Exploratory Data Analysis

Discover the essential foundation of data science that transforms raw information into actionable insights through systematic exploration and visualization.

Aranya Rayed | Team Hephaestus | The ManHattan Project

Problem Statement

Spotify is one of the world's leading music streaming platforms, hosting a vast catalog of tracks enriched with detailed audio features and popularity metrics. The company's **goal** is to **gain a deeper understanding** of what makes songs **resonate with listeners** in order to **enhance** recommendations, **guide** music production, and **optimize** artist promotion strategies.

Owner/User: Music Directors and Mixing Engineers, seeking insights to optimize song composition and production.

Consumer of Analysis: Producers, record labels, and streaming strategists who leverage data insights to make informed creative and business decisions.

My challenge was to act as data analyst for Spotify. Using the provided tracks dataset, I performed a comprehensive **Exploratory Data Analysis (EDA)** to uncover actionable insights. My analysis focuses on **answering key business questions** around music popularity, **identifying feature patterns** that drive success, and **highlighting opportunities** for innovation in the streaming industry.

Key Challenge Areas

- **Identifying the Drivers of Popularity**

  Understanding which audio features — such as danceability, energy, valence, loudness, and tempo — most strongly influence a song's popularity is critical. The challenge lies in separating features that consistently contribute to success from those that show only weak or context-specific correlations.

- **Uncovering Patterns and Clusters in Music Profiles**

  Songs often succeed not because of a single attribute, but due to specific combinations of features (e.g., high energy + high danceability + short duration). Identifying these "success recipes" or clusters of popular tracks requires multivariate analysis and careful interpretation.

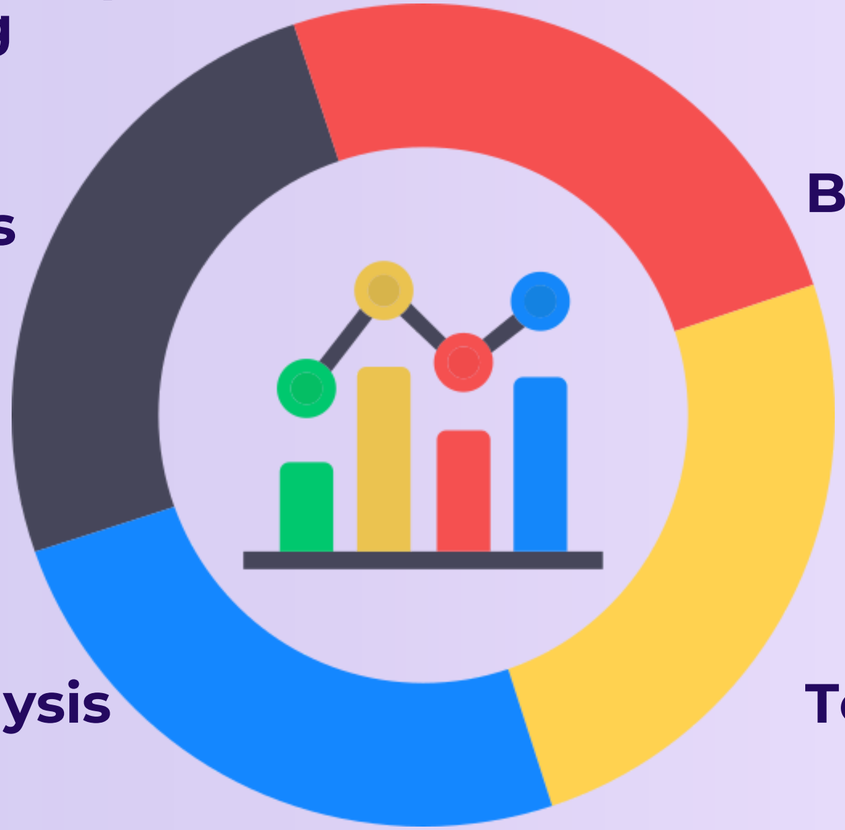
- **Tracking the Evolution of Music Trends Over Time**

  Listener preferences evolve, with noticeable shifts in song length, loudness, and even dominant languages across decades. A key challenge is to capture these long-term time series trends and translate them into actionable insights for music production and strategic decision-making.

Data Description & cleaning

Final Recommendations

Initial Analysis



Bottom N Analysis

Univariate Analysis

Top N Analysis

Bivariate Analysis

Temporal Analysis

Multivariate Analysis

Data Description

spotify_data_description.csv

	Column Name	Description
0	track_id	A unique identifier for the track on Spotify.
1	track_name	The title of the song.
2	artist_name	The name of the artist(s) who performed the song.
3	year	The release year of the song.
4	popularity	A measure of how popular a track is, ranging f...
5	artwork_url	A URL pointing to the album artwork for the tr...
6	album_name	The name of the album the track belongs to.
7	acousticness	A confidence measure indicating whether the tr...
8	danceability	A measure of how suitable a track is for danci...
9	duration_ms	The duration of the track in milliseconds.
10	energy	A perceptual measure of intensity and activity...
11	instrumentalness	Predicts whether a track contains no vocal con...
12	key	The key the track is in, represented as an int...
13	liveness	Detects the presence of an audience in the rec...
14	loudness	The overall loudness of a track in decibels (dB).
15	mode	Indicates the modality (major or minor) of a t...
16	speechiness	A measure detecting the presence of spoken wor...
17	tempo	The overall estimated tempo of a track in beat...
18	time_signature	An estimated overall time signature of a track
19	valence	A measure from -1.0 to 1.0 describing the musi...
20	track_url	A URL to the Spotify track.
21	language	The detected language of the song's lyrics.

spotify_tracks.csv

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 62317 entries, 0 to 62316
Data columns (total 22 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   track_id          62317 non-null  object
 1   track_name        62317 non-null  object
 2   artist_name       62317 non-null  object
 3   year              62317 non-null  int64
 4   popularity        62317 non-null  int64
 5   artwork_url       62317 non-null  object
 6   album_name        62317 non-null  object
 7   acousticness      62317 non-null  float64
 8   danceability      62317 non-null  float64
 9   duration_ms       62317 non-null  float64
 10  energy            62317 non-null  float64
 11  instrumentalness  62317 non-null  float64
 12  key               62317 non-null  float64
 13  liveness          62317 non-null  float64
 14  loudness          62317 non-null  float64
 15  mode              62317 non-null  float64
 16  speechiness       62317 non-null  float64
 17  tempo             62317 non-null  float64
 18  time_signature    62317 non-null  float64
 19  valence           62317 non-null  float64
 20  track_url         62317 non-null  object
 21  language          62317 non-null  object
dtypes: float64(13), int64(2), object(7)
memory usage: 10.5+ MB
```

Data Description

spotify_tracks.csv description

	year	popularity	acousticness	danceability	duration_ms	energy	instrumentalness	key	liveness	loudness	mode	speechiness	tempo	time_signature	valence
count	62317.000000	62317.000000	62317.000000	62317.000000	6.231700e+04	62317.000000	62317.000000	62317.000000	62317.000000	62317.000000	62317.000000	62317.000000	62317.000000	62317.000000	62317.000000
mean	2014.425935	15.358361	0.362292	0.596807	2.425270e+05	0.602496	0.146215	5.101658	0.194143	-65.103433	0.586052	0.087722	117.931247	3.857086	0.495226
std	9.645113	18.626908	0.314609	0.186209	1.129999e+05	0.246144	0.307804	3.553469	0.172030	2369.051478	0.493682	0.115150	28.509459	0.502660	0.264787
min	1971.000000	0.000000	-1.000000	-1.000000	5.000000e+03	-1.000000	-1.000000	-1.000000	-1.000000	-100000.000000	-1.000000	-1.000000	-1.000000	-1.000000	-1.000000
25%	2011.000000	0.000000	0.067100	0.497000	1.921600e+05	0.440000	0.000000	2.000000	0.093200	-10.727000	0.000000	0.036700	95.942000	4.000000	0.292000
50%	2017.000000	7.000000	0.286000	0.631000	2.362670e+05	0.639000	0.000025	5.000000	0.125000	-7.506000	1.000000	0.048900	117.991000	4.000000	0.507000
75%	2022.000000	26.000000	0.632000	0.730000	2.862400e+05	0.803000	0.015200	8.000000	0.243000	-5.456000	1.000000	0.089100	135.081000	4.000000	0.710000
max	2024.000000	93.000000	0.996000	0.986000	4.581483e+06	1.000000	0.999000	11.000000	0.998000	1.233000	1.000000	0.959000	239.970000	5.000000	0.995000

spotify_tracks.csv structure

	track_id	track_name	artist_name	year	popularity	artwork_url	album_name	acousticness	danceability	duration_ms	...	key	liveness	loudness	mode	speechiness	te
0	2r0ROhr7pRN4MXDMT1fEmd	Leo Das Entry (From "Leo")	Anirudh Ravichander	2024	59	https://i.scdn.co/image/ab67616d0000b273ce9c65...	Leo Das Entry (From "Leo")	0.0241	0.753	97297.0	...	8.0	0.1000	-5.994	0.0	0.1030	110.
1	4l38e6Dg52a2o2a8i5Q5PW	AAO KILLELLE	Anirudh Ravichander, Pravin Mani, Vaishali Sri...	2024	47	https://i.scdn.co/image/ab67616d0000b273be1b03...	AAO KILLELLE	0.0851	0.780	207369.0	...	10.0	0.0951	-5.674	0.0	0.0952	164.
2	59NoiRhnom3lTeRFaBzOev	Mayakiriye Sirikiriye - Orchestral EDM	Anirudh Ravichander, Anivee, Alvin Bruno	2024	35	https://i.scdn.co/image/ab67616d0000b27334a1dd...	Mayakiriye Sirikiriye (Orchestral EDM)	0.0311	0.457	82551.0	...	2.0	0.0831	-8.937	0.0	0.1530	169.
3	5uUqRQd385pvLxC8JX3tXn	Scene Ah Scene Ah - Experimental EDM Mix	Anirudh Ravichander, Bharath Sankar, Kabilan, ...	2024	24	https://i.scdn.co/image/ab67616d0000b27332e623...	Scene Ah Scene Ah (Experimental EDM Mix)	0.2270	0.718	115831.0	...	7.0	0.1240	-11.104	1.0	0.4450	169.
4	1KaBRg2xgNeCljmyxBH1mo	Gundellonaa X I Am A Disco Dancer - Mashup	Anirudh Ravichander, Benny Dayal, Leon James, ...	2024	22	https://i.scdn.co/image/ab67616d0000b2735a59b6...	Gundellonaa X I Am a Disco Dancer (Mashup)	0.0153	0.689	129621.0	...	7.0	0.3450	-9.637	1.0	0.1580	128.

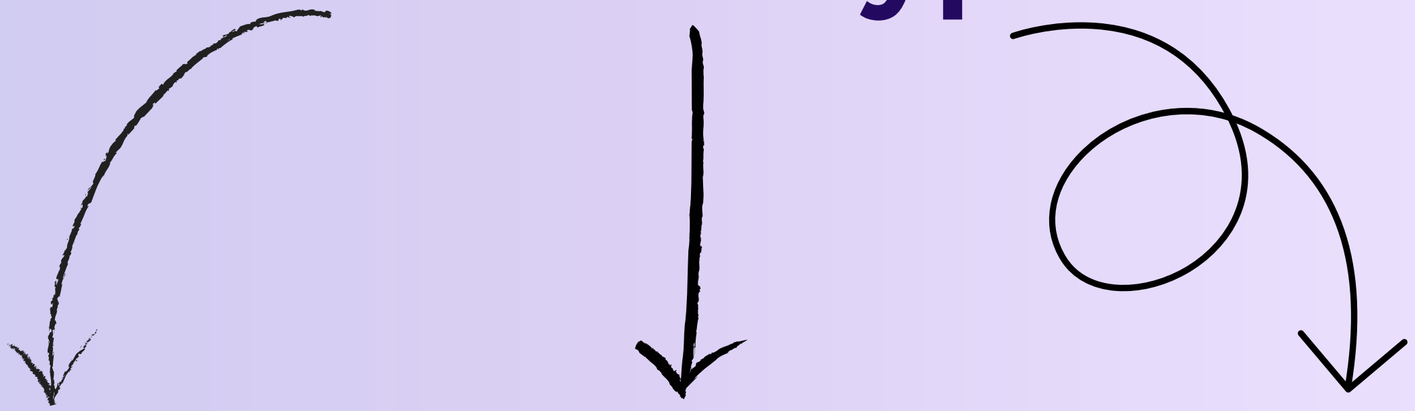
Data Cleaning

- Removing Duplicate data
- Fixing missing values
- Fixing Data types

spotify_tracks.csv
after data cleaning

```
Index: 62239 entries, 0 to 62316
Data columns (total 25 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   track_id            62239 non-null  object
 1   track_name          62239 non-null  object
 2   artist_name         62239 non-null  object
 3   year                62239 non-null  int64
 4   popularity          62239 non-null  int64
 5   artwork_url         62239 non-null  object
 6   album_name          62239 non-null  object
 7   acousticness        62239 non-null  float64
 8   danceability        62239 non-null  float64
 9   duration_ms         62239 non-null  int64
 10  energy              62239 non-null  float64
 11  instrumentalness    62239 non-null  float64
 12  key                 62239 non-null  int64
 13  liveness            62239 non-null  float64
 14  loudness            62239 non-null  float64
 15  mode                62239 non-null  int64
 16  speechiness         62239 non-null  float64
 17  tempo               62239 non-null  float64
 18  time_signature      62239 non-null  int64
 19  valence             62239 non-null  float64
 20  track_url           62239 non-null  object
 21  language            62239 non-null  category
 22  key_cat             62204 non-null  object
 23  mode_cat            62204 non-null  object
 24  time_signature_cat  62239 non-null  object
dtypes: category(1), float64(9), int64(6), object(9)
```

Variable Types



Categorical

language
key
mode
time_signature

Numerical

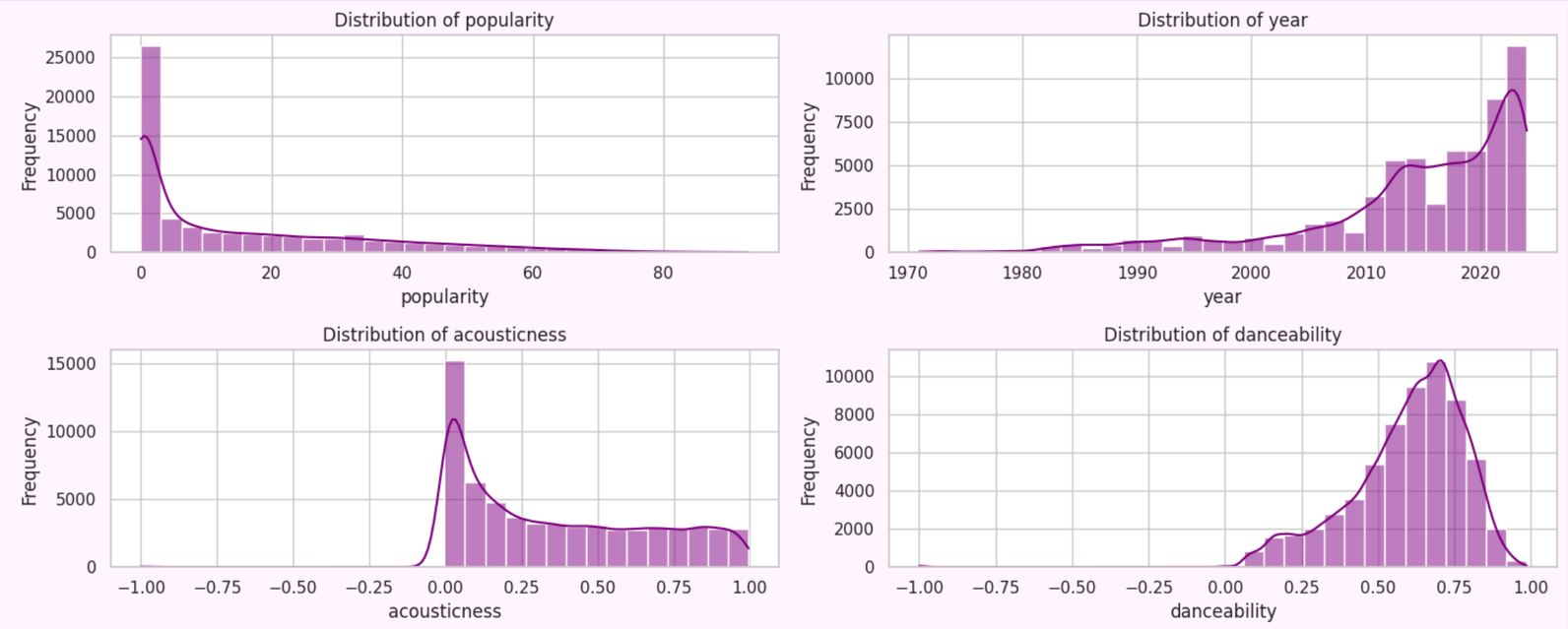
popularity
year
acousticness
danceability
duration_ms
energy
instrumentalness
liveness
loudness
speechiness
tempo
valence

Temporal

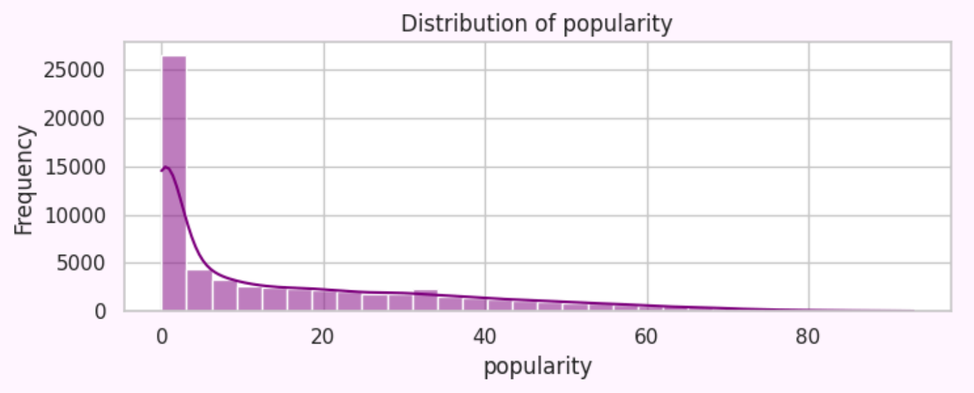
year

Univariate Analysis

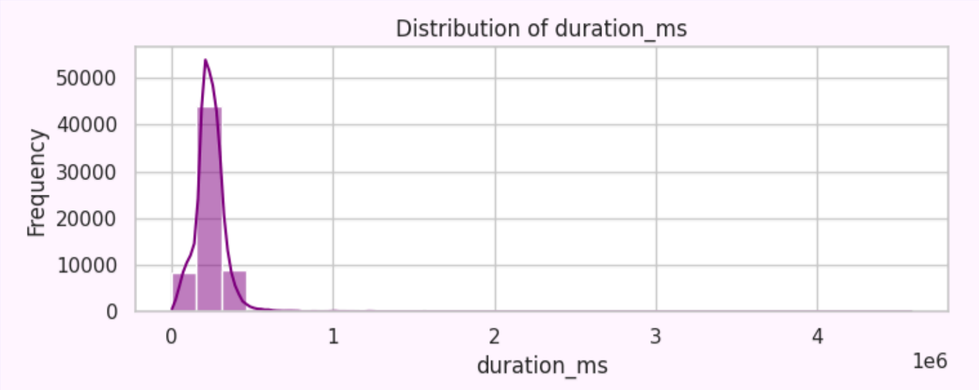
Univariate analysis is the simplest form of data analysis, where you look at only one variable at a time. The goal is to understand the distribution, central tendency, and spread of that single variable, without worrying about its relationship with others.

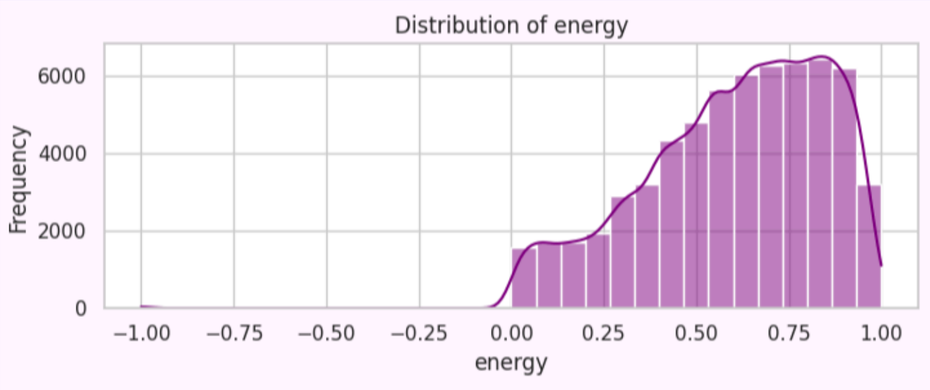


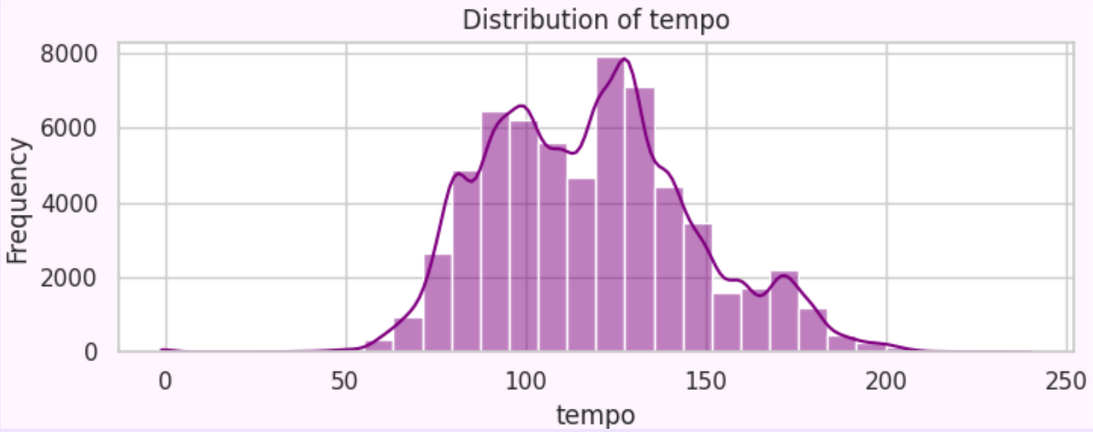
Numerical



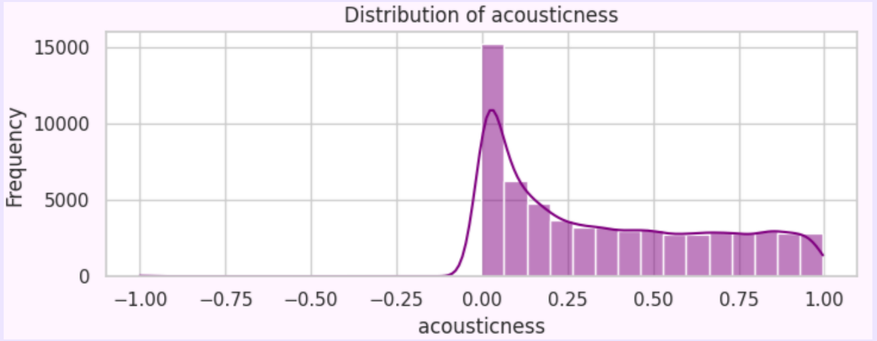
- Majority of tracks are not popular (median = 7); only a few achieve very high popularity.
- Typical songs last 3–5 minutes, but extreme outliers (up to 76+ minutes) distort the mean.
- Both distributions are centered around ~0.6, indicating that most tracks are moderately energetic and danceable.
- Clustered around ~118 BPM, with multiple peaks (likely genre-driven), fitting mainstream pop/EDM.

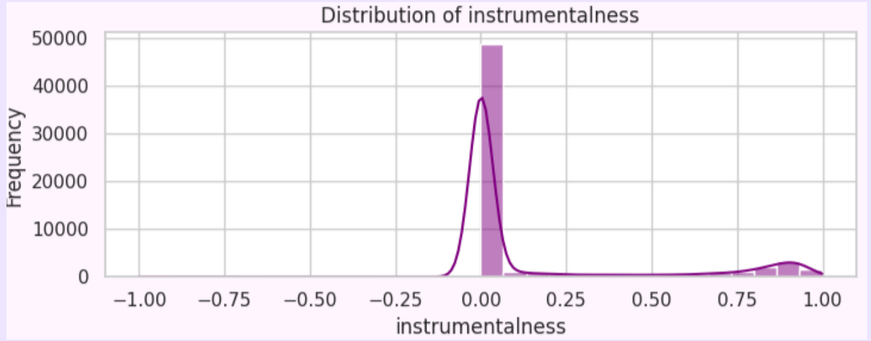


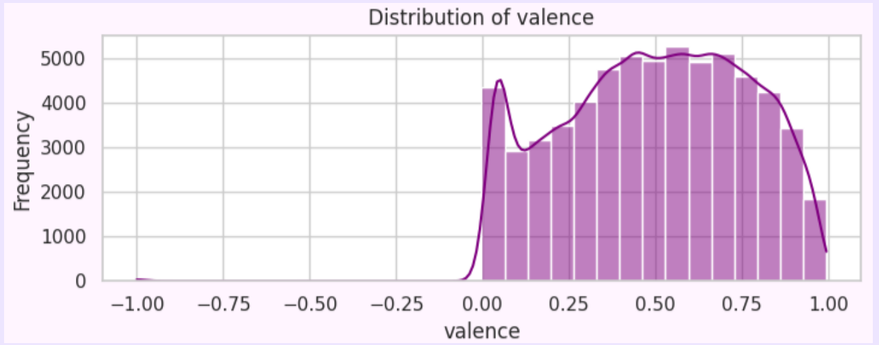


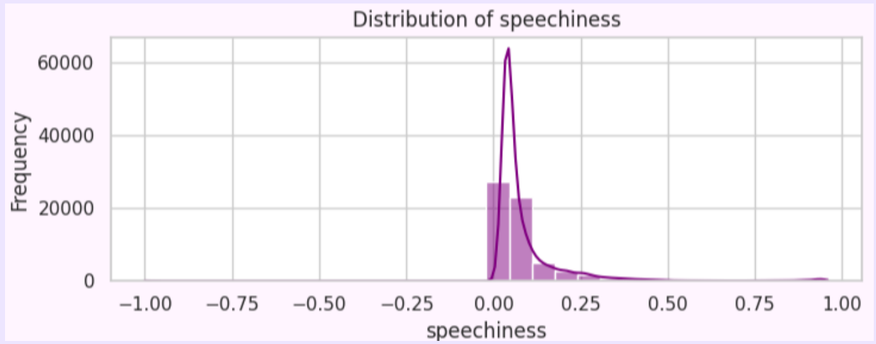


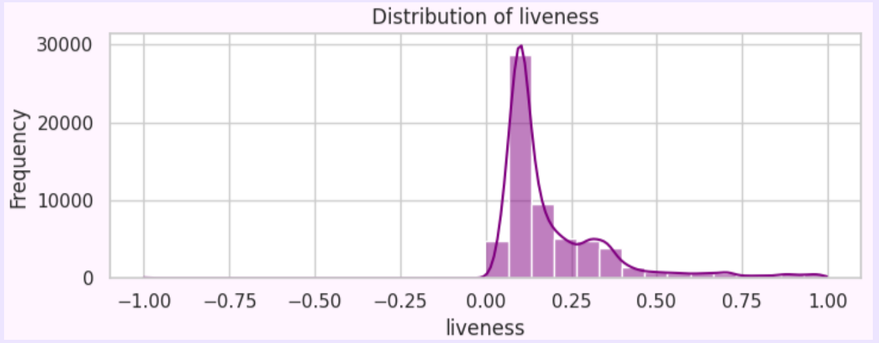
Numerical





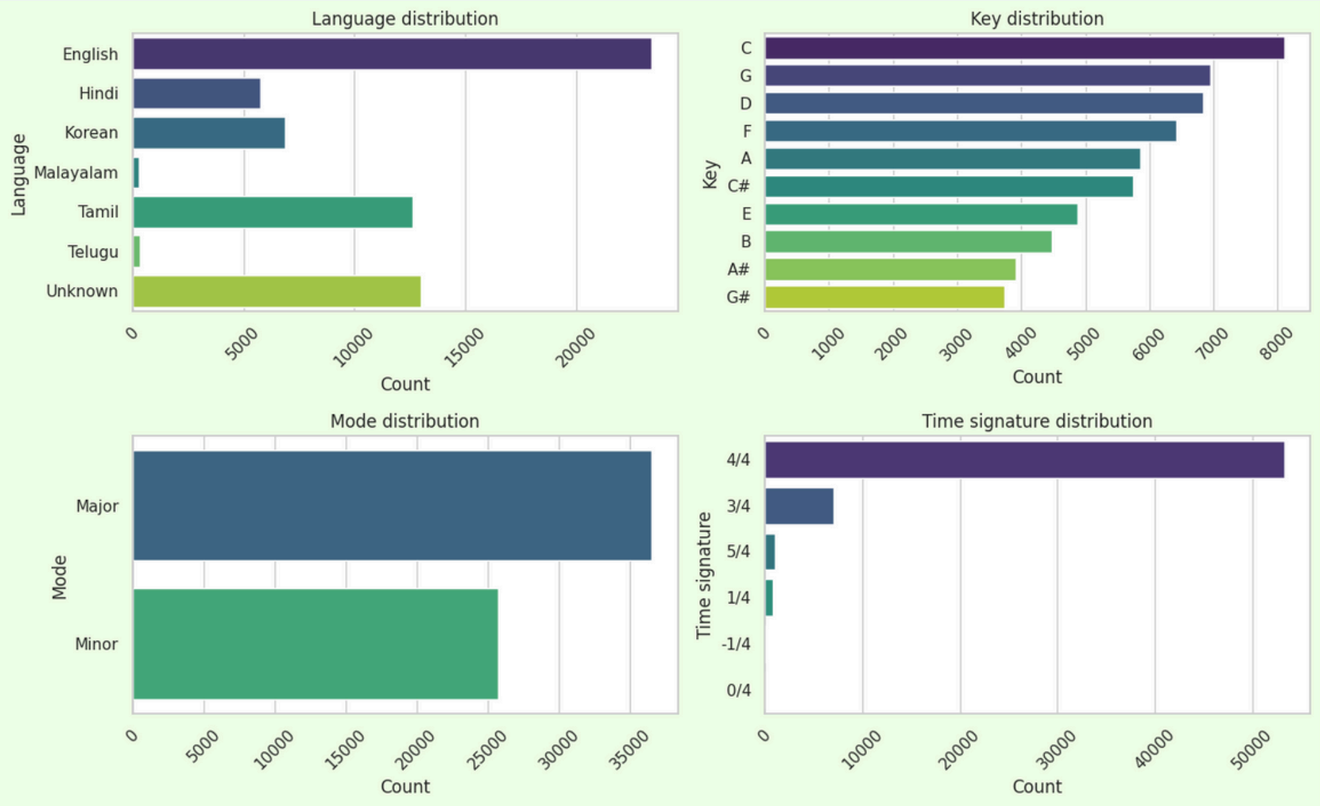






- Most tracks are non-acoustic and vocal-based, with only a small share being purely instrumental or acoustic.
- Fairly balanced mix of positive (“happy”) and negative (“sad/melancholic”) songs.
- Majority are studio recordings; live performances are rare.
- Most songs have low speech-like content, but outliers include rap or spoken-word tracks.

Categorical



- C (8,101) is the most common key.
- Other popular keys: G (6,956), D (6,832), F (6,411).
- Less frequent keys include A# (3,920) and G# (3,735).

  - Major mode is more prevalent (36,511 tracks) than Minor mode (25,693).
  - Rough ratio: 60% Major vs 40% Minor.

- 4/4 overwhelmingly dominates with 53,260 tracks while 3/4 (7,065) is also common.
- Other signatures like 5/4 and ¼ are rare.

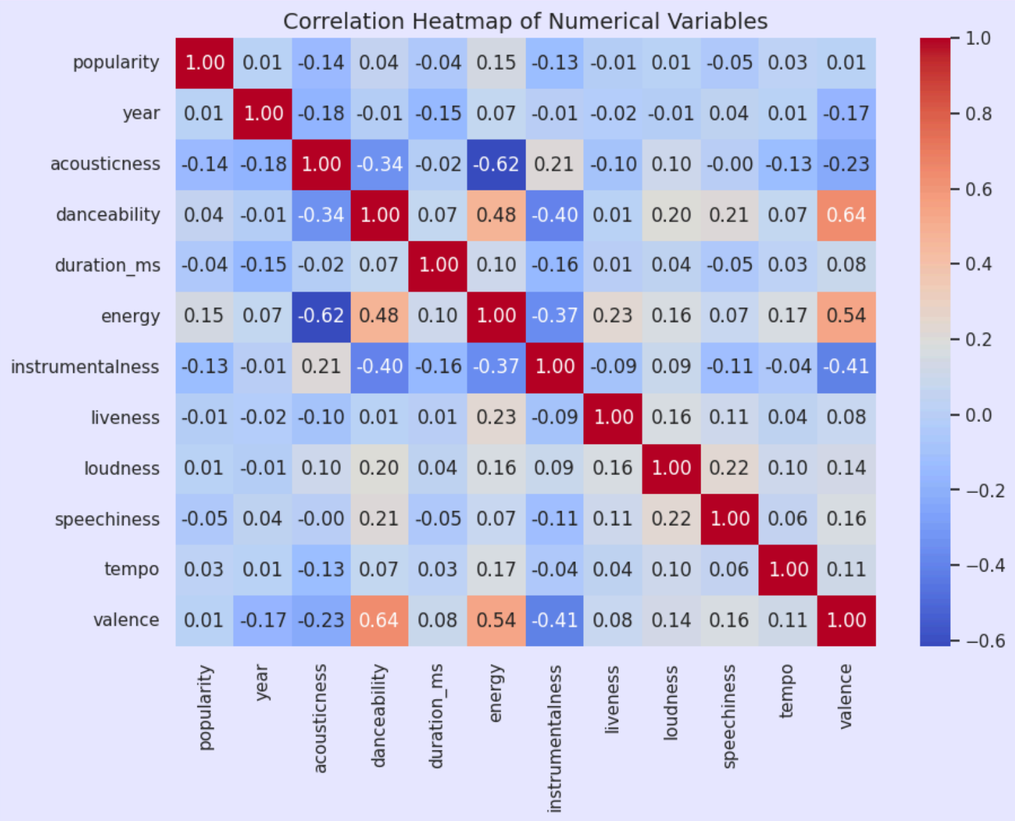
  - English dominates with 23,389 tracks.
  - Significant share of Unknown (13,005) entries.
  - Tamil (12,609) also has strong representation.
  - Other languages like Korean (6,893), Hindi (5,740) are moderately represented.
  - Telugu (321) and Malayalam (282) are minimal.

Suggestions

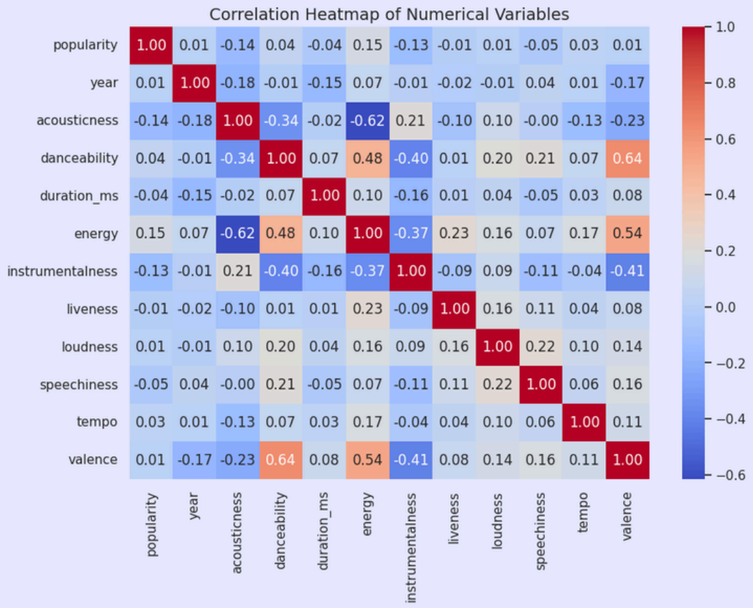
- **Boost Discoverability:** Most tracks remain obscure; focus on promoting mid-tier songs while leveraging the top percentile of highly popular tracks.
- **Leverage Diversity:** Use mood (valence), energy, and tempo variations to build targeted playlists (e.g., workout, chill, party, focus).
- **Highlight Niche Content:** Acoustic, instrumental, live, and non-4/4 tracks are rare — curating them can attract niche audiences.
- **Regional & Language Curation:** English dominates, but Tamil, Hindi, Korean, and others offer opportunities for regional and multilingual playlists.

Bivariate Analysis

Bivariate analysis is the study of the relationship between two variables at the same time. It helps identify correlations, associations, or dependencies between them. For example, analyzing how age affects spending or how payment method varies by customer segment.



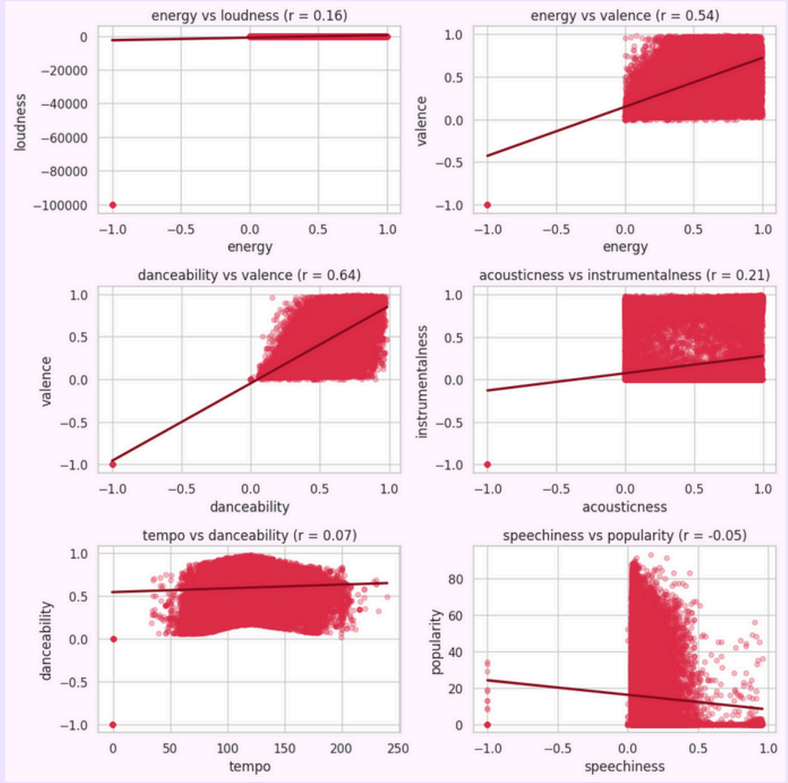
Numerical vs Numerical



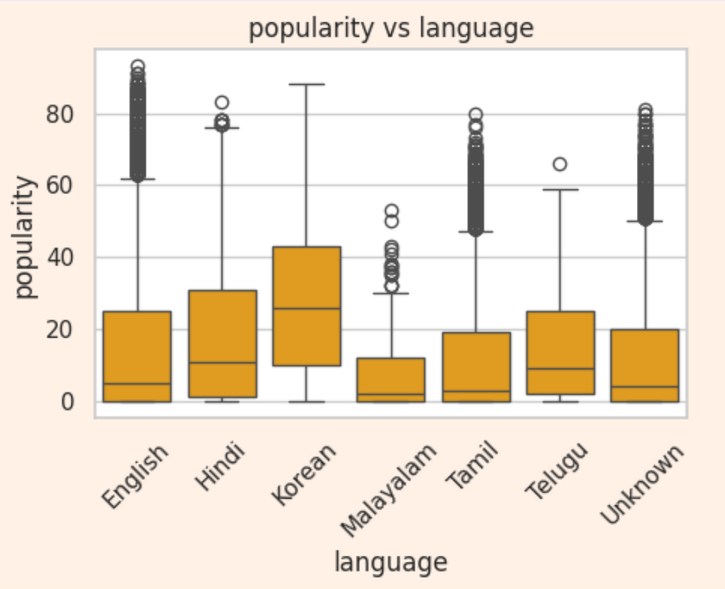
Correlation Heatmap

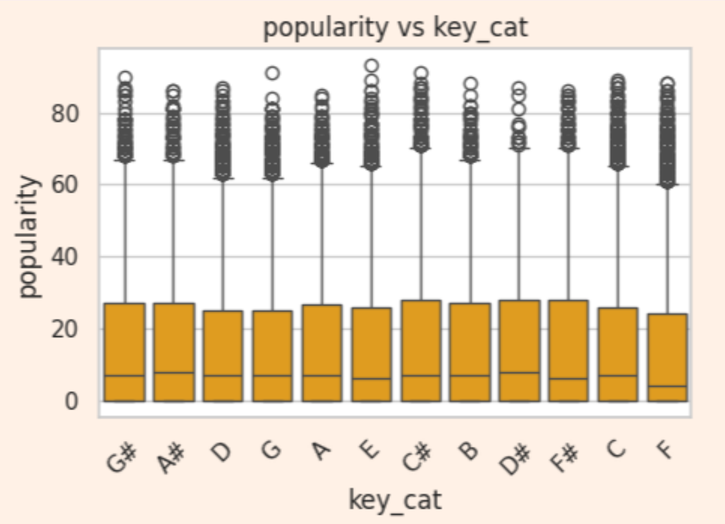
- Danceability vs Valence (r=0.64): Highly danceable songs are strongly associated with positive, “happy” moods.
- Energy vs Valence (r=0.54): Energetic songs tend to be happier in mood.
- Acousticness vs Energy (r=-0.62): Acoustic tracks are generally less energetic.
- Instrumentalness vs Valence (r=-0.41): Instrumental songs lean toward less positive/more melancholic moods.

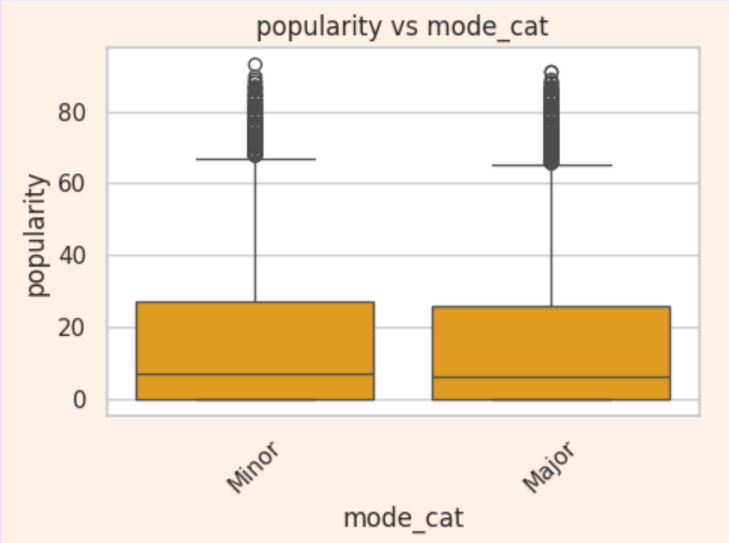
- **Danceability** and **valence** show the **strongest positive correlation** (r=0.64), suggesting happier-sounding songs are typically easier to dance to.
- A moderate **positive correlation** exists between **energy** and **valence** (r=0.54), indicating that high-energy tracks are often more positive in mood.
- There is almost **no linear relationship** between **tempo and danceability** (r=0.07), showing that a song's speed is a poor predictor of its danceability.
- **Speechiness** and **popularity** (r=-0.05) have a **negligible correlation**, with the vast majority of tracks having low speechiness scores.



Numerical vs Categorical

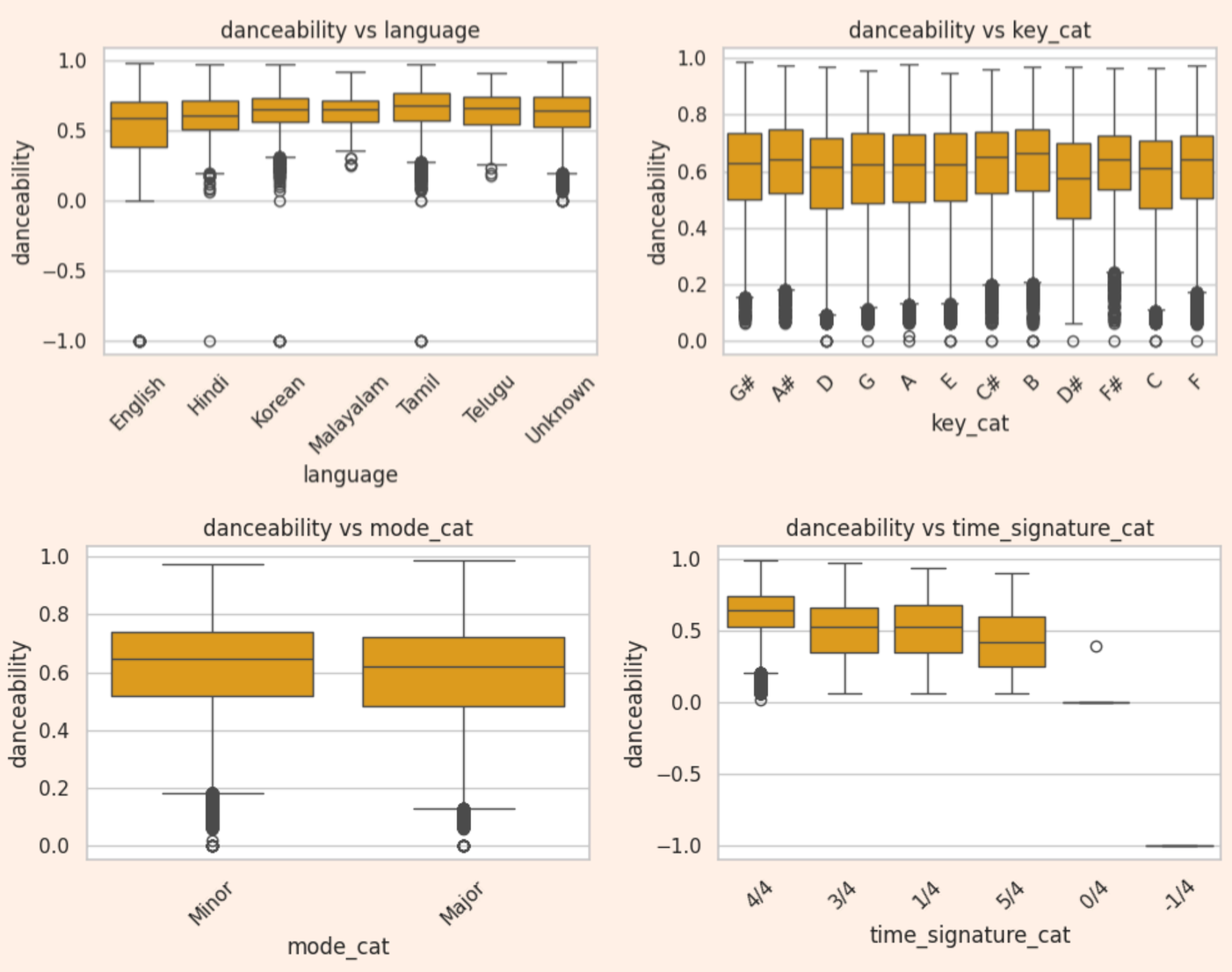






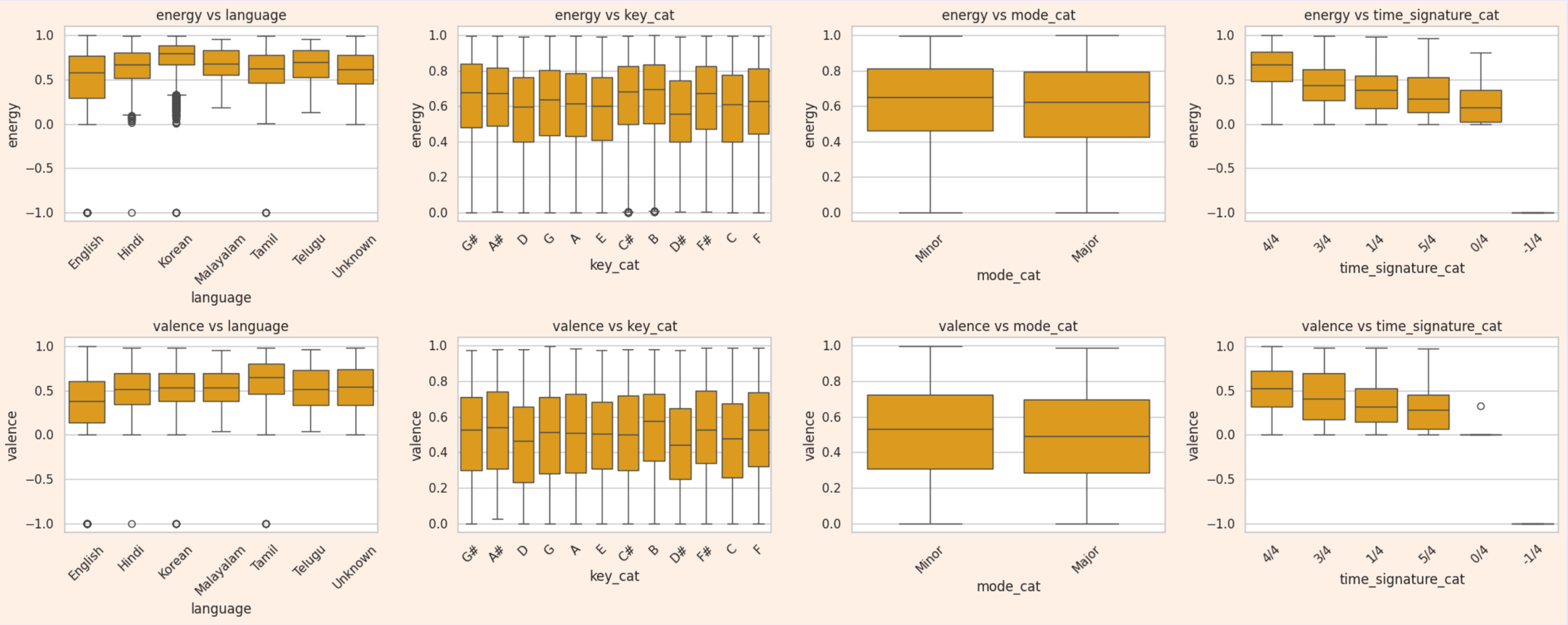
- Popularity shows little variation across keys, modes, or time signatures.
- Some language differences: Korean tracks have slightly higher popularity spread; Tamil/Hindi show lower medians.

Numerical vs Categorical



- Danceability is fairly consistent across languages, keys, and modes (median ~0.6).
- 4/4 songs are most energetic; unusual time signatures (0/4, -1/4) show very low energy.
- Danceable/energetic patterns reflect in valence: 4/4 has higher positivity, while rare time signatures (1/4, 5/4, -1/4) skew more negative.
- Mode (Major vs Minor) shows expected trend: Major tends toward higher valence, Minor more melancholic.

Numerical vs Categorical



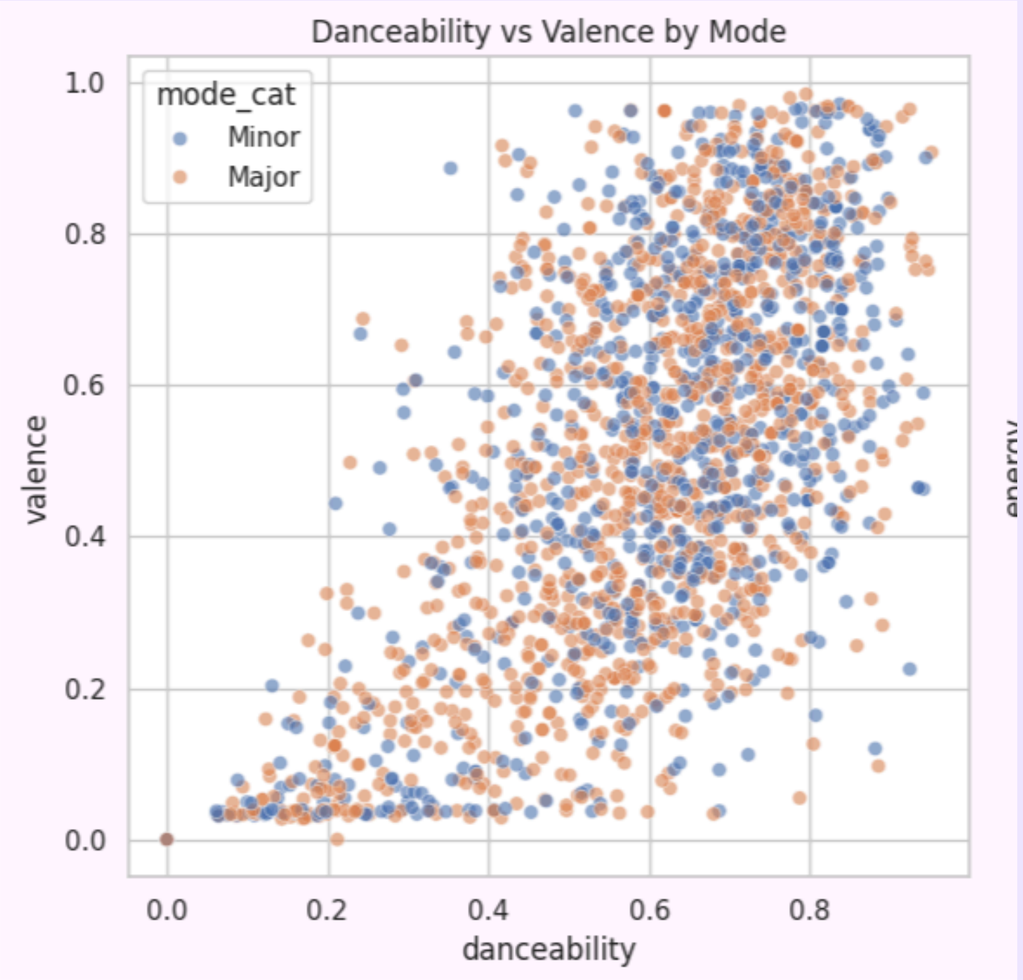
Categorical vs Categorical

- The 4/4 time signature is overwhelmingly dominant, being the most common by a large margin across every language, musical key, and mode in the dataset.
- While the Major mode is more common for most musical keys, there is a notable exception for the keys of F and B, where the Minor mode is more frequent.
- Among the specified languages, English and Hindi songs appear more frequently in a Major key than in a Minor key.

Suggestions

- **Create "Feel-Good" Hits:** The most effective formula for a positive listener experience is a combination of high danceability and high valence (musical positivity).
- **Use Energy to Boost Mood:** High energy levels, especially in Major key and 4/4 time, directly contribute to a song's positive feel, so avoid overly acoustic or instrumental sounds for upbeat tracks.
- **Prioritize Marketing for Popularity:** Since audio features alone have a weak link to popularity, focus on marketing, artist visibility, and cultural context as the primary drivers of success.
- **Stick to Proven Structures for Mainstream Appeal:** The 4/4 time signature in a Major or Minor key is the safest and most widely accepted musical structure for maximizing reach.
- **Use Niche Elements Strategically:** Reserve alternative time signatures (like 3/4 or 5/4) and less common structures for specific genres or to add creative diversity, not for songs aiming for mass popularity.

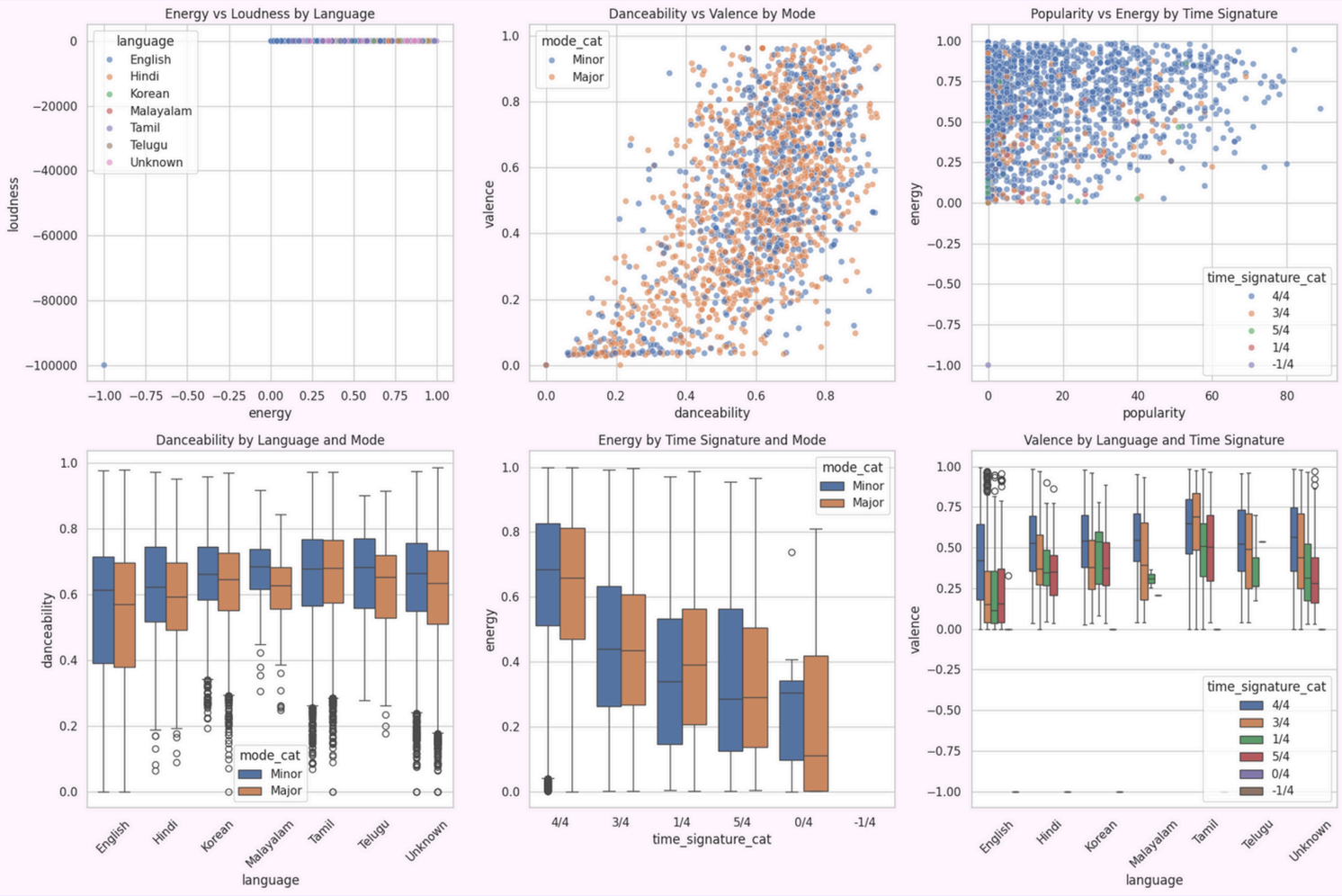
Multivariate Analysis



Multivariate Analysis means analyzing the relationship between three or more variables at the same time to understand patterns, interactions, and combined effects.

Where univariate looks at one variable and bivariate looks at two, multivariate goes deeper by studying how multiple variables(three or more) together influence outcomes.

Multivariate Analysis



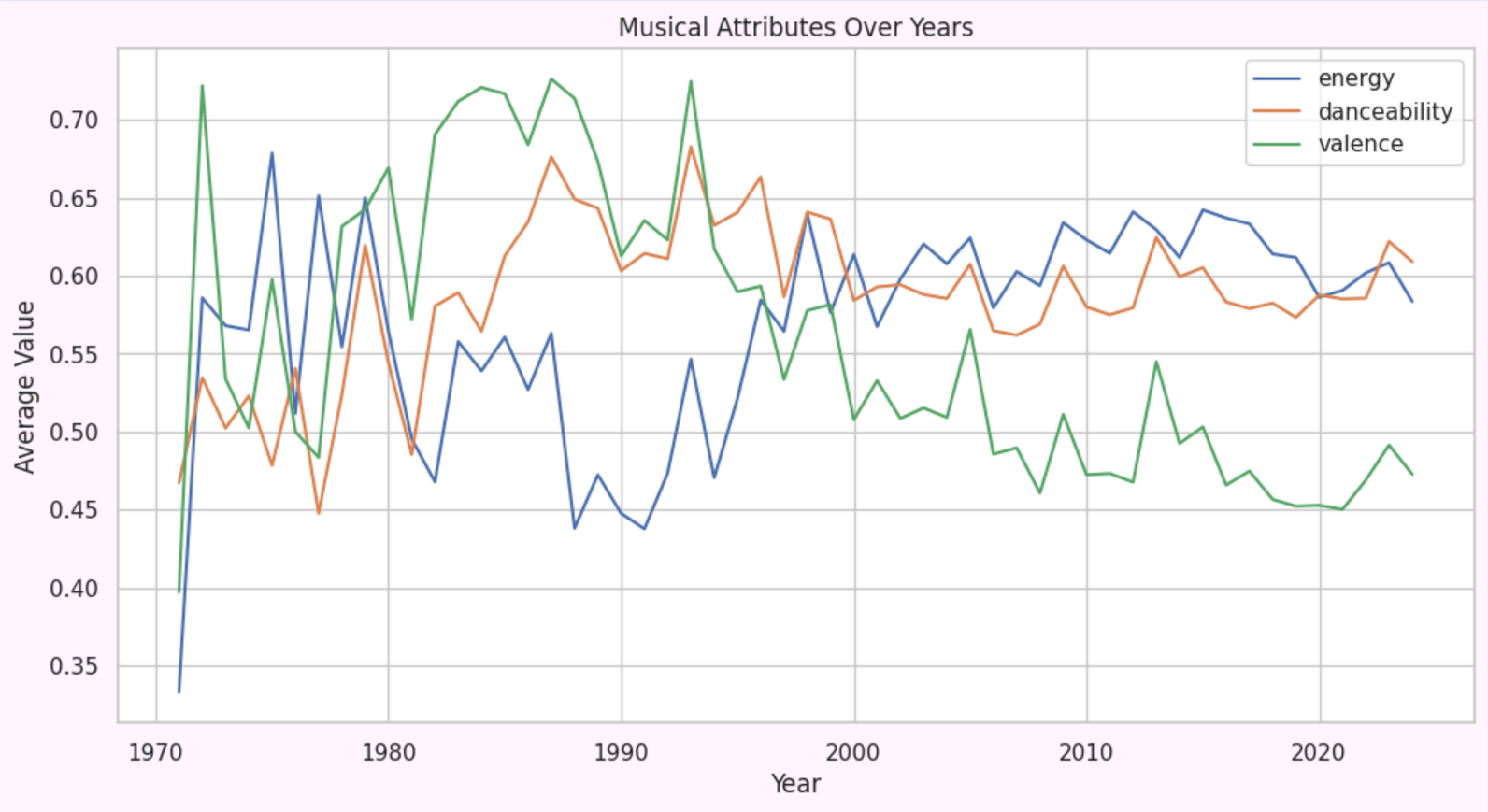
- Highly popular songs are consistently associated with high energy and an overwhelming majority are in a 4/4 time signature.
- For popular songs, high loudness aligns with high energy, but these plots don't provide information on their typical tempo or mode.
- The provided charts do not contain data on song duration or year, so it's not possible to identify any historical trends.

Suggestions

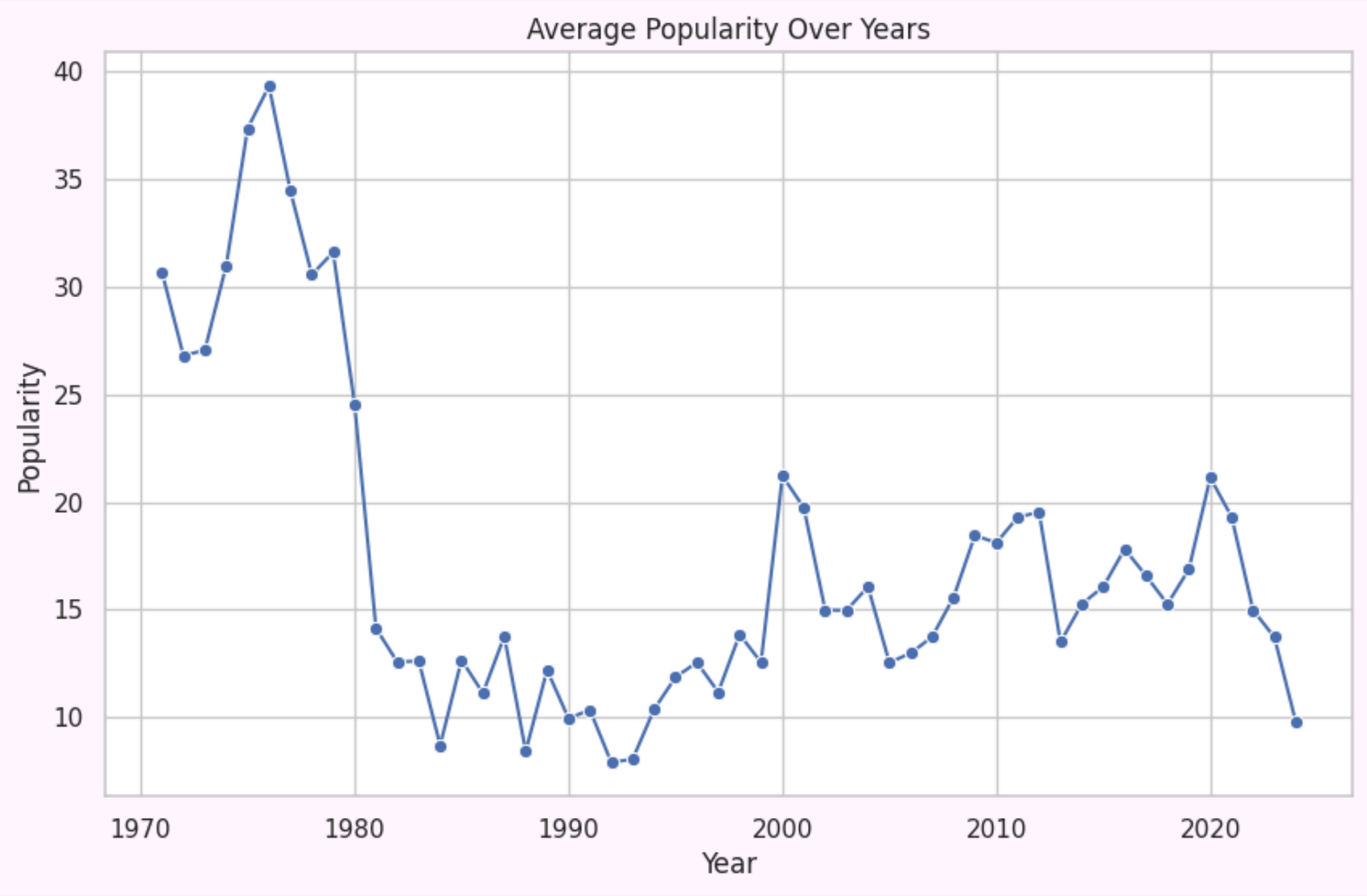
- **Prioritize 4/4 Major for Commercial Hits:** The most successful and common track structure combines a 4/4 time signature with a Major key, aligning with higher danceability, energy, and valence.
- **Create "Feel-Good" Tracks:** Leverage the strong link between high danceability and high valence (positivity) to produce engaging songs ideal for popular "party" or "happy" playlists.
- **Maintain Mood Diversity:** While Major keys are dominant, Minor keys are also vital for appealing to different listener moods. Aim for a ~70% Major to 30% Minor ratio in your music portfolio.
- **Audio Features Don't Guarantee Popularity:** Popularity has a very weak correlation with specific audio features, meaning the "perfect" sound doesn't ensure a hit.
- **Focus on Promotion Over Production Tweaks:** Since marketing and playlisting are stronger drivers of success, resources should be focused on distribution, collaborations, and promotion rather than just optimizing sonic traits.

Temporal Analysis

Temporal or Timeseries analysis is the study of data, events, or patterns over time. The word temporal comes from "time," so it's essentially about understanding how things change, evolve, or behave across different time periods.

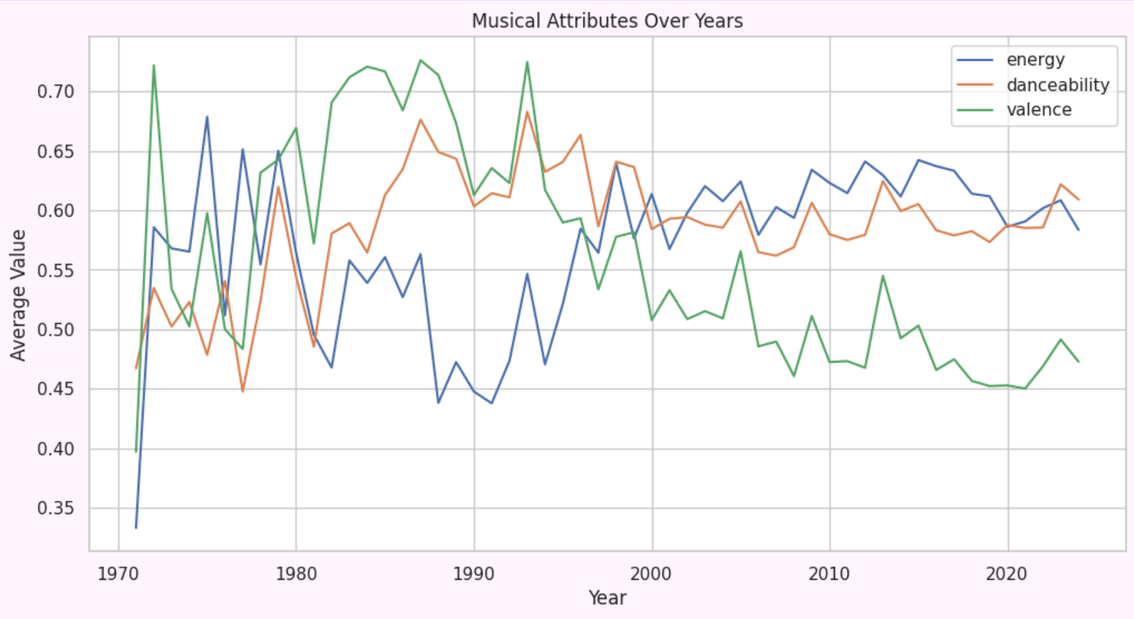


Yearly Trend in Popularity



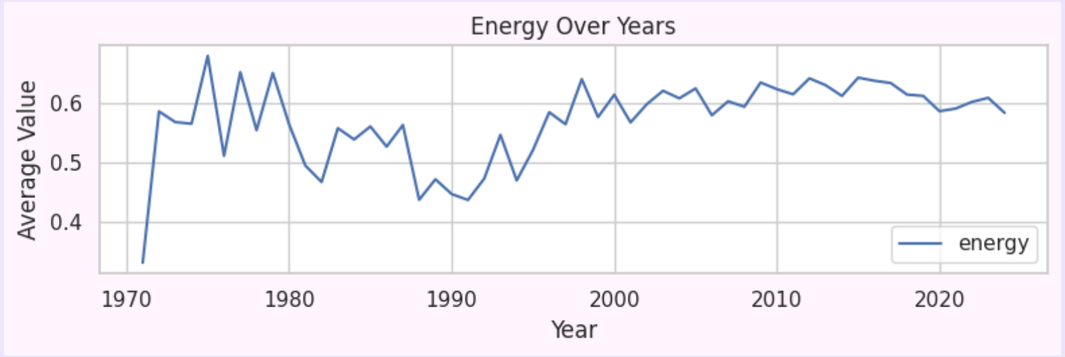
- Song popularity was highest from 1975-1978 before a sharp and prolonged decline after 1980 that lasted through the 90s.
- The 2000s saw a moderate recovery with the rise of digital music, which then stabilized into a consistent but moderate plateau throughout the 2010s.
- After 2020, average popularity has declined again, likely due to streaming oversaturation and shorter lifecycles for modern hit songs.

Yearly Trend in Energy, Danceability and Valence



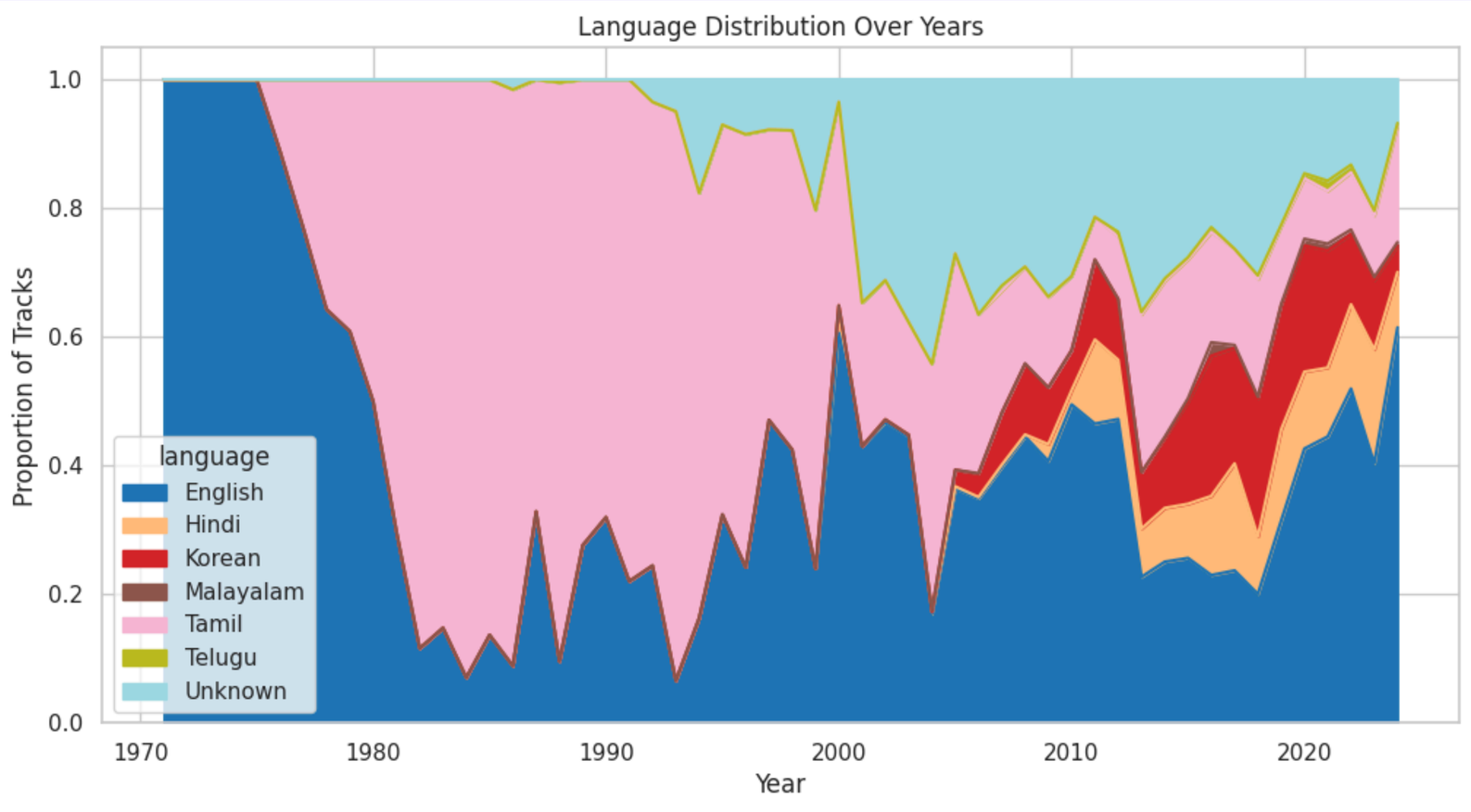
- Average energy in music shows a clear long-term upward trend since the 1970s and has stabilized at a high level for the past two decades.
- Danceability reached its peak in the late 1980s and early 1990s before declining and settling at more moderate levels in the modern era.

- In contrast, average musical valence (positivity) displays a significant long-term decline, with modern music being noticeably less "happy" sounding than in past decades.



Overall, these trends indicate that popular music has evolved to become more energetic, yet emotionally more somber or neutral and less overtly dance-focused than in previous eras.

Language Distribution over the Years



- English music overwhelmingly dominated the 1970s-80s, but its monopoly was later challenged by a surge in Tamil music, which peaked in popularity during the late 80s and 90s.
- The 2000s marked a significant shift towards a more diverse soundscape, with languages like Hindi, Malayalam, and especially Korean gaining steady representation, aligning with the global K-pop wave.
- The current music scene is a globalized mix where English still leads but with reduced dominance, sharing the stage with a strong presence from Hindi, Tamil, and Korean music.

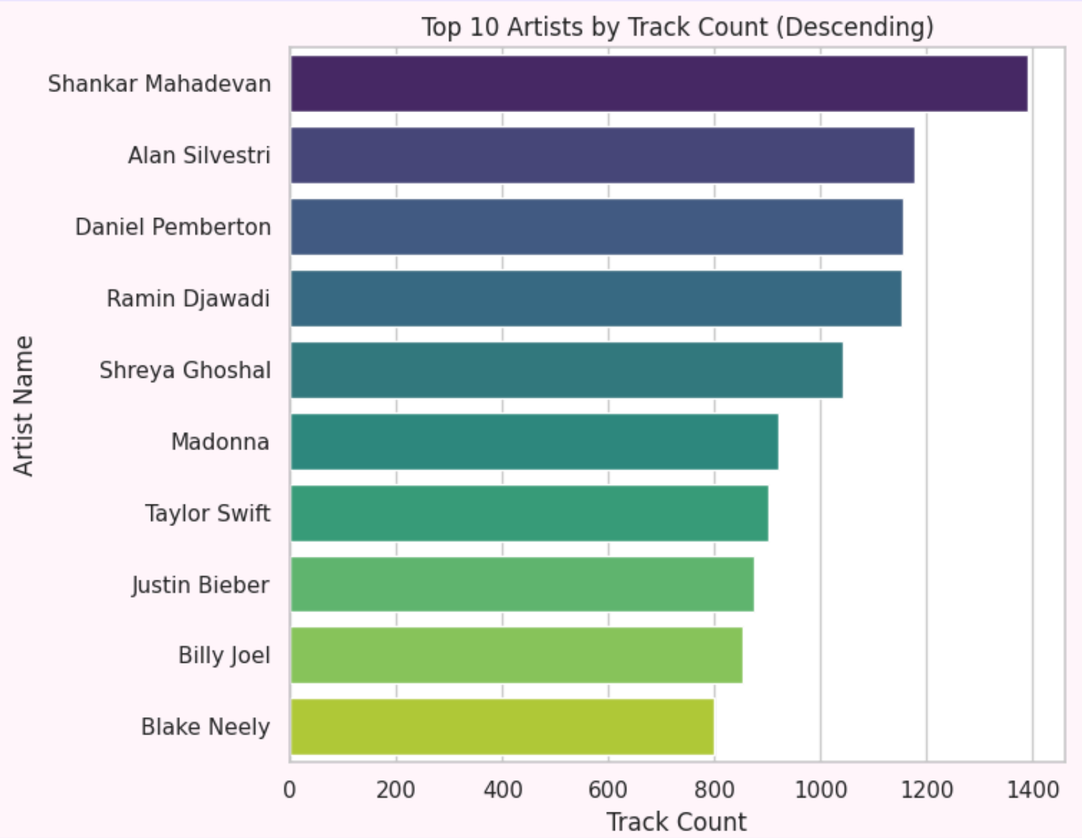
In essence, the data maps the music industry's journey from a largely monolithic, English-dominated market to a fragmented and globalized ecosystem where diverse languages now find significant listenership.

Suggestions

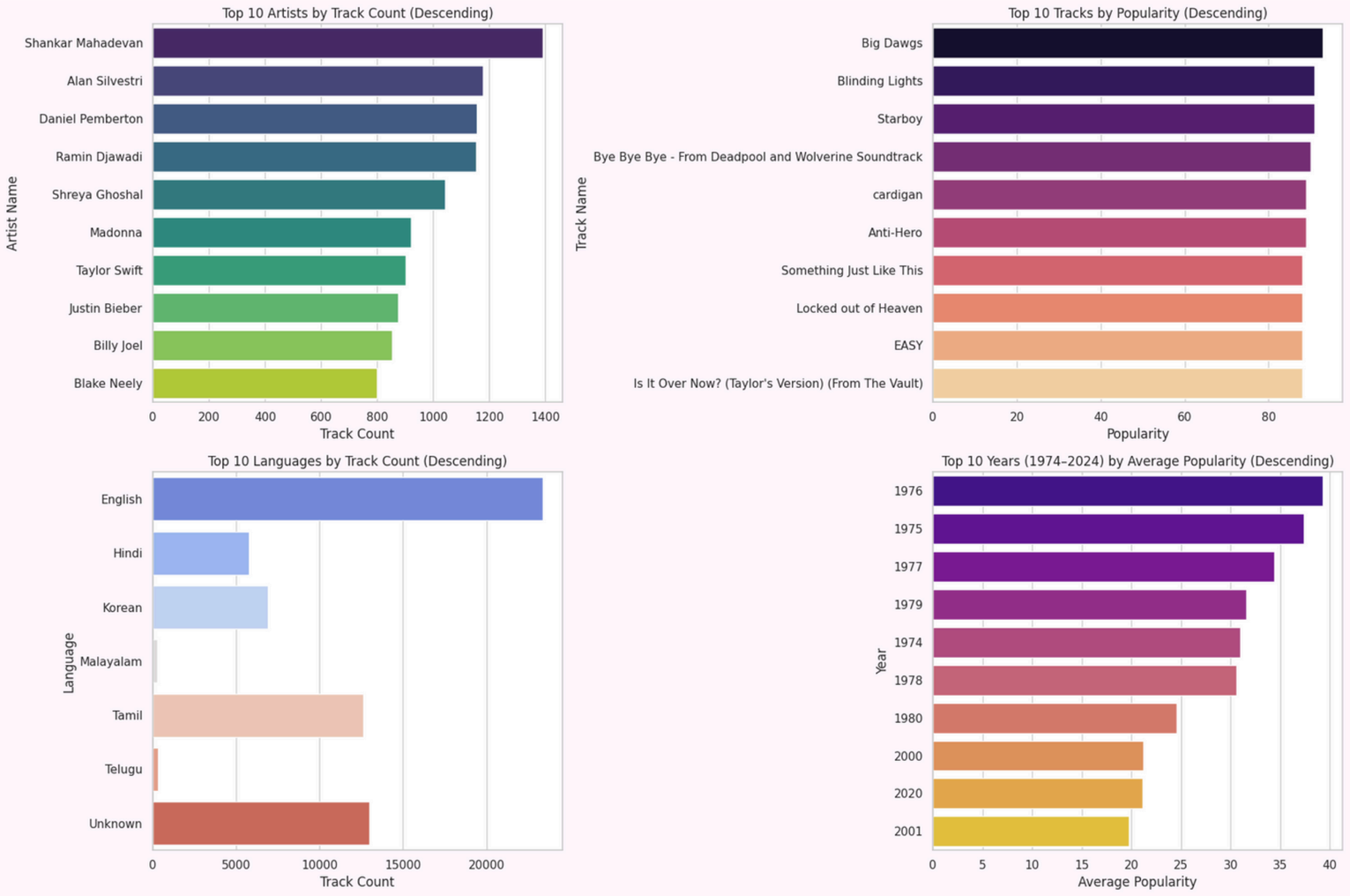
- **Tap into Nostalgia:** Artists should consider reviving classic styles or sampling music from the high-popularity eras of the 1970s and 80s to connect with modern audiences.
- **Balance Energy with Emotion:** Create tracks that are upbeat and energetic but also emotionally resonant, aligning with modern trends of high danceability but lower overall positivity (valence).
- **Invest in Multilingual Music:** Labels and platforms should actively promote regional language artists (especially in Tamil, Hindi, and Korean) and encourage cross-cultural collaborations.
- **Improve Music Discovery with Better Data:** Address gaps in music cataloging, such as the "Unknown" language category, through better metadata management to unlock new regional insights.
- **Adapt to a Globalized Market:** Marketing strategies must now balance global reach with the nurturing of local identity to succeed in today's diverse and multilingual music ecosystem.

Top N/Bottom N Analysis

Top N/Bottom N analysis is a method used to identify and examine the best-performing ('Top N') and worst-performing ('Bottom N') items in a dataset to understand the key drivers of their success or failure.



Top 10 Analysis



- Shankar Mahadevan is the most featured artist with the highest number of tracks, leading a diverse list that includes Indian artists, Western pop stars, and film score composers.

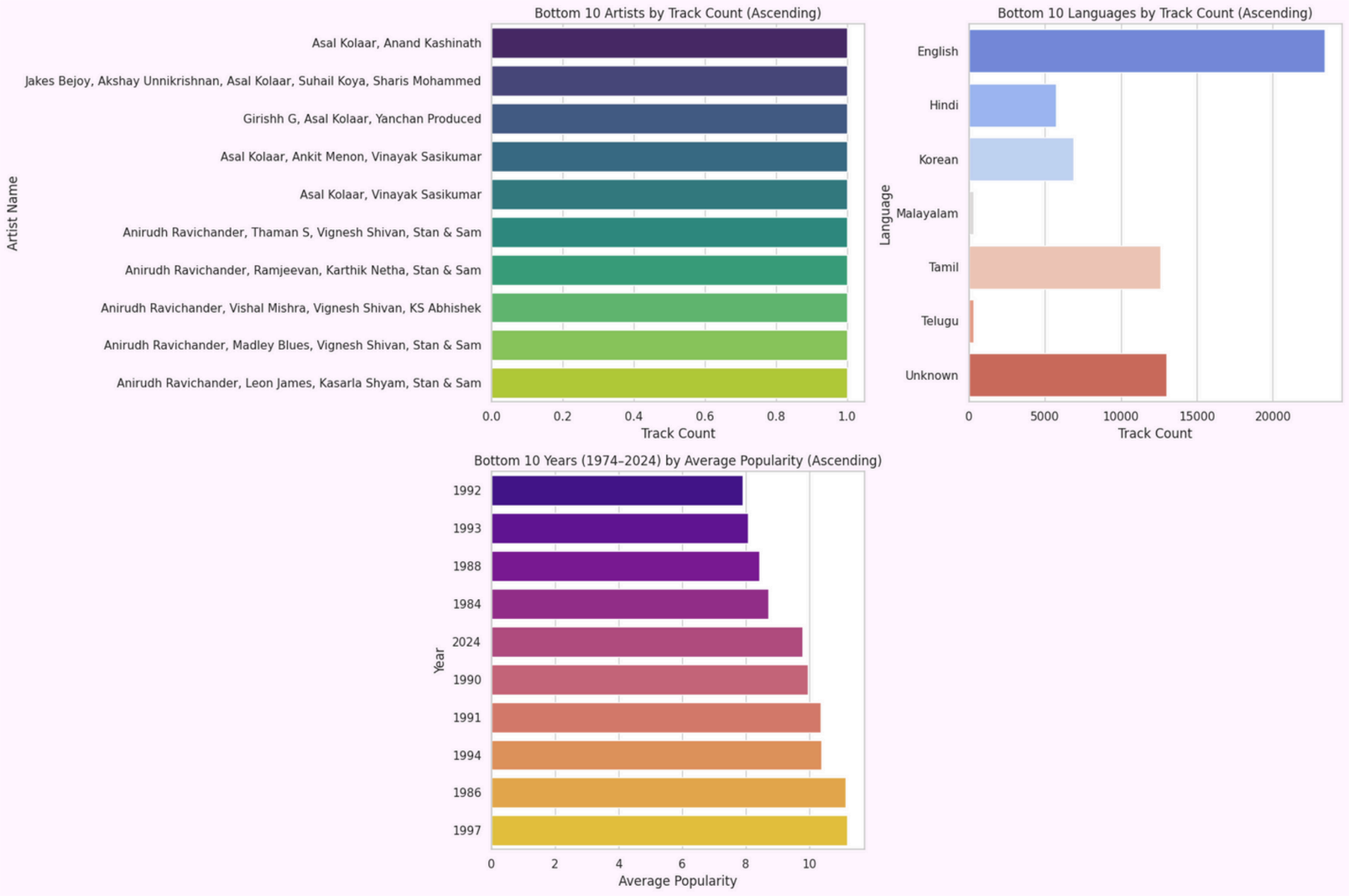
- The top tracks by popularity are major global hits from internationally renowned artists, with songs by The Weeknd, like "Blinding Lights" and "Starboy," scoring the highest.

- English is the most dominant language in the dataset by a significant margin, though a large number of tracks fall into an "Unknown" category, followed by Indian languages like Tamil and Hindi.

- The years with the highest average song popularity are concentrated in the mid-to-late 1970s, with 1976 and 1975 being the top two years.

This analysis highlights a contrast between the dataset's volume (led by prolific Indian artists) and its peak popularity (driven by global Western hits and music from the 1970s).

Bottom 10 Analysis



- The artists with the fewest songs are unique, one-off collaborations featuring multiple names, each with only a single track in the dataset.
- Among the major languages shown, Telugu and Malayalam have the lowest number of tracks, indicating a smaller catalog size compared to others.
- The years with the lowest average song popularity are heavily concentrated in the late 1980s and throughout the 1990s, marking this as the least popular era in the dataset.

This analysis pinpoints the late 80s/90s as the era with the lowest average popularity, while the rarest entries are unique artist collaborations and languages with smaller music catalogs.

Suggestions

- **Capitalize on the "Golden Era":** Since the mid-to-late 1970s is consistently the era with the highest average popularity, create curated playlists, marketing campaigns, or even new music inspired by the sounds of that decade to tap into powerful nostalgia.
- **Distinguish Between Prolific Artists and Popular Hits:** Recognize that the most prolific artists (like Shankar Mahadevan) and the most popular tracks (by artists like The Weeknd) serve different audience segments. Develop a dual strategy that supports the deep catalogs of regional artists while separately promoting high-performing global hits.
- **Explore the Untapped 90s Catalog:** The late 1980s and 1990s were identified as the least popular era. This presents an opportunity to find "hidden gems" or create content that reintroduces this decade's music to new listeners, potentially carving out a unique niche.
- **Invest in Regional and Niche Language Markets:** While English is dominant, the analyses show a large catalog of Indian languages. Investing in better curation and promotion for languages with smaller catalogs, like Telugu, can capture dedicated and underserved audiences.
- **Improve Data on Collaborations and Languages:** The large "Unknown" language category and the complex, multi-artist collaborations in the bottom-tier suggest a need for better data management. Cleaning up this metadata will improve music discovery and provide clearer insights into listening habits.

Final Recommendations

- **Capitalize on Nostalgia:** Actively promote and create content around the high-popularity music of the 1970s through retro playlists and marketing campaigns to engage listeners.
- **Engineer "Feel-Good" Hits:** Prioritize creating tracks with high danceability and valence for mainstream appeal, but balance this with emotionally resonant, moodier tones to align with modern listening trends.
- **Adopt a Global-Local Music Strategy:** Continue investing in English-language hits for global reach while actively growing and promoting catalogs for rising regional languages like Hindi, Tamil, and Korean.
- **Focus Promotion Beyond Audio Features:** Since popularity isn't guaranteed by sonic traits, shift strategic focus to artist branding, high-visibility collaborations, and targeted playlist placement to drive success.
- **Unlock the Value of Niche Catalogs:** Develop curation strategies to highlight "hidden gems" from the less popular 1990s and promote emerging, one-off artists to drive discovery and user engagement.
- **Prioritize Metadata Cleanup:** Invest resources in fixing data inconsistencies, especially the "Unknown" language category, to improve music discoverability and unlock more accurate insights.
- **Implement a Dual Content Strategy:** Balance the promotion of guaranteed "blockbuster" artists and tracks with the creation of discovery pathways for "underdog" content to maximize both engagement and long-term listener loyalty.

Thank You.