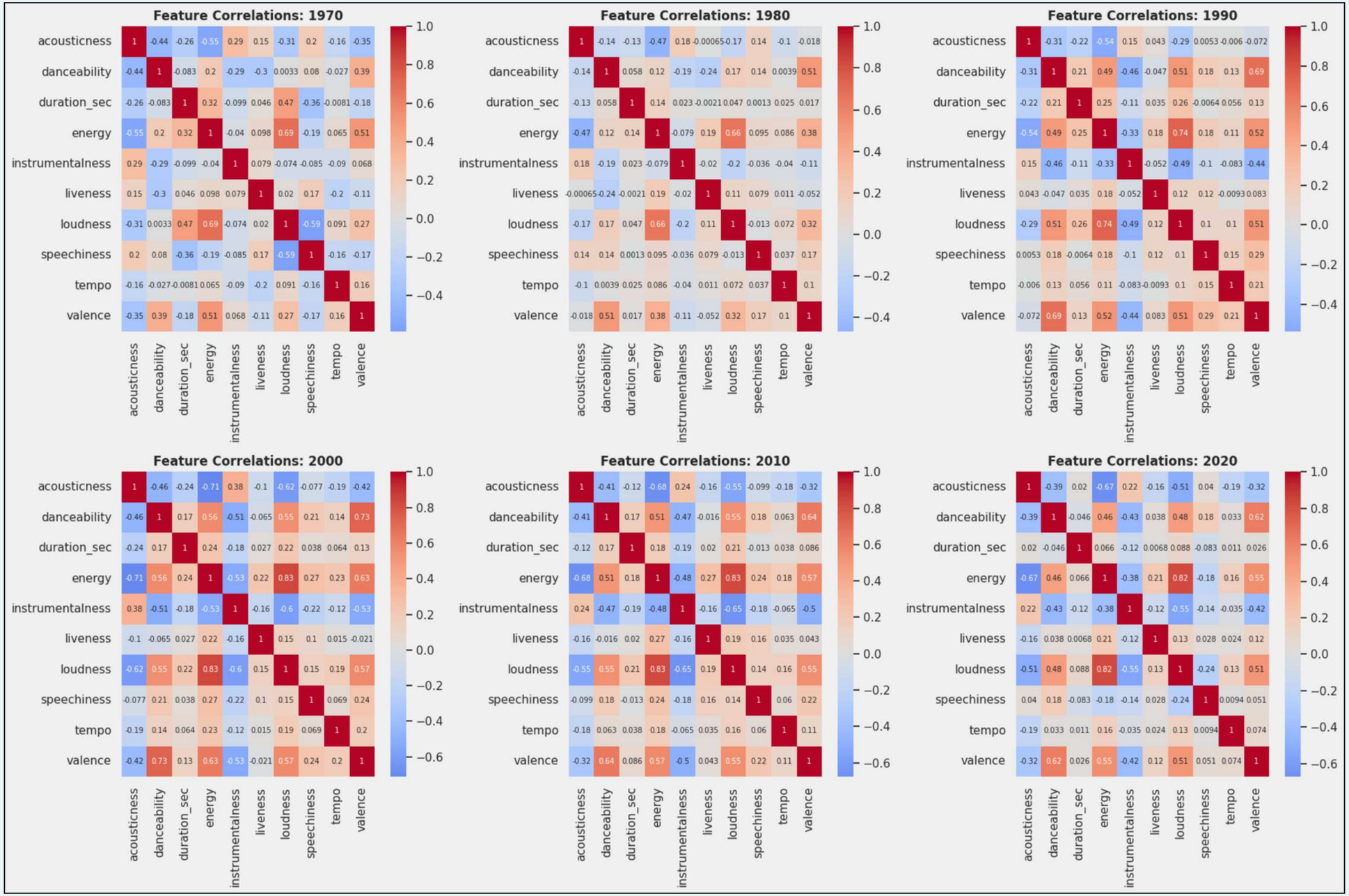
# Overview

- **Core structure is uniform across languages :** Keys hover near mid-index values, **major mode dominates, and time signature centers on 4/4** for English, Hindi, Korean, Malayalam, Tamil, and Telugu—song theory choices are broadly standardized.
- **Movement and punch vary by language :** Danceability rises from English to Tamil/Telugu, energy peaks for Korean, and valence is highest in Tamil, implying Tamil is best for upbeat feel-good sets while Korean excels at high-octane curation.
- **Format differences :** Hindi and Tamil have the longest average durations, Korean the shortest; tempos cluster tightly ~115–122 BPM across all languages, showing speed norms are similar while runtime strategy differs.
- **Texture and vocals :** English catalogs skew more instrumental than Indian languages; Tamil/Malayalam/Telugu show higher acousticness and moderate speechiness, whereas Korean has lower acousticness but slightly higher speechiness—use English for instrumental/sync beds, Tamil/Malayalam for organic moods, and Korean for modern vocal-driven energy.

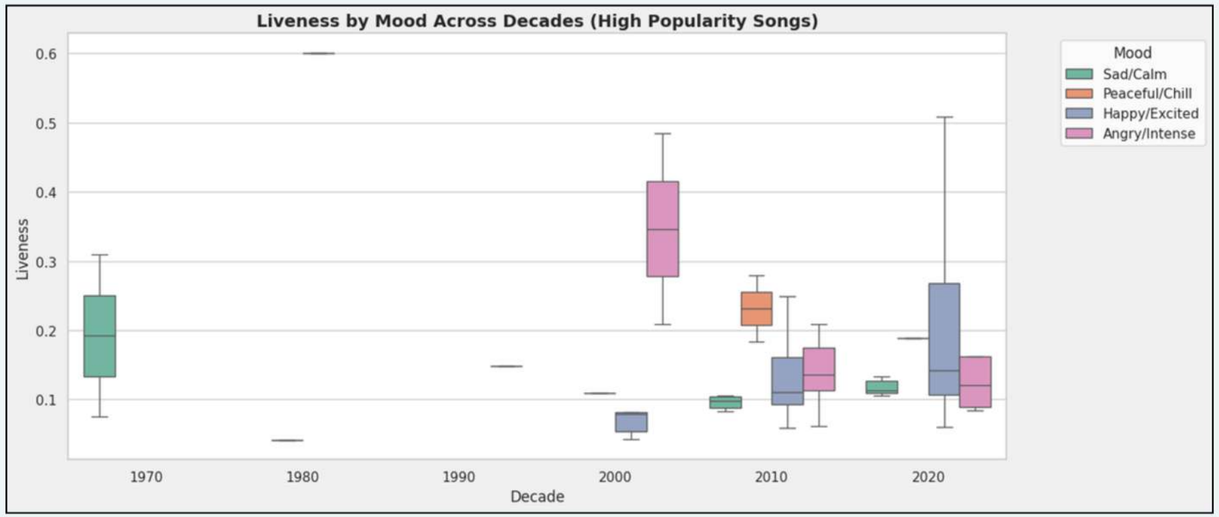
# Decade Wise Analysis

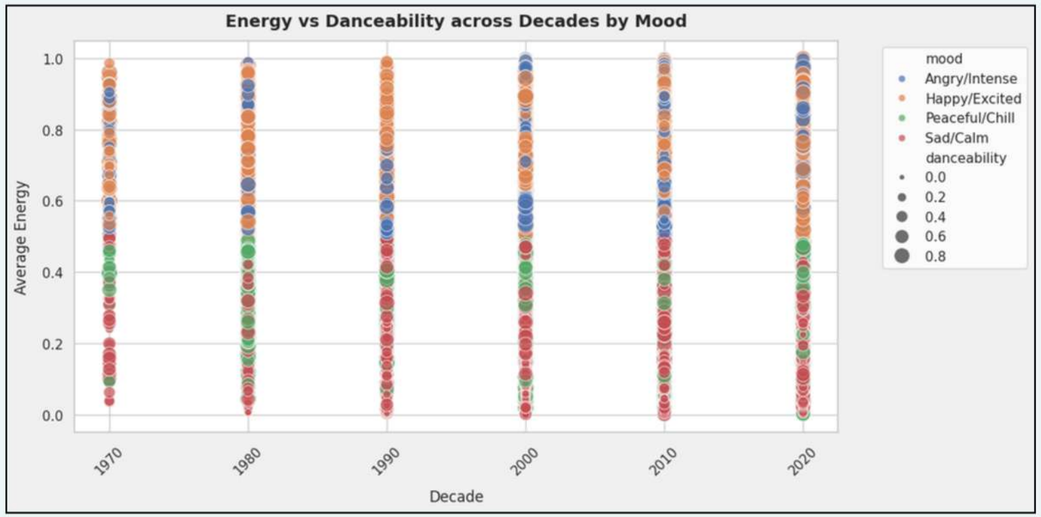


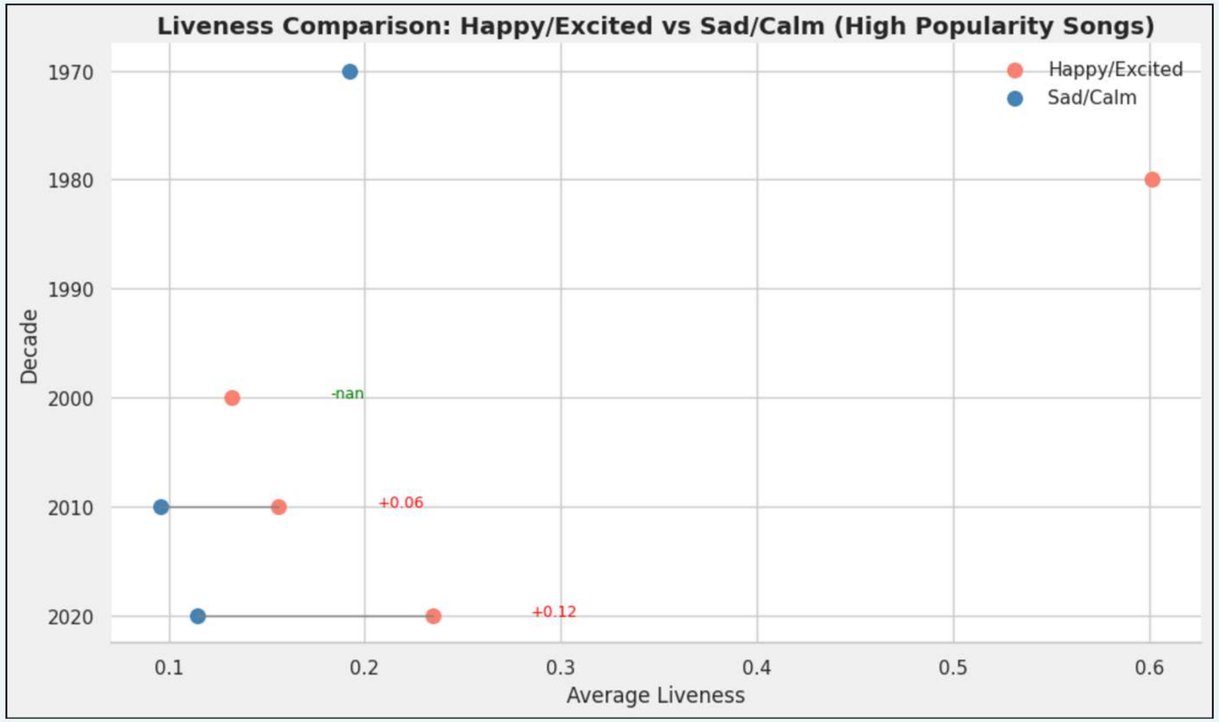
# Decade Wise Analysis

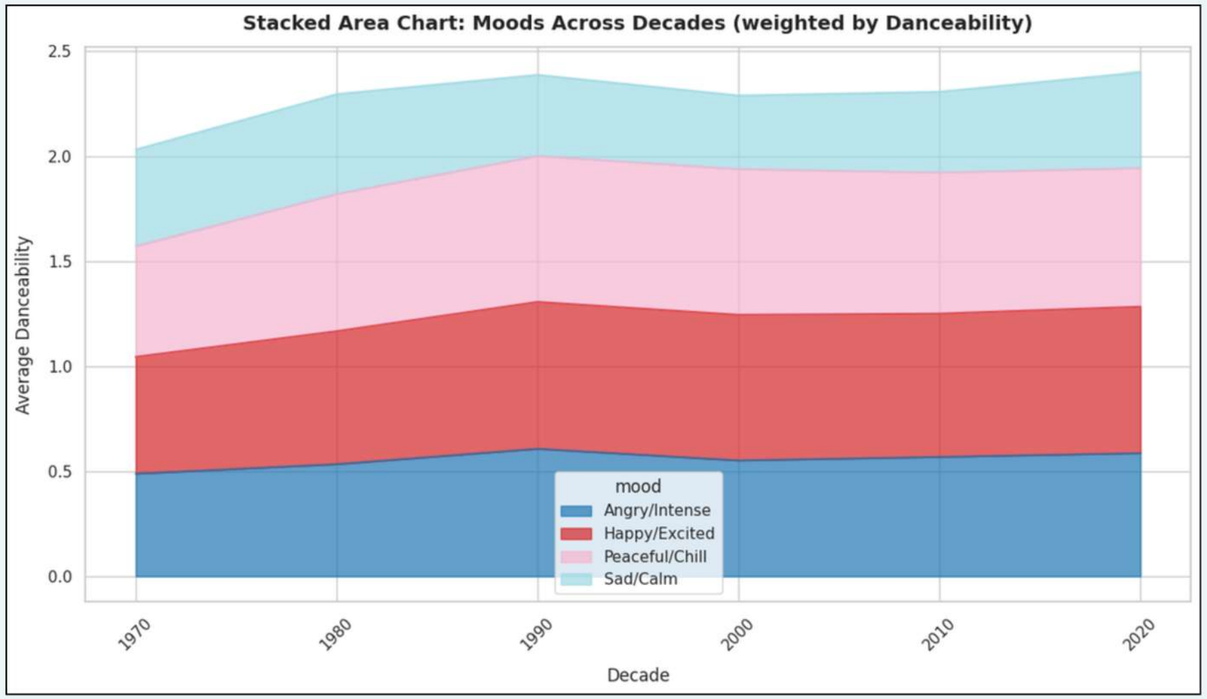


# Decade Wise Analysis

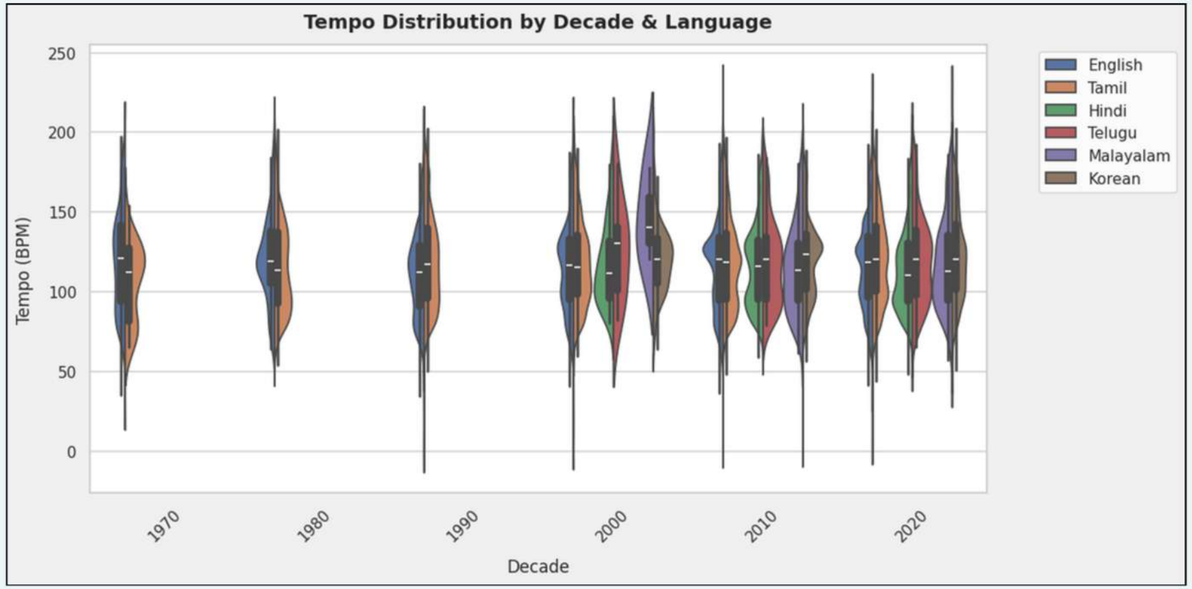


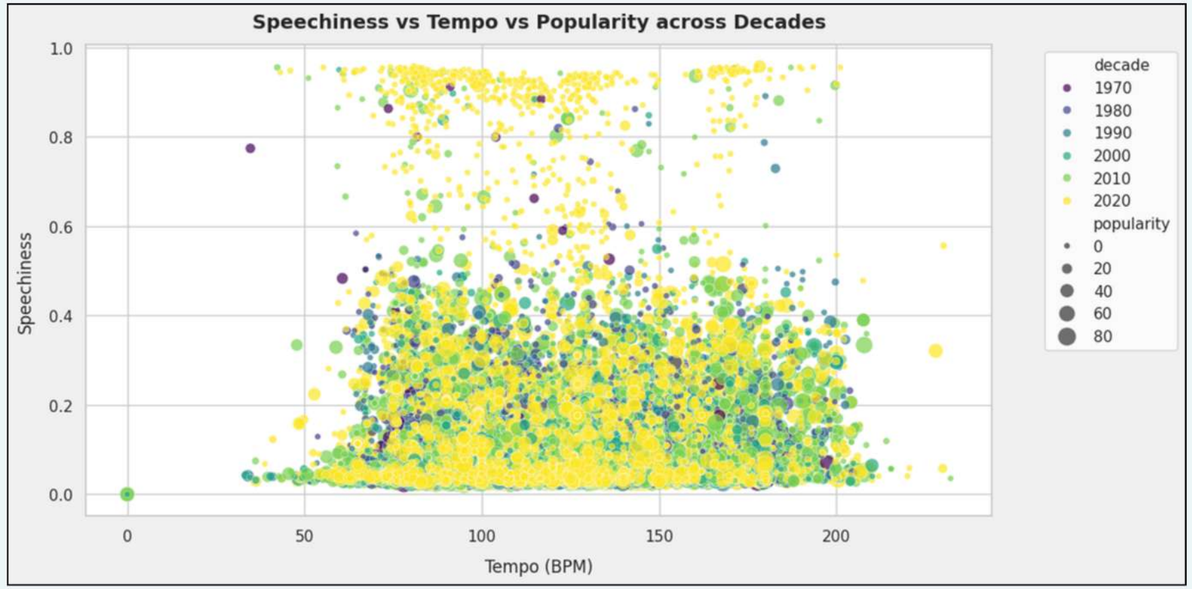



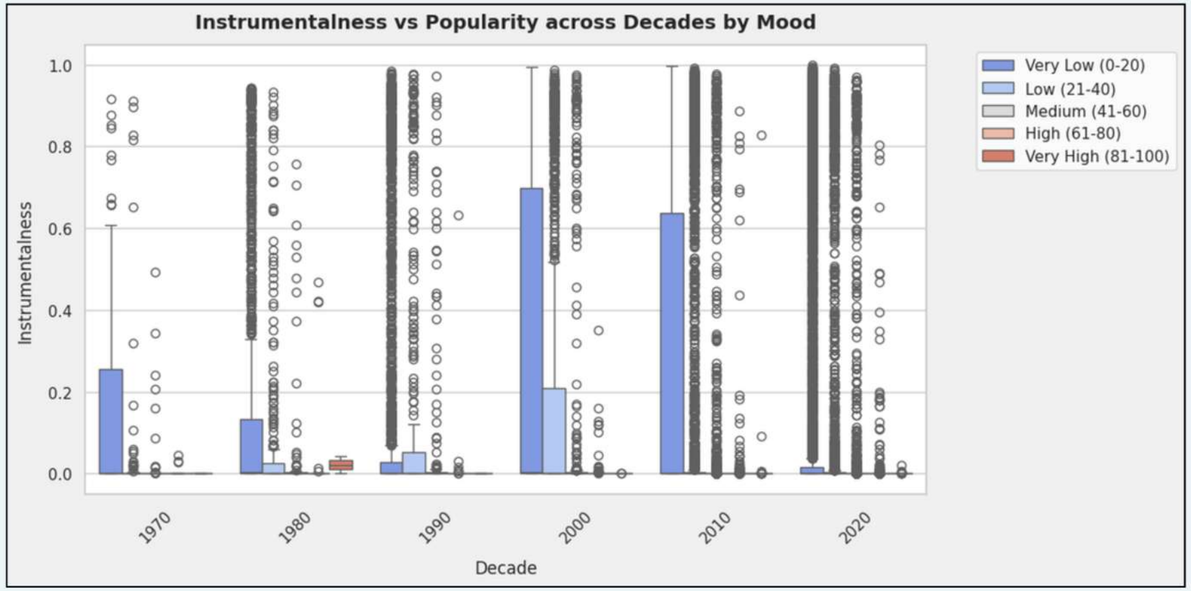



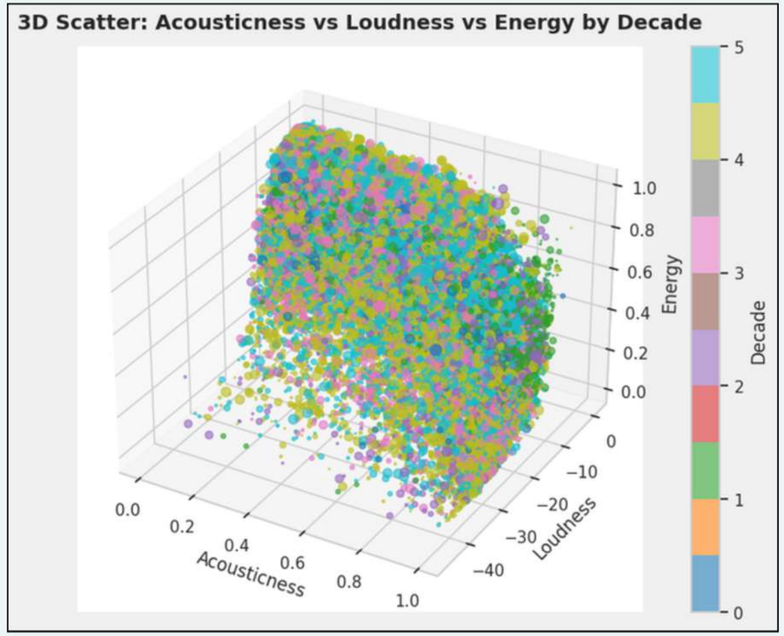



# Decade Wise Analysis



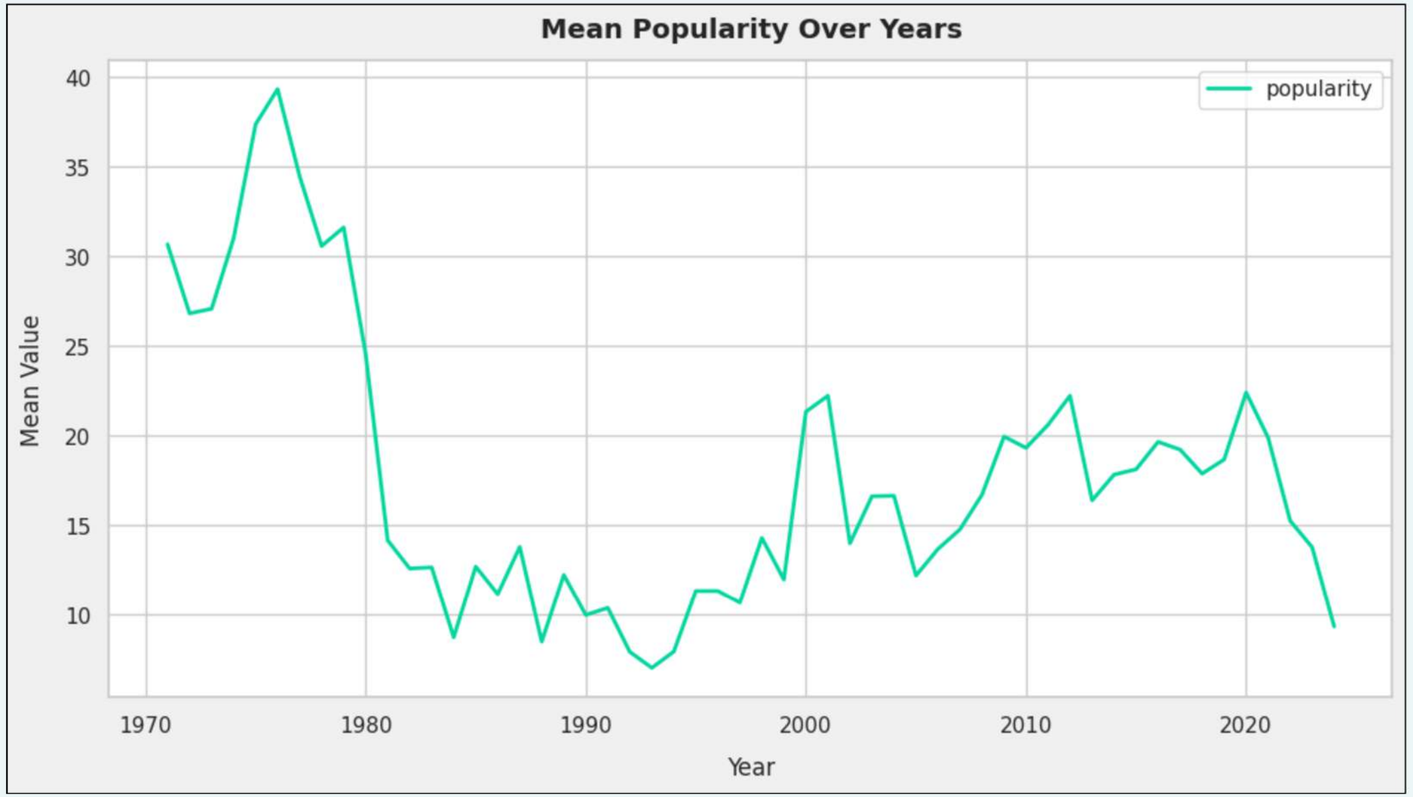
# Overview

- **High-popularity favors high energy over high valence :** In every segment, the densest clusters of top-decade points sit at energy ≥0.6 while valence spreads wide, showing hits are driven by intensity, not necessarily "happiness."
- **Live feel trends down with time for hits :** Average liveness in high-popularity songs drops from early decades to 2010s/2020s for both Happy/Excited and Sad/Calm, indicating tighter studio polish and fewer live artifacts in recent charting tracks.
- **Mood-by-decade liveness :** Angry/Intense shows the widest liveness ranges in the 2000s–2020s, while Peaceful/Chill remains low and stable, suggesting aggressive genres tolerate more room/noise than chill formats.
- **Danceability as a multiplier :** Across decades and moods, points with larger markers (higher danceability) consistently align with higher energy, reinforcing that movement plus punch is the clearest common recipe across eras.
- **Rap proxy vs tempo :** High-speechiness cuts (rap/spoken) appear across a wide BPM band, but the largest, most popular bubbles cluster between ~85–105 and ~120–140 BPM, implying two sweet spots for hit rap pacing.
- **Tempo norms by language :** Medians sit tightly around ~110–125 BPM for all languages each decade, with Korean and Tamil showing slightly higher ceilings—so tempo alone rarely differentiates; arrangement and groove do.
- **Instrumentals underperform at scale :** Across decades, **high instrumentalness is concentrated in the very-low and low popularity bands**, with scarce presence in the high tiers, underscoring the advantage of vocal-led tracks for mainstream success.
- **Production triad :** Energy rises as loudness becomes less negative and acousticness falls; modern hits inhabit the low-acousticness, loud, high-energy corner, reinforcing the electronic/processed tilt of contemporary chart music.
- **Danceability has risen for high-movement moods :** Weighted by danceability, **Happy/Excited and Angry/Intense both trend upward since the 1970s, while Peaceful/Chill plateaus and Sad/Calm dips then recovers—party-leaning moods increasingly dominate motion.**
- **Stable feature physics across eras :** Correlations show a persistent pattern—energy correlates positively with loudness and negatively with acousticness; danceability aligns moderately with valence and tempo; instrumentalness opposes danceability/energy—modern decades strengthen the energy-loudness link.

Spotify

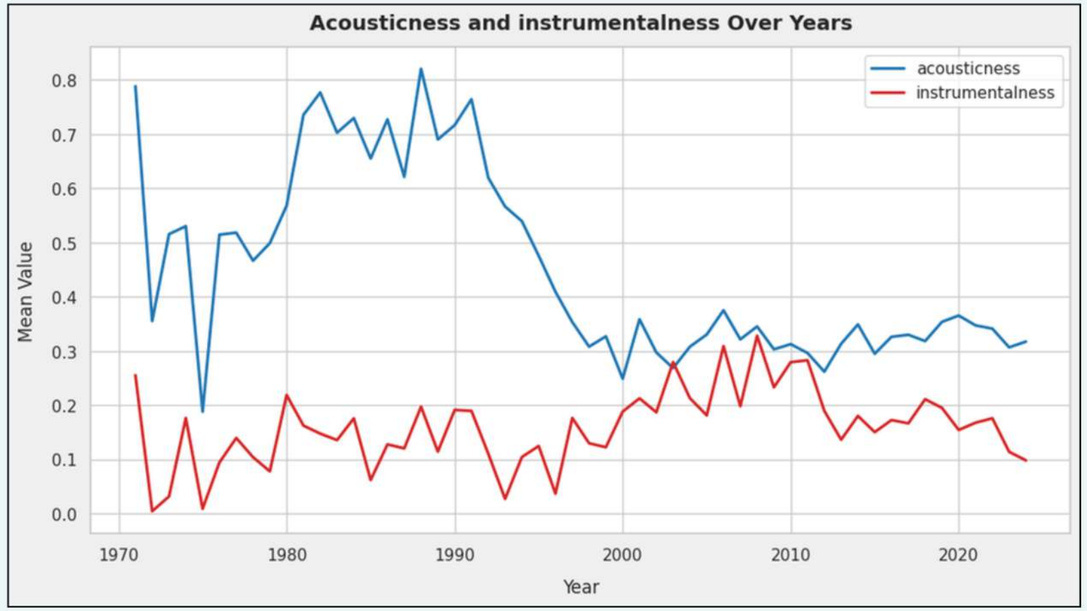
# *Timeseries Analysis*

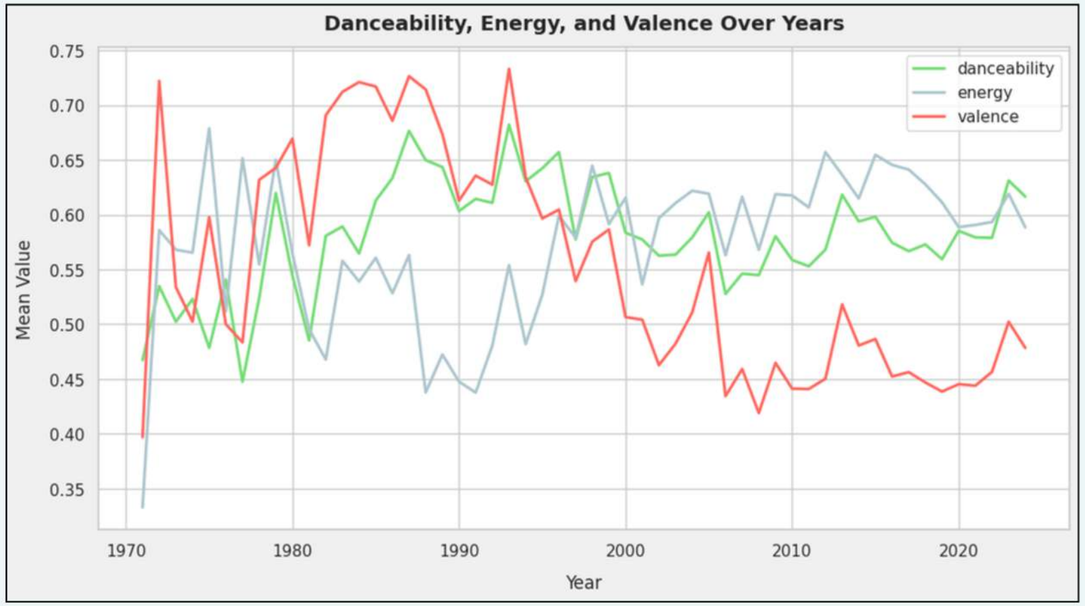
# Popularity Over Years

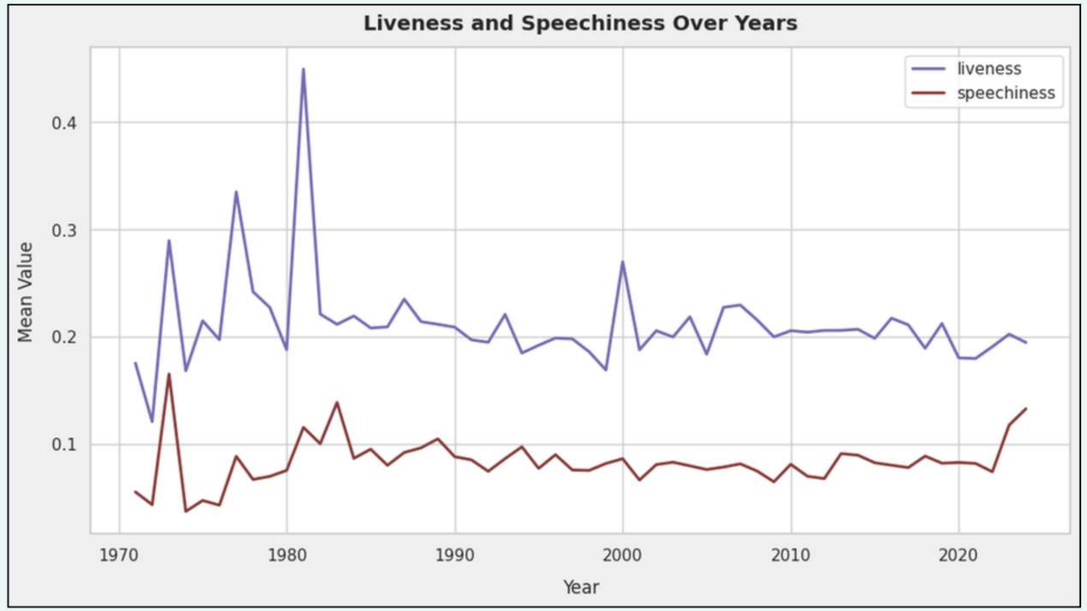


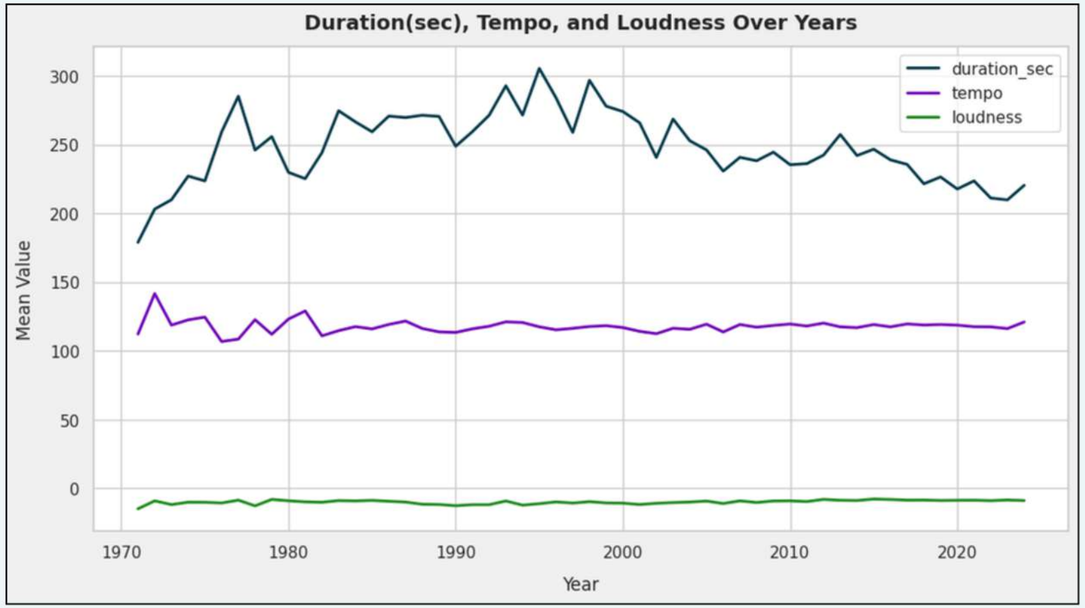

- **Mean popularity drifts downward from late-70s highs and stabilizes mid-2010s :** Average popularity peaks in the mid-to-late 1970s, declines through the 80s–90s, and hovers in the mid-teens since 2010 despite surging output, suggesting fragmentation and long-tail effects in streaming eras.

# Numerical KPI’s Over Years



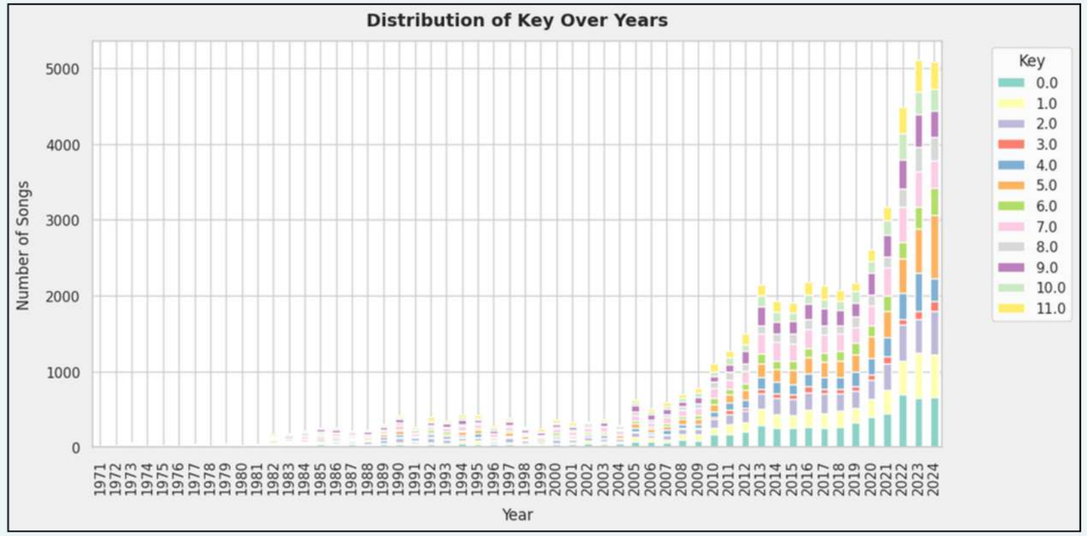



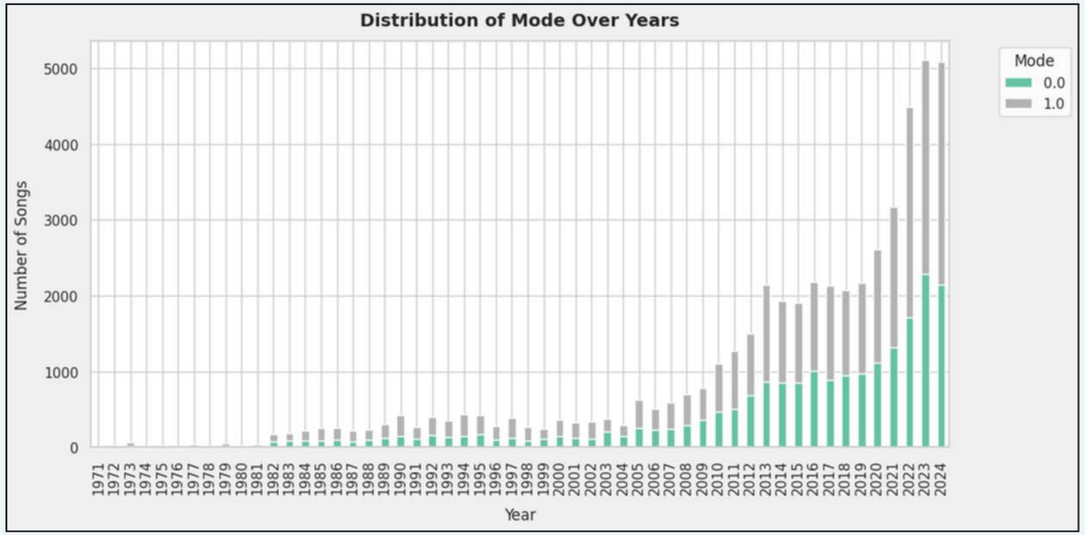



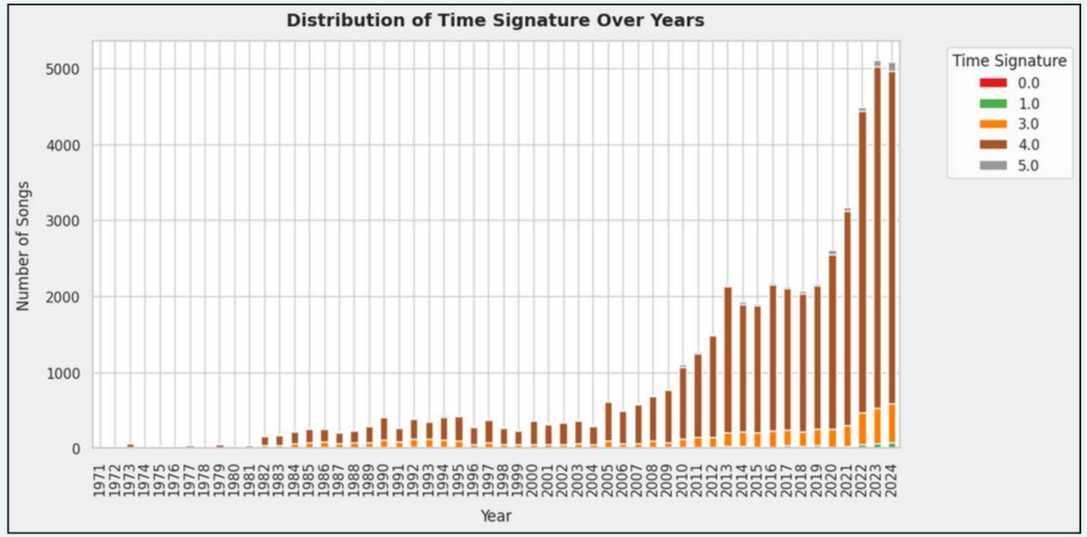
# Overview

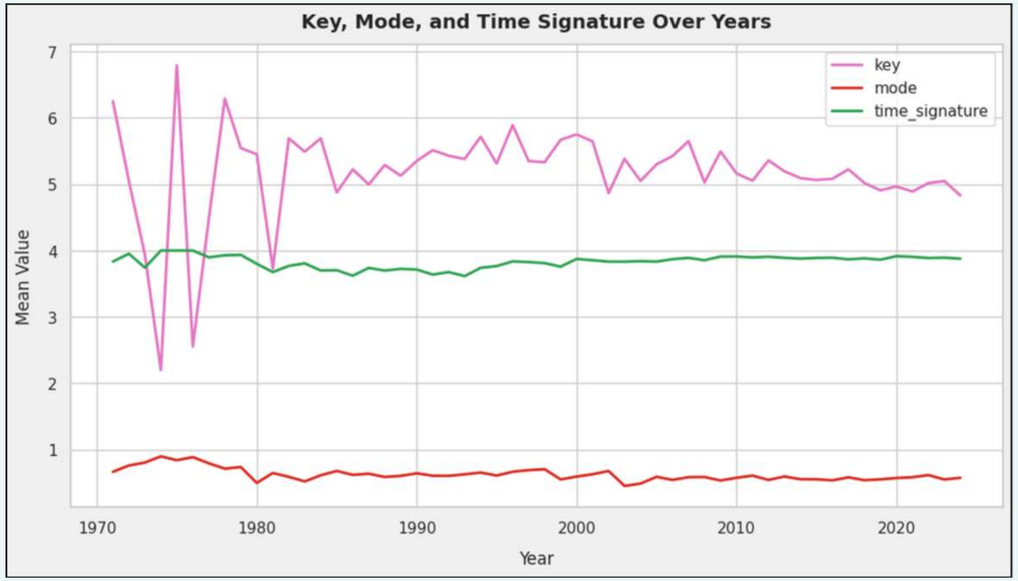
- **Songs have become shorter, slightly faster, and louder since the 1990s :** Average duration peaked in the late 90s then trended down, tempo holds in the 110–120 BPM band with a small uptick post-2010, and average loudness has become less negative, reflecting modern “hotter” masters.
- **Texture shifted from acoustic to produced, while vocals remained central :** **Acousticness fell steadily from the 90s onward** as electronic/processed production rose; instrumentalness rose into the 2010s but remains well below vocal norms, so mainstream continues to be vocal-led.
- **Movement and intensity are resilient; “happiness” softened :** Danceability stays stable to slightly rising, energy trends upward from the 2000s, whereas valence drifts downward from 90s highs—**today’s hits are energetic and danceable without needing bright mood.**
- **Live feel is flat-to-down, speechiness edges up recently :** Liveness has gradually leveled or dipped, consistent with studio-polished releases; **speechiness shows a late uptick after 2020, hinting at more talk/rap elements in recent tracks.**

# Numerical KPI’s Over Years



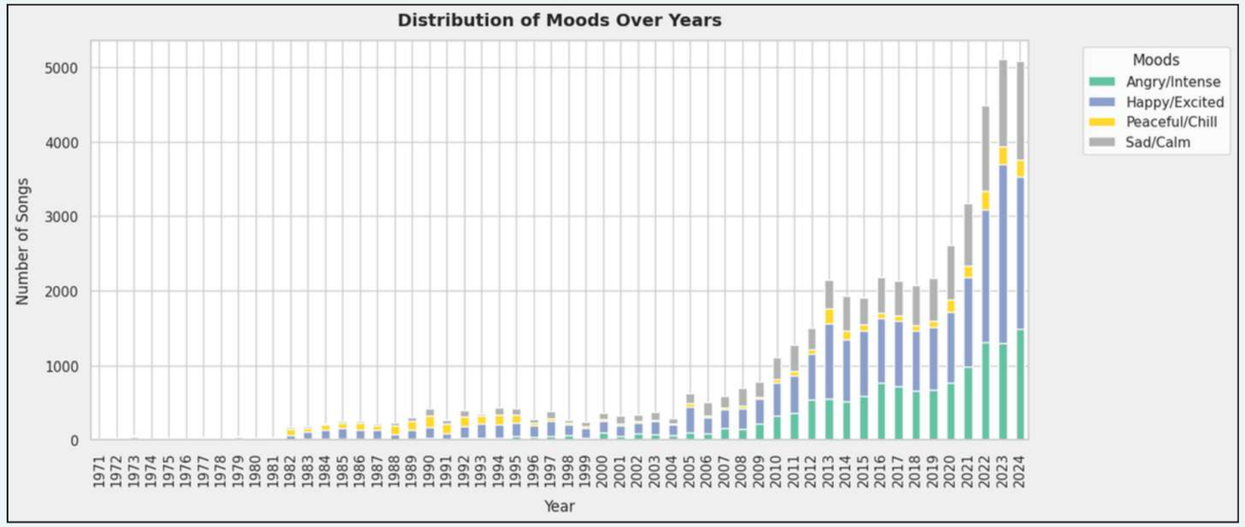



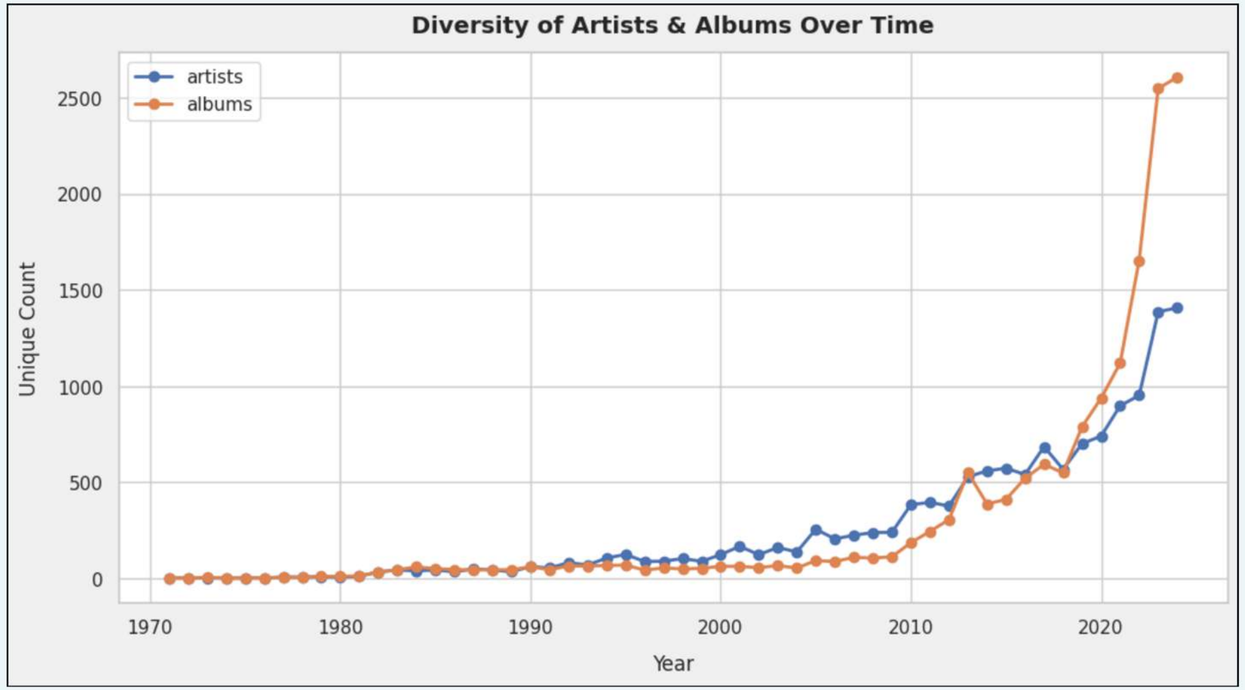



# Overview

- **Keys are broadly diversified but stable in aggregate :** All 12 pitch classes are represented each year with no decisive migration, and the mean key index stays near mid-scale over time—hitmaking doesn't hinge on specific keys.
- **Time signature is overwhelmingly 4/4 :** The 4/4 bar dominates every year with only small growth in 3/4 and 5/4, confirming that **rhythmic structure remains conventional despite genre changes.**
- **Major vs minor both scale with output :** Counts in both modes rise with overall volume, and the mean mode hovers around the major side without a strong directional trend—mood color from mode is secondary to production and groove.

# Categorical KPI's Over Years

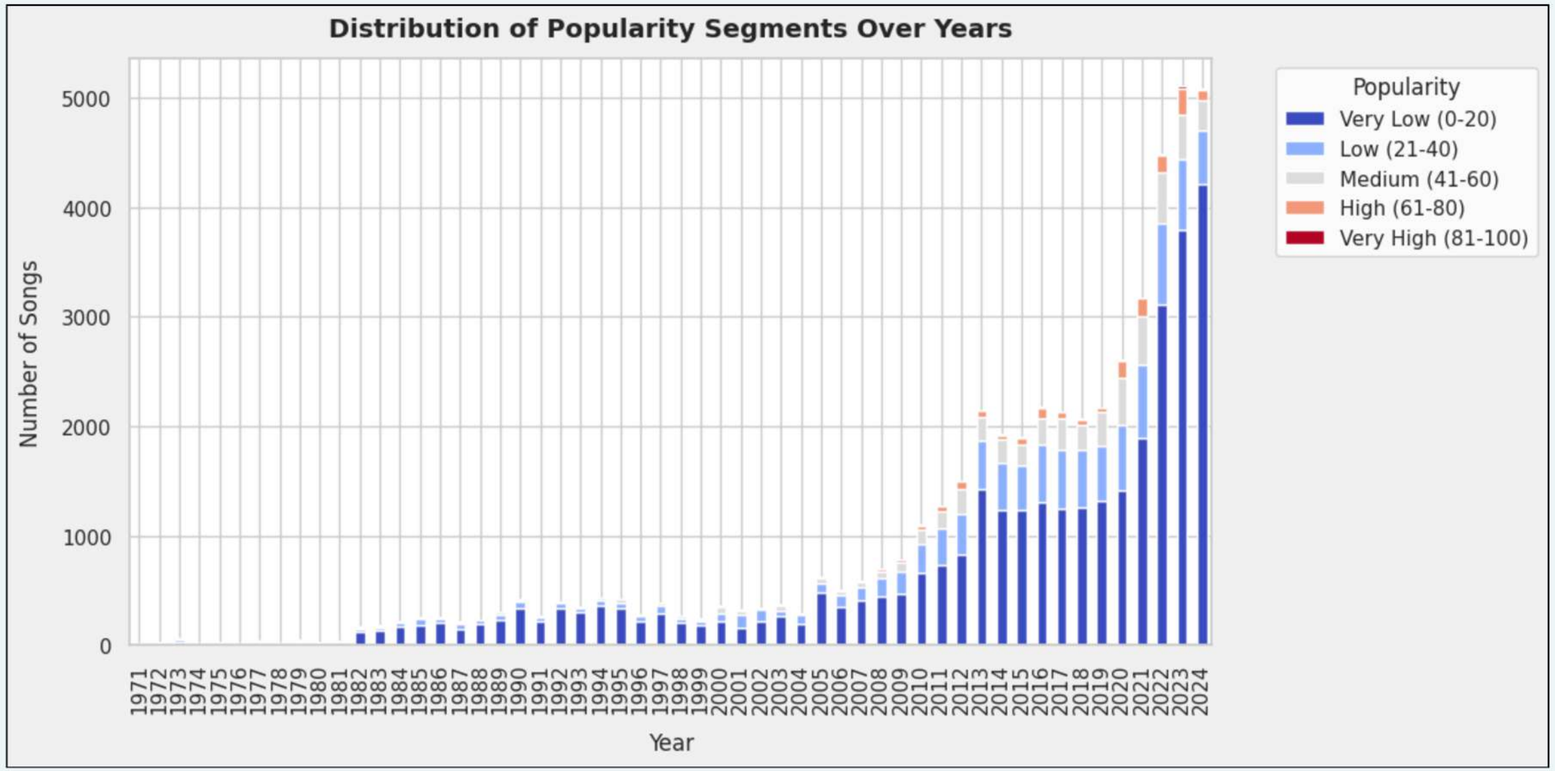



- **Catalog breadth has exploded since 2010 :** unique artists and albums grow super-linearly, with albums surging especially after 2021—supply diversity is rising faster than artist headcount, signaling prolific multi-release strategies and label catalog stuffing.

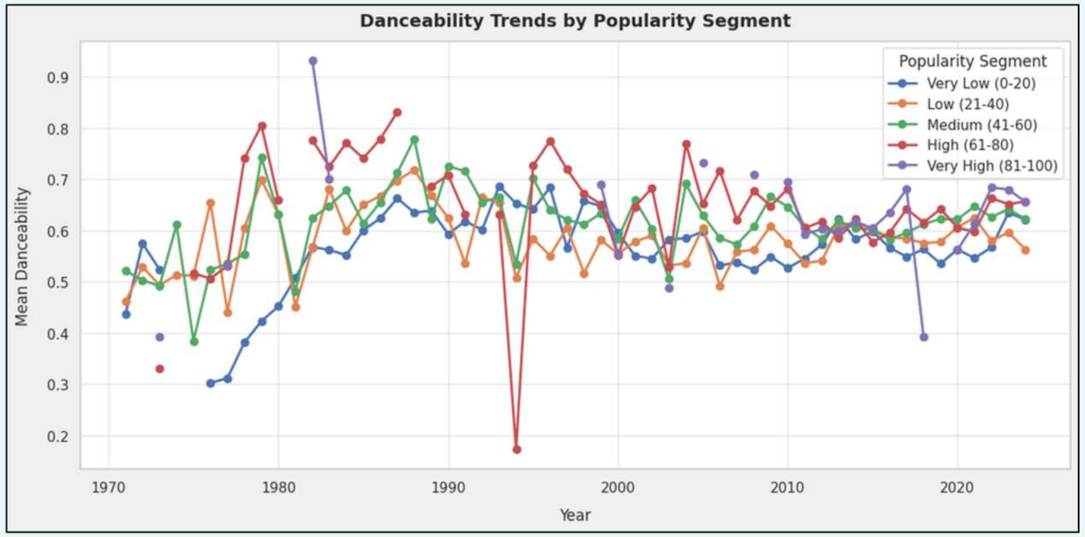
- **Mood mix scales with output rather than rotating dominance :** Happy/Excited and Angry/Intense grow the most in count, while Peaceful/Chill remains a thinner slice; this supports prioritizing energetic and upbeat programming as the catalog expands.

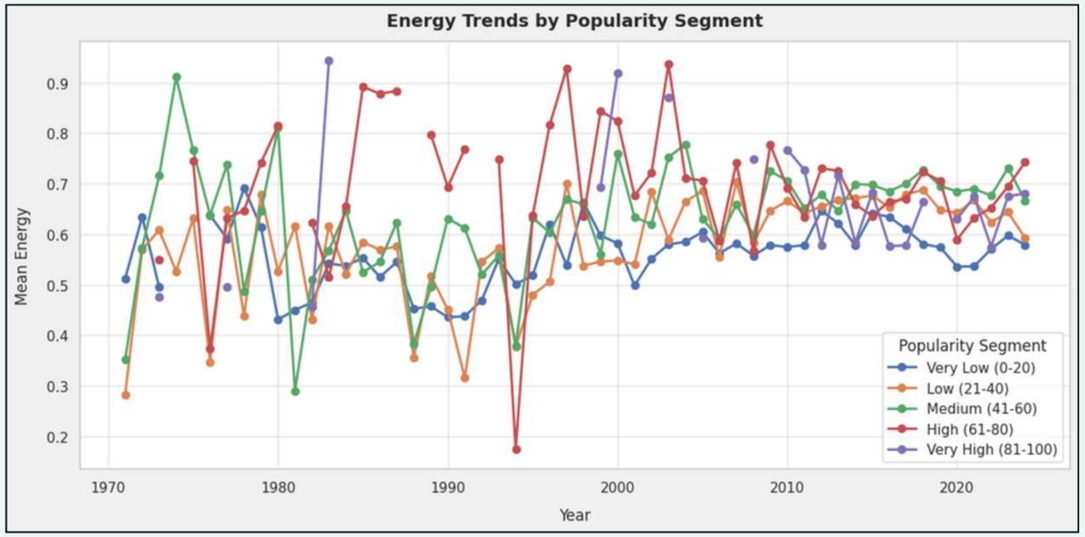
# Analysis Of Popularity Segments Over Years

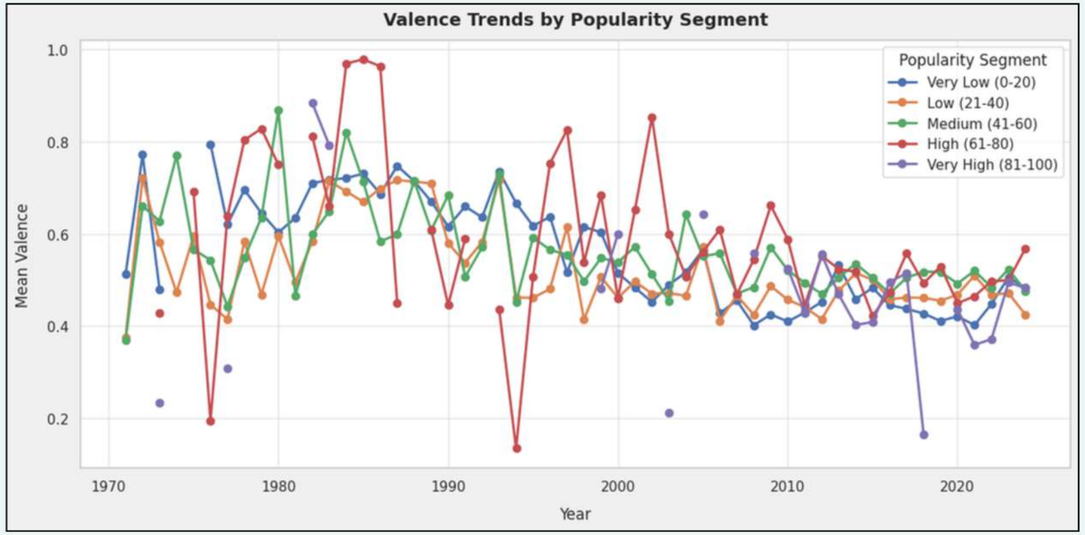



- **Volume exploded after 2005 but quality concentration didn't :** The count of released **tracks per year rises sharply post-2005** with especially large growth since 2015, **yet most additions sit in the very-low and low popularity buckets,** indicating catalog expansion outpacing hit density.

# Analysis Of Popularity Segments Over Years

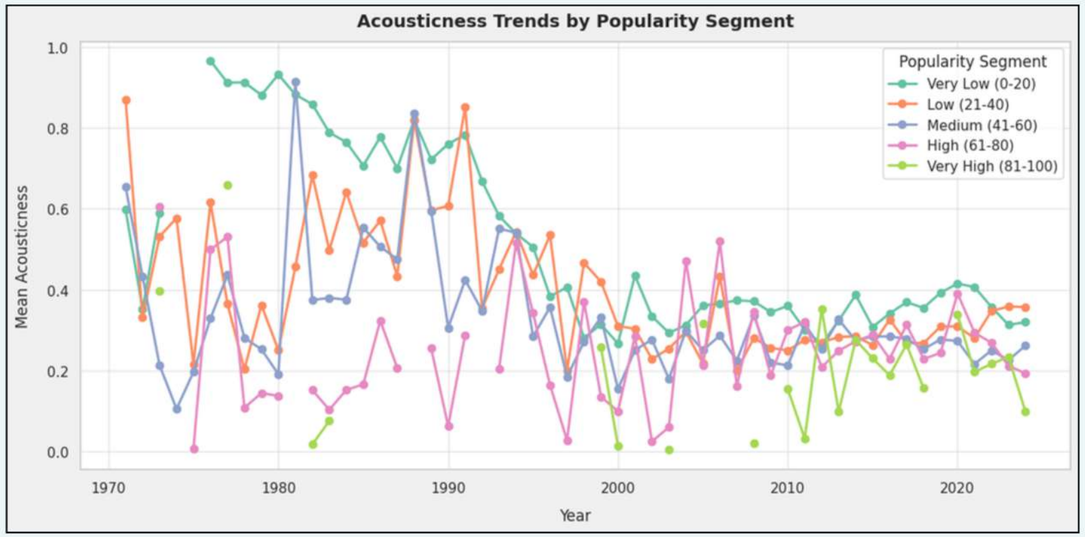


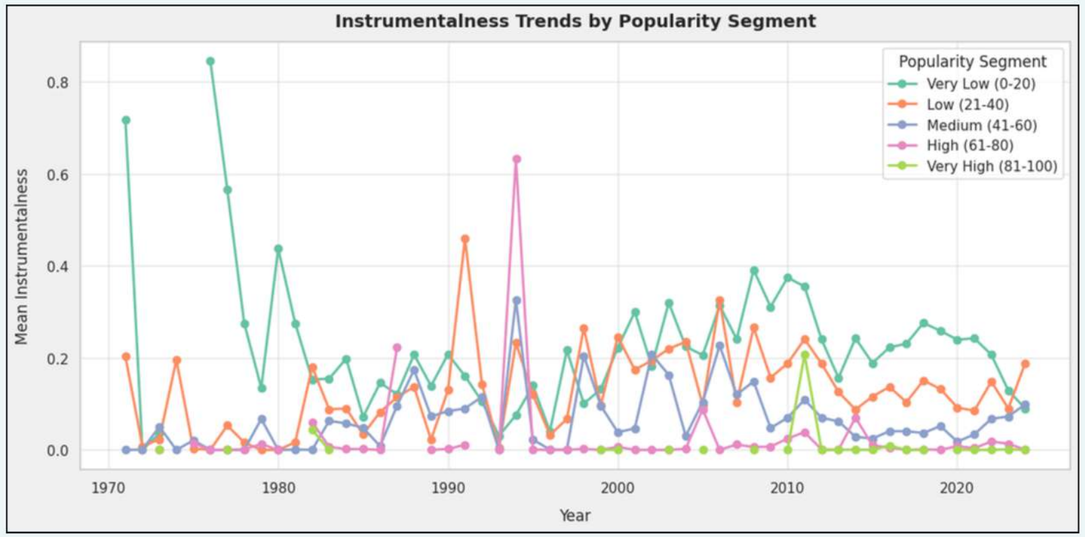



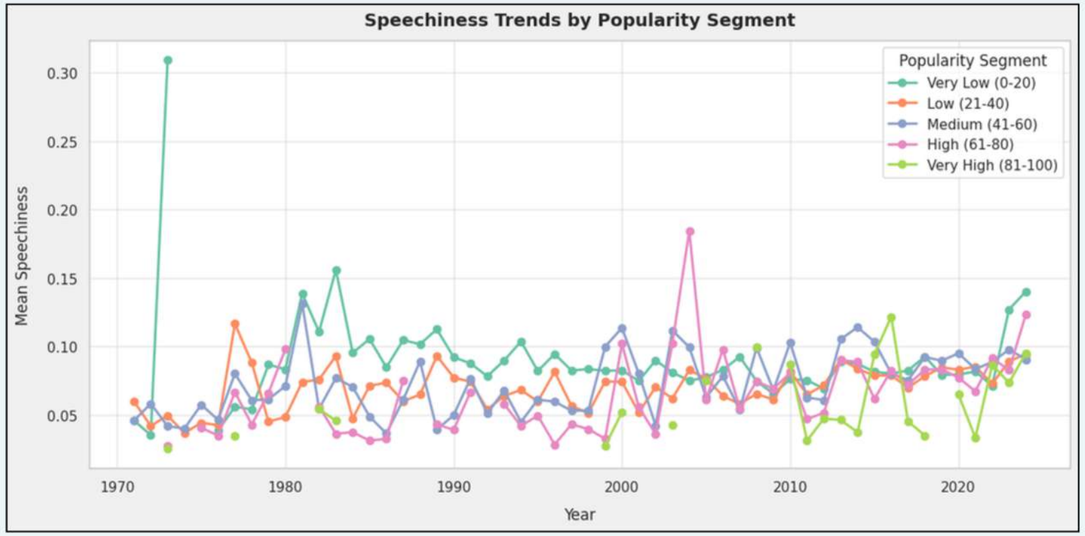




- Across all segments, energy trends upward since the 2000s while valence trends gently down, indicating modern hits win on intensity rather than brightness; danceability stays relatively stable to slightly rising, reinforcing movement as a constant.
- **Segment gaps narrow in recent years :** By the 2010s–2020s, Very High, High, and Medium segments converge around energy ~0.65–0.7 and danceability ~0.6–0.65, suggesting format homogeneity at the top and less differentiation by these features alone.
- **Early-era volatility vs. present consistency :** 1970s–1990s curves show large swings across segments in all three metrics, whereas 2010s onward are smoother—production standardization and playlist optimization likely compress feature variance.

# Analysis Of Popularity Segments Over Years

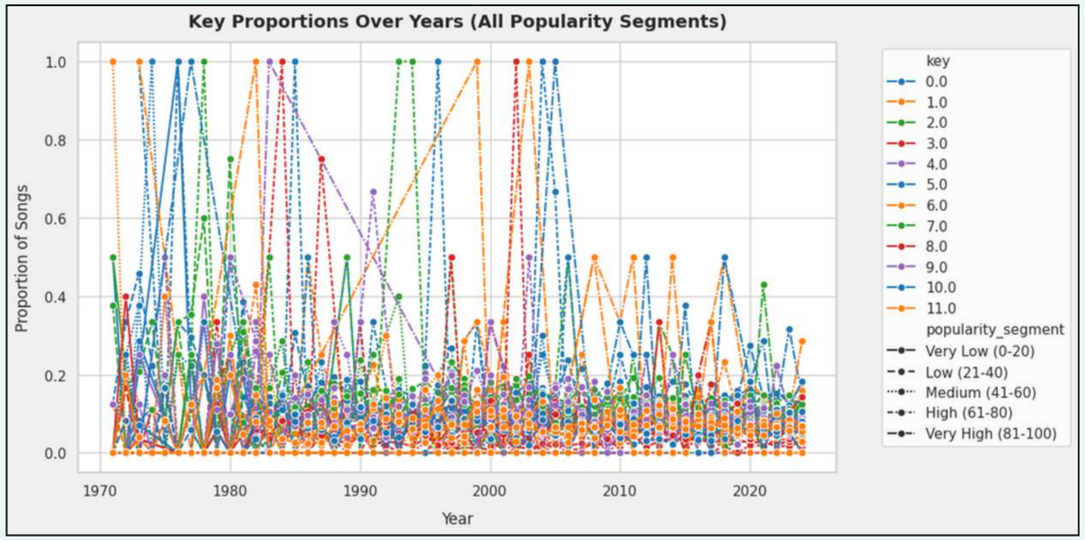


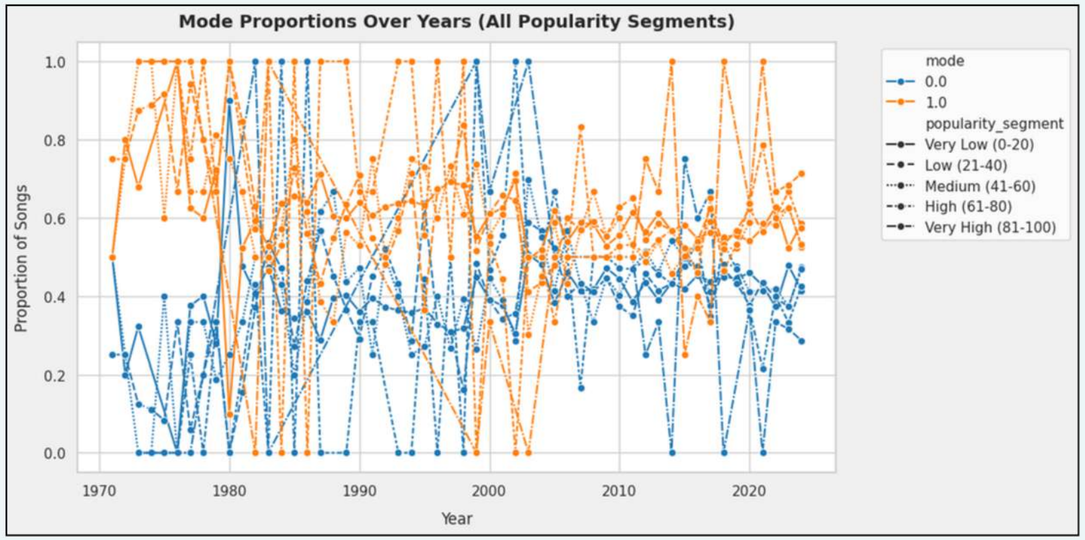



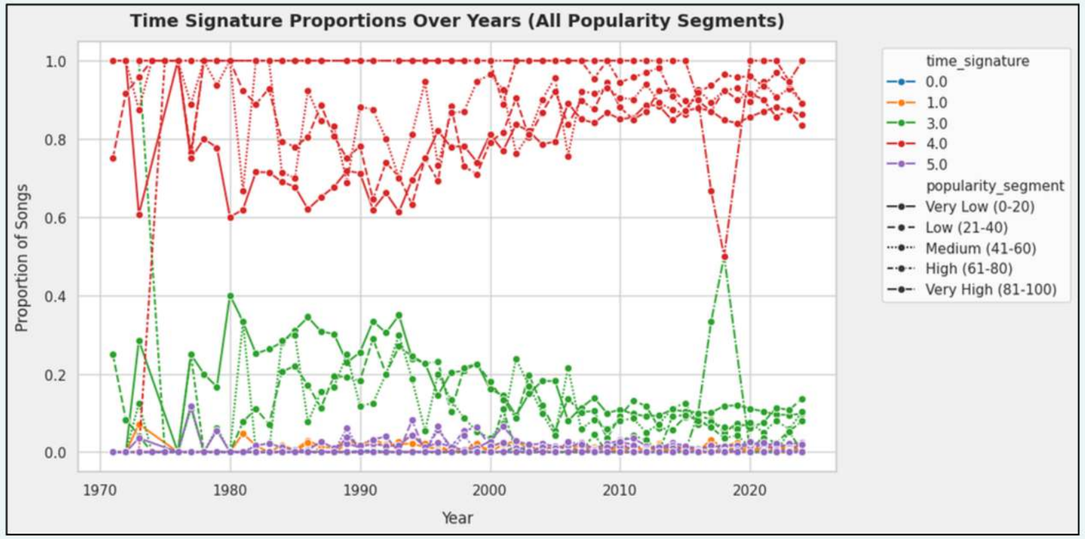




- **Speechiness up, instrumentals down at the top :** Since the 2000s, Very High and High segments maintain low instrumentalness while showing mild recent upticks in speechiness, indicating vocal- and rap-leaning hits, whereas very-low segments carry most instrumental content.
- **Acousticness convergence :** All segments trend toward lower acousticness from the 90s, then stabilize around ~0.25–0.35 post-2010, reflecting consistent, polished, non-acoustic production across popularity tiers.
- **Segment separation signal :** In recent years, High/Very High tracks combine low instrumentalness with moderate acousticness and slightly elevated speechiness relative to Medium/Low—use this mix when optimizing for mainstream success.

# Analysis Of Popularity Segments Over Years

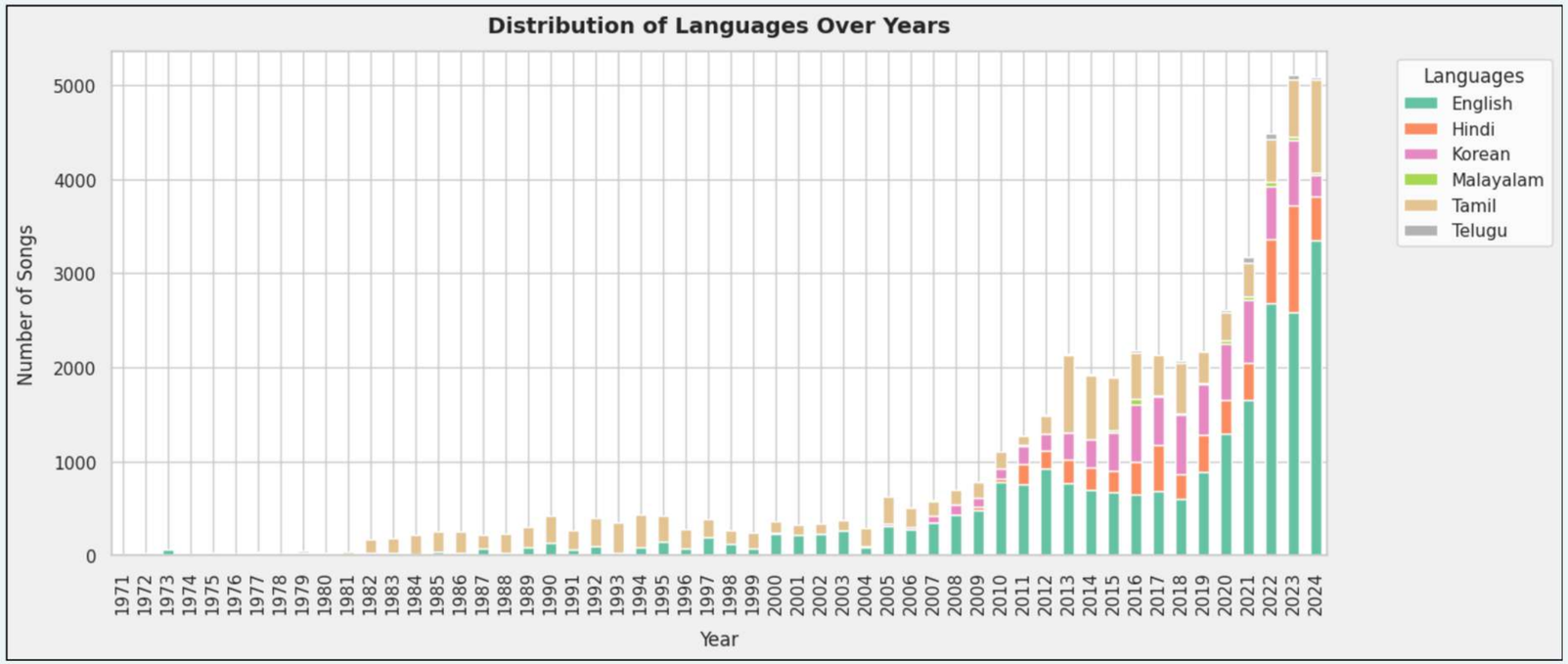


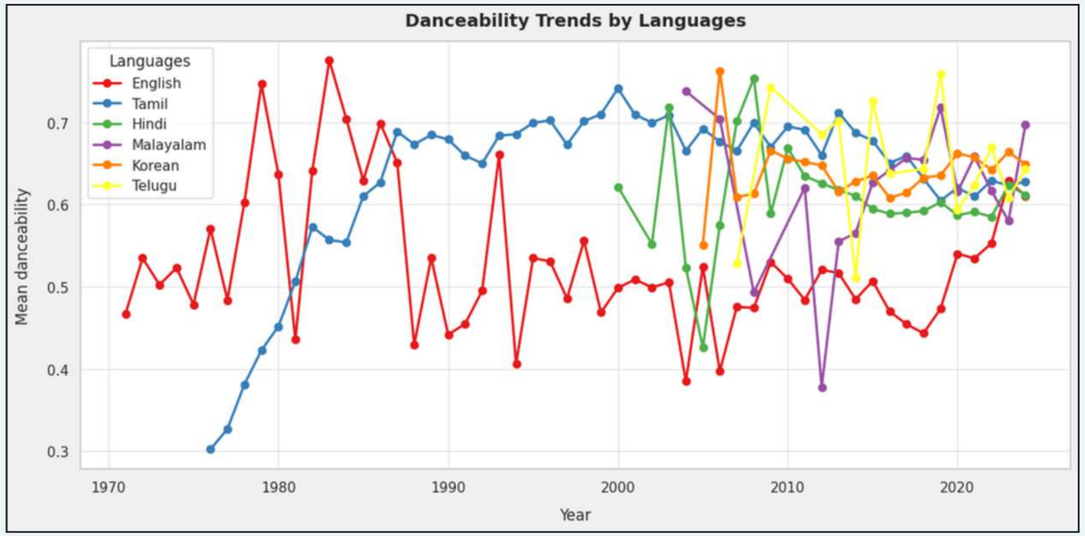

- **Key proportions show no dominance :** All keys split relatively evenly within each popularity segment across years—no single key greatly outpaces others, and this holds true from lower to very high popularity, signaling creative tonal variety in hits.
- **4/4 time signature is universal :** Over 80% of songs in every popularity bracket—especially medium, high, and very high—are in 4/4, with only minor representation for 3/4 and rare 5/4, confirming the centrality of standard meter in popular music.
- **Major/minor mode doesn't predict popularity :** Both modes (major = 1.0, minor = 0.0) are present in all popularity groups each year, with neither consistently leading—suggesting mode alone isn't a differentiator for mainstream appeal or niche success.

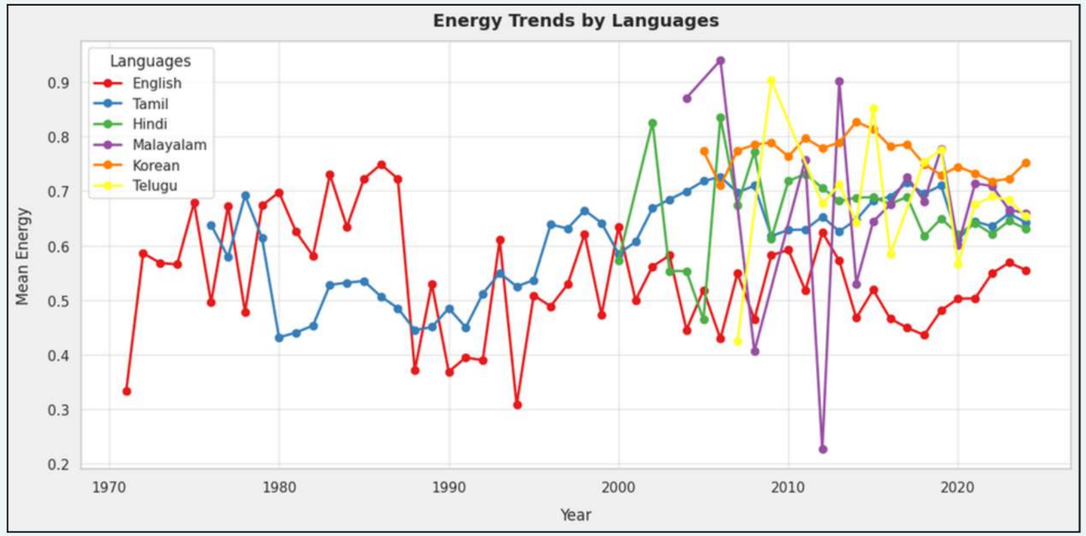
# Analysis Of Languages Over Years

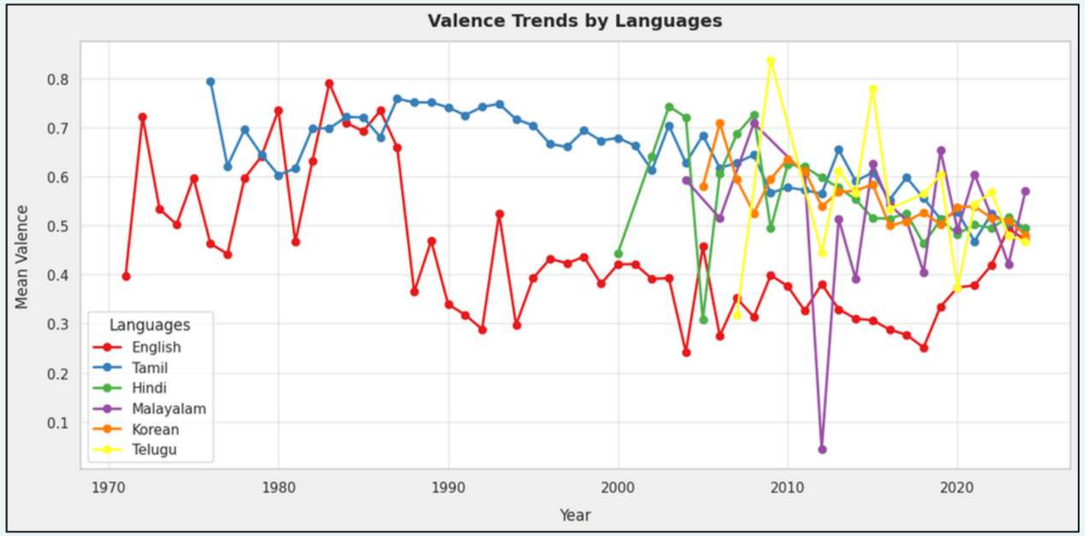



- Language expansion is led by English, then rapid gains from Hindi, Korean, and Tamil post-2015—global crossover is real, but English still contributes the largest absolute volume each year.

# Analysis Of Languages Over Years

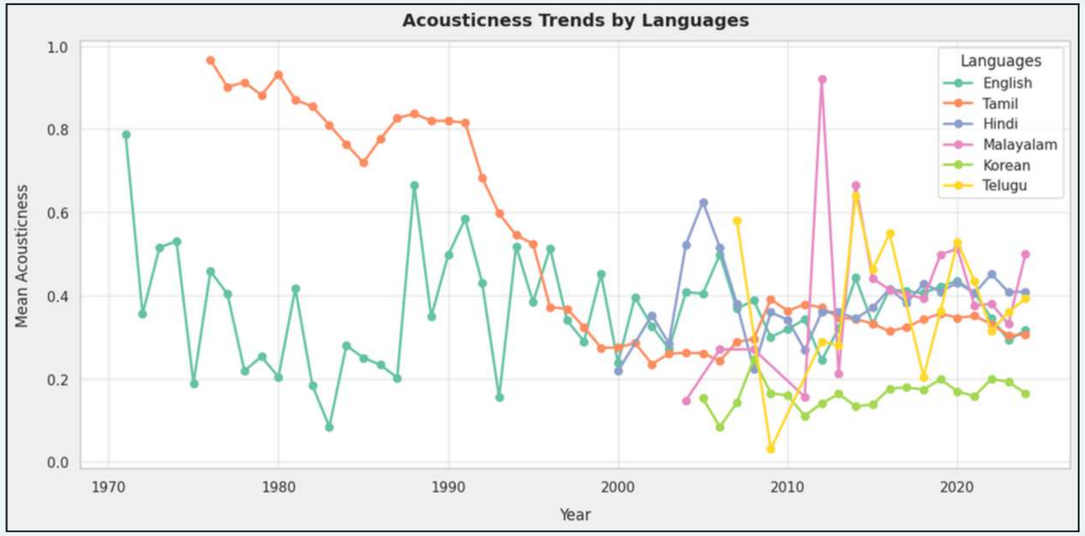


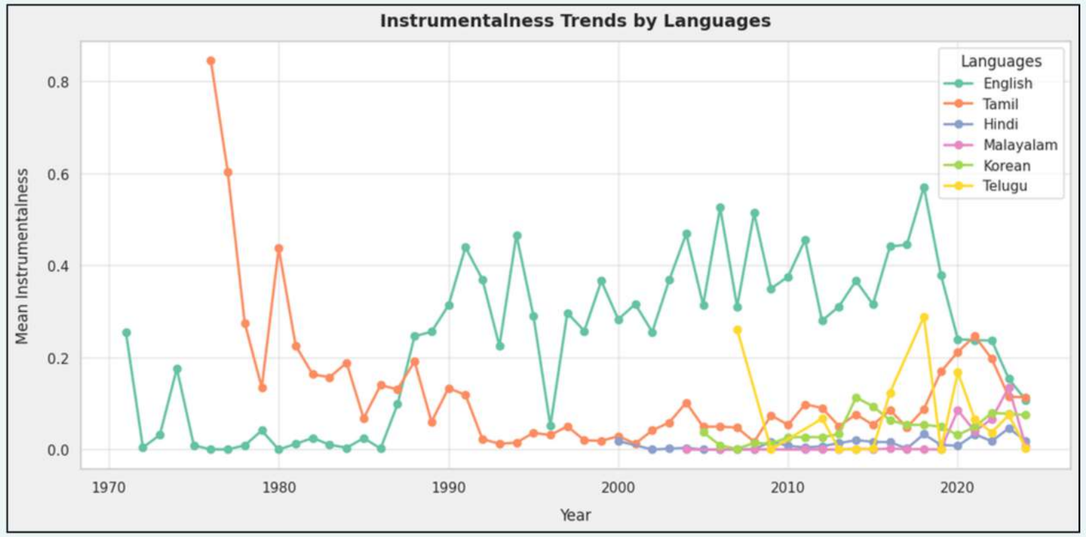



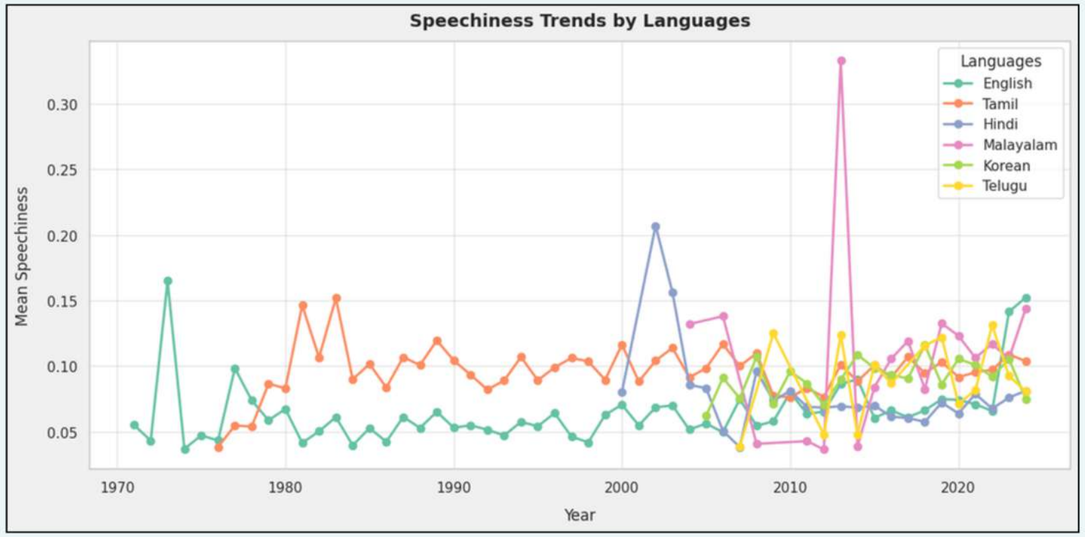




- **High-energy leaders :** Korean and Telugu maintain the highest recent energy (>0.7 mid-2010s; ~0.65–0.75 in 2020s), with Malayalam spiking earlier but normalizing—use these for hype and workout sets.
- **Danceability edge :** Tamil stabilizes around ~0.65–0.7 since the 1990s, with Korean and Telugu close behind post-2010; English trails at ~0.5–0.6—Tamil is the safest bet for sustained movement.
- **Valence split :** English trends lower in valence since the 2000s (0.3–0.5), while Tamil stays higher (0.6–0.7 historically) and Korean/Telugu cluster mid (~0.5–0.6)—pair English for moody energy, Tamil for feel-good, and Korean/Telugu for balanced uplift.

# Analysis Of Languages Over Years

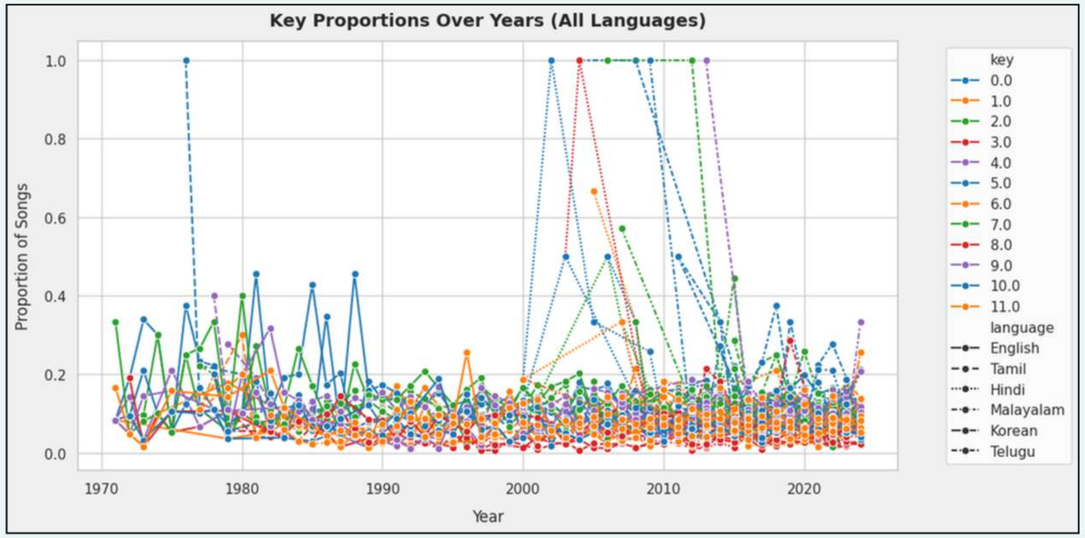


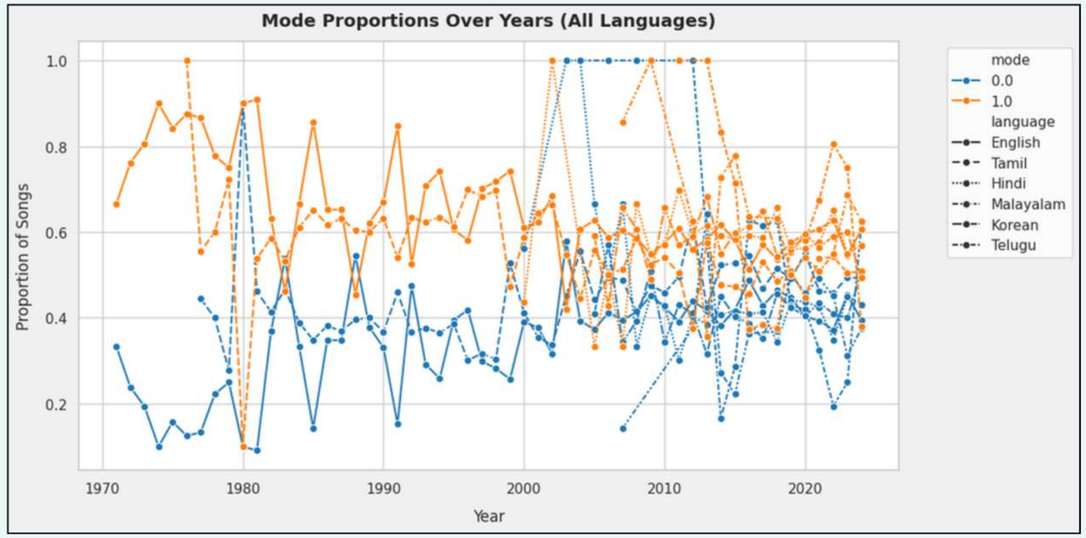



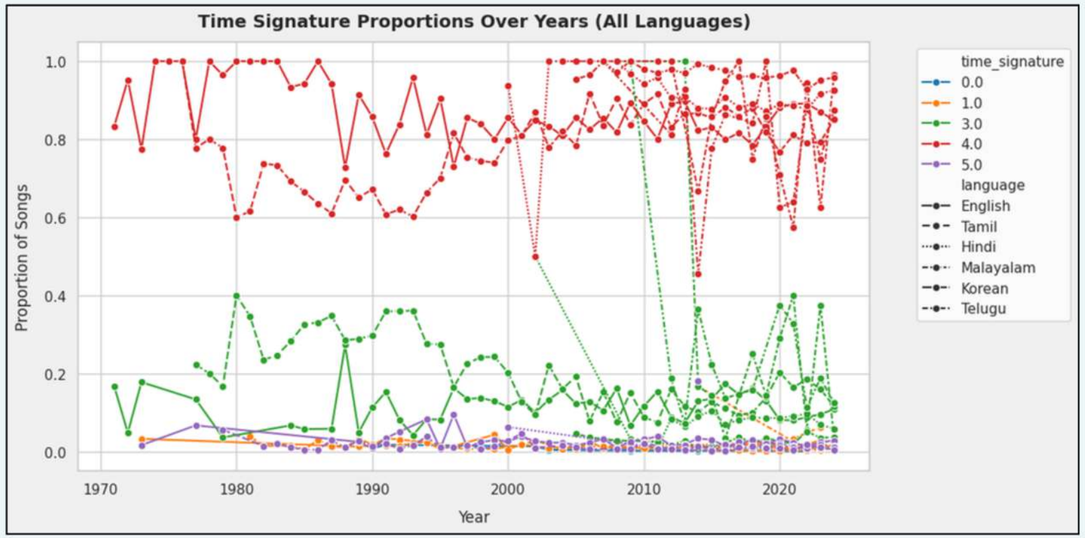




- **Rap/talk elements are rising across the board :** Since ~2010, speechiness drifts upward in Korean, Telugu, and Malayalam while English/Tamil stay mid-low; Hindi shows an early-2000s spike but normalizes—expect more vocal presence and rap features in non-English catalogs.
- **Instrumentals are rare outside English and trending lower :** English has the highest and most volatile instrumentalness, peaking around the 2010s, whereas Tamil/Hindi/Korean remain near zero—**vocal-led tracks dominate Asian-language hits.**
- **Production convergence on low acousticness :** Tamil falls from very high acousticness in the 80s–90s toward 0.3–0.4, English sits mid (0.3–0.5), and Korean is consistently low (~0.15–0.25); recent Telugu/Malayalam fluctuate but center near modern, processed textures—use Korean for slick electronic polish, Tamil for balanced organic-modern blends.

# Analysis Of Languages Over Years



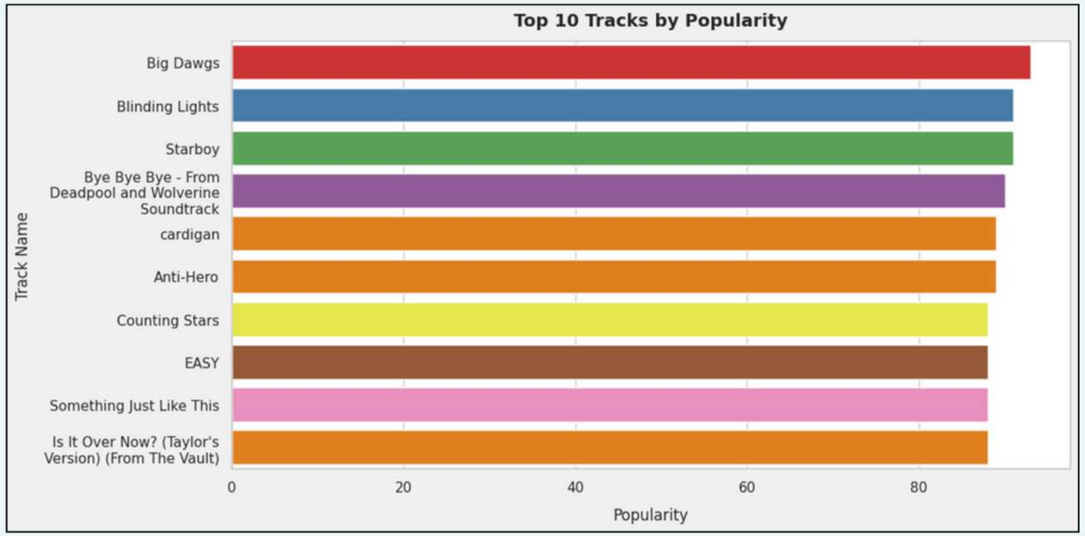



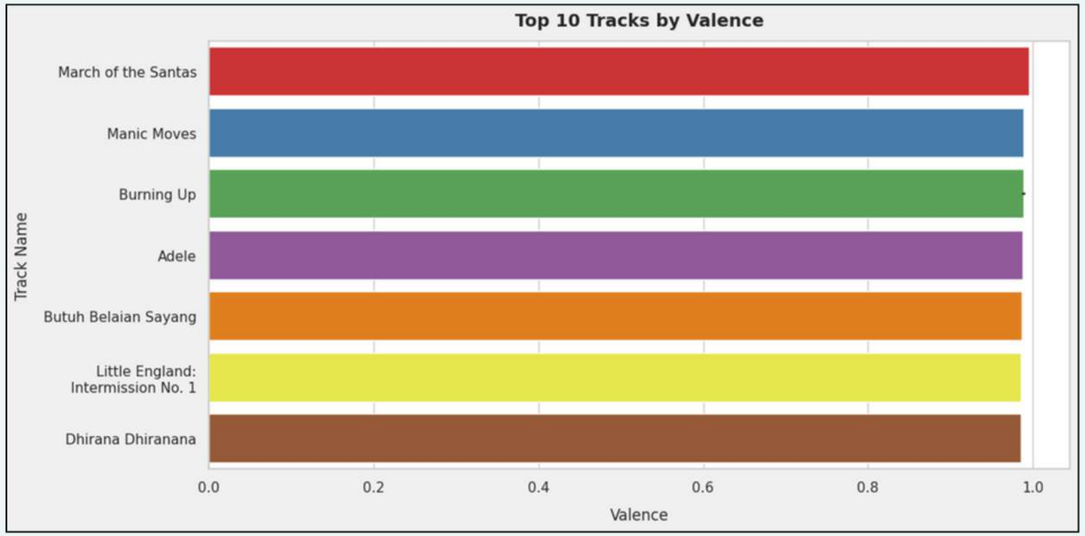

- **No language or era dominates key choice :** The proportional use of each key remains widely distributed among all languages with no clear long-term bias; each language's catalog is musically diverse, and there's no "hitmaking" key in any language.
- **Major mode (mode=1) is favored, but the split is stable :** Across all languages, about 55–65% of songs tend to be in major, with the remainder in minor, and this balance remains steady since the 1990s—feel-good and moody tracks coexist without dramatic shifts in preference.
- **Universal dominance of 4/4 :** All languages overwhelmingly default to 4/4 time, with minor but persistent use of 3/4 and 5/4 in some catalogs; rhythm conventions are global, forming the backbone of mainstream music regardless of language.

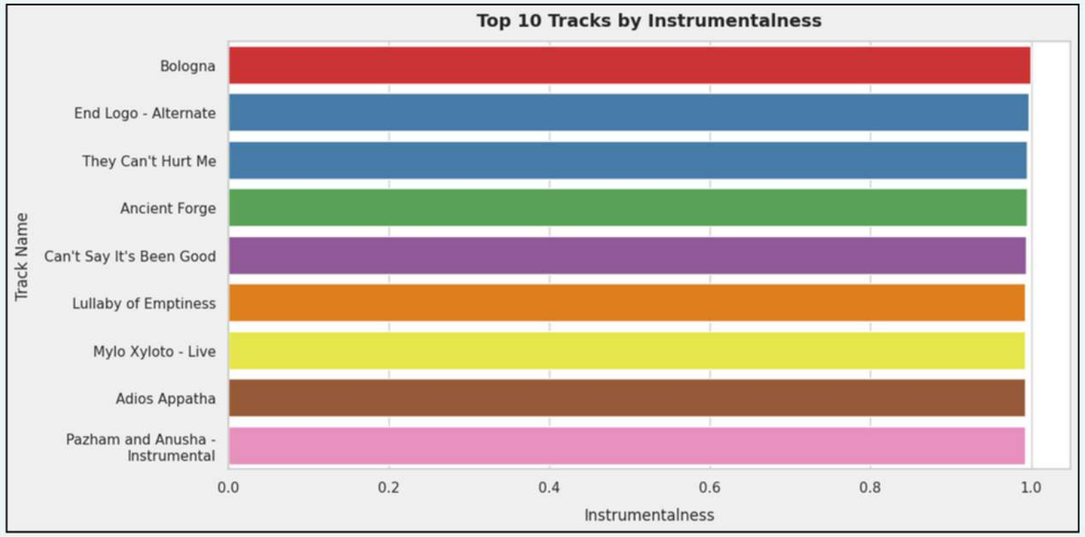
Spotify

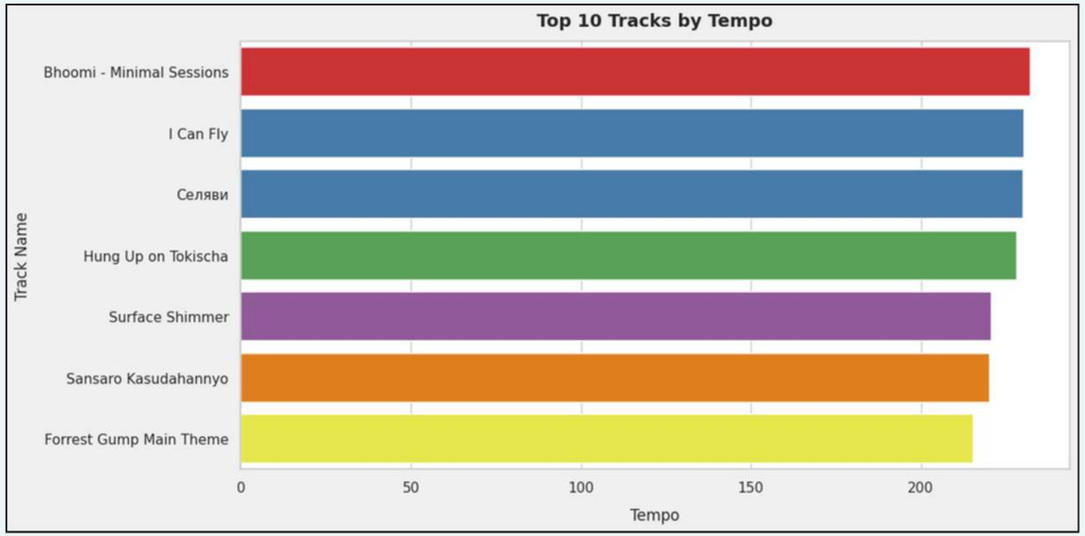
# *Top N & Bottom N Analysis*

# Top 10 Tracks By Popularity, Valence, Instrumentalness & Tempo



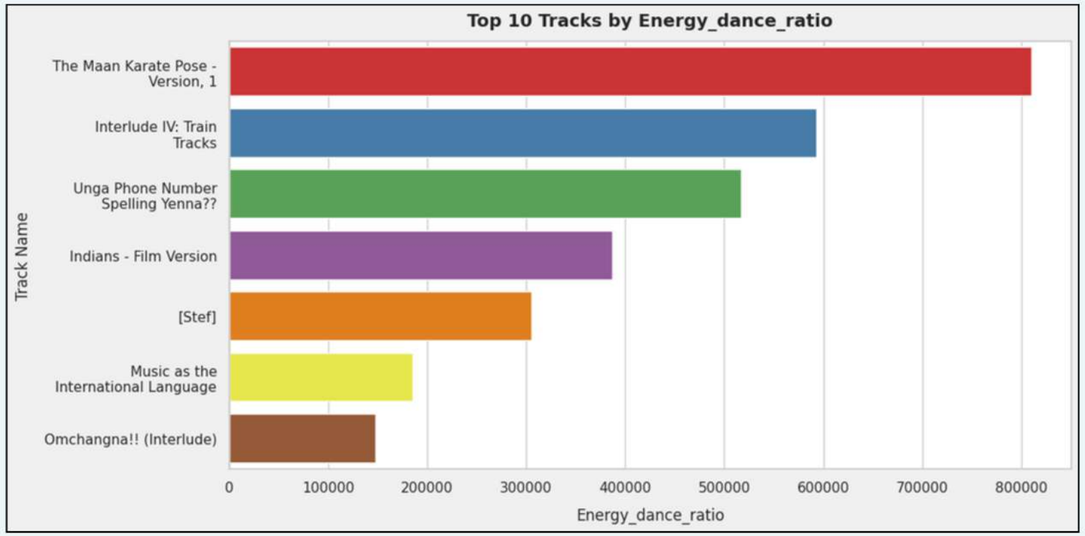



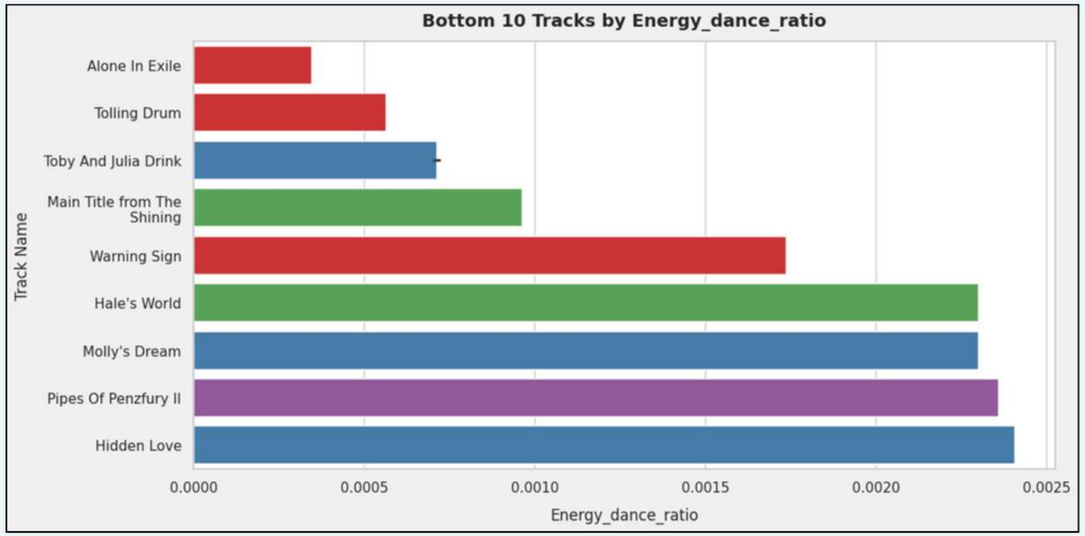



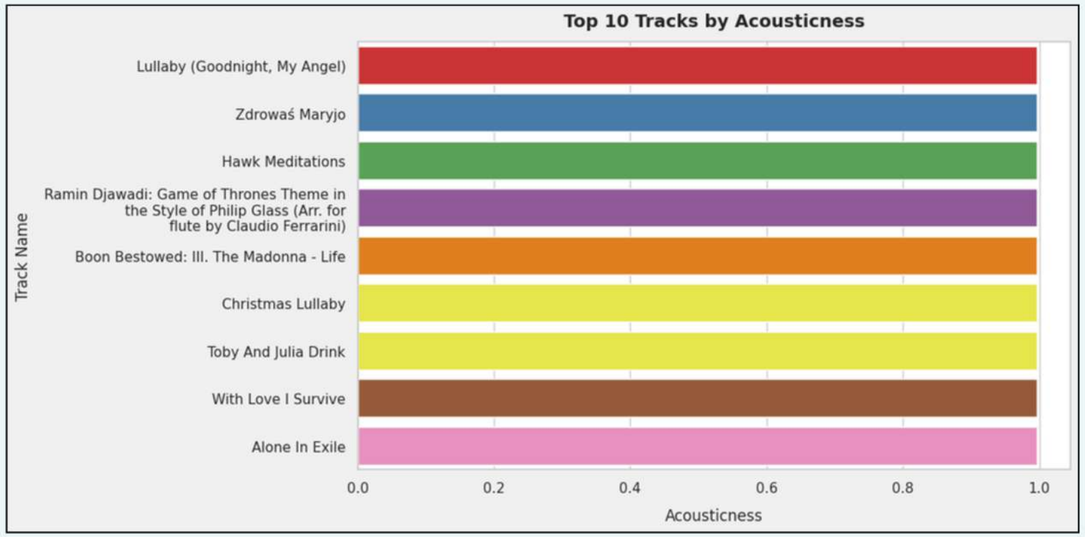
# Overview

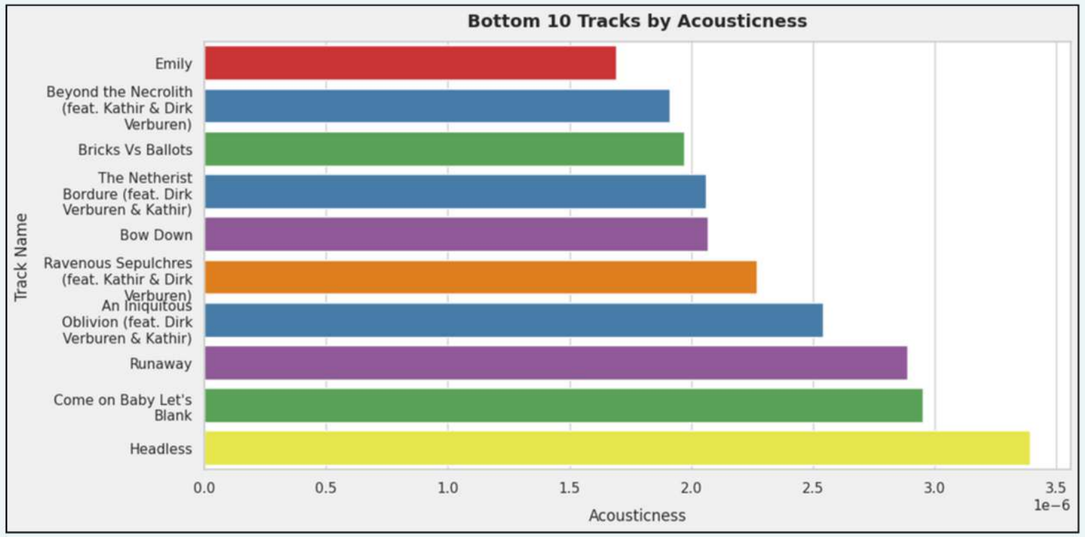
- The top tracks for instrumentalness are purely or almost entirely instrumental, with scores near 1.0, indicating no or minimal vocals.
- The top tracks for valence are consistently near the maximum (1.0), reflecting extreme happiness or positivity in musical mood.
- Tracks ranking highest in tempo are uniformly fast-paced, each with tempos exceeding 200 BPM, making them especially energetic and suited for dance or action-oriented playlists.
- The most popular tracks list aligns with global chart success, featuring international hits and crowd favorites that maintain high, but not extreme, values across musical features—indicating broad mainstream appeal.

# Top & Bottom 10 Tracks By Energy-Dance Ratio & Acousticness



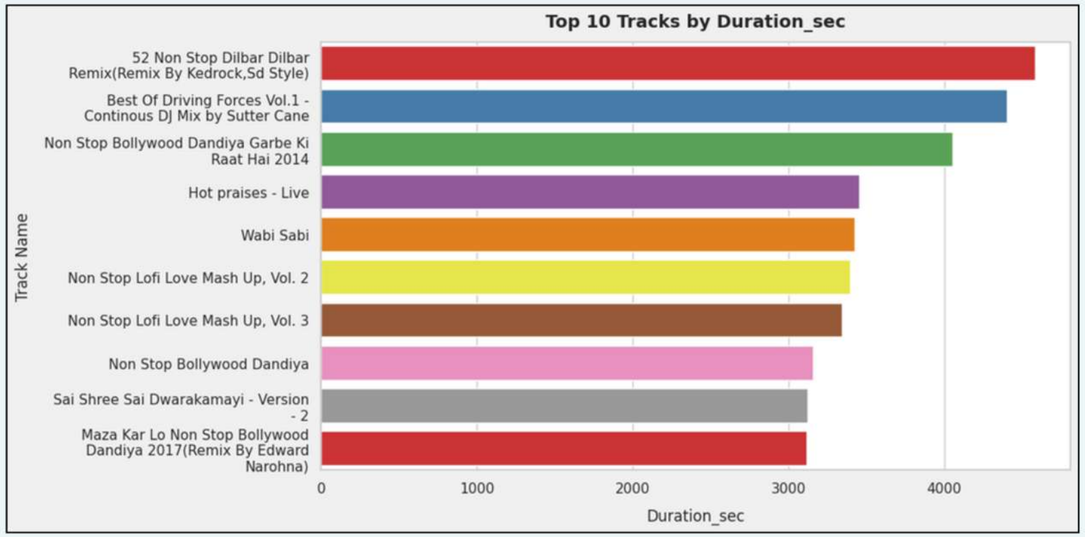



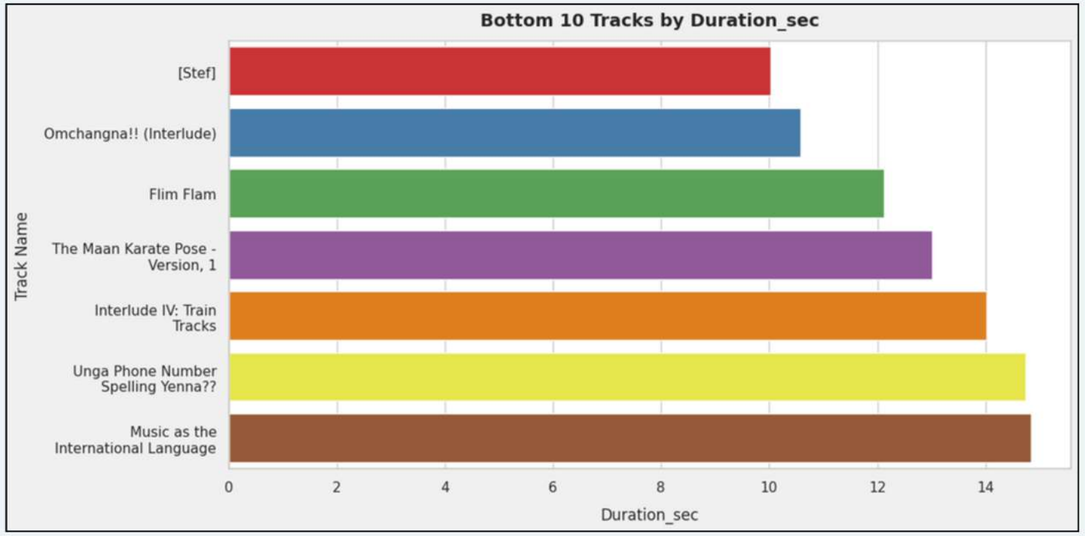



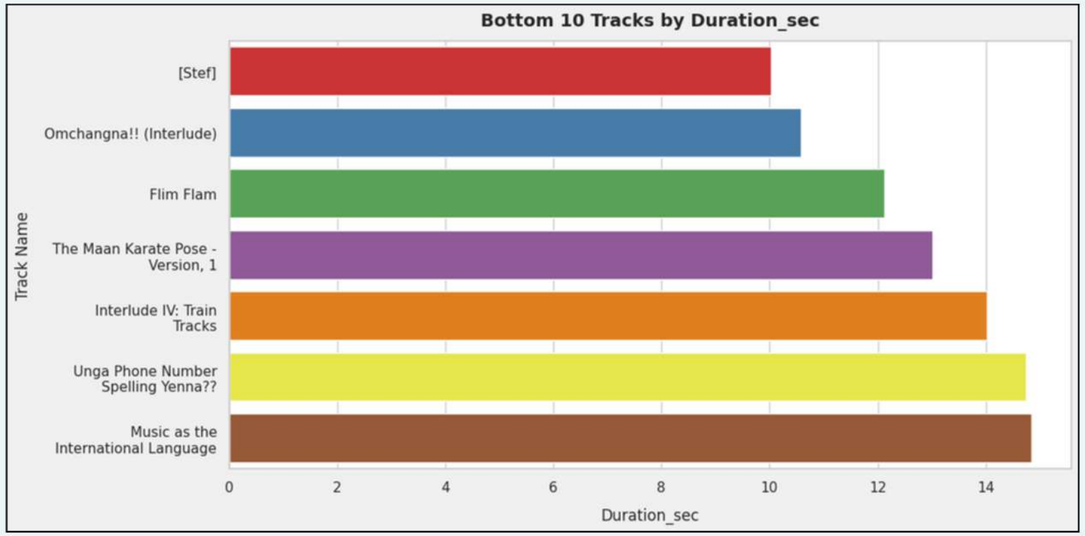
# Overview

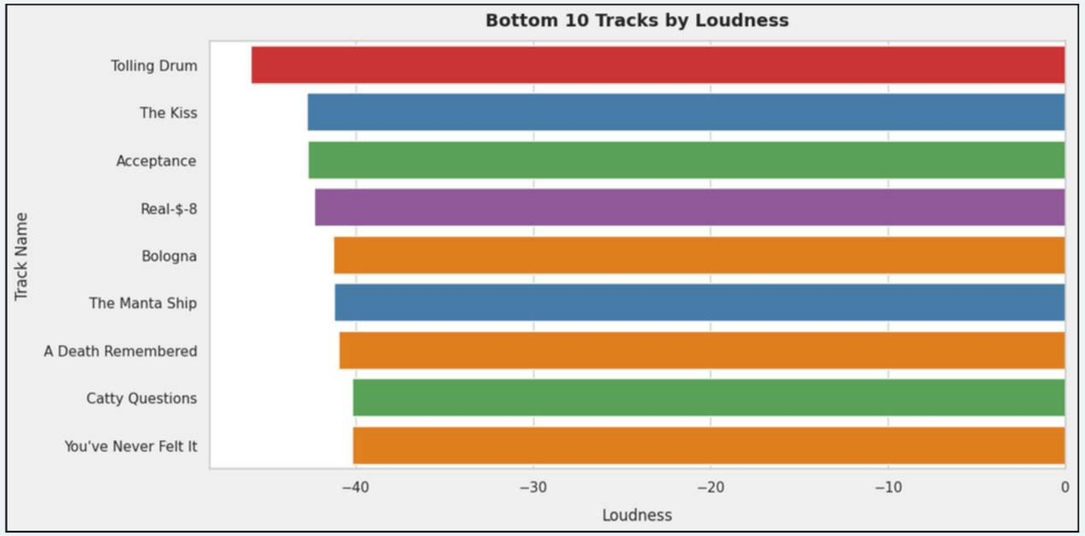
- Tracks with the lowest acousticness are highly electronic or synthetic, showing almost no “unplugged” qualities—these are typically modern, heavily-produced works in genres like electronic, industrial, or synthetic pop.
- The highest energy-to-danceability ratios belong to short tracks and interludes; these tracks have extremely concentrated bursts of energy relative to dance features, standing apart from typical songs where the ratio is far lower.
- Tracks with the lowest energy-to-danceability ratios are slow, atmospheric, or cinematic—these tend to be instrumental, lullaby, or film score tracks, where mood or ambiance dominates over movement.
- The most acoustic tracks are classical or meditative—flute, lullabies, or soft instrumental pieces, with “organic” sound and minimal production, ideal for relaxation, sleep, or introspection.

# Top & Bottom 10 Tracks By Duration & Loudness



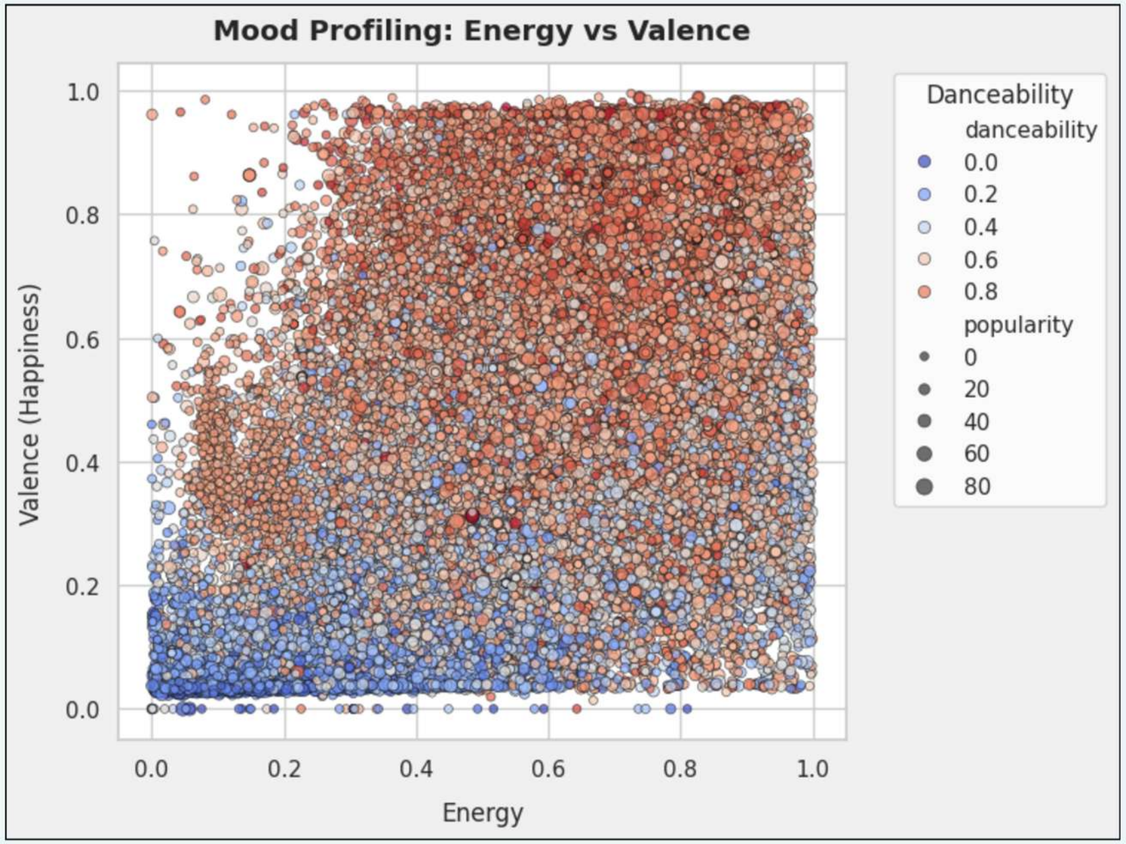
# Overview

- The loudest tracks range widely, with the highest reaching near maximum loudness and the lowest not far behind, illustrating some presence of loudness maximization in popular music for impact.
- The quietest tracks fall into very low loudness values (strong negative range), indicating ambient or minimally produced sounds, often soft or atmospheric, contrasting strongly with loud mainstream hits.
- Long-duration tracks are mostly remixes, live recordings, or continuous mashups, explaining unusual lengths well over typical song duration caps around 3-5 minutes.
- Very short tracks are almost all interludes, skits, or brief thematic pieces, showing a stark contrast with long tracks and illustrating diversity in album construction throughout the dataset.

Spotify

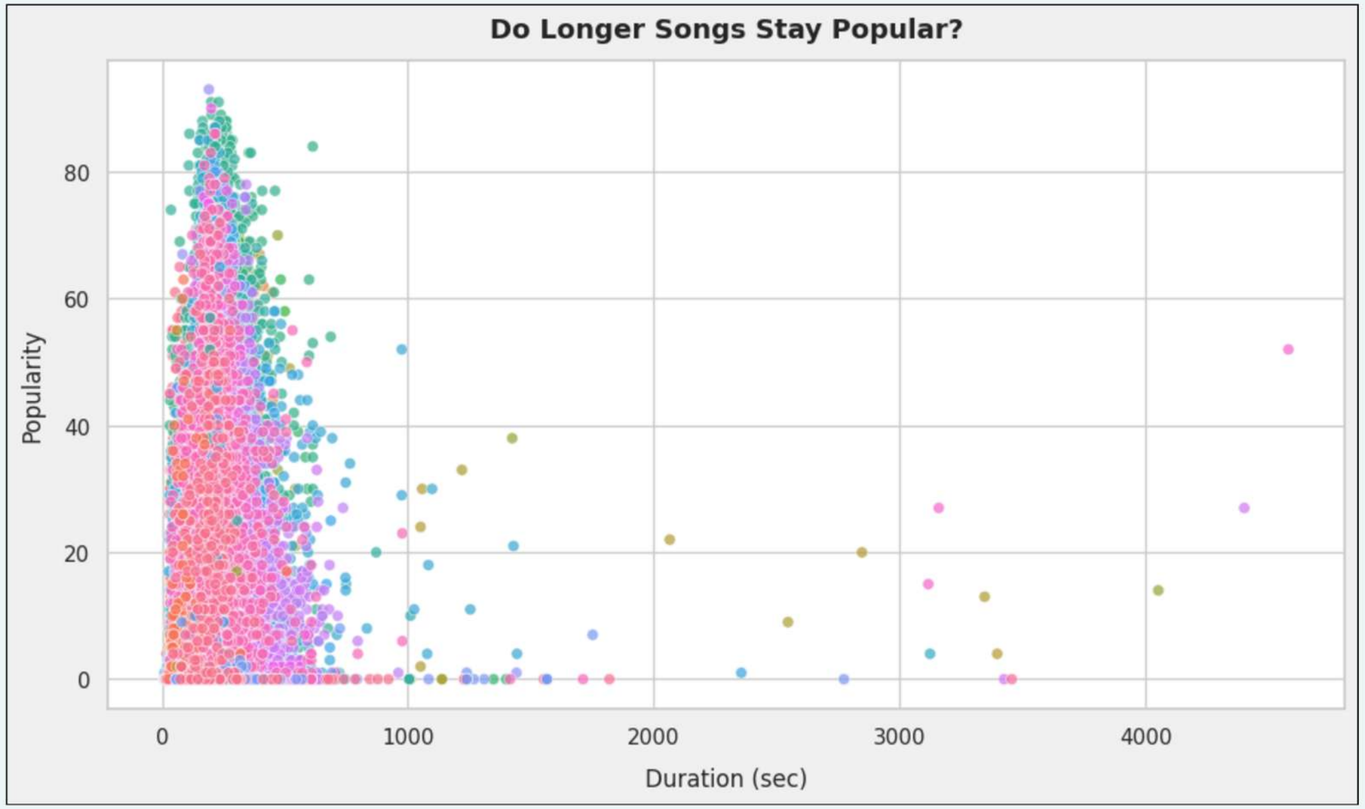
# *Some Hypothetical Analysis*

# Mood Profiling : Energy vs Valence



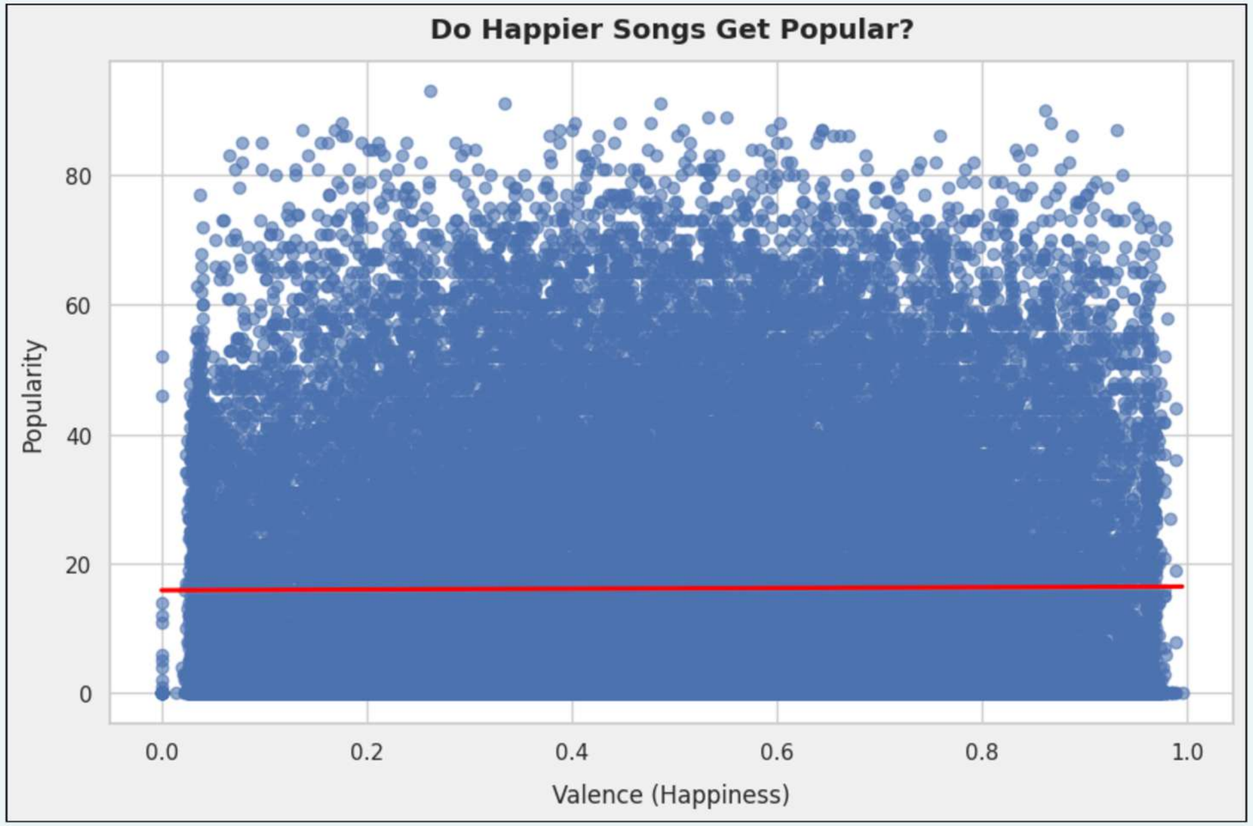

- Mood profile visualization shows the music catalog spanning the full ranges of energy and valence, with higher danceability tracks more densely populating the region of moderate to high energy and valence, reflecting a preference for upbeat and danceable tunes among popular songs.
- The plot confirms a wide distribution of tracks from low-energy, low-valence (calm or sad) to high-energy, high-valence (excited, happy) moods, highlighting the diversity of emotional content in the catalog.
- **Popular songs (indicated by darker/bigger points) tend to cluster at medium to high levels of both energy and valence, reinforcing** that energetic, happy tracks dominate mainstream popularity. Lower danceability tracks (lighter blue points) are scattered more broadly across the mood space, often with lower energy and valence, suggesting niche or less commercially-oriented content.

# Do Longer Songs Stay Popular?



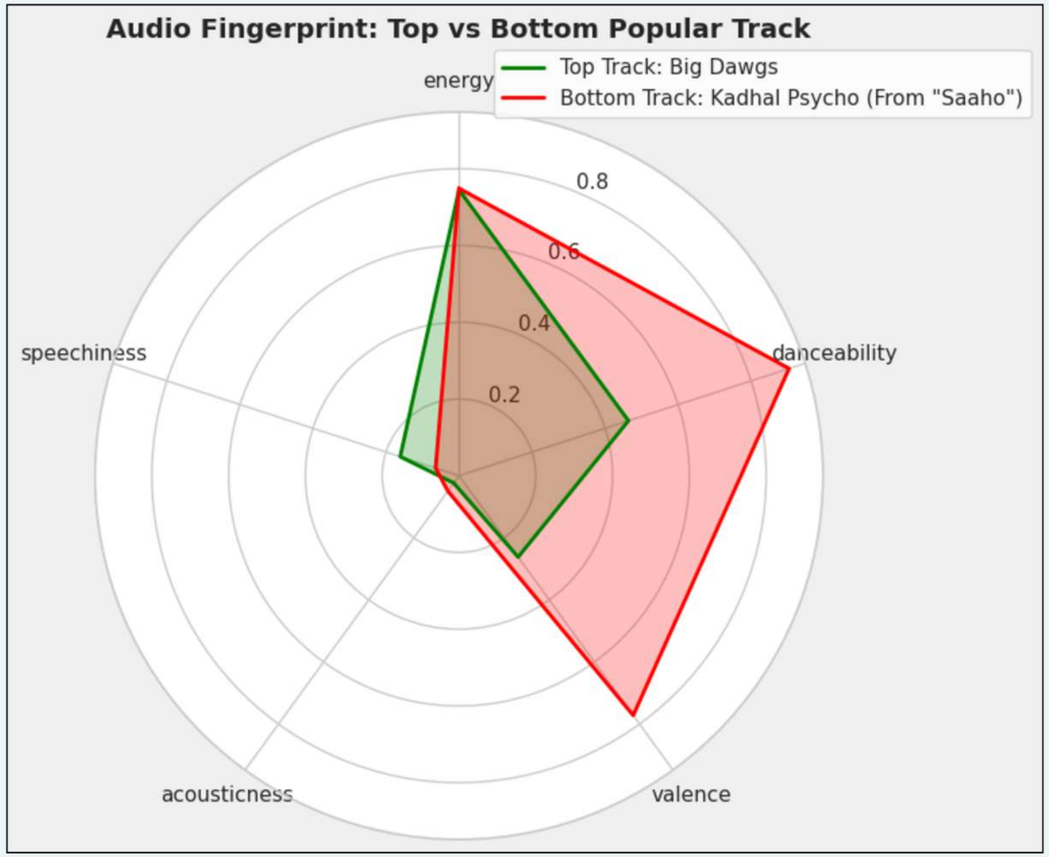

- The cluster of tracks with the highest popularity (80–100) overwhelmingly falls within a duration of roughly 2 to 5 minutes (120–300 seconds), showing that mainstream hits are almost never long-form tracks.
- As duration increases beyond approximately 7–8 minutes (500 seconds), the likelihood of high popularity drops sharply—very long tracks (e.g., DJ mixes, live sets) exist but almost never reach high chart status.
- **Short, conventional durations dominate popular music,** suggesting listener, radio, and playlist-friendly lengths are essential for commercial success, **while longer durations are generally associated with niche, experimental, or non-mainstream content.**

# Do Happier Songs Get Popular?




- Popularity is widely distributed across the full range of valence (happiness) values, indicating that both happy and sad songs can succeed—no single happiness level guarantees popularity.
- There is no strong upward or downward trend; highly popular tracks (popularity 80–100) are found at all valence levels, from the most somber to the most joyful.
- The average popularity line (in red) remains almost flat relative to valence, confirming that **"happiness" alone is not a key predictor for a song's commercial success**—other factors (like artist profile, marketing, or genre) play a more decisive role.

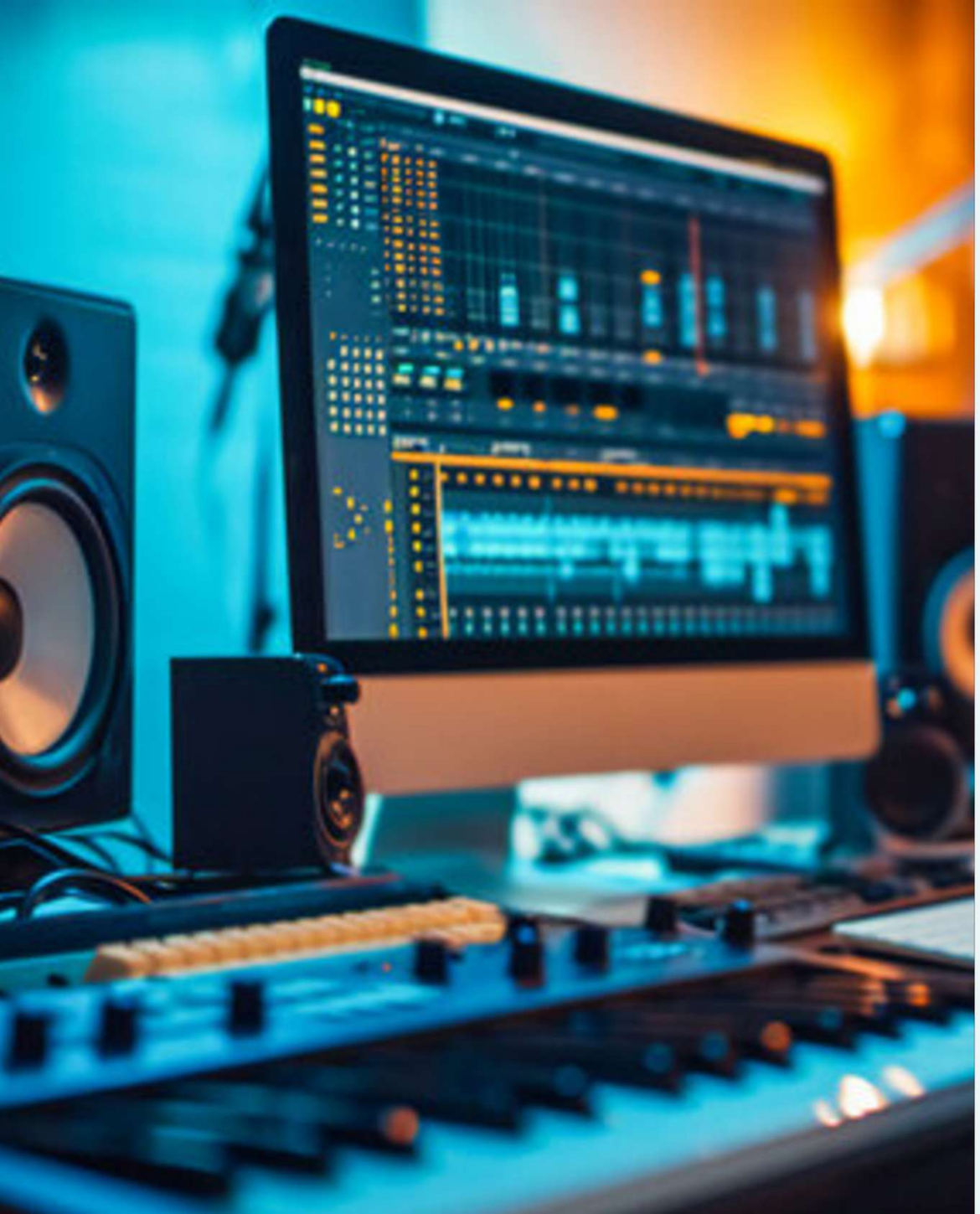
# Audio Fingerprint: Top vs Bottom Popular Track




- The radar plot compares the audio features of the most popular track ("Big Dawgs") vs the least popular track ("Kadhal Psycho" from "Saaho").
- Both tracks have very high energy, though "Kadhal Psycho" is slightly higher.
- The least popular track ("Kadhal Psycho") has much higher danceability and valence (happiness) compared to the popular track.
- The popular track ("Big Dawgs") has higher speechiness, suggesting more vocal or rap elements.
- Acousticness is low for both tracks, indicating studio-produced sound rather than acoustic or unplugged quality.
- **Overall, danceability and valence alone do not guarantee popularity; other features like speechiness and balance in energy are important differentiators.**

Spotify®

# *Final Recommendation*



# Strategic Music Direction

# Basic Recommendation

- **Target the hit zone with concise, punchy records: 2.7–4.3 minutes**, mid-to-fast grooves around 100–130 BPM, and assertive masters clustering near −6 to −4 dBFS where top-popularity points concentrate.
- Keep **energy and danceability balanced** in the mid-high band (energy ≈0.6–0.85, danceability ≈0.55–0.8), as higher segments enrich "High Energy & Danceability" profiles even though danceability alone isn't decisive.
- Favor vocal-led productions with very low instrumentalness and low speechiness, since top tiers compress speechiness toward low values and thin out instrumental content.
- Maintain low-to-moderate liveness for a polished studio feel, as popular tracks cluster away from live-leaning recordings.
- Default to 4/4 for scale; key and mode choices have negligible predictive power, so select for singer comfort and hook writing rather than "hit keys."
- Keep **acousticness modest** to low for mainstream formats due to its negative ties with energy, loudness, and a weak negative relation with popularity; reserve high-acoustic textures for calm/focus products.
- For rap-leaning cuts, work within the two proven tempo clusters (≈85–105 BPM and ≈120–140 BPM) and bias toward lower speechiness for mass playlists.
- Master for perceived punch: leverage the strong energy–loudness linkage and the slight positive relation between loudness and popularity while avoiding under-powered mixes that correlate with bottom tiers.
- **Avoid extremes that underperform: very long durations (>6–7 minutes),** very high liveness, high instrumentalness, ultra-low energy, or outlier ratios (unstable energy/dance balance).

# Production & Mixing Guidelines

- **Loudness :** aim in the dataset's hit corridor (approx. −6 to −4 dBFS), using limiting and transient control to sit on the "upper envelope" where higher popularity appears more often.
- **Spectral and vocal focus :** prioritize vocal intelligibility with controlled speechiness and minimal masking, as vocal-forward, **low-speechiness mixes dominate higher popularity bands.**
- **Texture :** reduce acousticness where energetic, processed impact is the goal; **high acousticness maps to calmer, less energetic outcomes better suited to niche contexts.**

# Song Arrangement & Composition Strategy

- Structure around replayability : short intros, early hooks, and tight sections at mid-tempo, reflecting that duration–popularity is flat overall but top hits cluster in radio-friendly lengths.
- Groove calibration : mid-to-high energy does the heavy lifting while danceability supports movement; extremely high energy does not automatically increase danceability.
- Meter and harmony : keep 4/4 and use major/minor flexibly; neither key nor mode materially shifts hit odds in the dataset.

# Language & Mood Optimization

- Lean Tamil for feel-good party (high danceability/valence), Korean for high-octane energy, Hindi for mainstream film/romance blends, and English for scale plus instrumental/sync-friendly beds.
- Prioritize Angry/Intense and Happy/Excited for scale while using Sad/Calm and Peaceful/Chill to grow session length and retention.

# Release Strategy & Audience Growth

- **Plan for the long tail :** popularity is heavily right-skewed—deploy collaborations, editorial pitching, and short-form hooks for the tail, and concentrate spend on proven High tracks to lift into Very High.
- Exploit top-artist and album clusters for cross-promos, remasters, and carousels, since their averages are consistently elevated.

# Priority Actions & Testing Framework

- Iterate on loudness, groove tightness, vocal presence, and succinct runtimes; these levers show clearer associations with higher segments than key/mode/time signature.
- Segment tests by language-mood lanes to match strengths (e.g., Tamil party, Korean hype) and improve early traction.

# Key Risks & Considerations

- High instrumentalness, high acousticness, high liveness, or very long durations markedly correlate with lower tiers; treat as deliberate niche choices rather than mainstream strategies.

# Algorithmic Optimization Insights

- **Cold-start priors :** boost candidates with mid-high energy and danceability, low instrumentalness and liveness, low speechiness, standard 4/4, and radio-length durations; down-weight key/mode/time signature due to minimal predictive power.
- **Hidden-gem surfacing :** scan the long tail for tracks already in the hit zone on loudness, energy, danceability, duration, and liveness but with low popularity, then allocate exploration budget and editorial trials.
- **Model hygiene :** winsorize or log-scale unstable energy_dance ratios and regularize collinear features (energy–loudness, acousticness vs energy/loudness, danceability–valence).
- **Language-mood routing :** map Tamil to party/high danceability, Korean to high energy, Hindi to mainstream film/romance, and English to scale plus instrumental/sync; route early impressions accordingly.
- **Rap pacing logic :** emphasize ≈85–105 and ≈120–140 BPM clusters with lower speechiness bias for mass playlists; reserve high-speechiness outliers for niche editorial.
- **Diversity guardrails :** maintain 4/4 dominance for scale while preserving non-4/4, key, and mode diversity as novelty signals rather than ranking drivers.
- **Evaluation :** success metrics should reflect uplift on completion, saves, and repeat plays for mid-to-high energy/danceability priors versus baseline, with safety checks against length and liveness extremes.

# THANK YOU

FOR YOUR ATTENTION

--- Presented by ---
RITTIK DAS
B.Tech, CSE, 2nd Year