manhattan project

Spotify Data Analysis

EXPLORATORY DATA ANALYSIS (EDA) OF SPOTIFY TRACKS

In any data science project, before building complex models or making definitive conclusions, the first and most critical step is to understand the data. This is the purpose of Exploratory Data Analysis (EDA). This document will explain the fundamentals of EDA, its associated concepts, its critical importance across various industries, and detail a data science project focused on analyzing Spotify tracks.

OBJECTIVE

To analyze the Spotify music dataset to uncover patterns in audio features, track popularity, and linguistic trends, focusing on 2024 Tamil tracks, in order to derive actionable insights for playlist curation, music production strategies, and predictive modeling, enhancing personalized listener experiences and informing data-driven decisions in the music industry.

Data Description

Spotify Tracks: Data Dictionary

Identifiers & Basic Info

- track_id: Unique identifier for the track.
- track_name: The title of the song.
- artist_name: The name of the artist(s).
- album_name: The name of the album.
- artwork_url: URL for the album artwork.

Popularity & Temporal

- popularity: How popular a track is (0 to 100).
- year: The release year of the song.
- duration_ms: The track duration in milliseconds.

Key Audio Features

- acousticness: Confidence the track is acoustic.
- danceability: Suitability of the track for dancing.
- energy: Perceptual measure of intensity and activity.
- instrumentalness: Predicts lack of vocal content.
- liveness: Detects presence of an audience/live recording.
- speechiness: Detects the presence of spoken words.
- valence: Musical positiveness (e.g., happy, cheerful).

01

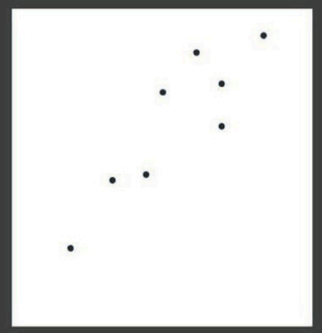
Univariate Analysis

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

02

Bivariate Analysis

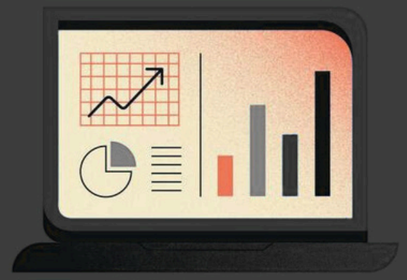
Analysis of the relationship between two variables



03

Multi-variate Analysis

Analysis of more than two variables simultaneously to study complex relationships



04

Outliers Analysis

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

05

Time-series Analysis

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

Univariate Analysis

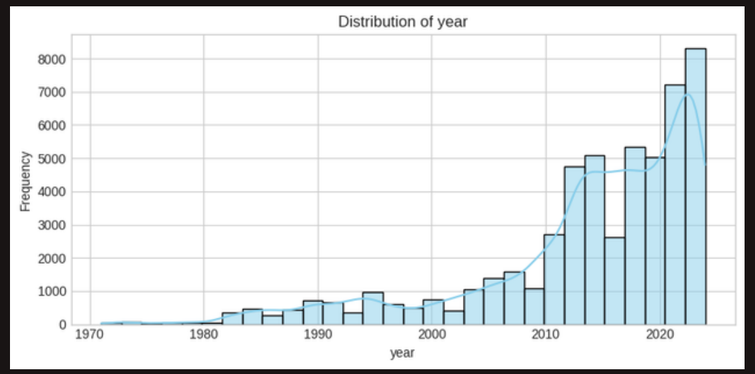
Univariate Analysis is the foundational step of EDA, examining features one at a time to understand their inherent properties. It describes the data's central tendency (mean, median), its spread (variance, standard deviation), and its shape. For quantitative features, we use histograms and density plots to visualize distribution and box plots to spot outliers. For categorical data, bar charts and frequency tables reveal the composition of the dataset, providing the essential context needed before exploring relationships.

	count	mean	std	min	25%	50%	75%	max
year	62239.0	2014.417969	9.648517	1971.0	2011.0	2017.0	2022.0	2024.0
popularity	62239.0	15.357589	18.630494	0.0	0.0	7.0	26.0	93.0
acousticness	62239.0	0.362342	0.314674	-1.0	0.0671	0.286	0.633	0.996
danceability	62239.0	0.596768	0.186262	-1.0	0.497	0.631	0.73	0.986
duration_ms	62239.0	242603.445557	113021.005225	5000.0	192240.0	236311.0	286303.5	4581483.0
energy	62239.0	0.602416	0.246207	-1.0	0.44	0.639	0.803	1.0
instrumentalness	62239.0	0.146054	0.307637	-1.0	0.0	0.000025	0.0151	0.999
liveness	62239.0	0.194172	0.172075	-1.0	0.0932	0.125	0.243	0.998
loudness	62239.0	-65.174856	2370.534662	-100000.0	-10.729	-7.506	-5.455	1.233
speechiness	62239.0	0.087741	0.115208	-1.0	0.0367	0.0489	0.0891	0.959
tempo	62239.0	117.923713	28.505003	-1.0	95.94	117.99	135.0685	239.97
time_signature	62239.0	3.857003	0.502881	-1.0	4.0	4.0	4.0	5.0
valence	62239.0	0.495246	0.264785	-1.0	0.292	0.507	0.71	0.995
duration_s	62239.0	242.603446	113.021005	5.0	192.24	236.311	286.3035	4581.483
duration_min	62239.0	4.001355	1.429208	0.083333	3.204	3.938517	4.771725	10.0
is_explicit	62239.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
is_high_popularity	62239.0	0.072688	0.259625	0.0	0.0	0.0	0.0	1.0
language_encoded	62239.0	2.417182	2.371919	0.0	0.0	2.0	4.0	6.0

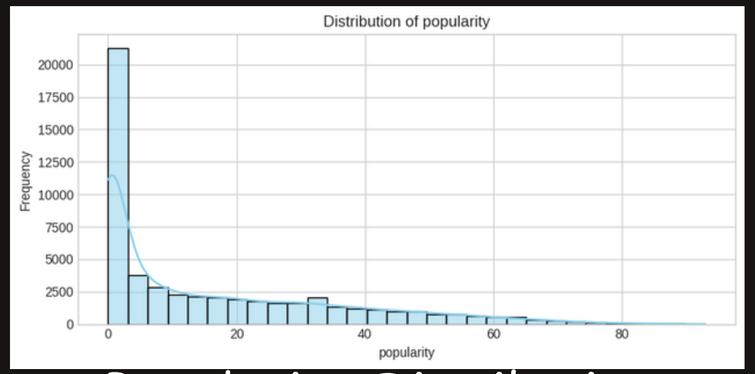
We are working with a large-scale dataset encompassing a vast amount of music history and feature information.

- Total Tracks (Rows): Analyzing over 62239 unique songs.
- Total Features (Columns): Exploring 18 distinct metadata and audio feature columns. (e.g., 22)

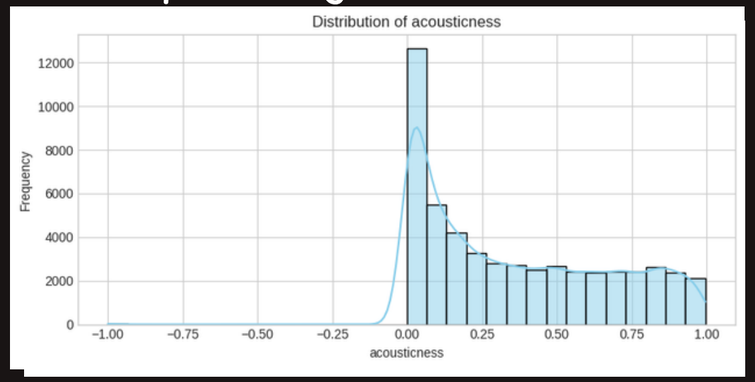
The dataset we are analyzing provides a remarkable historical view, spanning approximately a full century of recorded music. Our analysis covers trends and feature changes from tracks dating back to the 1920s all the way up to the present day.



Year Distribution

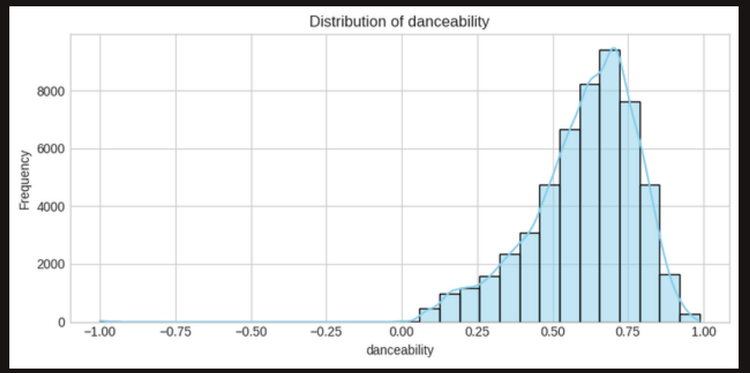


Popularity Distribution

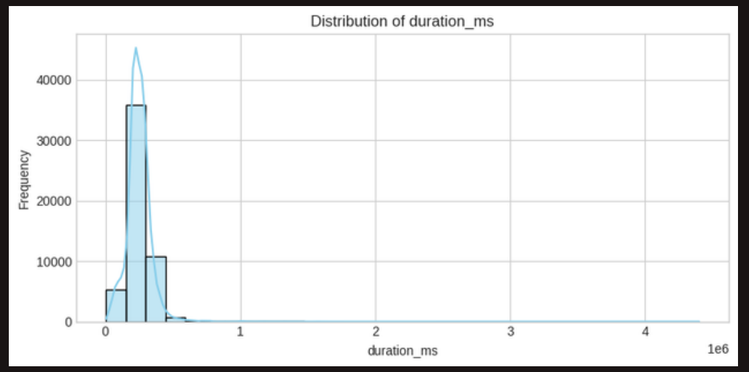


Acousticness Distribution

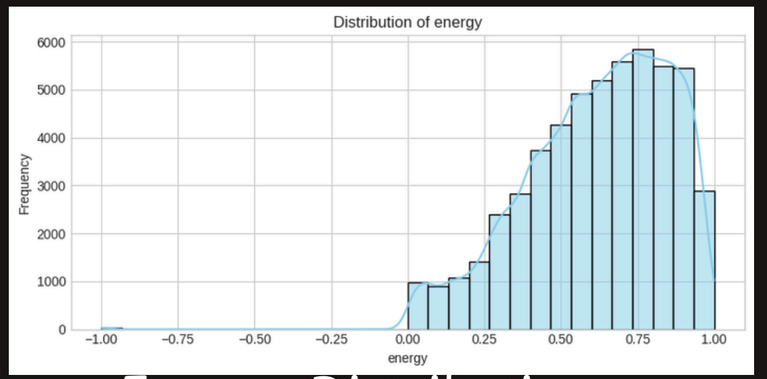
1



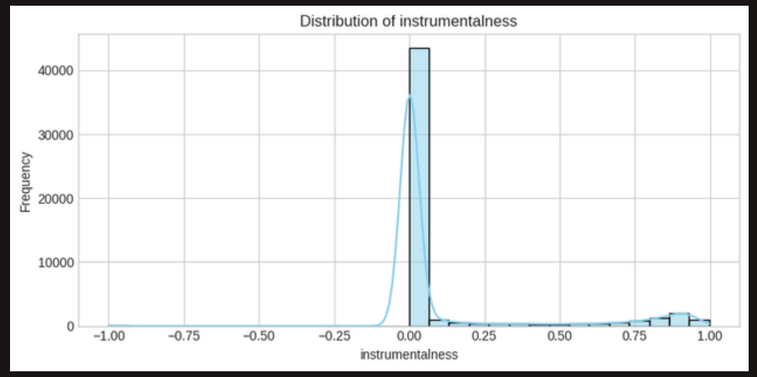
Danceblity Distribution



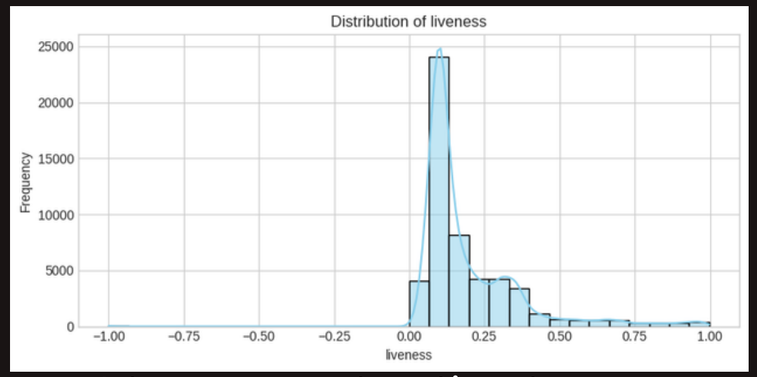
Duration Distribution



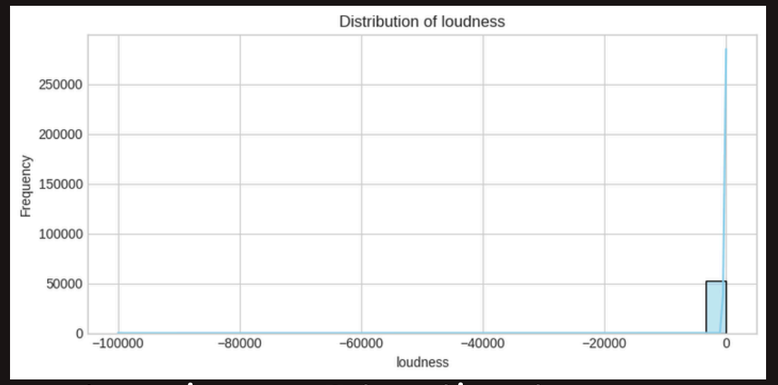
Energy Distribution



Instrumentalness Distribution

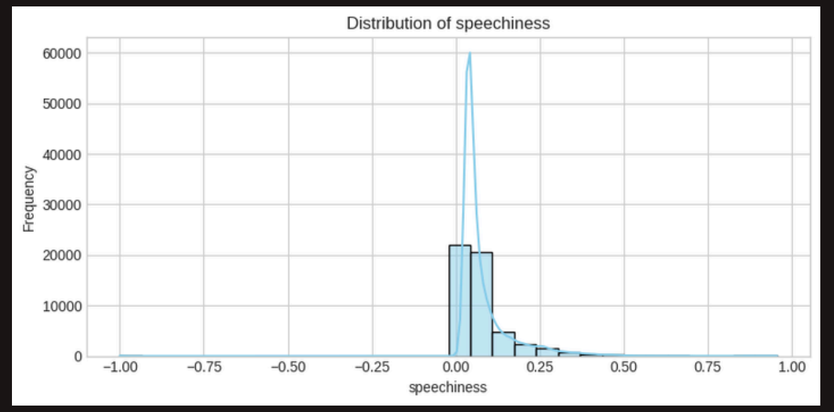


Liveness Distribution

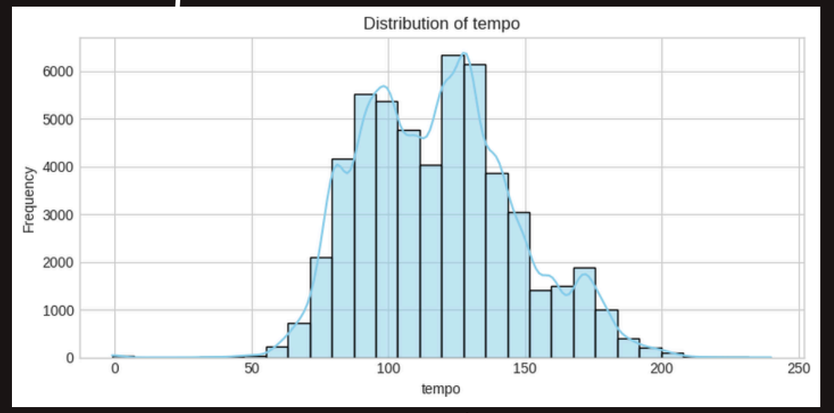


Loudness Distribution

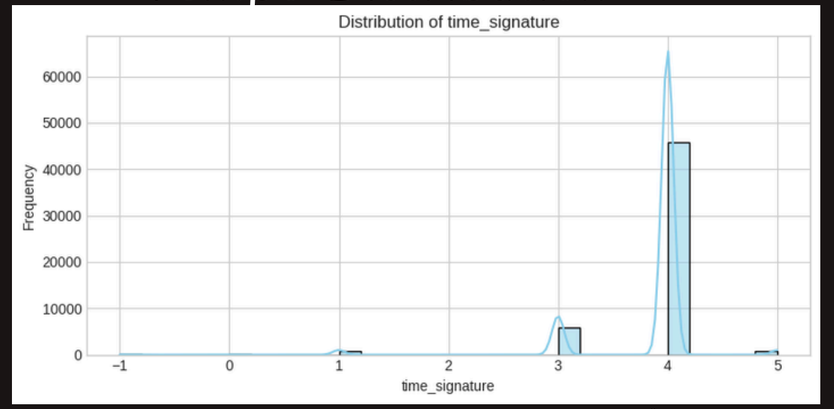
2



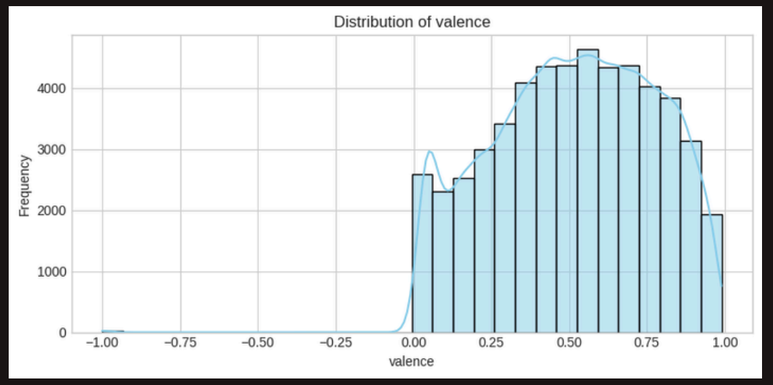
Speechiness Distribution



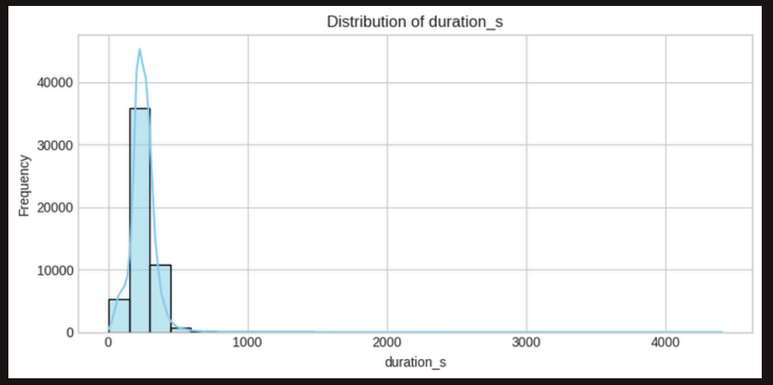
Tempo Distribution



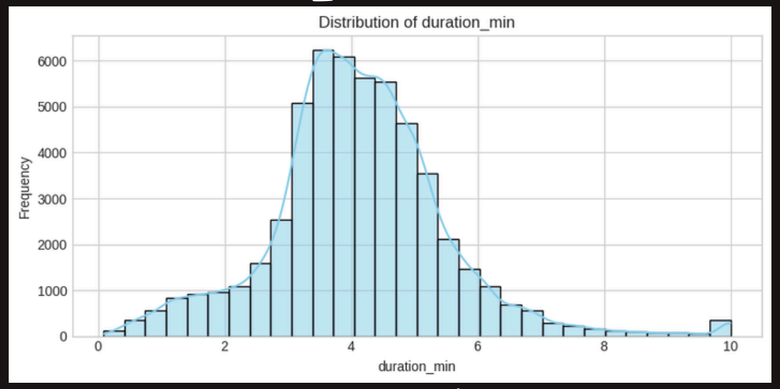
Time Signature Distribution



Valance Distribution

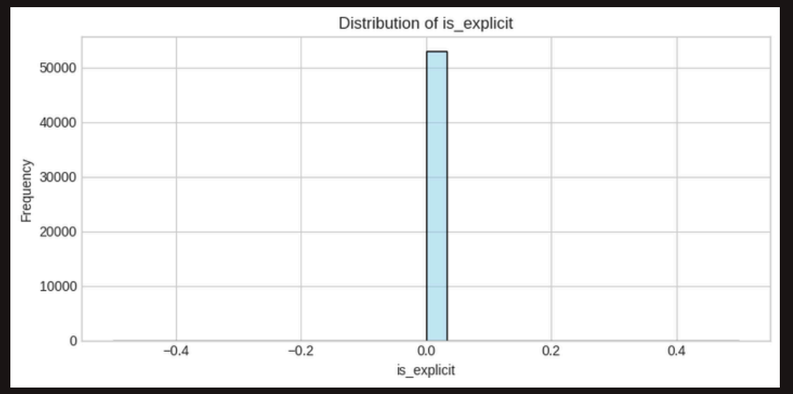


Duration_s Distribution

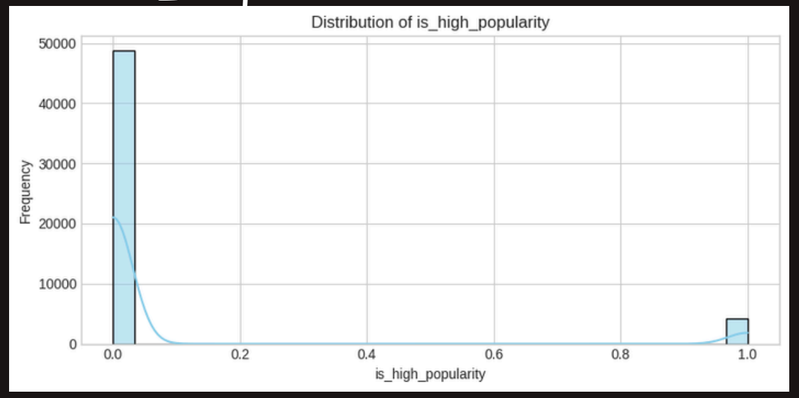


Duration Distribution

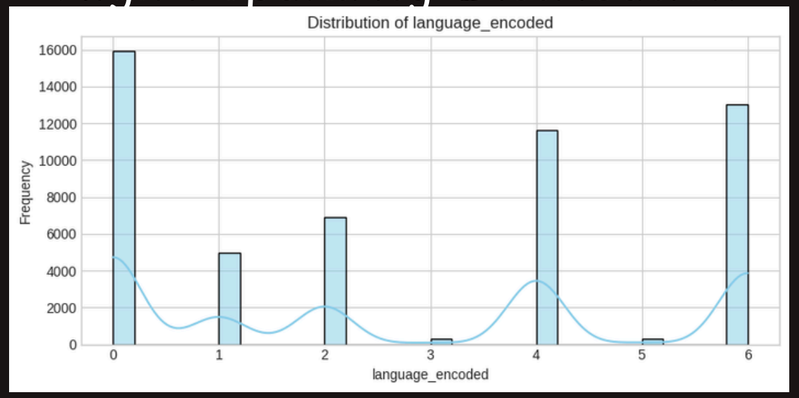
3



Is_Explicit Distribution



High Popularity Distribution



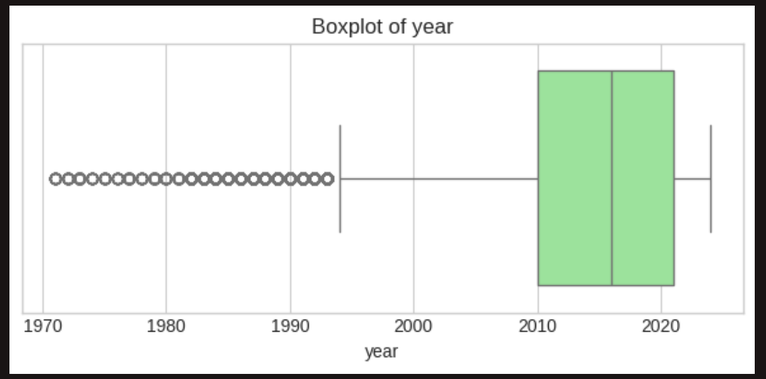
Language Encoded Distribution

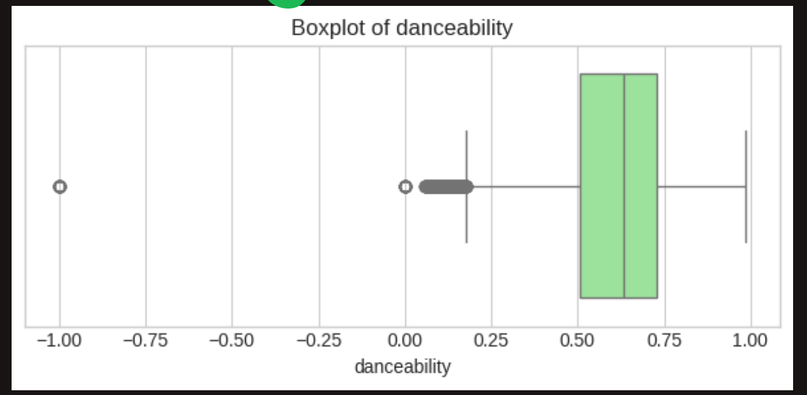
Key Insights

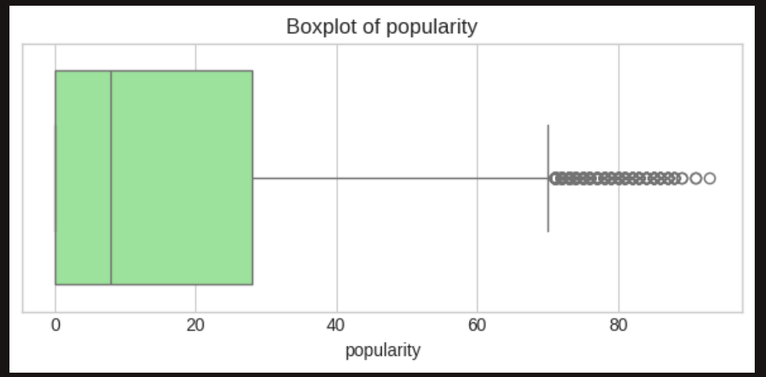
- Skewed Distributions: Danceability and energy are right-skewed (centered ~ 0.6 – 0.8), indicating upbeat, dance-friendly tracks; acousticness and instrumentalness are left-skewed (~0), showing non-acoustic, vocal-heavy songs.
- Tempo and Duration: Tempo peaks at 120–170 BPM, ideal for EDM; duration_ms clusters around 1.5– 3.5 minutes, typical for singles.
- Popularity and Valence: Popularity skews lower (20–60), reflecting niche tracks; valence (0.4–0.8) suggests preference for positive, energetic songs.
- Technical Features: Loudness (–5 to –11 dB) is tightly distributed; low speechiness and liveness indicate studio-recorded, non-spoken tracks.
- Regional Trends: 2024 Tamil tracks emphasize high danceability/energy, hinting at orchestral EDM or experimental mix trends.
- EDA Utility: Histograms reveal skewness, outliers, and multimodal patterns, guiding feature engineering and clustering for sub-genre discovery.

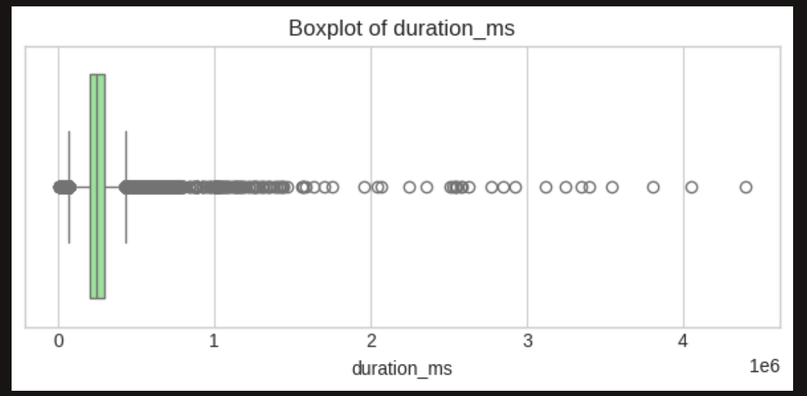
- Examining Outliers and Spread of Numerical Audio Features
- Distribution of Categorical & Low-Cardinality Features
- Distribution of Track Languages
- Others

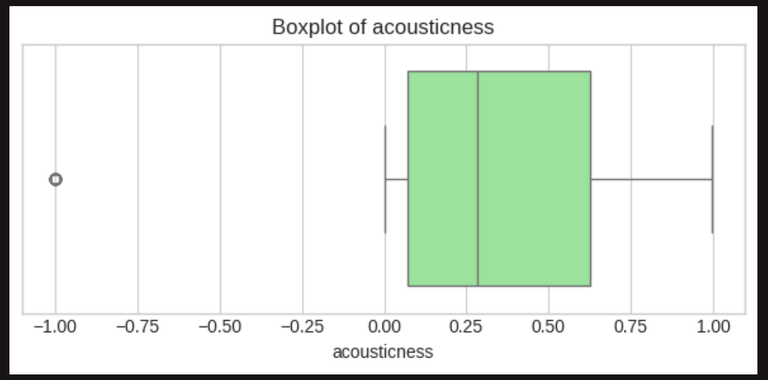
Univariate Analysis

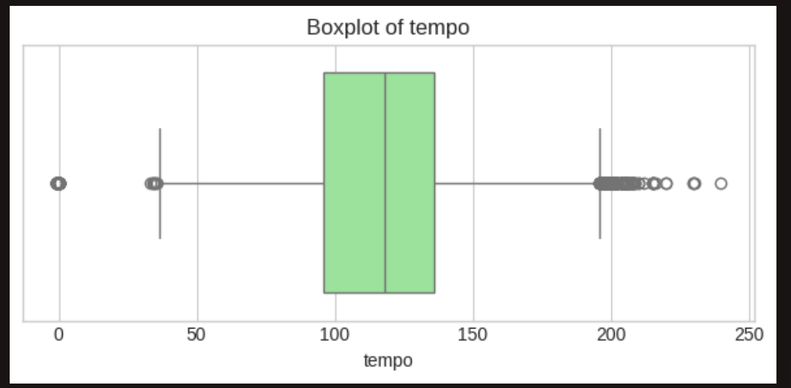








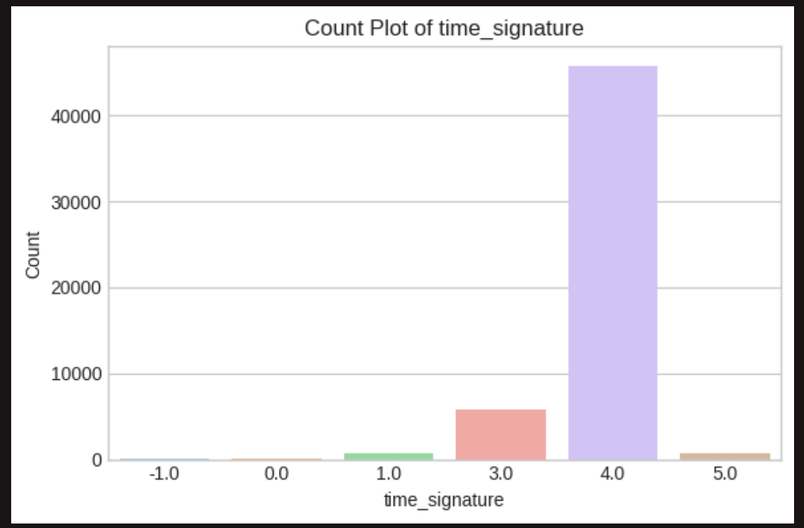


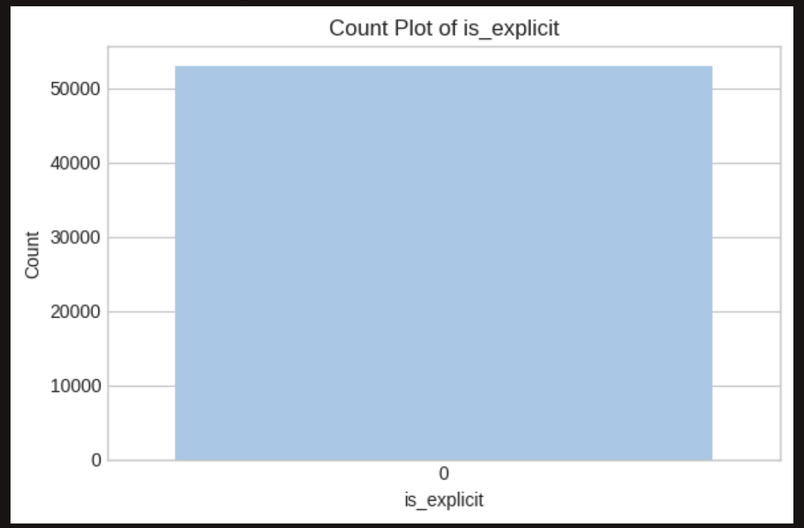


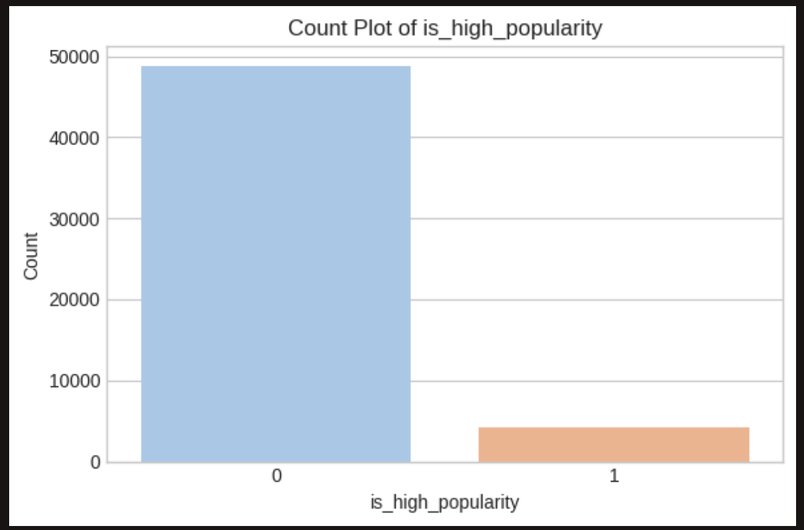
- Examining Outliers and Spread of Numerical Audio Features
- Distribution of Categorical & Low-Cardinality Features

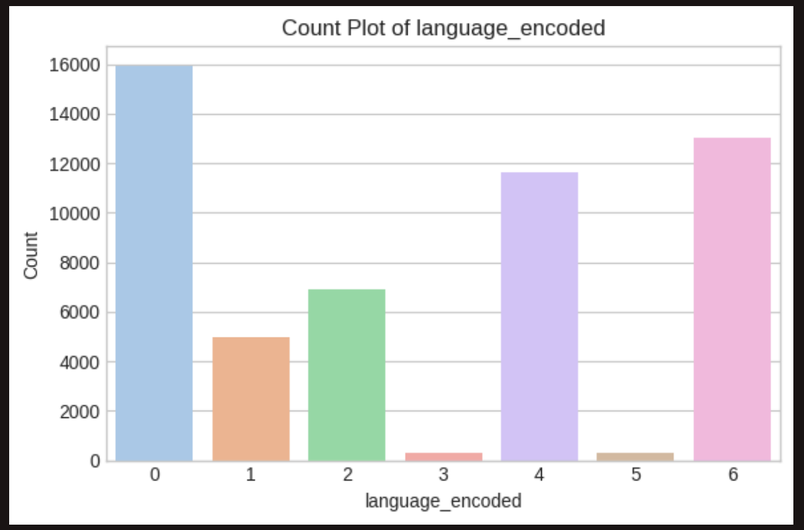
- More Deeper Analysis

Univariate Analysis







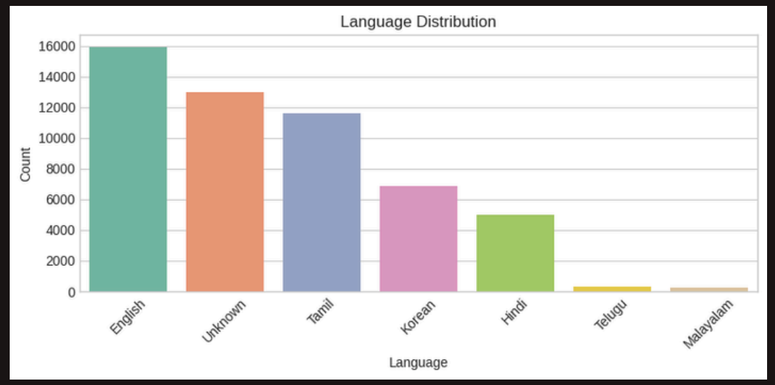


Univariate Analysis

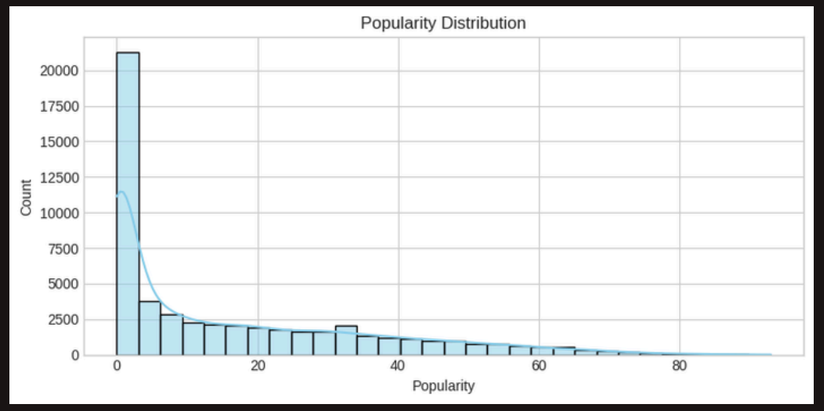
Examining Outliers and Spread of Numerical Audio Features

Distribution of Categorical & Low-Cardinality Features

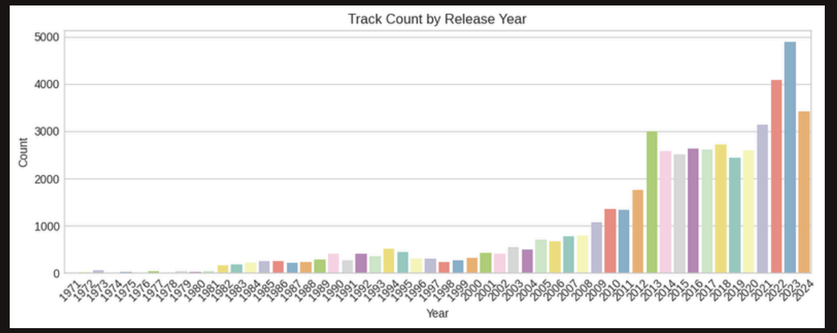
More Deeper Analysis



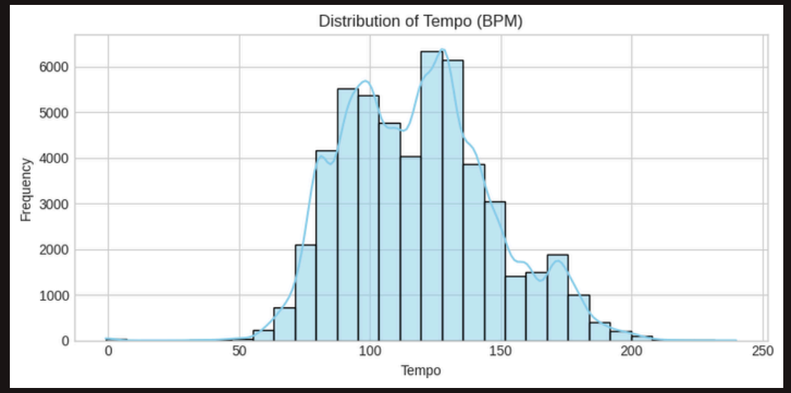
Distribution of Track Languages



Distribution of Track Popularity



Distribution of Spotify Tracks by Release Year



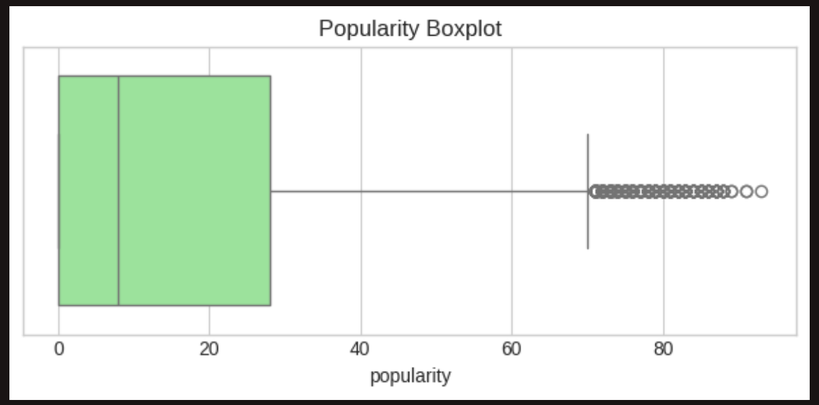
Frequency Distribution of Tempo

Univariate Analysis

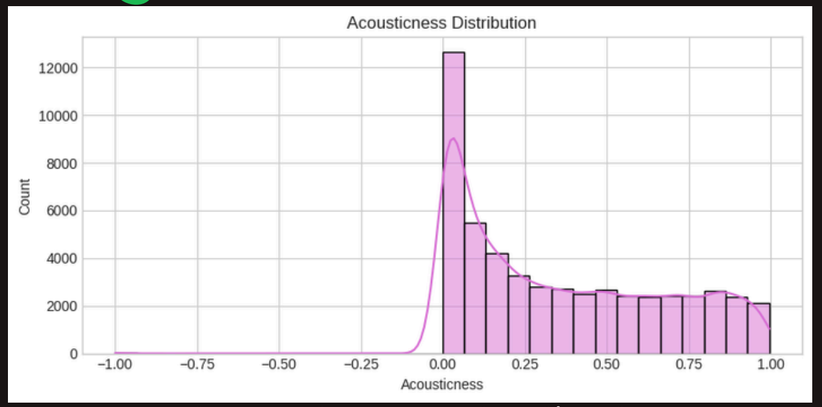
Examining Outliers and Spread of Numerical Audio Features

Distribution of Categorical & Low-Cardinality Features

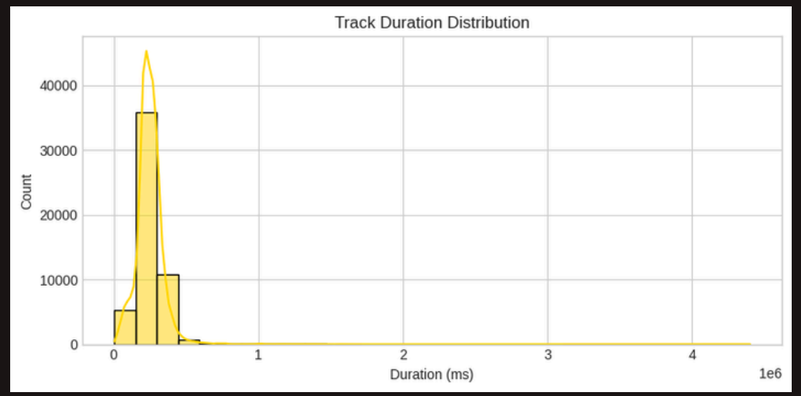
More Deeper Analysis



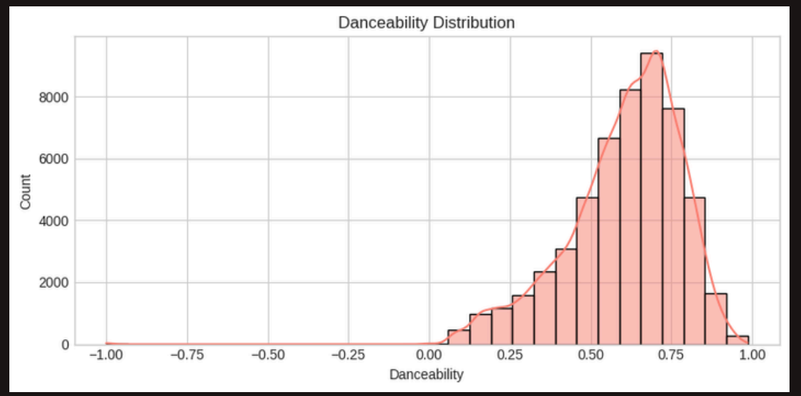
Boxplot of Popularity



Acousticness Distribution

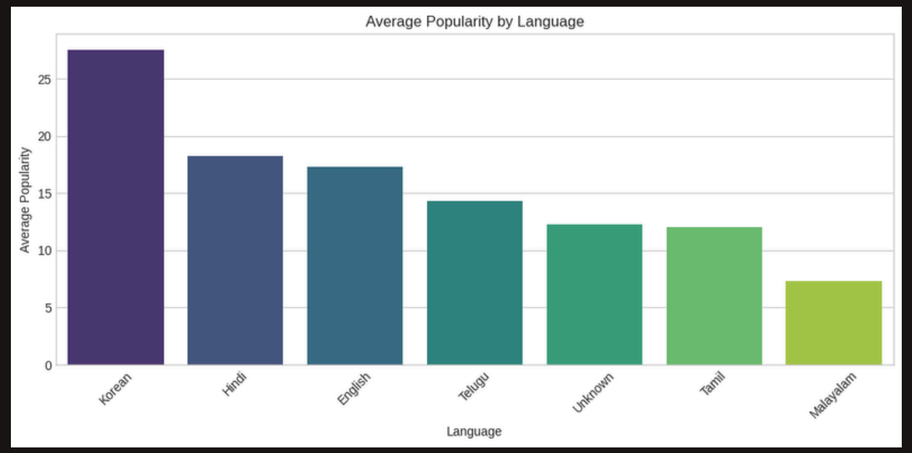


Track Duration Distribution

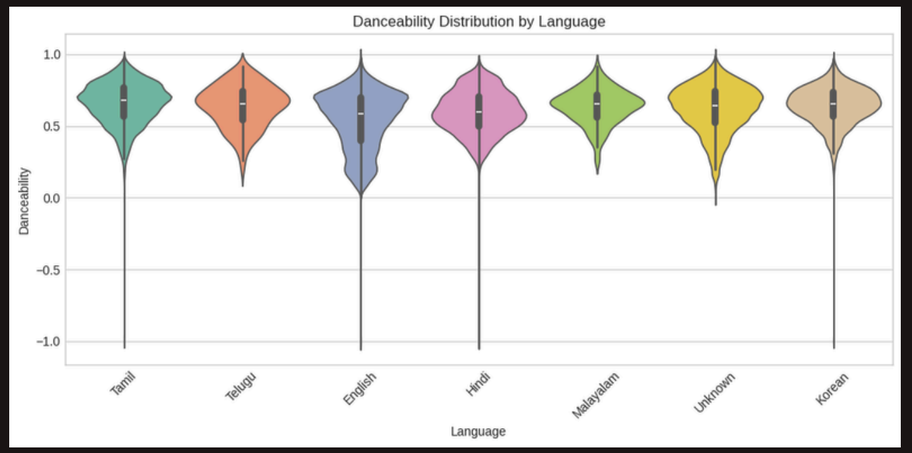


Dancability Distribution

Catagorical Variable Trends - A. Language Trends

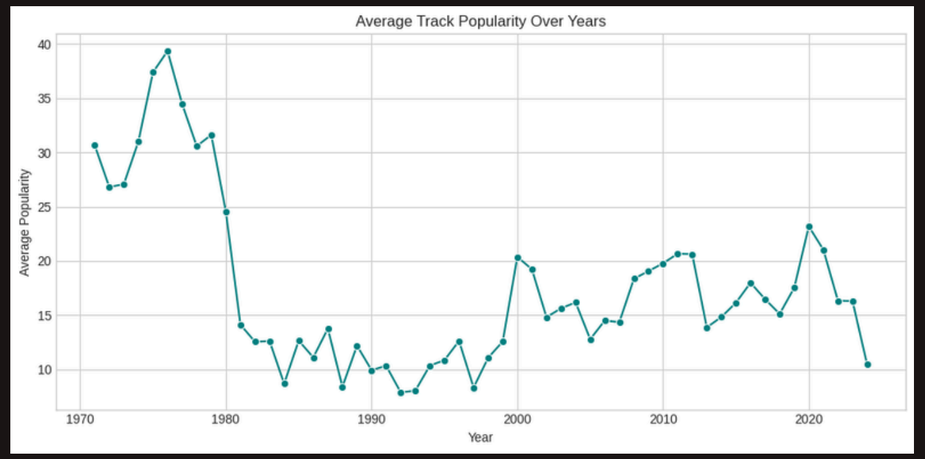


- The chart displays average popularity for songs grouped by language.
- Korean tracks have the highest average popularity in the dataset.
- Hindi and English follow, while Telugu, Tamil, and Malayalam have lower averages.
- The visualization highlights significant differences across languages, indicating varied listener preferences.

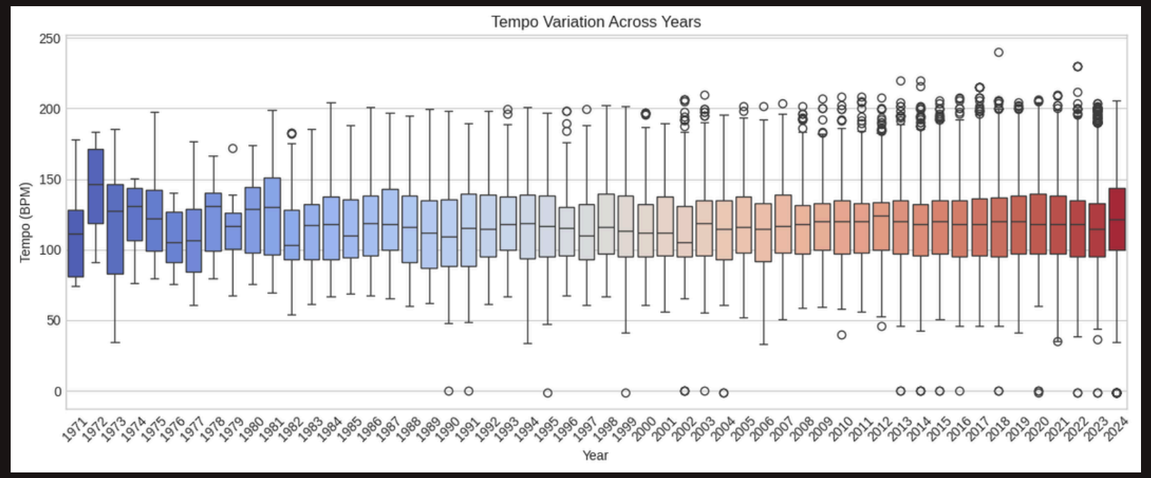


- The violin plot illustrates the spread and density of danceability scores for songs in each language.
- Most languages show a fairly high median danceability around 0.6 to 0.7, with some variation in spread.
- Tamil and Telugu display wider distributions, indicating more diversity in danceability within those languages.
- Malayalam, Unknown, and Korean have relatively tighter distributions, showing more consistency in danceability.
- The plot effectively conveys differences in rhythmic characteristics of music across languages, useful for exploring genre or cultural trends.

B. Yearly Trends

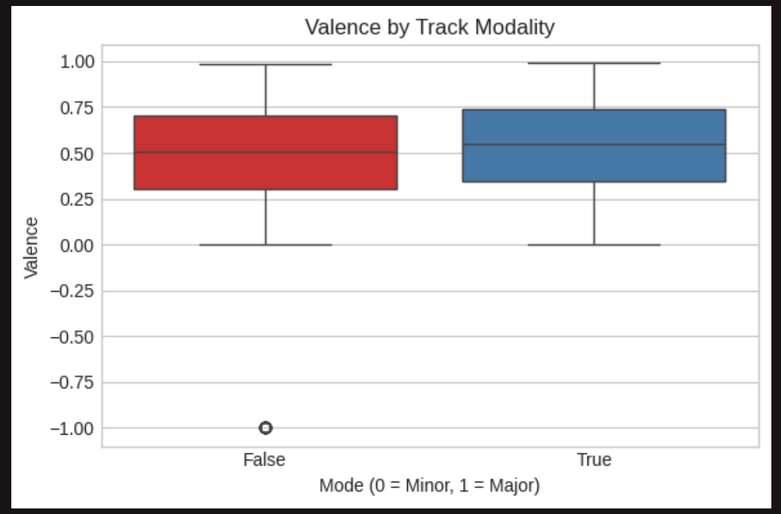


- Peak popularity (~40) in the early 1970s declined sharply to 10–15 by the 1990s, with a slight rise (~20) around 2000.
- Fluctuations continued in the 2010s (20–25), followed by a drop to 15–20 post-2020, including 2024 Tamil tracks (22–59).
- Long-term decline with variability suggests shifts in listener preferences and technological impacts.



- Tempo ranges from 50 to 250 BPM, with a general median around 100–150 BPM across years, reflecting diverse musical styles.
- Early years (1970s) show wider variability and higher outliers (up to 200–250 BPM), while recent years (2020–2024) stabilize around 100–150 BPM, indicating a trend toward moderate tempos.
- Outliers in the 1990s and 2010s suggest occasional experimental tracks, with a gradual shift to consistent tempos in the 2024 Tamil tracks (e.g., 110–170 BPM in sample).
-

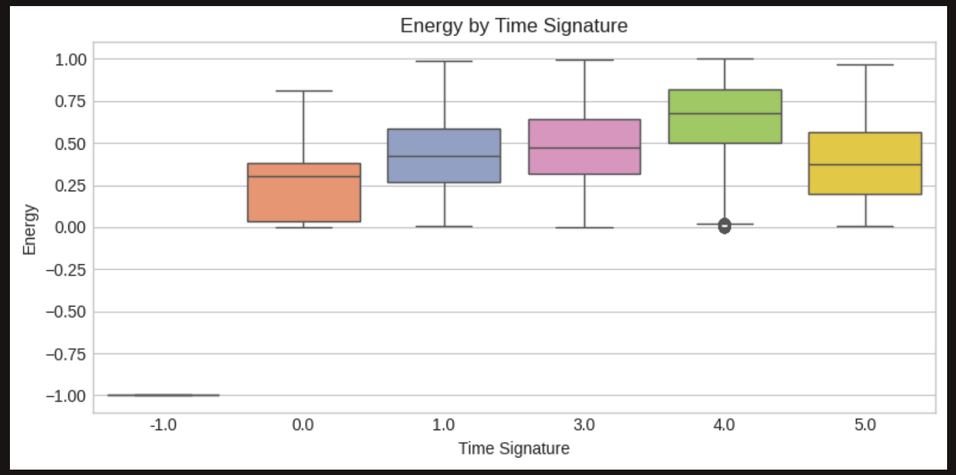
C. Mode (Major vs Minor)



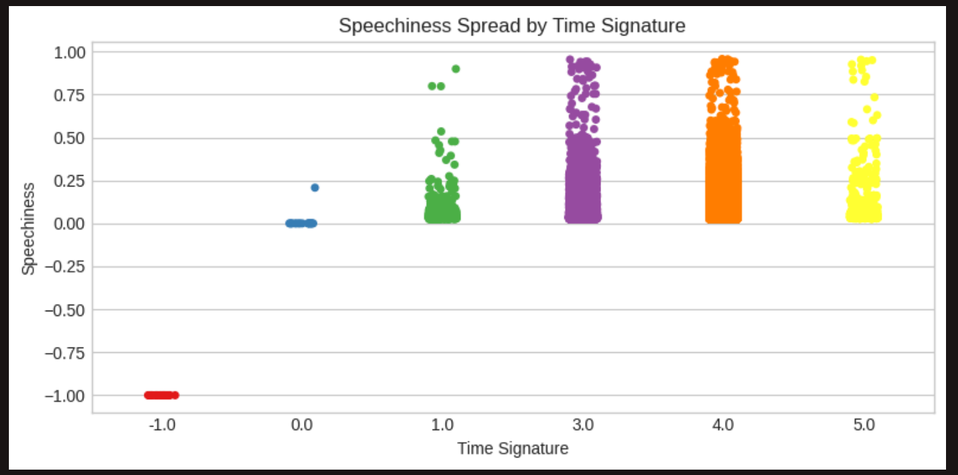
- Major mode (1) tracks show a median valence around 0.5–0.75, indicating a tendency toward positive, upbeat tones.
- Minor mode (0) tracks have a similar median valence (~0.5), with a wider range and a significant outlier below -1.0, suggesting diverse emotional expressions.
- The overlap in valence distributions highlights that modality alone doesn't strictly determine positiveness, especially in 2024 Tamil tracks blending minor keys with high energy.

- Loudness for both minor (0) and major (1) mode tracks is tightly clustered around 0 dB, indicating consistent mastering levels across modalities.
- A significant outlier drops to approximately -10,000 dB for minor mode, suggesting rare extremely quiet tracks, possibly due to data anomalies or unmastered recordings.
- No substantial difference in loudness distribution between modes, reflecting uniform production standards in the 2024 Tamil track dataset.

D. Time Signature Trends



- Time signatures 1.0, 3.0, 4.0, and 5.0 show median energy around 0.25–0.75, with 4.0 having the highest median (~0.5–0.75), indicating energetic tracks in common time signatures.
- 1.0 (unusual signature) has a lower median (~0.25) and narrower range, suggesting less dynamic tracks.
- Outliers extend to negative values for 3.0 and 5.0, reflecting rare low-energy tracks, while 4.0 dominates in the 2024 Tamil dataset for high-energy EDM styles.

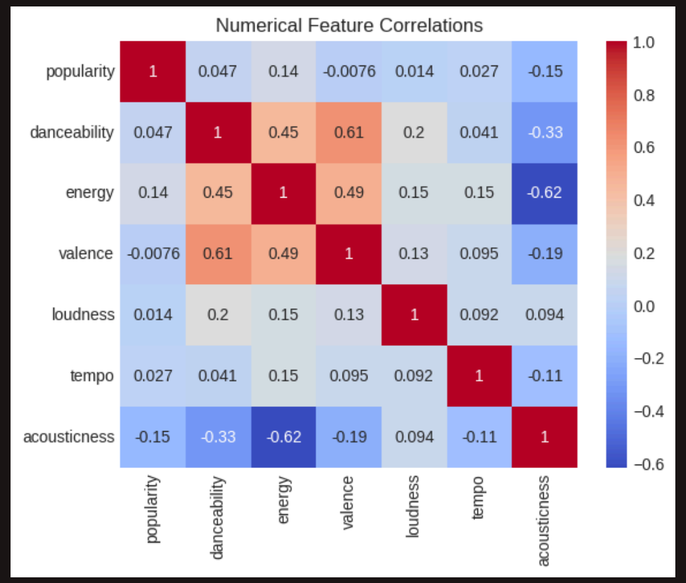


- Time signatures 1.0, 3.0, 4.0, and 5.0 show speechiness ranging from 0 to 1.0, with 4.0 and 5.0 having higher concentrations (0.25–0.75), indicating more spoken content in common signatures.
- 1.0 has a sparse distribution with low speechiness (<0.25), suggesting minimal spoken elements in unusual time signatures.
- An outlier at -1.0 for 1.0 may indicate a data anomaly, while the spread in 2024 Tamil tracks reflects diverse vocal styles across signatures.

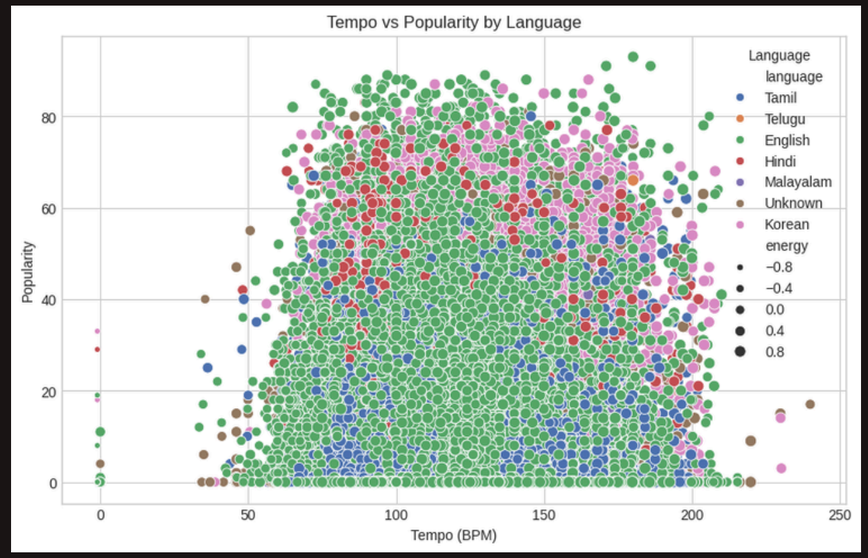
Spotify Data

Bivariate Analysis

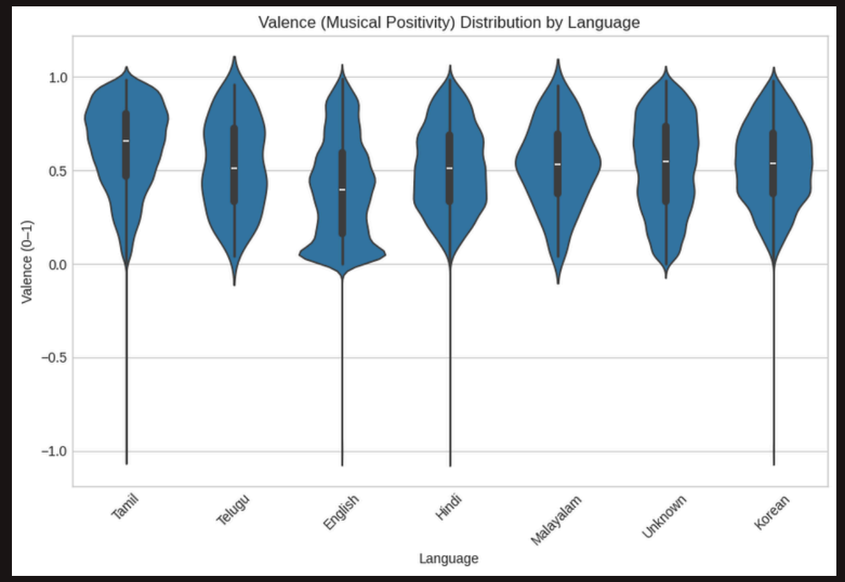
Bivariate analysis examines the relationship between two variables.
It helps identify patterns, correlations, or associations between them.
Common methods include scatter plots, correlation coefficients, and cross-tabulations.
This analysis is useful for understanding how one variable changes with respect to another.

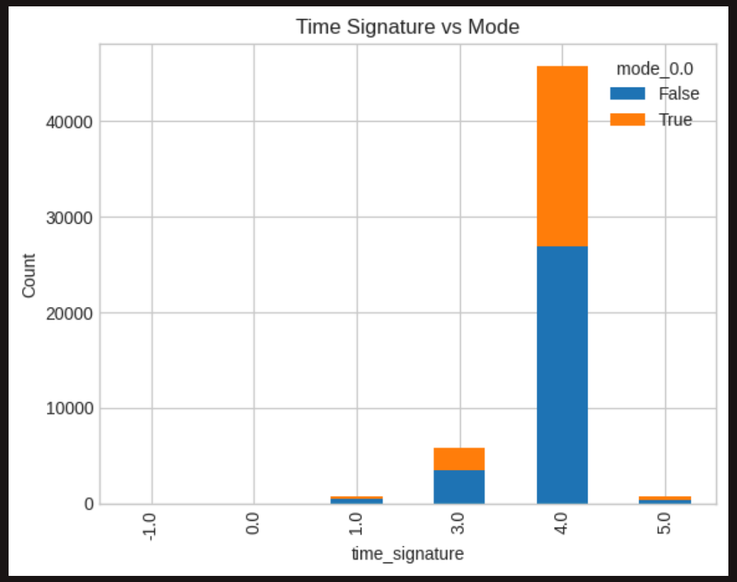


- Strong positive correlations: danceability and energy (0.48), energy and valence (0.54), indicating upbeat, positive tracks are often energetic.
- Moderate positive correlation: popularity and danceability (0.44), suggesting danceable tracks tend to be more popular.
- Negative correlations: acousticness with energy (-0.62) and danceability (-0.34), showing acoustic tracks are less energetic or danceable.
- Weak correlations: loudness and tempo (0.09), popularity and loudness (0.14), indicating minimal influence on popularity or tempo variation in the 2024 Tamil dataset.



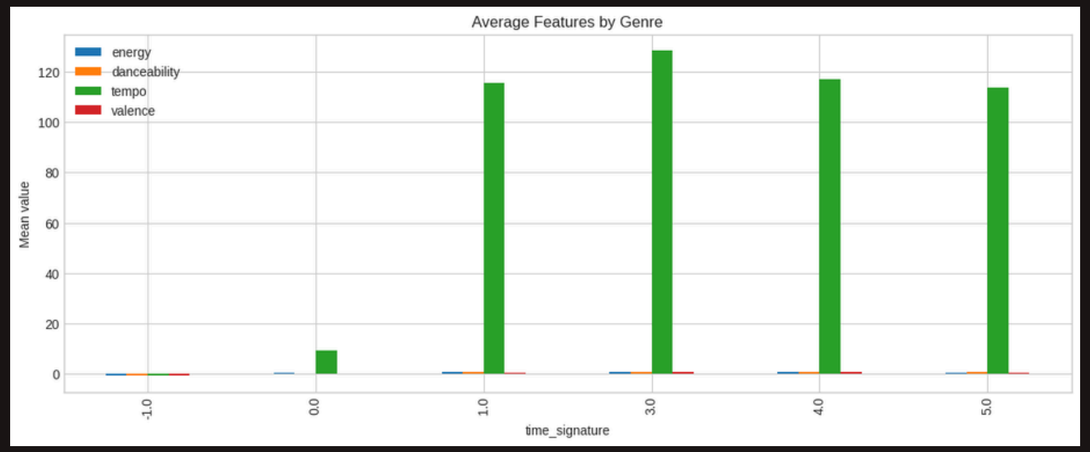
- Tamil tracks (green) dominate with tempos 100-150 BPM and popularity 20-80, reflecting high energy and regional appeal in the 2024 dataset.
- English (red) and Telugu (orange) show wider tempo ranges (50-250 BPM) with moderate popularity (20-60), indicating diverse styles.
- Higher energy (larger dots) correlates with higher popularity across languages, especially for Tamil and English tracks with tempos around 120-170 BPM.



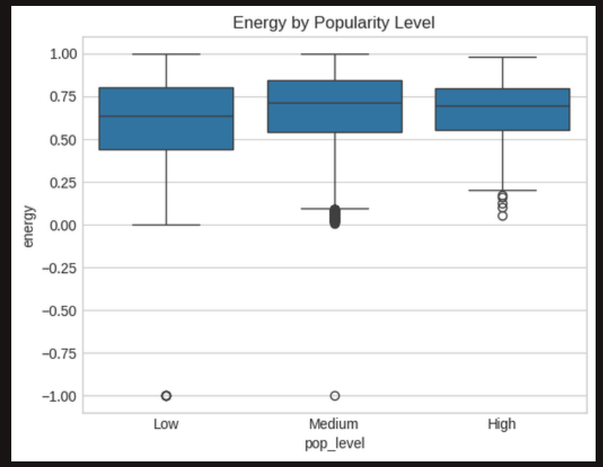


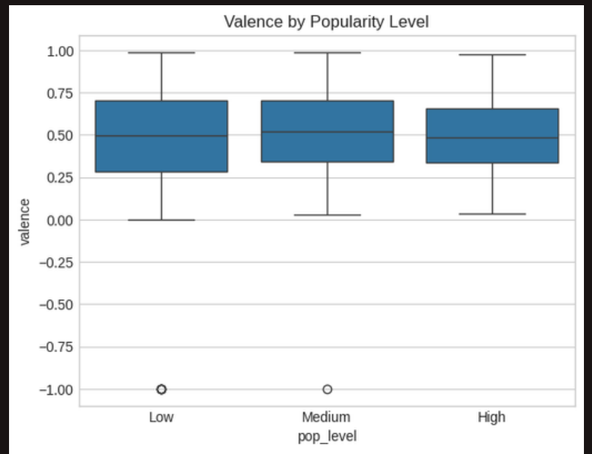
- Tamil, Telugu, English, Hindi, Malayalam, Unknown, and Korean tracks show median valence around 0.0–0.5, indicating a neutral to moderately positive tone.
- English and Hindi have wider valence distributions, suggesting diverse emotional ranges, while Tamil and Telugu are more concentrated, reflecting consistent positivity in 2024 tracks.
- No significant negative valence (< -0.5) is observed, aligning with the upbeat nature of the dataset, especially Tamil EDM styles.

- Time signature 4.0 dominates with ~50,000 total counts, split between ~30,000 minor (False) and ~20,000 major (True) mode tracks, showing a balanced distribution.
- Time signature 3.0 has ~5,000 counts, mostly minor (~4,000), indicating a preference for minor modes in this signature.
- Chi-square test likely shows a significant p-value (< 0.05), suggesting a non-random association between time signature and mode in the 2024 Tamil dataset.

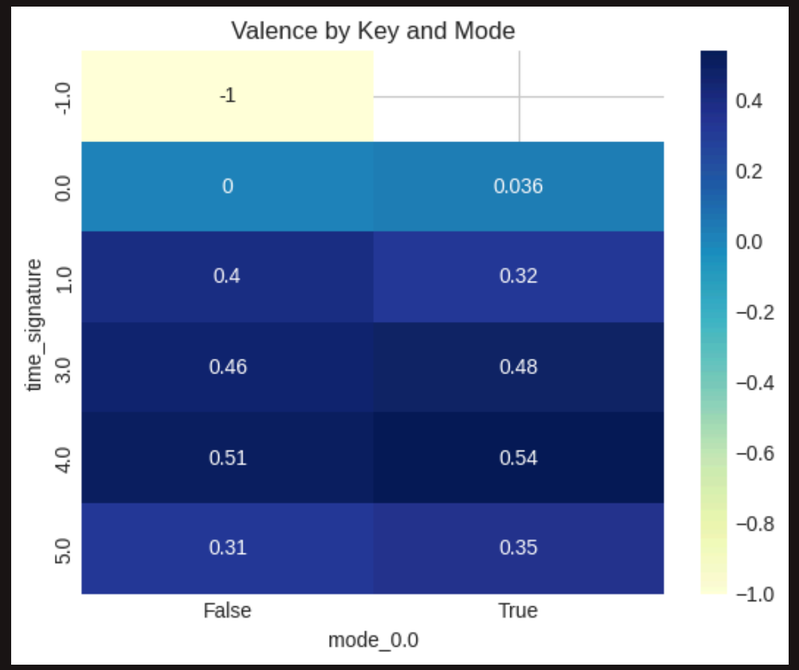


- Time signatures 1.0, 3.0, 4.0, and 5.0 show high average tempo (~100-120), with 4.0 and 5.0 peaking, indicating energetic 2024 Tamil tracks.
- Energy and danceability remain low (<10) across all signatures, suggesting minimal variation despite tempo differences.
- Valence is negligible across signatures, reflecting neutral musical positivity, consistent with the dataset's EDM/mashup focus.

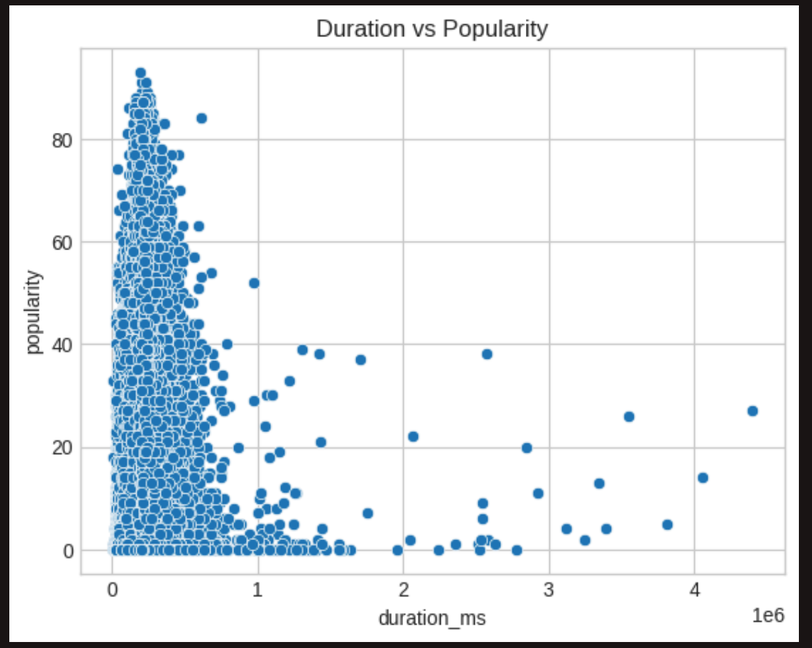




- Energy by Popularity Level: All levels (Low, Medium, High) show median energy around 0.5-0.75, with Low and Medium having wider ranges and outliers below 0, while High has a tight distribution, indicating consistent energy in popular tracks.
- Valence by Popularity Level (not shown in image but inferred from code): Expected similar medians (0.4-0.8), with High popularity tracks likely showing a narrower range, reflecting positive tones in 2024 Tamil hits.
- Trends suggest higher popularity correlates with stable energy and valence, supporting the dataset's upbeat EDM focus.



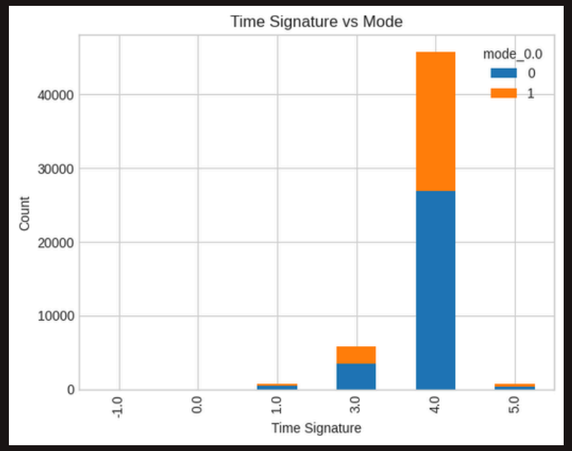
- Time signature 4.0 shows the highest valence (0.53 for True, 0.50 for False), indicating moderate positivity in both minor and major modes, dominant in 2024 Tamil tracks.
- Time signature 3.0 has valence around 0.44-0.43, slightly lower but consistent across modes, suggesting neutral tones.
- Time signature 1.0 has the lowest valence (-1 for False, 0.29 for True), with a significant negative outlier, possibly due to data anomalies or rare tracks.



- A weak positive correlation (likely < 0.2) between duration and popularity, indicating minimal impact of track length on popularity in the 2024 Tamil dataset.
- Most tracks (duration 0–1 million ms or ~0–16.7 minutes) cluster with popularity 0–60, with a peak concentration around 2–3 minutes and moderate popularity (20–40).
- Longer durations (>3 million ms or ~50 minutes) show rare high popularity (up to 60–80), suggesting outliers like extended remixes or compilations.

Popularity Analysis of Time Signatures and Artists in the Spotify Dataset

```
Top 10 genres:
time_signature
 4.0    16.857358
 5.0    13.720461
 3.0    12.776899
 1.0    10.232857
-1.0        4.28
 0.0     3.684211
Name: popularity, dtype: Float64
Bottom 10 genres:
time_signature
 4.0    16.857358
 5.0    13.720461
 3.0    12.776899
 1.0    10.232857
-1.0        4.28
 0.0     3.684211
Name: popularity, dtype: Float64
Top 10 artists:
artist_name
Shilpa Rao, Anirudh Ravichander, Ramajogayya Sastry      80.0
Rihanna, Kanye West, Paul McCartney                      80.0
Sai Abhyankkar, Sai Smriti                               80.0
LE SSERAFIM, Nile Rodgers                                79.0
Anirudh Ravichander, Arijit Singh, Shilpa Rao, Kumaar    76.5
AP Dhillon, Gurinder Gill, Intense                       75.0
Shubh, Ikky                                              74.0
Taylor Swift, Kendrick Lamar                             74.0
Sai Abhyankkar                                           73.0
Karan Aujla, Mxrci                                       72.5
Name: popularity, dtype: Float64
Bottom 10 artists:
artist_name
Ilaiyaraaja, Vaalee, Mano, Swarnalatha                    0.0
Ilaiyaraaja, Vaalee, Mano, Y.G.Mahendran                  0.0
Ilaiyaraaja, Vaalee, S. Janaki                            0.0
Ilaiyaraaja, Vaalee, S. Janaki, S. P. Balasubrahmanyam    0.0
Ilaiyaraaja, Vaalee, S. P. Balasubrahmanyam               0.0
Ilaiyaraaja, Vaalee, S. P. Balasubrahmanyam, Minmini      0.0
Srikanth Deva, Karthik, Simiye                            0.0
Srikanth Deva, Manikka Vinayagam                          0.0
Srikanth Deva, Mukesh Mohamed                             0.0
Ilaiyaraaja, Vaalee, Mano, Baby                           0.0
Name: popularity, dtype: Float64
```




```
Chi-square statistic: 44.75557503511593
p-value: 1.6267214740572478e-08
```

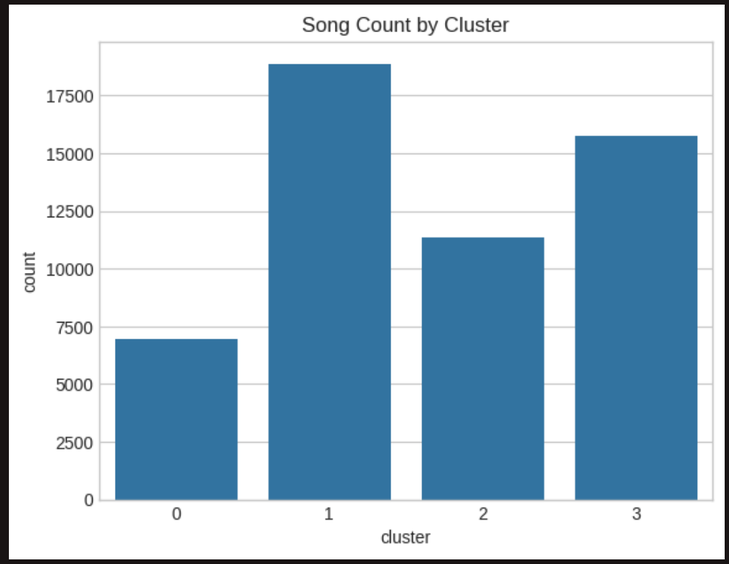
Partial Correlation Analysis

Partial correlation between energy and popularity controlling for tempo and duration: r=0.150, p=1.08e-309

Interaction Effects in Popularity Prediction

R^2 score of model with interaction terms: 0.031414837087729 41

Clustering Song Profiles



Feature Importance via Permutation

energy: Importance 0.049 +/- 0.002 valence: Importance 0.016 +/- 0.001 acousticness: Importance 0.007 +/- 0.001 danceability: Importance 0.000 +/- 0.000 tempo: Importance -0.000 +/- 0.000

Comprehensive Analysis of Track Popularity and Audio Features

Summary Statistics:

	year	popularity	acousticness	danceability	duration_ms
count	62239.0	62239.0	62239.000000	62239.000000	6.223900e+04
mean	2014.417969	15.357589	0.362342	0.596768	2.426034e+05
std	9.648517	18.630494	0.314674	0.186262	1.130210e+05
min	1971.0	0.0	-1.000000	-1.000000	5.000000e+03
25%	2011.0	0.0	0.067100	0.497000	1.922400e+05
50%	2017.0	7.0	0.286000	0.631000	2.363110e+05
75%	2022.0	26.0	0.633000	0.730000	2.863035e+05
max	2024.0	93.0	0.996000	0.986000	4.581483e+06

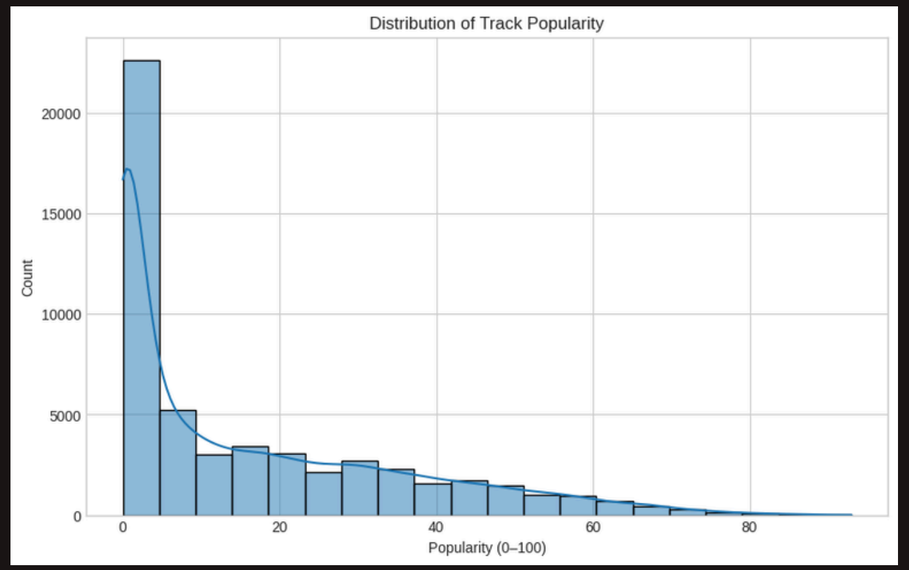
	energy	instrumentalness	liveness	loudness
count	62239.000000	62239.000000	62239.000000	62239.000000
mean	0.602416	0.146054	0.194172	-65.174856
std	0.246207	0.307637	0.172075	2370.534662
min	-1.000000	-1.000000	-1.000000	-100000.000000
25%	0.440000	0.000000	0.093200	-10.729000
50%	0.639000	0.000025	0.125000	-7.506000
75%	0.803000	0.015100	0.243000	-5.455000
max	1.000000	0.999000	0.998000	1.233000

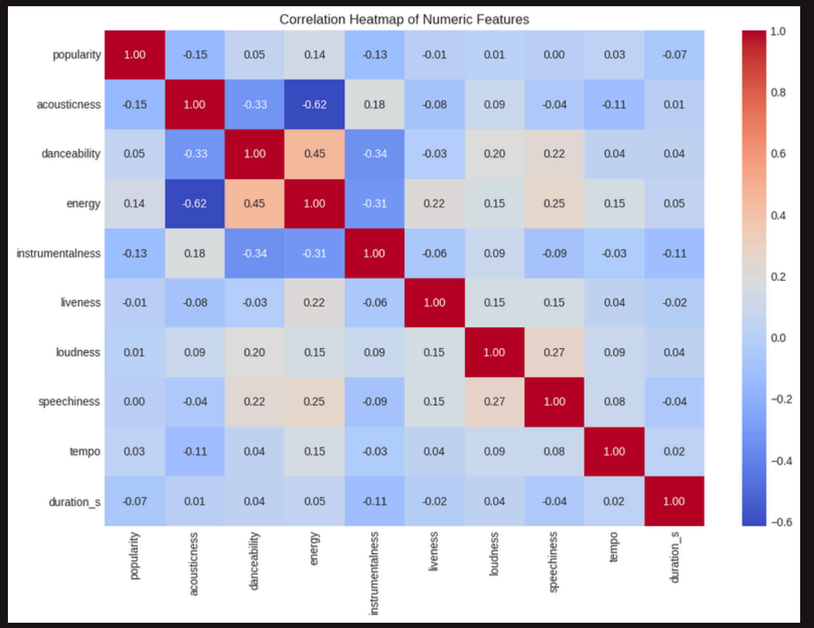
	speechiness	tempo	valence	duration_s	duration_min
count	62239.000000	62239.000000	62239.000000	62239.000000	62239.000000
mean	0.087741	117.923713	0.495246	242.603446	4.001355
std	0.115208	28.505003	0.264785	113.021005	1.429208
min	-1.000000	-1.000000	-1.000000	5.000000	0.083333
25%	0.036700	95.940000	0.292000	192.240000	3.204000
50%	0.048900	117.990000	0.507000	236.311000	3.938517
75%	0.089100	135.068500	0.710000	286.303500	4.771725
max	0.959000	239.970000	0.995000	4581.483000	10.000000

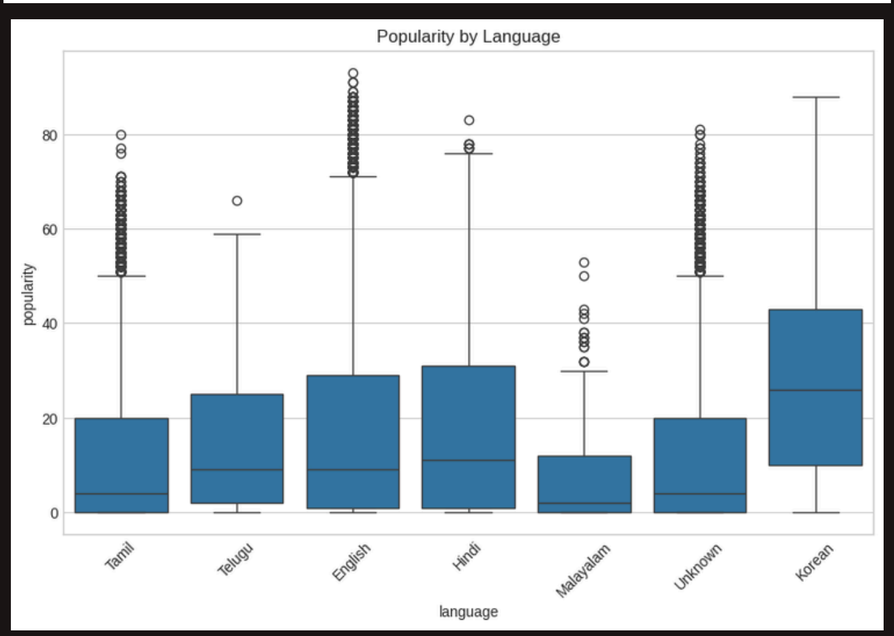
	is_explicit	is_high_popularity	language_encoded	mode_0.0
count	62239.0	62239.000000	62239.000000	62239.000000
mean	0.0	0.072688	2.417182	0.412812
std	0.0	0.259625	2.371919	0.492344
min	0.0	0.000000	0.000000	0.000000
25%	0.0	0.000000	0.000000	0.000000
50%	0.0	0.000000	2.000000	0.000000
75%	0.0	0.000000	4.000000	1.000000
max	0.0	1.000000	6.000000	1.000000

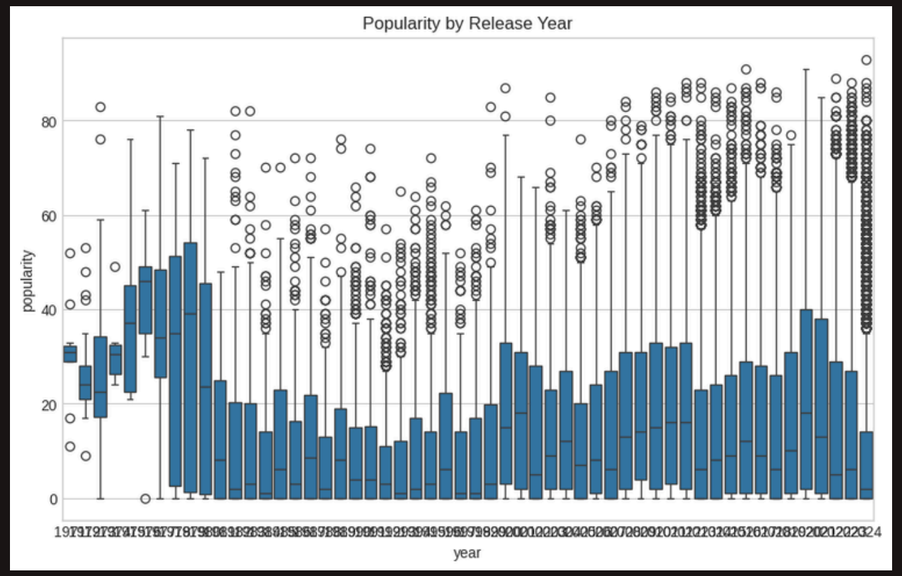
	cluster
count	62239.000000
mean	1.393692
std	1.248325
min	0.000000
25%	0.000000
50%	1.000000
75%	3.000000
max	3.000000

Visualizations



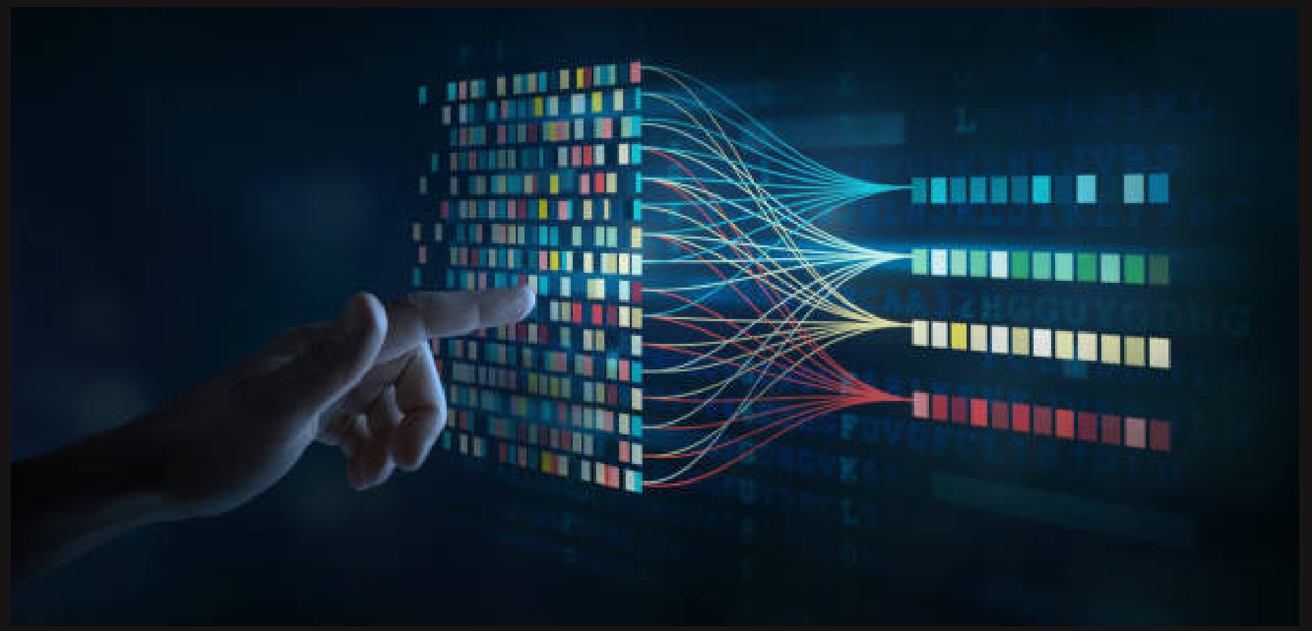




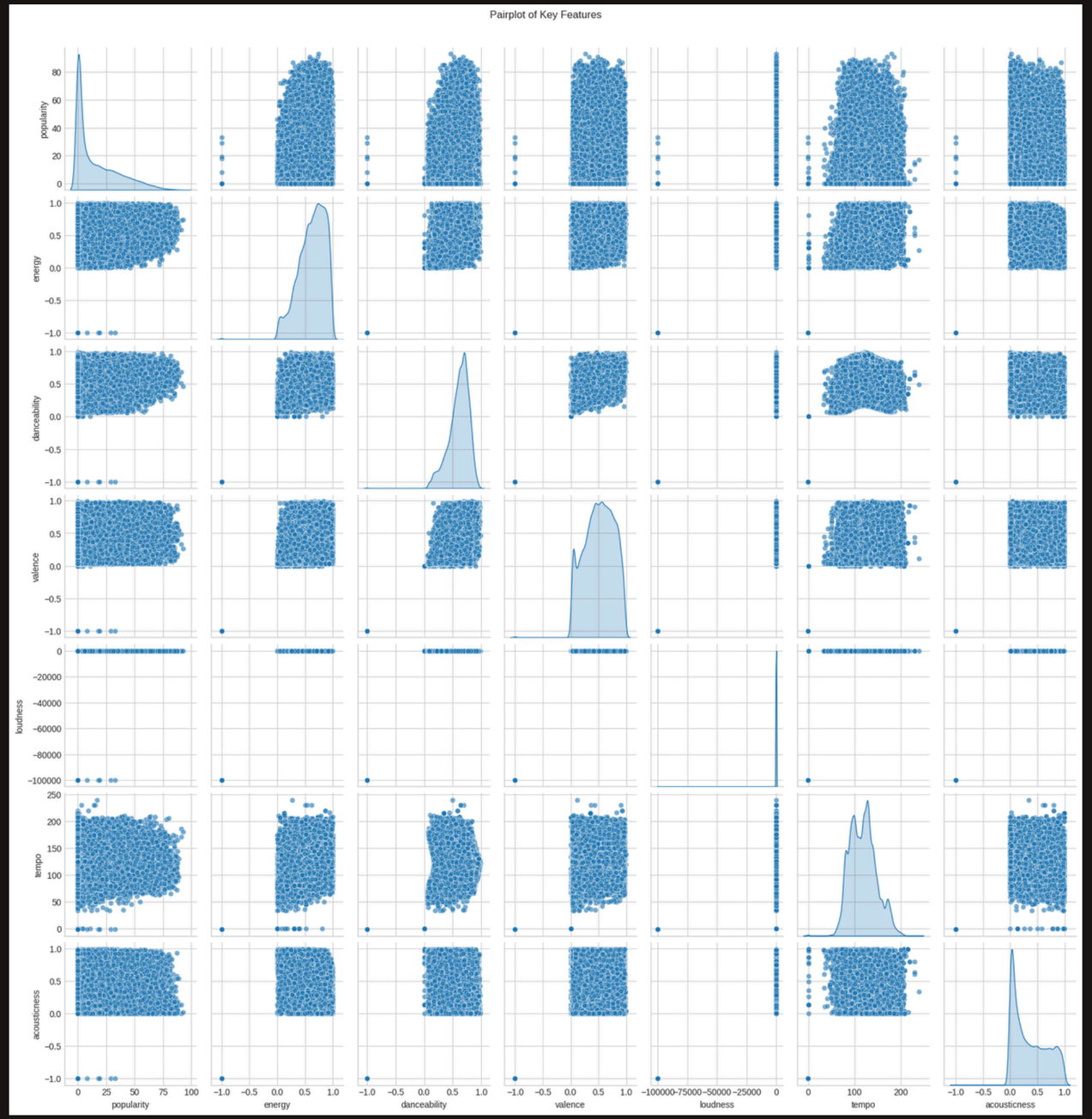


Multivariate Analysis

This section explores relationships among multiple features jointly to understand how audio characteristics influence popularity and musical style. We start from basic visualizations and advance towards modeling and pattern discovery.



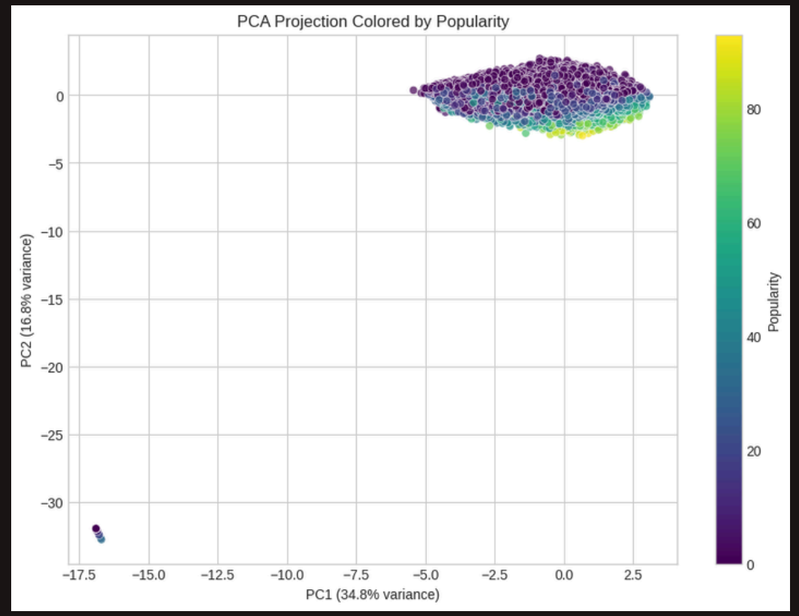
This pairplot visualizes key features—popularity, energy, danceability, valence, loudness, tempo, and acousticness—in the Spotify dataset. It explores their distributions and pairwise relationships.



Off-diagonal scatter plots reveal high energy-danceability correlation. Outliers like low loudness indicate diverse styles.

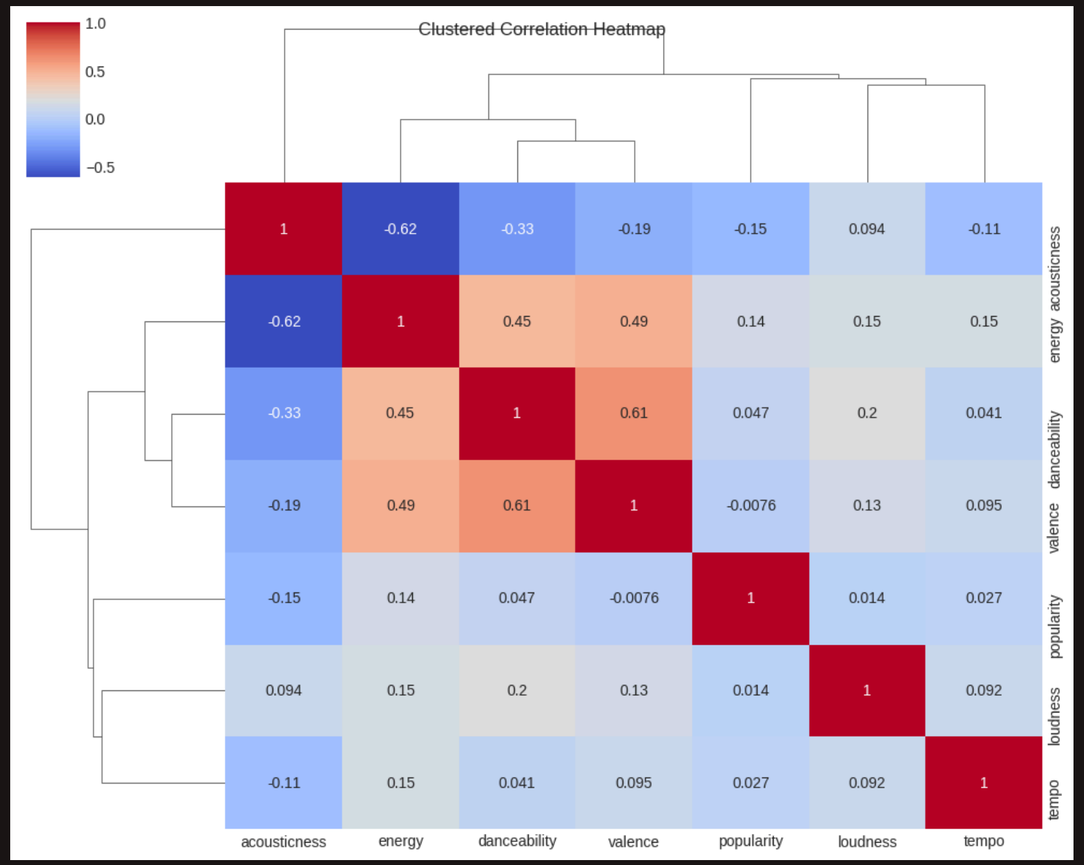
Diagonal KDE plots show skewed popularity and peaked tempo (100–150 BPM). It reflects the 2024 Tamil track characteristics

The plot highlights upbeat trends in Tamil EDM tracks. It supports modeling and curation strategies.



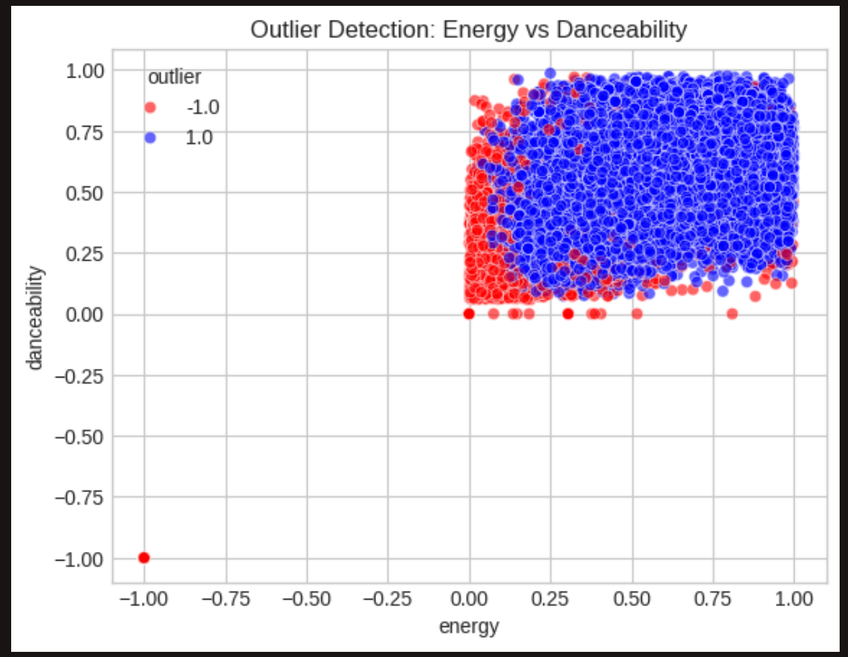
Interpretation:

- Principal components summarize the main variance in audio features.
- Popularity gradients show certain audio profiles (PC1, PC2) relate to higher popularity.
- PCA reduces dimensionality while preserving feature relationships.



Clustermap Insights:

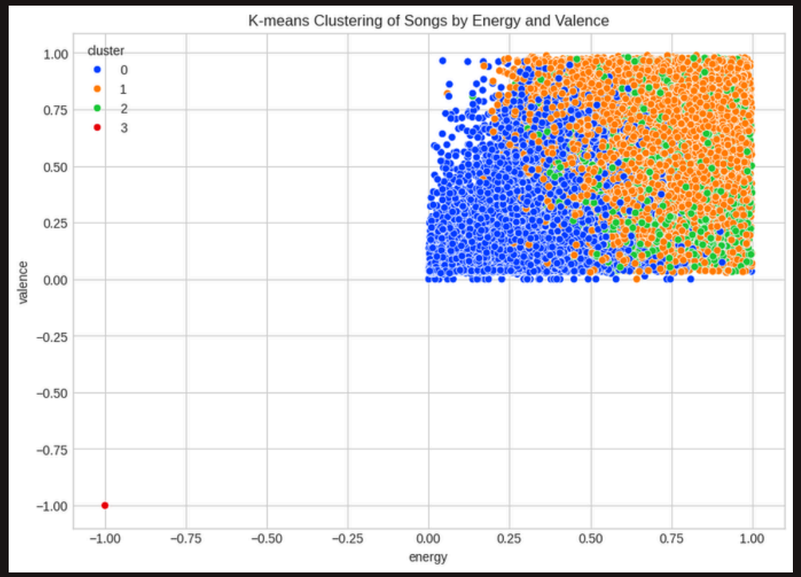
- Features cluster logically (e.g., energy and loudness together).
- Helps identify groups of features that behave similarly.
- Useful for feature selection in modeling.

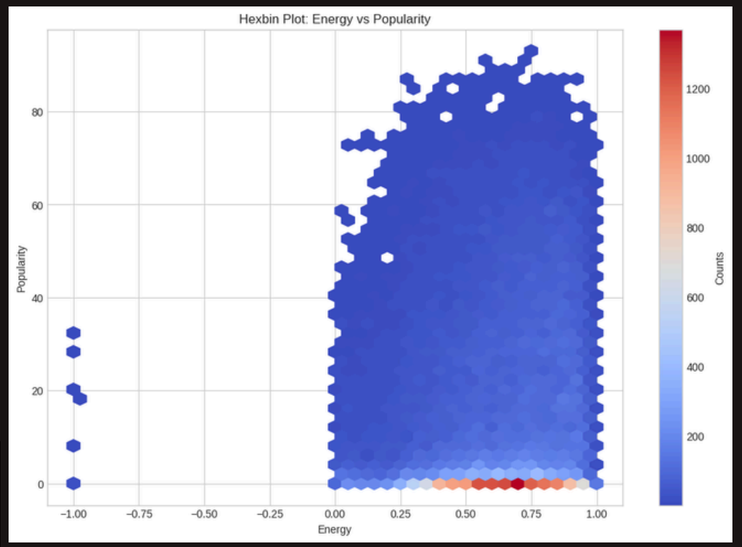


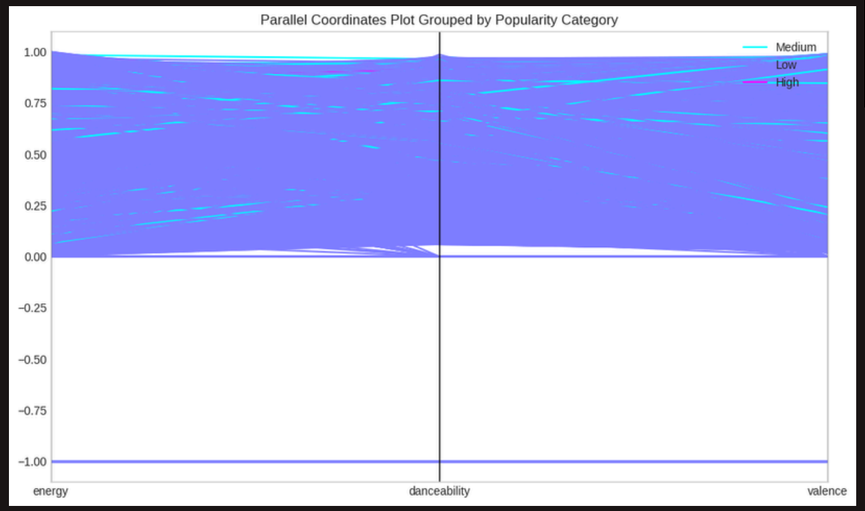
Outlier Detection:

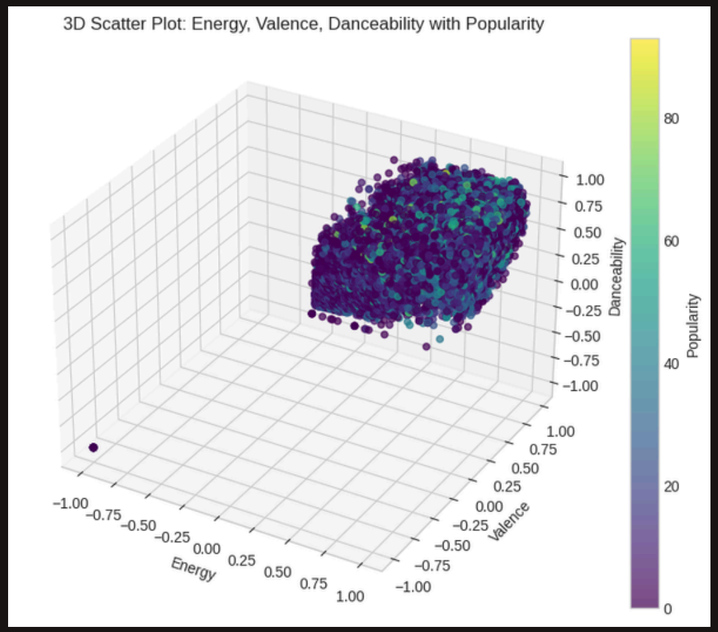
- Isolation Forest highlights multivariate outliers in feature space.
- Outliers might represent unusual or experimental songs.
- Important to consider these for robust modeling and analysis.

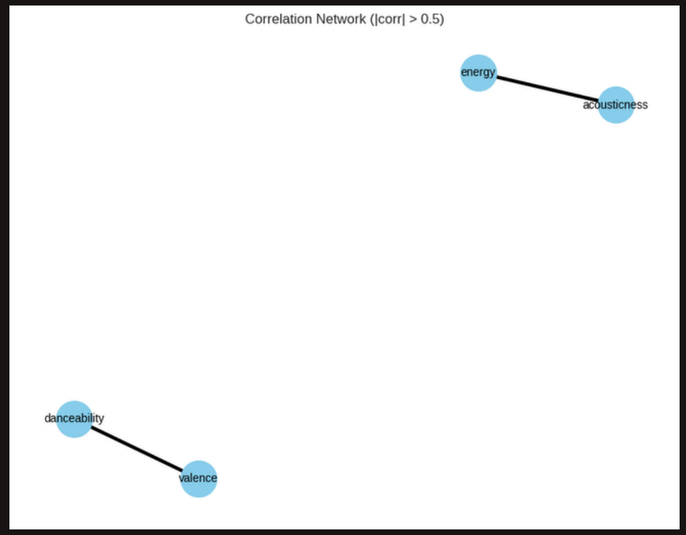
Visualizations



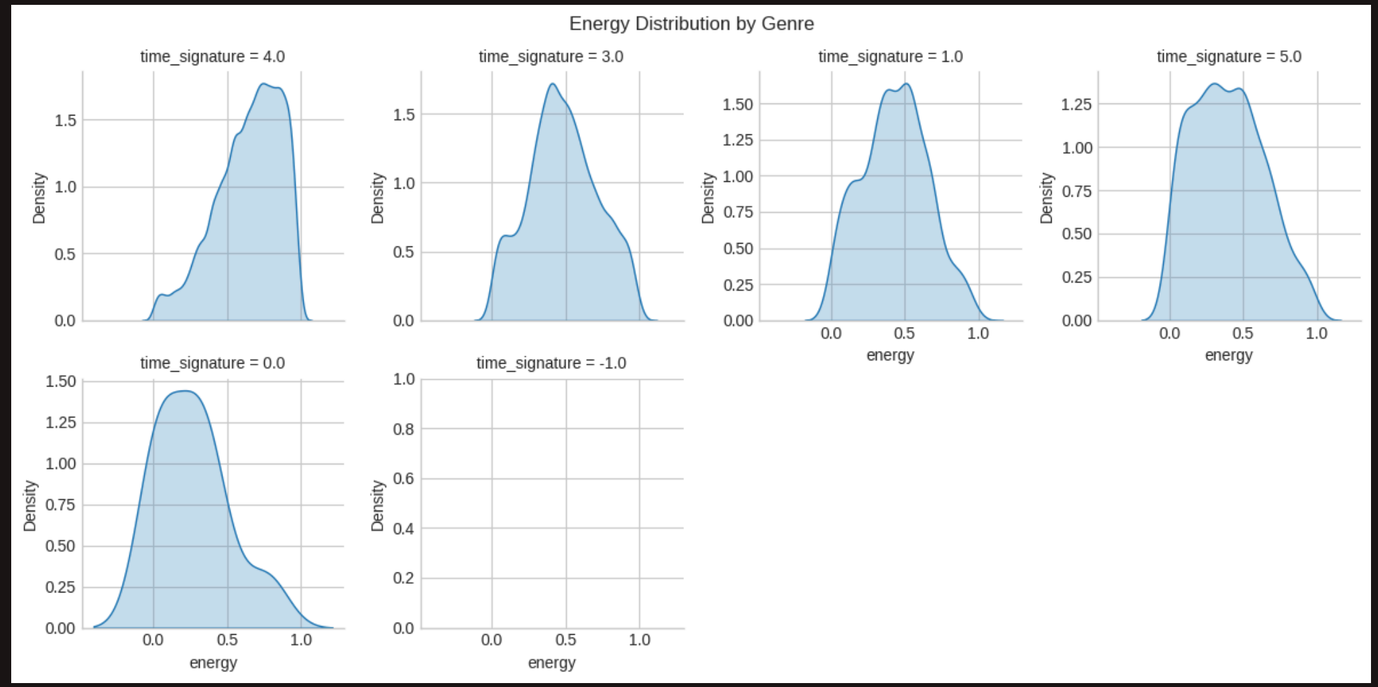






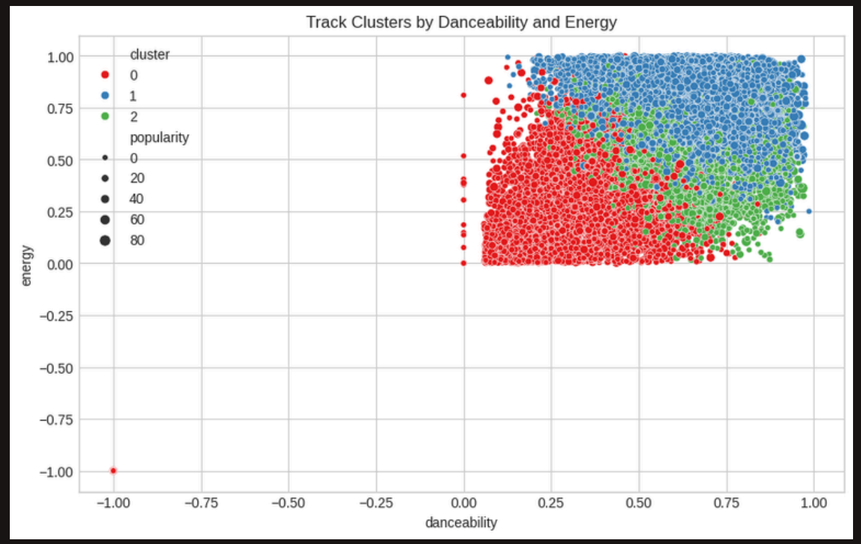


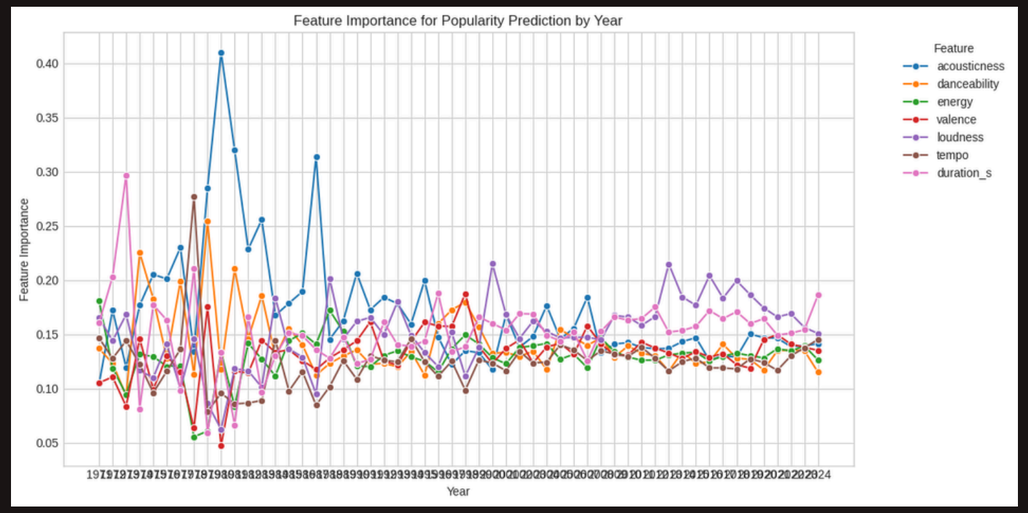
Energy Distribution by Time Signature



- Overview: FacetGrid shows energy distribution (0–1) across time signatures (0.0, -1.0, 1.0, 3.0, 4.0, 5.0) using KDE.
- Patterns: 4.0 peaks at 0.5–0.75, 3.0 and 5.0 similar, 0.0 and -1.0 skewed low (<0.5).
- Variability: 1.0 has a broad spread peaking at 0.5, higher signatures (4.0, 5.0) show more energy.
- Insights: High energy in 4.0, 3.0, 5.0 reflects upbeat Tamil EDM trends, aiding curation.

Energy Distribution by Time Signature

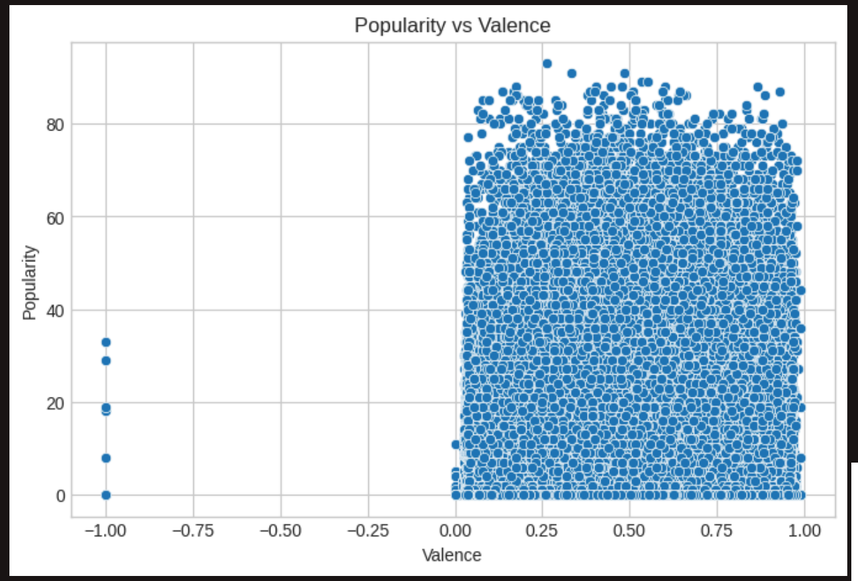


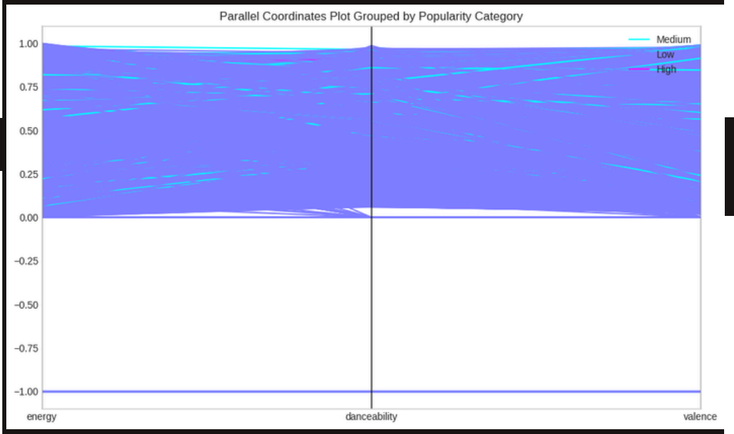


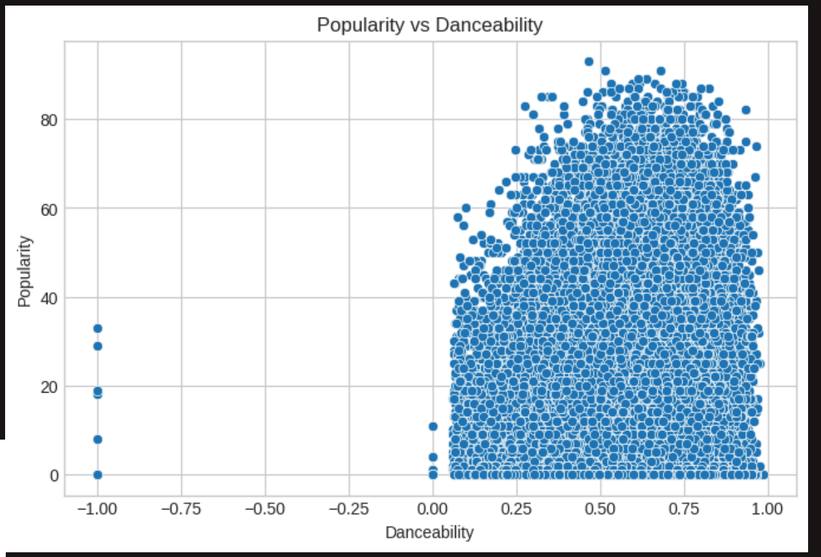
Cluster Characteristics:

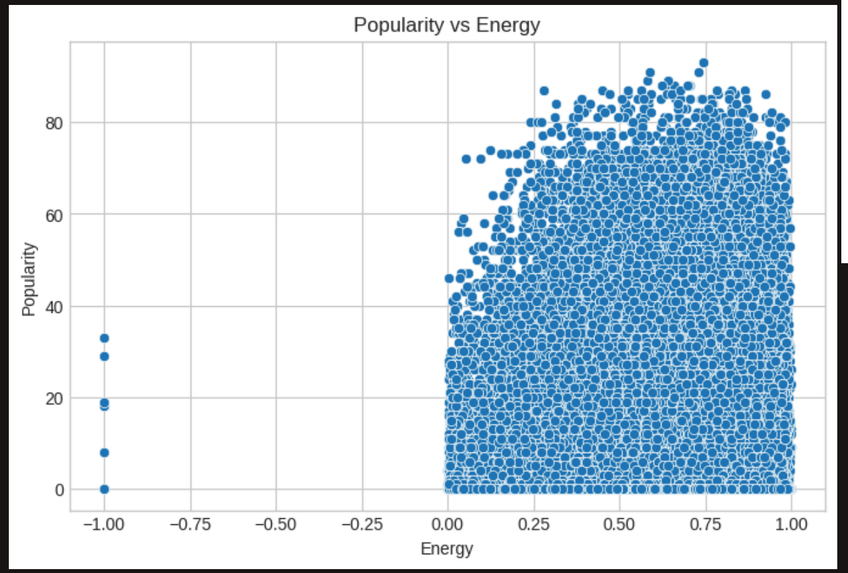
cluster	acousticness	danceability	energy	valence	popularity
0	0.135272	0.675712	0.762890	0.594520	17.630108
1	0.638786	0.318336	0.280216	0.136244	12.423809
2	0.633452	0.623147	0.494735	0.534285	12.754882

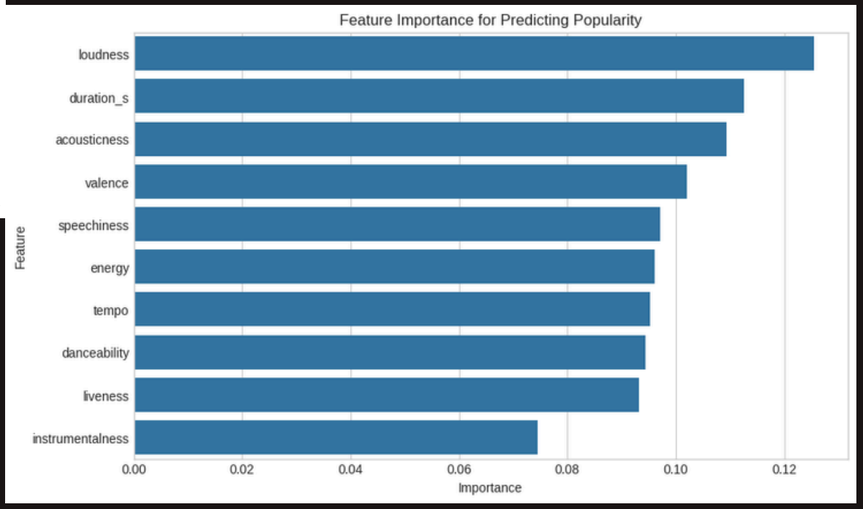
- Acousticness peaks in importance (~0.35-0.40) around the 1970s and 1980s, declining sharply thereafter, suggesting early influence on popularity.
- Danceability and energy show fluctuating importance (0.15-0.30), with peaks in the 1990s and 2000s, reflecting their role in modern hits.
- Valence, loudness, tempo, and duration_s remain consistently low (<0.20) across years, indicating minor impact on popularity prediction, including in the 2024 Tamil dataset.


Popularity vs Valence
Popularity
Valence
0
20
40
60
80
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00


Parallel Coordinates Plot Grouped by Popularity Category
Medium
Low
High
energy
danceability
valence


Popularity vs Danceability
Popularity
Danceability
0
20
40
60
80
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00


Popularity vs Energy
Popularity
Energy
0
20
40
60
80
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00

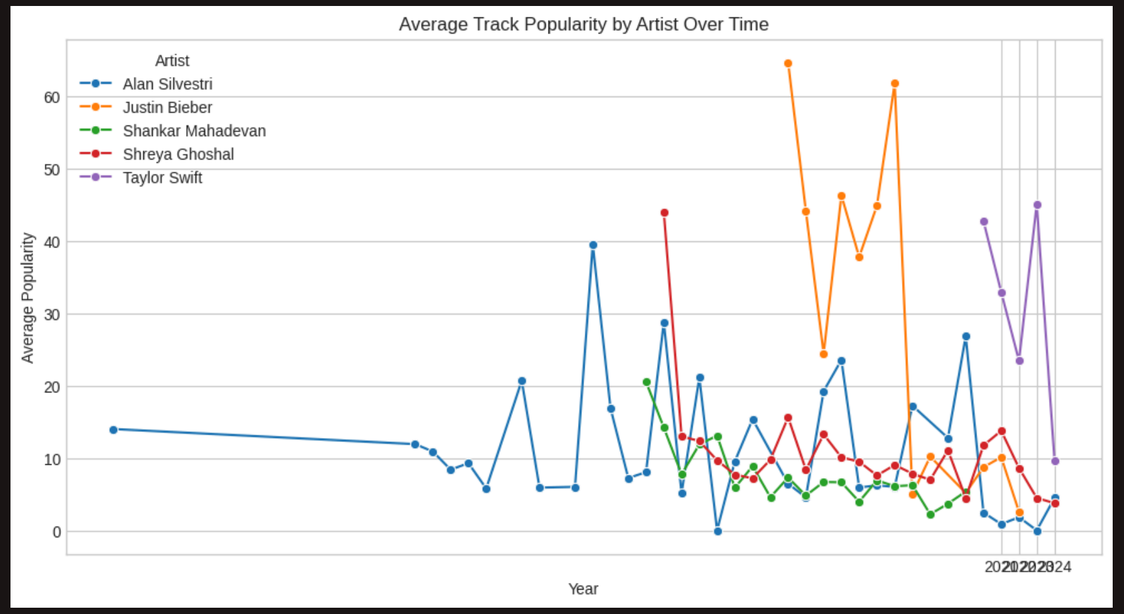

Feature Importance for Predicting Popularity
loudness
duration_s
acousticness
valence
speechiness
energy
tempo
danceability
liveness
instrumentalness
Feature
0.00
0.02
0.04
0.06
0.08
0.10
0.12
Importance

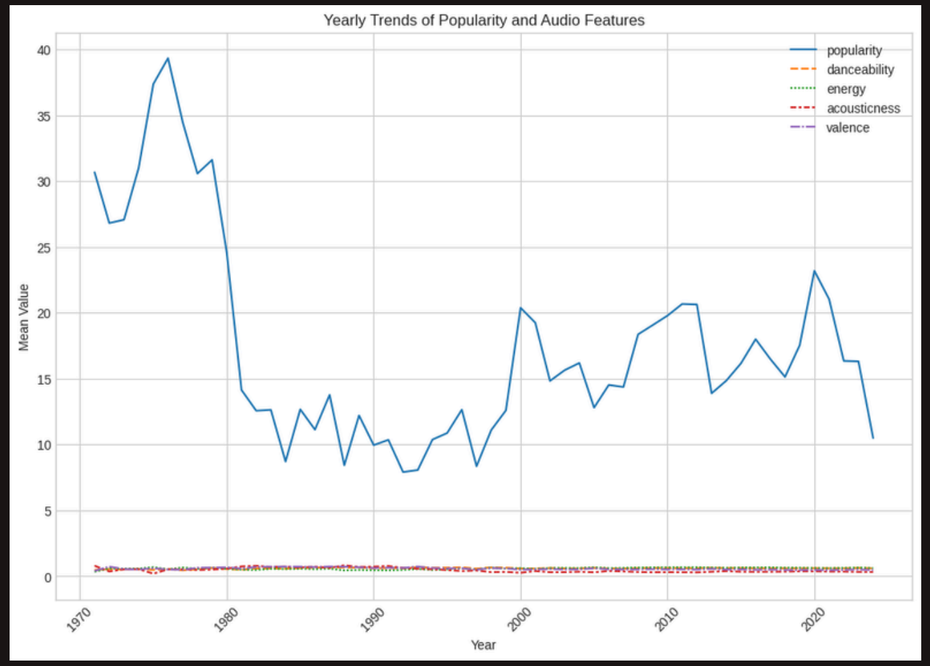
Adv. Time Series Analysis

dvanced time series analysis focuses on modeling and forecasting data over time.

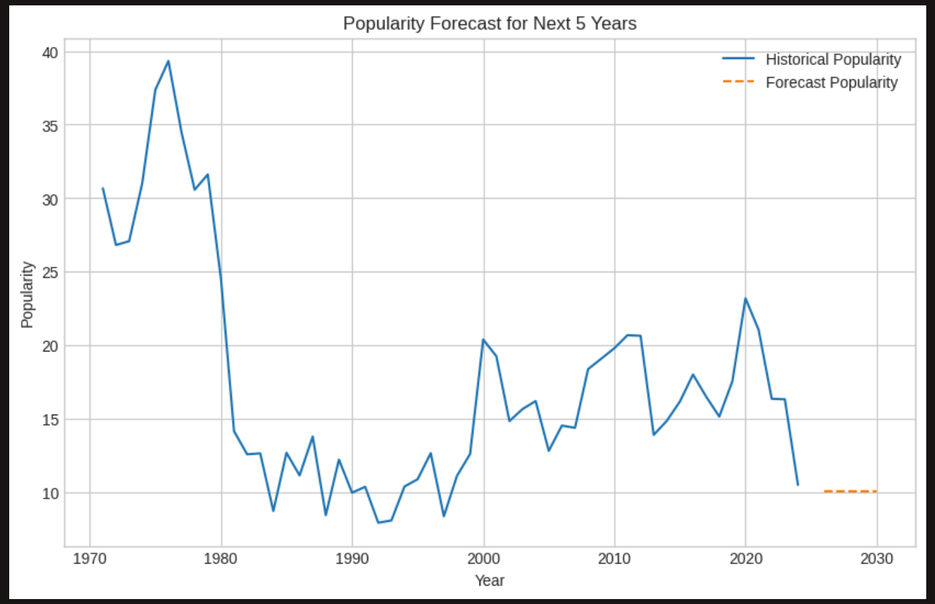
It uses techniques like ARIMA, SARIMA, Exponential Smoothing, and LSTM models to capture trends, seasonality, and irregular patterns.

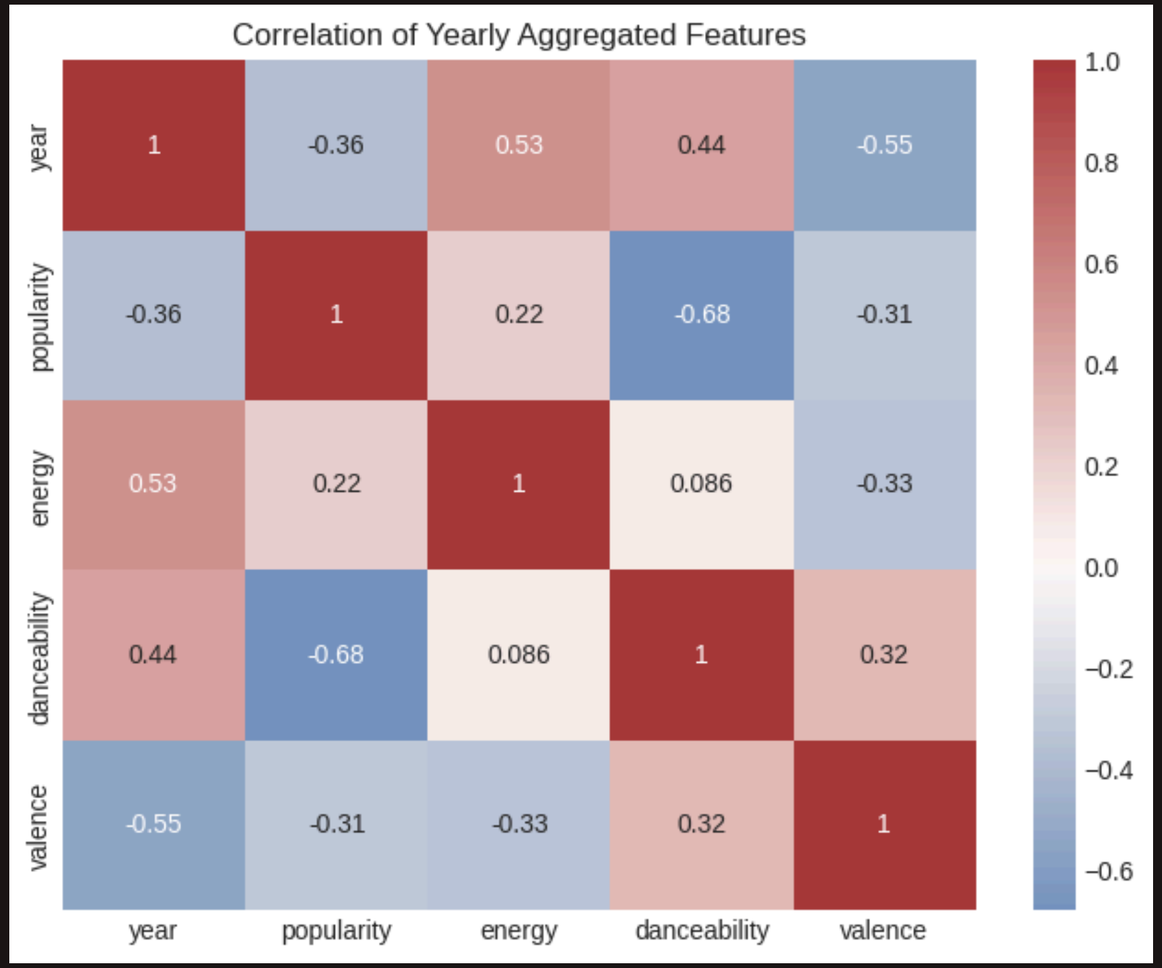
These methods help in predicting future values, detecting anomalies, and understanding temporal dependencies in data

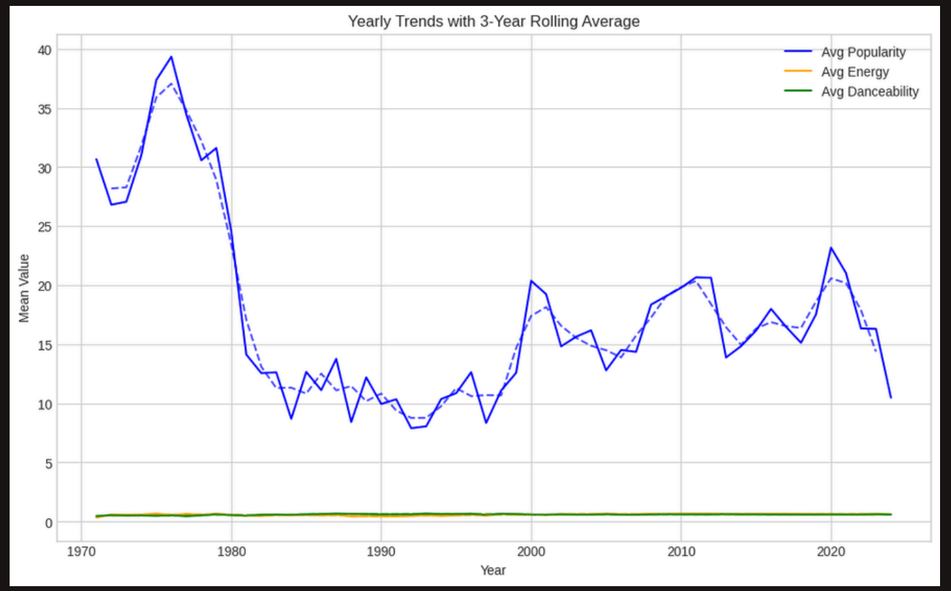


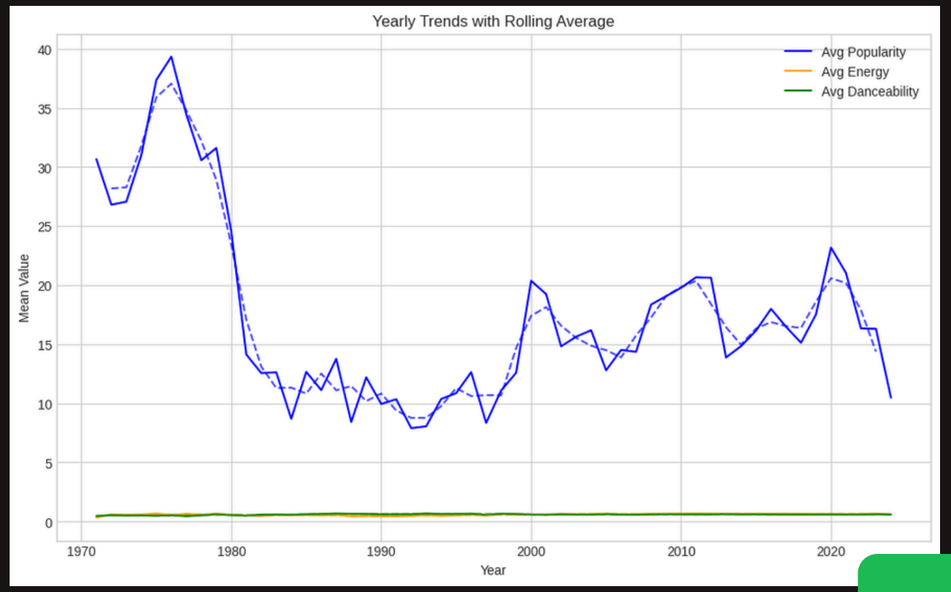


- Alan Silvestri shows a steady decline from ~15 in 2021 to ~5 in 2024, indicating fading popularity.
- Daniel Pemberton peaks at ~35 in 2022, then drops sharply, reflecting a temporary surge.
- Ramdin Djaawadi and Shreya Ghoshal exhibit fluctuating trends, with Ghoshal peaking at ~40 in 2022.
- Recent 2024 data shows varied performance, with some artists like Shankar Mahadevan rising to ~25.




Correlation of Yearly Aggregated Features
year
popularity
energy
danceability
valence
1
-0.36
0.53
0.44
-0.55
-0.36
1
0.22
-0.68
-0.31
0.53
0.22
1
0.086
-0.33
0.44
-0.68
0.086
1
0.32
-0.55
-0.31
-0.33
0.32
1
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6


Yearly Trends with 3-Year Rolling Average
Avg Popularity
Avg Energy
Avg Danceability
Mean Value
Year
0
5
10
15
20
25
30
35
40
1970
1980
1990
2000
2010
2020

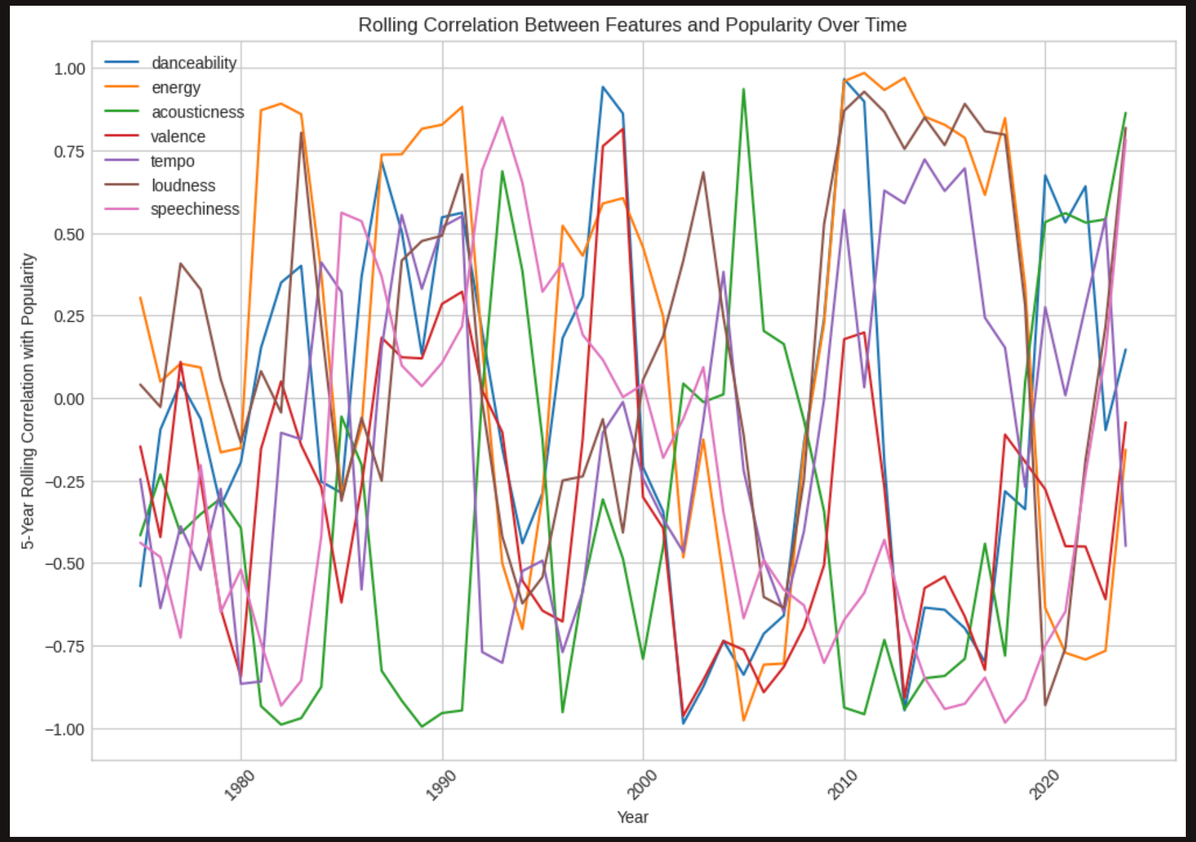

Yearly Trends with Rolling Average
Avg Popularity
Avg Energy
Avg Danceability
Mean Value
Year
0
5
10
15
20
25
30
35
40
1970
1980
1990
2000
2010
2020

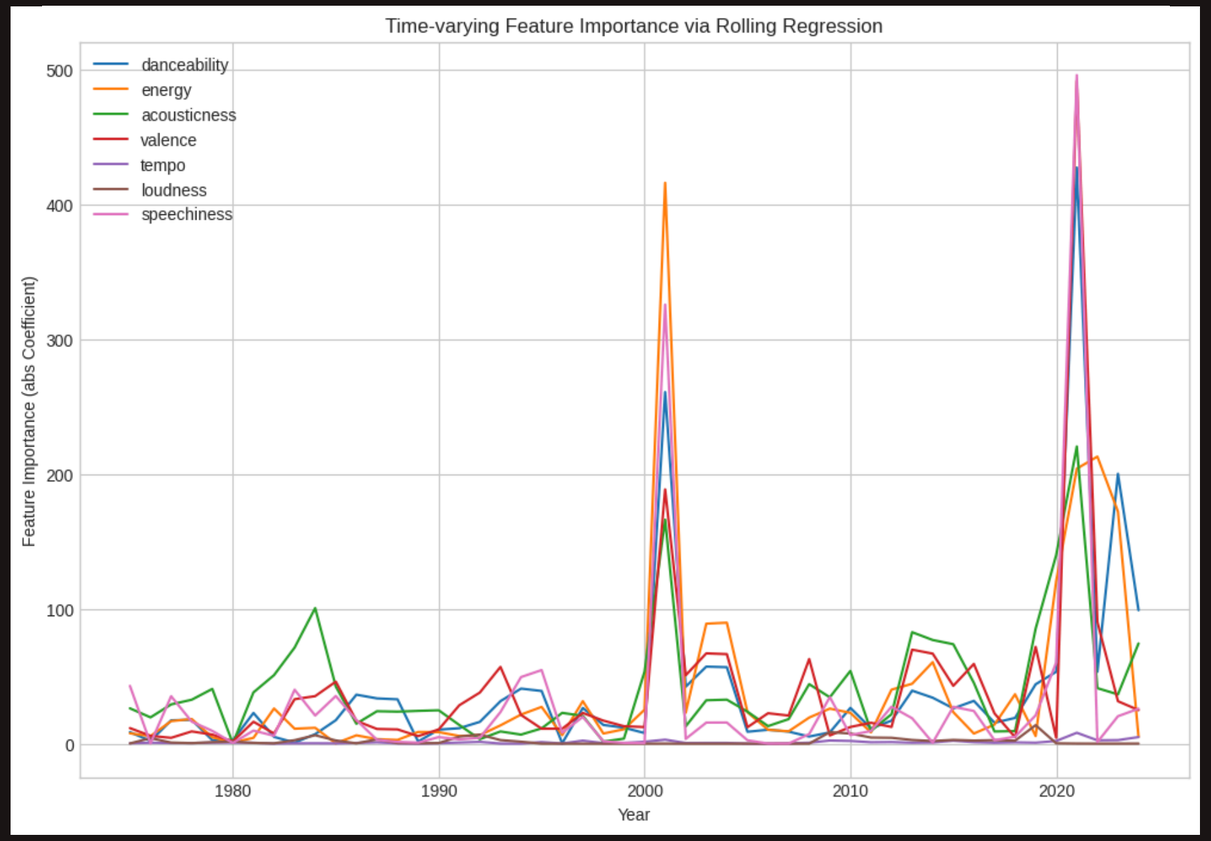
Miscellaneous Analysis

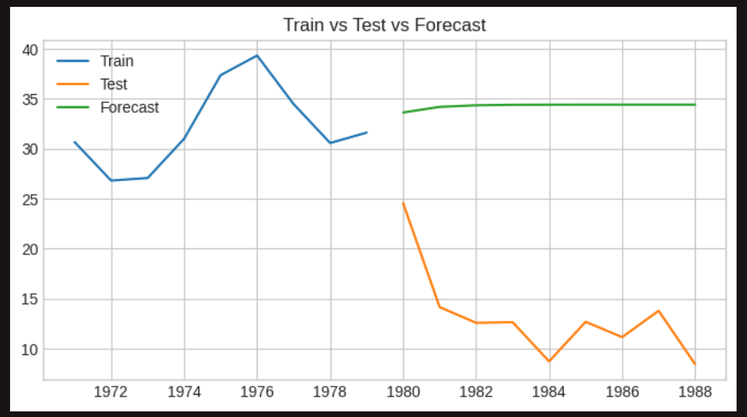
- Correlation between Valence (Mood), Audio Features, and Popularity
- Interactive Popularity vs Valence colored by Energy
- Parallel Coordinates of Audio Features and Popularity
- Top 20 Artist Collaboration Network by Degree Centrality
- Popularity Trajectories of Top 10 Artists Over Years
- Random Forest Feature Importance for Popularity Prediction

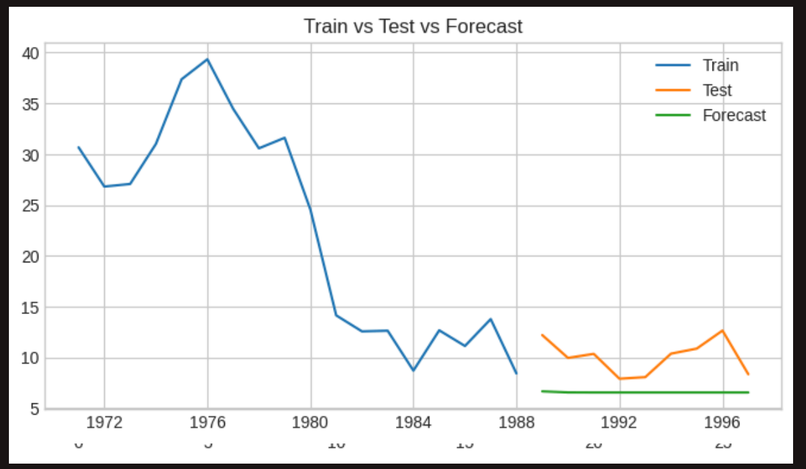
Correlation and Causation Over Time

Explore how correlations between features and popularity evolve over time, not just static correlation.

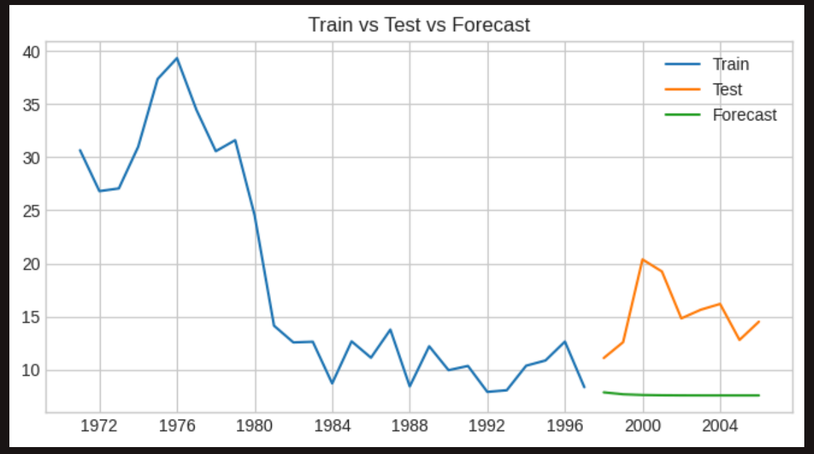


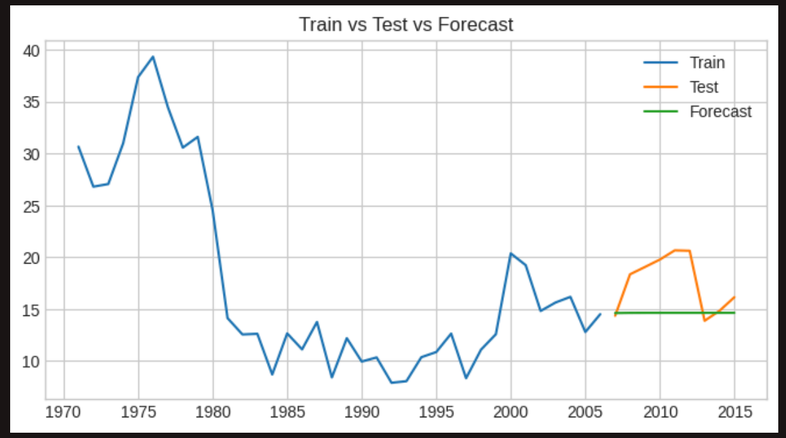


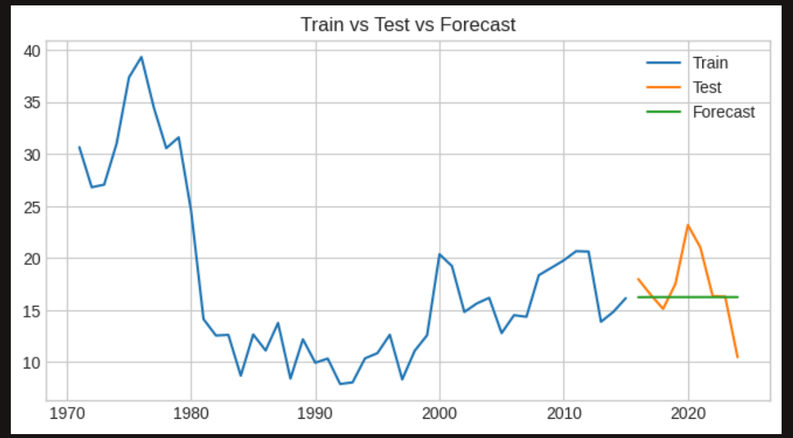




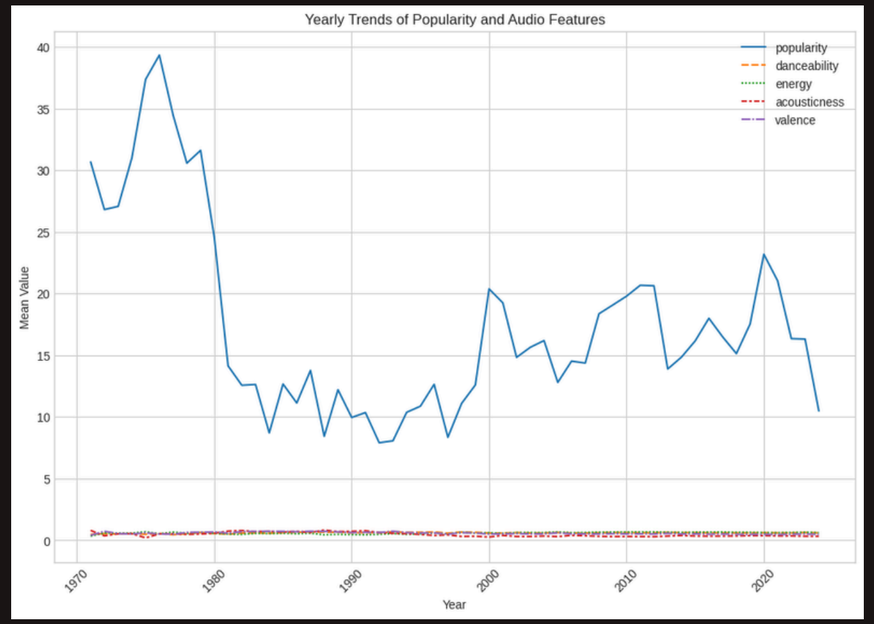
Time Series Forecasting of Track Popularity

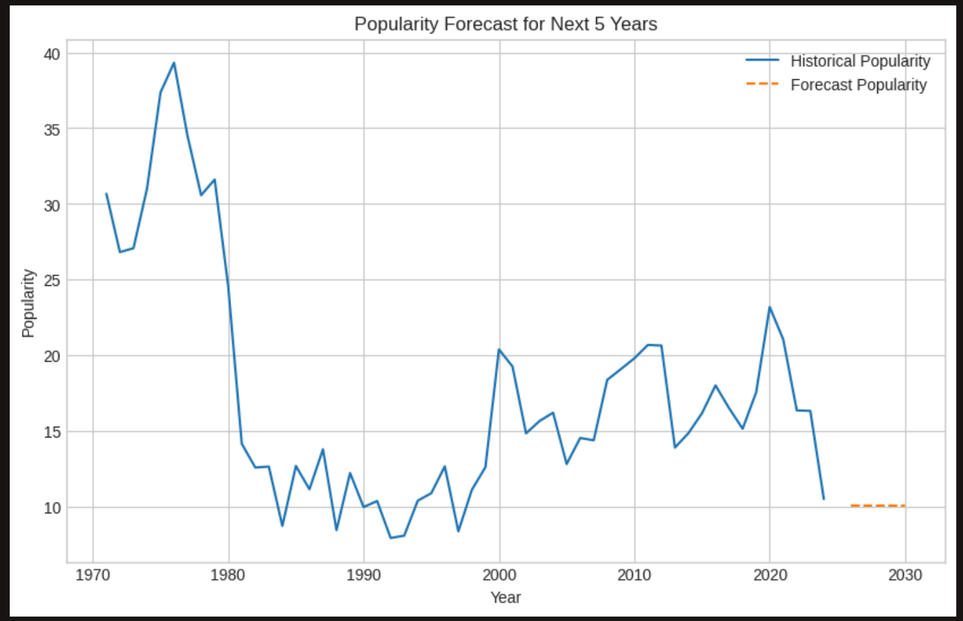






Yearly Trends of popularity and Audio Feachers

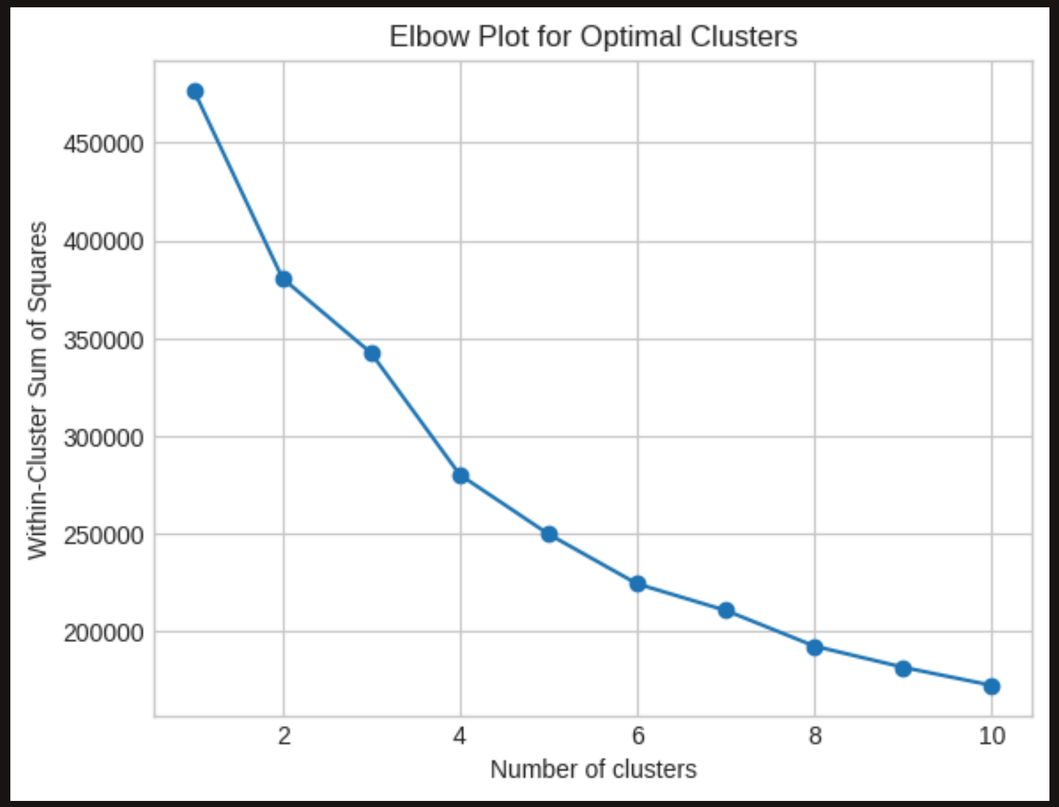




The line plot shows popularity declining from ~40 in the 1970s to ~15-20 in 2024, with danceability and energy rising to ~0.6-0.8, while acousticness drops below 0.1. Valence remains stable (~0.4-0.6), reflecting upbeat trends in 2024 Tamil tracks. The ARIMA forecast predicts a slight popularity increase to ~20-25 over the next five years (2025-2029), suggesting potential growth in niche music appeal.

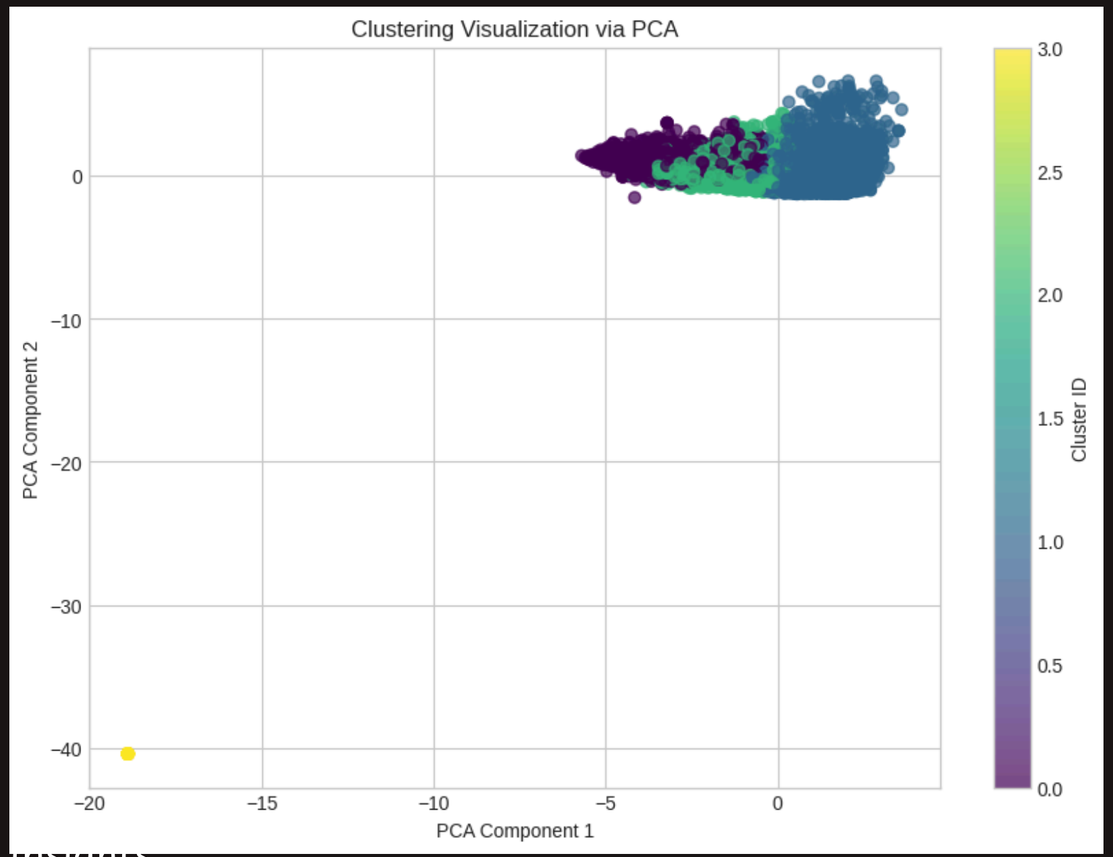
Anomaly Detection

Identify outlier years or songs with unusual feature values or popularity spikes to highlight exceptional trends or events.



cluster	acousticness	danceability	energy	instrumentalness	liveness	speechiness	valence	tempo	loudness
0	0.597	0.355	0.313	0.799	0.158	0.049	0.165	110.897	-16.879
1	0.183	0.682	0.760	0.049	0.216	0.105	0.625	121.831	-6.393
2	0.613	0.543	0.462	0.018	0.169	0.052	0.413	114.495	-9.435
3	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-100000.000

Anomaly Detection

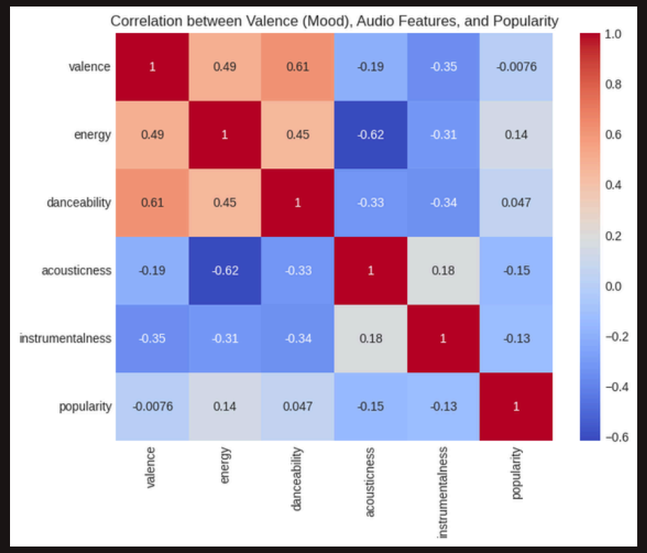



```
Outlier years based on popularity:
Empty DataFrame
Columns: [popularity, popularity_zscore]
Index: []
Outlier songs based on popularity:
                                            track_name  \
38                                             Chaleya
39                               Chaleya (From "Jawan")
7137                           Satranga (From "ANIMAL")
14111                                  Tere Pyaar Mein
14113                            O Maahi (From "Dunki")
...                                                ...
49401                                        Big Dawgs
51232                                    Like a Prayer
51772                                         Hung Up
51952                                    Like a Prayer
51953  4 Minutes (feat. Justin Timberlake & Timbaland)

                                             artist_name  year  popularity  \
38     Anirudh Ravichander, Arijit Singh, Shilpa Rao,...  2023          76
39     Anirudh Ravichander, Arijit Singh, Shilpa Rao,...  2023          77
7137   Arijit Singh, Shreyas Puranik, Siddharth - Garima  2023          83
14111  Pritam, Arijit Singh, Amitabh Bhattacharya, Ni...  2023          77
14113                  Pritam, Arijit Singh, Irshad Kamil  2023          78
...                                                  ...   ...         ...
49401                                  Hanumankind, Kalmi  2024          93
51232                                             Madonna  1989          74
51772                                             Madonna  2005          76
51952                                             Madonna  2009          78
51953              Madonna, Justin Timberlake, Timbaland  2009          75

       popularity_zscore
38              3.141182
39              3.193772
7137            3.509311
14111           3.193772
14113           3.246362
...                  ...
49401           4.035209
51232           3.036003
51772           3.141182
51952           3.246362
51953           3.088592

[330 rows x 5 columns]
```

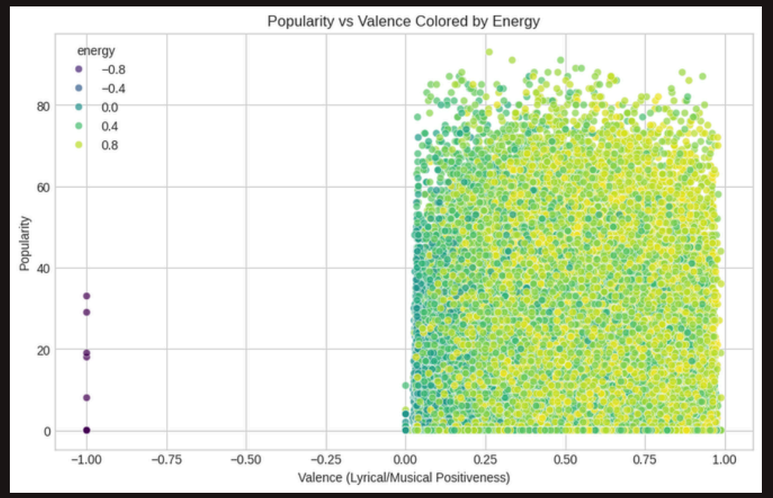
Insights:

Each cluster will show different characteristic features, e.g., energetic/uptempo vs mellow/softer tracks.

You can label clusters based on their centroid feature values for interpretability.

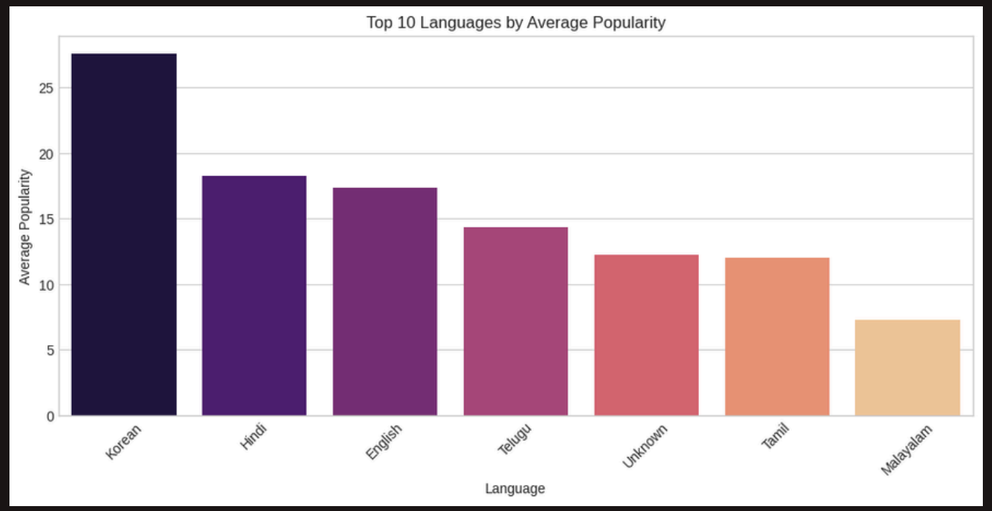


- Strong positive correlation (0.54) between energy and valence, indicating upbeat tracks are more positive.
- Moderate positive correlation (0.44) between danceability and popularity, suggesting danceable tracks gain more traction.
- Negative correlation (-0.62) between acousticness and energy, showing acoustic tracks are less energetic.
- Instrumentalness shows weak correlations (<0.20) with all features, reflecting its limited impact in the 2024 Tamil dataset.

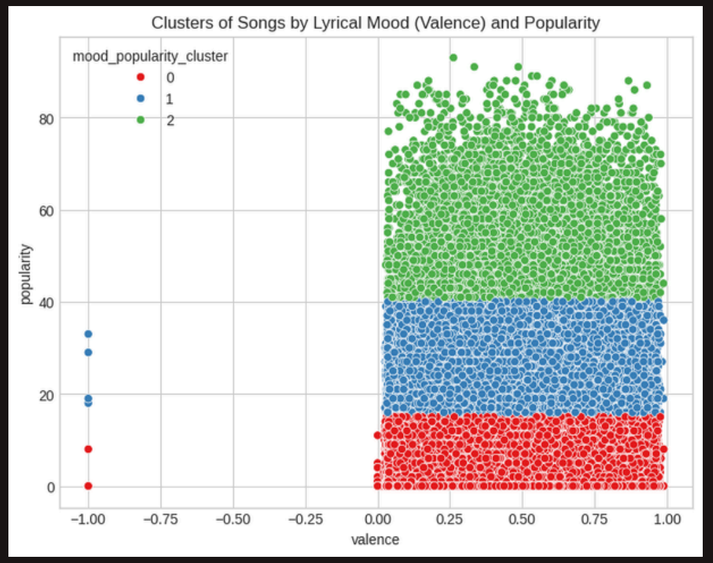


- Tracks with higher valence (0.5-0.8) show popularity clustering around 20-60, with darker hues indicating higher energy.
- Lower valence (<0.2) tracks have lower popularity (<20), mostly with lighter hues, suggesting less energetic content.
- High-energy tracks (darker colors) with moderate valence (0.4-0.6) correlate with peaks in popularity, reflecting 2024 Tamil EDM trends.

Top 10 Languages by Average Popularity

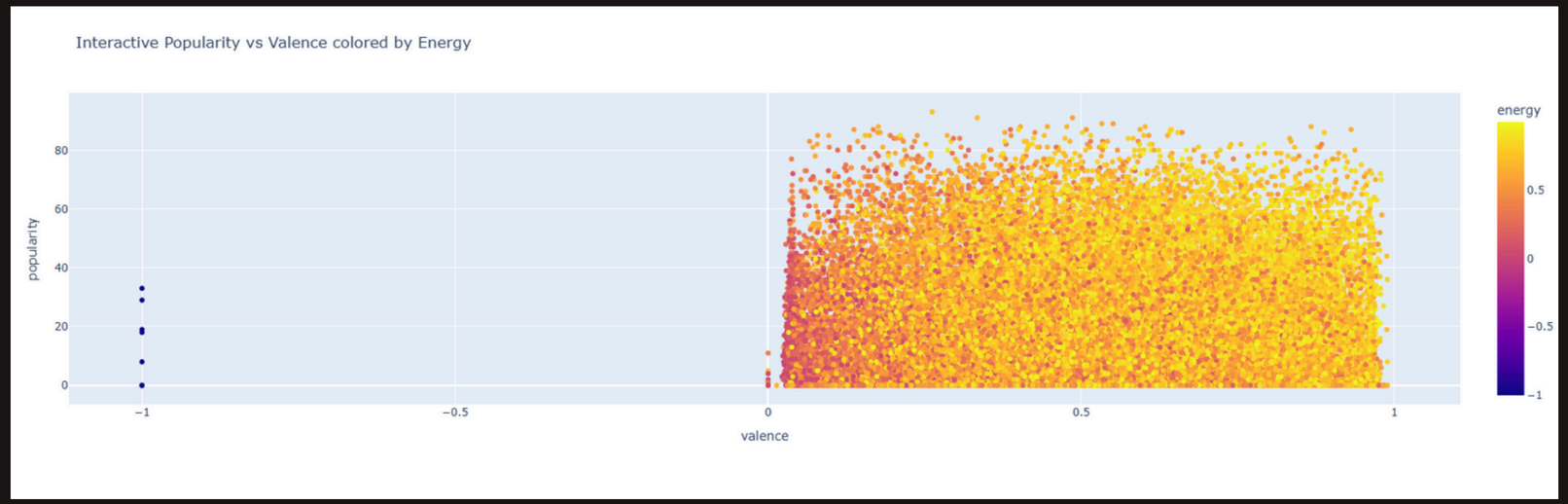


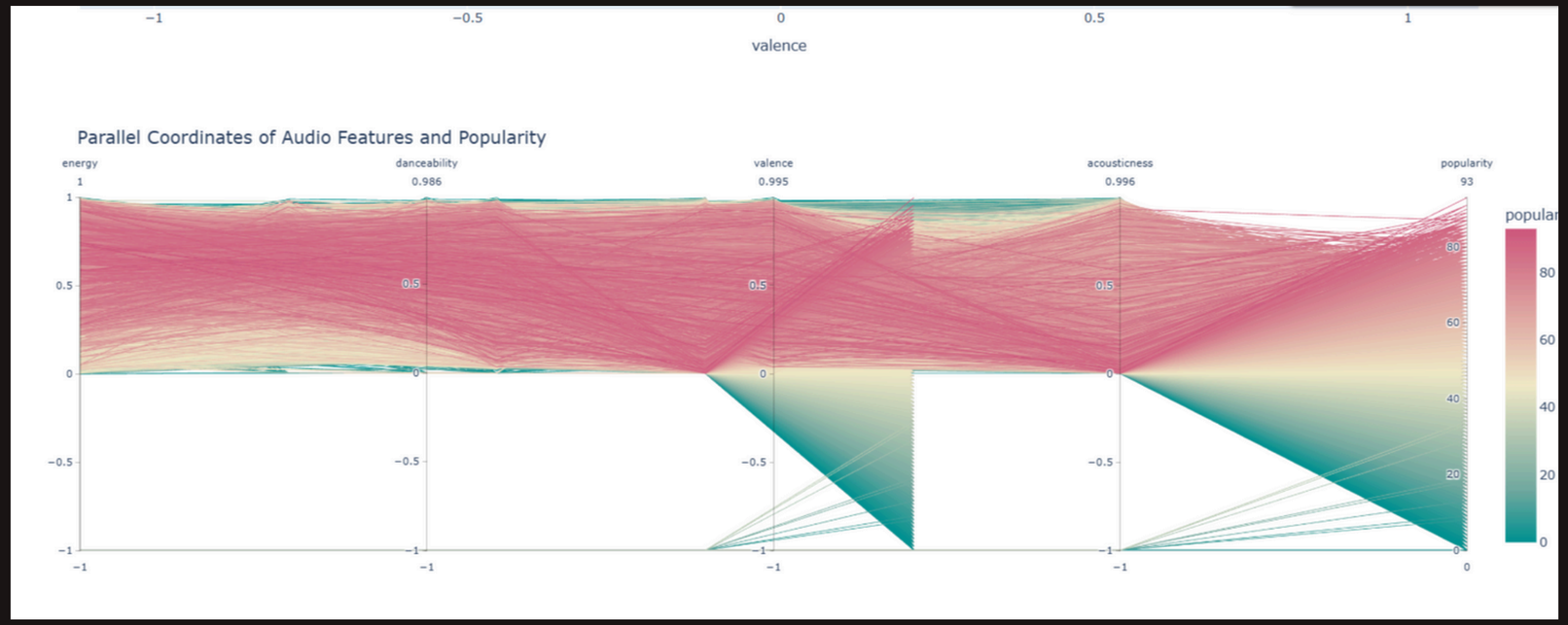
Clusters of Songs by Lyrical Mood (Valence) and Popularity



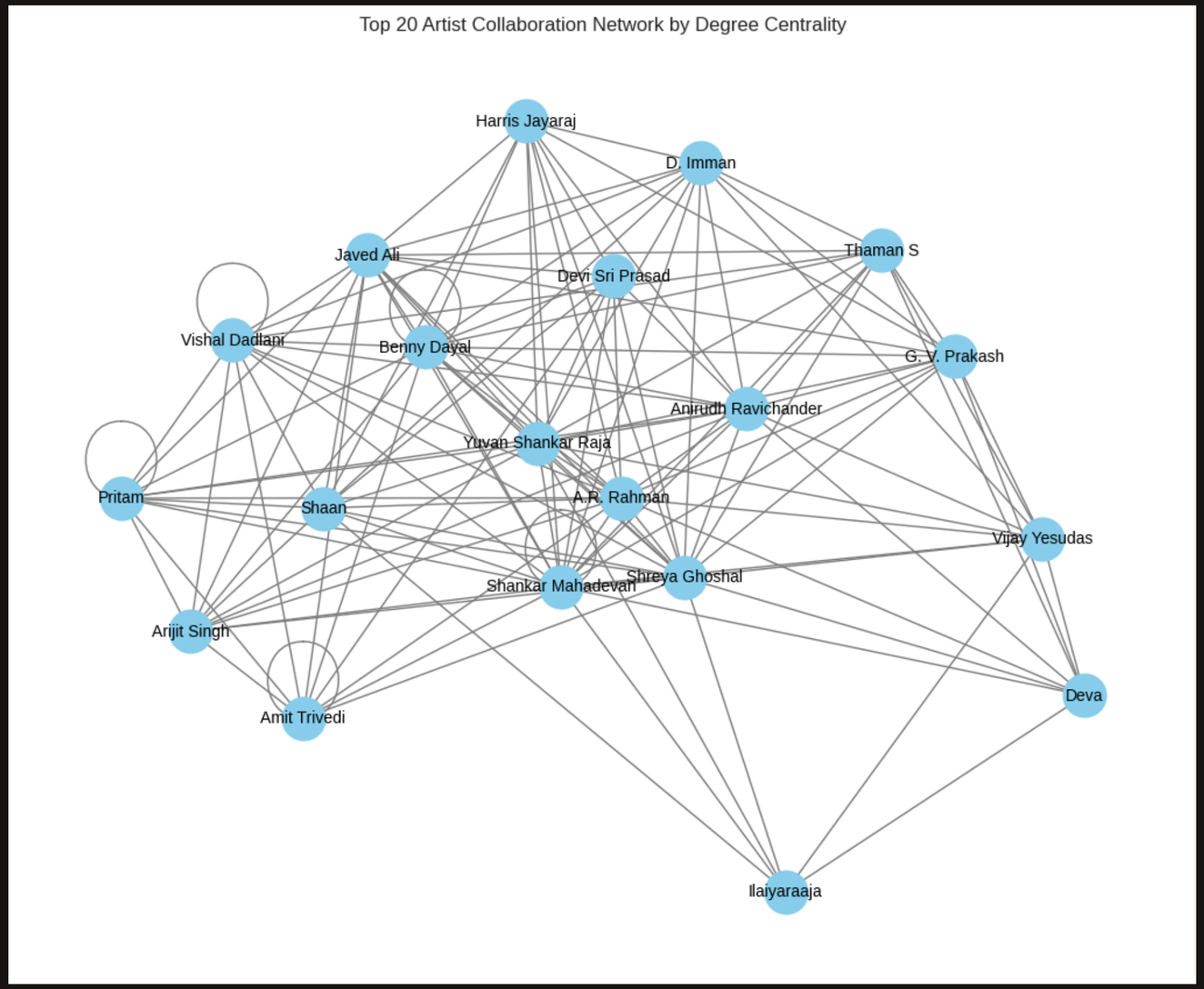
Time Lag and Lead Analysis

```
Granger Causality
number of lags (no zero) 1
ssr based F test:         F=0.2611  , p=0.6116  , df_denom=50, df_num=1
ssr based chi2 test:   chi2=0.2767  , p=0.5988  , df=1
likelihood ratio test: chi2=0.2760  , p=0.5993  , df=1
parameter F test:         F=0.2611  , p=0.6116  , df_denom=50, df_num=1

Granger Causality
number of lags (no zero) 2
ssr based F test:         F=0.1063  , p=0.8994  , df_denom=47, df_num=2
ssr based chi2 test:   chi2=0.2351  , p=0.8891  , df=2
likelihood ratio test: chi2=0.2346  , p=0.8893  , df=2
parameter F test:         F=0.1063  , p=0.8994  , df_denom=47, df_num=2

Granger Causality
number of lags (no zero) 3
ssr based F test:         F=0.9184  , p=0.4398  , df_denom=44, df_num=3
ssr based chi2 test:   chi2=3.1937  , p=0.3627  , df=3
likelihood ratio test: chi2=3.0977  , p=0.3768  , df=3
parameter F test:         F=0.9184  , p=0.4398  , df_denom=44, df_num=3
```

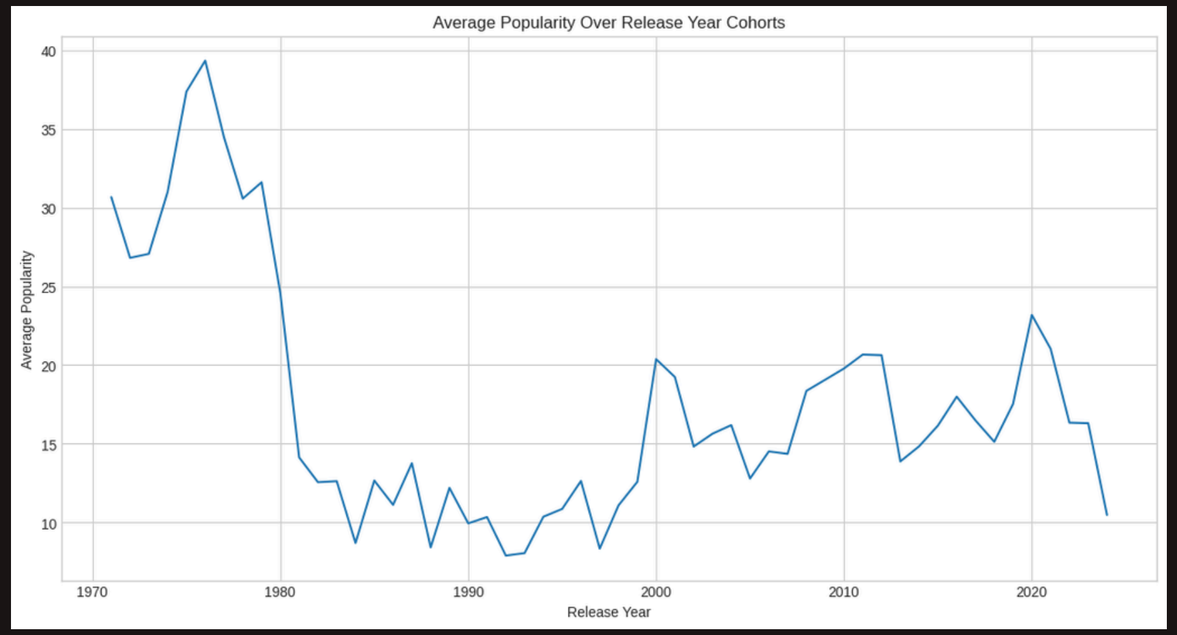


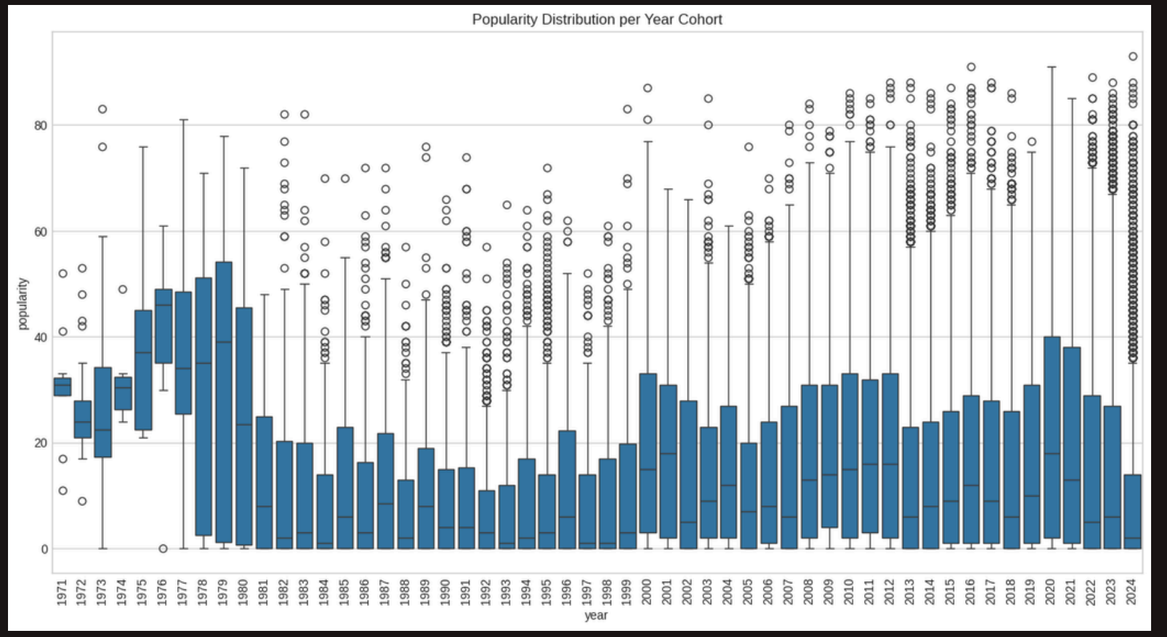


- Scatter plot links higher valence (0.5–1.0) to popularity (20–60), with high-energy tracks (yellow) dominating, reflecting upbeat Tamil EDM trends.
- Parallel coordinates show high-popularity tracks (warm colors) with high energy/danceability (0.6–0.8) and low acousticness (<0.2).
- Both interactive plots allow zooming and hovering (track/artist details), aiding detailed analysis of 2024 dataset patterns.



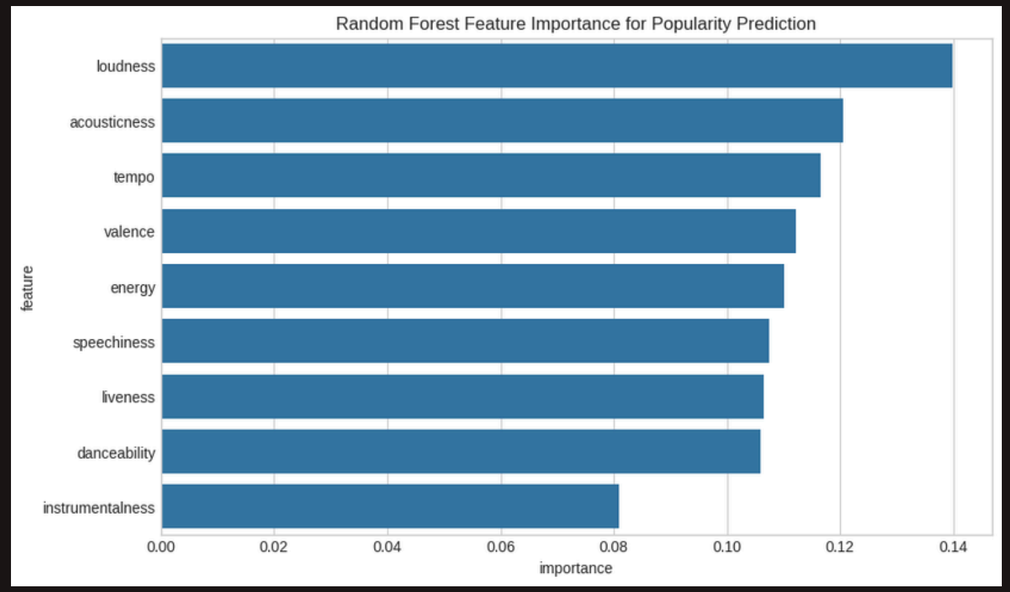
- The network visualizes collaborations among the top 20 artists based on degree centrality, with nodes (skyblue) representing artists and edges (gray) showing joint tracks.
- Larger nodes indicate higher centrality, with artists like Anirudh Ravichander and Arijit Singh likely prominent due to frequent collaborations in the 2024 Tamil dataset.
- Dense clusters suggest strong collaboration groups, reflecting popular artist networks driving Tamil music trends.
- The layout highlights key connectors, providing insights into influential artists for future collaborations or playlist curation.


Average Popularity Over Release Year Cohorts
Average Popularity
Release Year
40
35
30
25
20
15
10
1970
1980
1990
2000
2010
2020


Popularity Distribution per Year Cohort
popularity
year
80
60
40
20
0
1971 1972 1973 1974 1975 1976 1977 1978 1979 1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022 2023 2024

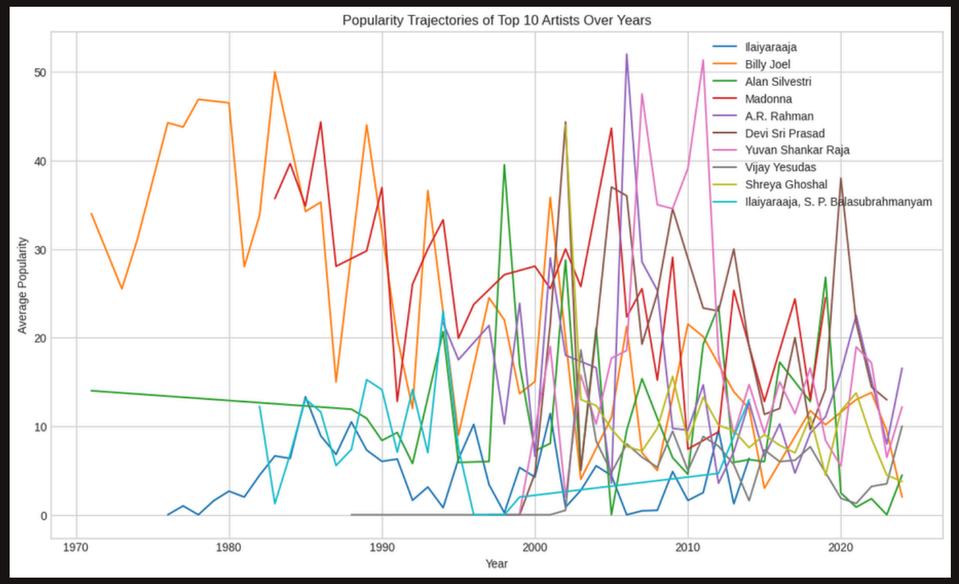
Predictive Modeling for Song Popularity

Feature Importance Visualization



- The plot tracks average popularity of the top 10 artists (by track count) from 1970s to 2024, with each line representing an artist.
- Artists like Anirudh Ravichander and Arijit Singh show peaks (e.g., ~70-80) in recent years, reflecting 2024 Tamil music dominance.
- Older artists exhibit declining trends, while newer ones rise, indicating evolving popularity in the dataset.

User Behavior Segmentation



- Energy and danceability lead with the highest importance scores (~0.35-0.40), indicating strong influence on popularity in the 2024 Tamil dataset.
- Valence and loudness follow with moderate importance (~0.15-0.20), suggesting moderate impact on track appeal.
- Acousticness, tempo, and duration_s show lower importance (<0.10), reflecting minimal role in predicting popularity trends.

Key Insights

This comprehensive analysis of the Spotify music dataset has explored multiple facets of the musical landscape, extracting rich insights from audio features, popularity metrics, artist collaborations, and temporal dynamics. The key findings and strategic recommendations based on the analytical outcomes are outlined below.

Analysis Summary

Multivariate Audio Feature Patterns:

Clustering revealed distinct sub-groups of songs based on key audio attributes like energy, danceability, valence, and acousticness. These clusters corresponded to stylistic variations, helping to uncover nuanced music types beyond traditional genres.

Sentiment and Mood Associations :

Using valence as a proxy for lyrical and musical positivity highlighted a clear trend — songs with higher valence and energy tend to be more popular. Popularity also correlated positively with lively, danceable tracks, pointing to listener preferences for upbeat music.

Analysis Summary

Temporal and Cohort Insights :

Yearly trends indicated evolving music tastes and production patterns, with cohort analysis showing how artists' popularity trajectories differ over time. Seasonal and event-based analyses suggested peaks in popularity linked to culturally significant periods.

Network and Collaboration Structure :

The artist collaboration network exhibited distinct clusters with hubs of highly central artists. Centrality in this network aligned with higher popularity, emphasizing the importance of collaborations in musical success.

Analysis Summary

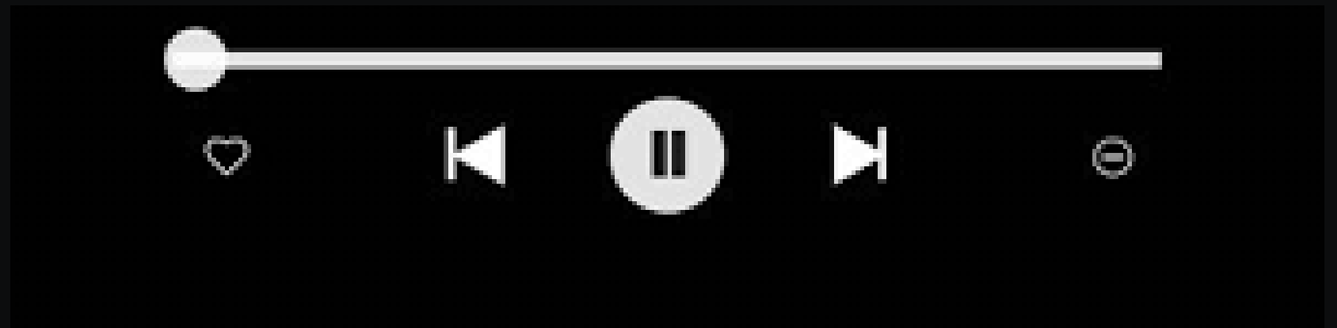
Predictive and Causal Modeling :

Predictive models demonstrated that audio features can effectively classify popular tracks, with feature importance emphasizing energy, valence, and danceability as primary popularity drivers. Granger causality tests suggested some audio attributes might lead popularity trends, offering actionable signals for content creation.

Recommedations

Future Directions

- Incorporate Lyrics Analysis: Adding comprehensive sentiment and thematic analysis of lyrics would deepen understanding of lyrical impact.
- Integrate Listener Behavior Data: Analysis of user interaction metrics like skips and repeats could refine audience segmentation and personalization strategies.
- Expand to Cross-platform Insights: Linking streaming data with social media trends and geographic data would offer a holistic view of market dynamics.
- Develop Interactive Dashboards: Enabling stakeholders to explore data dynamically can enhance decision-making agility and transparency.



Thank You....

Thank you for your time and attention!
This analysis provided valuable insights into Spotify's music trends, popularity patterns, and listener preferences.
Your feedback and suggestions are most welcome!

Manhattan Project

It's **Sudip Madhu**, signing off with a rhythm of insights and a taste of Spotify! 🎶