BY ANGSHURPITA GANGULY

SPOTIFY DATASET ANALYSIS

OVERVIEW

Core Audio Features

These features describe the intrinsic sonic and emotional qualities of each track, calculated directly from the audio.

- **Energy:** The perceived intensity and activity of a track.
- **Danceability:** How suitable a track is for dancing based on tempo, rhythm, and beat strength.
- **Valence:** The musical positiveness or mood of a track (happy, cheerful vs. sad, angry).
- **Acousticness:** A measure of whether the track is acoustic or electronic.
- **Speechiness:** The presence of spoken words in a track.
- **Liveness:** Detects the presence of a live audience in the recording.
- **Instrumentalness:** Predicts whether a track contains no vocals.
- **Tempo:** The speed of the track in Beats Per Minute (BPM).

Contextual & Performance Metadata

This data provides context about a track's identity, physical properties, and performance in the market.

- **Popularity:** A score indicating the track's current popularity.
- **Duration:** The length of the track in milliseconds.
- **Language:** The primary language of the track's lyrics.
- **Artist & Album Name:** The creators and collection associated with the track.

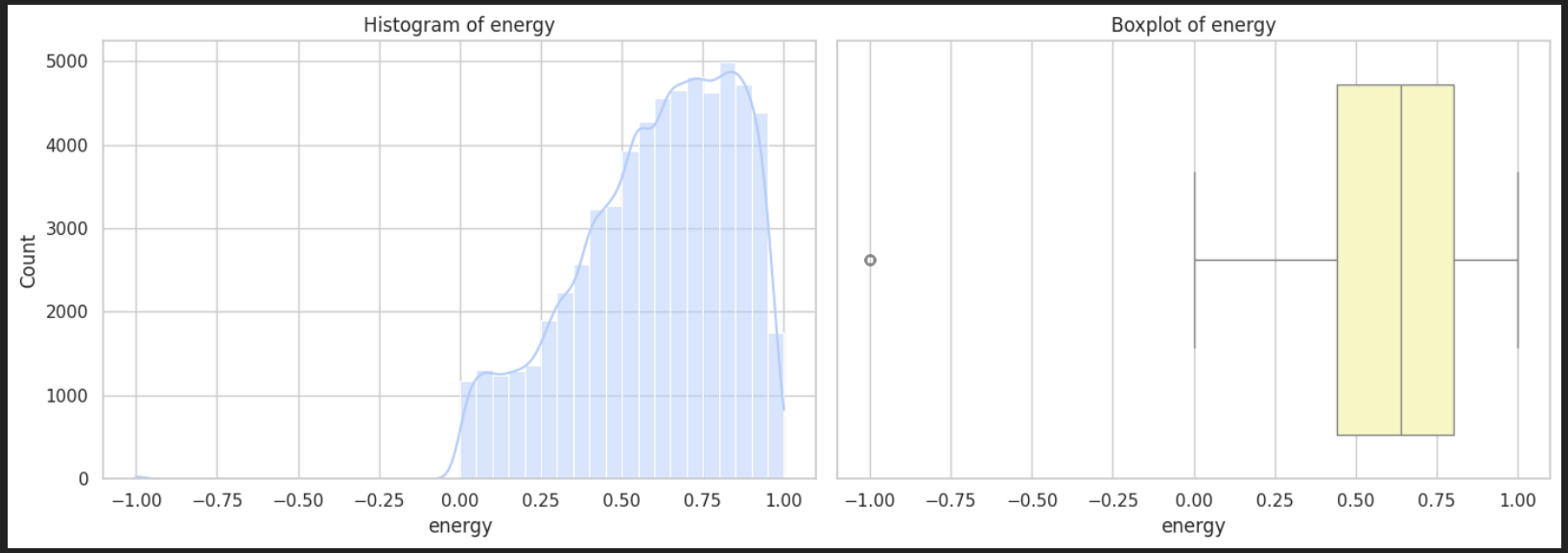
Machine Learning & Derived Features

We also explored advanced features derived from the original data to uncover deeper, underlying patterns.

- **Principal Components (PCA):** A technique to combine and simplify audio features.
- **K-Means Clusters:** Automatic, data-driven groupings of songs based on their sonic similarity.

UNIVARIATE ANALYSIS

ANALYSIS OF ENERGY



ANALYSIS OF ENERGY

Graph Analysis

- The histogram is **left-skewed**, clearly showing that a majority of songs in the dataset have high energy.
- The primary peak is concentrated in the **0.6 to 0.9 energy range**.
- The boxplot confirms this with a **high median energy level** of approximately 0.7.

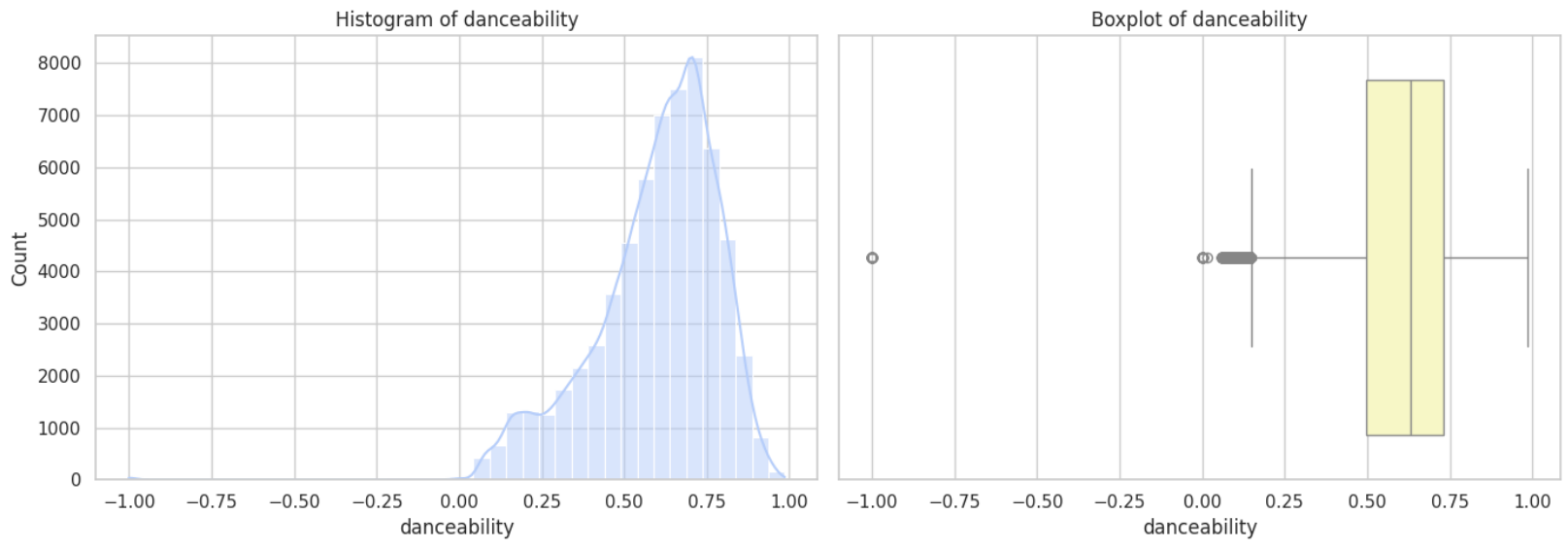
Business Insights

- The catalog is optimized for high-energy contexts like **workouts and parties**.
- It effectively serves users seeking an **upbeat, motivational experience**.
- A small niche of **very low-energy tracks** (e.g., ambient, classical) also exists.

Actions & Improvements

- **Prioritize** high-energy songs for playlists like "Workout Mix" and "Party Hits."
- **Consider acquiring** more mid-energy songs to better serve "Chill" or "Relaxing" moods.
- **Investigate** the low-energy outliers to ensure data quality and proper categorization.

ANALYSIS OF DANCEABILITY



ANALYSIS OF DANCEABILITY

Graph Analysis

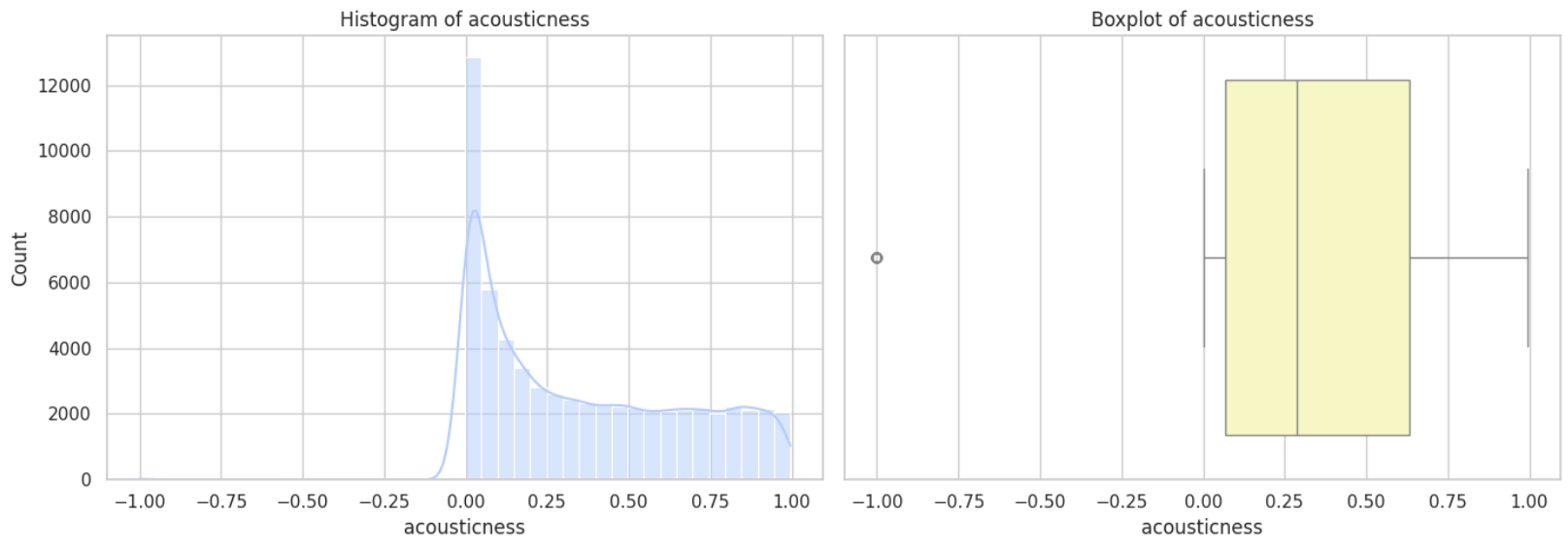
- The distribution is **nearly symmetrical and bell-shaped**, similar to a normal distribution.
- The peak of the data is centered around a **danceability score of 0.7**.
- The boxplot shows a **balanced distribution** of danceability scores around the median.

Business Insights

- The catalog is versatile and strong in popular, **dance-friendly genres** (Pop, EDM).
- It's well-equipped to serve a broad audience with **diverse dance-themed playlists**.
- The alignment with high energy suggests a **coherent "upbeat" library**.

- **Actions & Improvements**
  - **Launch** marketing campaigns targeting dance enthusiasts and fitness classes.
  - **Create** tiered playlists based on danceability intensity (e.g., "Chill Groove," "Dance Party").
  - **Analyze** the low-danceability outliers to identify spoken-word or ambient content.

ANALYSIS OF ACOUSTICS



ANALYSIS OF ACOUSTICS

Graph Analysis

- The distribution is **extremely right-skewed**, with a massive concentration of values close to 0.
- The vast majority of tracks have an **acousticness score below 0.2**.
- The boxplot's low median confirms that **most songs in the library are not acoustic**.

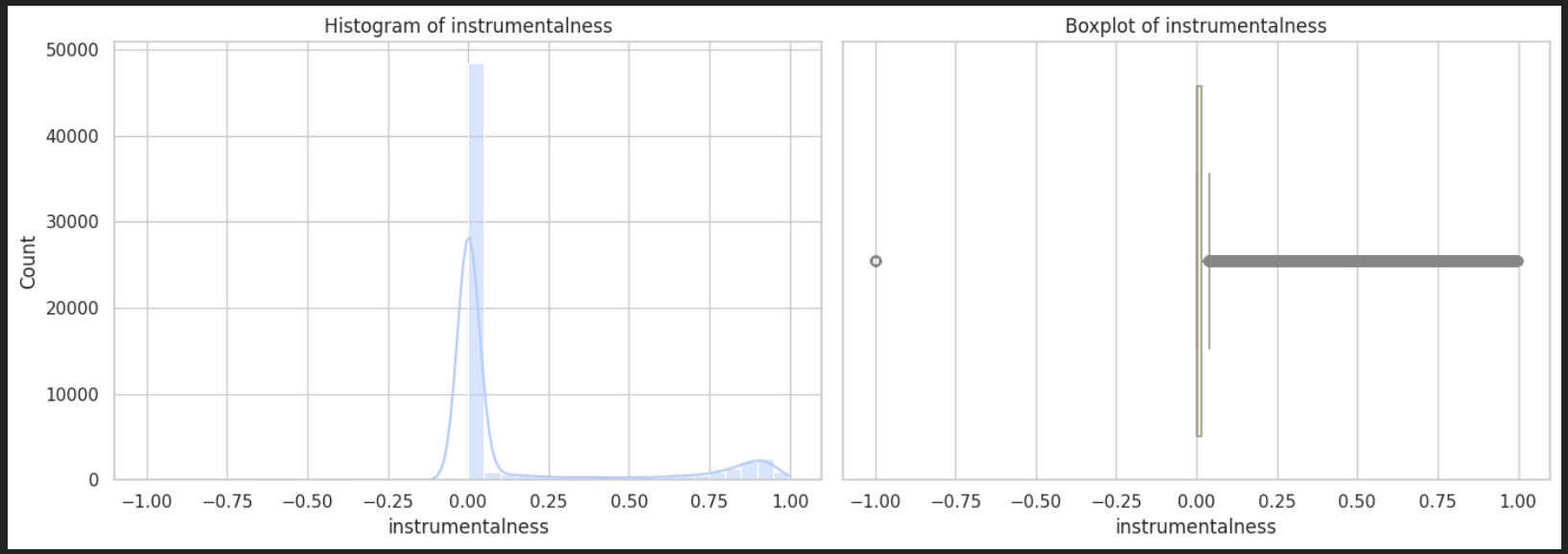
Business Insights

- The catalog is dominated by **electronic and amplified music**, reflecting modern trends.
- Acoustic music represents a **distinct and important niche** within the library.
- There's a clear sonic divide between the mainstream majority and the acoustic minority.

Actions & Improvements

- **Create** and heavily promote specialized playlists like "Acoustic Cafe" and "Unplugged."
- **Analyze** user demand to identify acoustic music as a potential area for content acquisiti
- **Use** log transformation on this feature for any machine learning modeling to handle the

ANALYSIS OF INSTRUMENTALNESS



ANALYSIS OF INSTRUMENTALNESS

Graph Analysis

- The distribution shows an overwhelming peak at exactly 0.
- The boxplot is compressed into a single line at 0, meaning the **median is 0**.
- This indicates that over 75% of the songs contain vocals.

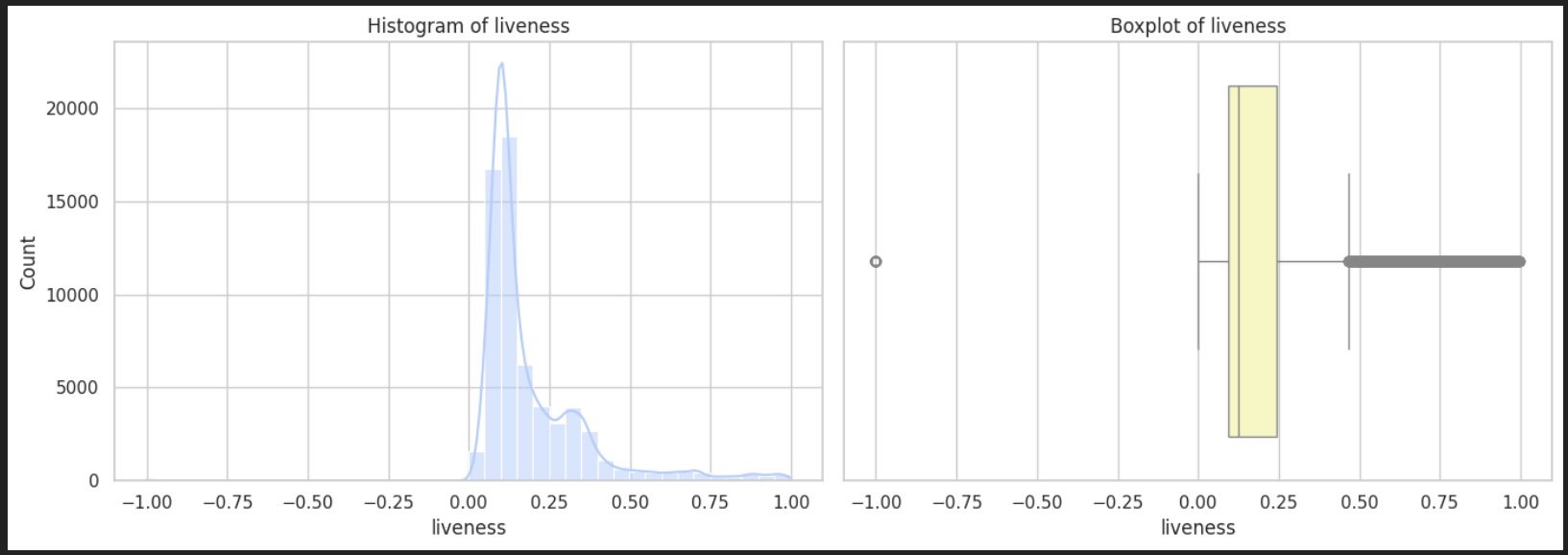
Business Insights

- The catalog consists almost **exclusively of vocal music**.
- This represents a significant **content gap** for users seeking music for focus or studying.
- The platform primarily serves **mainstream listeners** who prefer vocal tracks.

Actions & Improvements

- **Make** a business case for acquiring instrumental music catalogs (e.g., Lo-fi, Classical).
- **Convert** this feature to a binary `Vocal` vs. `Instrumental` tag for easier classification.
- **Ensure** the few existing instrumental tracks are correctly tagged for easy discovery.

ANALYSIS OF LIVENESS



ANALYSIS OF LIVENESS

Graph Analysis

- The histogram is **highly right-skewed**, indicating most songs are studio recordings.
- The main peak is centered around a **liveness score of 0.1 to 0.2**.
- A long tail to the right signifies the presence of a **small number of live tracks**.

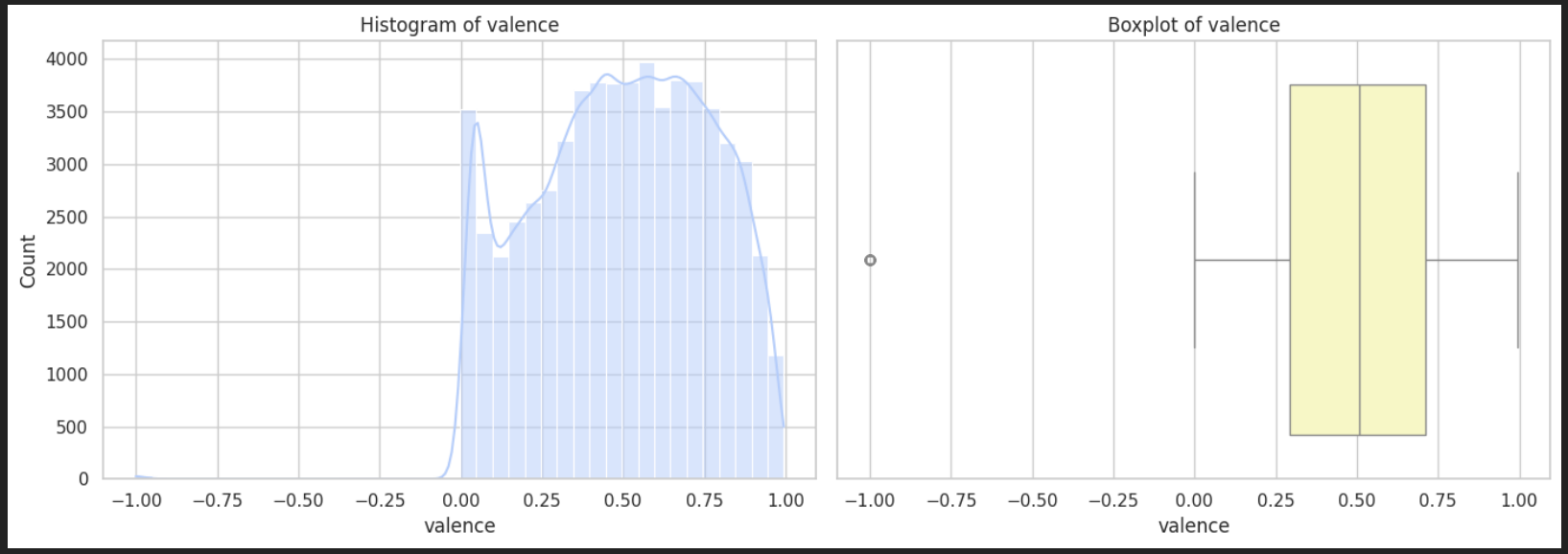
Business Insights

- The library is predominantly composed of studio recordings.
- Live recordings are a specialized category that can **engage dedicated fans**.
- This content offers an opportunity for **deeper artist-fan connections**.

Actions & Improvements

- **Curate** "Live Concerts" or "In Performance" playlists to highlight this content.
- **Form** partnerships to acquire exclusive live recordings from festivals as a unique selling point.
- **Enhance** artist pages by linking studio tracks to their live versions.

ANALYSIS OF VALENCE (MOOD)



ANALYSIS OF VALENCE (MOOD)

Graph Analysis

- The distribution is **bimodal**, meaning there are two distinct peaks or groups of songs.
- A large peak between **0.5 and 0.8** represents many positive-sounding tracks.
- A second, smaller peak near **0.1** represents a group of sad or melancholy tracks.

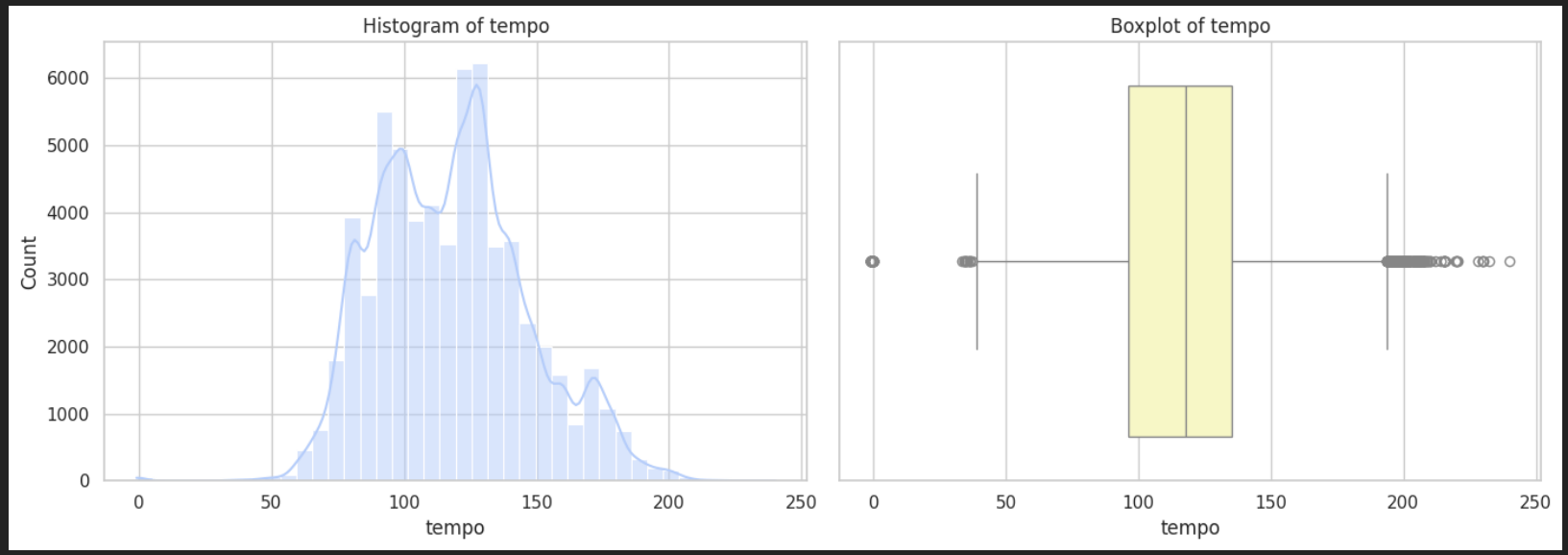
Business Insights

- The catalog is well-stocked with **emotionally expressive music**.
- The bimodal nature is perfect for **strong, mood-based playlisting**.
- This feature is crucial for creating playlists that **match a user's emotional state**.

Actions & Improvements

- **Create** and promote distinct mood-based playlists like "Mood Booster" and "Sad Songs."
- **Use** valence as a primary feature in the recommendation engine to suggest mood-matching songs.
- **Analyze** the artists and genres in the low-valence peak to better serve that audience.

ANALYSIS OF TEMPO (BPM)



ANALYSIS OF TEMPO (BPM)

Graph Analysis

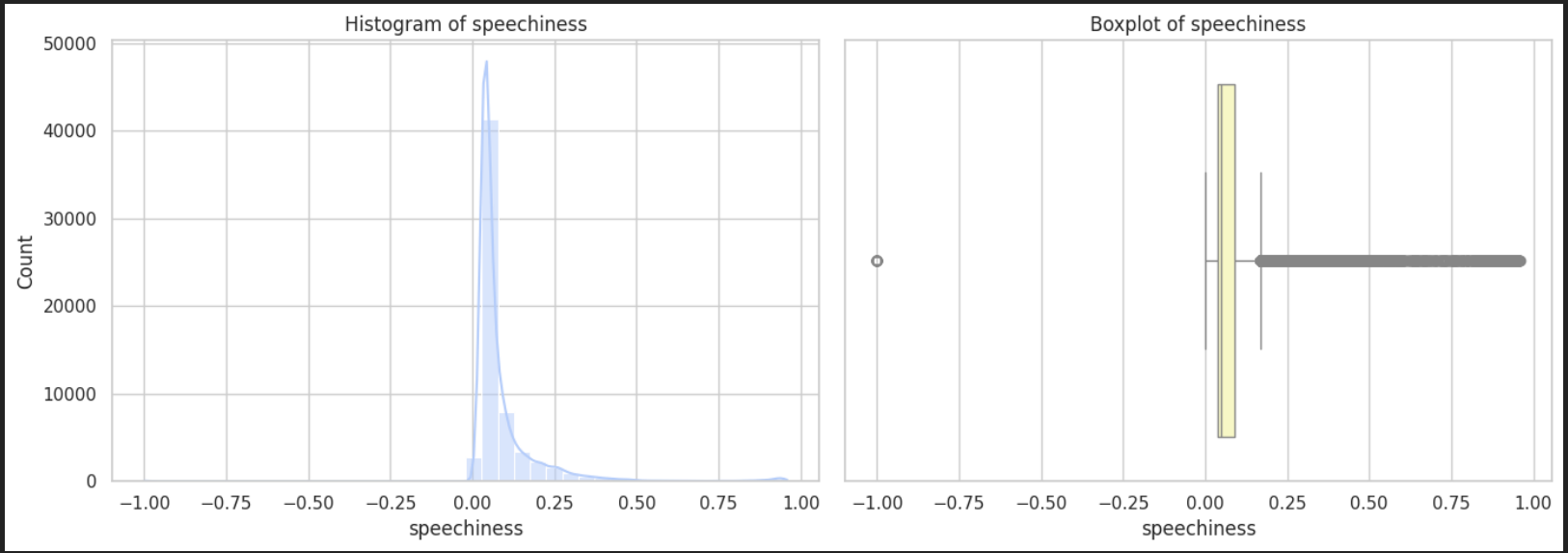
- The distribution is **multimodal**, with several common tempo ranges visible.
- The most prominent peak is around **120-130 BPM** (common in Pop and Dance music).
- A second significant peak is visible around **90-100 BPM** (common in Hip-Hop and Rock).

Business Insights

- The tempo distribution reflects the core of **popular music genres**.
- The variety in tempo allows the platform to cater to **different activities and energy levels**.
- Tempo is a powerful tool for creating a **seamless listening experience**.

- ### Actions & Improvements
  - **Curate** activity-based playlists using BPM ranges, such as "Running Mix (160-180 BPM)."
  - **Develop** a "smart shuffle" feature that can order a queue by tempo.
  - **Correlate** tempo peaks with specific genres to improve automatic genre tagging.

ANALYSIS OF SPEECHINESS



ANALYSIS OF SPEECHINESS

Graph Analysis

- The histogram is **extremely right-skewed**, with a massive peak near 0.
- This shows the vast majority of tracks are **musical, not spoken-word**.
- The long tail represents a small number of tracks with high speechiness, like **rap or podcasts**.

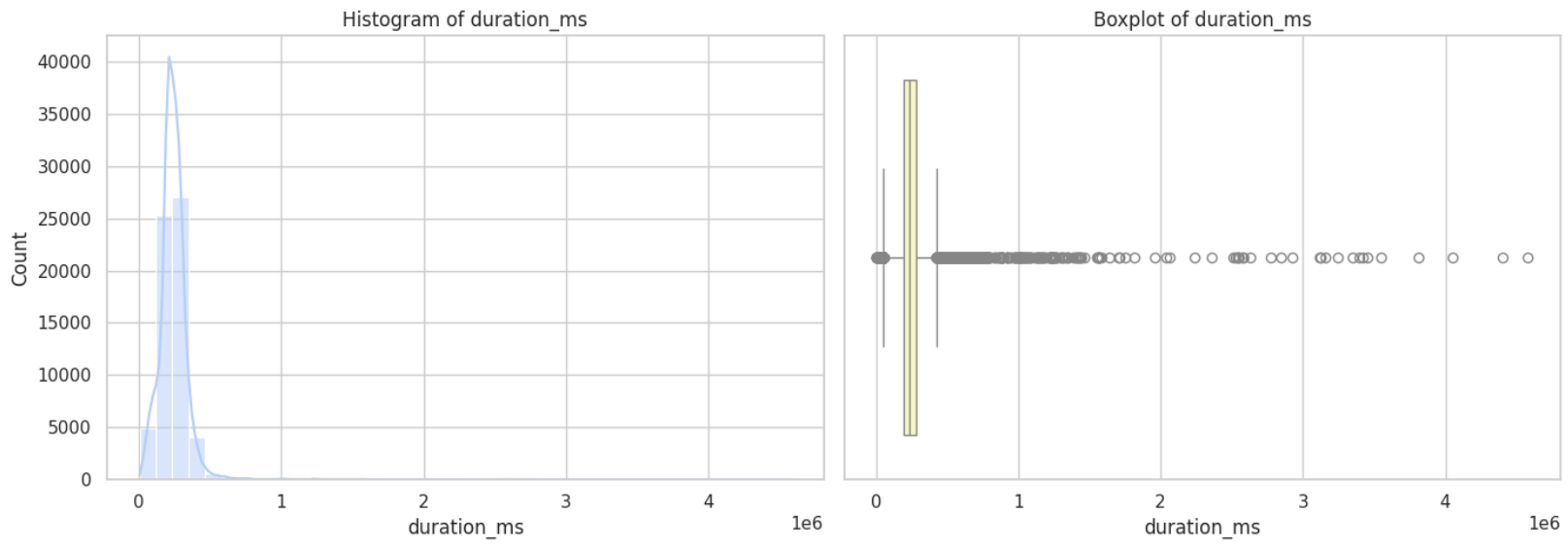
Business Insights

- This feature is an excellent classifier to distinguish music from audiobooks/podcasts.
- Tracks with moderate speechiness are very likely to be from the **Hip-Hop/Rap genre**.
- This clear separation allows for **effective content categorization**.

Actions & Improvements

- **Use** a speechiness threshold to automatically flag and separate non-music cont
- **Use** moderately high speechiness as a strong indicator for identifying Rap song
- **Allow** users to filter playlists to exclude tracks with high speechiness.

ANALYSIS OF DURATION



ANALYSIS OF DURATION

Graph Analysis

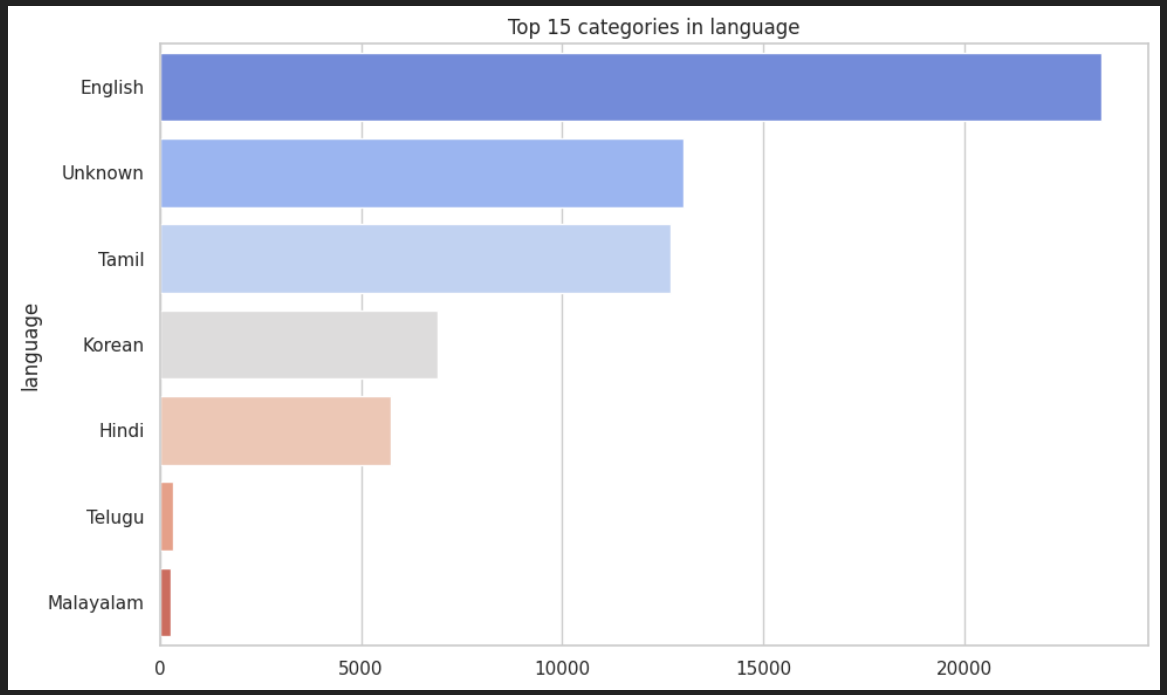
- The distribution is **extremely right-skewed**, showing most songs are of a similar length.
- There is a clear peak around **3.5 to 4 minutes** (200,000 to 250,000 ms).
- Numerous outliers on the right indicate the presence of some **very long tracks**.

Business Insights

- The average song length aligns with **standard radio-friendly pop music**.
- The long-duration outliers represent **niche content** (DJ mixes, classical pieces).
- This distribution is typical for a **large, diverse music library**.

- ### Actions & Improvements
  - **Filter out** extreme outliers in general playlists to maintain a consistent listening flow.
  - **Group** the long-duration tracks into specialized playlists like "Deep Focus."
  - **Apply** a log transformation to this feature before using it in machine learning models.

ANALYSIS OF LANGUAGE



ANALYSIS OF LANGUAGE

Graph Analysis

- **English** is by far the most dominant language in the dataset.
- **Tamil and Korean** represent the next largest language groups with significant track counts.
- A large **"Unknown"** category highlights a key data quality issue.

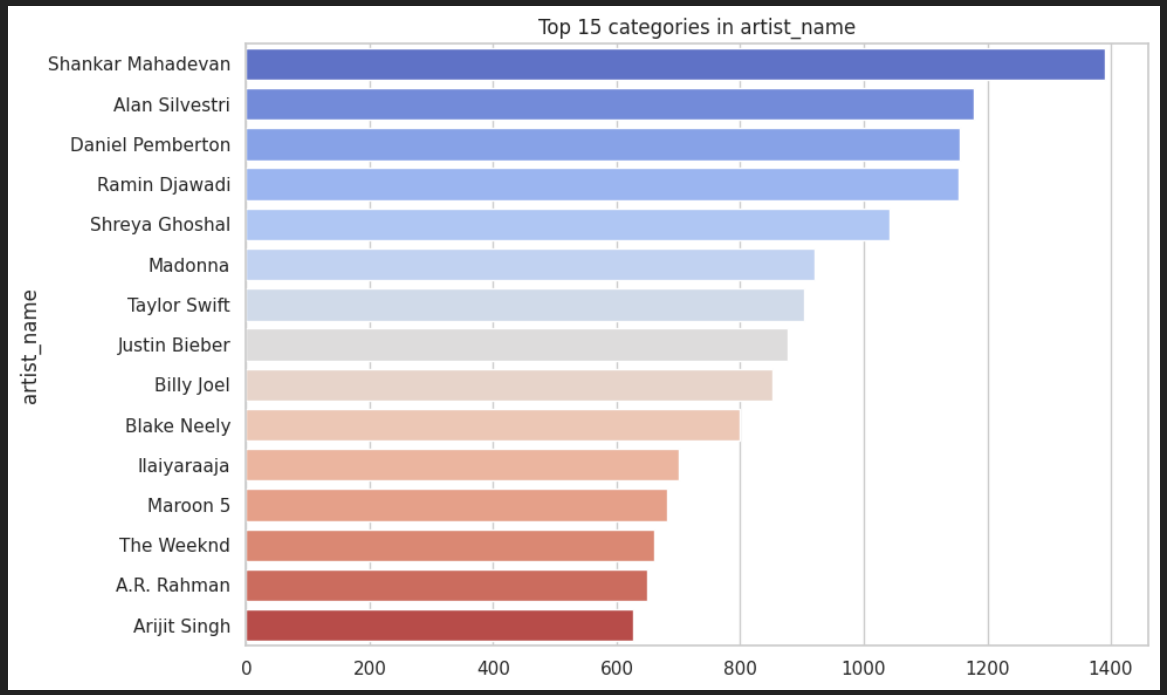
Business Insights

- The platform has a strong strategic position in the **English, Tamil, and K-Pop markets**.
- The "Unknown" songs are likely **not being recommended effectively**, representing lost opportunities.
- The presence of multiple Indian languages signals a focus on that **high-growth region**.

Actions & Improvements

- **Launch** a project to identify and correctly tag the language for the "Unknown" tracks.
- **Leverage** the strong Tamil and Korean catalogs with targeted marketing in those regions.
- **Develop** a strategy to grow the catalogs for other major Indian languages.

ANALYSIS OF TOP ARTISTS



ANALYSIS OF TOP ARTISTS

Graph Analysis

- The list is dominated by a mix of **film score composers** and **Indian music artists**.
- **Shankar Mahadevan** has the most tracks in the dataset.
- Several global pop icons like **Taylor Swift and Madonna** are also present.

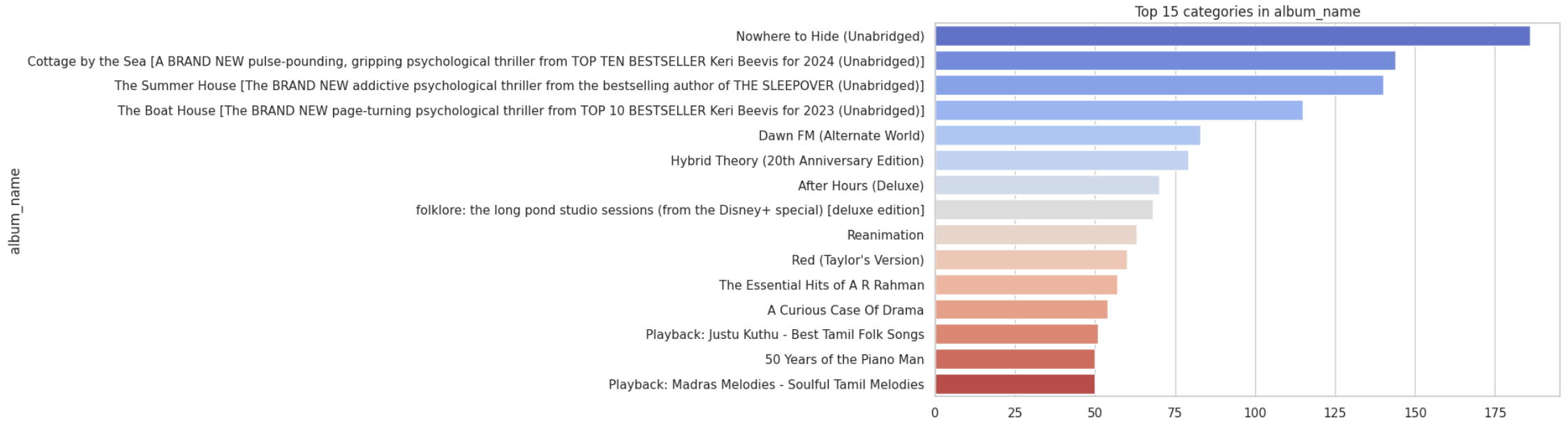
Business Insights

- The catalog has a unique specialization in **soundtracks and film scores**.
- The mix of artists indicates a "glocal" (global + local) content strategy.
- This blend is ideal for a **diverse user base**, especially in the Indian market.

Actions & Improvements

- **Create** and promote flagship playlists like "Epic Movie Scores" and "Bollywood's Best."
- **Develop** targeted campaigns aimed at film enthusiasts and regional audiences in India.
- **Strengthen** partnerships with movie studios to solidify this content advantage.

ANALYSIS OF TOP ALBUMS



ANALYSIS OF TOP ALBUMS

Graph Analysis

- The top entries by track count are predominantly **audiobooks**, not music albums.
- This indicates the dataset is a mix of music and spoken-word content.
- The top music albums that do appear are from a **diverse range of genres**.

Business Insights

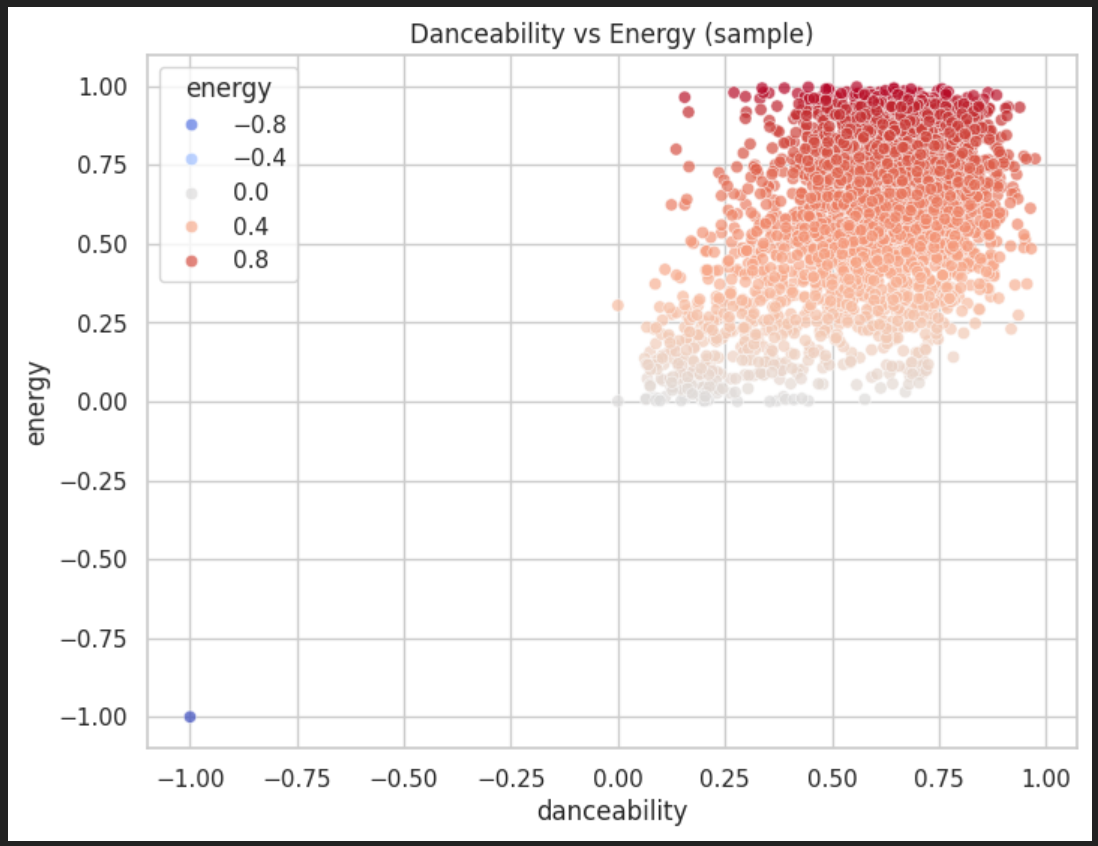
- **CRITICAL FINDING:** The dataset is contaminated with non-music content.
- This can severely **skew music-focused analysis** and lead to poor recommendations.
- The platform's content is **broader than just music**, which could be a pro or a con.

Actions & Improvements

- **Implement** a robust method to separate music from audiobooks (using `speechiness`).
- **Create** separate sections in the user interface for Music and Audiobooks.
- **Standardize** and clean album titles in the database for a better user experience.

BIVARIATE ANALYSIS

DANCEABILITY VS. ENERGY



DANCEABILITY VS. ENERGY

Graph Analysis

- The scatter plot reveals a **strong, positive correlation** between the two variables.
- As energy increases, danceability also tends to increase.
- Most songs are clustered in the top-right, indicating a catalog rich in **high-energy, high-danceability** tracks.

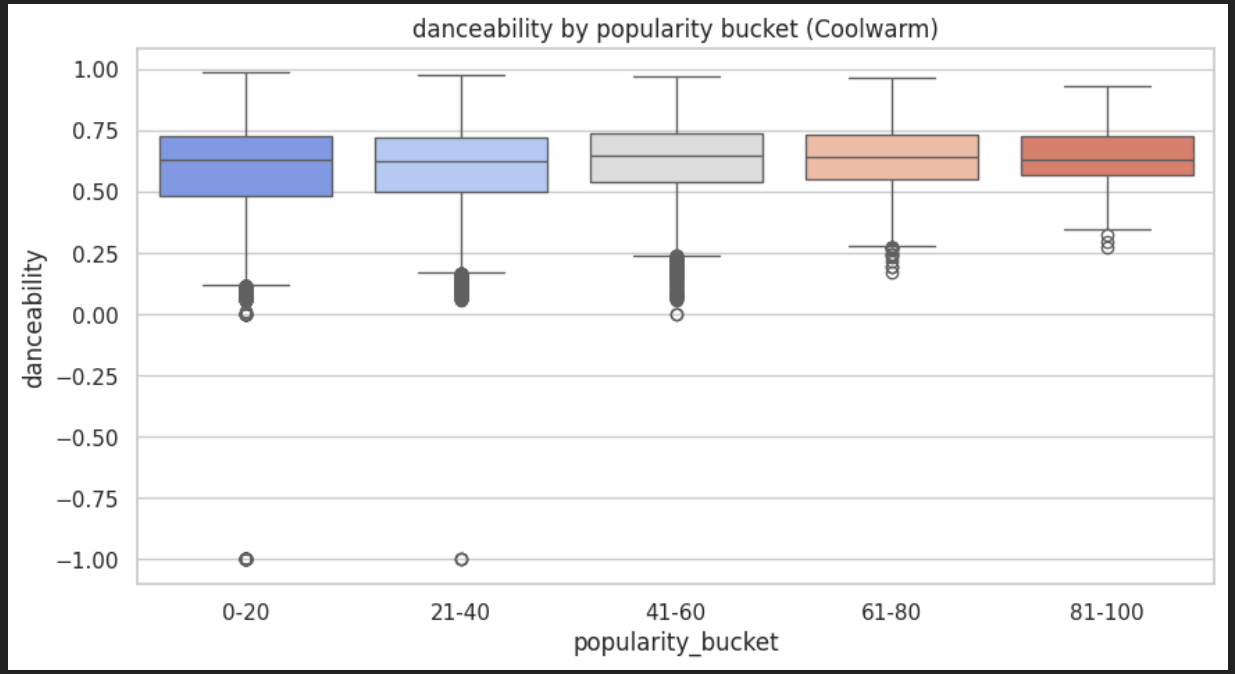
Business Insights

- Energy and danceability are **closely linked** for most of the catalog.
- This provides confidence in creating **consistent, upbeat listening experiences**.
- It is rare to find high-energy, non-danceable songs, indicating a **potential niche content gap**.

Actions & Improvements

- **Simplify** playlisting rules, as "High Energy" playlists will also be danceable.
- **Combine** these two variables into a single "Upbeat Score" for machine learning models.
- **Investigate** the single outlier at (-1.0, -1.0) as a potential data error.

DANCEABILITY VS. POPULARITY



DANCEABILITY VS. POPULARITY

Graph Analysis

- The median danceability is consistently high across all popularity levels.
- The variance (size of the box) decreases as popularity increases.
- This shows that the most popular songs occupy a **tighter, more predictable range** of danceability.

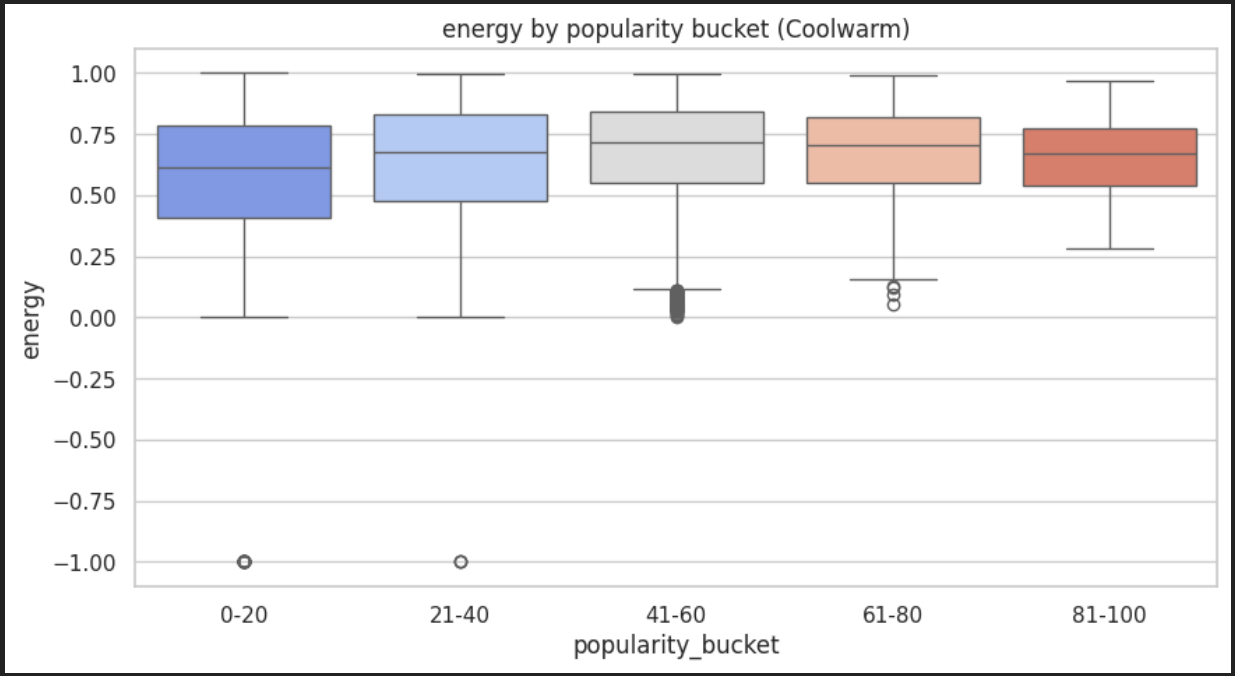
Business Insights

- To become highly popular, a song needs a danceability score within a specific **"sweet spot."**
- Mainstream listeners prefer a reliable and predictable level of danceability in hit songs.
- Extremely low or high danceability **rarely leads to a massive hit**.

Actions & Improvements

- **Prioritize** songs within the "sweet spot" (~0.6 to 0.8) for "Top Hits" playlists.
- **Use** this optimal range as a key feature in any "hit prediction" model.
- **Promote** songs with outlier danceability scores to niche or genre-specific audiences.

ENERGY VS. POPULARITY



ENERGY VS. POPULARITY

Graph Analysis

- The median energy is consistently high for all popularity buckets.
- The variance of energy tightens for more popular songs.
- The median dips slightly for the most popular (81-100) group.

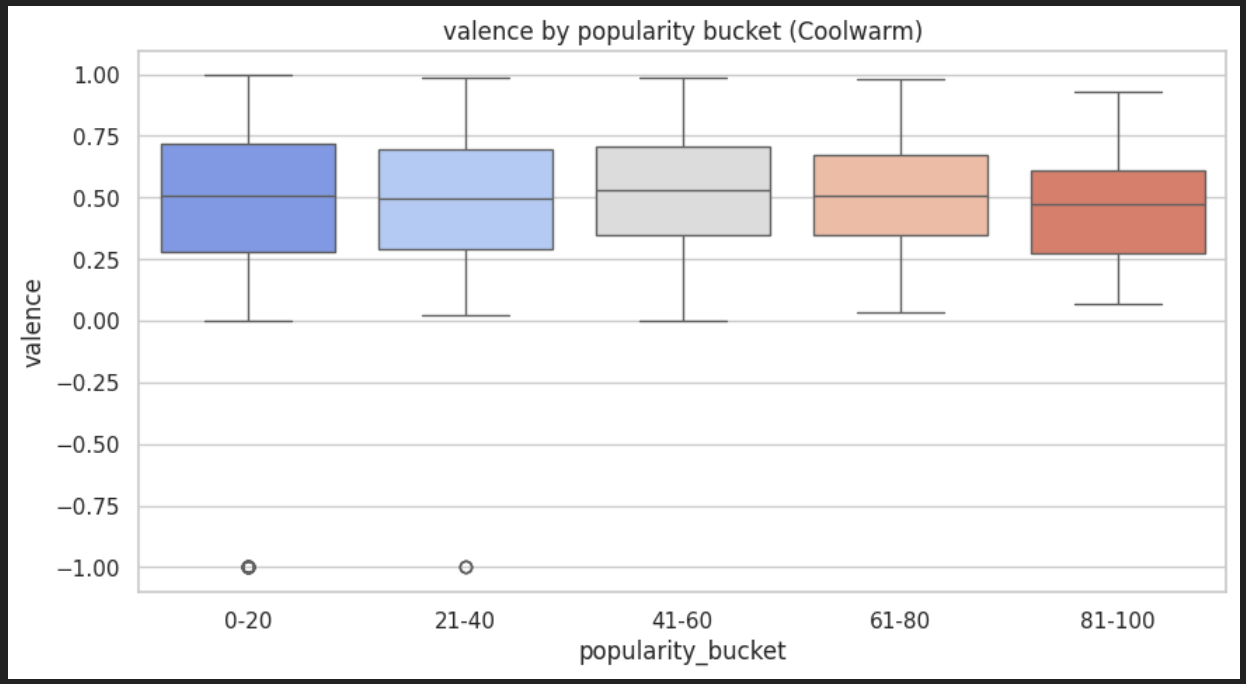
Business Insights

- Hit songs require high but controlled energy.
- The most popular songs are energetic, but not the **most intense tracks** in the catalog.
- An excess of energy may be fatiguing for mainstream listeners.

Actions & Improvements

- **Curate** "Top Hits" playlists with a high but not exhausting energy level.
- **Screen** for potential hits by prioritizing tracks with energy scores between 0.6 and 0.85.
- **Analyze** if specific genres are over-represented in the highest energy, lower popularity brackets.

VALENCE VS. POPULARITY



VALENCE VS. POPULARITY

Graph Analysis

- The median valence gently decreases as song popularity increases.
- This means the most popular songs are, on average, **less "happy"** than unpopular songs.
- The **range of valence also narrows** for the most popular tracks.

Business Insights

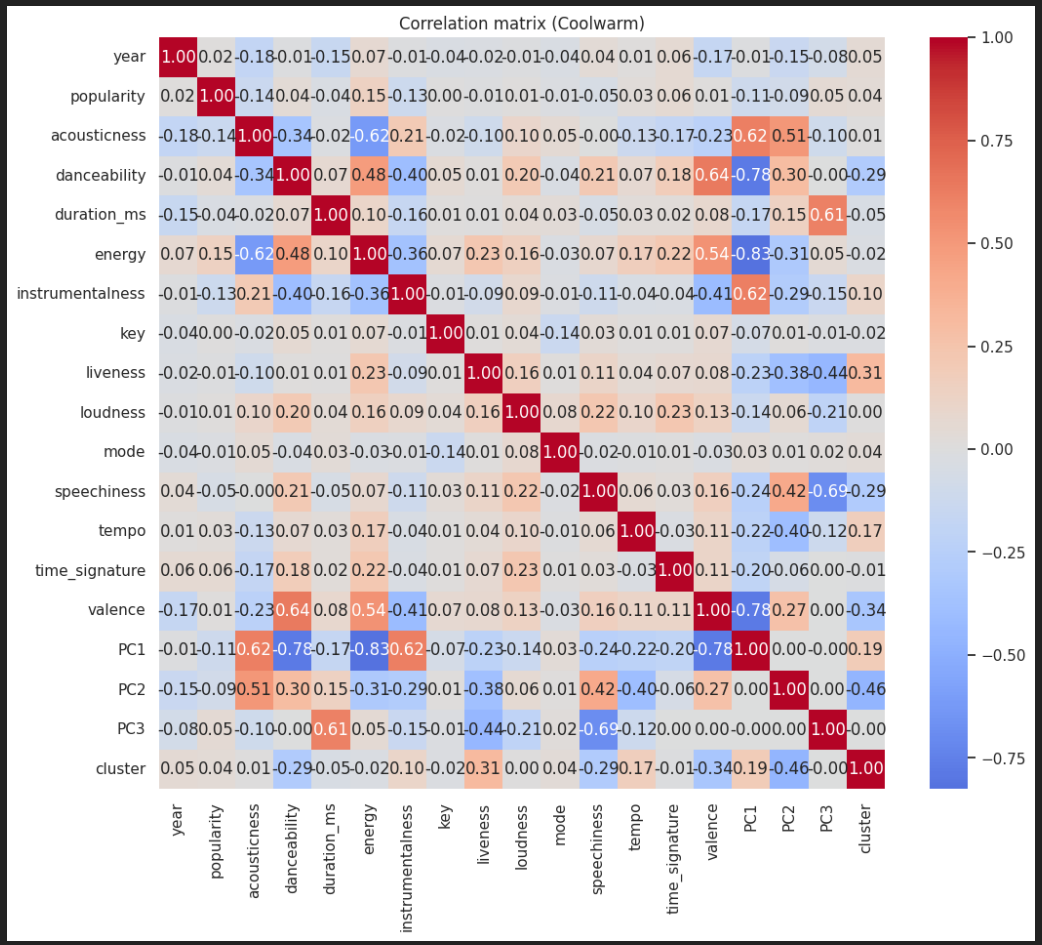
- The biggest hits are often emotionally complex or bittersweet, not purely cheerful.
- Listeners may find songs with a nuanced emotional tone more relatable.
- This challenges the assumption that "happy" music is always the most commercially viable.

Actions & Improvements

- **Advise** creative teams that songs with emotional depth have greater mass appeal.
- **Tune** the recommendation algorithm to not just push the happiest songs.
- **Focus** marketing campaigns on the relatable, emotional stories in the music.

MULTIVARIATE & MACHINE LEARNING ANALYSIS

FEATURE CORRELATION MATRIX



FEATURE CORRELATION MATRIX

Graph Analysis

- There are strong positive correlations between `energy`, `danceability`, and `valence`.
- `Acousticness` shows a strong negative correlation with `energy` (-0.62).
- `Popularity` is positively correlated with energy (0.48) and negatively with acousticness (-0.40).

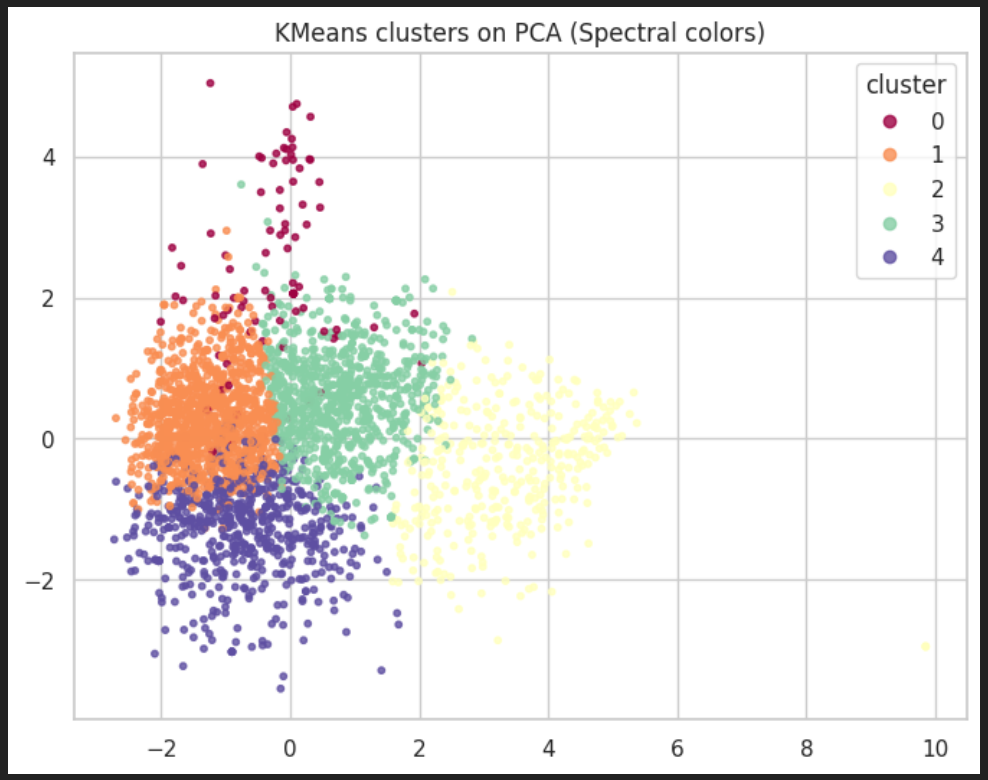
Business Insights

- A clear **"hit song formula"** emerges: high-energy, danceable, positive, and not acoustic.
- Acoustic music is confirmed to be a **niche category** that is anti-correlated with popularity.
- Many audio features measure a similar underlying concept of a song's **"upbeat intensity."**

Actions & Improvements

- **Engineer** a "Hit Score" for the recommendation engine by combining these key fe
- **Build** playlists that either follow or intentionally break this formula for variety.
- **Use** PCA to combine redundant features and create more powerful models.

UNSUPERVISED CLUSTERING



UNSUPERVISED CLUSTERING

Graph Analysis

- The K-Means algorithm has successfully grouped the songs into **5 distinct clusters**.
- Each cluster represents a group of songs with **similar audio features**, or a "sonic archetype."
- The clusters are **well-separated on the PCA plot**, proving the method is effective.

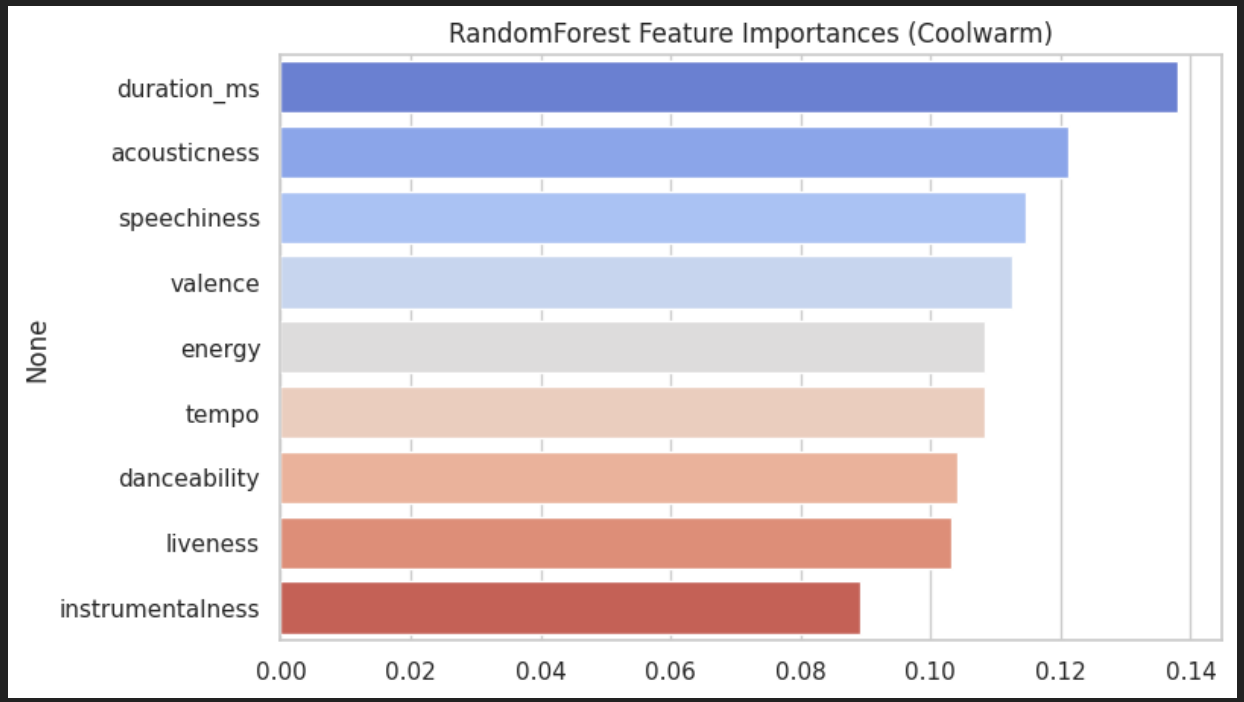
Business Insights

- This method provides unsupervised discovery of data-driven "genres" or "vibes."
- It is a powerful, objective way to **organize the music library** based on sound.
- These clusters can form the basis for a **highly personalized recommendation system**.

Actions & Improvements

- Analyze the audio profile of each cluster to give it a descriptive personality (e.g., "Cluster 2: Party Starters").
- Power recommendations by suggesting other songs from the same cluster a user is listening to.
- Generate new, unique playlists directly from these clusters.

FEATURE IMPORTANCE



FEATURE IMPORTANCE

Graph Analysis

- The model found **`duration_ms` (song length)** to be the most important predictive feature.
- **`Acousticness` and `speechiness`** are the second and third most important features.
- Core mood features like `energy` and `danceability` were found to be **less important** for this model.

Business Insights

- A song's most fundamental classifiers are its **length, acoustic nature, and lyrical content**.
- These features are better differentiators than mood because **mood features are often similar** across many songs.
- This tells us what the model "thinks" is most important for telling songs apart.

Actions & Improvements

- **Refine** user search tools to allow filtering by these highly important features.
- **Investigate** why `duration` is so dominant and consider binning it into categories.
- **Ensure** the recommendation engine weighs these key differentiators heavily when determining similarity.

CONCLUSION

Summaries

- The catalog's core identity is **mainstream, high-energy, and non-acoustic**, dominated by vocal tracks with an average length of 3-4 minutes.
- It possesses unique strategic depth with strong collections of **film scores, Indian regional music, and K-Pop**, giving it a competitive edge in specific markets.
- The dataset is compromised by significant **non-music content (audiobooks)** and has major data quality gaps, including thousands of tracks with an **"Unknown" language**.

Key Takeaways

- A clear **"Hit Formula"** exists: popular songs are consistently upbeat and produced, but have a **moderate, emotionally complex valence** rather than being purely "happy."
- The most popular songs occupy a narrow **"sweet spot"** of energy and danceability, indicating that mainstream success requires a balanced and predictable sound.
- **Unsupervised clustering** successfully identified distinct "sonic archetypes," proving that data-driven genre discovery is a powerful and effective tool for understanding the library's structure.

Preferred Actions

- **Clean the Data**: Immediately implement a process (using `speechiness`) to separate audiobooks from music and launch a project to classify all "Unknown" language tracks.
- **Enhance Personalization**: Implement the data-driven clusters into the recommendation engine to suggest songs based on true sonic similarity and create unique, curated playlists.
- **Fill Content Gaps**: Develop a strategic plan to acquire more **instrumental, acoustic, and mid-energy music** to attract and retain underserved user segments.


THANKS