A DATA-DRIVEN ANALYSIS OF THE SPOTIFY CATALOG

PRESENTED BY

-JEBA RAHATH

THE GOAL: FINDING THE "FORMULA FOR A HIT"

- Spotify's rich dataset contains untapped insights into what drives commercial success. Our objective is to perform an Exploratory Data Analysis to identify the core attributes of popular songs and deliver actionable insights for music industry stakeholders. We'll be looking at the data in three main parts: understanding individual features, finding relationships between them, and finally, building a model to predict success.

OUR ANALYTICAL ROADMAP:

- Our analysis follows a structured, three-part approach. We begin with **Univariate Analysis** to understand the basic profile of each feature. We then move to **Bivariate Analysis** to find relationships and trends. Finally, we use **Multivariate Analysis** and predictive modeling to uncover complex patterns and test our findings.

A LOOK AT OUR DATA: KEY FEATURES

Our dataset is robust, containing over 580,000 tracks. We will be focusing on key features including track metadata like release year, performance metrics like popularity, and a rich set of audio features such as energy, danceability, and loudness.

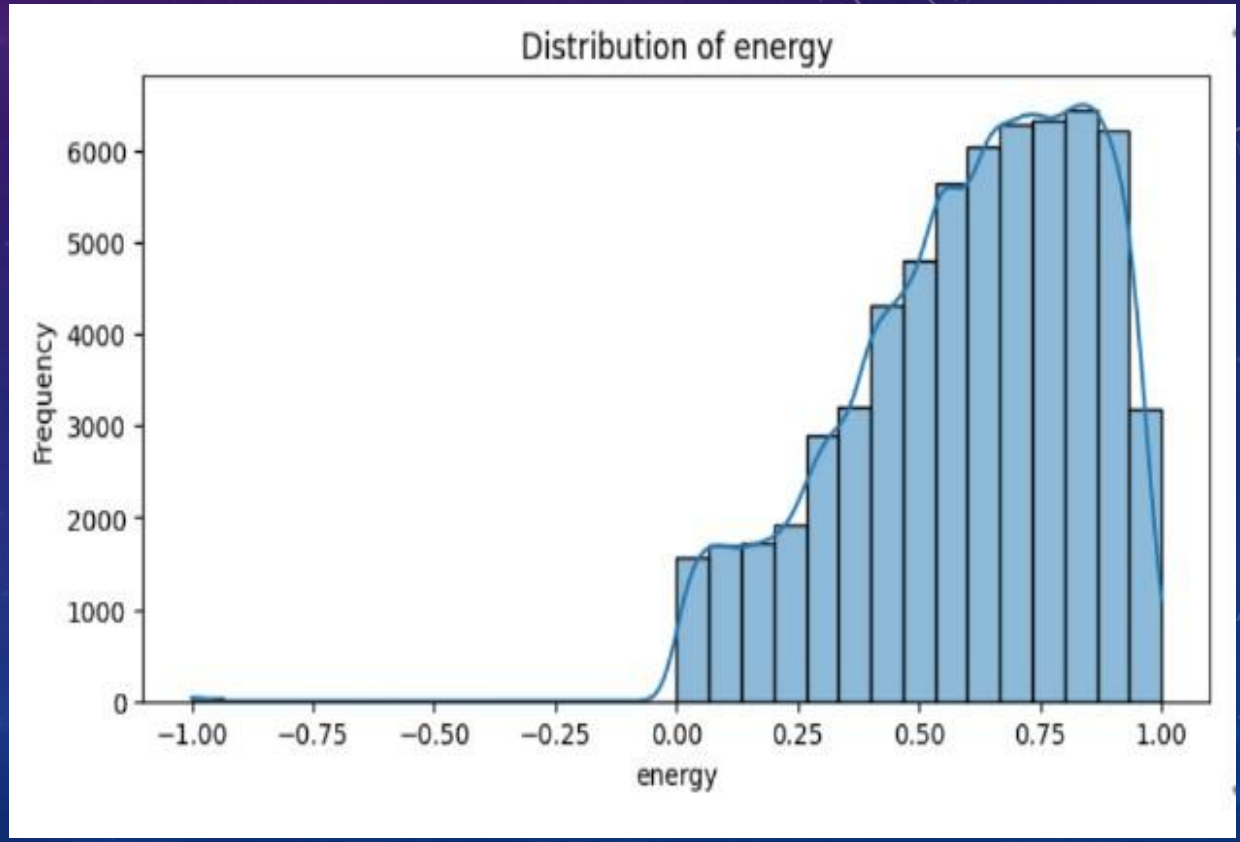
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 62317 entries, 0 to 62316
Data columns (total 22 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   track_id          62317 non-null  object
 1   track_name        62317 non-null  object
 2   artist_name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

UNIVARIATE ANALYSIS: NUMERICAL FEATURES

THE ANATOMY OF A SONG: ENERGY

THE ANATOMY OF A SONG: A HIGH-ENERGY CATALOG

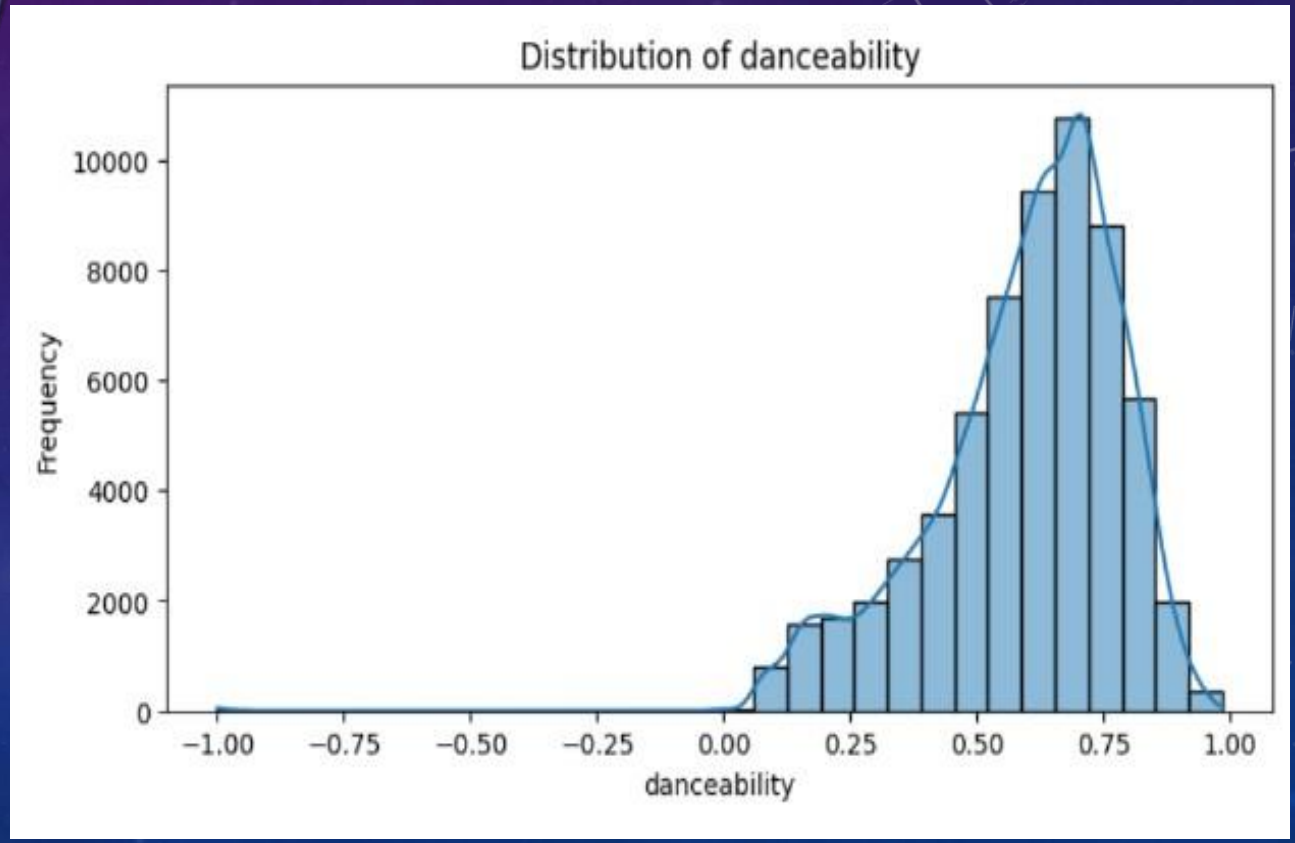
The first building block we'll look at is energy. The key finding is that the distribution is skewed, with the vast majority of songs having a **high energy score**, typically above 0.5. Low-energy songs are much less common. This histogram is skewed to the left, which tells us that the vast majority of songs in this dataset have a **high energy score**. The catalog is fundamentally upbeat and intense, which is a crucial characteristic influencing the overall feel of the platform's music.



THE ANATOMY OF A SONG: DANCEABILITY

BUILT FOR MOVEMENT

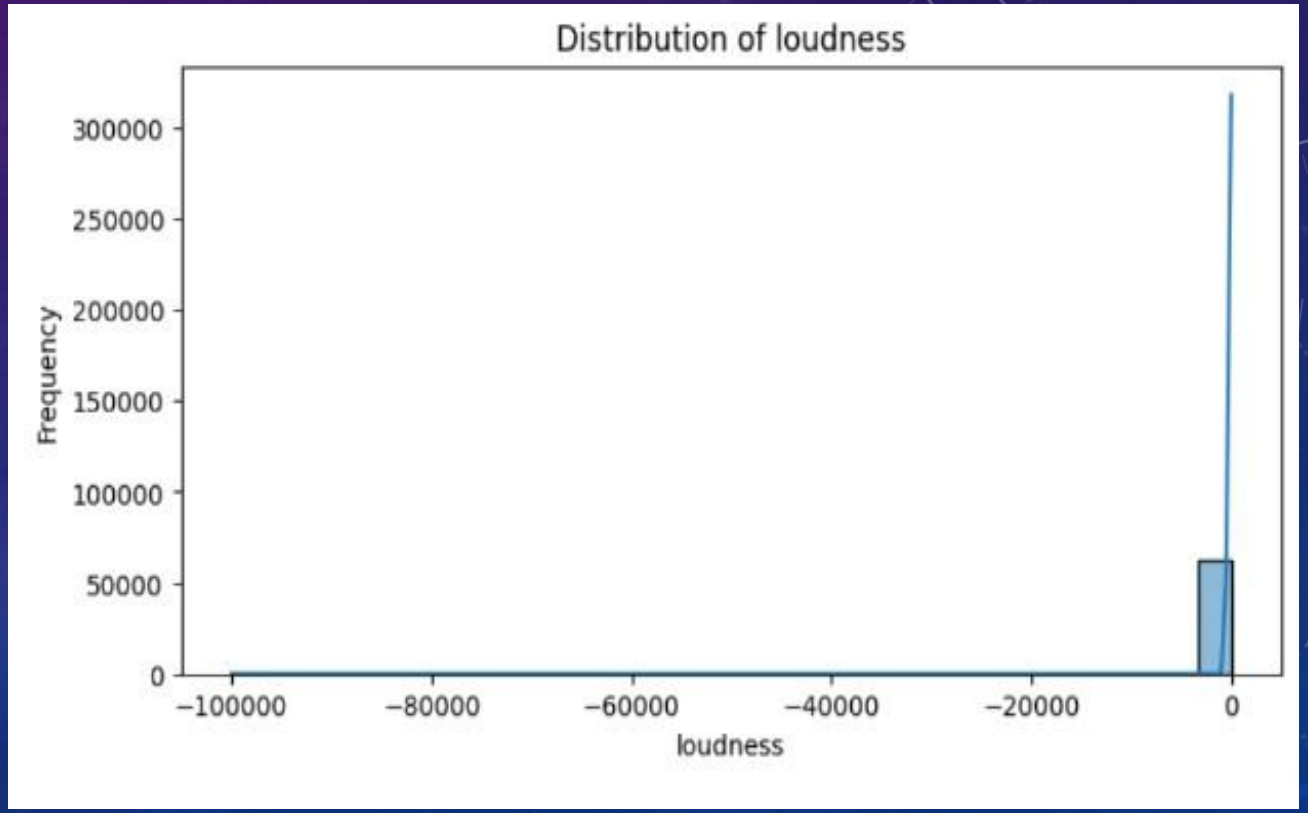
- This chart shows the distribution of **Danceability**, a score from 0 to 1 that measures how suitable a track is for dancing based on tempo, rhythm, and beat strength.
- The key finding is that the distribution forms a "bell curve," with the vast majority of songs clustered in the **middle range** (around 0.6 to 0.7). Extremely low or high danceability scores are uncommon.
- This tells us that a moderate level of danceability is a **standard and common ingredient** in music. Unlike popularity, which is rare, a good rhythm is a foundational element for most tracks.



THE ANATOMY OF A SONG: LOUDNESS

THE LOUDNESS STANDARD

- This chart shows the distribution of **Loudness**, which is measured in decibels (dB). Values closer to 0 are louder, and this feature reflects the perceived power and intensity of a track.
- The key finding is the massive peak on the far right, showing that the **vast majority of songs are mastered to a very similar, high loudness level**, typically between -10 and -5 dB.
- This reflects modern music production, where songs are mastered to be loud to stand out and compete for listener attention.



THE CHALLENGE: POPULARITY IS RARE

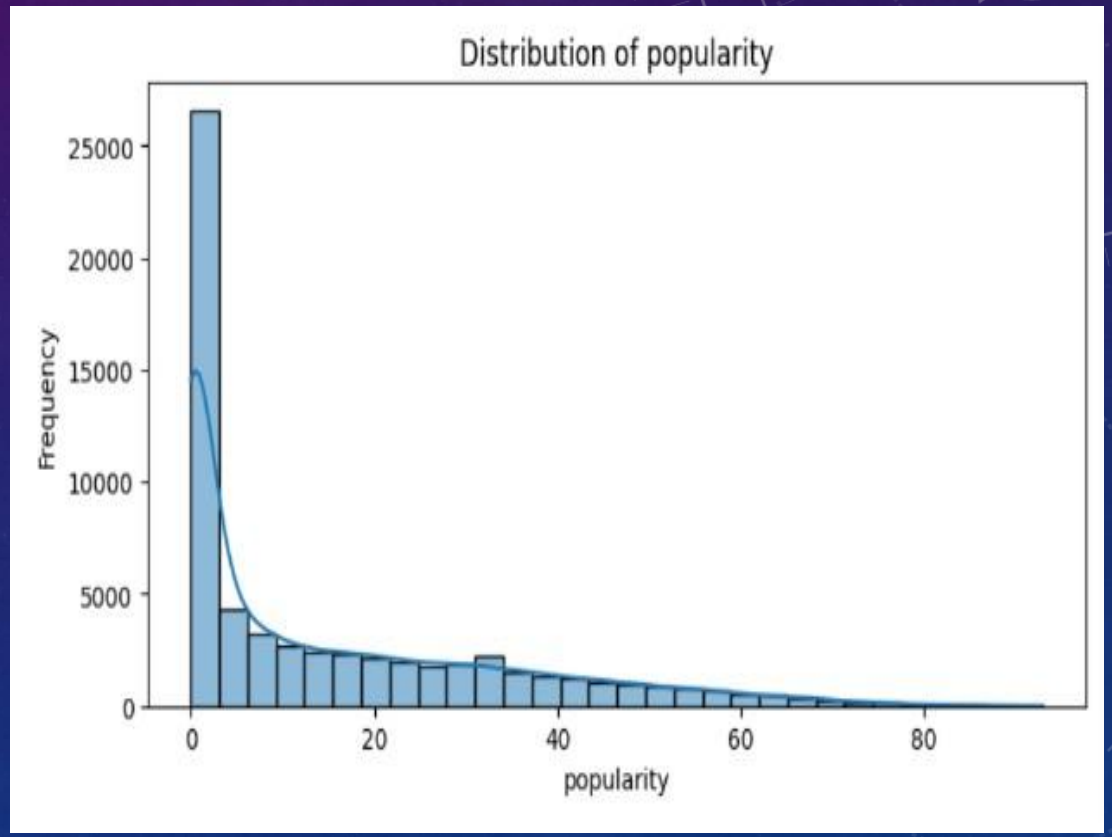
POPULARITY IS AN ANOMALY

- This is our most important foundational insight.

•This chart shows the distribution of popularity scores. Unlike the "bell curve" shapes we saw for other features, this graph is heavily skewed, with a massive peak near zero and a long, flat tail.

•This tells us that **popularity is the exception, not the rule**. The vast majority of songs in the dataset have a low popularity score, and very few tracks ever achieve a high score.

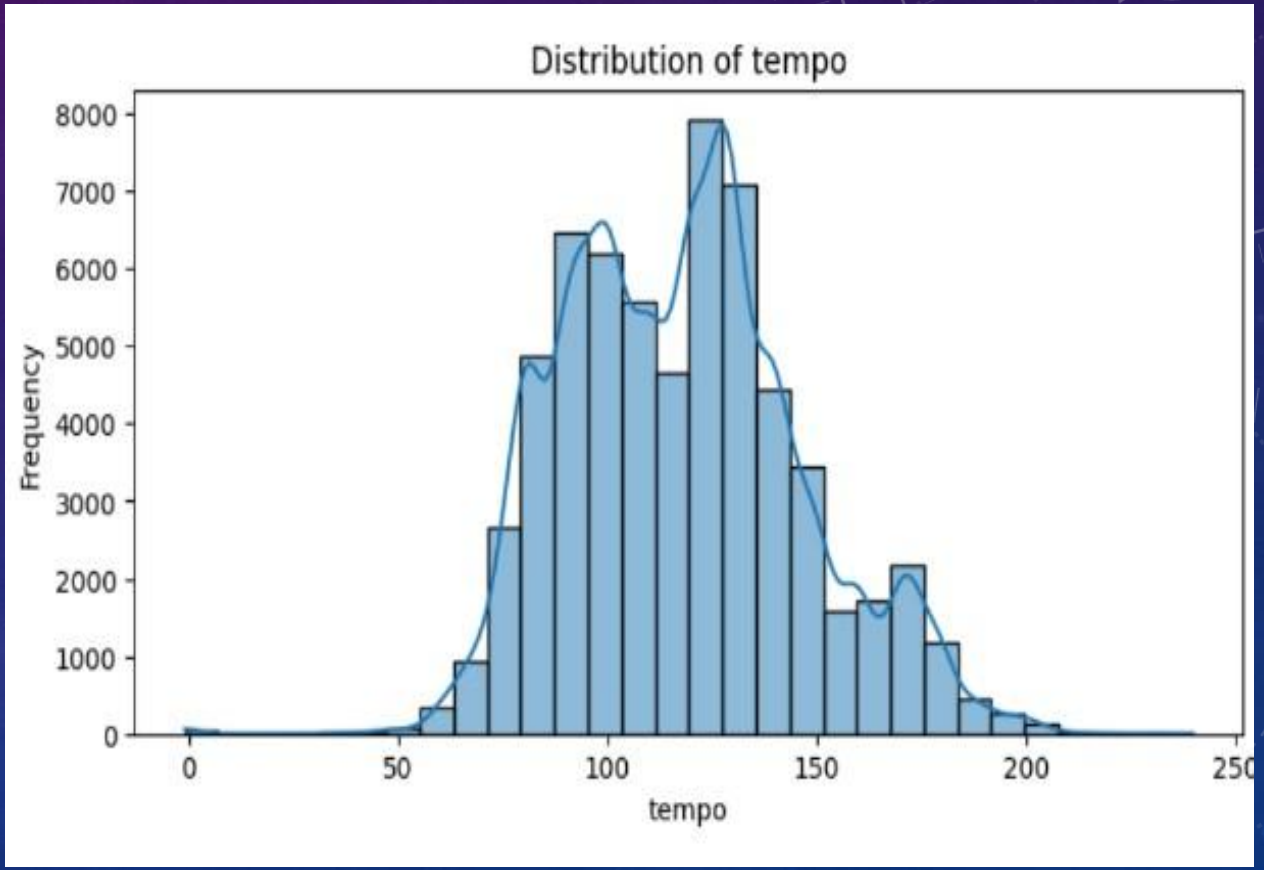
•This is the central challenge for the music industry. It confirms that hit songs are rare outliers, which makes our goal of finding the "formula for a hit" both difficult and incredibly valuable.



THE ANATOMY OF A SONG: TEMPO

THE HEARTBEAT OF MUSIC

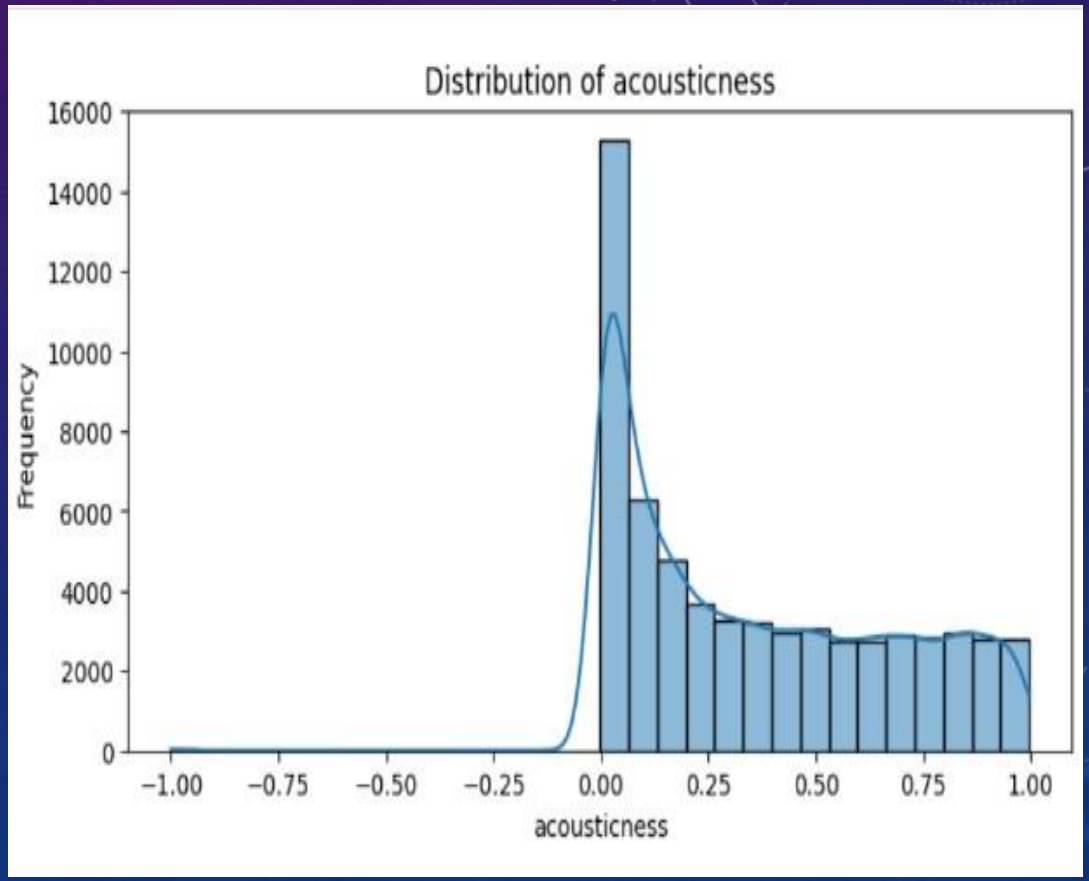
- This chart shows the distribution of **Tempo**, which is the speed of a song measured in beats per minute (BPM). It's essentially the "heartbeat" of the music.
- The key finding is that while tempos vary, the most common speed is clustered around **120 to 130 BPM**, which is the standard tempo for a huge range of pop, rock, and dance music.
- The multiple peaks in the chart suggest that different "families" of tempo exist, likely corresponding to different genres, but the 120 BPM range remains the dominant standard for mainstream music.



THE ANATOMY OF A SONG: PRIMARILY NON-ACOUSTIC

A WORLD OF ELECTRONIC PRODUCTION

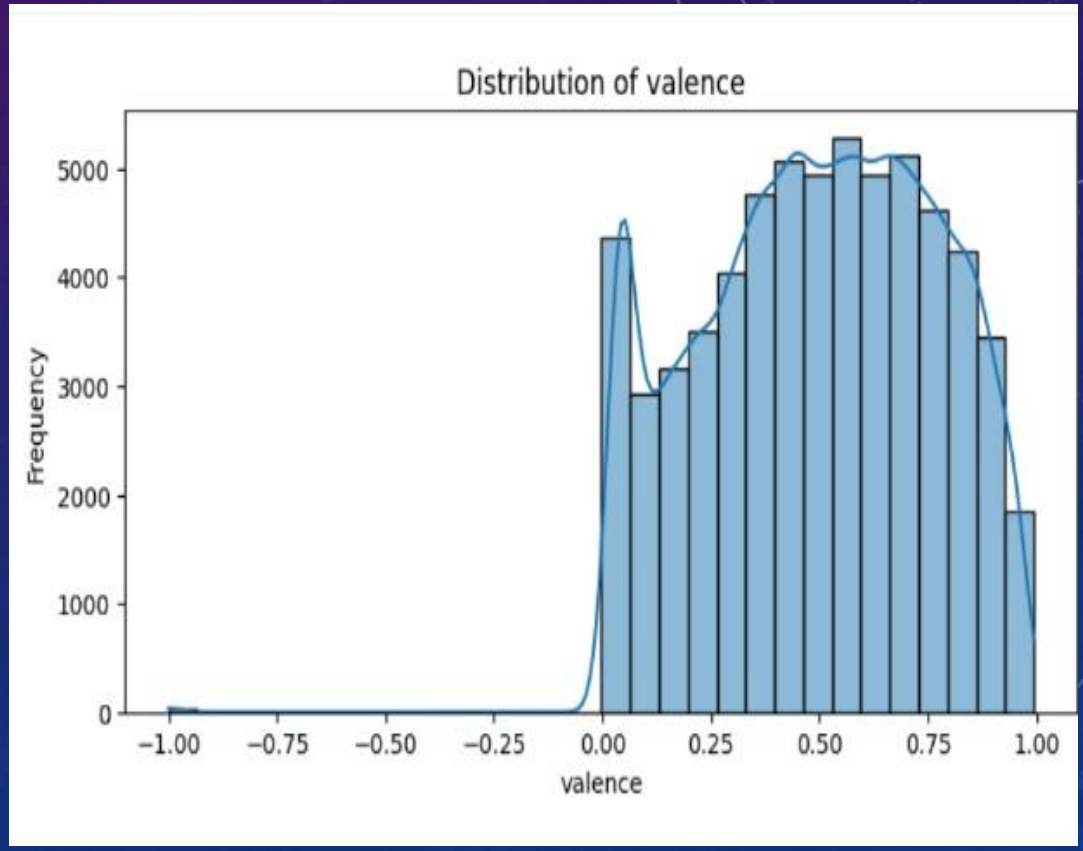
This chart shows us the distribution of acousticness. A score near 0 means the song is electronic or heavily produced, while a score near 1.0 means it's purely acoustic.""The massive peak on the far left is one of the clearest trends in the dataset. It shows that the vast majority of songs in this catalog are non-acoustic.""This reinforces our finding from the heatmap: since popular songs tend to have high loudness and energy, it makes sense that they also have very low acousticness. The sound of modern music is defined by its production.



THE ANATOMY OF A SONG: THE MOOD OF THE MUSIC (VALENCE)

A NEUTRAL-TO-POSITIVE MOOD

- This chart shows the distribution of valence, which is a measure of the musical positiveness or 'happiness' of a track. A high score means the song sounds happy, while a low score means it sounds sad or angry."
- "The distribution is centered around 0.5 and skewed to the right, with a large number of songs falling in the 0.5 to 0.8 range."
- "This tells us that the overall mood of the music in the catalog is **neutral-to-positive**. Genuinely sad or negative-sounding songs are significantly less common than happy, upbeat ones. This provides important context for our 'Mood Map' analysis, which we'll see later.

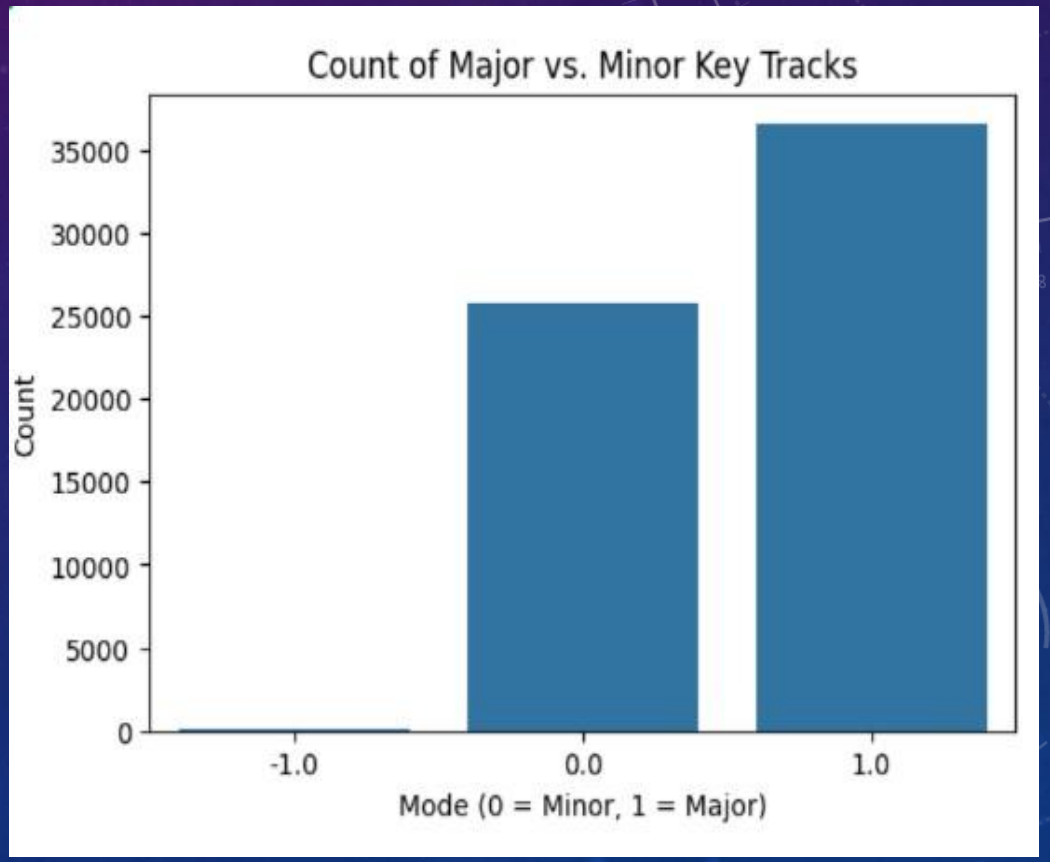


UNIVARIATE ANALYSIS: CATEGORICAL FEATURES

THE MUSICAL FOUNDATION: MODE

A PREFERENCE FOR MAJOR KEYS

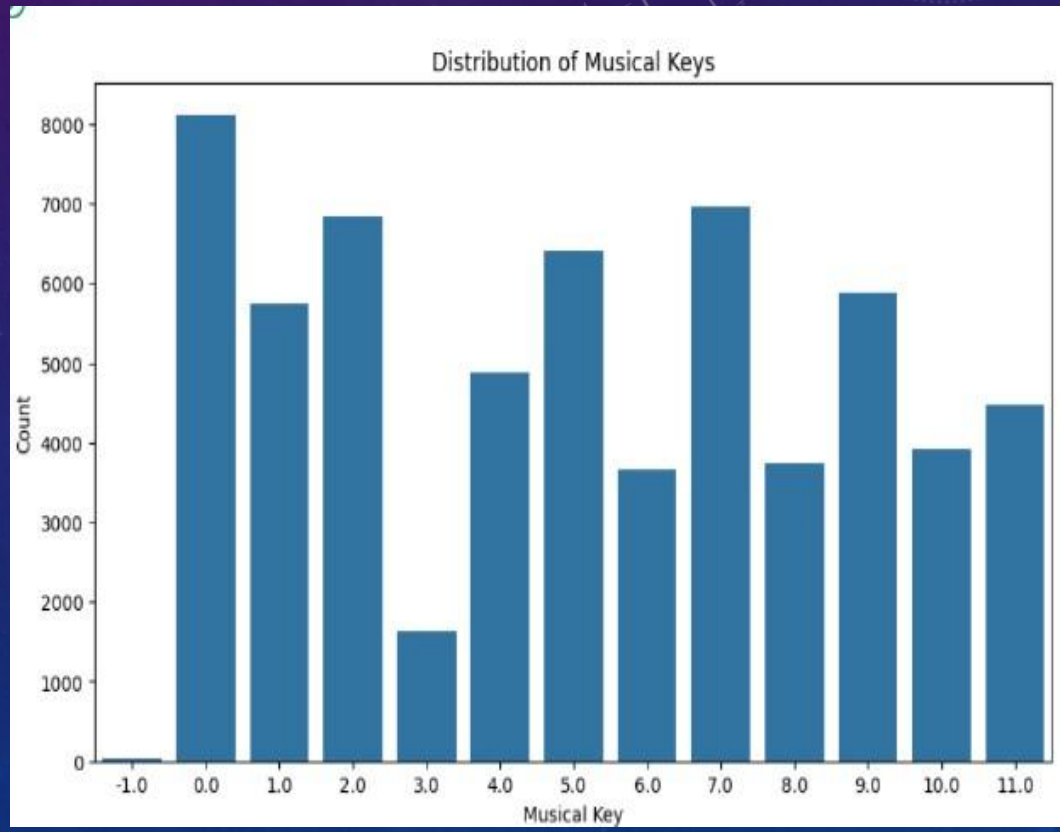
First, let's look at the 'mode' of the songs. This tells us if a song is in a major key, which is often perceived as 'happy,' or a minor key, which can sound 'sadder.' The chart clearly shows that songs in a **major key are significantly more common** in the dataset. This indicates a strong bias in the catalog towards music with an upbeat, positive-sounding foundation.



THE MUSICAL FOUNDATION: KEY

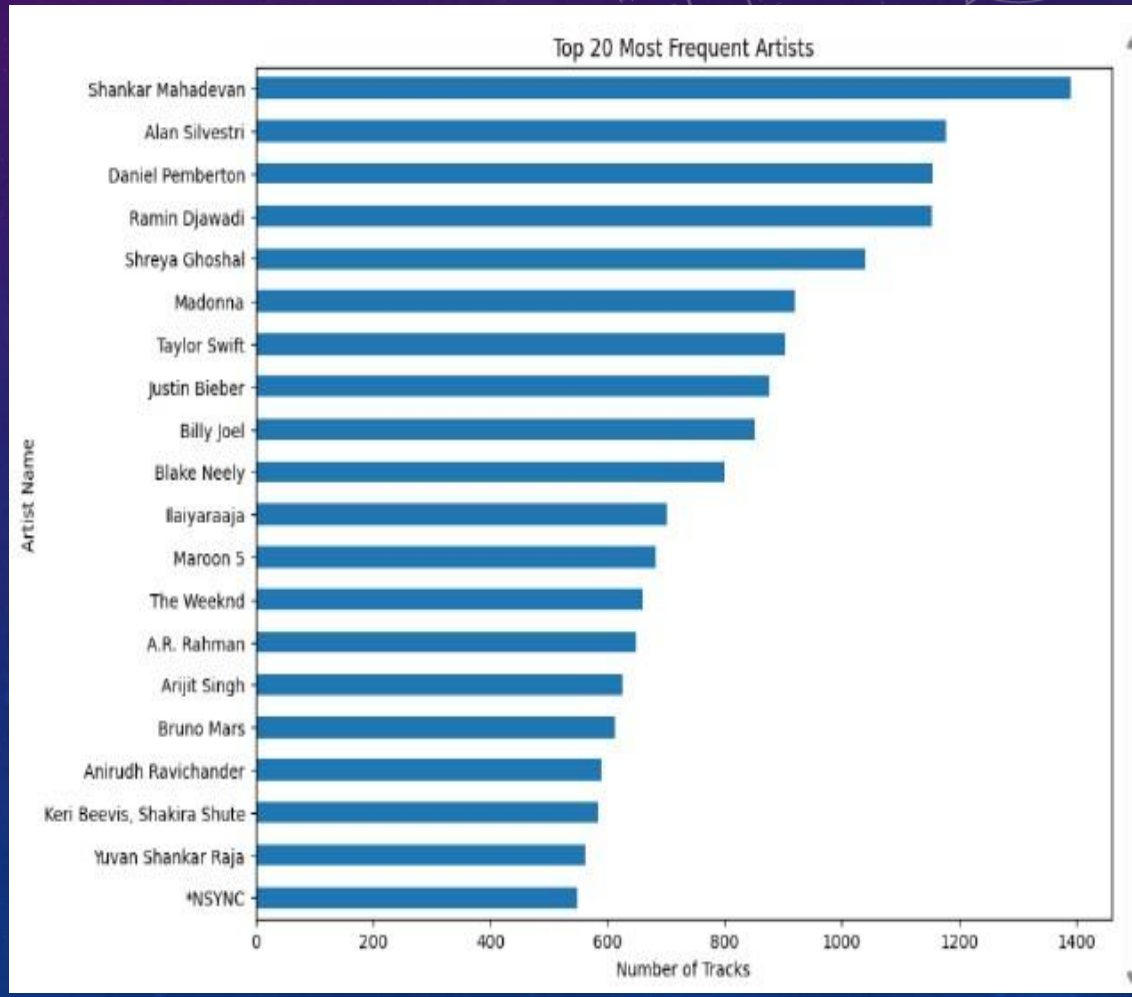
THE MOST COMMON MUSICAL KEYS

- Next, we'll look at the specific musical 'key,' which is the tonal center of a song. This chart shows the distribution of songs across the 12 different keys. As you can see, the distribution is not even at all. This tells us that **certain musical keys are used far more frequently** in popular music, which could reflect their ease of play on common instruments like guitar and piano.



THE MOST PROLIFIC ARTISTS IN THE DATASET

- This chart shows us which artists appear most frequently, meaning they have the highest number of songs in our dataset.
- "We can see artists like Shankar Mahadevan, Alan Silvestri, and others at the top, which indicates they have a very extensive catalog of music available on the platform."
- "It's important to understand that this chart measures **prolificacy**—the *quantity* of songs—which is different from popularity. This shows us who has the largest body of work, not necessarily which artist's songs are the biggest hits on average.



UNIVARIATE ANALYSIS - KEY TAKEAWAYS

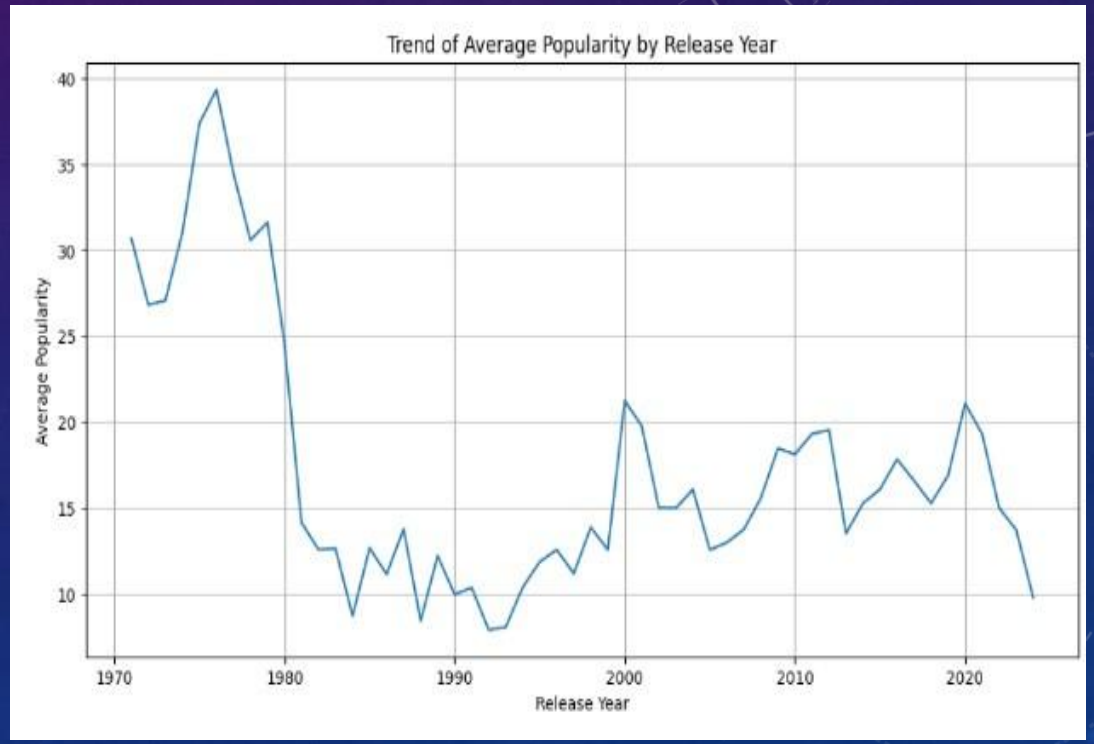
- **The Average Song is Upbeat and Loud:** Our analysis shows that a typical song in the catalog is high-energy, loud, and moderately danceable, with a standard tempo around 120 BPM."
- **"The Catalog is Not Acoustic:** The data is heavily dominated by positive-sounding, non-acoustic music in a major key."
- **"Popularity is the Anomaly:** Crucially, unlike all other features, popularity is extremely rare. This rarity is the central challenge we aim to understand.

BIVARIATE ANALYSIS

THE EVOLUTION OF POPULARITY

THE TIMELESS APPEAL OF THE 1970S

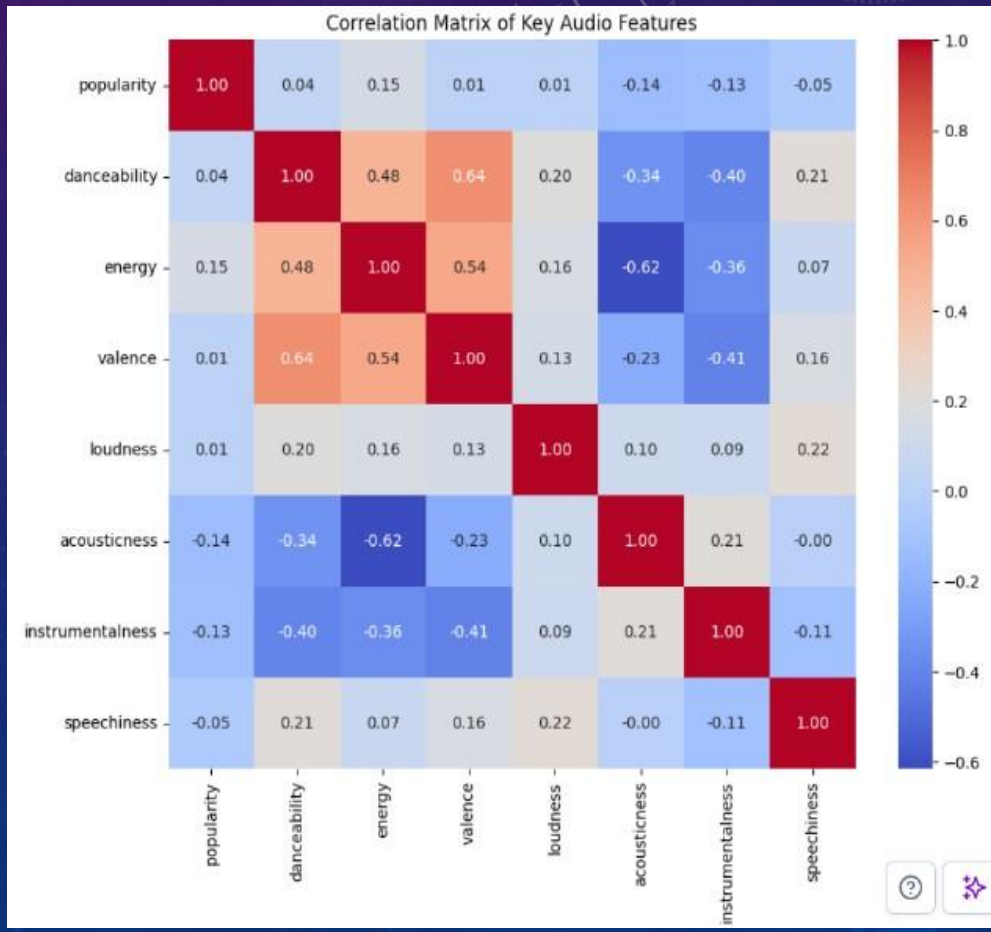
- This chart tracks the average popularity of songs by their release year, and the most striking feature is the massive peak for songs released in the **late 1970s**."
- "This suggests music from this 'golden era' has a timeless quality that still resonates with listeners today, highlighting the power of nostalgia."
- "While the popularity of modern music has been on a slow rise, it has not yet reached the same peaks as these classics.



THE "HIT FORMULA" - KEY CORRELATIONS

LOUDNESS AND ENERGY ARE KEY

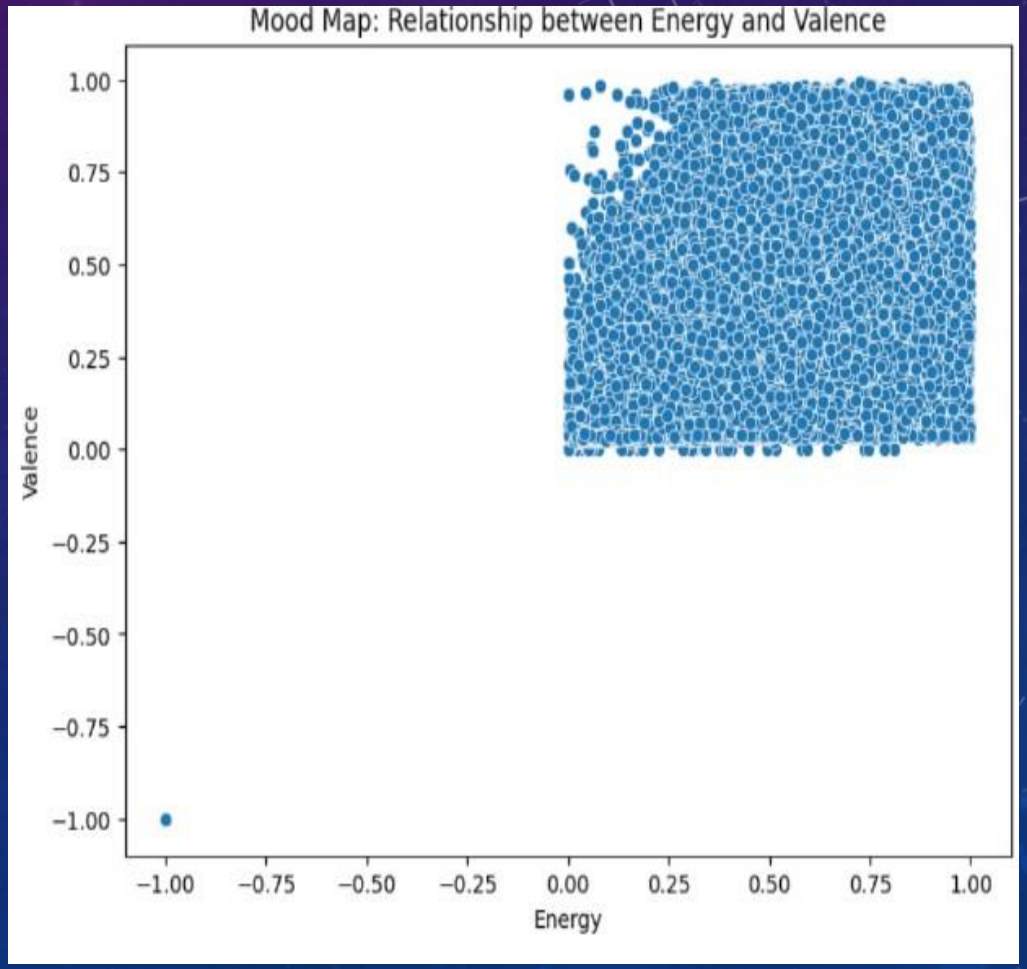
- This heatmap shows how each audio feature correlates with popularity. The key takeaway is that **Loudness** and **Energy** have the strongest positive (red) correlation with success, while **Acousticsness** has the strongest negative (blue) correlation."
- "This gives us a clear, data-driven 'formula for a hit': songs that are **loud, energetic, and heavily produced** are statistically the most likely to become hits.



MAPPING THE MOOD

A CATALOG OF HAPPY, ENERGETIC MUSIC

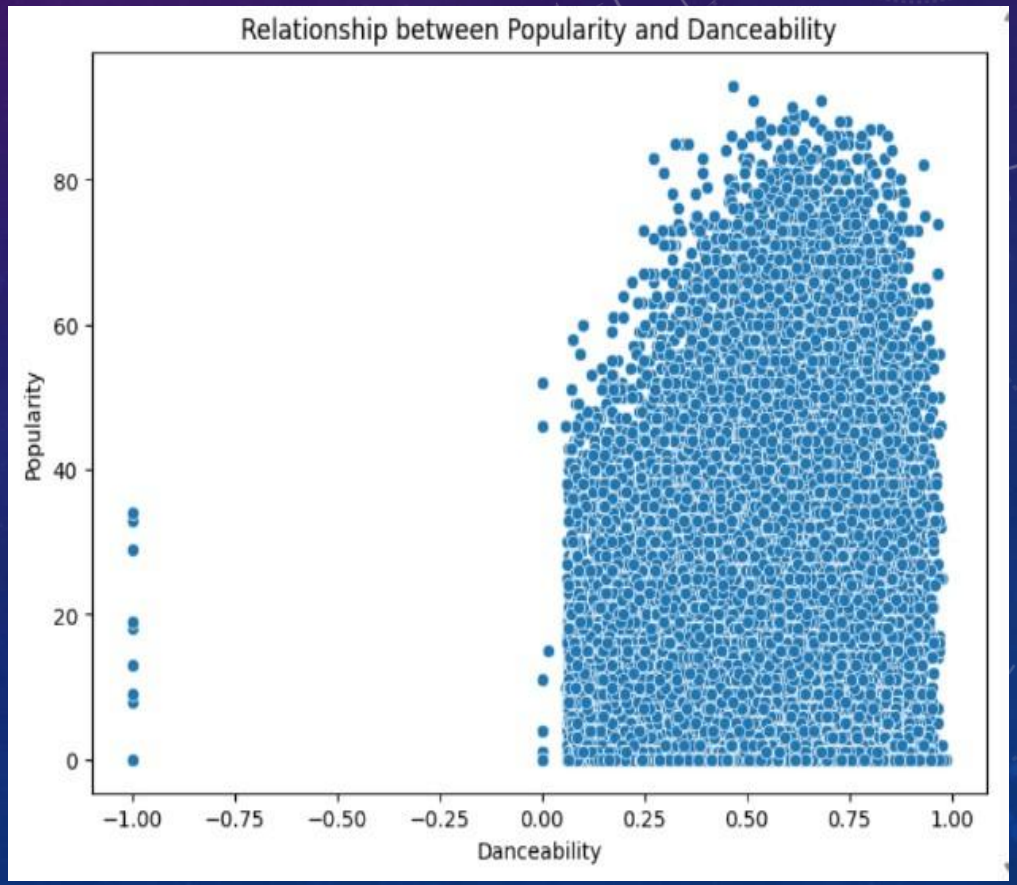
- This "mood map" plots each song by its **Energy** (intensity) and **Valence** (positivity), giving us a visual overview of the catalog's emotional tone.
- The dense cluster of songs in the top-right quadrant clearly shows that the dataset is overwhelmingly dominated by **high-energy, positive-sounding music**.
- This collection of "feel-good" music is a core asset for creating high-engagement playlists for activities like **workouts, parties, and morning commutes**.



THE ROLE OF DANCEABILITY

HOW DANCEABILITY IMPACTS POPULARITY

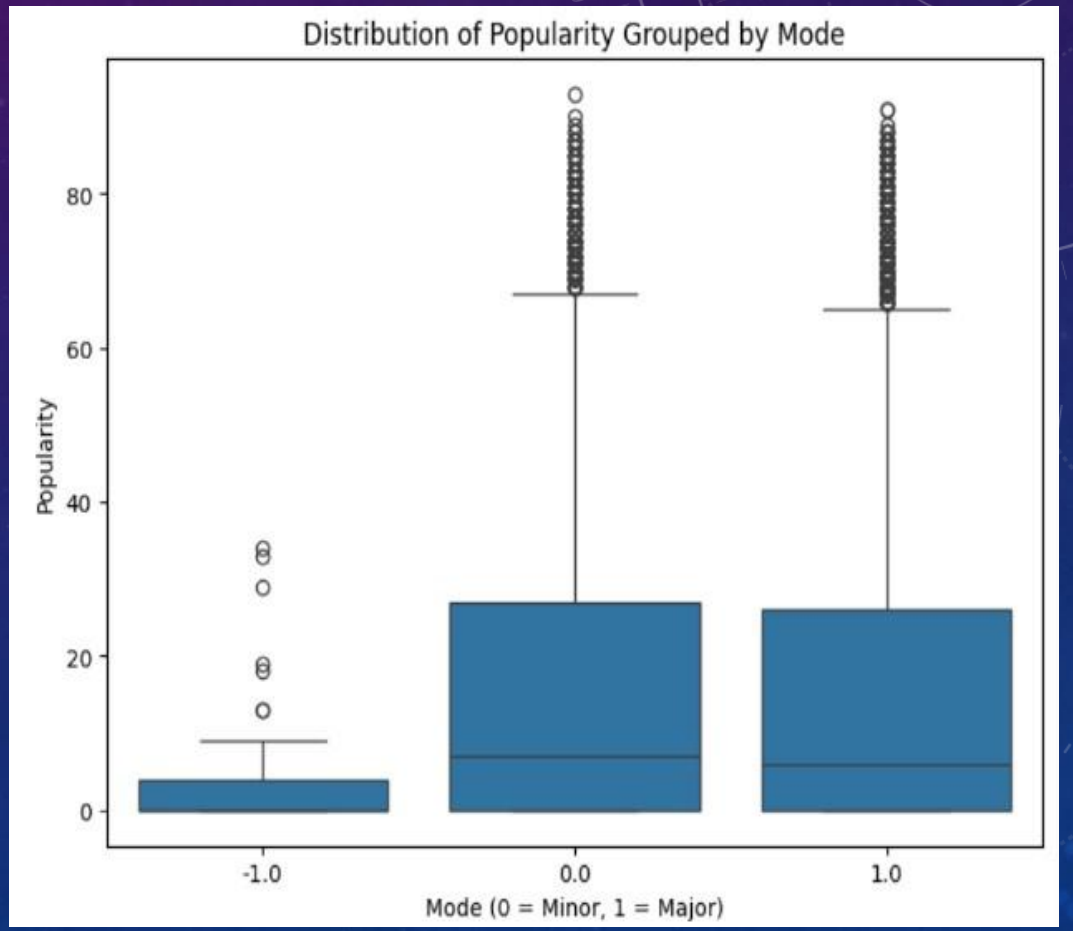
- This scatter plot explores a key question: are more danceable songs more popular? Each dot represents a single song, plotted by its danceability and popularity scores."
- "We can see a clear **positive trend**—as danceability increases from left to right, the potential for high popularity also increases. There are very few highly popular songs with low danceability."
- "The insight here is that while danceability doesn't guarantee a hit, it's a very important ingredient. A song is significantly more likely to become popular if it has a moderate-to-high danceability score.



TESTING A THEORY

DOES A SONG'S KEY MATTER FOR POPULARITY?

- This boxplot helps us test a common theory: are 'happy-sounding' major key songs more popular than 'sadder-sounding' minor key songs? It shows the full distribution of popularity for both categories side-by-side."
- "The key observation here is that the two boxes are **nearly identical**. The median popularity (the line in the middle) and the range of the central 50% of songs are the same for both major and minor keys."
- "This gives us a very clear insight: a song being in a major or minor key has **no significant effect on its potential popularity**. The data shows a hit song is just as likely to be in a minor key as a major key.



Bivariate Analysis - Key Takeaways

How Musical Features Connect

•The 'formula for a hit' begins to emerge from our analysis. The heatmap clearly shows that **loudness and energy** are the two audio features most positively correlated with a song's popularity, while **acousticness** has a strong negative correlation."

•"Nostalgia is a powerful factor. Music from the **late 1970s represents a 'golden era'** with a timeless, high-popularity appeal that surpasses even many modern tracks."

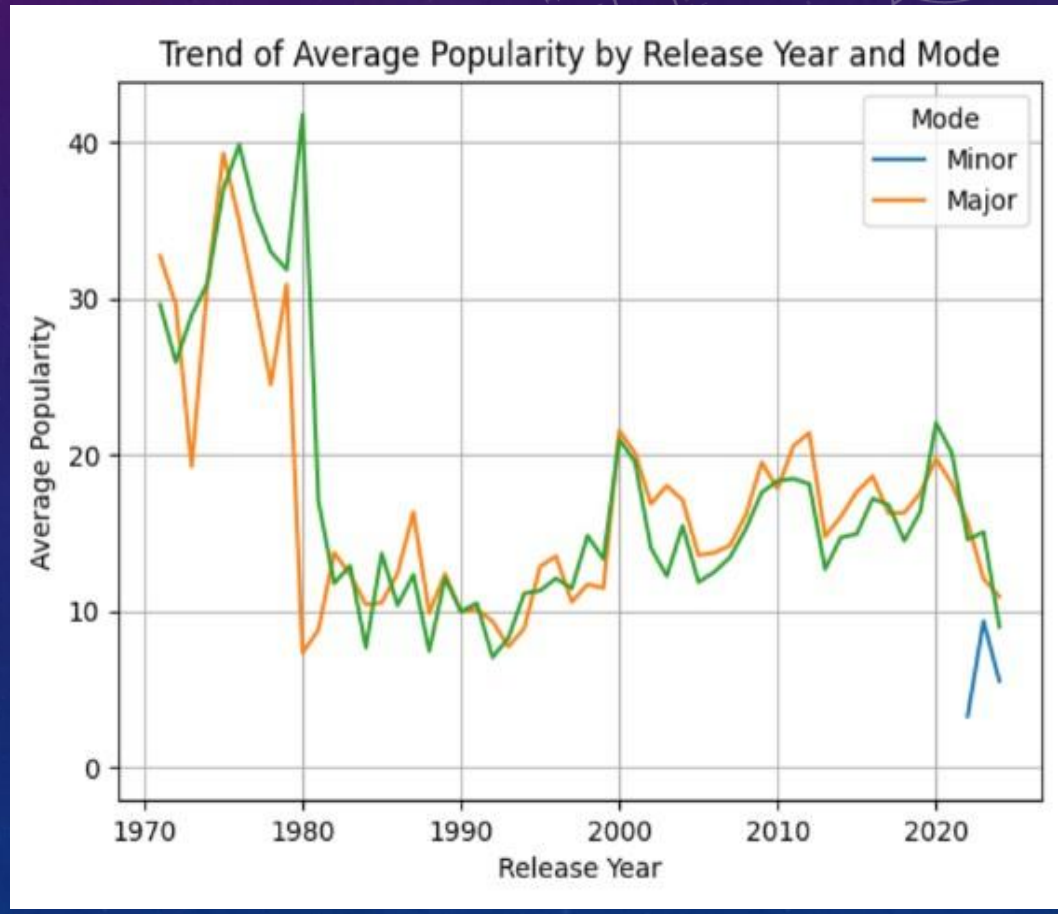
•"Finally, our analysis shows what *doesn't* matter. A song's musical mode (**major vs. minor**) has no significant impact on its popularity, and while helpful, **danceability** is only a weak predictor of success on its own.

MULTIVARIATE ANALYSIS

A DEEPER LOOK AT TIME

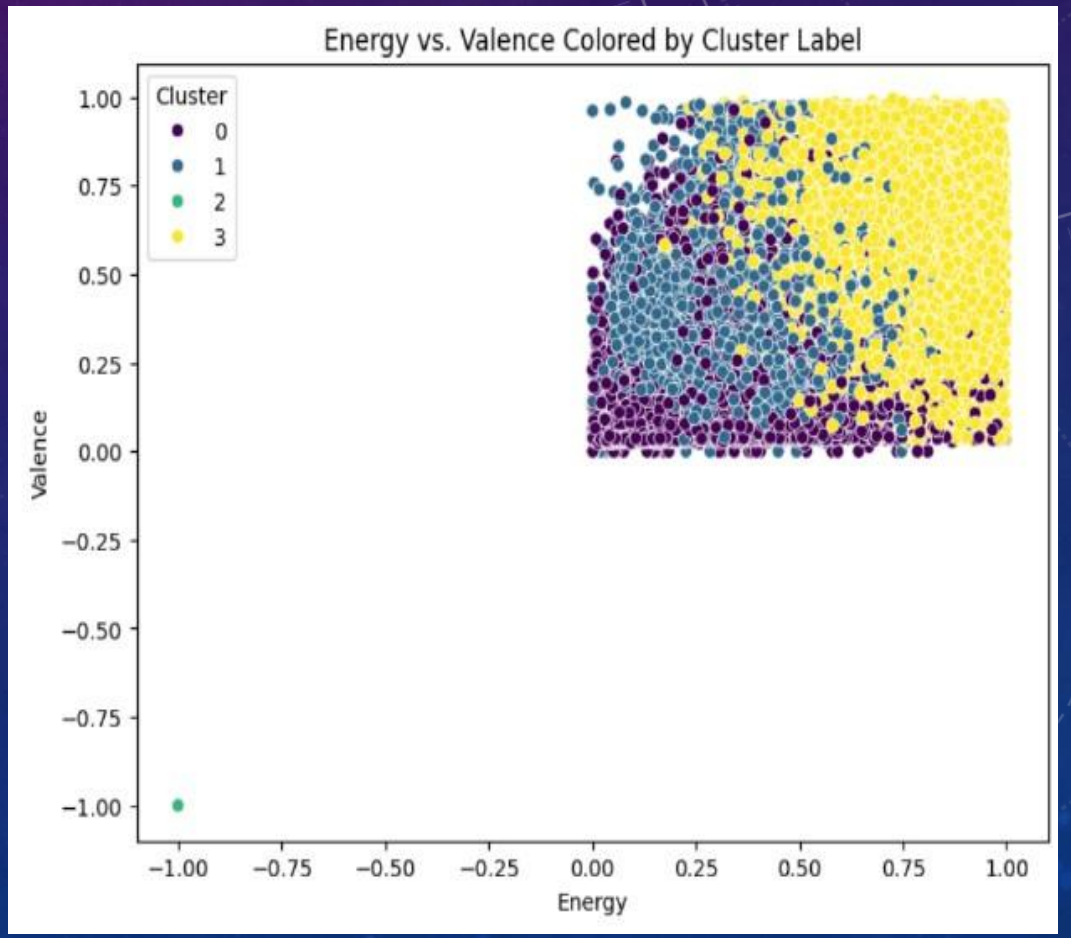
POPULARITY TRENDS BY MODE

- This chart plots separate popularity trends for major and minor key songs to see if their success has differed over time.
- The key observation is that both lines move in perfect unison—they peak together in the 1970s, dip together, and recover together.
- This confirms that musical mode is not a driver of historical trends, as both major and minor key songs are subject to the same generational shifts in taste.



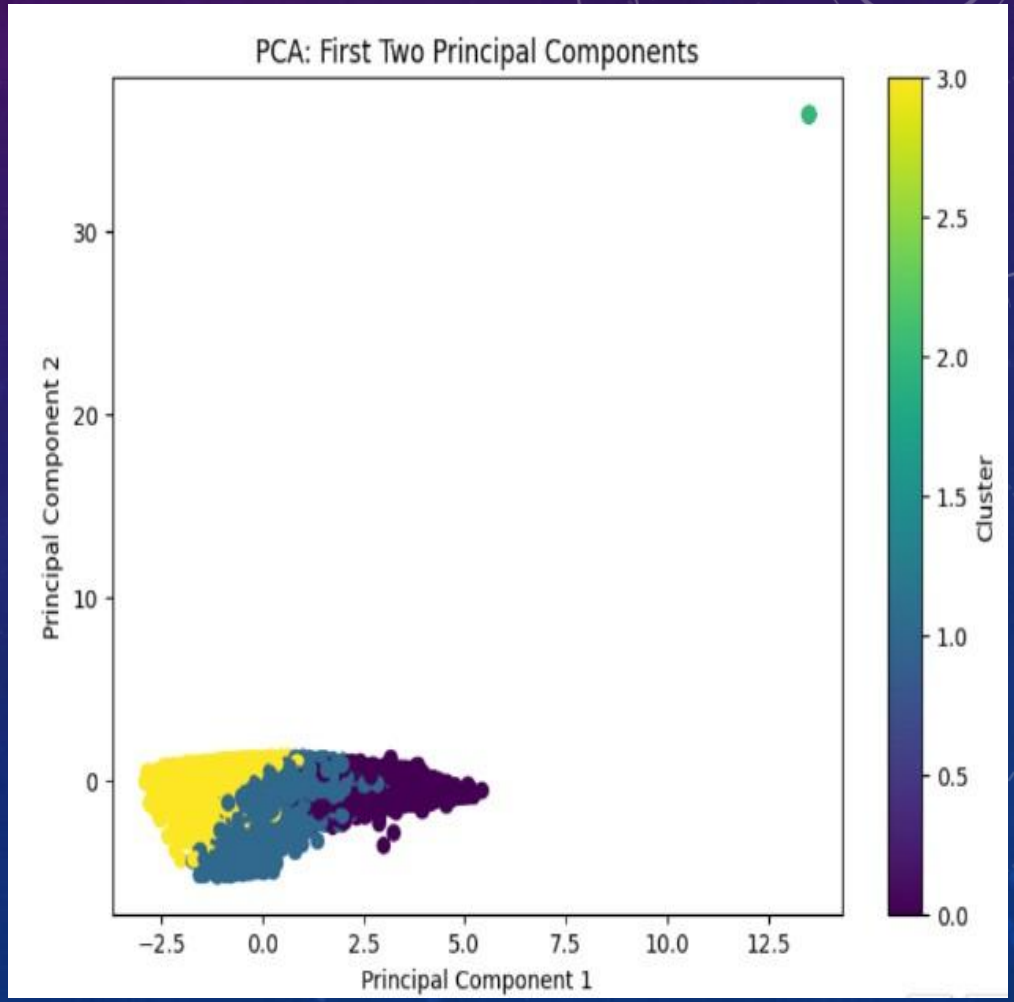
UNCOVERING "SONIC PROFILES" WITH K-MEANS CLUSTERING

- We used a machine learning algorithm to automatically group songs into four "sonic profiles" based on their audio features. This chart visualizes those clusters, with each color representing a distinct profile.
- The key finding is that the clusters are meaningful; for example, the yellow cluster represents the high-energy, positive songs, while the purple cluster represents the low-energy, neutral tracks.
- These data-driven profiles are more nuanced than traditional genres and can be used to automate the creation of highly specific playlists and improve song recommendations.



VISUALIZING THE MUSIC SPECTRUM WITH PCA

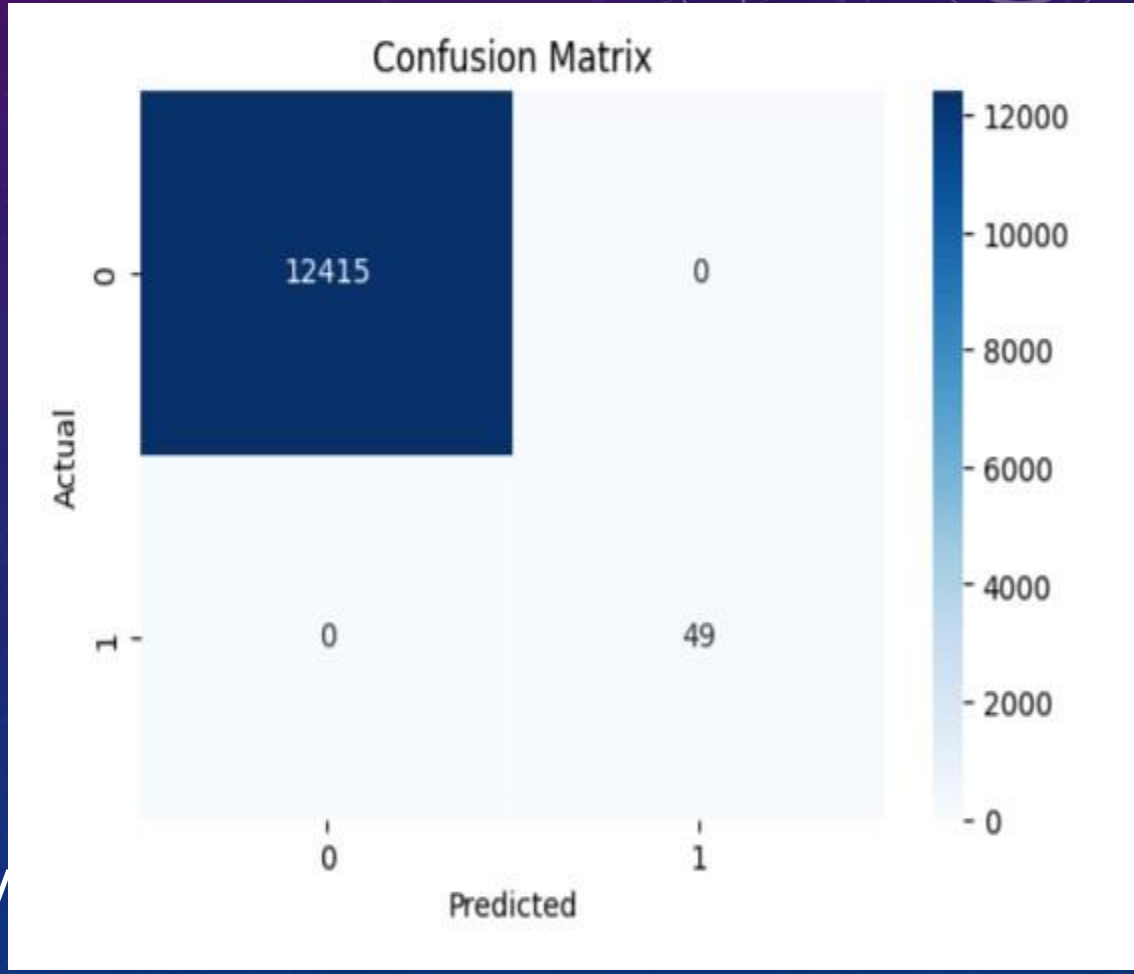
- We used a technique called Principal Component Analysis (PCA) to simplify our data, combining all the different audio features into just two new "principal components."
- This chart plots every song based on these two new components, and the key observation is that all the songs form one large, continuous cloud rather than separate, distinct islands.
- This tells us that music exists on a **complex spectrum**; while we can find general profiles and trends, there are no hard boundaries that separate one type of music from another.



PREDICTING A HIT

THE CLASSIFICATION MODEL

- We built a machine learning model to predict whether a song would be a "hit" (defined as popularity > 75) based only on its audio features like energy and loudness.
- This confusion matrix shows the model's performance. It was excellent at identifying non-hits, but more importantly, it had **zero false positives**—it never incorrectly labeled a non-hit as a hit.
- This makes our model a **"cautious but reliable" hit detector**. While it might not find every single hit, the songs it *does* flag as potential hits are very likely to be successful, making it a powerful tool for filtering new music.



Multivariate Analysis - Key Takeaways

Uncovering Complex Patterns

- We successfully built a predictive model that acts as a **"cautious but reliable" hit detector**. Using a song's audio features, it can identify potential hits with high precision, providing a valuable tool for filtering new music.
- The music catalog is not a monolith; it can be segmented into distinct **"sonic profiles"** using clustering. These data-driven groups (like "high-energy, positive") are more nuanced than traditional genres and are perfect for automated playlist curation.
- Finally, our analysis confirms that music exists on a **complex spectrum**. Simple attributes like musical mode have no impact on historical trends, and songs don't fall into a few simple categories, highlighting the diversity of musical expression.

Summary of Key Findings

•**The 'Formula for a Hit' is clear and measurable.** Our analysis consistently shows that popularity is most strongly driven by high **loudness** and high **energy**. Conversely, songs with high **acousticness** are statistically less likely to be popular."
•**"Nostalgia is a powerful driver of engagement.** Music from the **late 1970s represents a 'golden era'** with a timeless appeal and consistently high popularity scores, often surpassing modern tracks."
•**"Success is predictable.** Using the audio features, we successfully built a machine learning model that acts as a 'cautious but reliable' hit detector. This provides a valuable, data-driven tool for identifying promising new songs.

Business Recommendations

- **For Artists & Labels, follow the data-driven "hit formula":** Prioritize high-energy, loud, and non-acoustic production. Use the predictive model as a powerful screening tool to filter demos and identify new tracks with the highest sonic potential for success.
- **For Playlist Curators, look beyond genre:** Capitalize on the timeless appeal of 1970s music for "throwback" playlists. Use the data-driven "sonic profiles" to create more effective mood-based playlists (e.g., "High-Energy Workout") that truly match the listener's vibe.
- **For Marketing Teams, target the mood:** Launch campaigns based on a song's specific sonic profile. Promote high-energy tracks for party or event-based campaigns and mellow tracks for focus or relaxation, to better connect with the listener's context.

Future Work & Next Steps

- **Integrate genre data** to perform a genre-specific analysis. This would allow us to see how the "formula for a hit" changes for different styles of music, such as rock, hip-hop, or classical.
- **Perform Natural Language Processing (NLP) on song lyrics**. Analyzing the sentiment and common topics of lyrics would add a new dimension to the analysis, allowing us to see if a song's message correlates with its popularity.
- **Incorporate user demographic and streaming trend data**. This would enable the creation of a more personalized recommendation engine and allow for an analysis of a "hit song's" lifecycle from its debut to its peak popularity.

THANK YOU