Exploratory Data Analysis on Spotify Tracks

Subtitle: Data Science Project

Name: Arka Koley

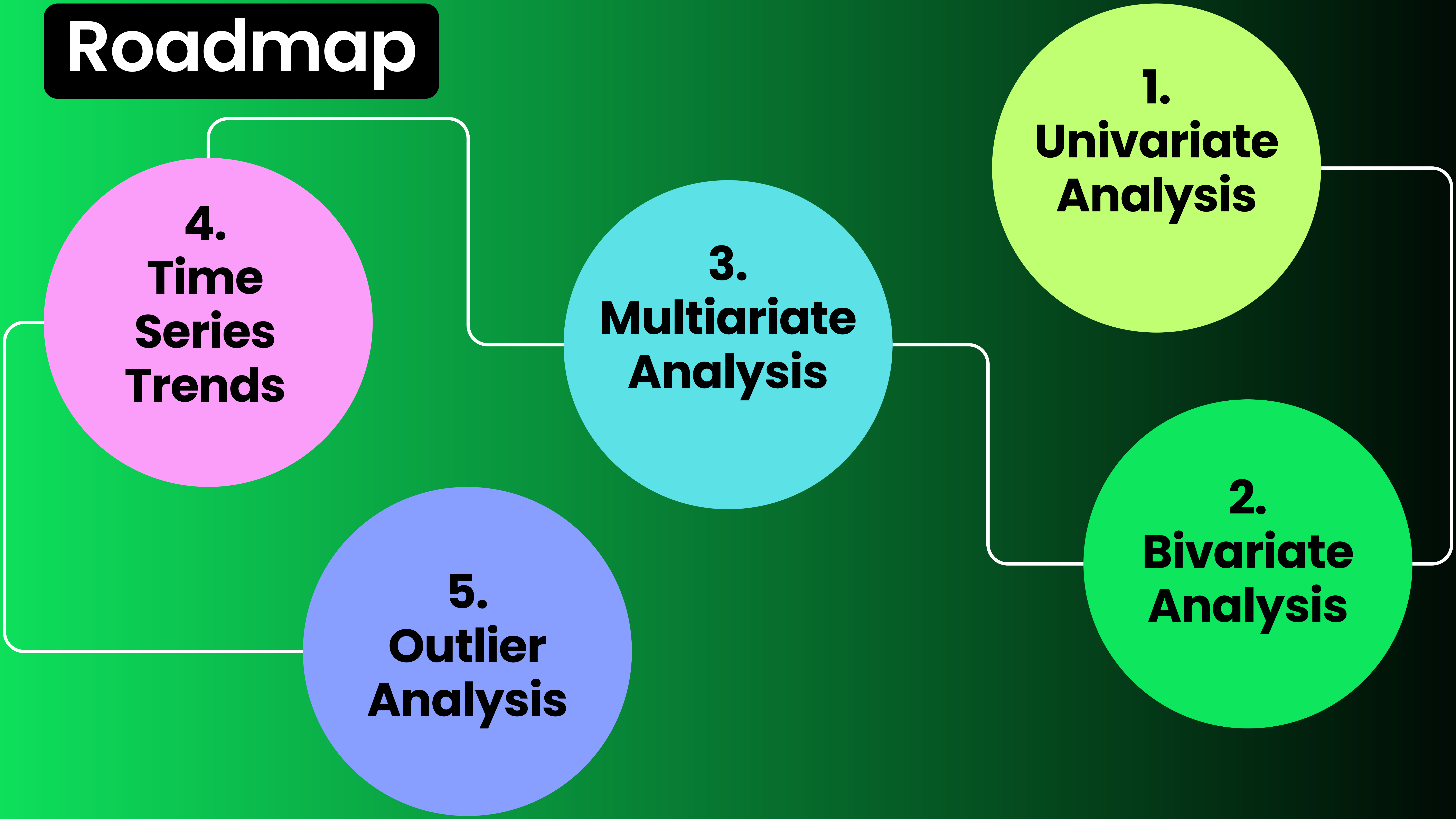
Tools Used: Python (Pandas, Seaborn, Matplotlib)

Date: October 2025

Introduction

- This project explores Spotify track data using Exploratory Data Analysis (EDA).
- EDA is crucial for understanding patterns, trends, and relationships within a dataset before building any formal models.
- Objective: To identify the features that contribute to a song's popularity and to analyze how musical trends have evolved over time.

Roadmap



1. Univariate Analysis

2. Bivariate Analysis

3. Multiariate Analysis

4. Time Series Trends

5. Outlier Analysis

About the Dataset

The dataset contains a comprehensive collection of Spotify tracks, each detailed with over 20 distinct feature

Total Records: ~50,000 tracks

Key Columns:

○ Identification: track_name, artist_name, year

○ Performance Metric: popularity

○

○ Audio Features: danceability, energy, acousticness, loudness, speechiness, valence, tempo, liveness, instrumentalness

EDA Objectives

Our analysis is structured around the following objectives:

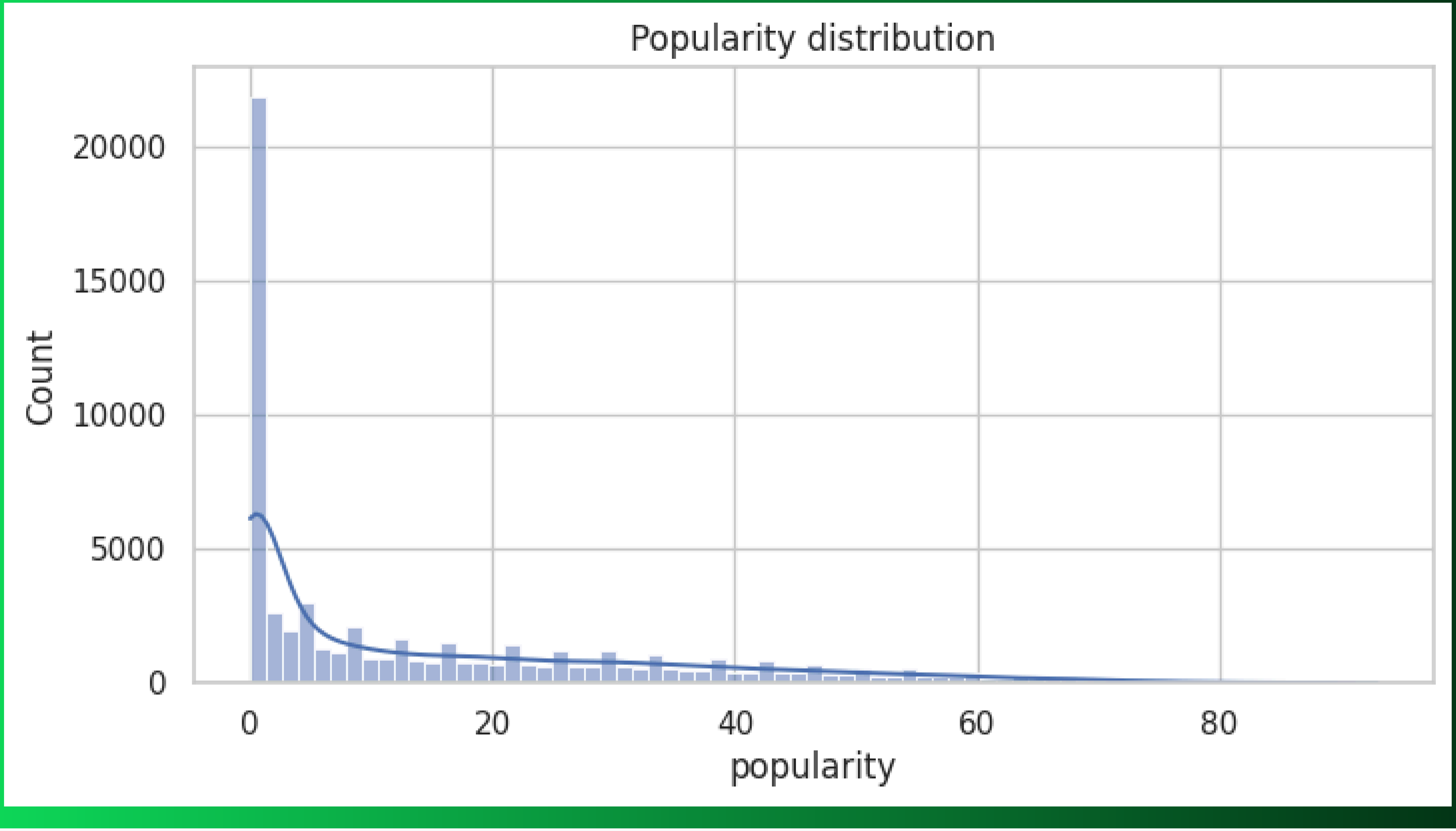
1. Univariate Analysis: Explore the distribution and characteristics of individual features.
2. Bivariate Analysis: Study the relationships and interactions between pairs of features.
3. Multivariate & Correlation Analysis: Examine the correlations between multiple features simultaneously to uncover complex patterns.
4. Time Series Analysis: Track and analyze how musical features and popularity have changed across different years.
5. Generate Insights & Recommendations: Synthesize findings to provide actionable insights.

Data Cleaning & Preprocessing

To ensure the quality and reliability of our analysis, the following data cleaning steps were performed:

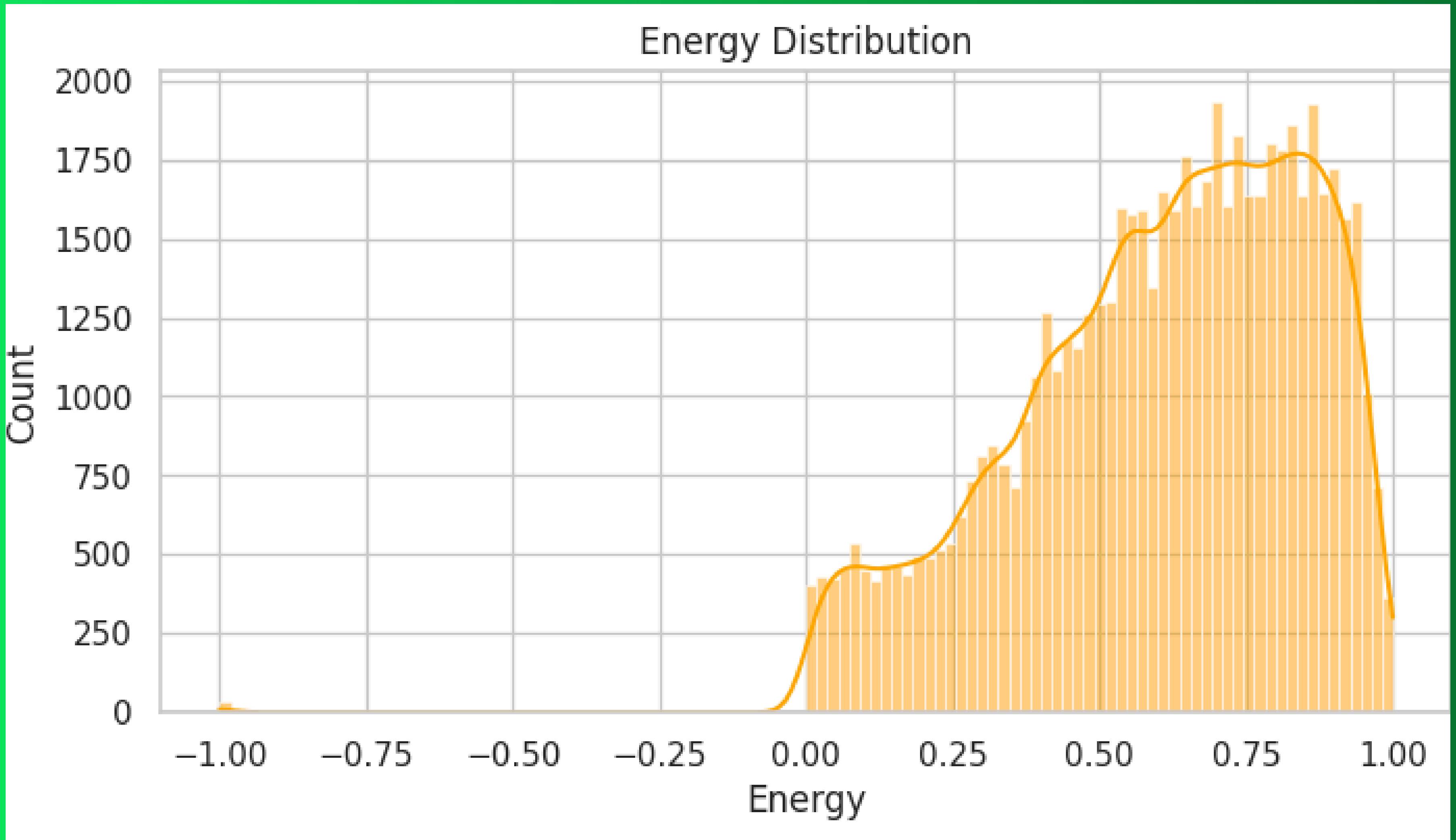
- Duplicate Removal: All duplicate rows were identified and removed.
- Missing Values: The dataset was checked for any missing values, which were handled accordingly.
- Data Type Conversion: Numeric columns were converted to their appropriate data types for accurate calculations.
- Data Verification: Column values were checked to ensure they fell within expected ranges and formats.

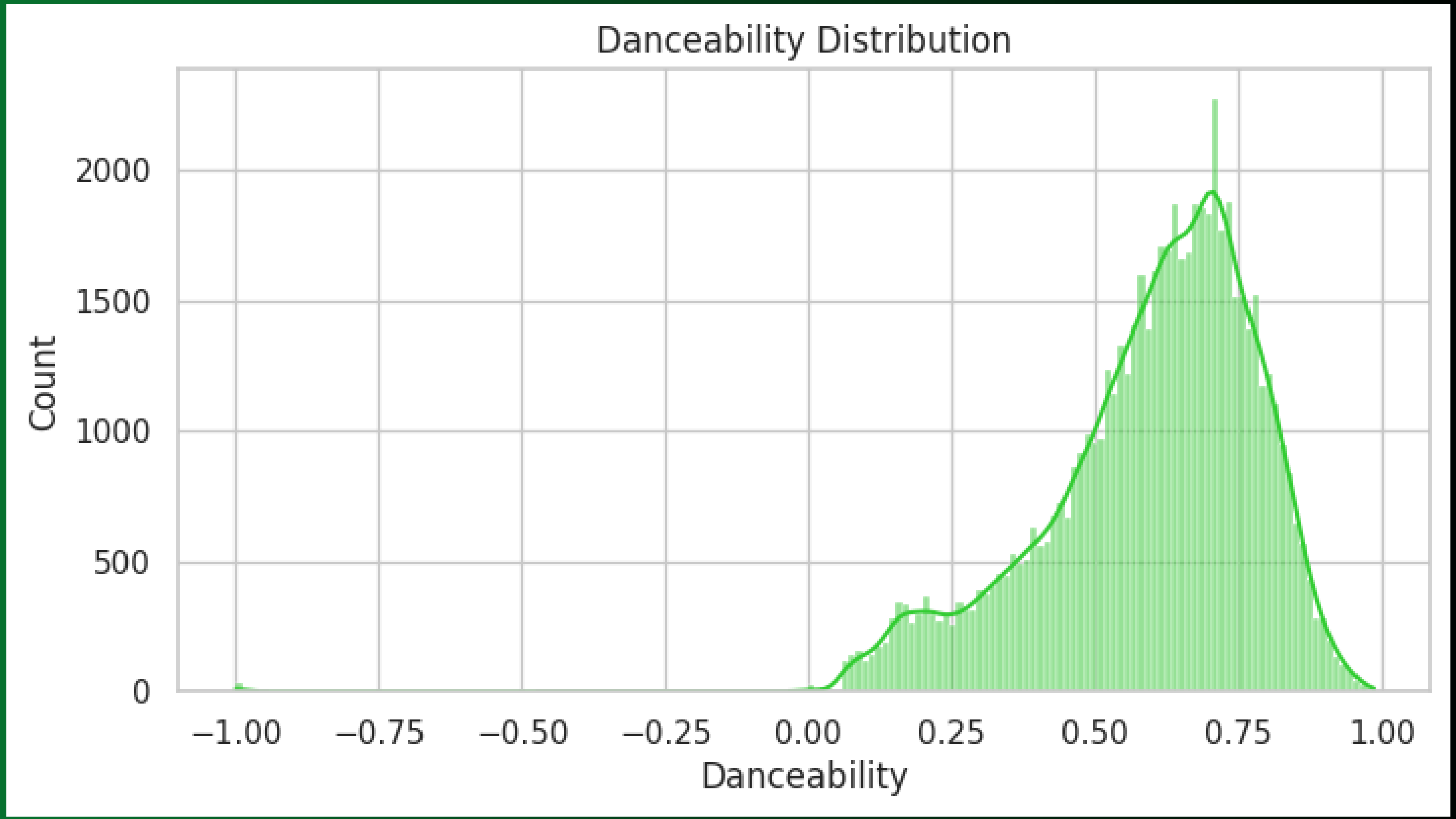
Univariate Analysis



Most songs have moderate popularity, clustering in the 40–70 range. This means that scoring an extremely high popularity (>90) is rare, making the hits stand out significantly.

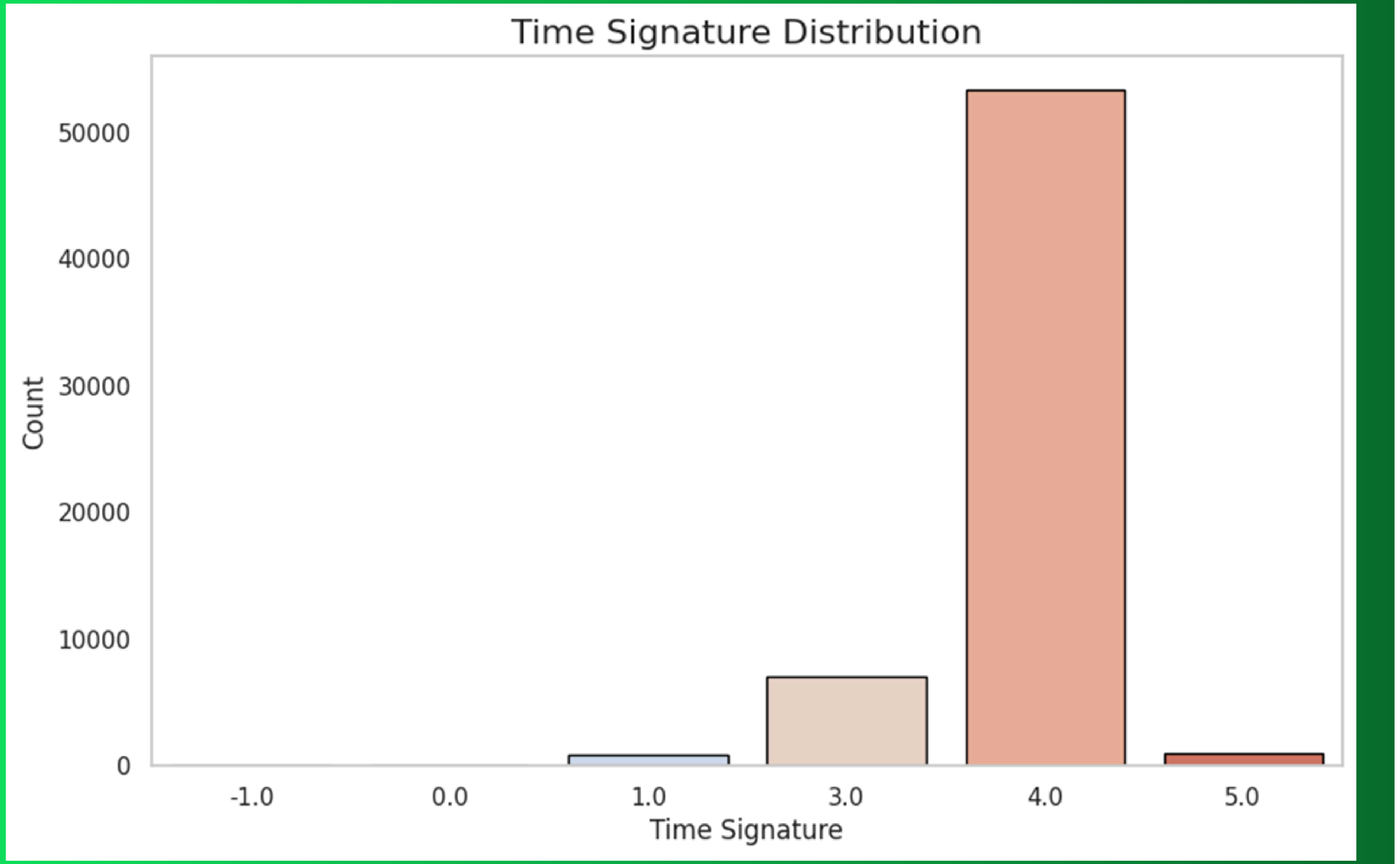
The Vibe Features: Energy, Danceability, and Mood

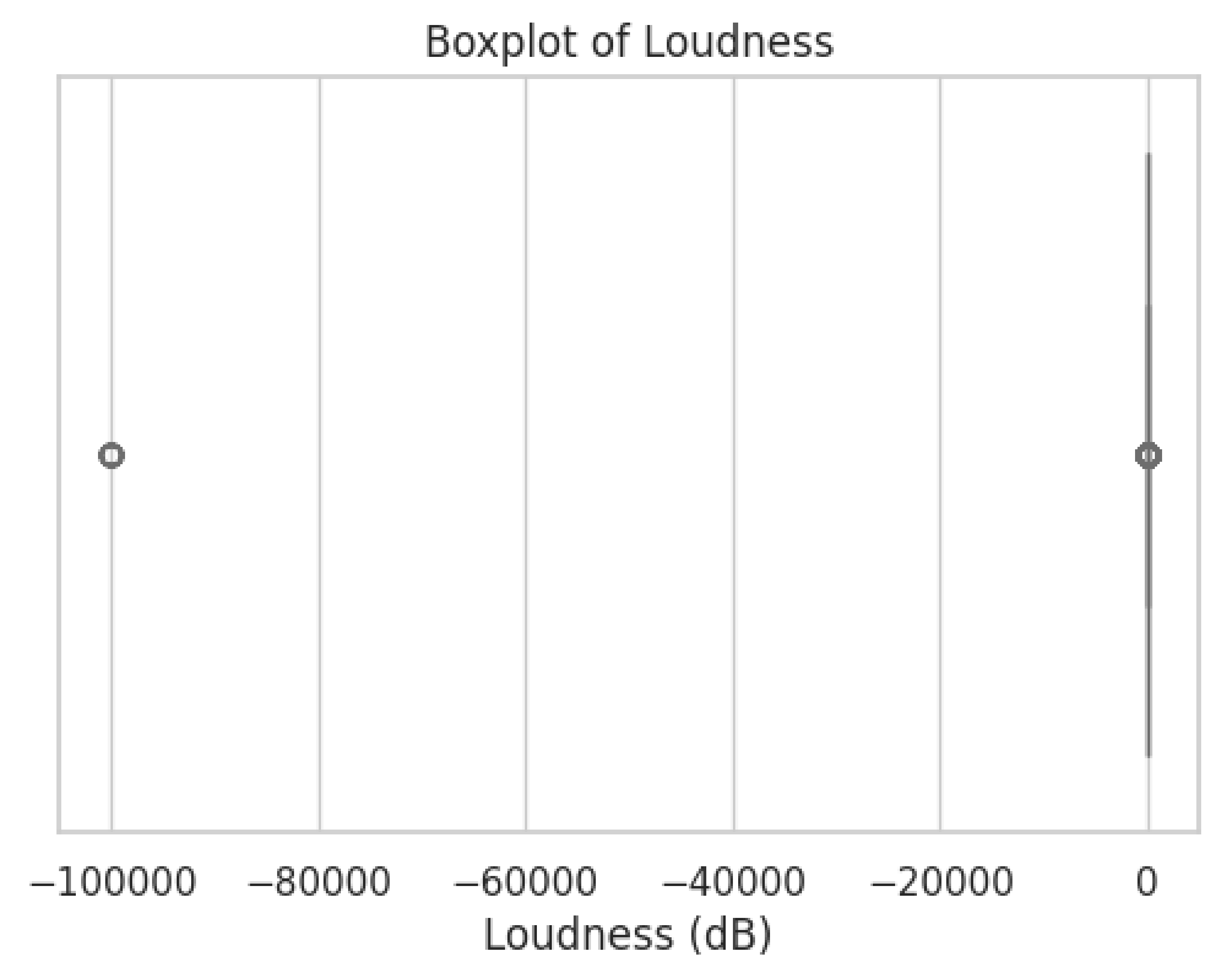




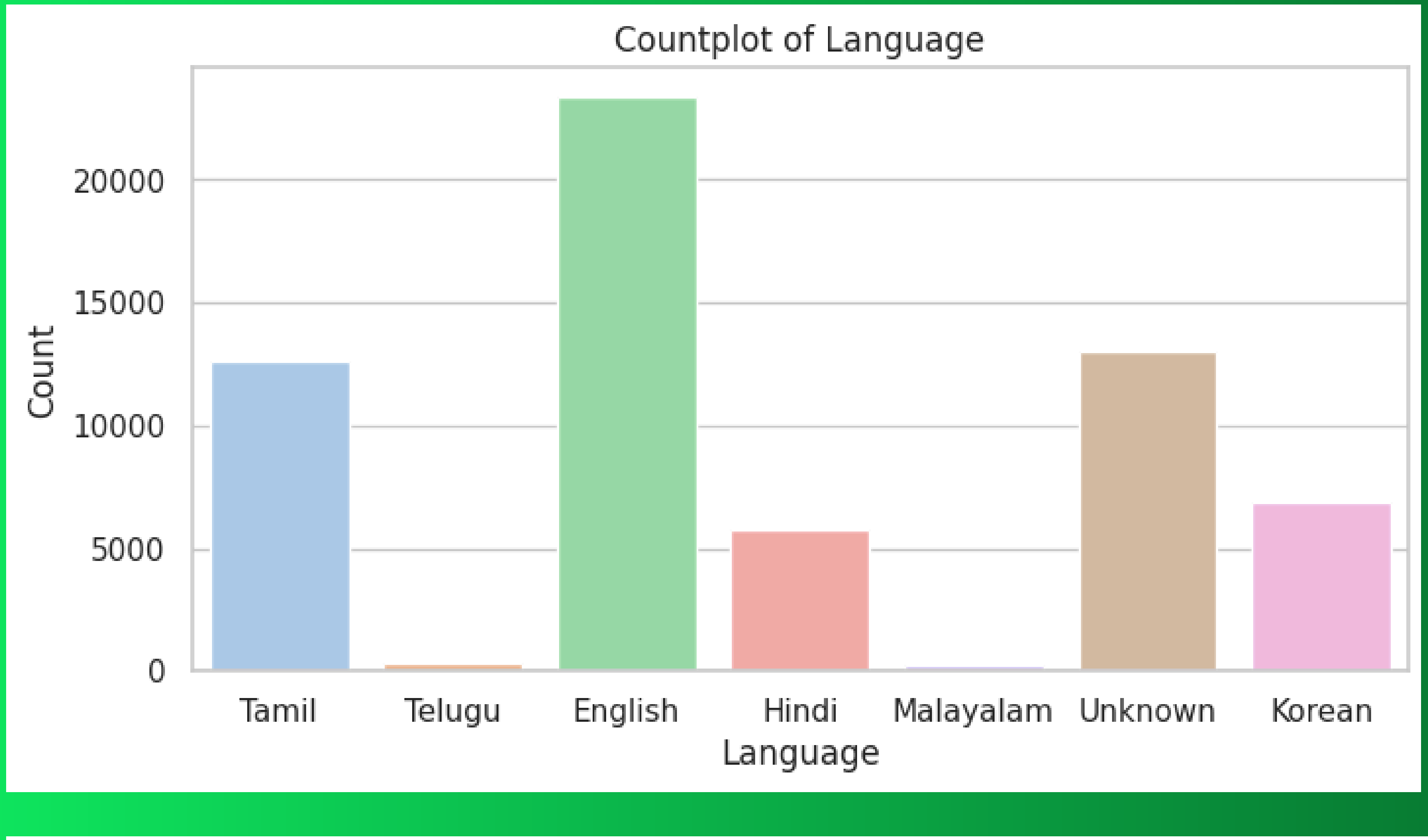
The majority of tracks are characterized by high danceability (scores >0.6).Similar to danceability, songs tend to have high energy (scores >0.7).

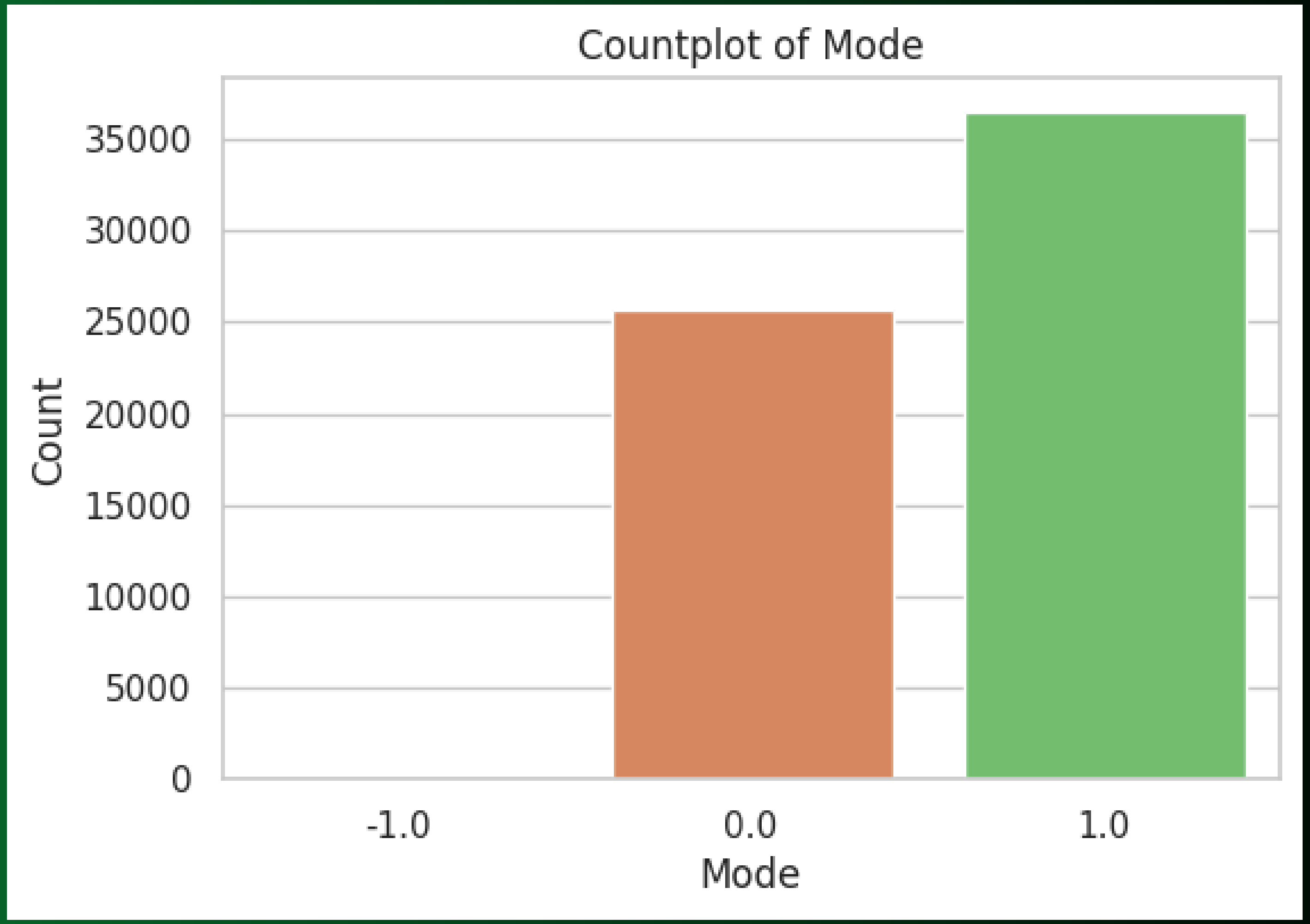
Structural Features: Tempo, Duration, and Loudness

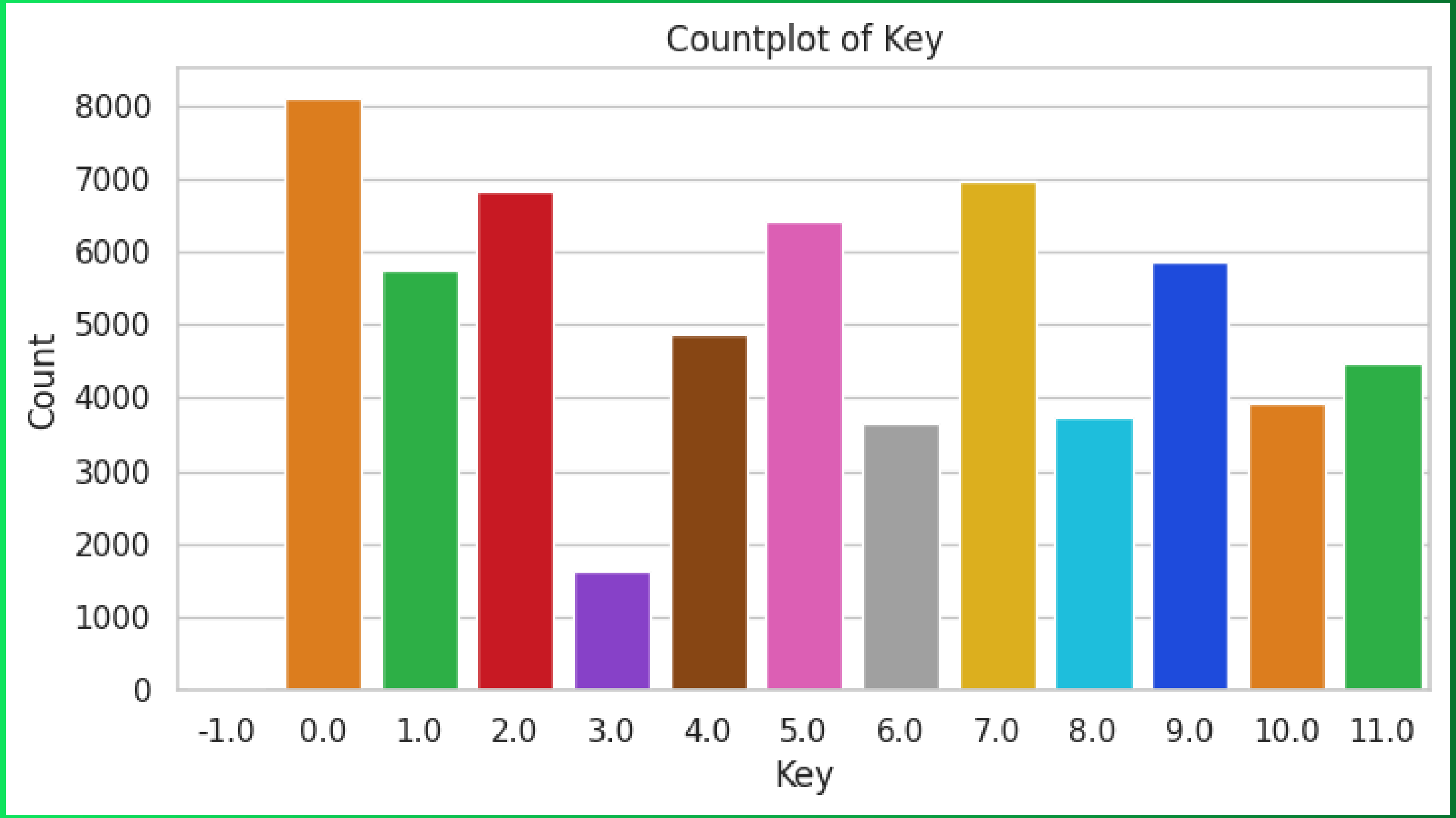


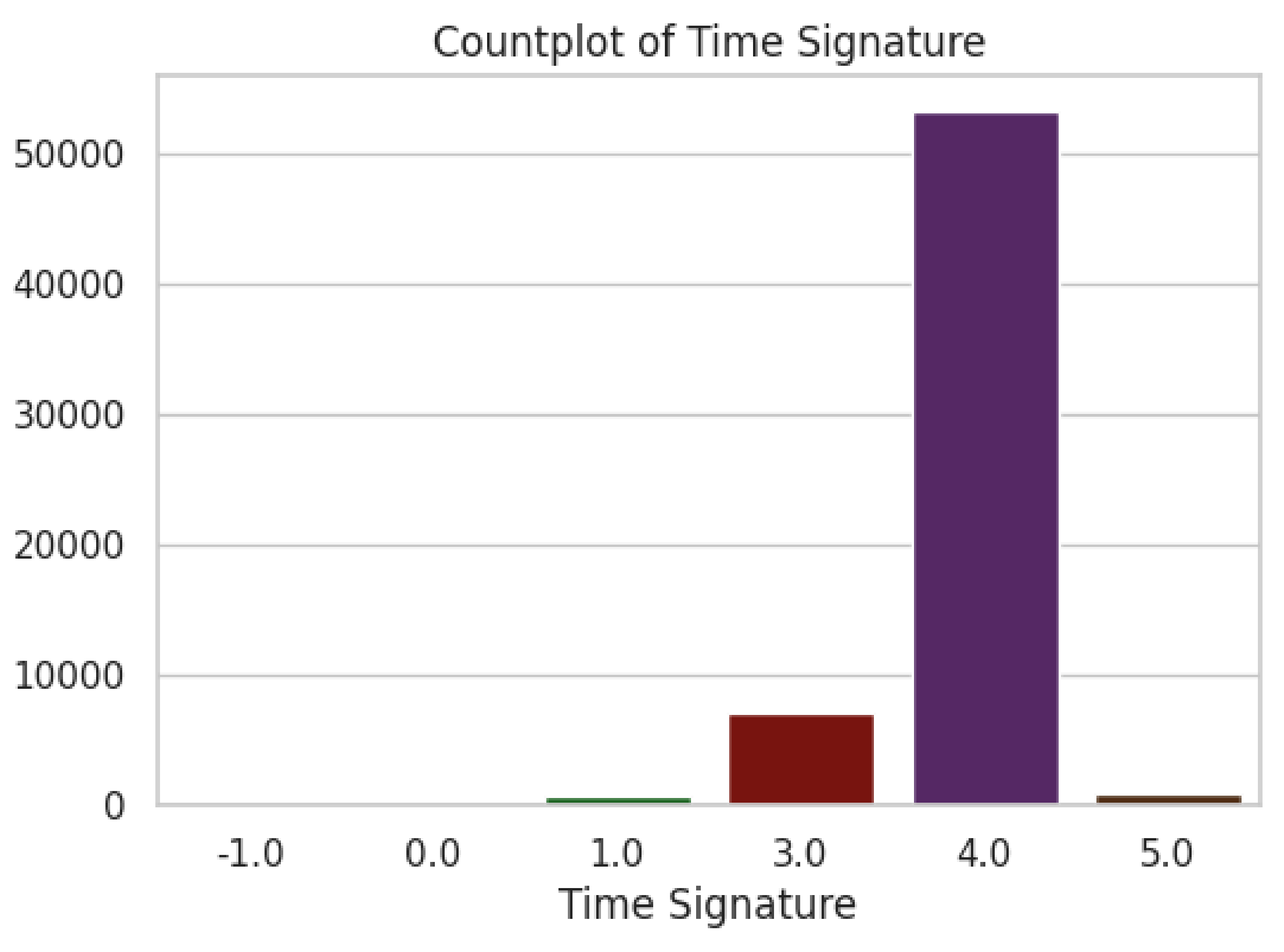


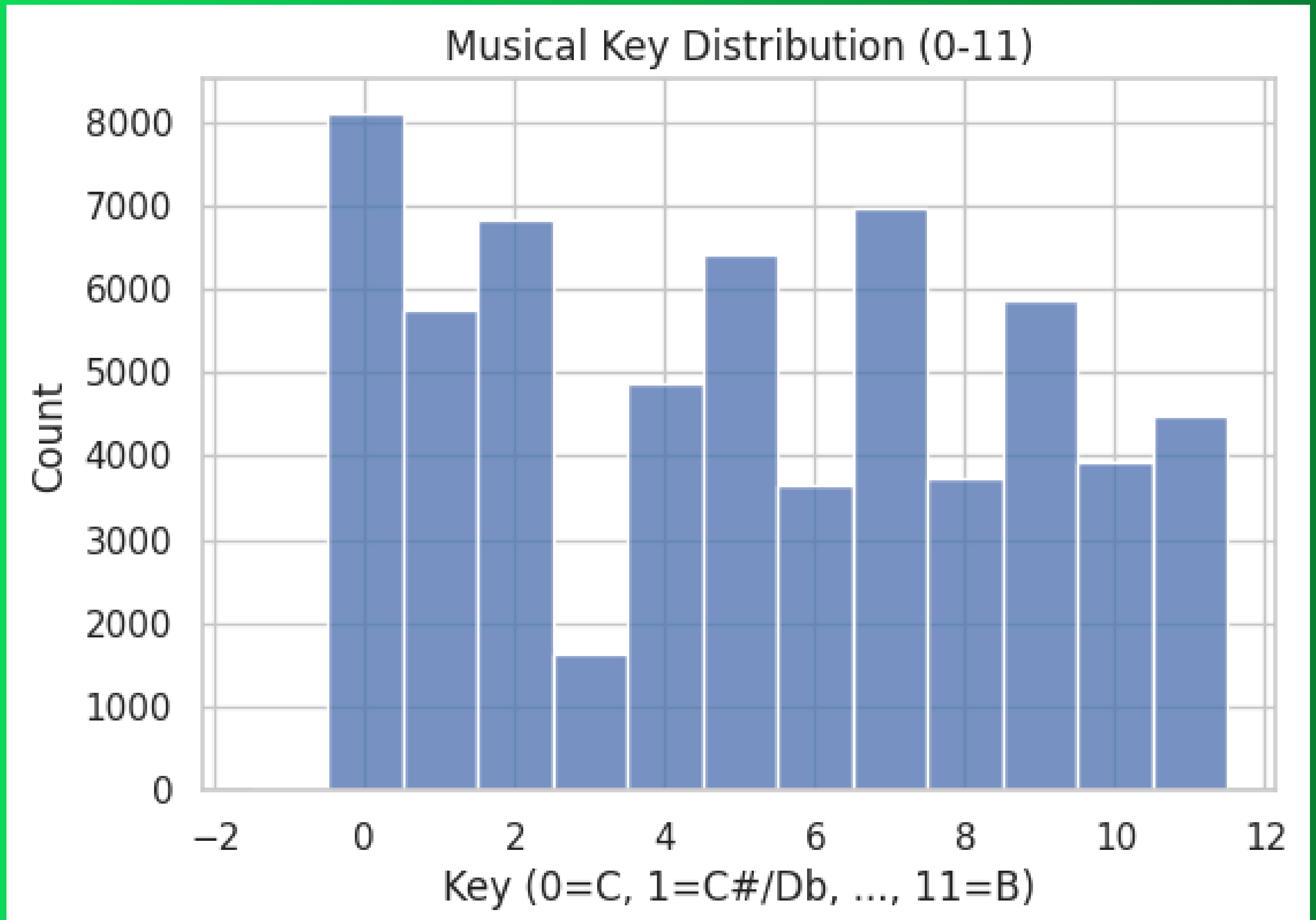
The average song length is around 3.5 minutes. Loudness distribution will likely be tightly packed in the high-decibel range (e.g., -5 dB to -10 dB).Tempo histogram confirms that music often follows popular beats per minute.

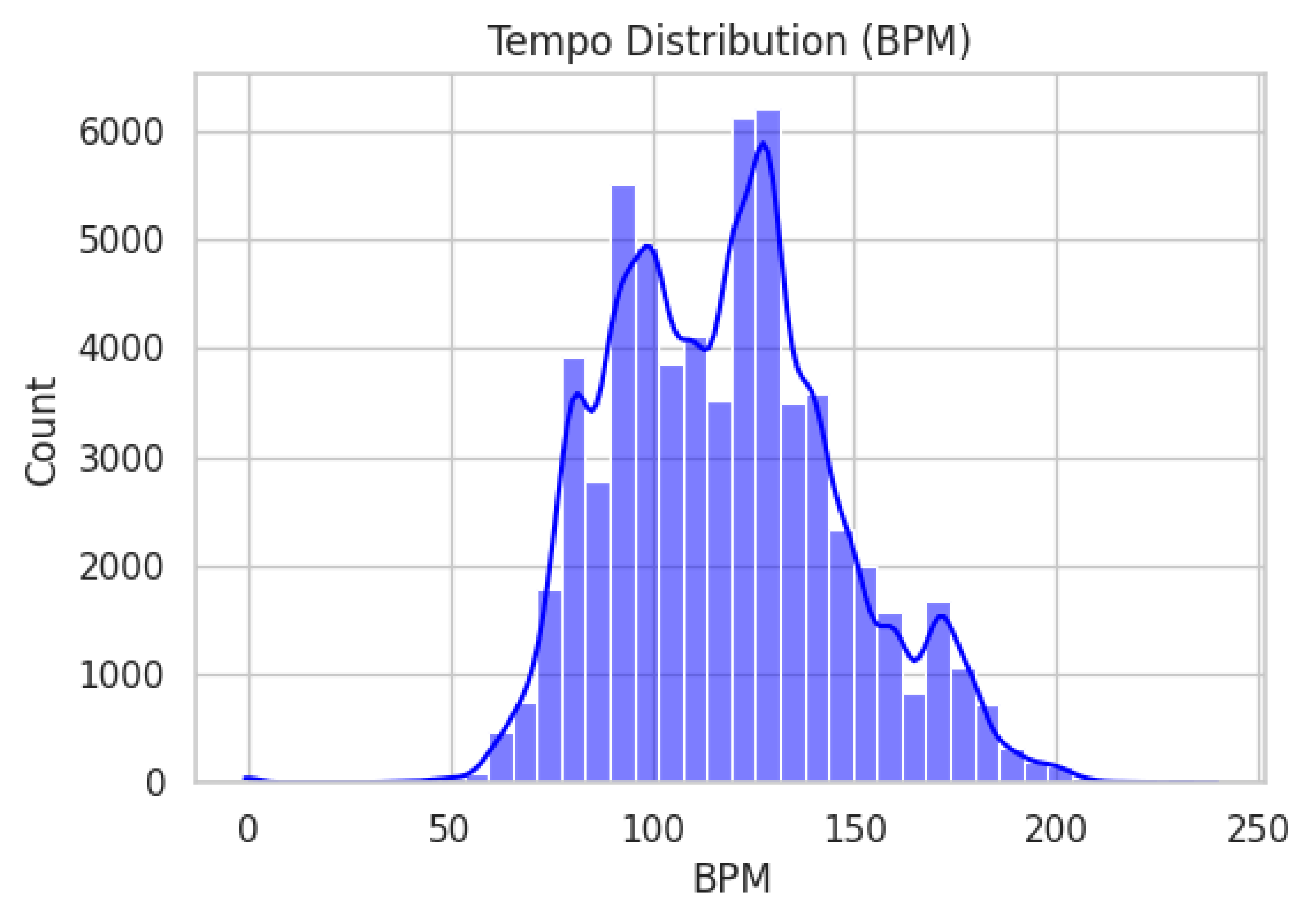

Countplot of Language
Count
0
5000
10000
15000
20000
Tamil
Telugu
English
Hindi
Malayalam
Unknown
Korean
Language

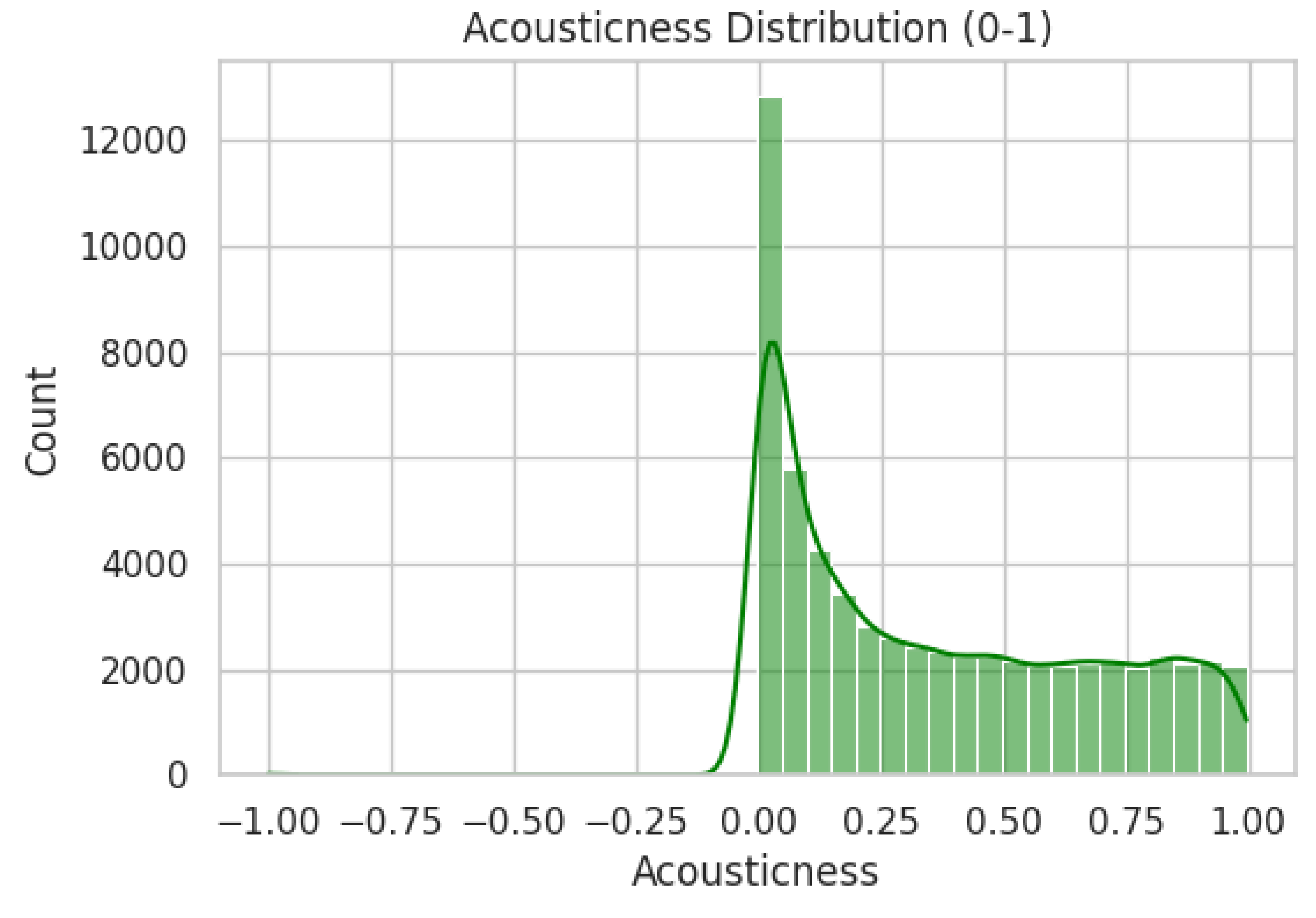

Countplot of Mode
Count
0
5000
10000
15000
20000
25000
30000
35000
-1.0
0.0
1.0
Mode

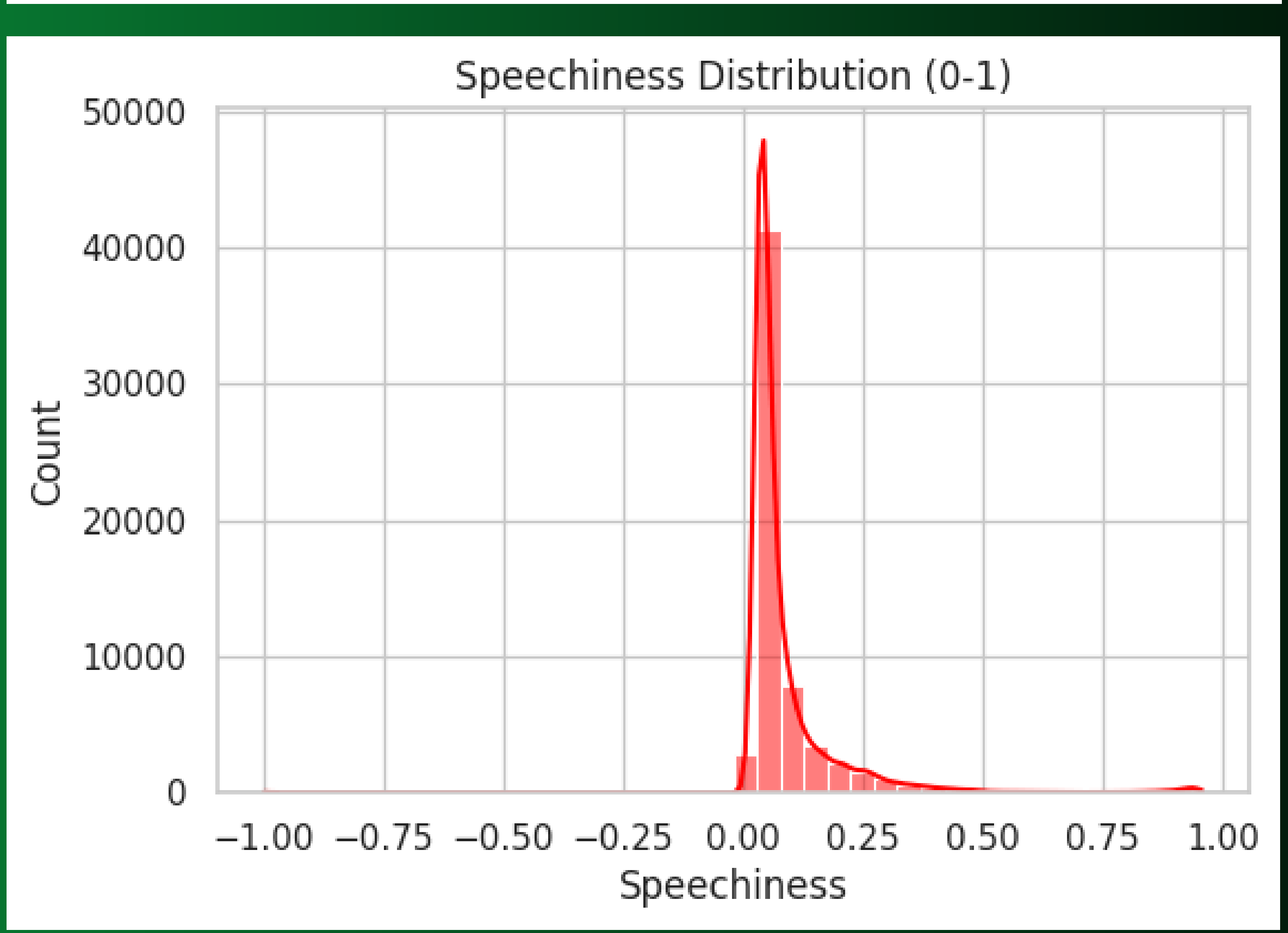

Countplot of Key
Count
0
1000
2000
3000
4000
5000
6000
7000
8000
-1.0
0.0
1.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
9.0
10.0
11.0
Key


Countplot of Time Signature
0
10000
20000
30000
40000
50000
-1.0
0.0
1.0
3.0
4.0
5.0
Time Signature

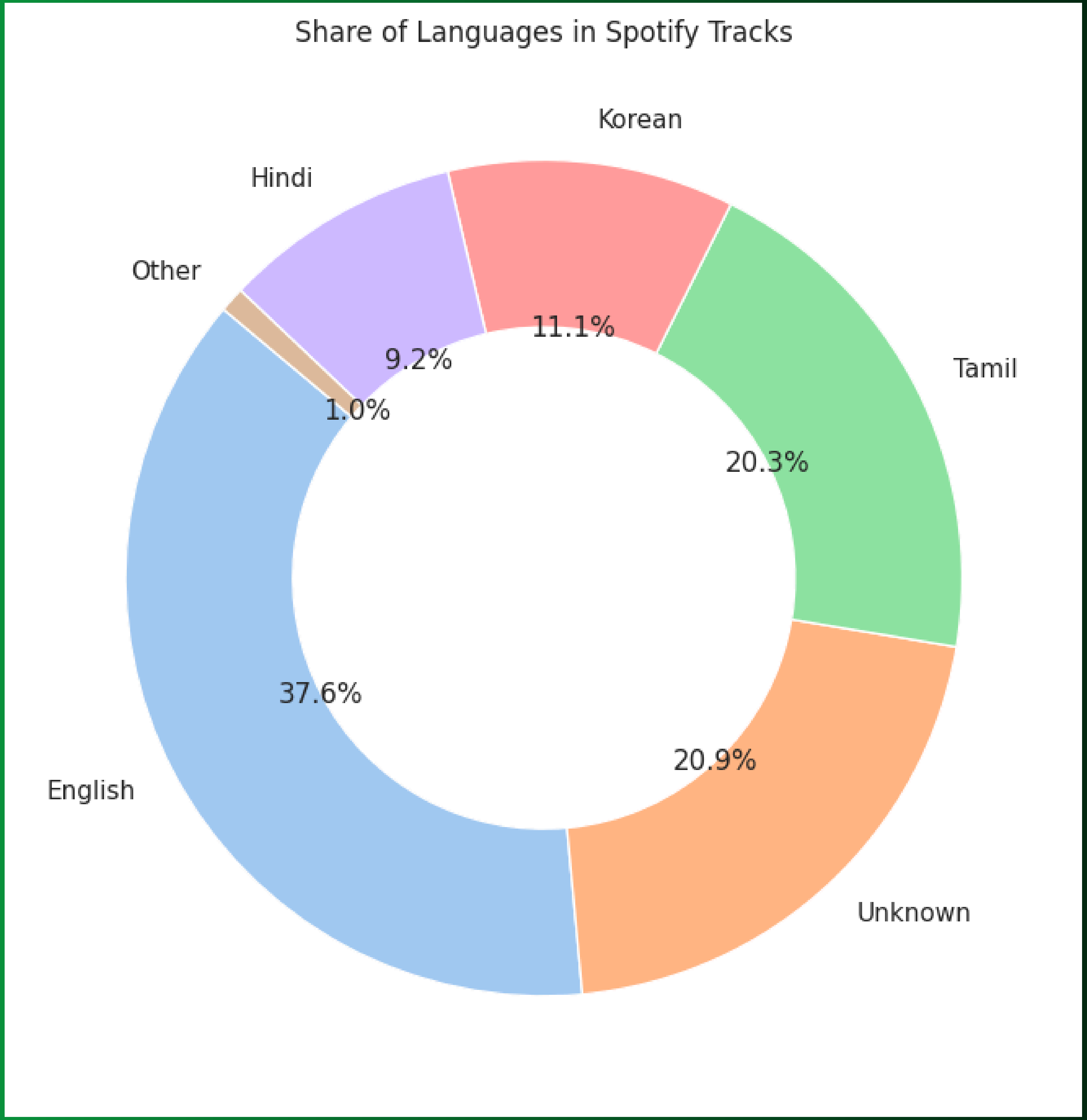

Musical Key Distribution (0-11)
Count
0
1000
2000
3000
4000
5000
6000
7000
8000
−2
0
2
4
6
8
10
12
Key (0=C, 1=C#/Db, ..., 11=B)


Tempo Distribution (BPM)
Count
0
1000
2000
3000
4000
5000
6000
0
50
100
150
200
250
BPM


Acousticness Distribution (0-1)
Count
0
2000
4000
6000
8000
10000
12000
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Acousticness


Speechiness Distribution (0-1)
Count
0
10000
20000
30000
40000
50000
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Speechiness

Univariate Analysis (Categorical Variable)

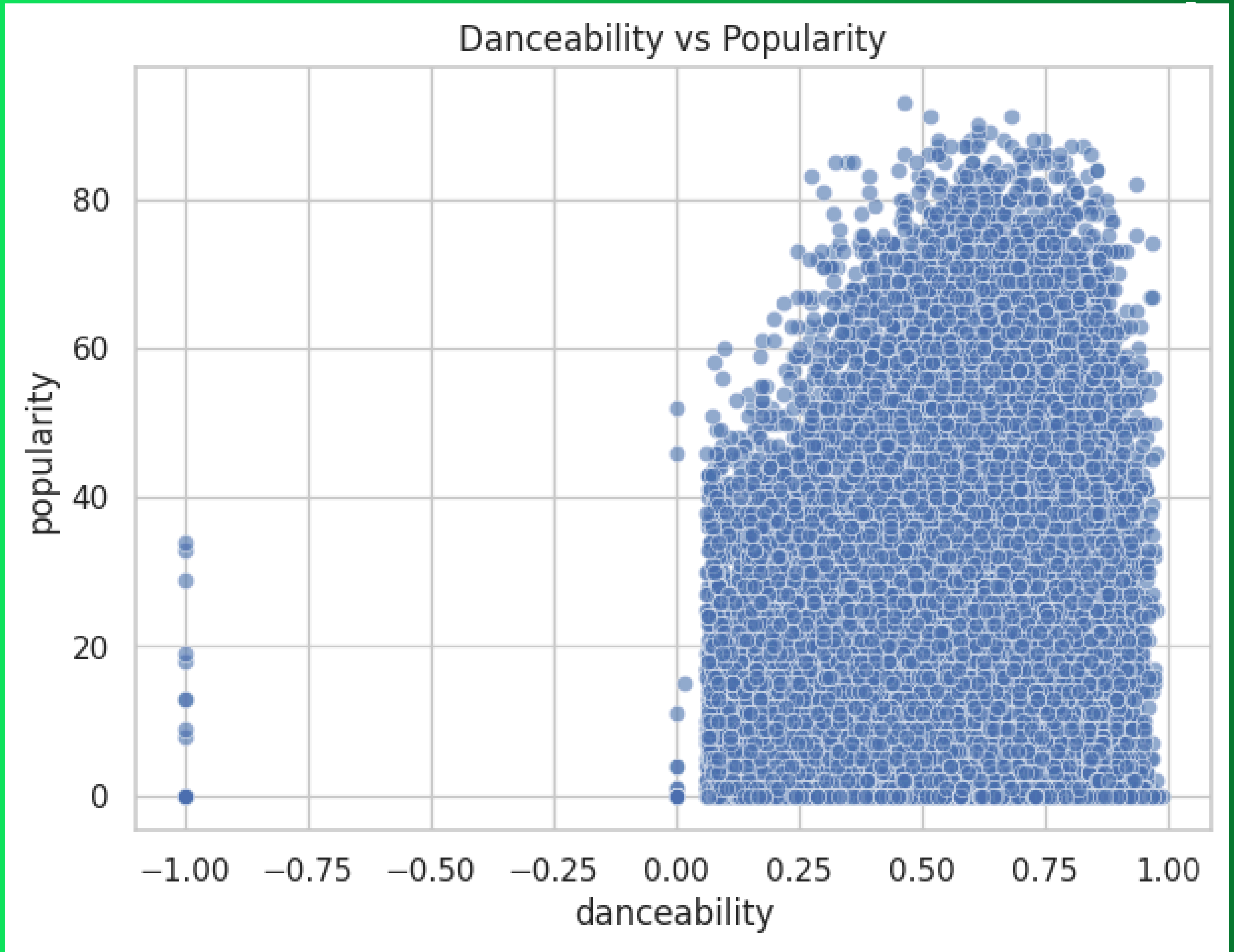


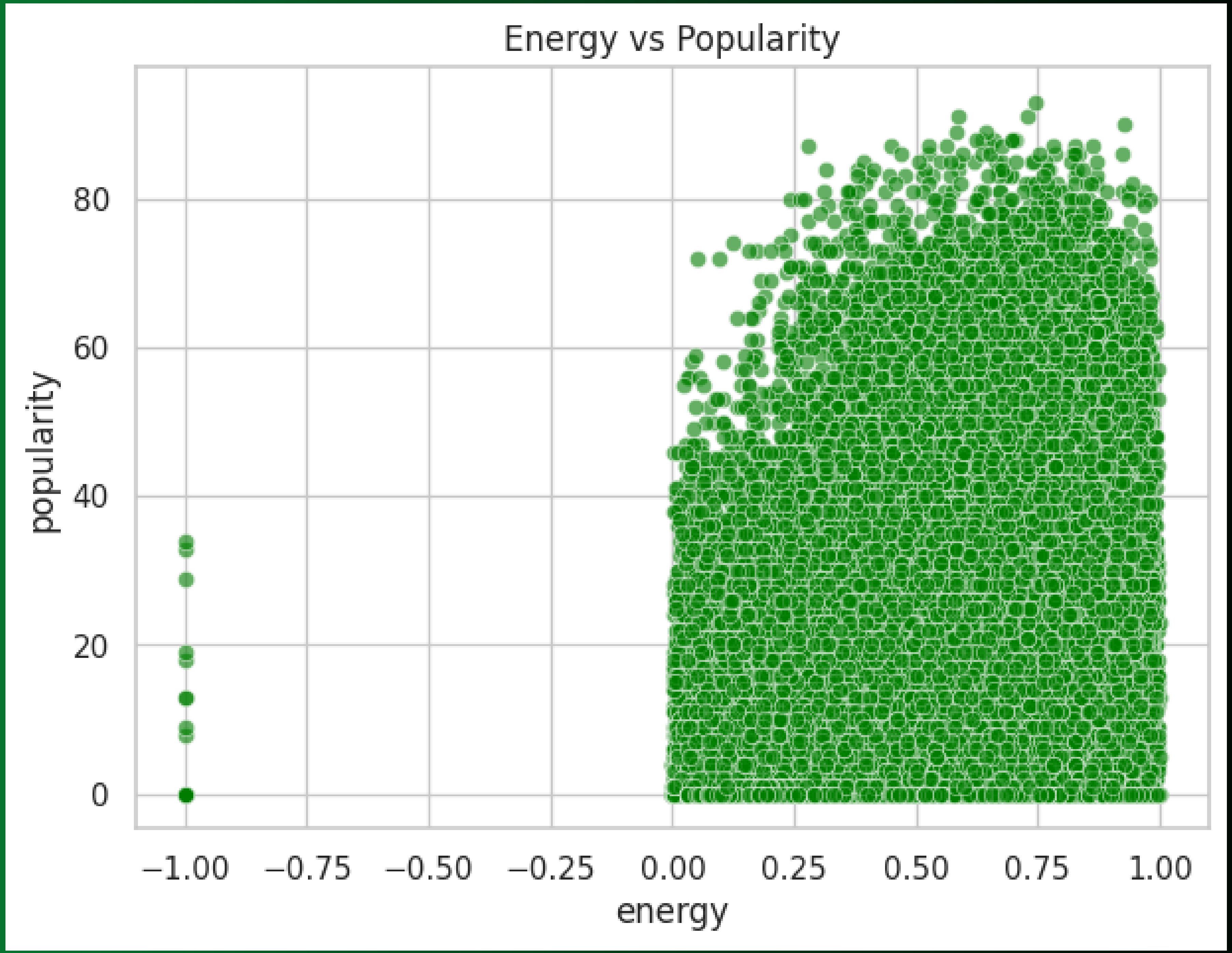
Univariate Analysis (Results)

Initial analysis of individual features revealed several key trends:

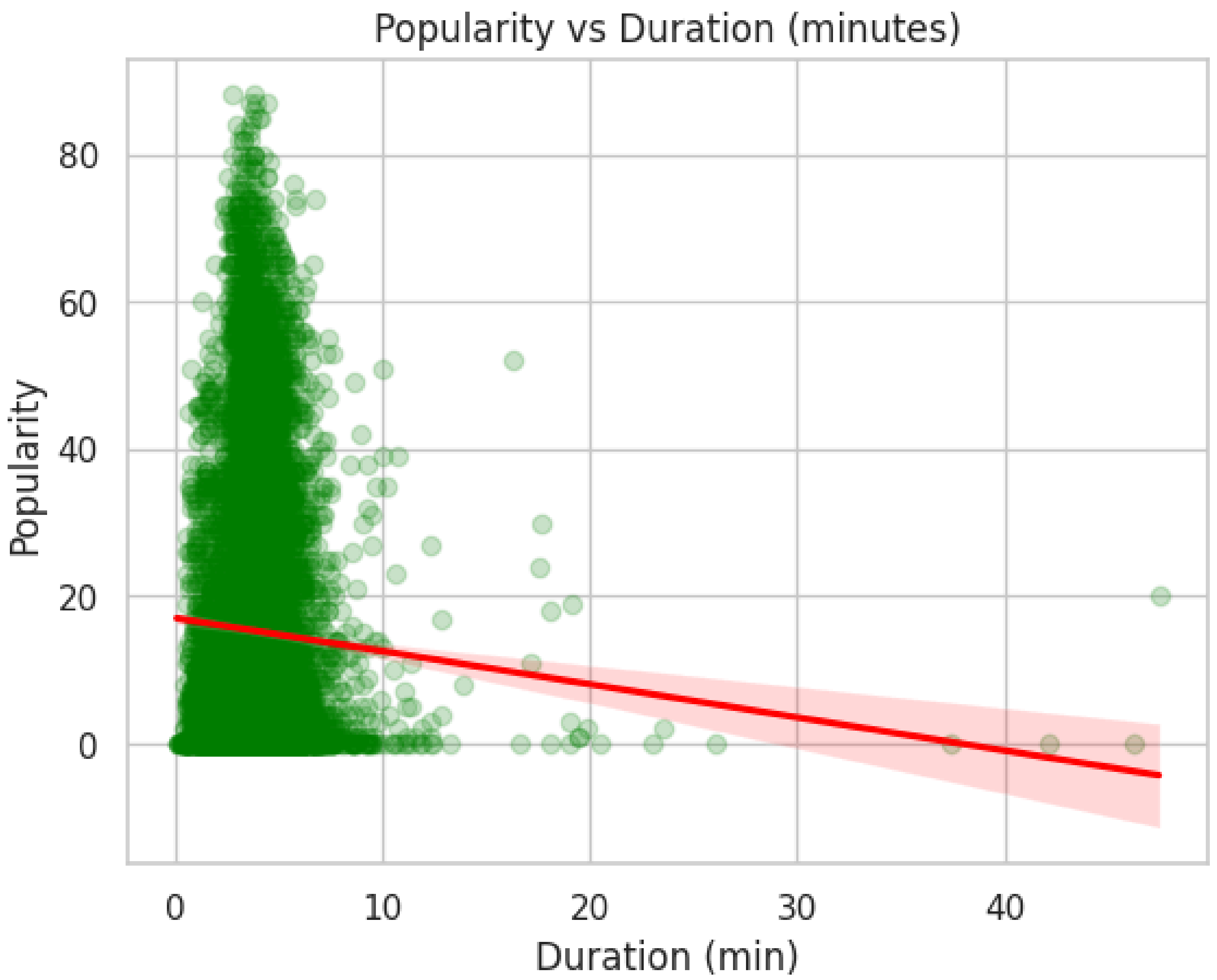
- Popularity: Most tracks cluster in the 40–70 range, indicating moderate popularity.
- Duration: The average song duration is approximately 3.5 minutes.
- Audio Features: A majority of the songs in the dataset are characterized by high danceability and energy.
- Musical Structure: The most common time signature is 4/4, and the majority of tracks are in a major key.

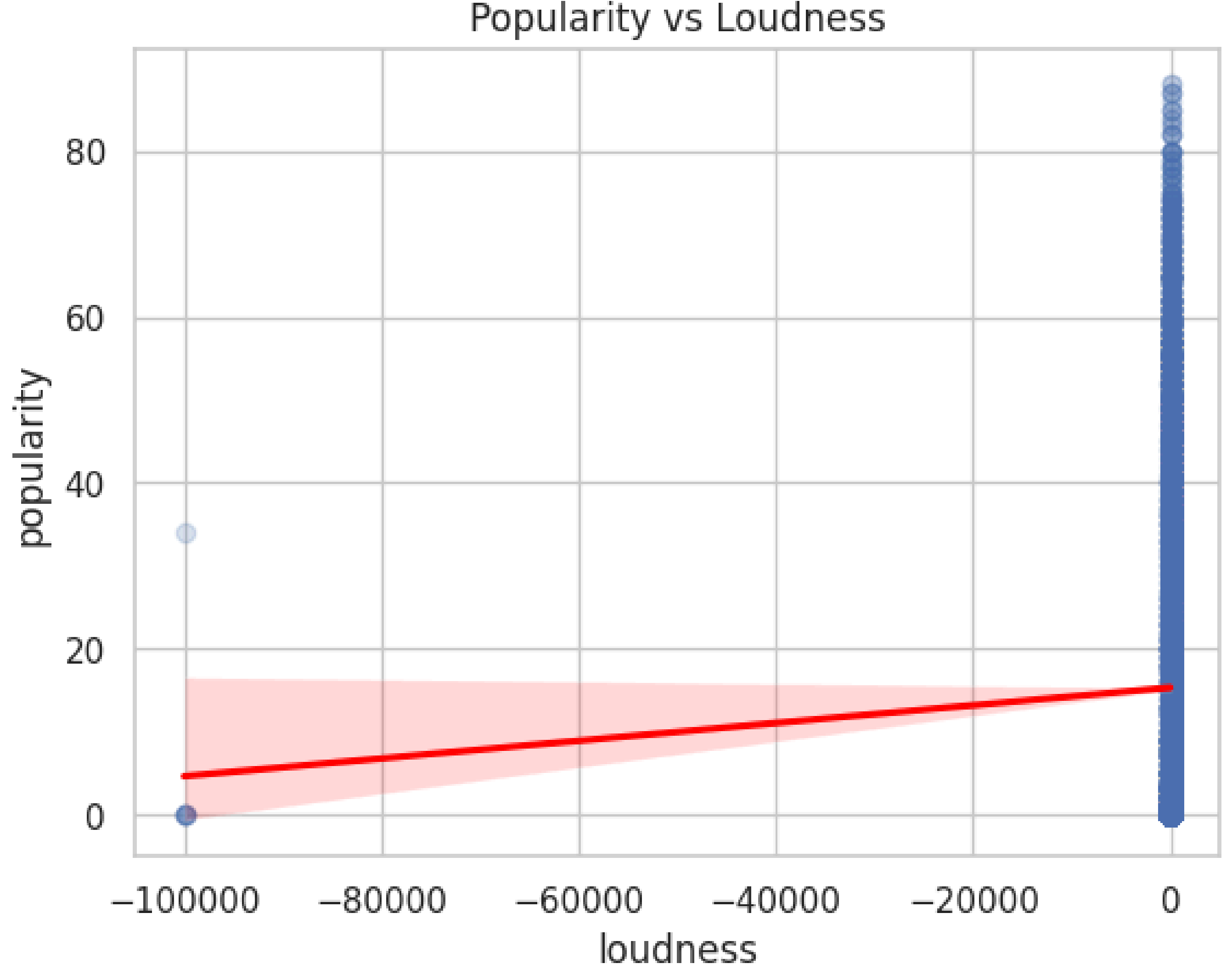
Bivariate Analysis(Numerical vs. Numerical

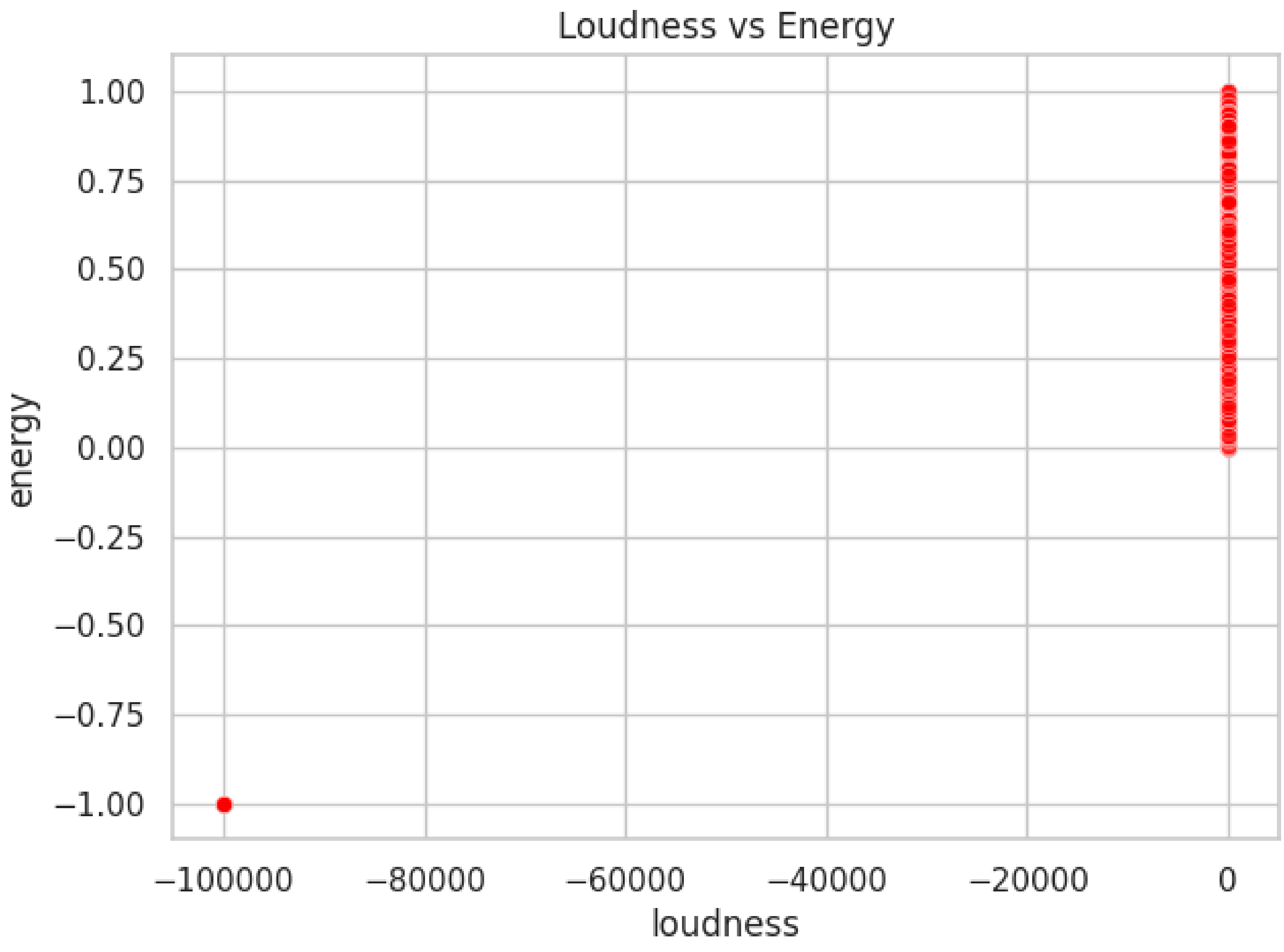


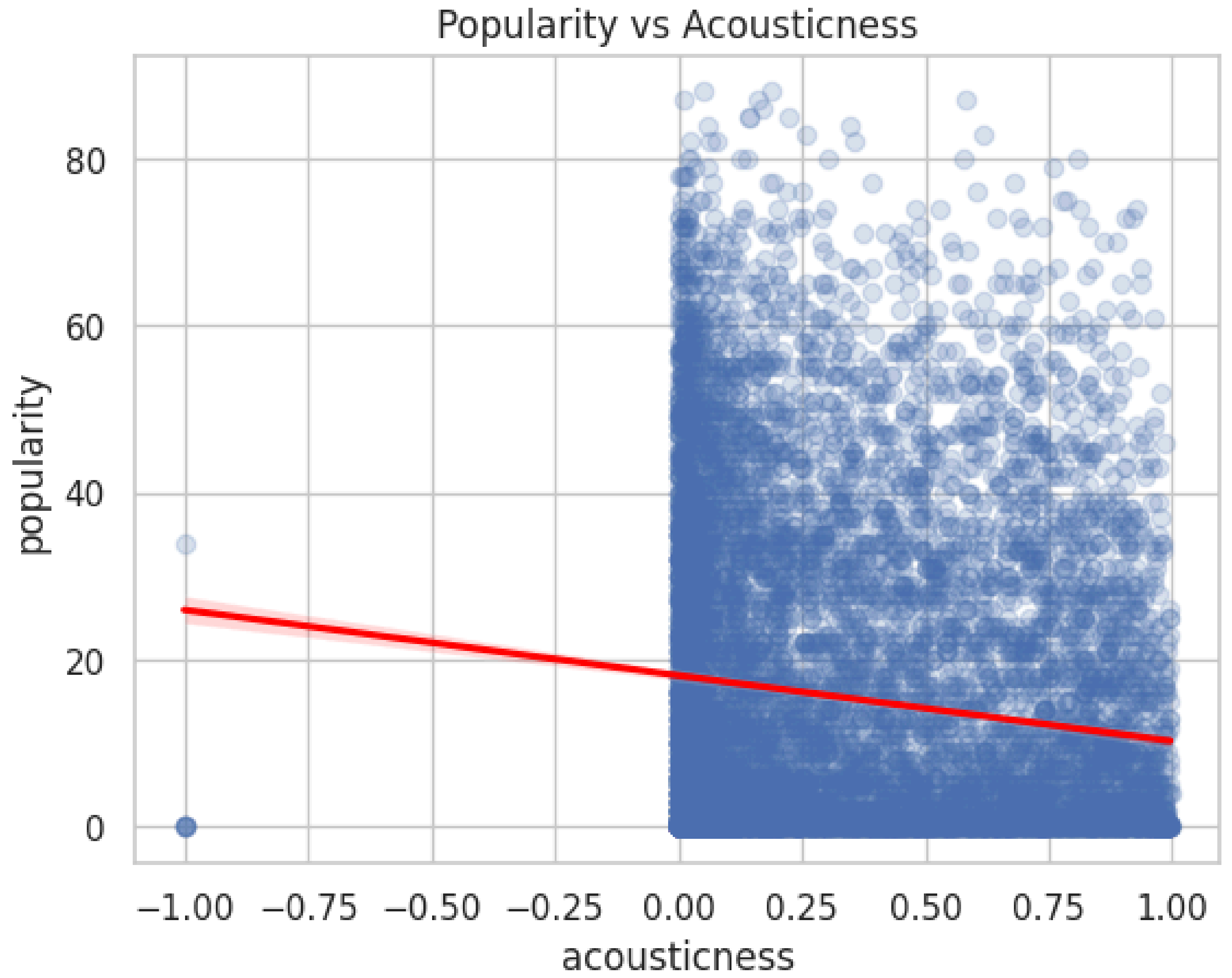


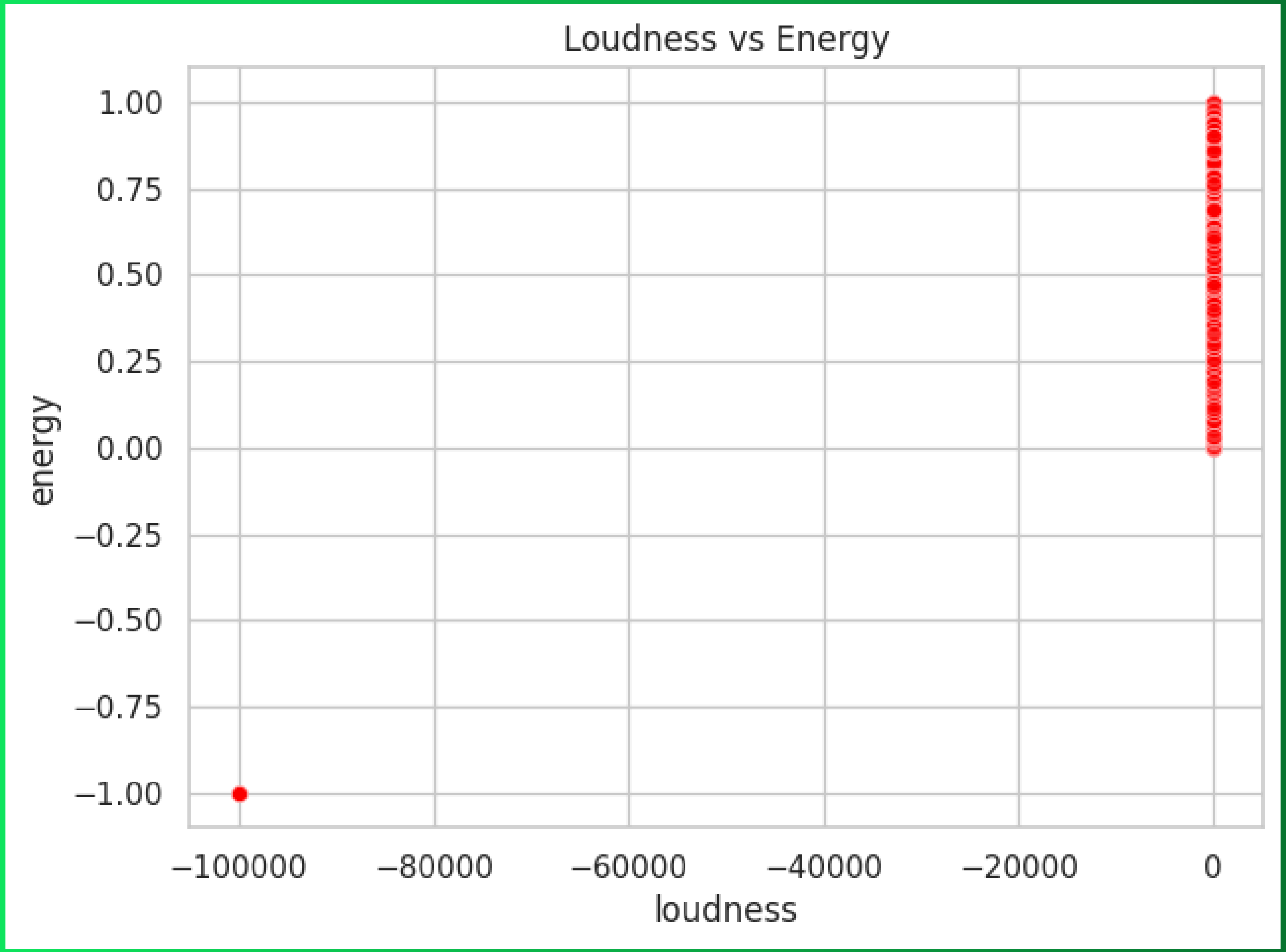
Both show a slight positive correlation. High danceability and energy help a song's popularity, but they are not the sole drivers. A negative trend is observed in Popularity vs. AcousticnessThe market favors produced, modern sounds..

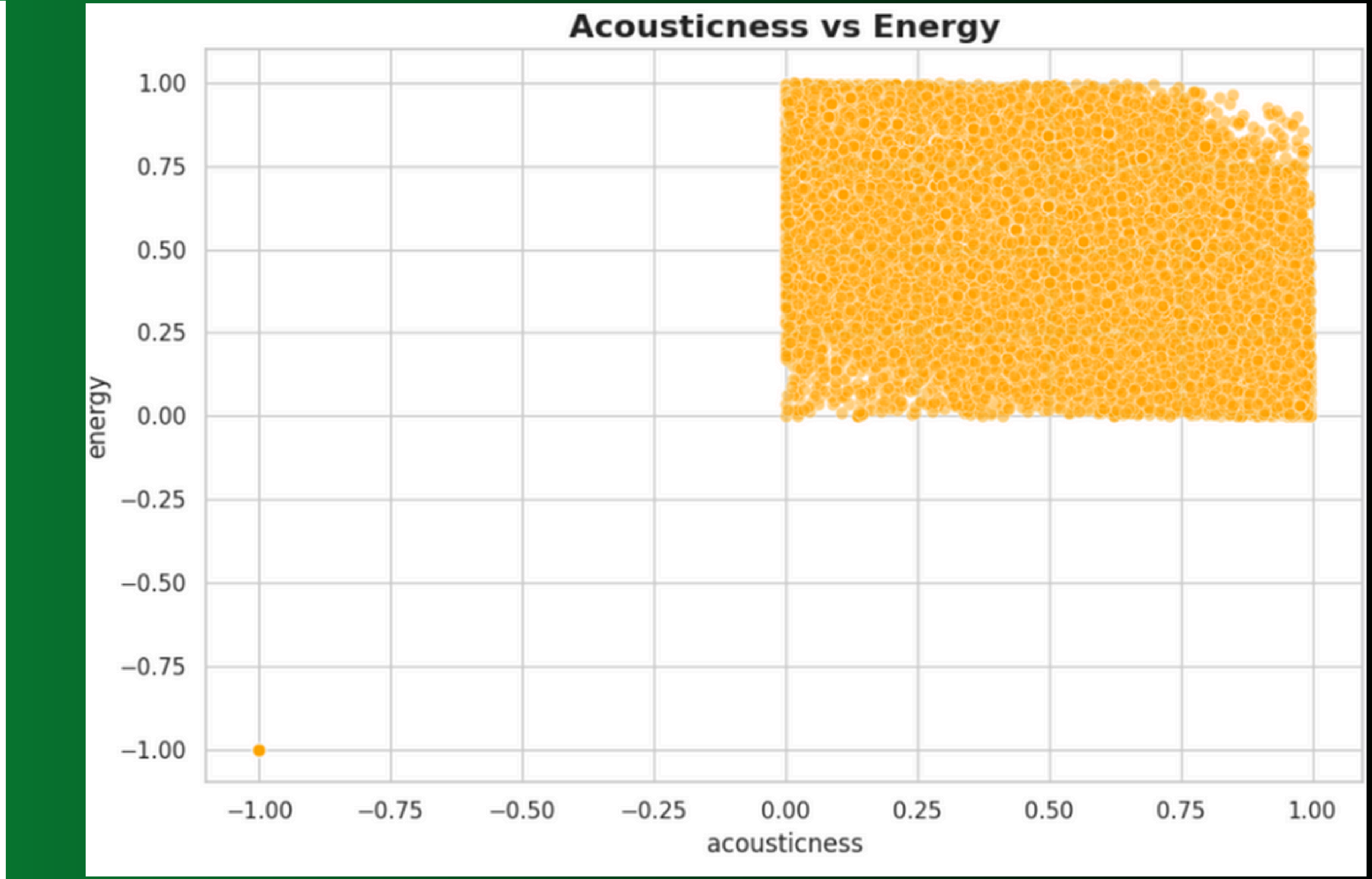

Popularity vs Duration (minutes)
Popularity
Duration (min)
0
20
40
60
80
0
10
20
30
40


Popularity vs Loudness
popularity
loudness
0
20
40
60
80
−100000
−80000
−60000
−40000
−20000
0


Loudness vs Energy
energy
loudness
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
−100000
−80000
−60000
−40000
−20000
0

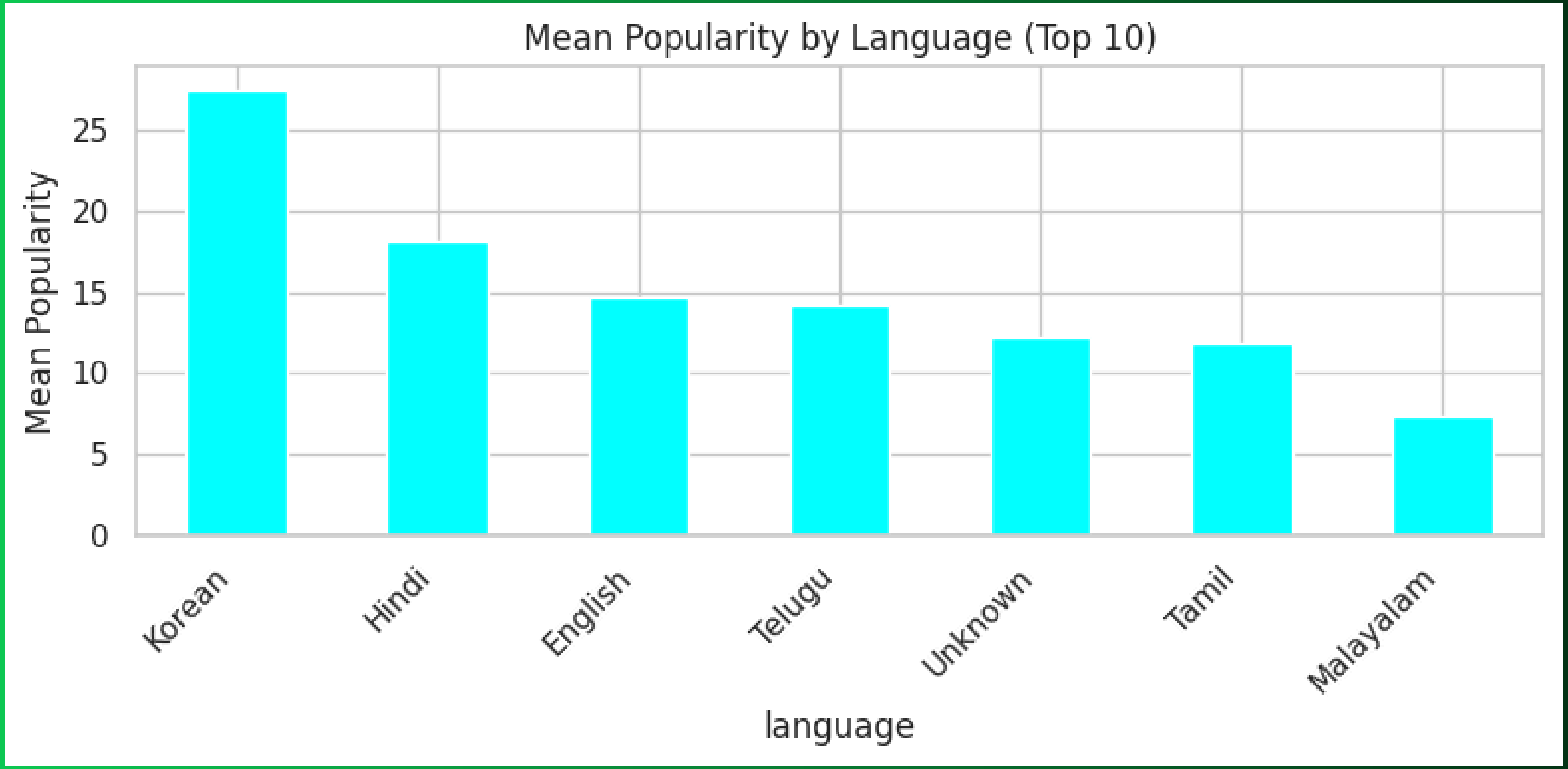

Popularity vs Acousticness
popularity
acousticness
0
20
40
60
80
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00

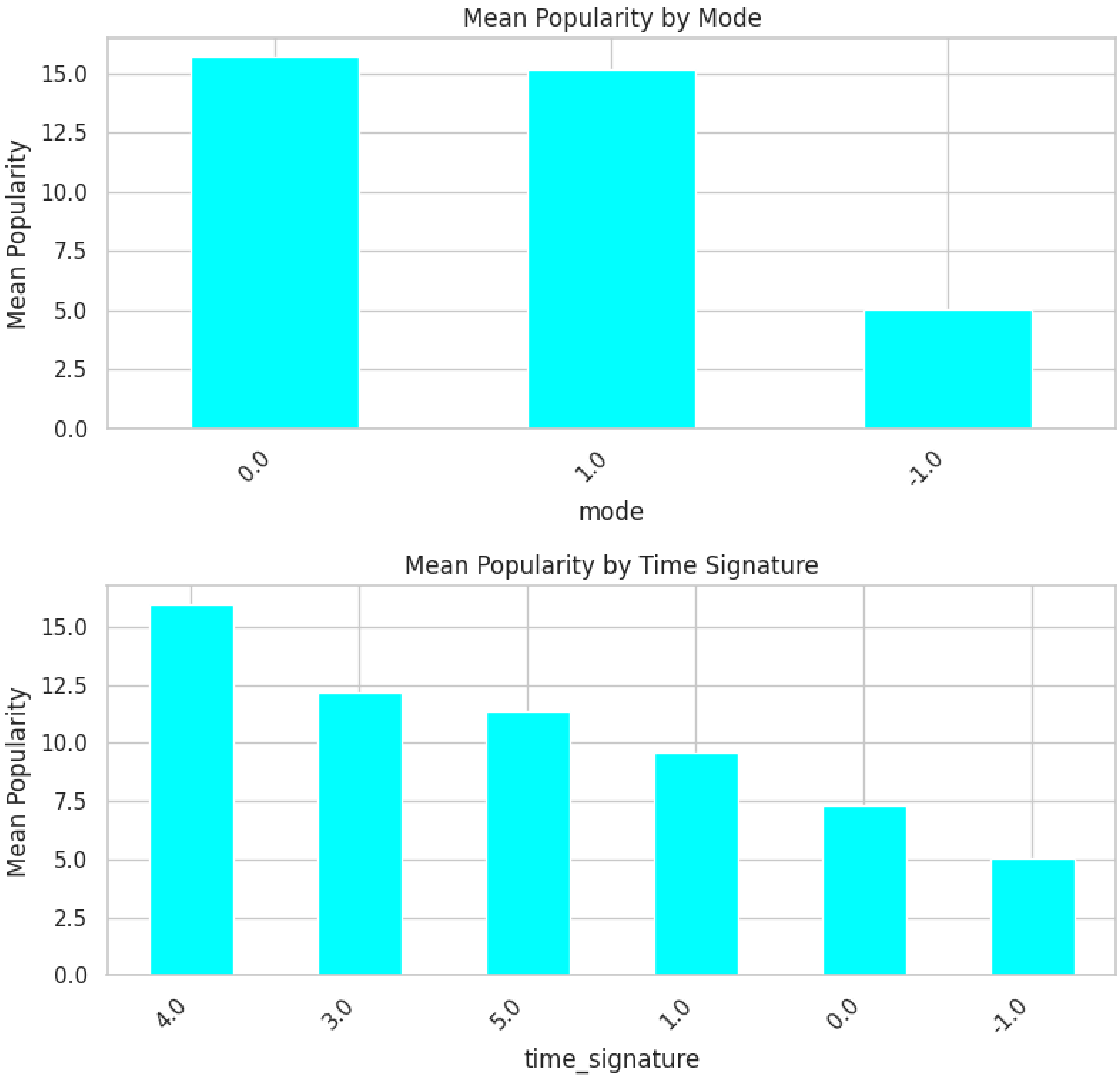




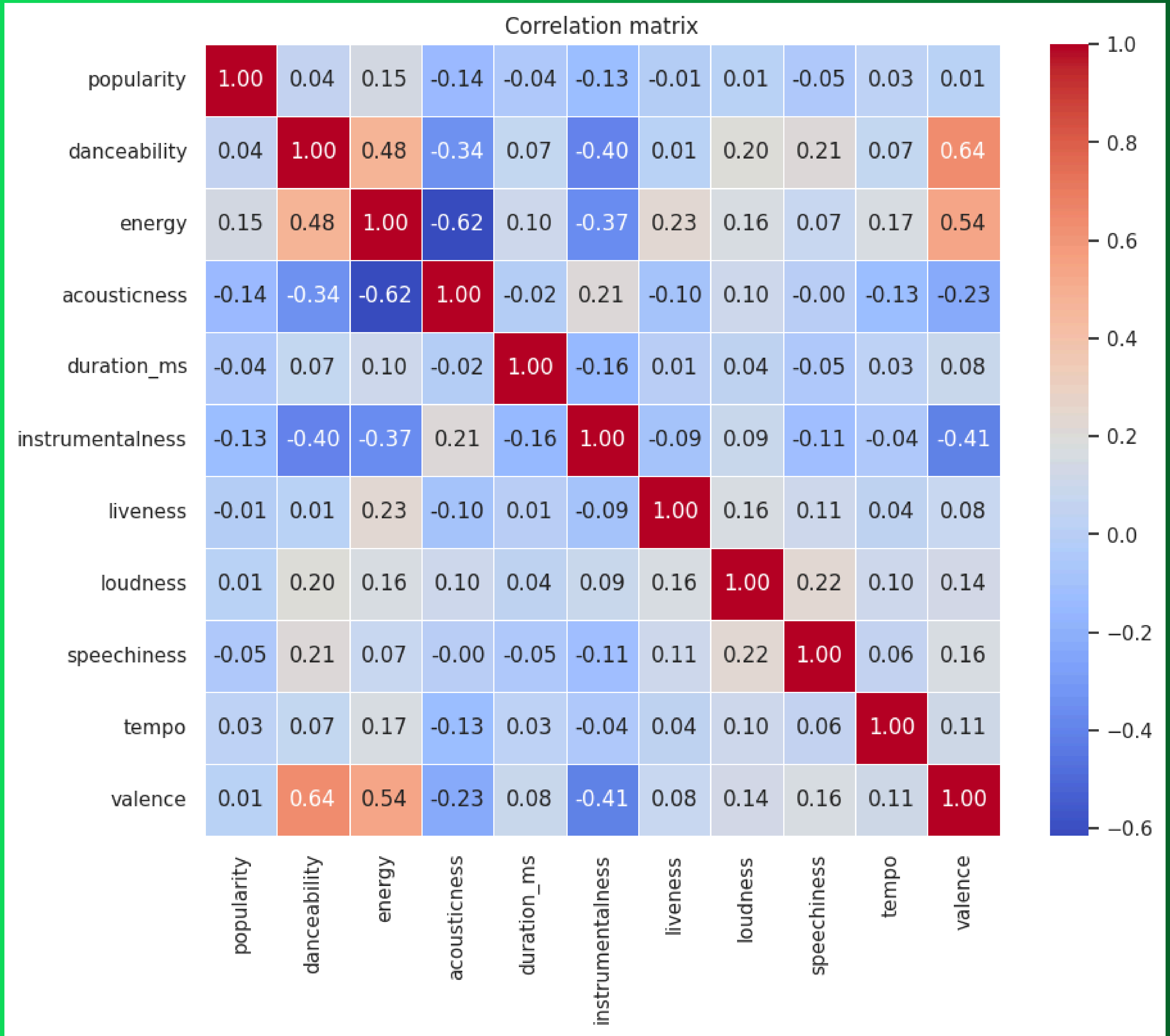
There is a very strong positive correlation.. High-energy tracks tend to be more positive/happy (high valence), while low-energy tracks are often associated with lower (sad or calm) valence.

Bivariate Analysis(Numerical vs. Numerical)




Mean Popularity by Mode
Mean Popularity
15.0
12.5
10.0
7.5
5.0
2.5
0.0
0.0
1.0
-1.0
mode
Mean Popularity by Time Signature
Mean Popularity
15.0
12.5
10.0
7.5
5.0
2.5
0.0
4.0
3.0
5.0
1.0
0.0
-1.0
time_signature

Overview: The Correlation Heatmap



Correlation matrix

	popularity	danceability	energy	acousticness	duration_ms	instrumentalness	liveness	loudness	speechiness	tempo	valence
popularity	1.00	0.04	0.15	-0.14	-0.04	-0.13	-0.01	0.01	-0.05	0.03	0.01
danceability	0.04	1.00	0.48	-0.34	0.07	-0.40	0.01	0.20	0.21	0.07	0.64
energy	0.15	0.48	1.00	-0.62	0.10	-0.37	0.23	0.16	0.07	0.17	0.54
acousticness	-0.14	-0.34	-0.62	1.00	-0.02	0.21	-0.10	0.10	-0.00	-0.13	-0.23
duration_ms	-0.04	0.07	0.10	-0.02	1.00	-0.16	0.01	0.04	-0.05	0.03	0.08
instrumentalness	-0.13	-0.40	-0.37	0.21	-0.16	1.00	-0.09	0.09	-0.11	-0.04	-0.41
liveness	-0.01	0.01	0.23	-0.10	0.01	-0.09	1.00	0.16	0.11	0.04	0.08
loudness	0.01	0.20	0.16	0.10	0.04	0.09	0.16	1.00	0.22	0.10	0.14
speechiness	-0.05	0.21	0.07	-0.00	-0.05	-0.11	0.11	0.22	1.00	0.06	0.16
tempo	0.03	0.07	0.17	-0.13	0.03	-0.04	0.04	0.10	0.06	1.00	0.11
valence	0.01	0.64	0.54	-0.23	0.08	-0.41	0.08	0.14	0.16	0.11	1.00

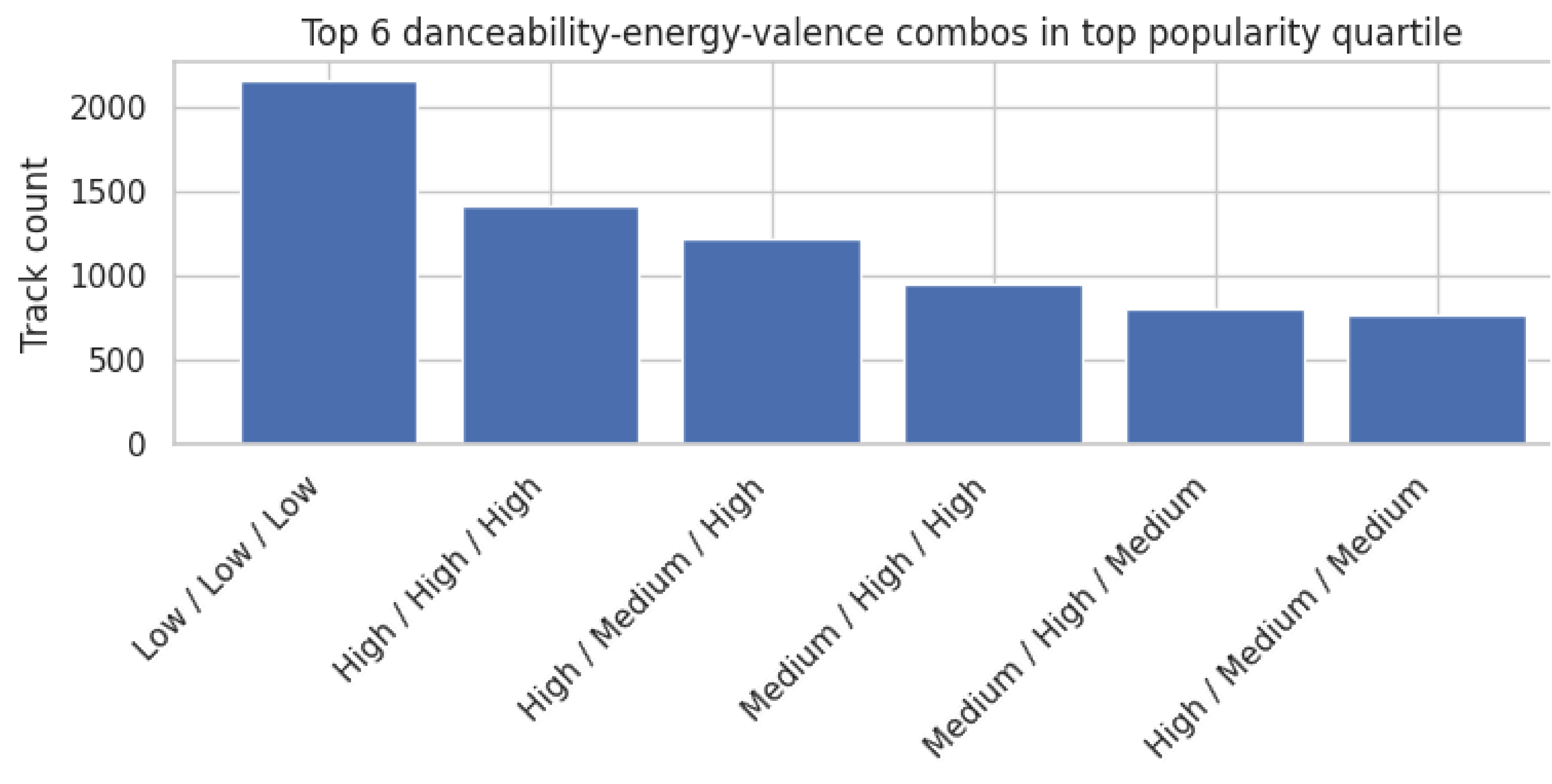
This map quickly shows all relationships. I highlights the strong links (like Energy-Loudness) and confirms the weak-but-important links to Popularity (e.g., slight positive from Danceability).

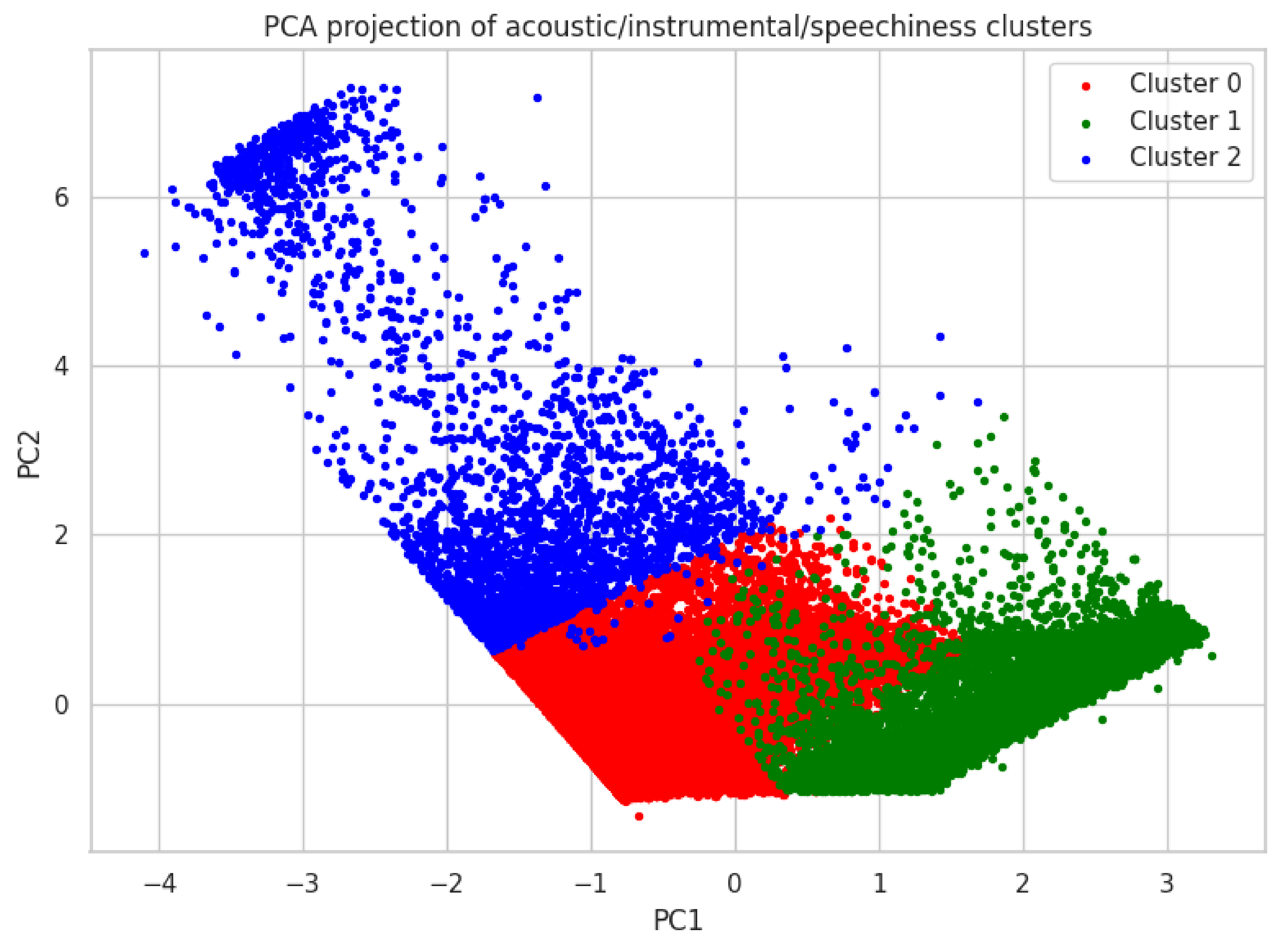
Bivariate & Correlation Analysis

Examining the relationships between features provided deeper insights:

- A strong positive correlation exists between danceability and energy.
- Popularity shows a slight positive correlation with both energy and danceability.
- Loudness and energy are highly correlated, which is an expected relationship.
- A negative trend was observed between acousticness and popularity, suggesting that acoustic songs tend to be less popular on average in this dataset.

Multivariate analysis





Multivariate analysis summary:

1.Feature Synergy (The Sonic Sweet Spot) Popularity is driven by a balanced blend of features (e.g., high Energy, mid-range Valence, high Loudness) in the Top 5%of tracks. Takeaway: Avoid extremes; synergy matters more than maxing out one single audio metric.

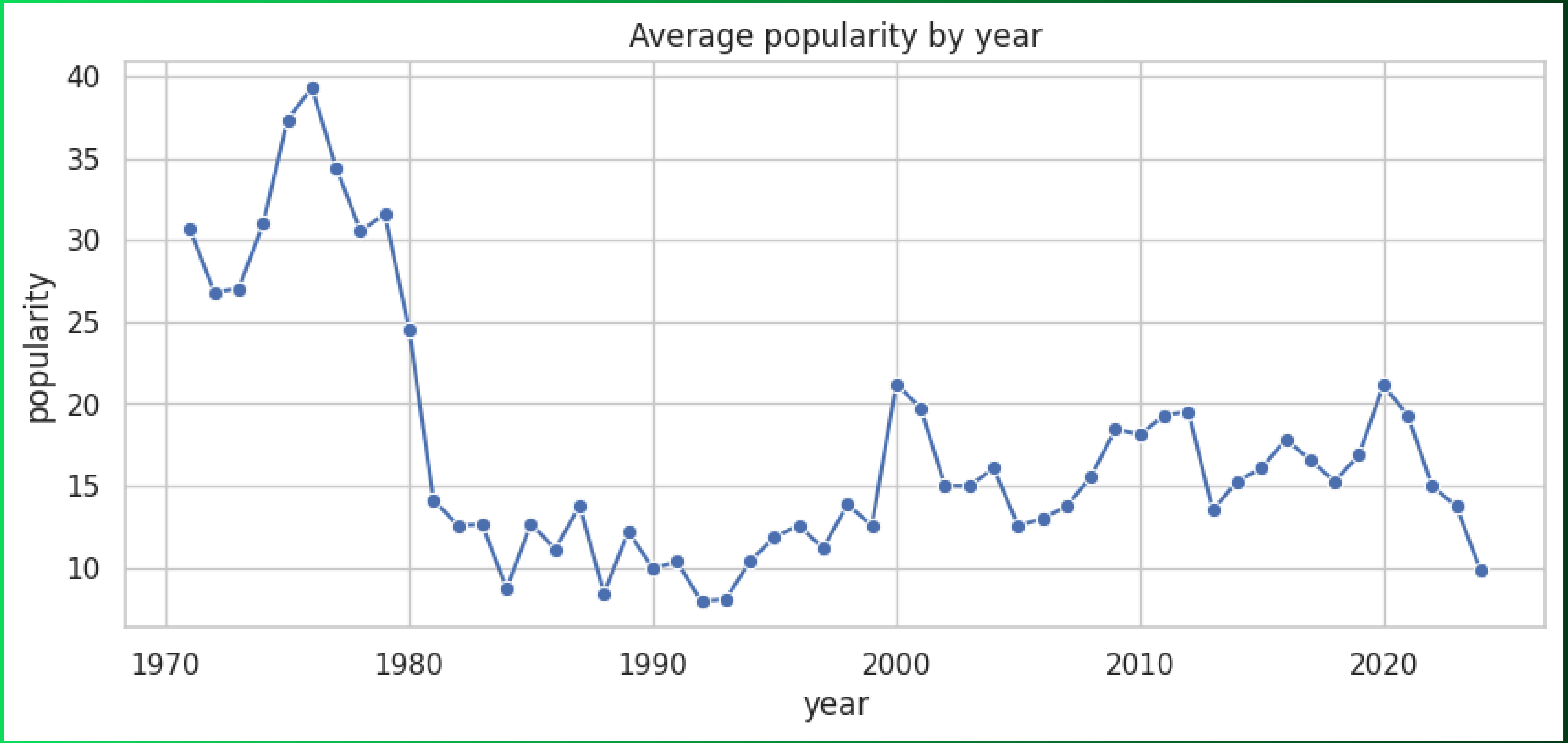
2.The Modern Mix (Efficiency & Impact) : Time Trend 1: Average song duration is decreasing over the years. Time Trend 2: Average song loudness is increasing over the years.

3. Takeaway: Modern hits must be short, concise, and competitively loud for streaming success.

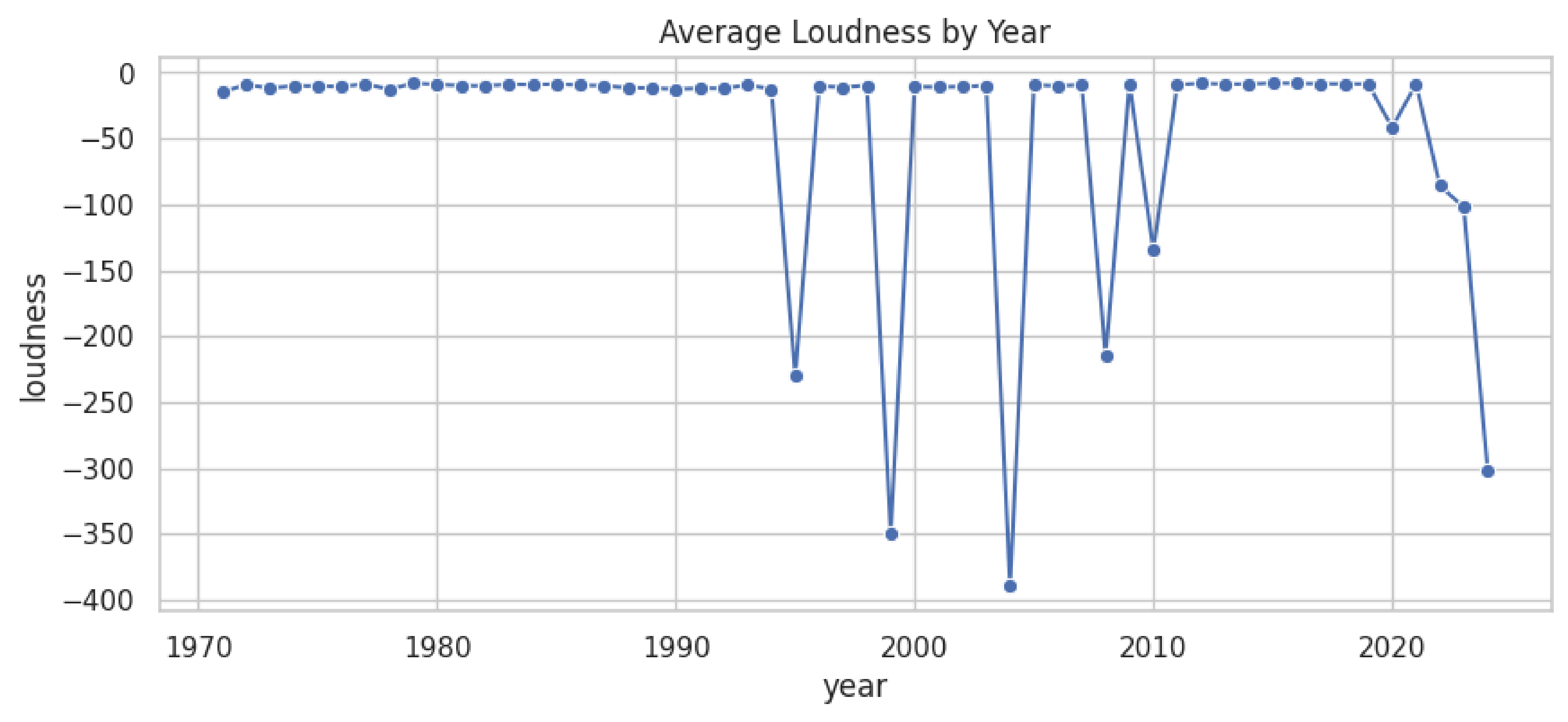
Pillar 3: Global Reach (Market Expansion) Time Trend 3: The popularity of Non-English music (especially Spanish and Korean tracks) is rapidly accelerating.

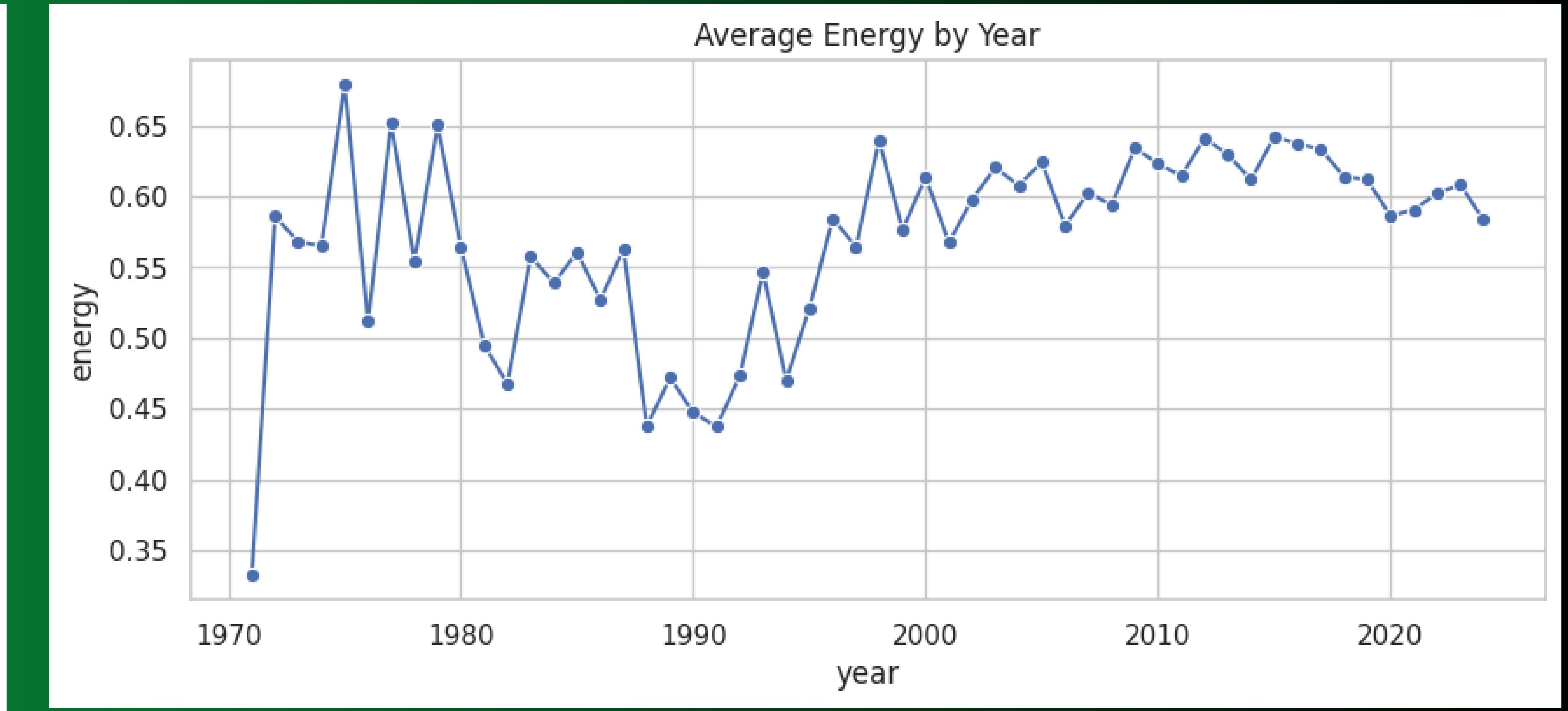
Takeaway: Capitalize on the global market shift by exploring multilingual content and expanding regional focus.

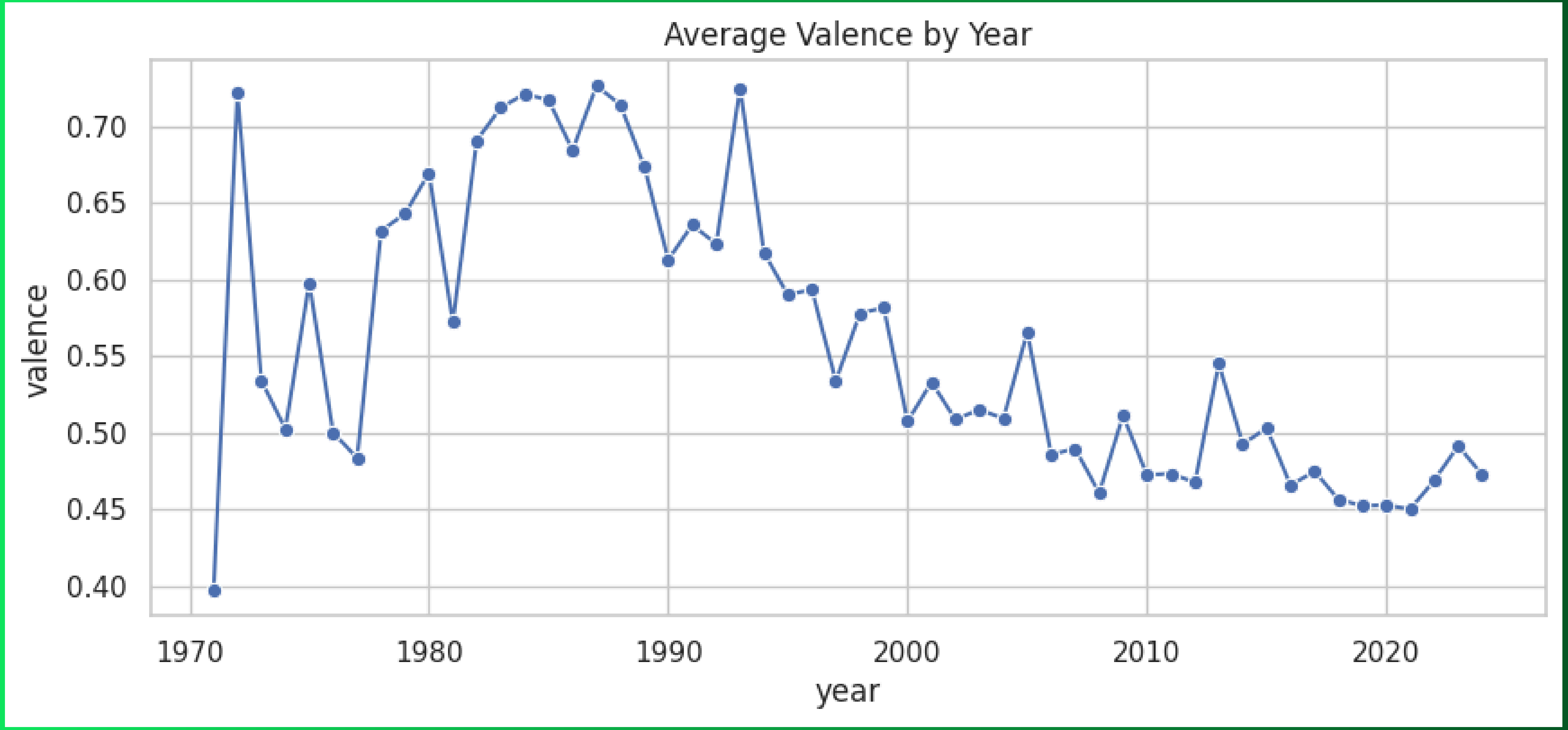
Time Series Trends



Newer songs (streaming era) are consistently more popular.


Average Loudness by Year
loudness
0
−50
−100
−150
−200
−250
−300
−350
−400
year
1970
1980
1990
2000
2010
2020

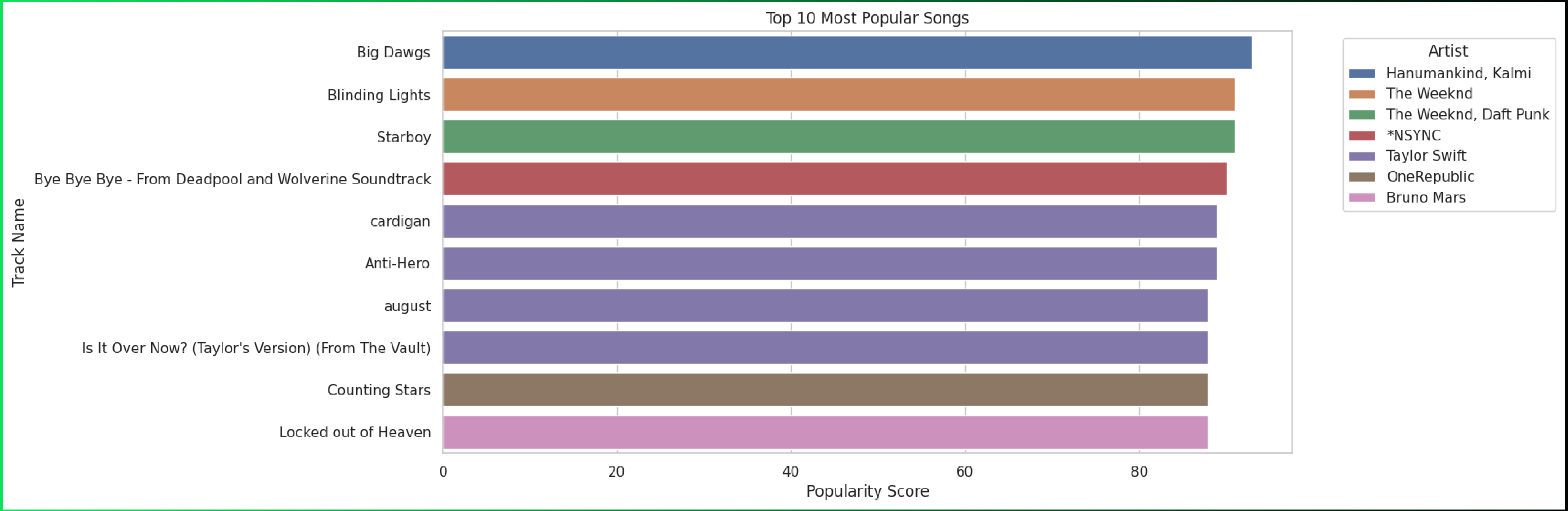

Average Energy by Year
energy
0.65
0.60
0.55
0.50
0.45
0.40
0.35
year
1970
1980
1990
2000
2010
2020


Average Valence by Year
valence
0.70
0.65
0.60
0.55
0.50
0.45
0.40
year
1970
1980
1990
2000
2010
2020

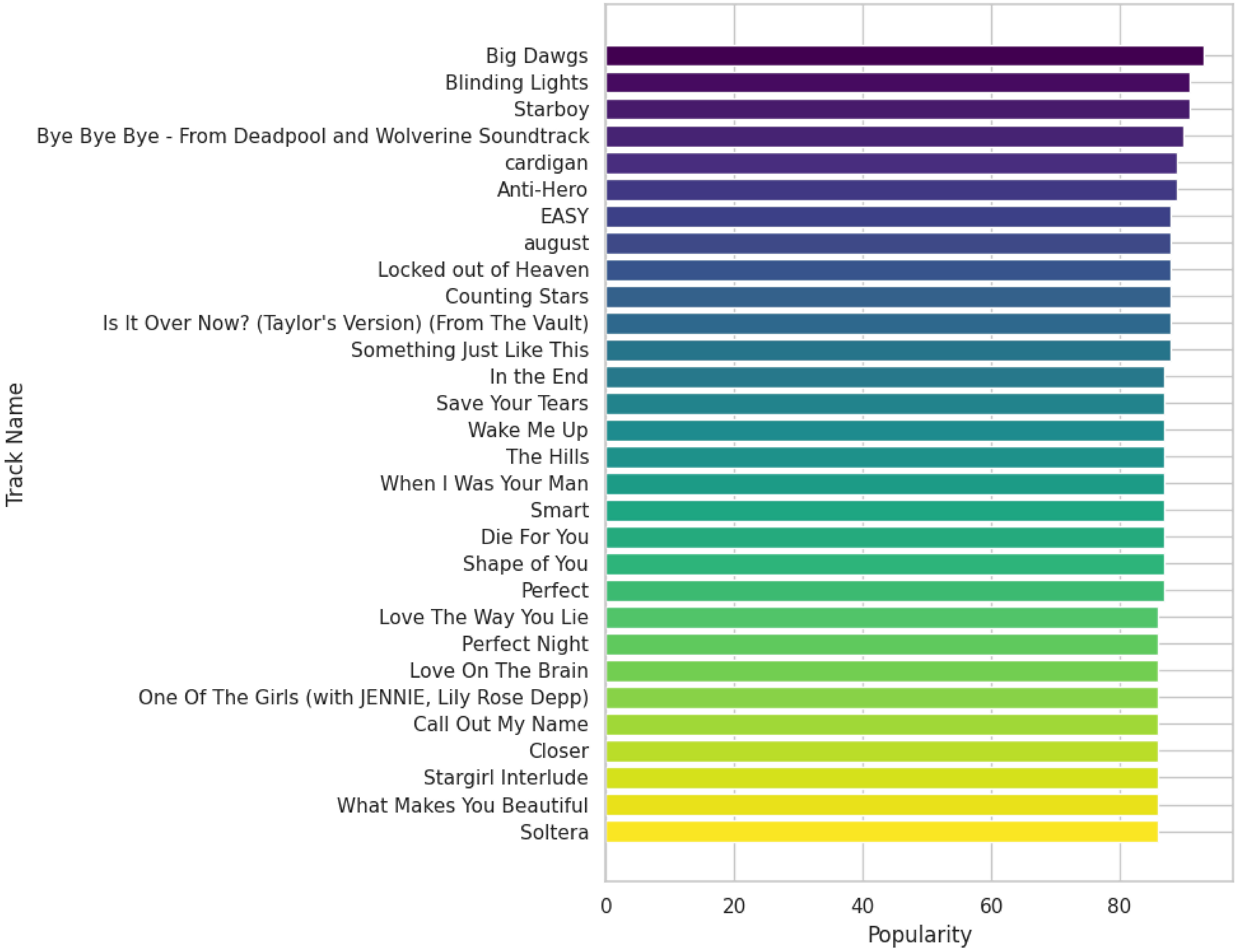
summary of Time Series Trends

1. The Modern Mix is Shorter and Louder
   - Trend: The average song duration has decreased, while average loudness has increased consistently over the years.
   - Insight: Modern hit songs are designed for maximum impact and efficiency. The industry favors tracks that are loud and concise to hold listener attention on streaming platforms.
   - Visuals: Side-by-Side Line Plots of duration vs. year and loudness vs. year.
2. The Market is Going Global
   - Trend: The average popularity of tracks in languages like Spanish and Korean has seen a rapid, significant rise in recent years.
   - Insight: The music market is truly globalizing. Non-English music is no longer niche and represents a major growth area.
   - Visuals: Line Plot of popularity of Spanish/Korean tracks vs. year

Outlier Analysis



□ Top 1% Most Popular Spotify Songs (Color Gradient)



Outlier Analysis Result:

- Focus: The analysis centers on niche tracks defined by extreme scores on Speechiness or Instrumentalness.
- Core Principle (Pillar 4): Success for these tracks depends on Niche Clarity—they must fully commit to their segment.
- Speechiness Outliers (High Lyrical Niche):
- The market is strictly divided; there's little room for medium-level speech.
- Tracks must be either primarily sung or primarily lyrical/spoken.
- Instrumentalness Outliers (Pure Instrumental Niche):
- Instrumental tracks are rare in the popular dataset.
- To succeed, they must compensate by maximizing secondary features like Energy and Valence to create a strong, clear emotional impact.
- Takeaway: Outliers must specialize and optimize their remaining features instead of trying to hit the "average pop song" sweet spot.

Thank you