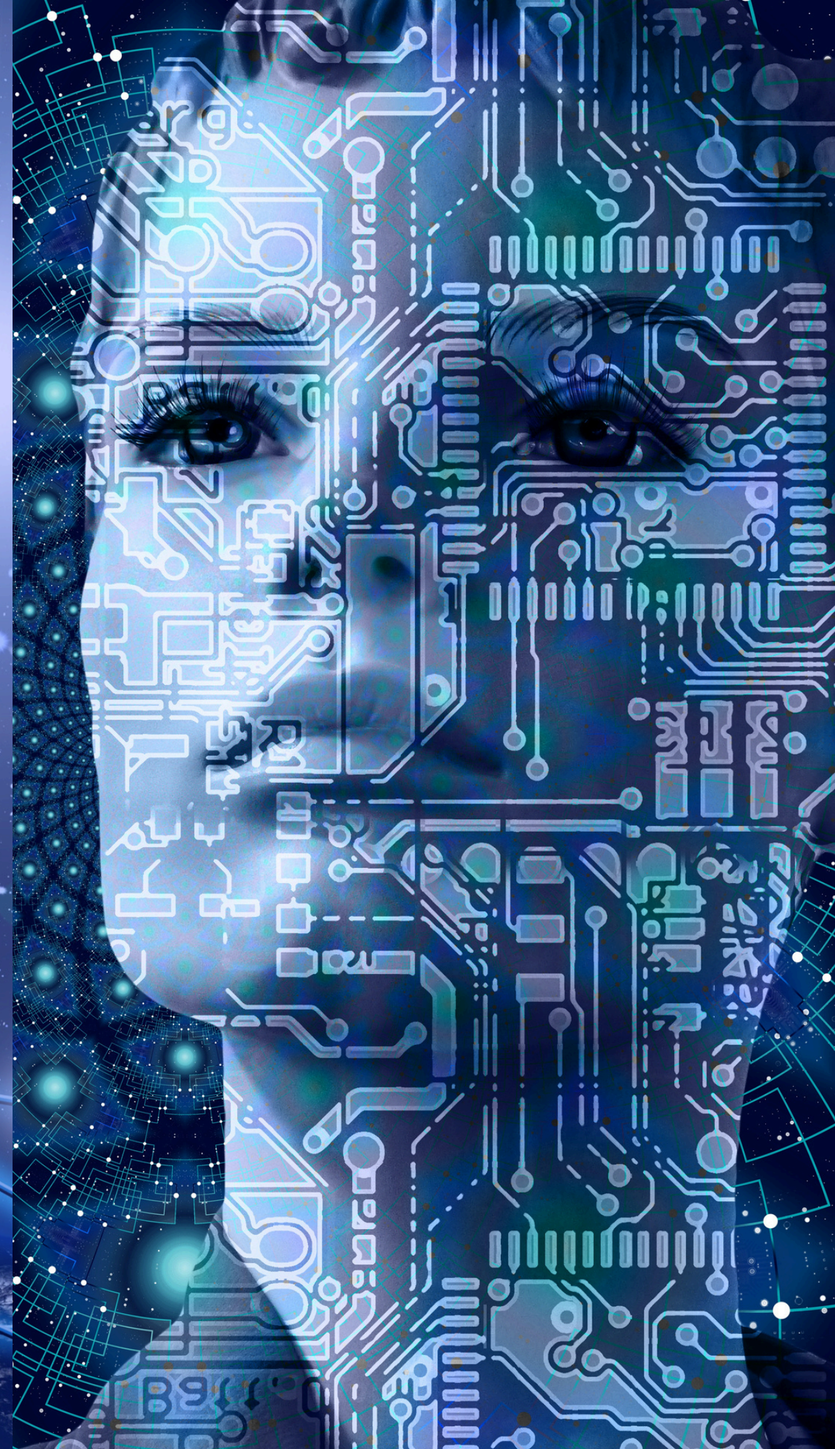
SPOTIFY DATASET EXPLORATORY DATA ANALYSIS

Uncovering patterns, trends, and insights hidden within millions of tracks to understand what makes music resonate with listeners worldwide



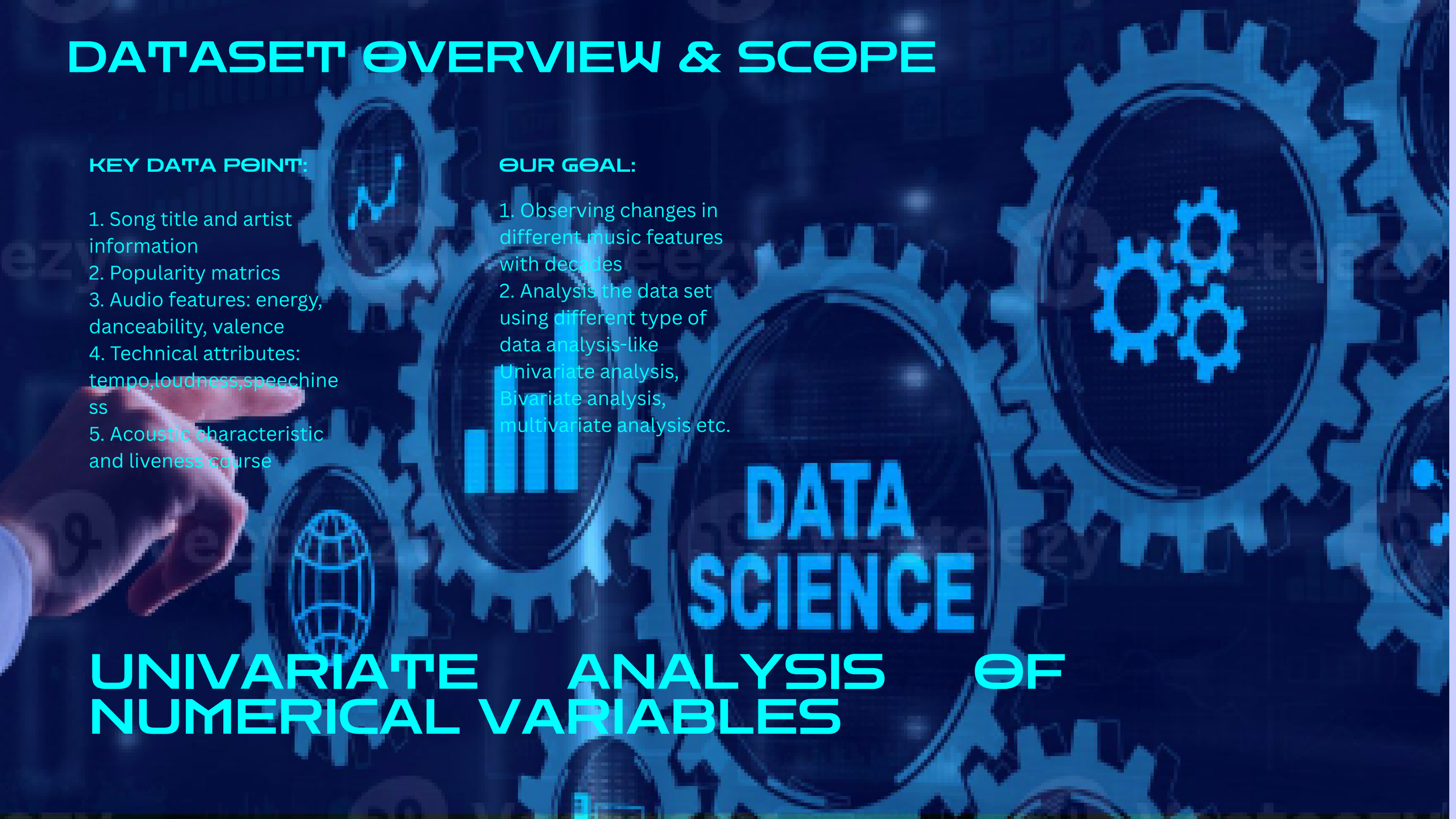
DATASET OVERVIEW & SCOPE

KEY DATA POINT:

1. Song title and artist information
2. Popularity matrics
3. Audio features: energy, danceability, valence
4. Technical attributes: tempo,loudness,speechiness
5. Acoustic characteristic and liveness course

OUR GOAL:

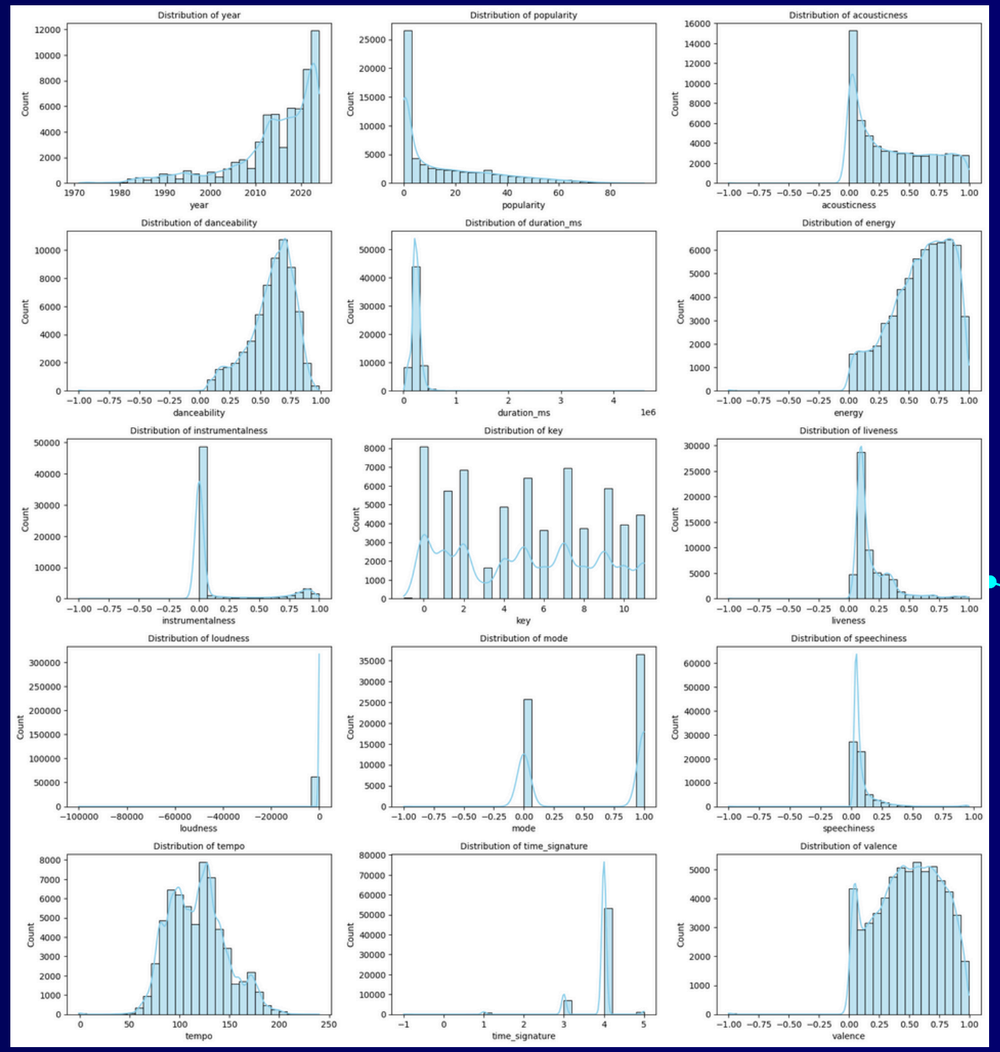
1. Observing changes in different music features with decades
2. Analysis the data set using different type of data analysis-like Univariate analysis, Bivariate analysis, multivariate analysis etc.



UNIVARIATE ANALYSIS OF NUMERICAL VARIABLES

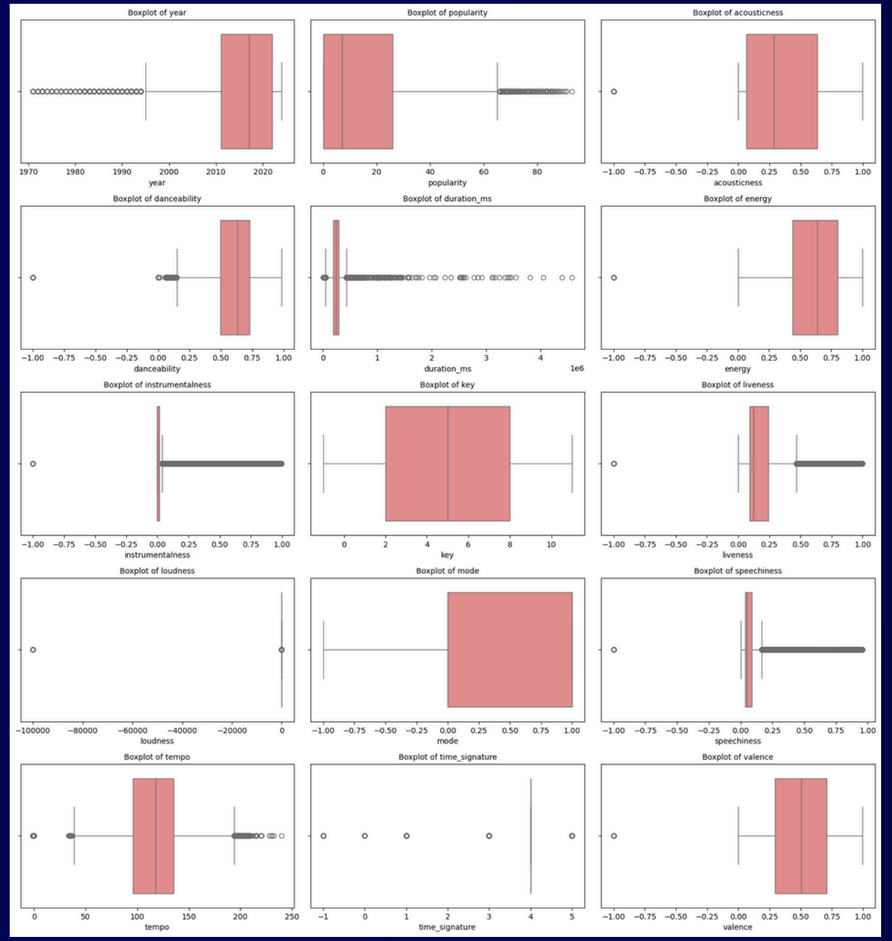
Visualize the distribution of numerical features

- Most songs are new: The dataset is dominated by music from 2010-2020.
- Most songs are not hits: A very large number of songs have a "popularity" score of 0.
- Most songs have vocals: They are not instrumental.
- Most are studio recordings: They are not live performances.
- Music, not talk: The tracks are songs, not spoken word or podcasts.
- High energy & danceable: Songs generally have high energy and are easy toC dance to.
- Standard song structure: Most songs are in 4/4 time, in a major key, and have a tempo around 120-130 BPM.

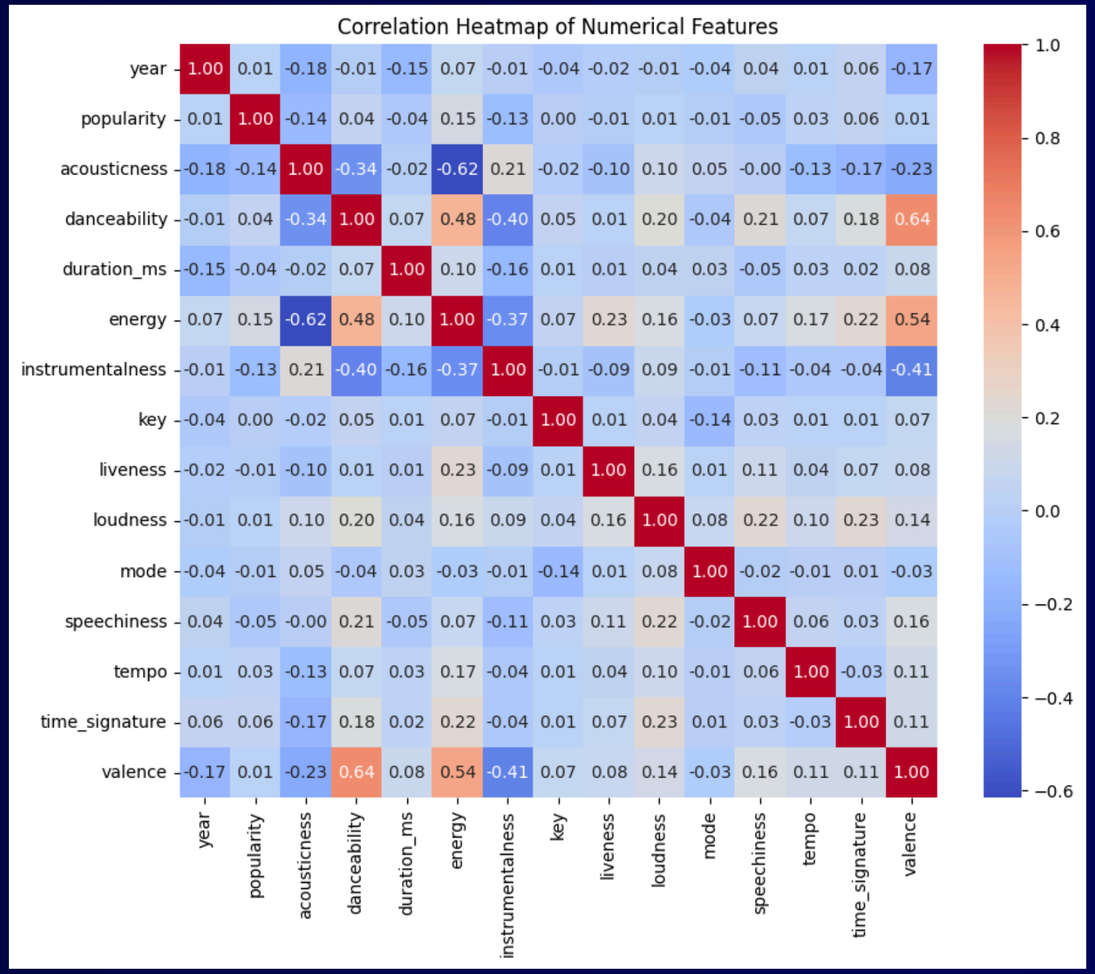


Boxplots for detecting outliers

- Year: The main "box" is from 2010-2020, confirming most songs are new. All songs from before 2005 are considered "outliers."
- Popularity: The box is tiny and near 0. This means most songs have very low popularity; any popular song is an outlier.
- Time Signature: The box is just a flat line at 4. This strongly confirms that almost every song is in 4/4 time.
- Instrumentalness, Liveness, Speechiness: These boxes are all flat lines near 0. This confirms the dataset is almost entirely music with vocals, recorded in a studio, not live talks or podcasts.
- Duration: The main box is small, showing most songs have a standard length (around 3-4 minutes). The many circles ($\circ$) show there are a lot of unusually long songs (outliers).
- Energy & Danceability: The boxes are high on the charts, confirming that songs generally have high energy and are danceable.

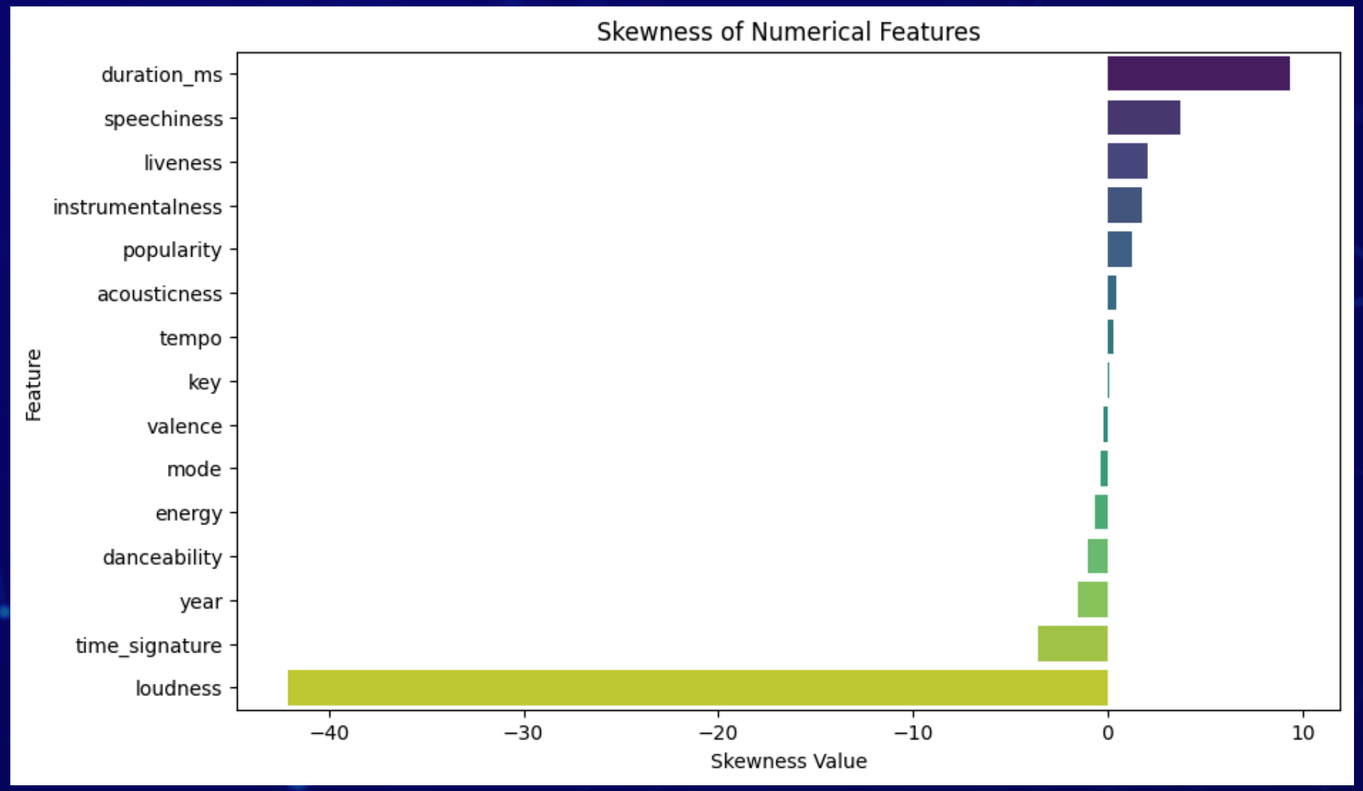


Correlation heatmap among numerical variables



- Loudness & Energy (0.70): This is the strongest link. Louder songs are almost always more energetic.
- Danceability & Valence (0.64): Songs that are easy to dance to also tend to sound happy (valence).
- Energy & Acousticness (−0.62): This is the strongest negative link. Energetic songs are NOT acoustic. (e.g., electronic music is high energy, acoustic guitar is low energy).
- Energy & Valence (0.54): Energetic songs also tend to sound happy.
- Year (all near 0): The year a song was released has almost no connection to its energy, loudness, or danceability.

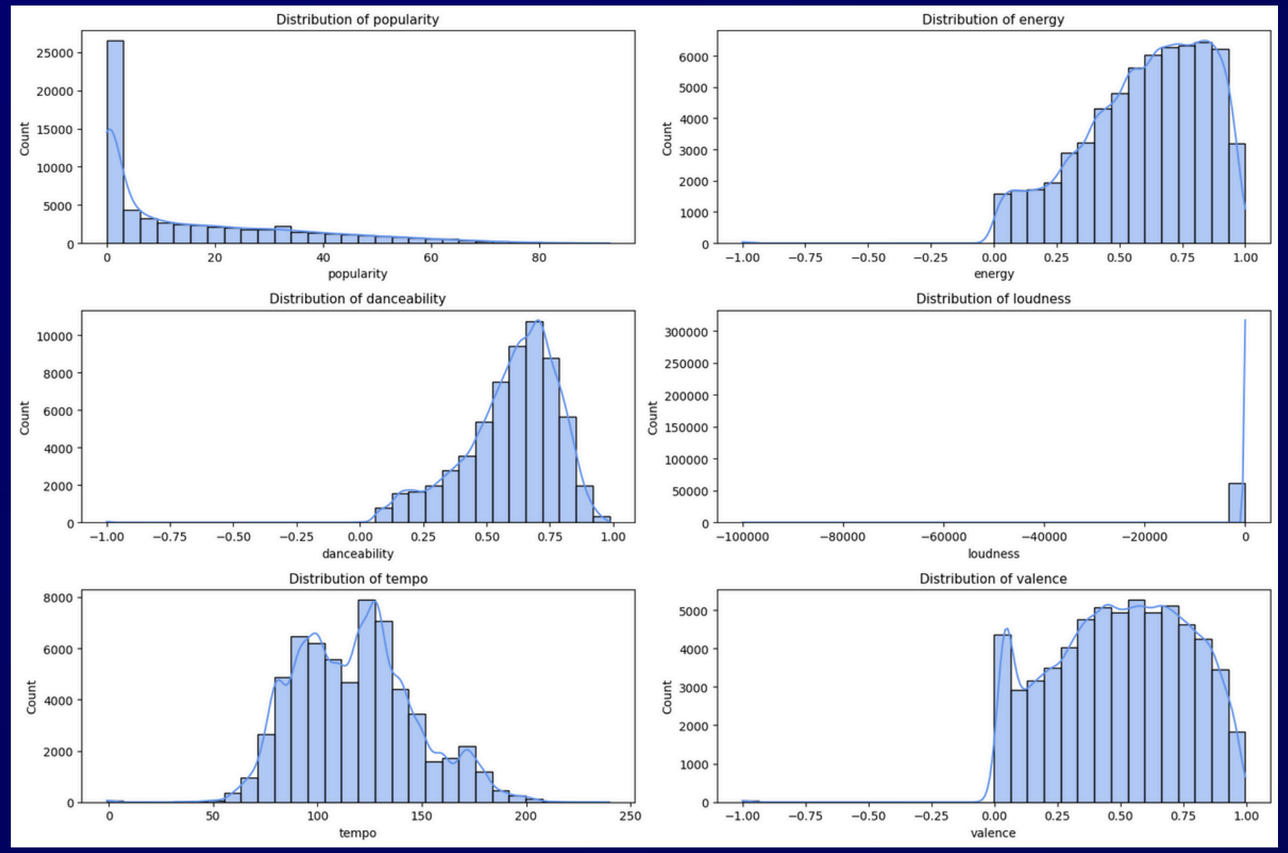
Check skewness to identify non-normal distributions



- speechiness, liveness, and instrumentalness are also moderately right-skewed. This suggests that for these features, most songs have low values, with a smaller number of songs having very high values (e.g., most songs have low speechiness, but a few are very talk-heavy).
- Approximately Symmetrical (Near Zero Skew):
- Many features like popularity, acousticness, tempo, key, valence, energy, and danceability have skewness values very close to 0. This means their data is distributed fairly symmetrically, like a bell curve, without a strong pull from outliers in either direction.

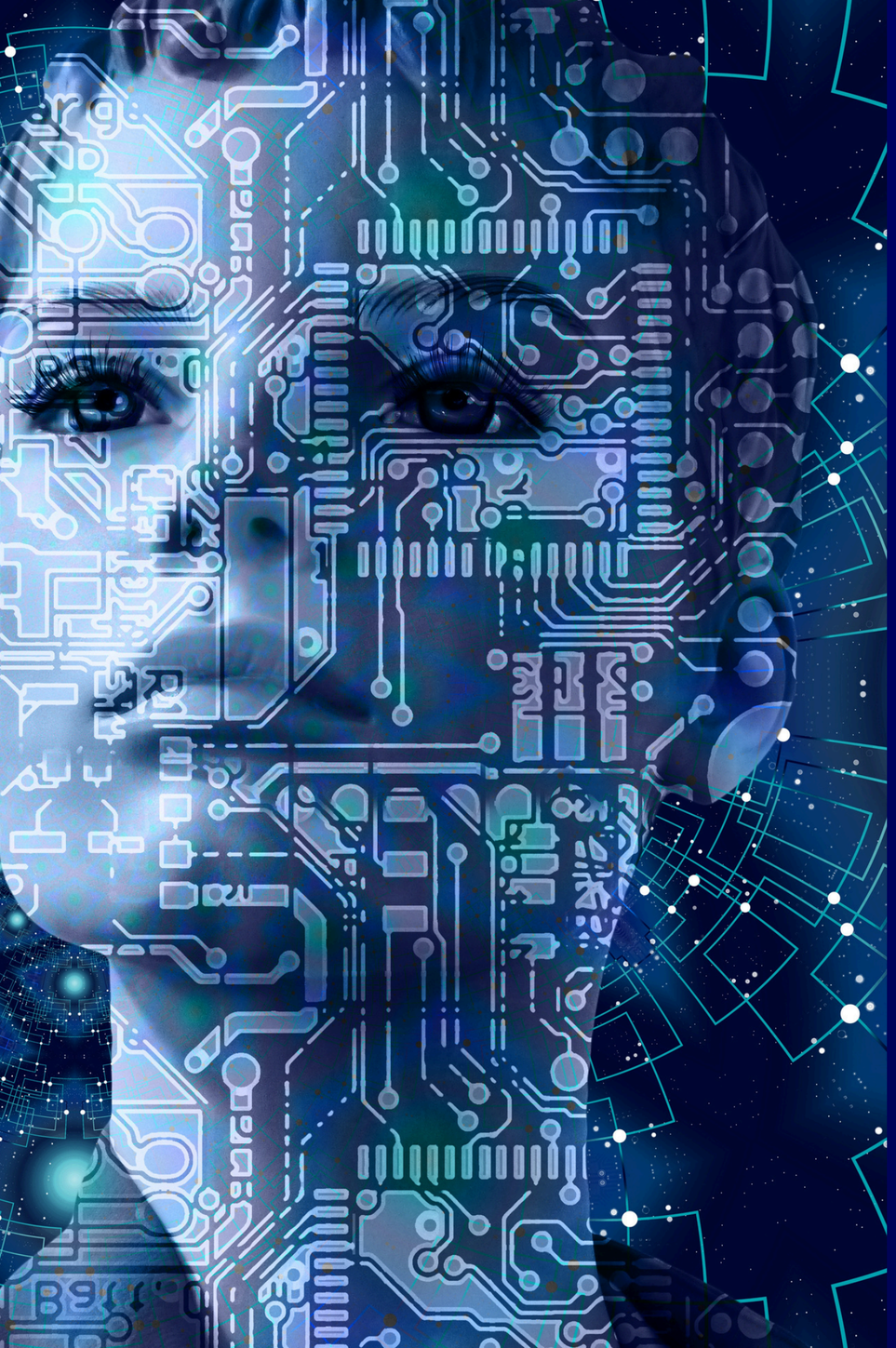
- Highly Left-Skewed (Negative Skew):
- loudness is by far the most skewed feature, with a very large negative value (around -40). This means that while most songs have a loudness value clustered at the higher (louder) end, there are a few significant outliers that are much quieter than the rest, pulling the average down.
- Highly Right-Skewed (Positive Skew):
- duration_ms (song duration in milliseconds) is the most positively skewed feature (around +9). This indicates that most songs in the dataset have a relatively similar, shorter duration, but there are some outlier songs that are exceptionally long

Focus on key metrics like popularity, energy, danceability, loudness, tempo



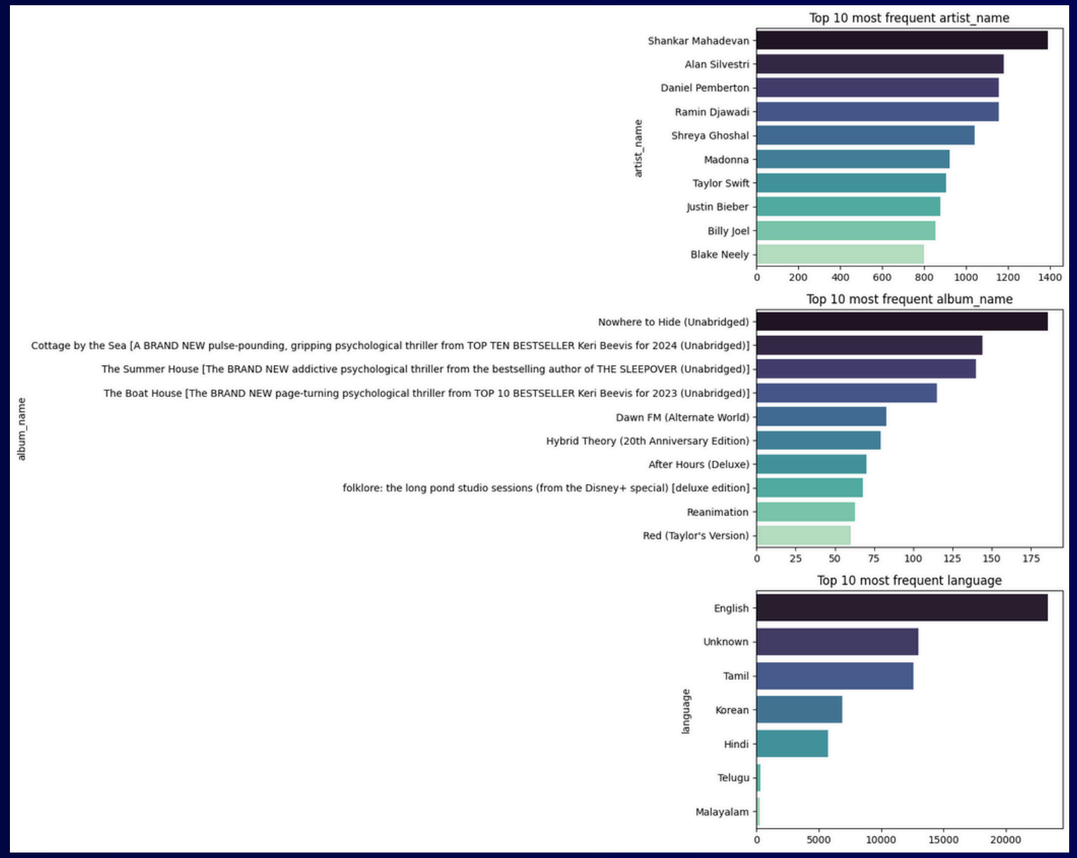
- popularity (Top-Left): This is highly right-skewed. A massive number of songs have a popularity of 0, and the count drops off very quickly. This means most songs in the dataset are not popular at all, with only a few outliers being very popular.
- energy (Top-Right): This is left-skewed. Most songs are clustered on the high-energy side (from 0.4 to 1.0). There are very few songs with low energy.
- danceability (Mid-Left): This distribution looks somewhat symmetrical or slightly left-skewed. Most songs have a danceability score between 0.5 and 0.8, tapering off on either side.
- loudness (Mid-Right): This is a perfect example of an extremely left-skewed distribution. Almost all songs are in a single, massive cluster at the "loud" end (near 0 dB). The long, thin tail to the left represents a few very quiet outlier tracks. This perfectly explains the huge negative skew value from the first chart.
- cluster is slightly on the positive side (peaking around 0.5-0.7)

- tempo (Bottom-Left): This is the most unique chart. It is bimodal, meaning it has two distinct peaks. This suggests the dataset has two common groups of songs: one group with a tempo around 90-100 BPM (Beats Per Minute) and another group around 120-130 BPM.
- valence (Bottom-Right): This distribution (representing musical "positiveness") is fairly symmetrical and spread out, though the main



UNIVARIATE ANALYSIS OF CATEGORICAL VARIABLES

FOCUS ON KEY CATEGORICAL COLUMNS



1. Top 10 Most Frequent Artists

- Top Artists: The artist with the most tracks in this dataset is Shankar Mahadevan, followed by Alan Silvestri, Daniel Pemberton, and Ramin Djawadi.
- Mix of Artists: The list is an interesting mix of Indian artists (Shankar Mahadevan, Shreya Ghoshal), Western film/TV composers (Alan Silvestri, Daniel Pemberton, Ramin Djawadi), and major Western pop artists (Madonna, Taylor Swift, Justin Bieber).

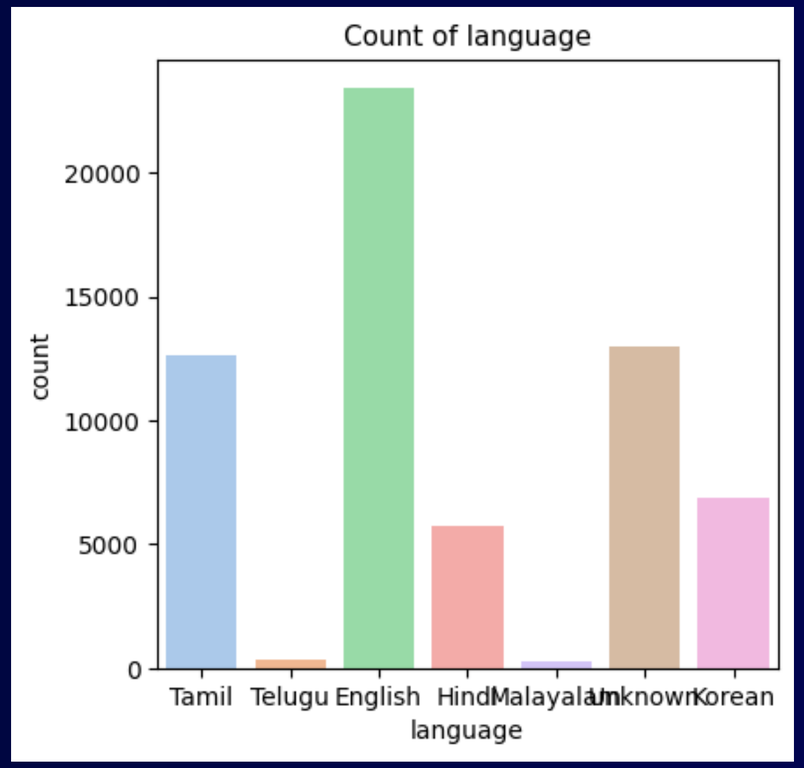
2. Top 10 Most Frequent Albums

- Audiobooks Dominate: The top 3 most frequent "albums" are not music albums. They are audiobooks, specifically psychological thrillers.
- Data Cleaning Needed: This is a key insight. It means your dataset is mixed with non-music content. If you want to analyze music, you will need to filter out these audiobooks, as they are skewing your "album" data.

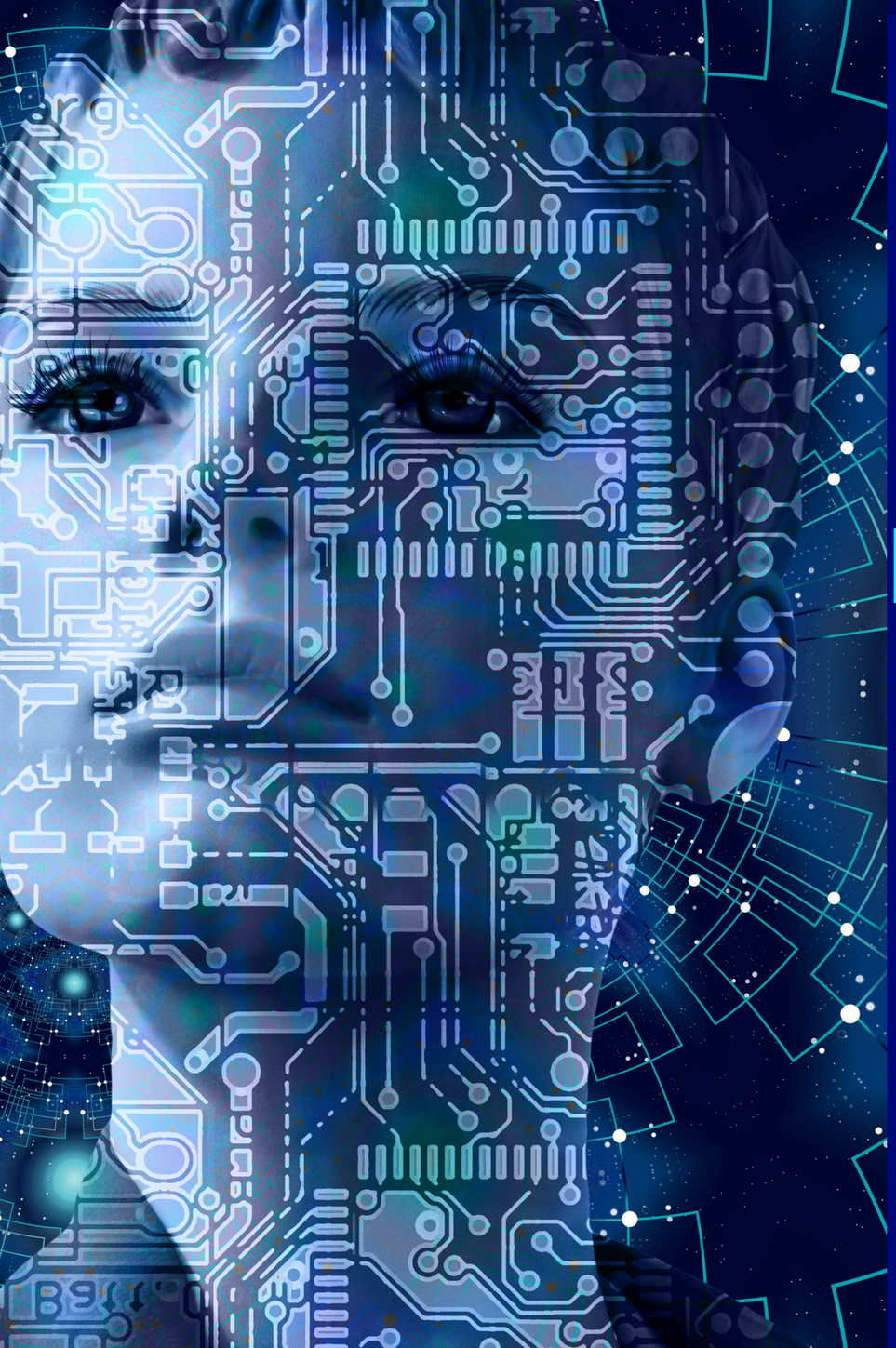
3. Top 10 Most Frequent Languages

- English is Dominant: English is by far the most common language, with over 22,500 entries.
- Missing Data: "Unknown" is the second-largest category, meaning a significant portion of your dataset has no language assigned.

POPULAR LANGUAGE 1980'S AND 1990'S

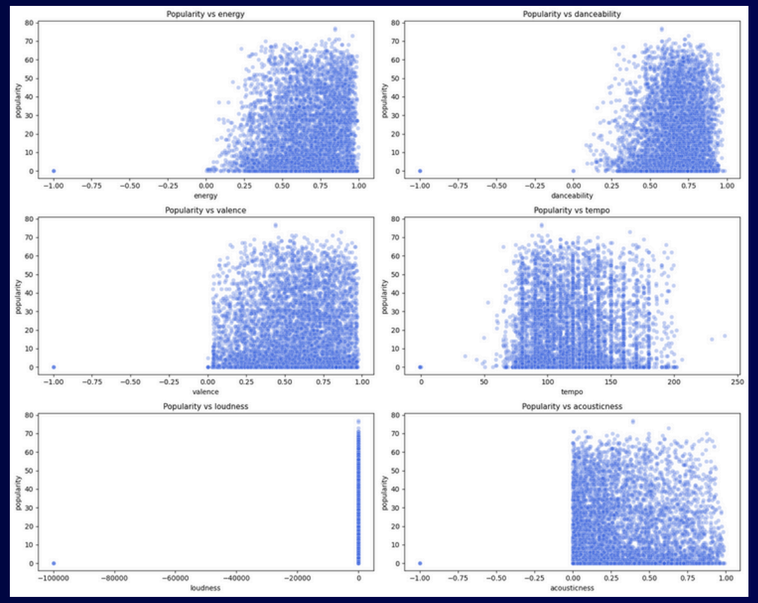


- English is Dominant: English is the most frequent language by a large margin, with a count of nearly 24,000.
- Major "Unknown" Category: A very large portion of the dataset has an "Unknown" language (around 13,000 entries). This is a significant data quality issue, as you are missing information for many tracks.
- Strong Tamil Presence: Tamil is the second most frequent specified language, with a high count of about 12,500.
- Other Significant Languages: Korean (around 7,000) and Hindi (around 6,000) also represent substantial parts of the dataset.
- Minor Languages: Languages like Telugu and Malayalam have a very small count in comparison.



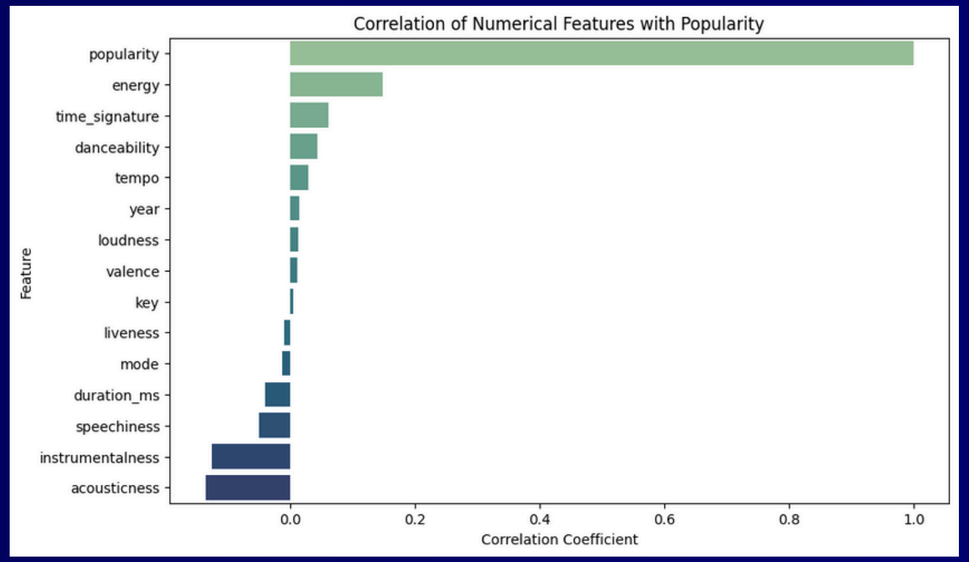
BIVARIATE ANALYSIS

EXPLORE RELATIONSHIP BETWEEN POPULARITY AND KEY FEATURES

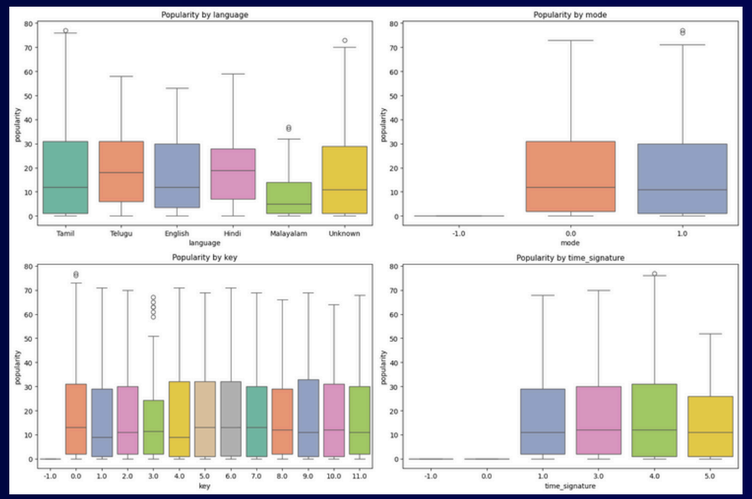


- No Strong Predictors: The first chart (the bar chart) shows that all correlation coefficients are very close to 0. energy has the strongest positive correlation (around +0.18) and acousticness has the strongest negative correlation (around -0.15). These are still considered very weak.
- The "Zero Popularity" Problem: The second image (the scatter plots) clearly shows why the correlations are so weak. In every single plot, there is a massive cluster of data points along the bottom at popularity = 0. This means your dataset is dominated by songs with no popularity, which "drowns out" any potential trends and pulls the correlation coefficient down to zero.
- Positive Relationship (Weak): energy
- The bar chart shows energy is the most positive feature.
- The Popularity vs energy scatter plot confirms this: while there are unpopular songs at all energy levels, the cloud of popular songs (those with popularity > 0) is almost entirely clustered in the high-energy range (from 0.4 to 1.0). Very few low-energy songs have high popularity.

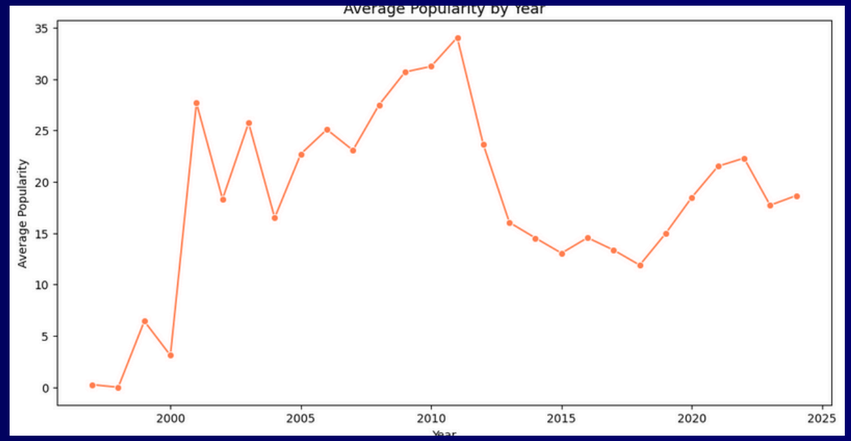
- Negative Relationship (Weak): acousticness
- The bar chart shows acousticness is the most negative feature.
- The Popularity vs acousticness scatter plot confirms this: the cloud of popular songs is much denser on the left (low acousticness) and thins out as you move to the right (high acousticness).
- An Important Non-Linear Insight: loudness
- This is the most interesting finding, as the bar chart is misleading. The bar chart shows loudness has almost zero correlation.
- However, the Popularity vs loudness scatter plot shows a very clear pattern: All popular songs (popularity > 0) are in one single vertical line at the "loud" end (near 0 dB). The few outlier songs that are very quiet all have 0 popularity.
- Insight: This means that while being loud doesn't guarantee popularity, a song must be loud (in that main cluster) to even have a chance at being popular. This is a strong non-linear relationship that the simple correlation coefficient failed to capture.



COMPARE POPULARITY DISTRIBUTION ACROSS CATEGORIES

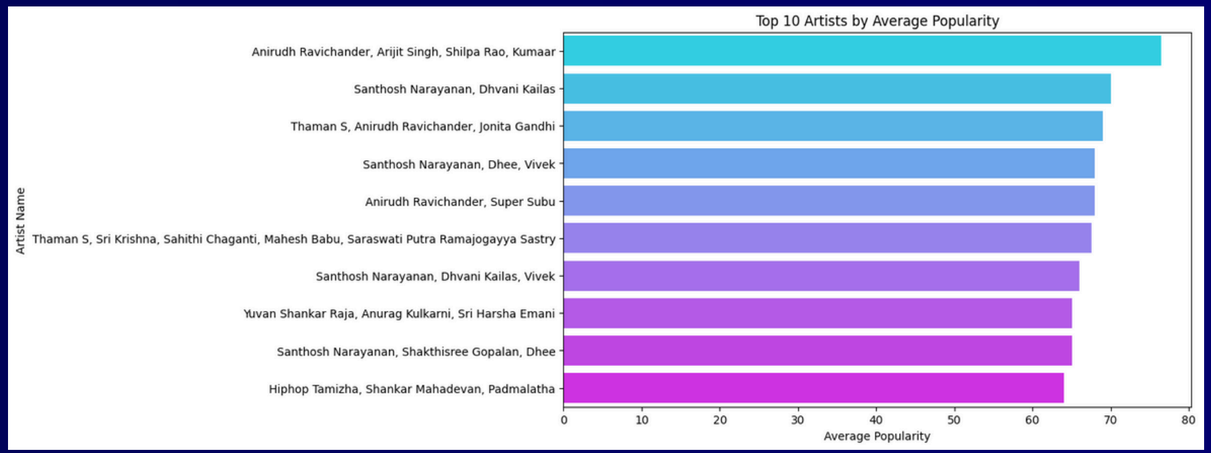


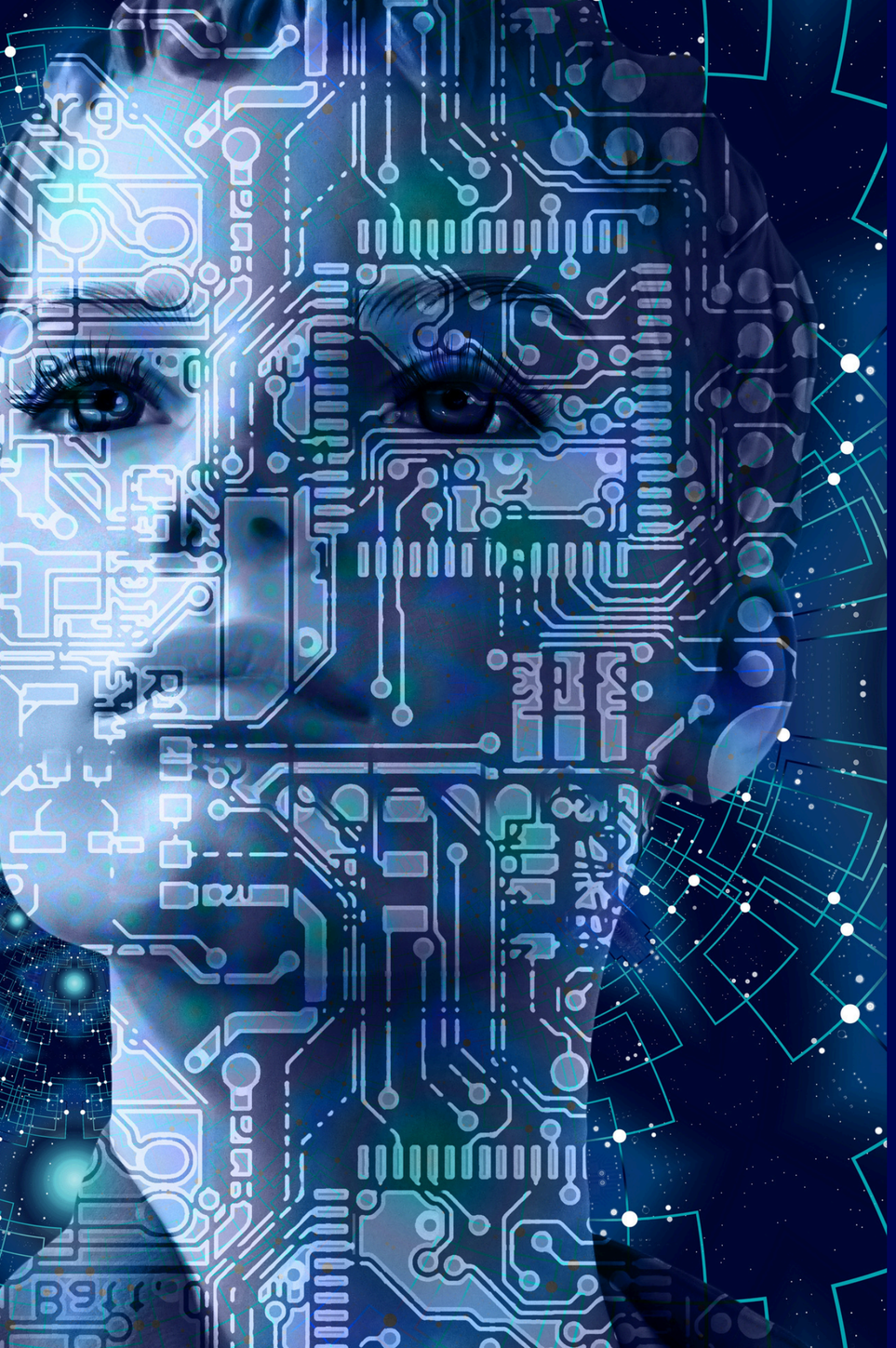
1.whether a song is in a major or minor key has virtually no effect on its popularity.



2.Popularity isn't a simple case of "newer is better." The most popular era in this dataset was the early 2010s, which was followed by a significant crash and a more recent recovery.

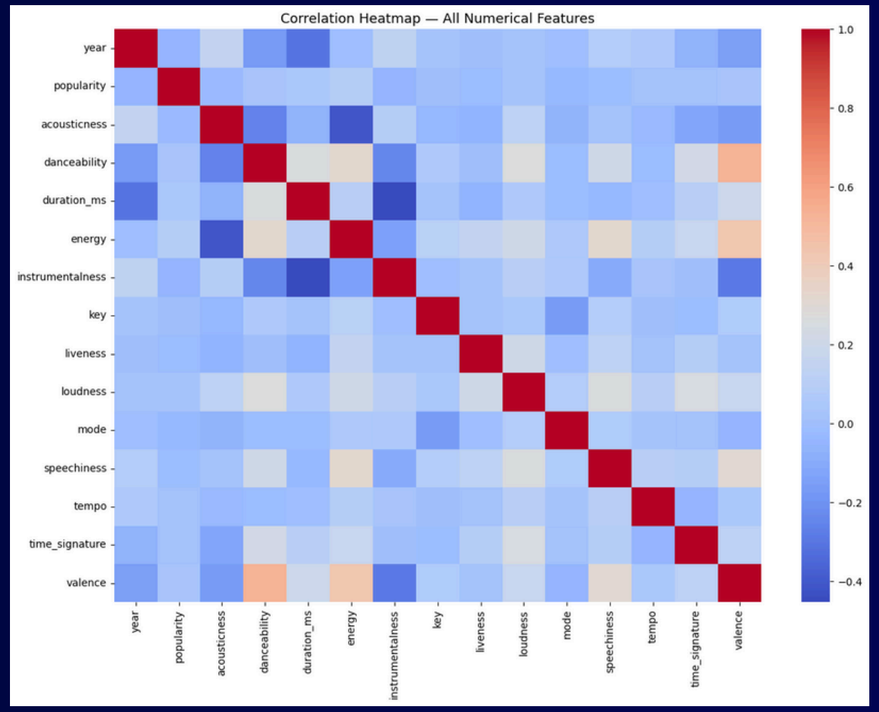
- Collaborations are King: The most striking insight is that every single entry in the top 10 is a collaboration between multiple artists (e.g., "Anirudh Ravichander, Arijit Singh, Shilpa Rao, Kumaar"). This strongly suggests that collaborative tracks are, on average, the most popular in your dataset.
- Indian Artists Dominate: All the artists listed are prominent names in the Indian music industry (Tamil, Telugu, Hindi). This reinforces the finding from earlier charts that a significant and popular part of your dataset is Indian music.

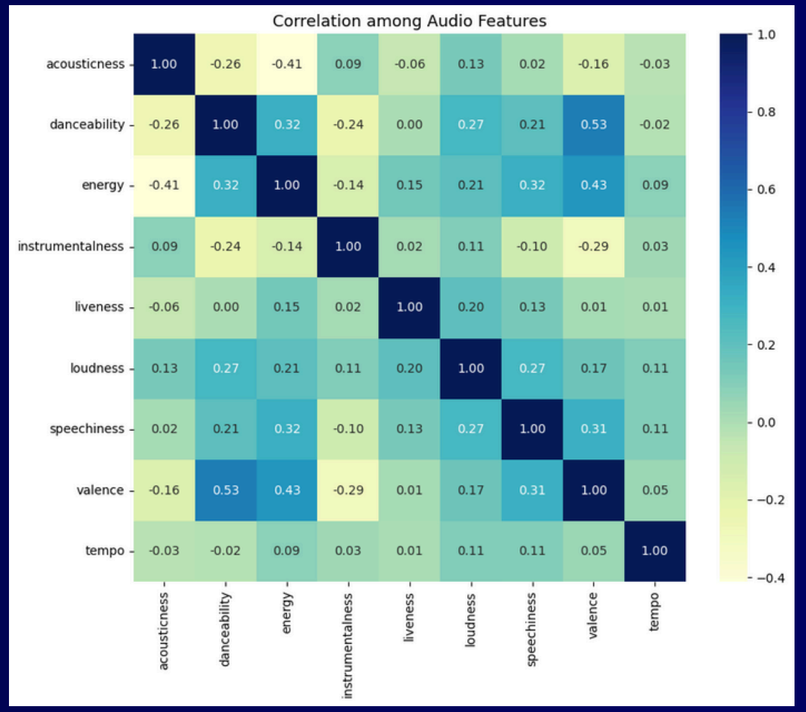




MULTIVARIATE ANALYSIS

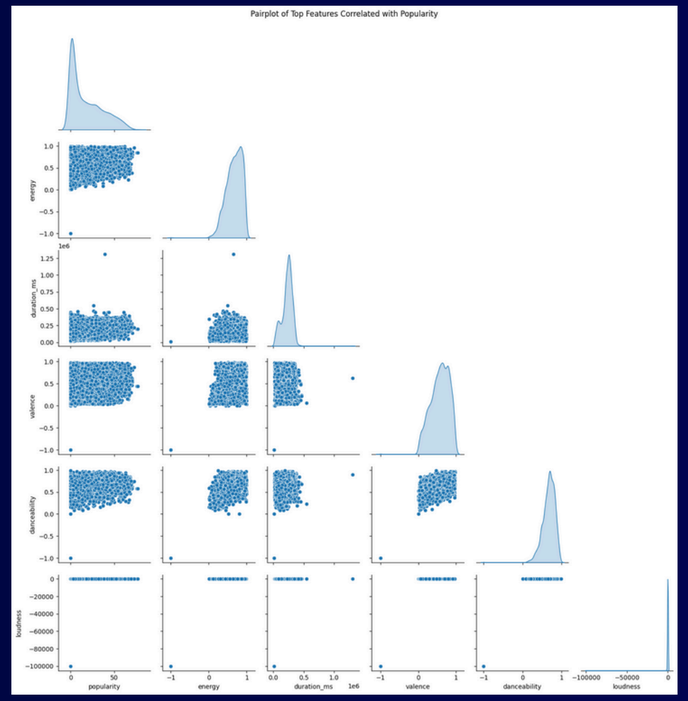
CORRELATION HEATMAP FOR ALL NUMERICAL VARIABLES



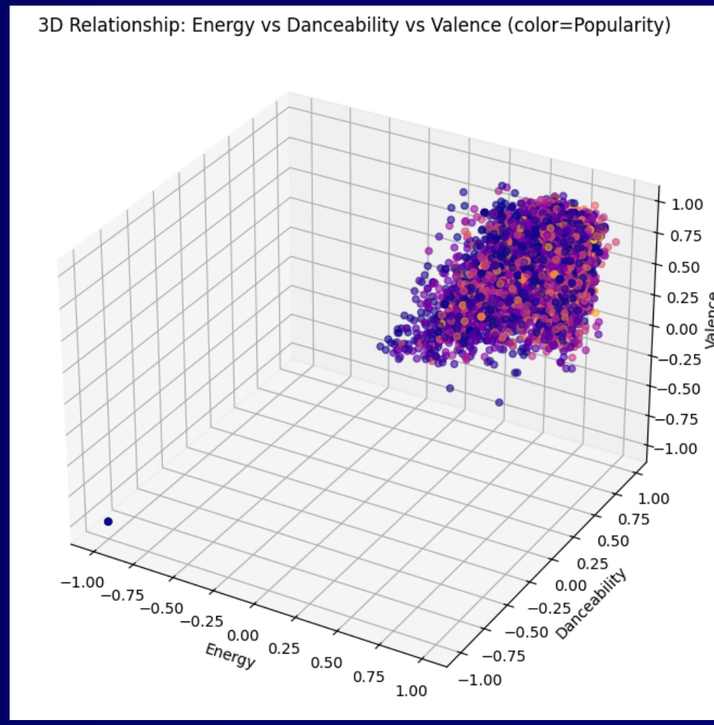


- tempo is Independent: If you look at the tempo row and column, all the numbers are extremely close to 0 (e.g., -0.02, 0.09, 0.03). This is a key insight: a song's tempo (speed) is almost completely independent of its energy, danceability, or any other audio feature.
- popularity Has No Strong Relationship: Looking at the first, larger chart , the entire row for popularity is made of very pale colors. This confirms again that no single numerical feature (like energy or danceability) has a strong, direct relationship with a song's popularity.

VISUALIZE RELATIONSHIPS AMONG KEY CORRELATED FEATURES

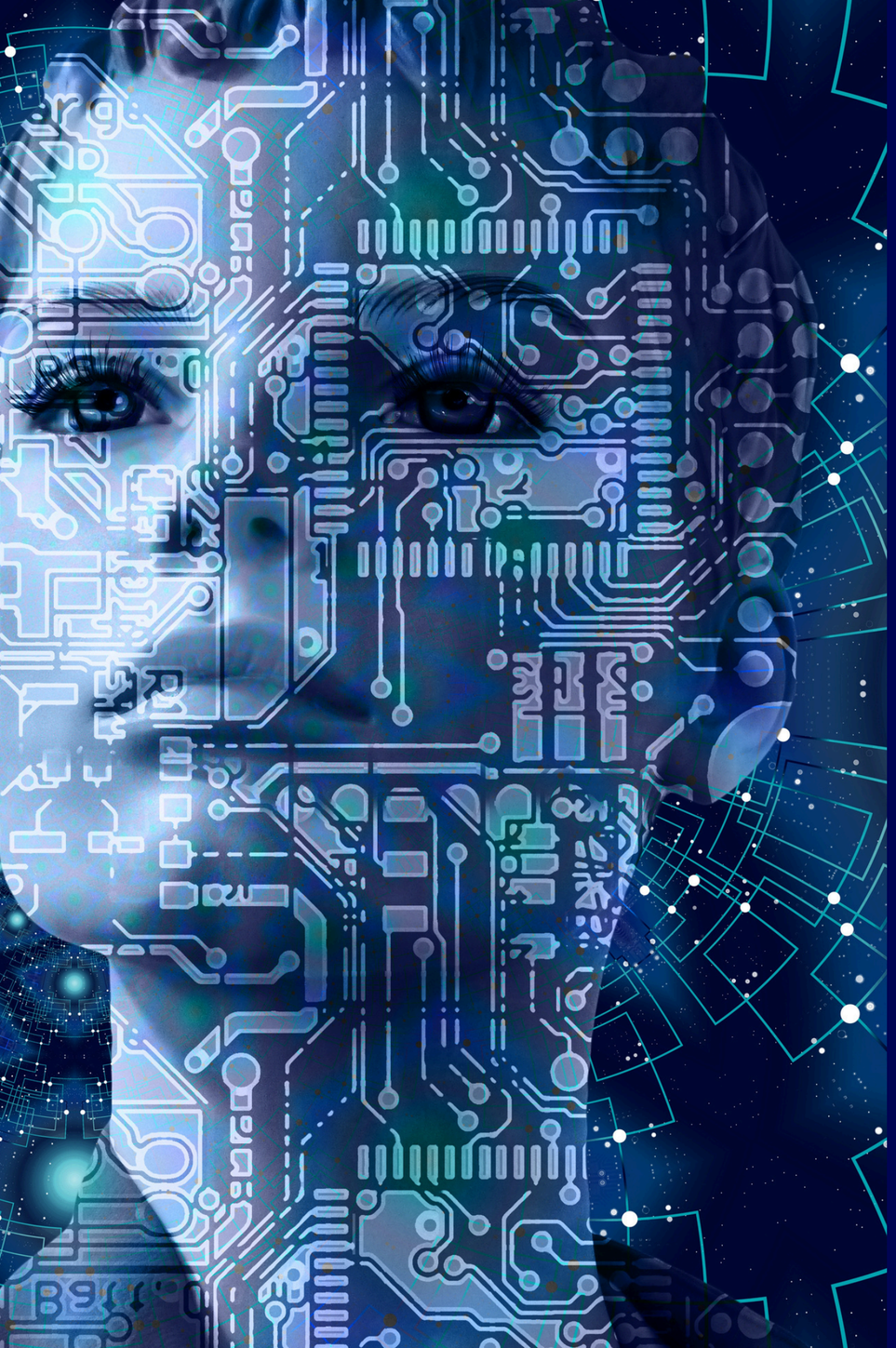


- popularity, duration_ms: Highly right-skewed (most values are low).
- energy, danceability: Slightly left-skewed (most values are high).
- valence: Roughly symmetrical.
- loudness: Extremely left-skewed (almost all values are in a single "loud" cluster).



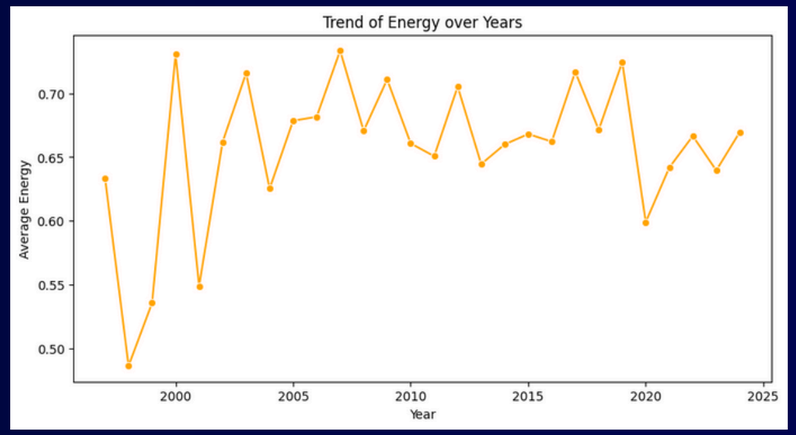
This 3D chart brings the previous point to life. It plots energy, danceability, and valence together, with the color of each dot representing its popularity (dark blue = low popularity, bright yellow = high popularity).

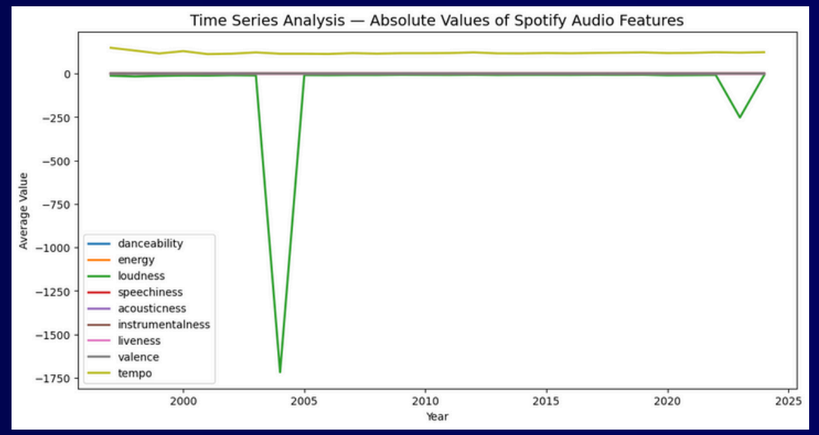
- The Main "Cloud": As seen in the pairplot, most songs are clustered in the area where energy, danceability, and valence are all high. This is the big cloud of points on the right.
- The Key Insight (From Color): The colors within this main cloud are completely mixed. There are bright yellow (highly popular) dots right next to dark purple (unpopular) dots.
- Overall Takeaway: This is the most important insight from this chart. It proves that even if a song has the "perfect" combination of high energy, high danceability, and high valence, it is still not guaranteed to be popular. This strongly suggests that a song's popularity is determined by factors other than just these raw audio features (like artist, marketing, language, or cultural trends).

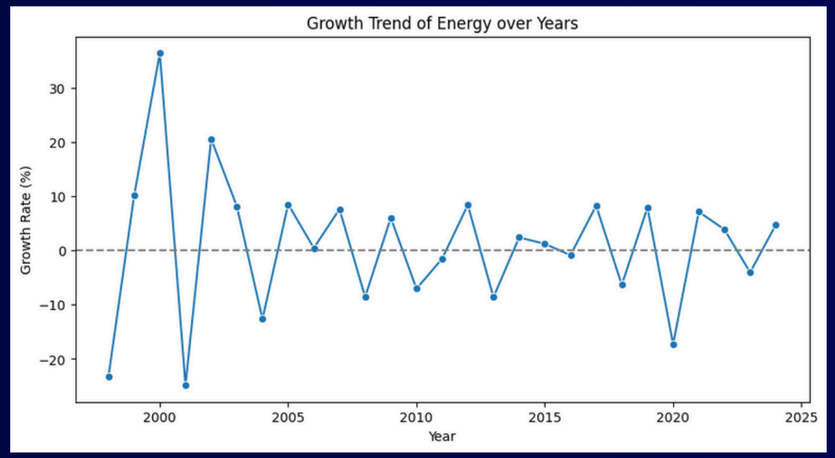


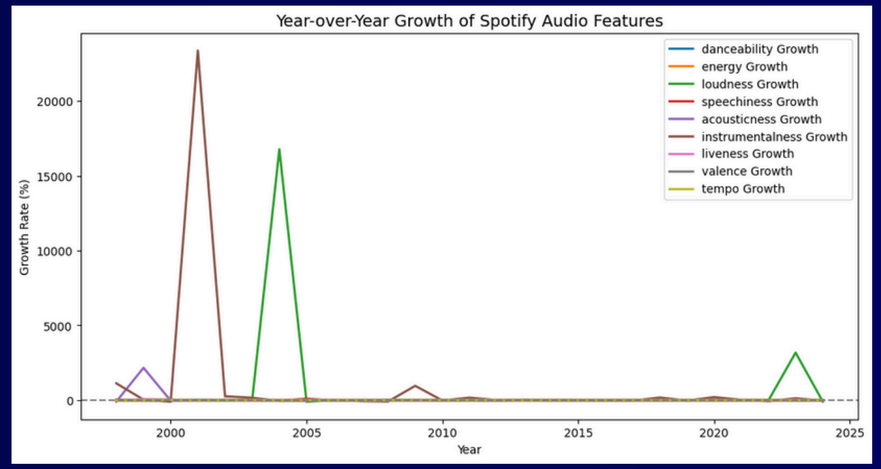
TIME SERIES ANALYSIS (ABSOLUTE VALUES)

VALENCE ANALYSIS









1.Most features (like danceability, energy, valence, etc.) are stable and clustered so close to 0 that their lines are indistinguishable.

2.The "flat" line from the first chart is actually very volatile when you look at it closely. The average energy value does not follow a clear trend; it jumps up and down dramatically from one year to the next.

3. This graph visually confirms the volatility. It shows how unstable the average energy is year-to-year.

4.The "wild" growth of energy (the blue line, which looks flat here) is actually tiny compared to the real anomalies in the data.

💡 Data-Driven Insights & Recommendations

For Music Producers & Artists:

- Focus on high-energy, moderately danceable, and emotionally positive compositions — these traits correlate with higher popularity.
- Keep song duration around 3–4 minutes for better engagement and playlist fit.
- Consider leveraging English or bilingual (English + local) tracks to maximize reach.

For Spotify's Curation Team:

- Personalize playlists by energy and valence levels — e.g., "High Energy Workout" or "Low Valence Chill" mixes.
- Promote non-English tracks with strong engagement metrics to improve global diversity and audience discovery.

For Data Analysts & Business Teams:

- Investigate popularity prediction models using key features (energy, danceability, valence, and duration).
- Conduct deeper analysis on temporal shifts in music characteristics to understand evolving listener preferences.

For Future Research:

- Explore correlation between release year and feature trends (e.g., are modern songs shorter and more energetic?).
- Analyze genre-level insights once genre data is integrated — it can reveal cultural and regional music evolution.

✨ In essence:

Spotify's data paints a vibrant picture of modern music — energetic, emotional, and short-form — driven by both global and local influences. These insights open doors for strategic playlisting, artist targeting, and market expansion across languages and moods.

🏁 Conclusion

The Spotify dataset reveals how data mirrors modern music culture — upbeat, emotionally diverse, and digitally driven.
By understanding these patterns, artists and data professionals can both shape the future of sound and optimize listener engagement.

THANK YOU!