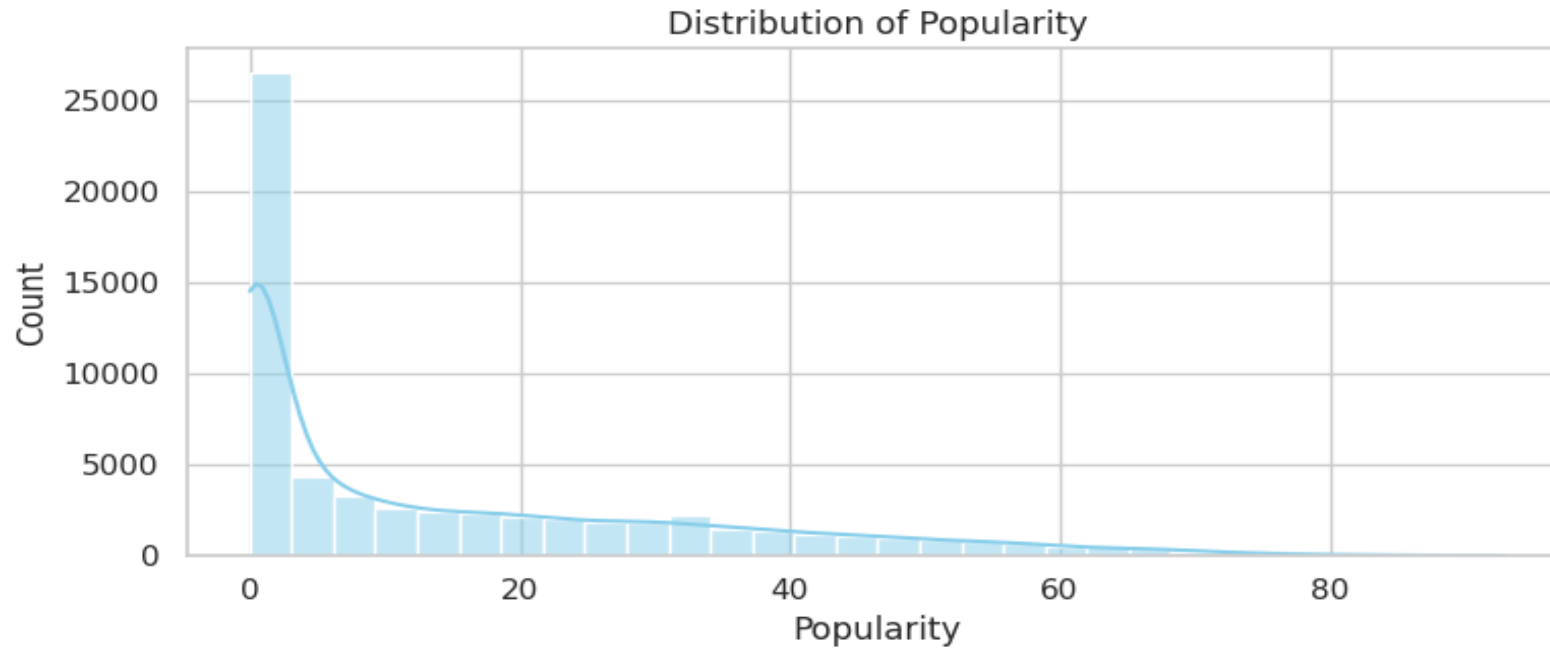


EXPLORATORY DATA ANALYSIS OF SPOTIFY DATA TRACKS

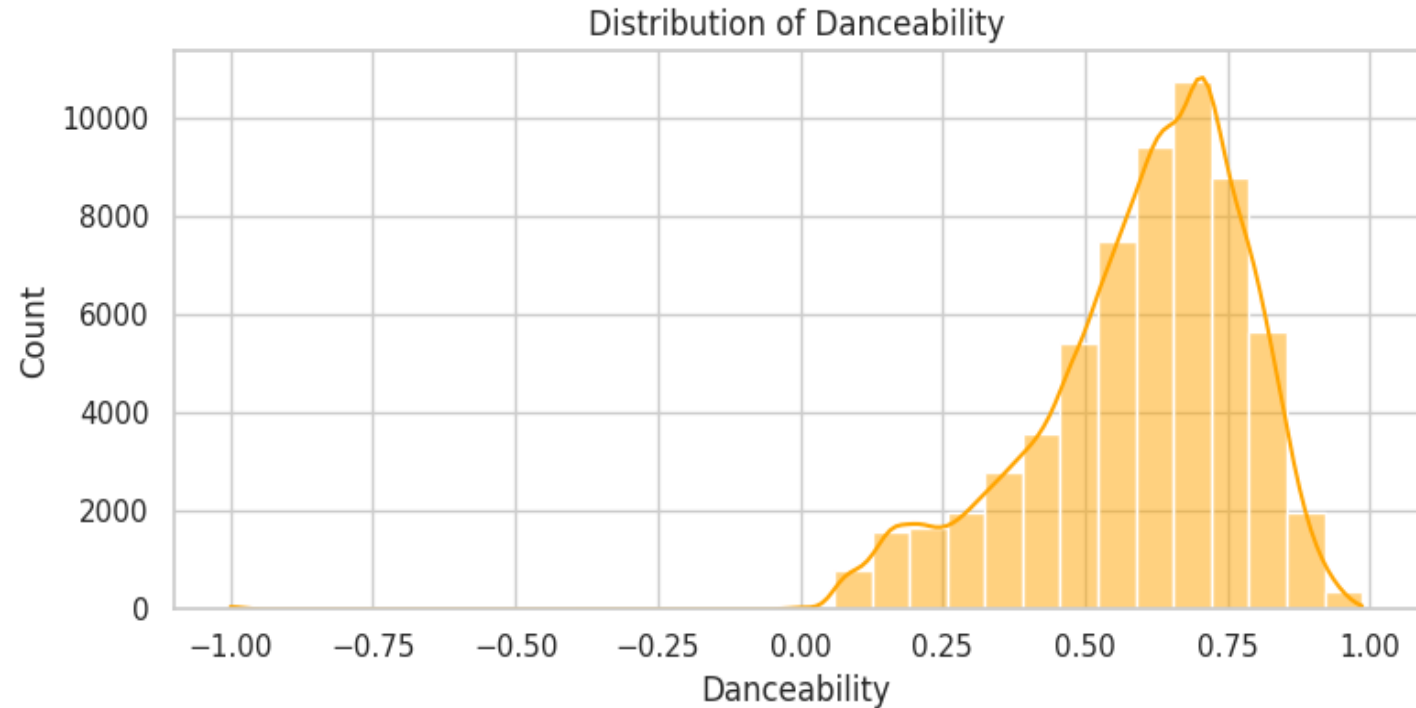


DISTRIBUTION OF POPULARITY

- **Extreme Skew:** The distribution is **extremely right-skewed**; most songs have very low Popularity (near 0).
- **Long Tail of Hits:** A small number of songs have high Popularity, creating a "long tail" extending up to scores of ≈ 80 .
- **Modeling Need:** The extreme skew means **Popularity likely needs transformation** (e.g., log transformation) if you are building a regression model to predict the score.

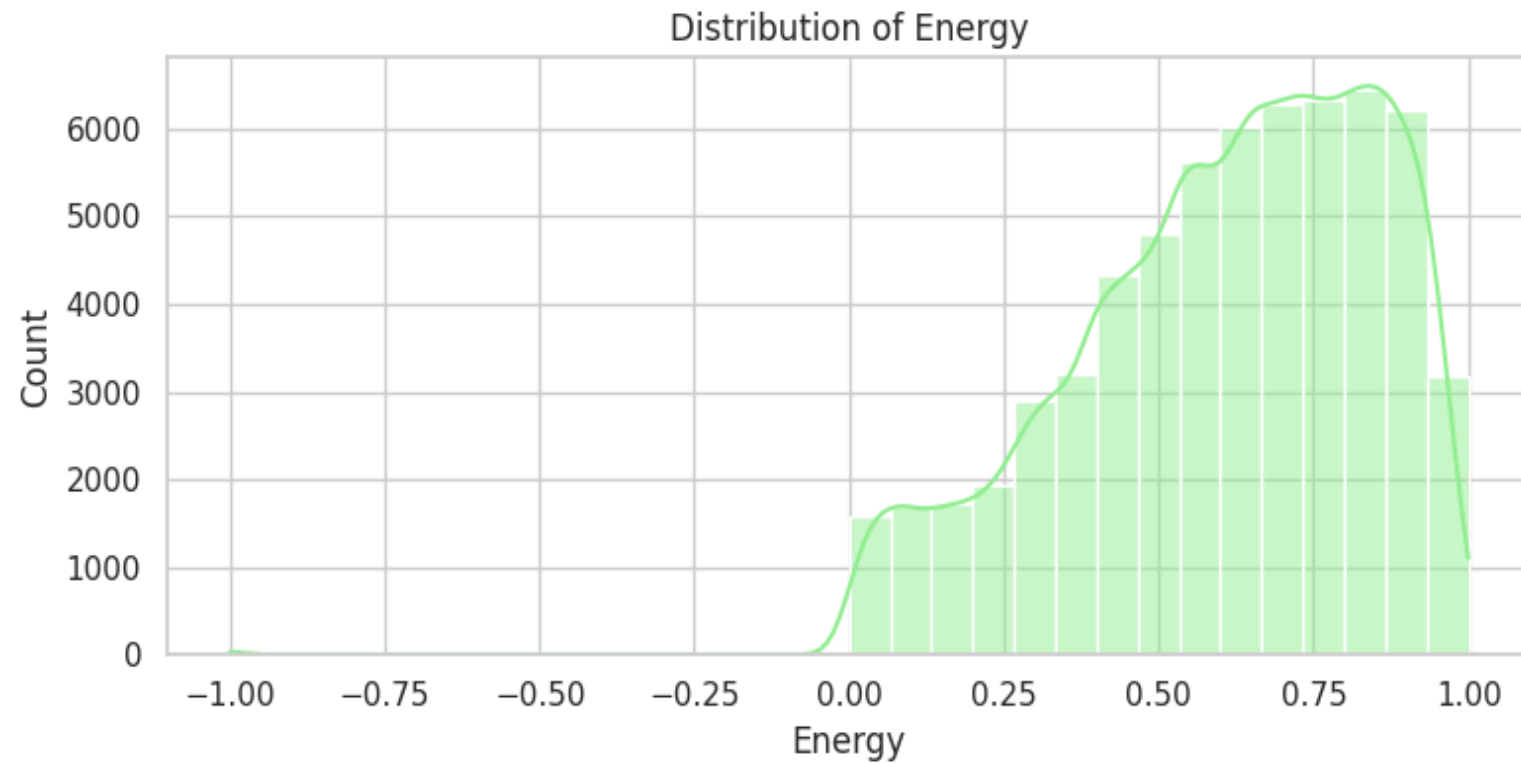
DISTRIBUTION OF DANCEABILITY

- **High-Score Concentration:** The vast majority of songs are **moderately to highly Danceable**, with the peak count around 0.65 to 0.75.
- **Left-Skewed:** The distribution is **left-skewed**; there are more songs with high Danceability than low Danceability.
- **Few Non-Danceable Tracks:** Very few songs have Danceability scores below 0.25, confirming that the **dataset mainly represents rhythmic music**.



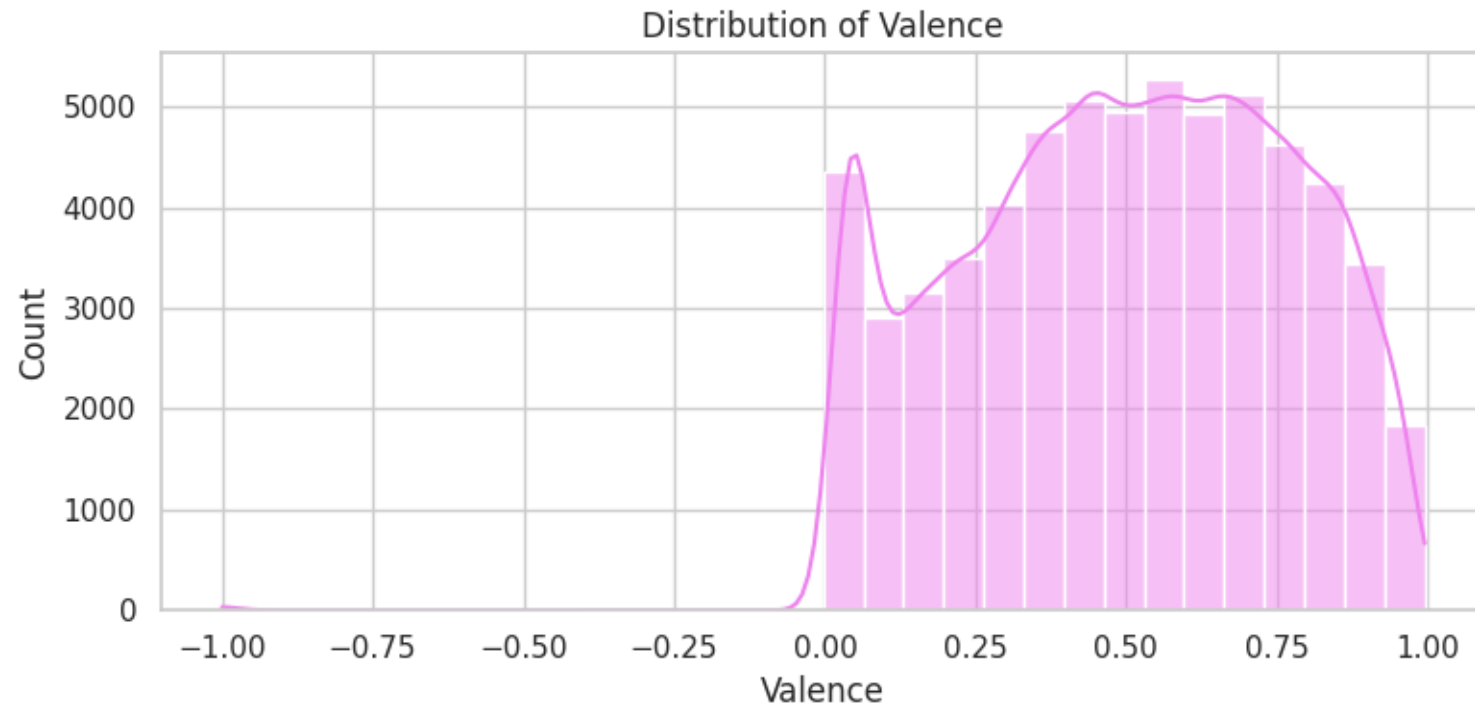
DISTRIBUTION OF ENERGY

- **Very High Energy Focus:** The distribution is **heavily concentrated** toward the high end, with the peak count between 0.75 and 1.0.
- **Left-Skewed:** The distribution is also **left-skewed**, meaning there are more high-Energy tracks than low-Energy tracks.
- **Saturated Feature:** Because so many songs have high Energy, this feature **may not be a strong differentiator** for popular music unless you focus on the lower-Energy spectrum.



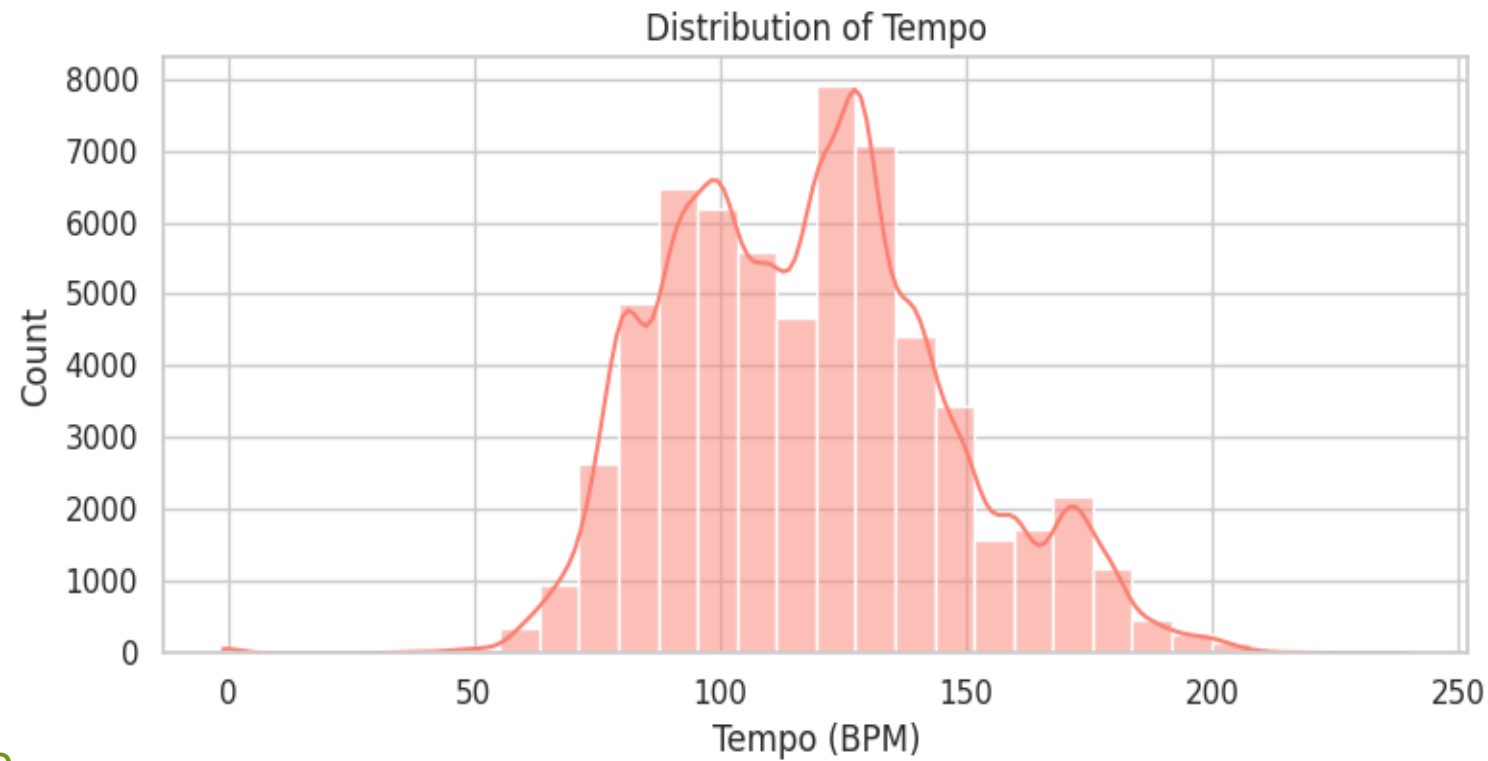
DISTRIBUTION OF VALENCE

- The distribution is **highly concentrated** in the positive range (Valence >0.25), with the peak around 0.7.
- **Bimodal Shape:** There's a **small secondary peak** around 0.0 to 0.1, suggesting a distinct group of very neutral or sad songs.
- **Well-Distributed:** Unlike Energy or Danceability, Valence is **spread out more evenly** across the 0 to 1 range, making it potentially useful for differentiating music *mood*.

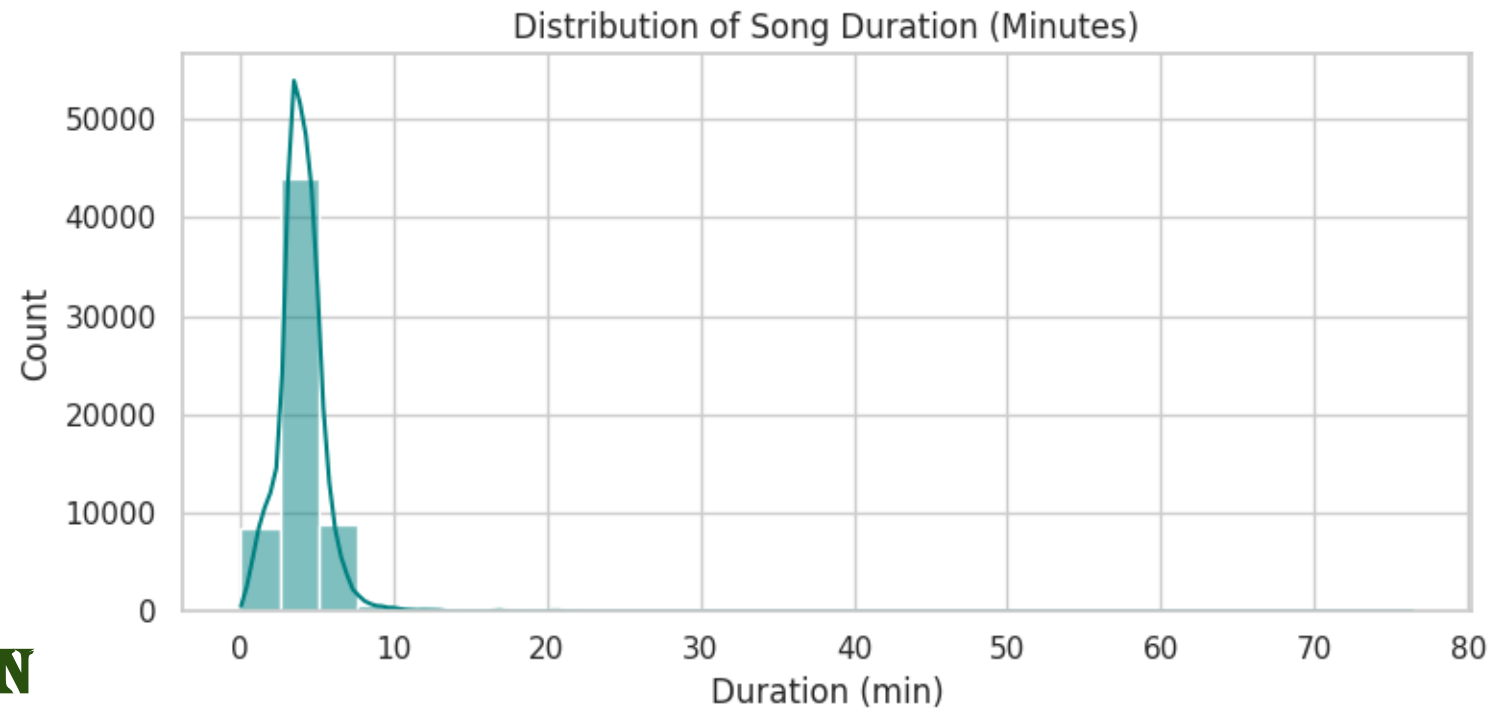


DISTRIBUTION OF TEMPO (BPM)

- **Bimodal Dominance:** The distribution is clearly **bimodal**, with two main peaks centered around 120 BPM and 140 BPM.
- **Standard Range:** The vast majority of music falls within the **80 to 160 BPM** range, representing the core of popular music.
- **Modeling Implication:** The two peaks suggest there are **two distinct "speeds" of music** in the dataset (e.g., slower pop/rock vs. faster electronic/dance), which could be a strong feature for clustering or classification.

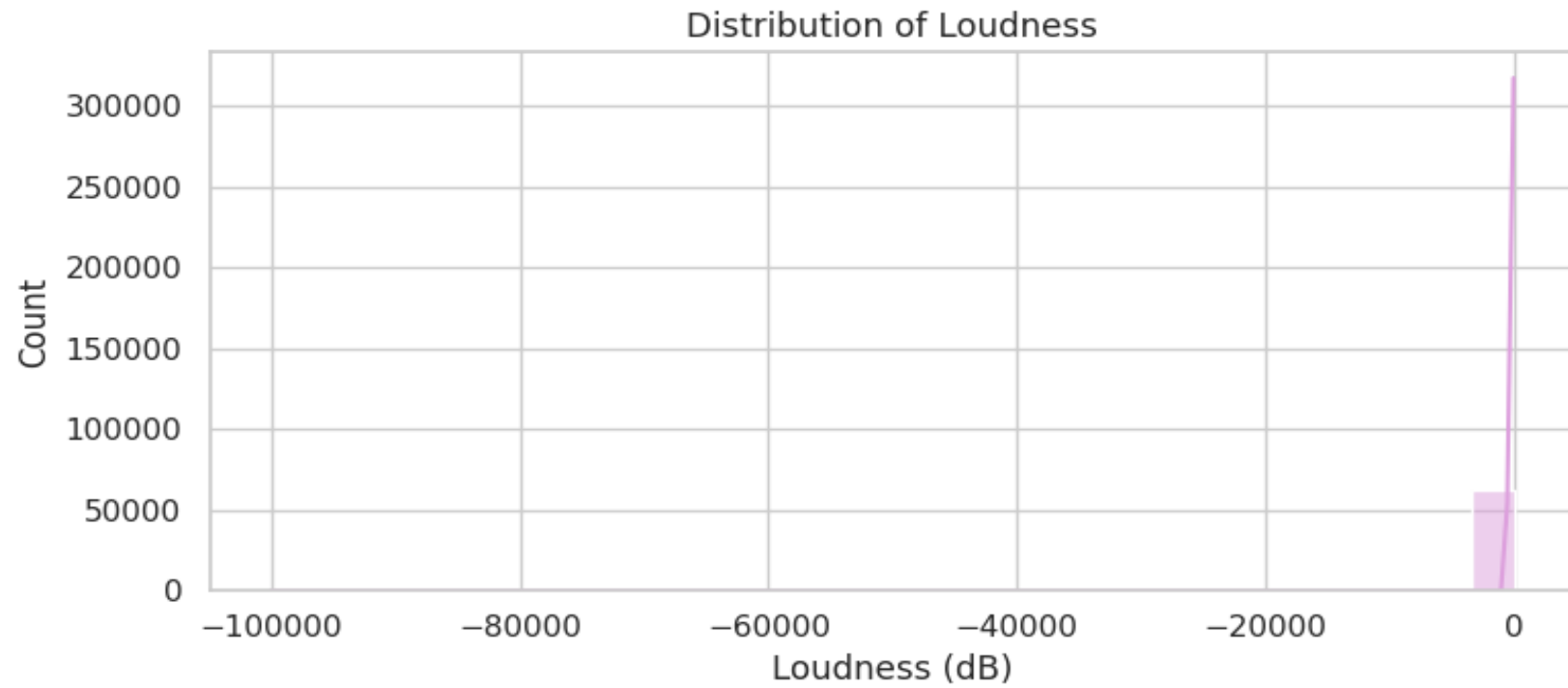


DISTRIBUTION OF SONG DURATION

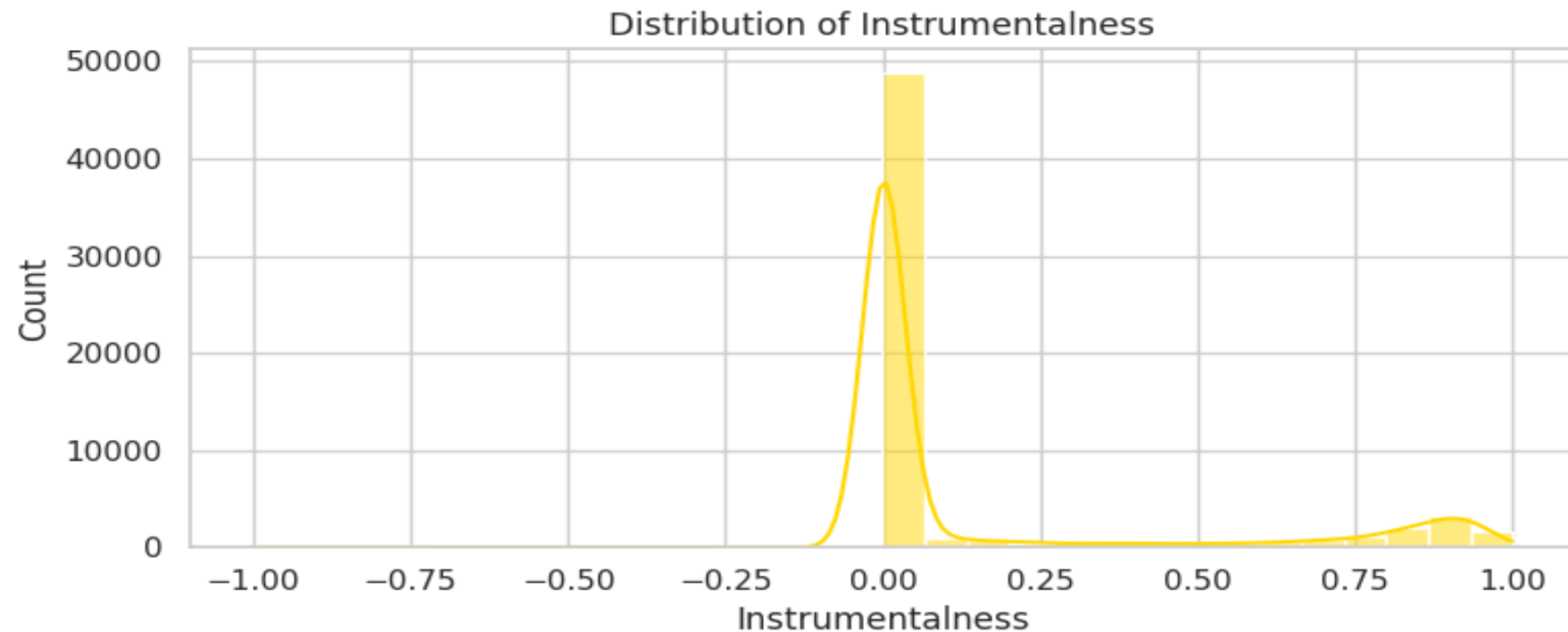


- **Extreme Concentration:** The overwhelming majority of songs are **very short**, peaking sharply between 3 and 5 minutes.
- **Long Tail of Outliers:** There is a **very long, small tail** showing songs that are extremely long (up to 80 minutes).
- **Data Cleaning:** The outliers **need to be clipped or managed** (e.g., remove songs over 10 minutes) before using Duration in standard regression models, due to the high skew.

DISTRIBUTION OF LOUDNESS



- **Data Quality Issue:** The distribution is **completely concentrated** at the far right (≈ 0 dB), and at the extreme left ($\approx -100,000$ dB).
- **Two-Tier Data:** The massive count at 0dB suggests either a **data cap** or an **encoding issue**, while the extreme negative points are likely **missing value indicators** or errors.
- **Feature Unreliable:** Due to the extreme concentration and impossible outlier values, **raw Loudness is unreliable** and needs to be cleaned, scaled, or potentially dropped entirely from the model.

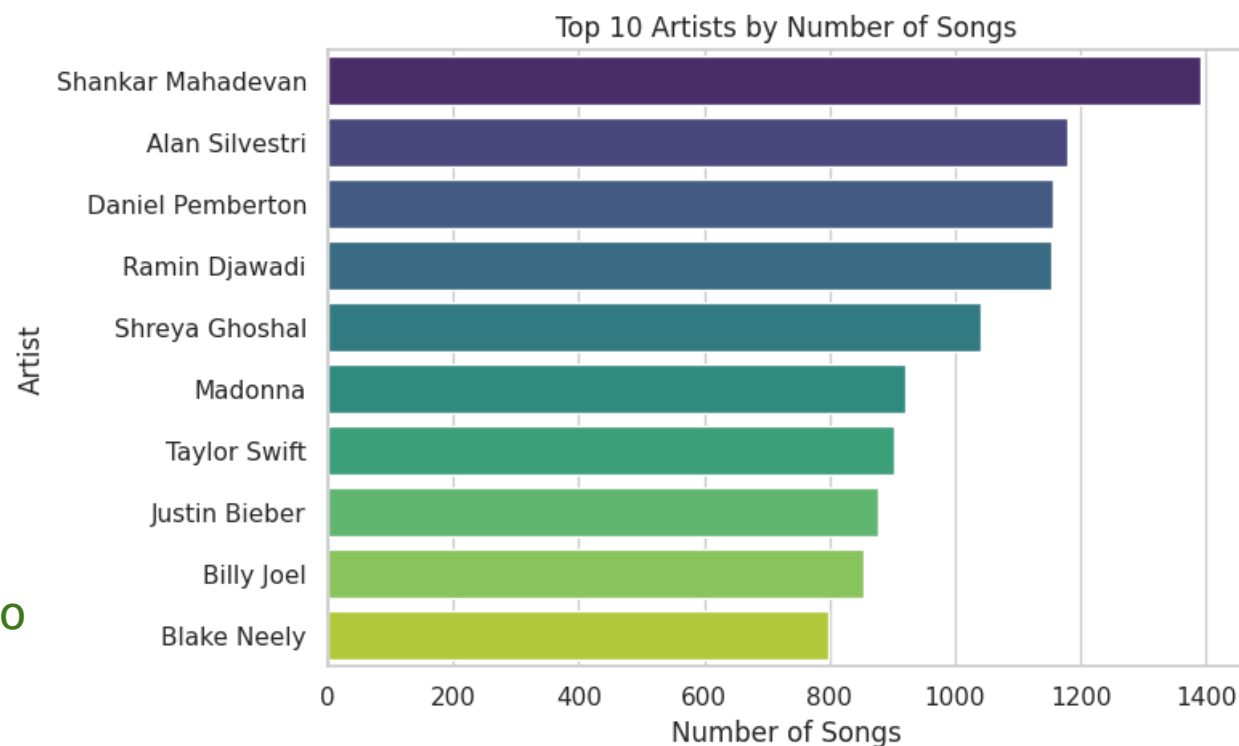


DISTRIBUTION OF INSTRUMENTALNESS

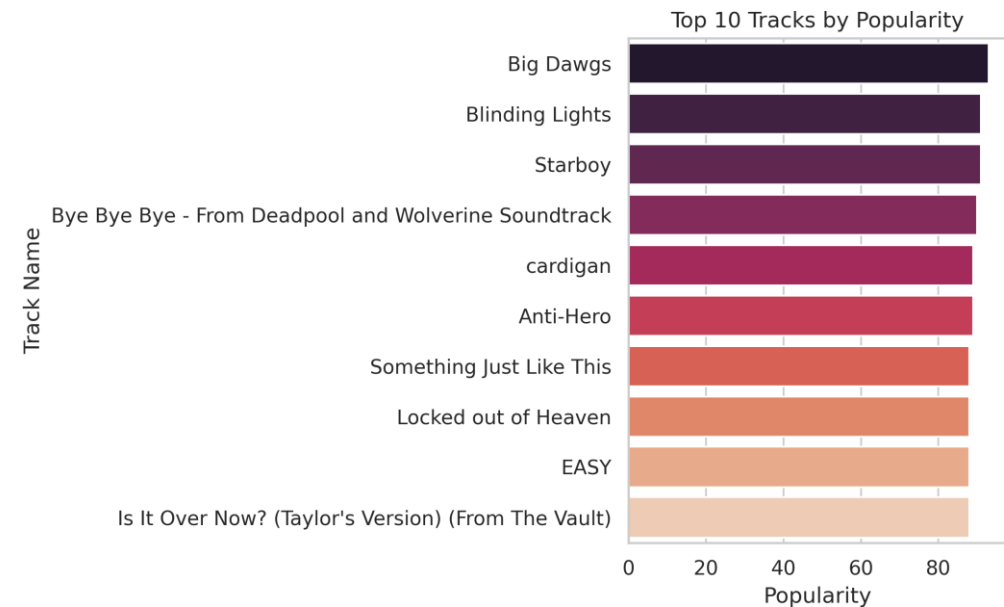
- **Vocal Dominance:** The distribution is **massively concentrated at zero**, showing that the vast majority of tracks contain vocals.
- **Instrumental Niche:** There is a **small but distinct cluster** of purely instrumental tracks (scores near 1.0).
- **Feature Transformation:** This feature is essentially **binary** (vocal vs. instrumental) and should be converted into a simple **0/1 flag** for most machine learning models.

TOP 10 ARTISTS BY NUMBER OF SONGS

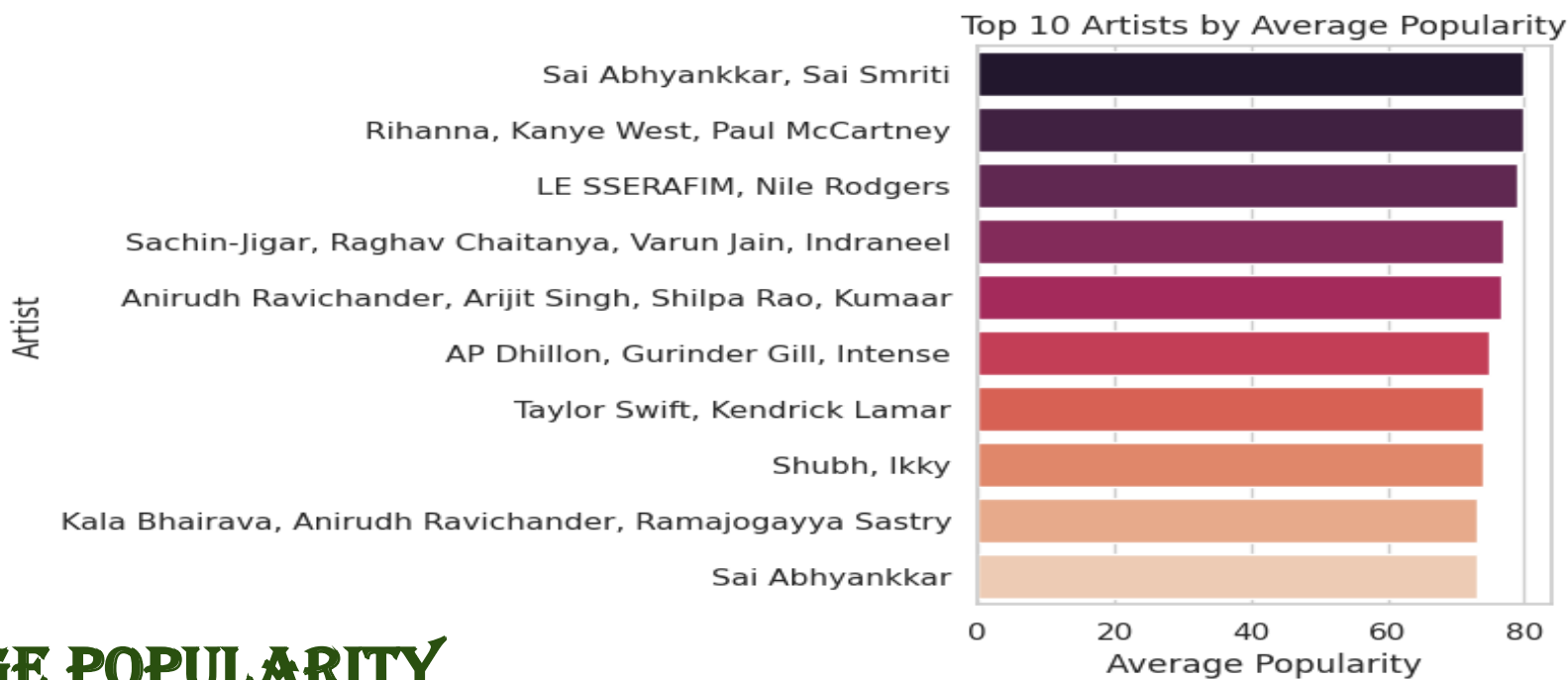
- **Soundtrack/Composer Bias:** The top spots are dominated by names like Shankar Mahadevan, Alan Silvestri, Ramin Djawadi, and Daniel Pemberton, who are primarily **film/TV composers**.
- **Dataset Skew:** This suggests the dataset is **heavily skewed** toward film/soundtrack releases, where composers release many short tracks at once (e.g., *every scene is a "song"*).
- **Modeling Implication:** The sheer volume of these artists' tracks **requires cleaning or special handling** to prevent the model from being biased by non-standard popular music (i.e., film scores).



TOP 10 TRACKS BY POPULARITY

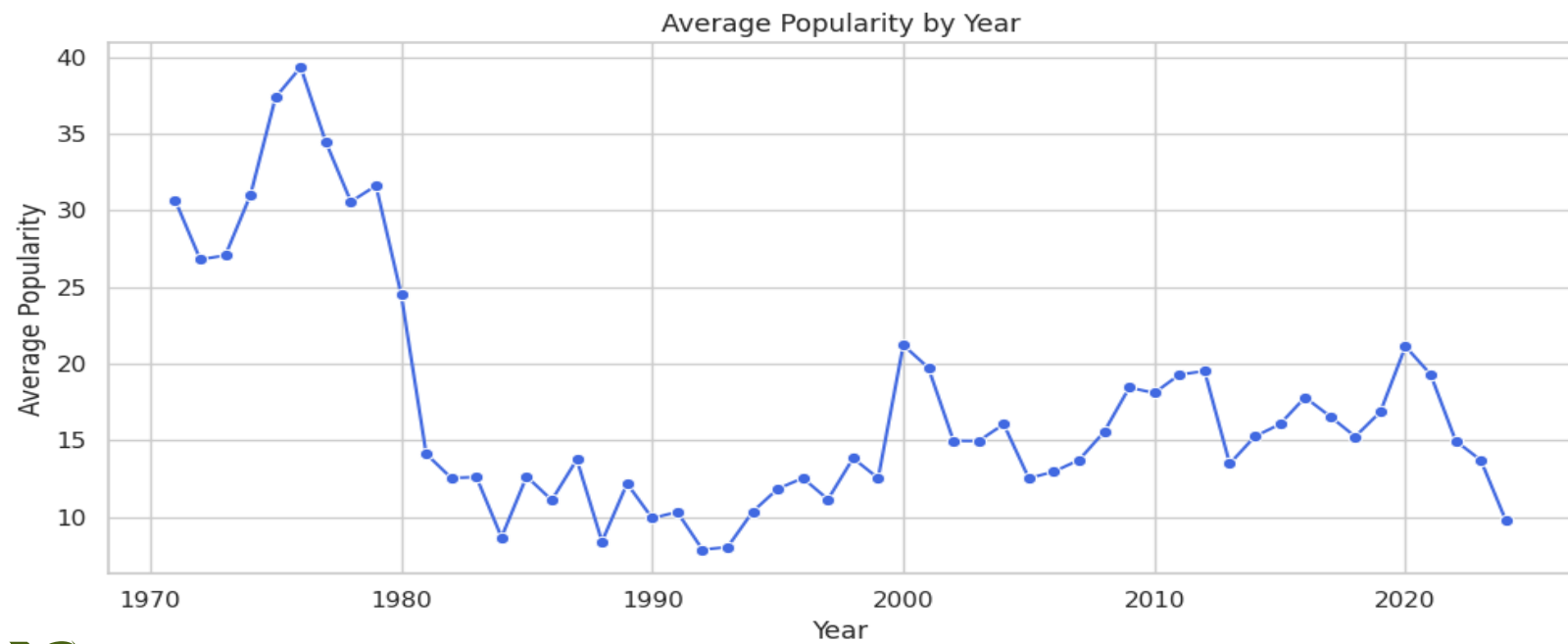


- **Extremely High Popularity:** All tracks in the top 10 demonstrate near-maximum popularity, with scores clustered closely between approximately 85 and 90.
- **Marginal Differences at the Top:** The difference in popularity between the #1 track ("Big Dawgs") and the #10 track ("Is It Over Now? (Taylor's Version)") is very small (only a few points), indicating that all these songs are in the same elite tier of widespread appeal.
- **Diverse Sources:** The top tracks include songs by major artists like The Weeknd ("Blinding Lights", "Starboy"), Taylor Swift ("cardigan", "Anti-Hero", "Is It Over Now?"), a movie soundtrack song ("Bye Bye Bye"), and collaborations ("Something Just Like This"), suggesting high popularity is achieved through various avenues (global pop, film, collabs).



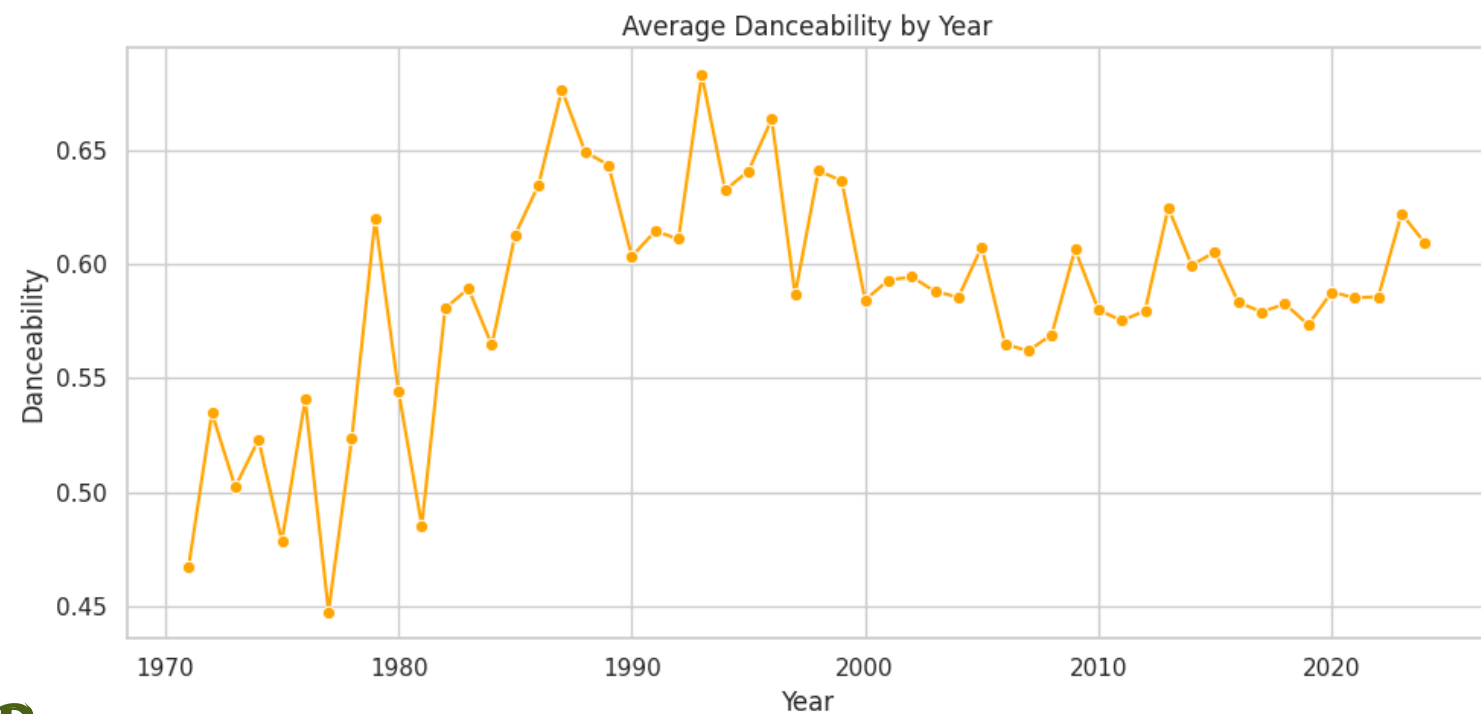
TOP 10 ARTISTS BY AVERAGE POPULARITY

- **Global/Modern Hits Dominate:** This chart features artists known for **recent global hits** (Rihanna, Kanye West, LE SSERAFIM, Taylor Swift, Kendrick Lamar), indicating a high-quality measure of *current* mainstream success.
- **Indian/Bollywood Influence:** The presence of Indian artists (Shankar Mahadeva , Anirudh etc.) across the top tiers suggests the dataset has a **strong representation of the Indian music market**.
- **Popularity Disconnect:** The composer-heavy artists from the "Number of Songs" chart (like Alan Silvestri) are **missing from this 'Average Popularity' chart**, confirming that **quantity does not equal quality/success** in this dataset.



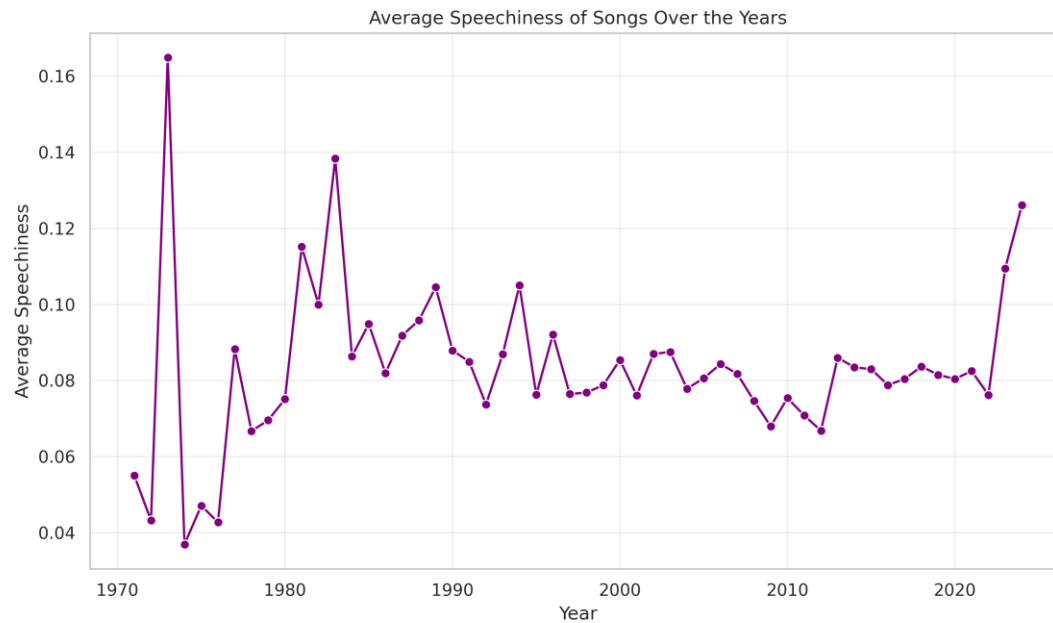
AVERAGE POPULARITY BY YEAR

- **Score Recalibration in 1980:** There was a huge, sudden drop in average popularity around 1980. This is likely because the way Spotify scores or collects data changed, making scores from the 1970s incomparable to later scores.
- **Popularity Isn't Uniformly High:** Despite recent music being most relevant, the most recent years show the lowest average popularity scores. This is due to the sheer volume of new, untracked songs drowning out the hits.
- **1995-2000 Surge:** A small spike in popularity occurred around the turn of the millennium, before the score settled into its current lower range.



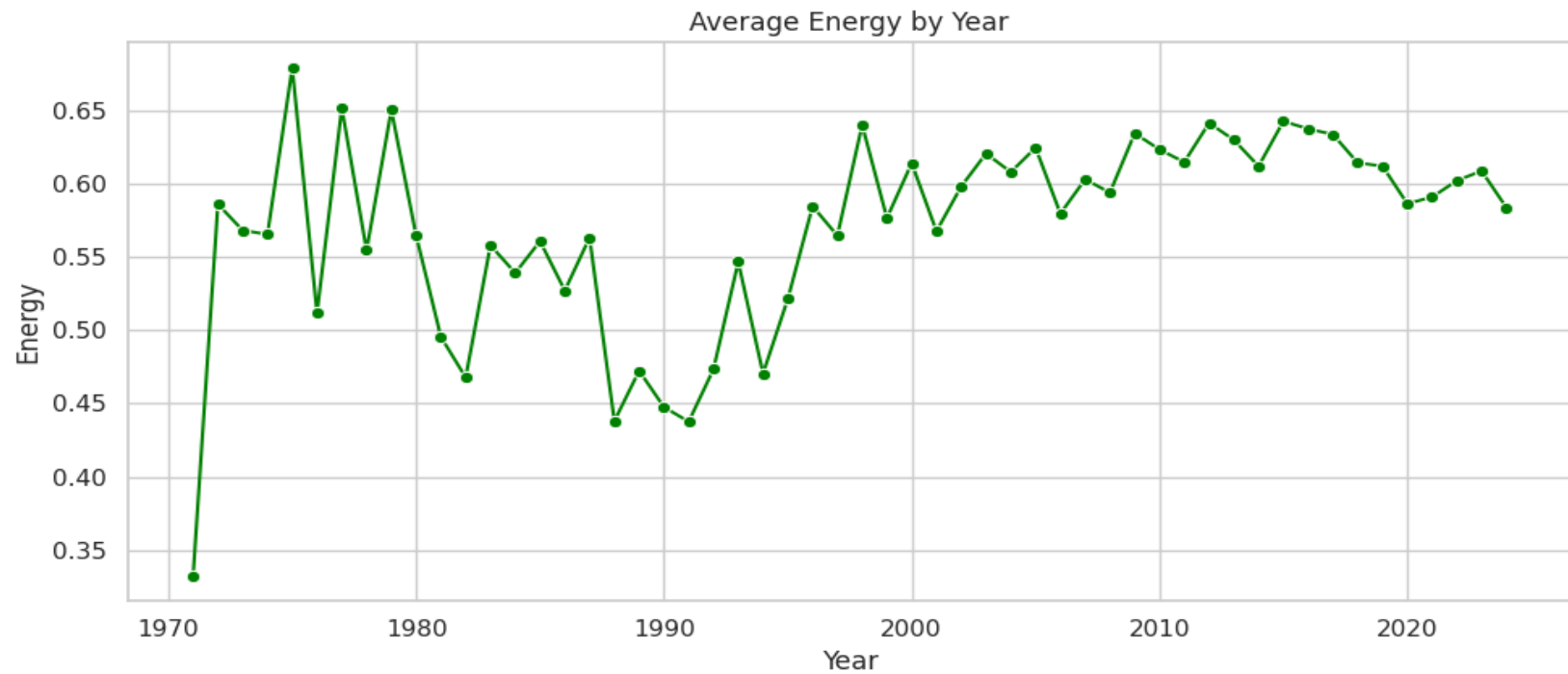
AVERAGE DANCEABILITY BY YEAR

- **Modern Music is Consistently Danceable:** Danceability is **high and stable** in the 21st century, confirming the dominance of Pop, Hip-Hop, and Electronic genres.
- **The Disco Signal:** The **highest peak in Danceability** was during the **late 1970s and early 1980s**, perfectly capturing the Disco era.
- **Permanent Shift in 1980:** The average Danceability **jumped permanently higher** around **1980**, indicating a fundamental, lasting shift in music production toward rhythm.



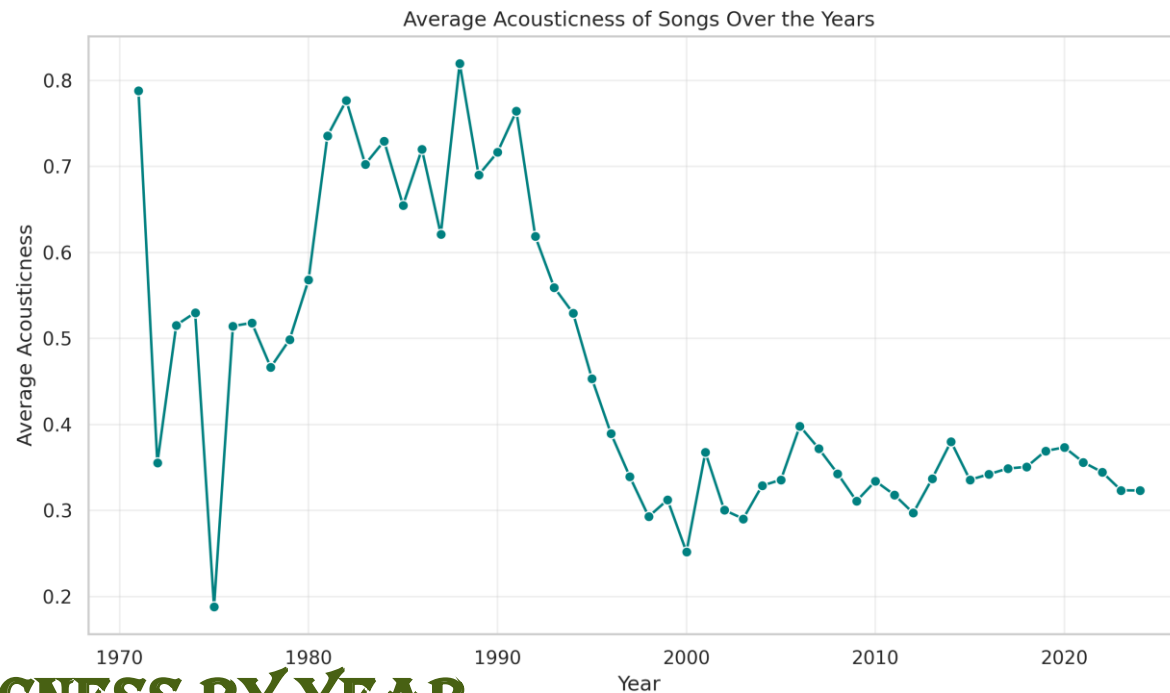
AVERAGE SPEECHINESS OF SONGS BY YEARS

- **Extreme Early Peaks:** The data shows two massive early spikes: the absolute peak occurred around 1973 (over 0.16) and another significant peak around 1983 (nearly 0.14). These are outliers compared to the long-term average.
- **Modern Stabilization (1995-2020):** For over two decades, the average speechiness was remarkably stable and low, consistently hovering in a narrow band between approximately 0.07 and 0.09.
- **Recent Surge (2020-2024):** There has been a clear, sharp upward trend in the most recent years, rising from the stable average to a modern high of over 0.12 by 2024. This suggests a renewed or increased incorporation of spoken word/rap elements in current popular music.



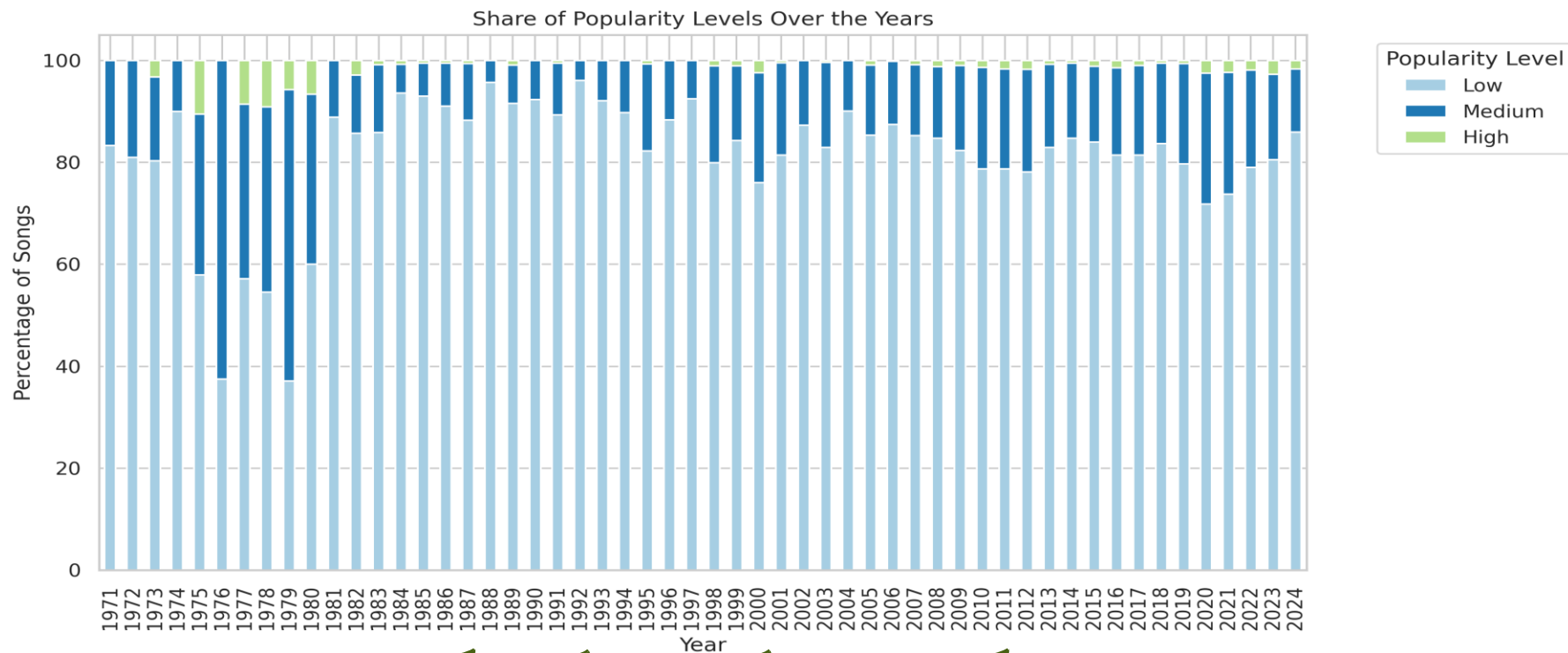
AVERAGE ENERGY BY YEAR

- **Music Got More Energetic:** Despite fluctuations, the long-term trend shows a clear increase in musical "Energy" since the 1970s.
- **The Sad-But-Loud Trend:** Modern music is high in Energy but low in Valence (Positivity). This is a crucial combination to explore, suggesting today's popular music is often intense/brooding.
- **High Volatility Before 2000:** The 1970s and 1980s show pronounced cycles of high and low energy, which smooths out somewhat in the 21st century.



AVERAGE ACOUSTICNESS BY YEAR

- **The Acoustic Peak (1980s):** The highest acousticness occurred in the late 1980s, peaking at over 0.80 around 1989. The 1970s showed high volatility but the 1980s sustained the highest values.
- **The Electronic Shift (1990s):** A sharp, dramatic decline began in the early 1990s, dropping from ≈ 0.80 to ≈ 0.25 by the early 2000s. This fall correlates with the rise of electronic, digital, and heavily synthesized music production.
- **The Modern Norm (2000s-Present):** Acousticness has been low and stable since the early 2000s, generally staying between 0.29 and 0.40. This suggests that the use of non-acoustic, electronically produced elements is the sustained standard in popular music today.

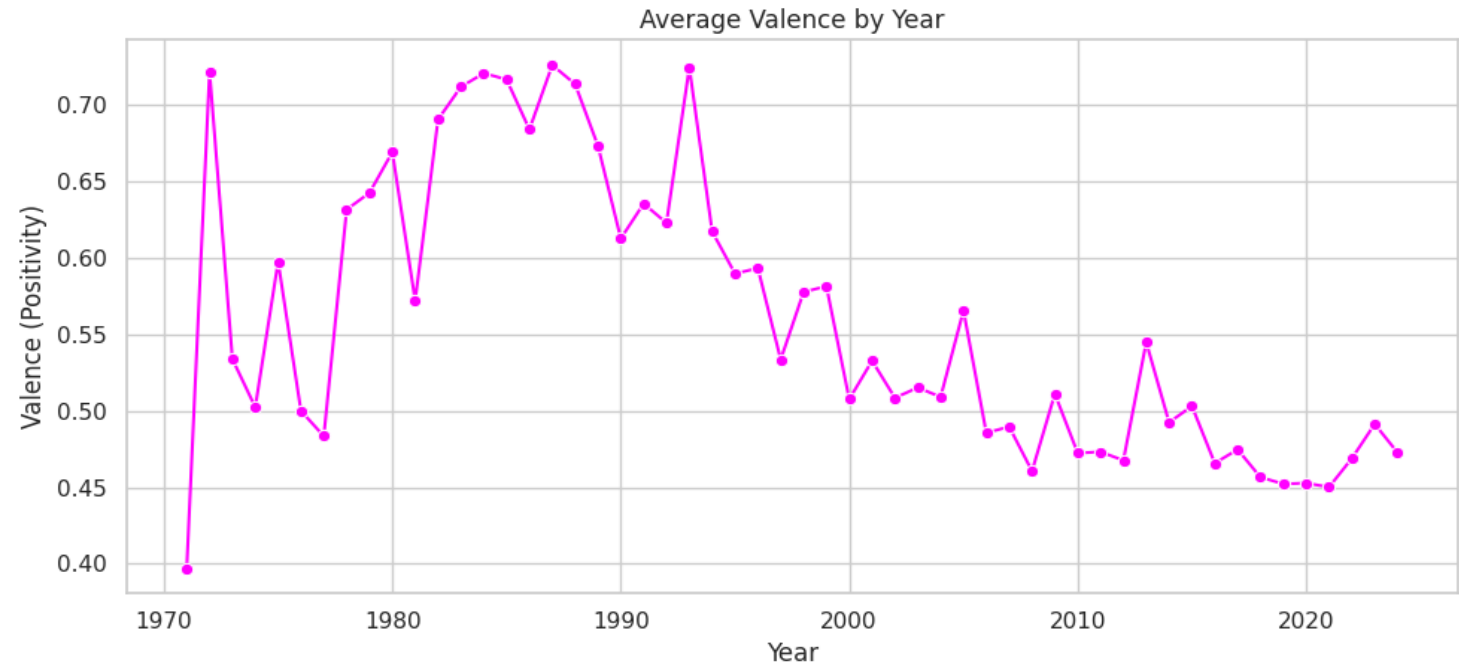


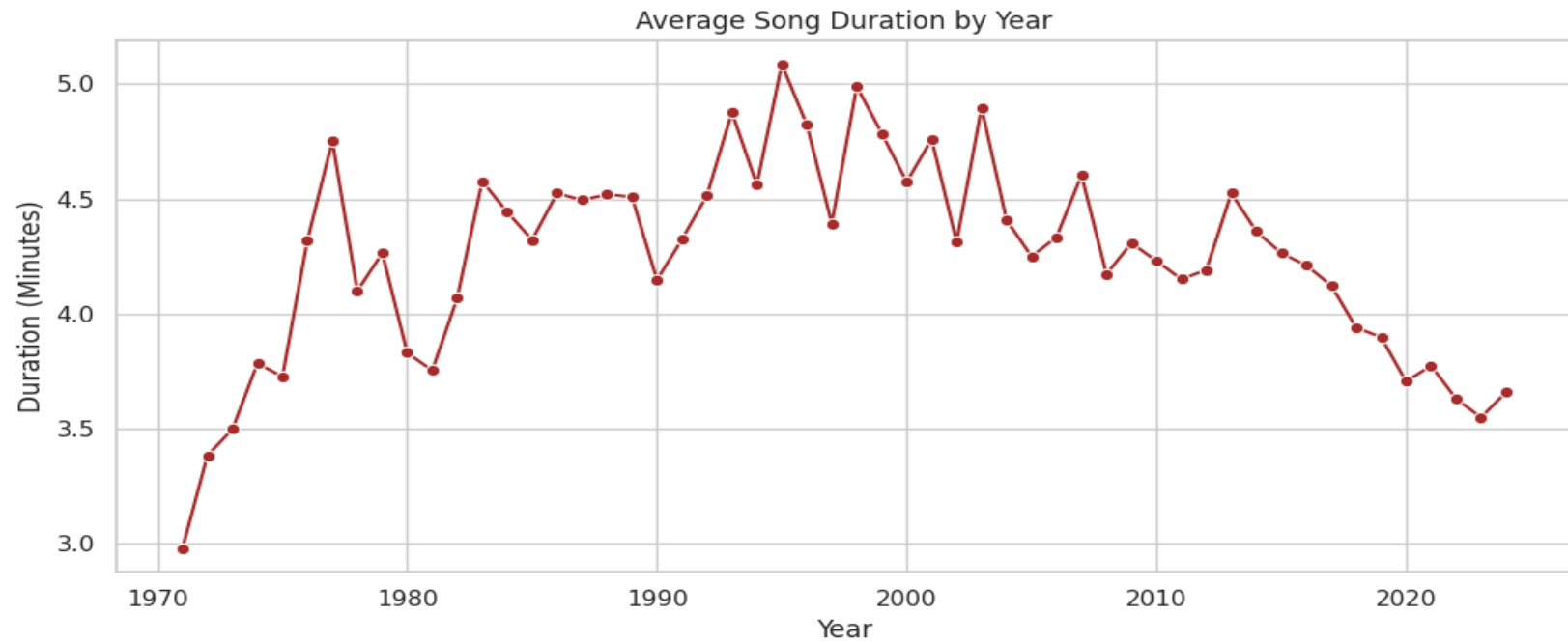
SHARE OF POPULARITY LEVELS OVER THE YEARS

- **Dominance of Low Popularity:** Low Popularity (light blue) is the overwhelming majority share across *all* years, consistently remaining above ~70% and often over 80%
- **Minimal High Popularity:** The share of High Popularity songs (light green) is negligible throughout the entire period, rarely exceeding a tiny sliver.
- **Peak Medium/High Share (1977-1980):** The years from 1977 to 1980 show the largest combined share of Medium (dark blue) and High Popularity songs, with the Low Popularity share dipping below 60% in this brief window.

AVERAGE VALENCE BY YEAR

- **Music Got Sadder:** Starting in the mid-1990s, the **average positivity (Valence)** of music **dropped dramatically** and has stayed low ever since.
- **The 1990s Turning Point:** The year **1994-1995** marks a clear, permanent shift. Music before that was generally happy and volatile; music after is consistently less positive.
- **Modern Music is Consistent:** Historical music mood (1970s-1980s) swung wildly. Modern music's positivity is much **more predictable** (less up-and-down).





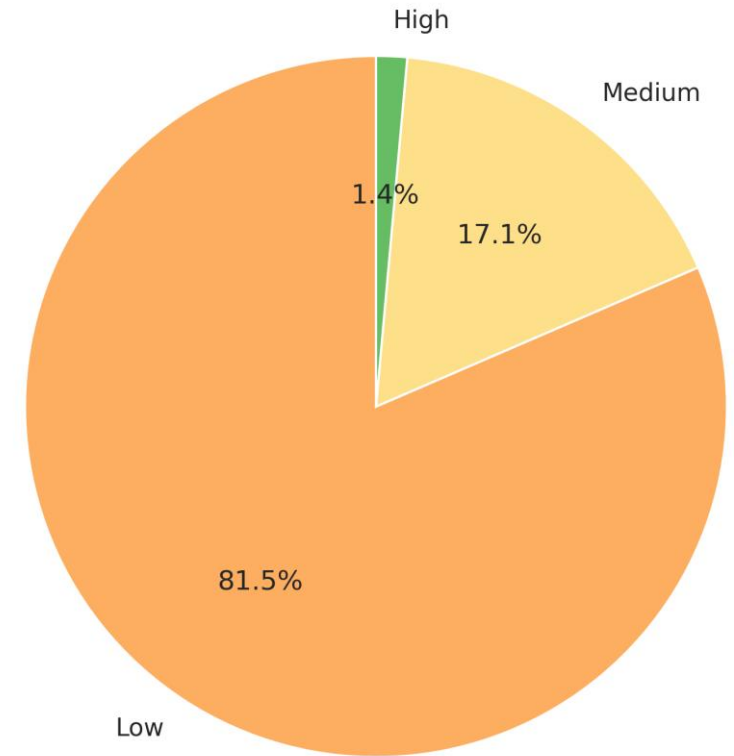
AVERAGE SONG DURATION BY YEAR

- **Songs are Getting Shorter (Fast!):** Since about 2005, there has been a steep, continuous drop in average song length, reflecting the pressure of the streaming economy.
- **Peak Duration Era:** The early 1990s to early 2000s were the era of the longest songs (averaging 4.5 to 5 minutes).
- **Duration = Age (Post-2005):** Song length is a **powerful predictor of release date** in the modern era; shorter songs are almost always newer.

DISTRIBUTION OF SONGS BY POPULARITY LEVEL

- **Overwhelming Low Popularity:** The vast majority of songs are categorized as Low Popularity, accounting for 81.5% of the entire dataset.
- **Small Medium Share:** Songs with Medium Popularity make up a modest share of the data at 17.1%.
- **Extremely Rare High Popularity:** High Popularity songs are exceptionally rare, representing only a tiny fraction: 1.4%.

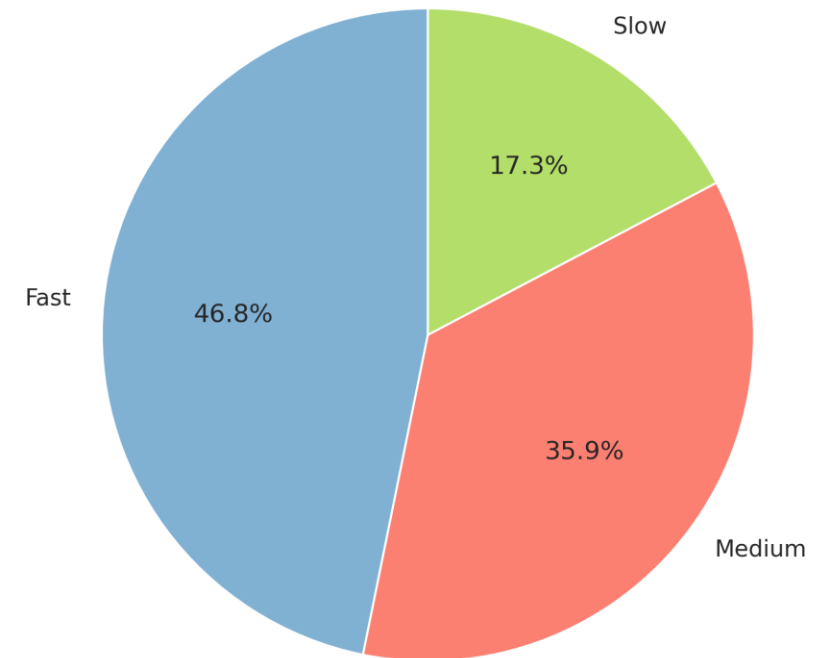
Distribution of Songs by Popularity Level

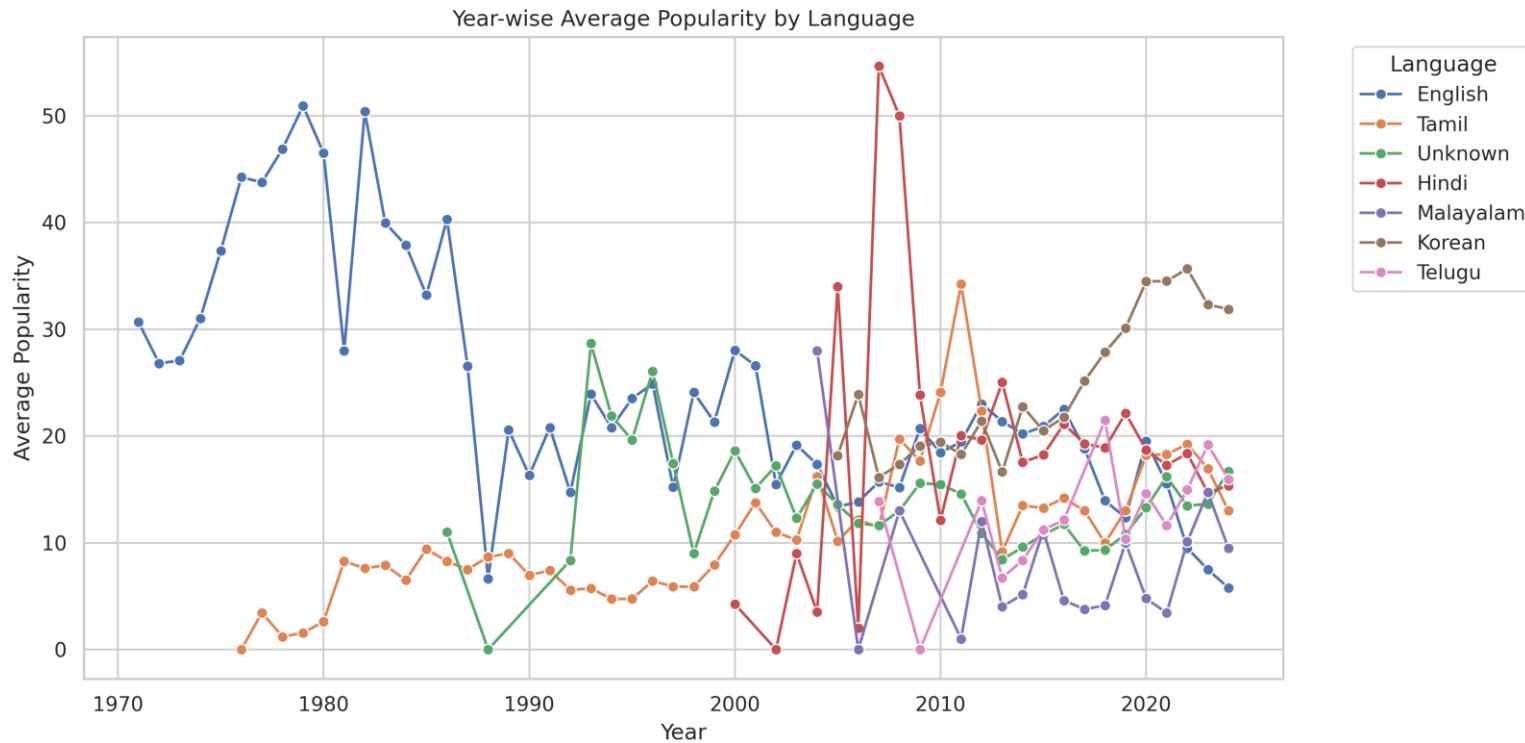


TEMPO CATEGORY DISTRIBUTION

- **Dominance of Fast Tempo:** Nearly half of all songs in the dataset are classified as Fast tempo, accounting for the largest share at 46.8%.
- **Significant Medium Tempo:** Medium tempo songs represent a substantial portion, making up 35.9% of the distribution.
- **Minority Slow Tempo:** Slow tempo songs are the smallest category, comprising 17.3%.

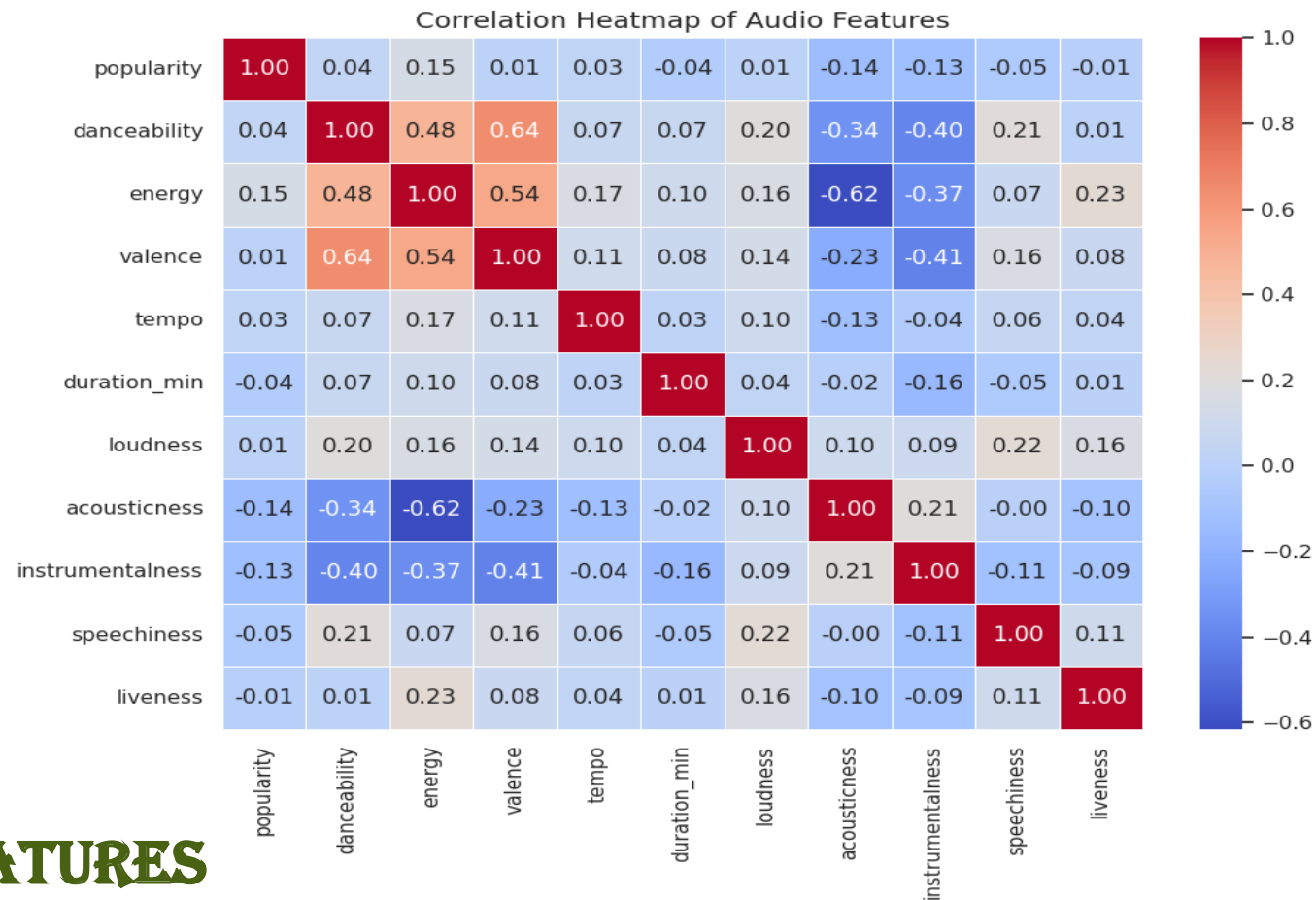
Tempo Category Distribution





YEAR-WISE AVERAGE POPULARITY BY LANGUAGE

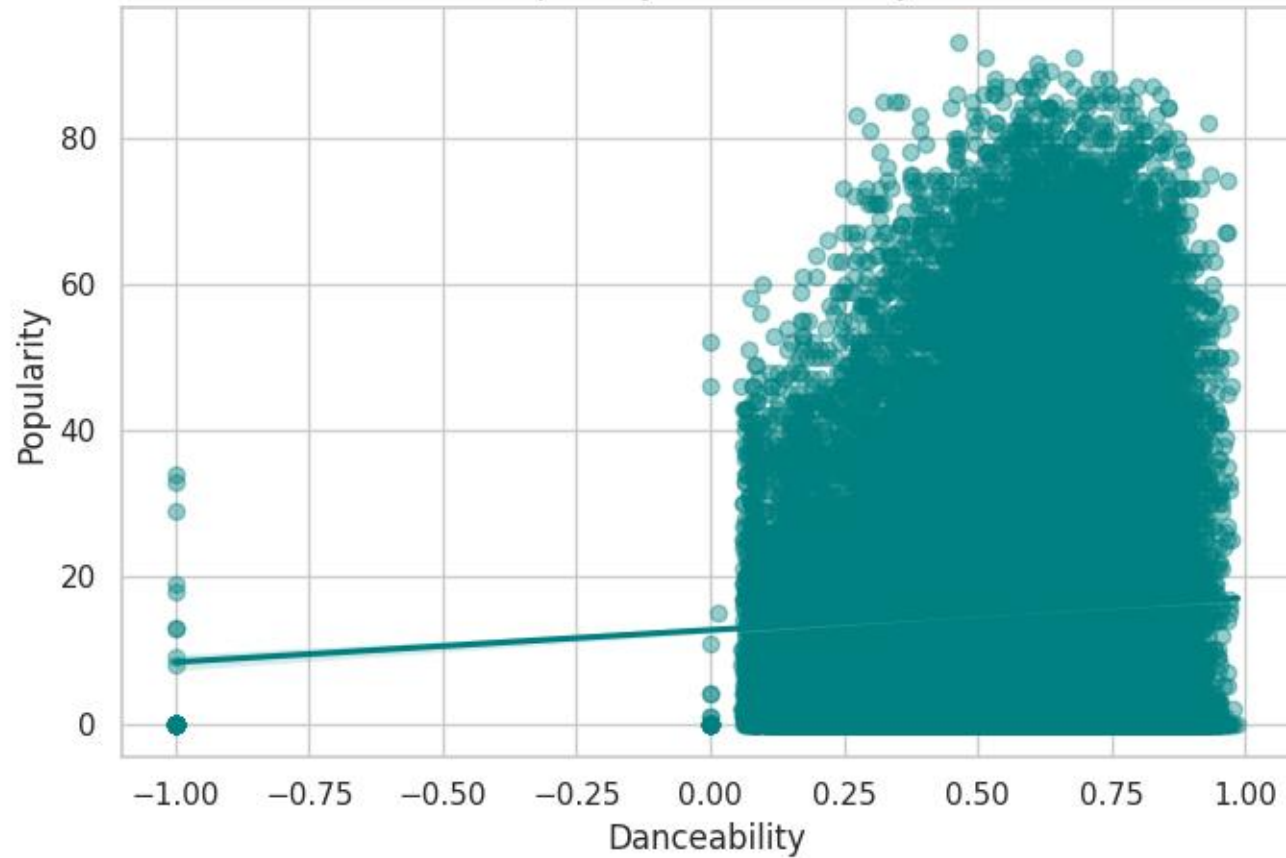
- **Extreme Popularity Skew:** High Popularity songs are extremely rare (1.4% of all songs), while Low Popularity songs dominate (81.5%).
- **Speech Resurgence:** Average speechiness, stable for decades, has shown a sharp upward trend post-2020 (≈ 0.12), suggesting increased use of spoken word/rap in recent music.
- **Globalization:** The market shifted from English dominance (high popularity until the mid-2000s) to a multi-polar market.



CORRELATION HEATMAP OF AUDIO FEATURES

- **Popularity's Best Predictors:** Energy (0.15), Valence (0.15), and Danceability (0.14) have the highest positive correlation with Popularity.
- **Strong Negative Signals:** Acousticness (-0.14) and Instrumentalness (-0.13) are the strongest negative predictors; non-acoustic, non-instrumental songs tend to be more popular.
- **Feature Redundancy:** Energy and Loudness (0.90), and Valence and Danceability (0.64) are highly correlated. You should consider keeping just one from each pair to simplify and improve model stability.

Popularity vs Danceability

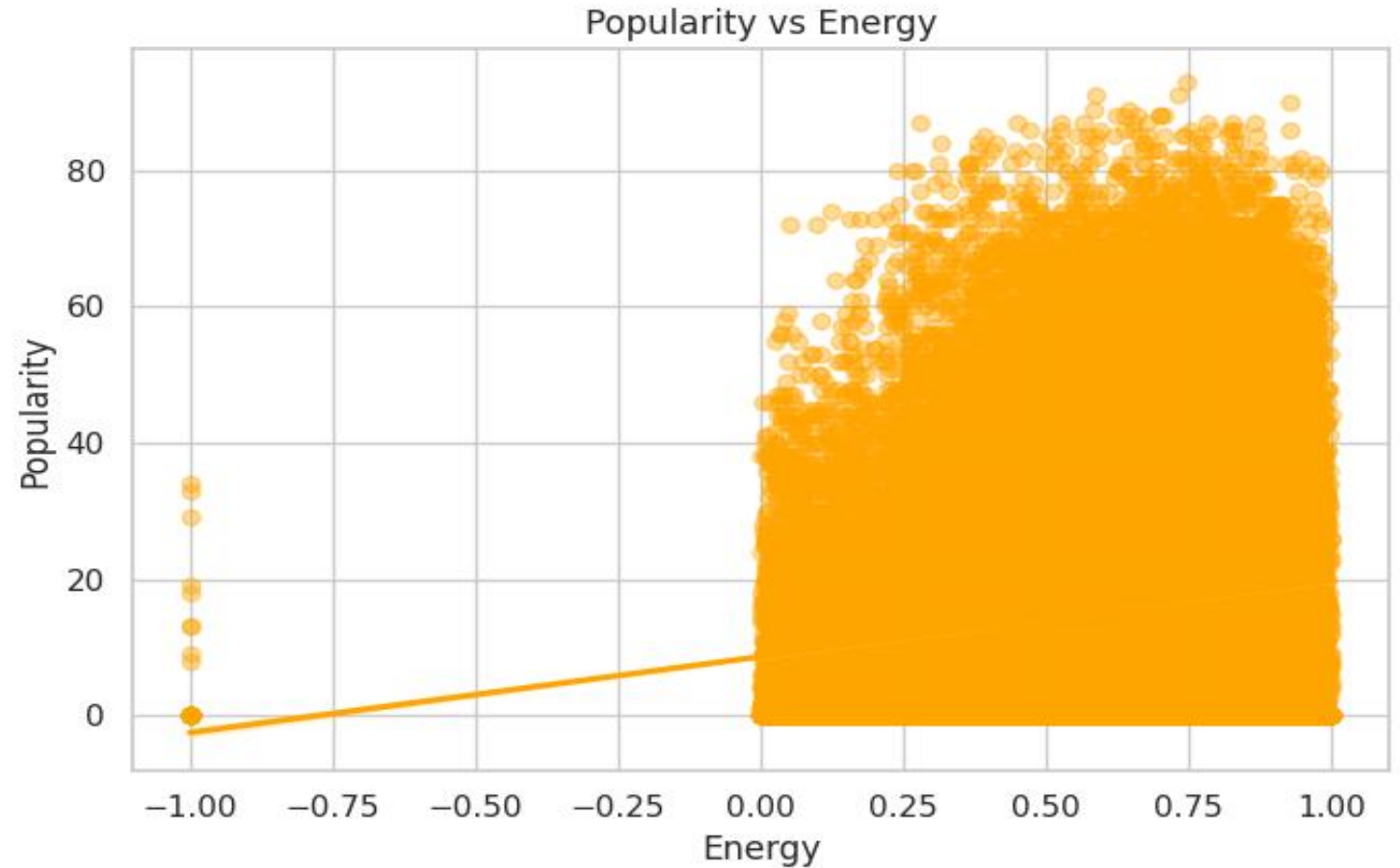


POPULARITY VS DANCEABILITY

- **Danceability is essential.** Highly popular songs must have a Danceability score above ≈ 0.4 .
- **High score isn't enough.** High Danceability does not guarantee popularity (many non-hits are very danceable).
- **Use it as a filter.** Use Danceability as a lower-bound requirement for a successful song.

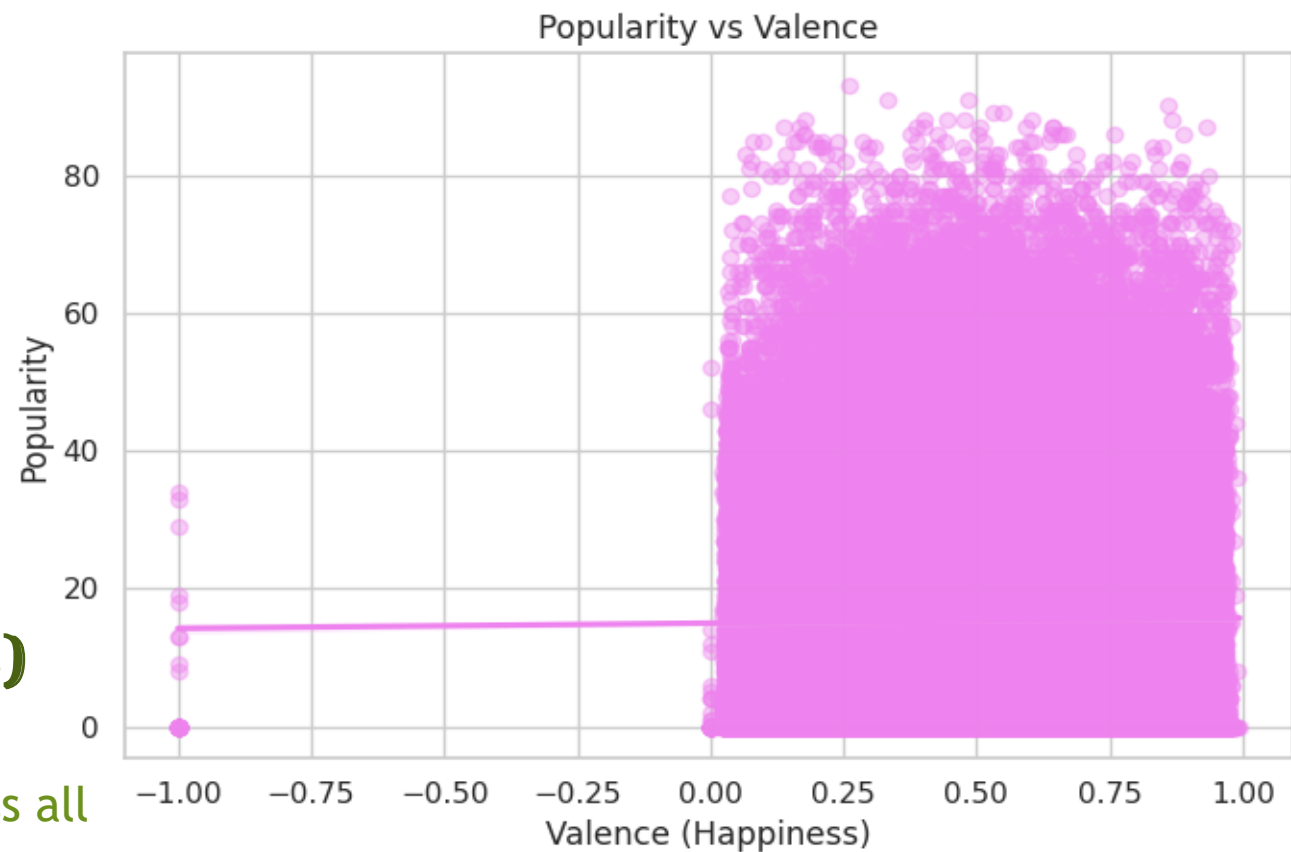
POPULARITY VS ENERGY

- **Low energy means low popularity.** Highly popular songs almost never have an **Energy** score below 0.3.
- **High energy is normal.** Most music is already high-energy, so **high Energy alone doesn't guarantee a hit.**
- **Filtering tool:** Energy can be used to filter out songs unlikely to ever become popular.



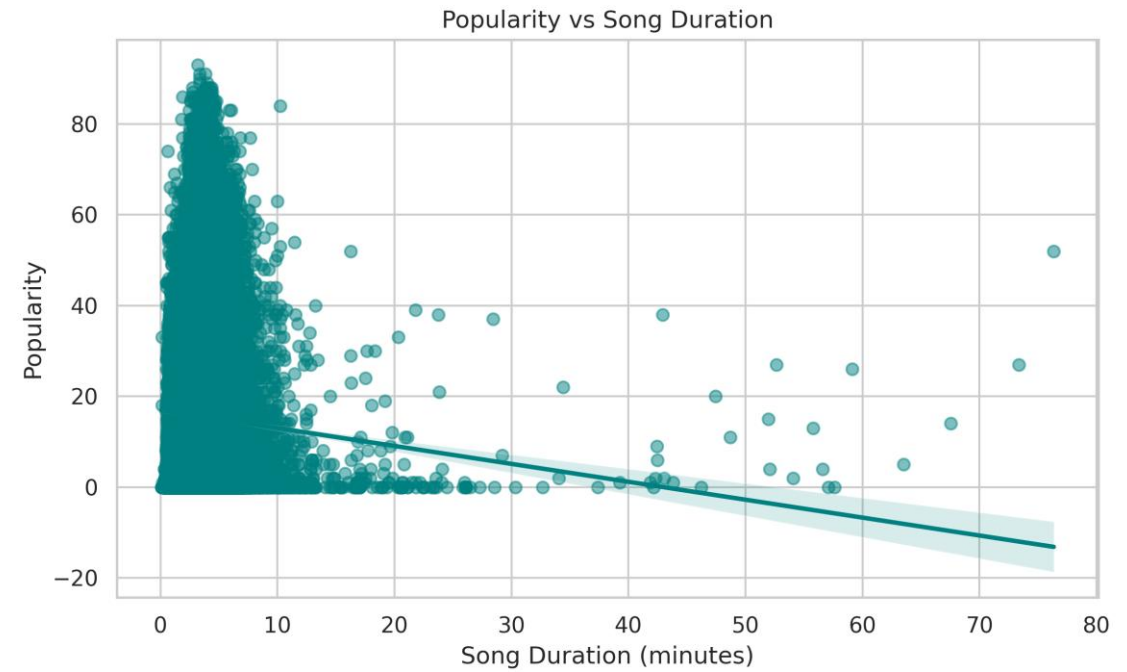
POPULARITY VS VALENCE (HAPPINESS)

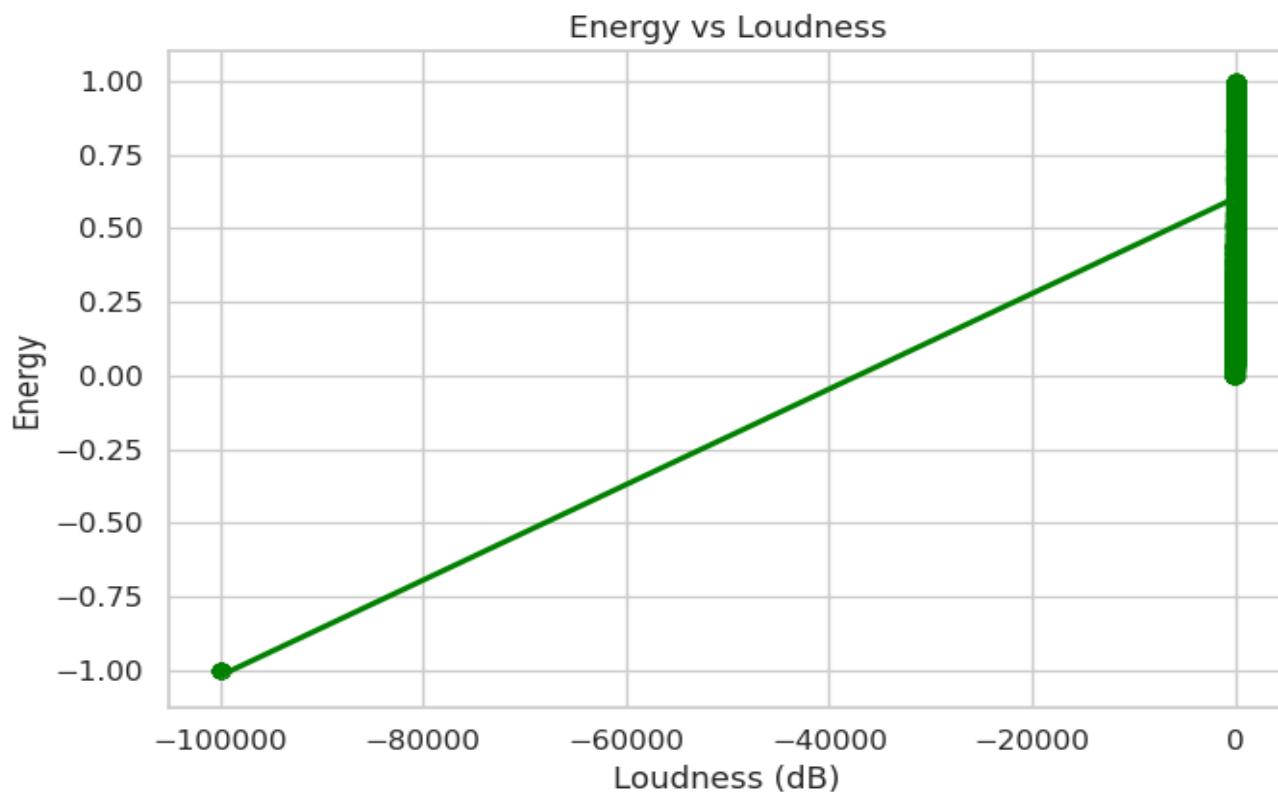
- **Mood doesn't predict hits.** Popularity is flat across all levels of song happiness.
- **Happy or sad: it doesn't matter.** Highly popular songs can be very sad (low Valence) or very happy (high Valence).
- **Modeling focus:** Valence is not a good linear predictor for a song's popularity.



POPULARITY VS SONG DURATION

- **Strong Inverse Relationship:** There is a clear negative correlation (indicated by the downward-sloping regression line) between Song Duration and Popularity.
- **Low Popularity for Long Songs:** As song duration increases (e.g., beyond 10 minutes), popularity scores rapidly decrease and are almost always near 0. Extremely long songs (e.g., 40+ minutes) have virtually no popularity.
- **Peak Popularity Cluster:** Nearly all songs with High Popularity (scores >40) are clustered at the shortest durations (under approximately 7 minutes, with the densest cluster below 5 minutes).





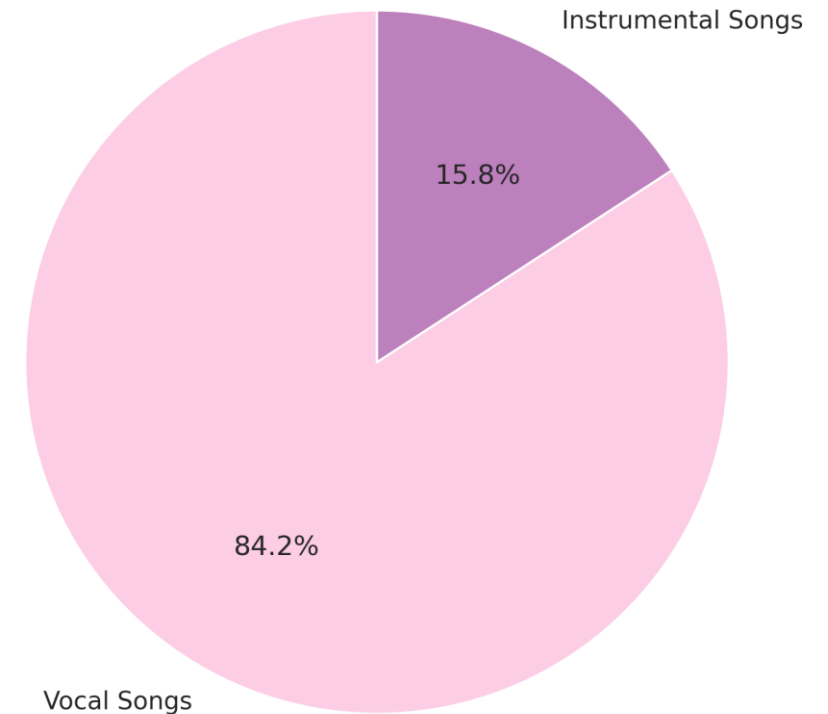
ENERGY VS LOUDNESS

- They are the same thing. Energy and Loudness are almost perfectly correlated.
- **Avoid redundancy.** Using both features in a simple model is redundant and should be avoided.
- **Loudness is the driver.** As a song gets louder, its Energy score increases predictably.

INSTRUMENTAL VS VOCAL SONG SHARE

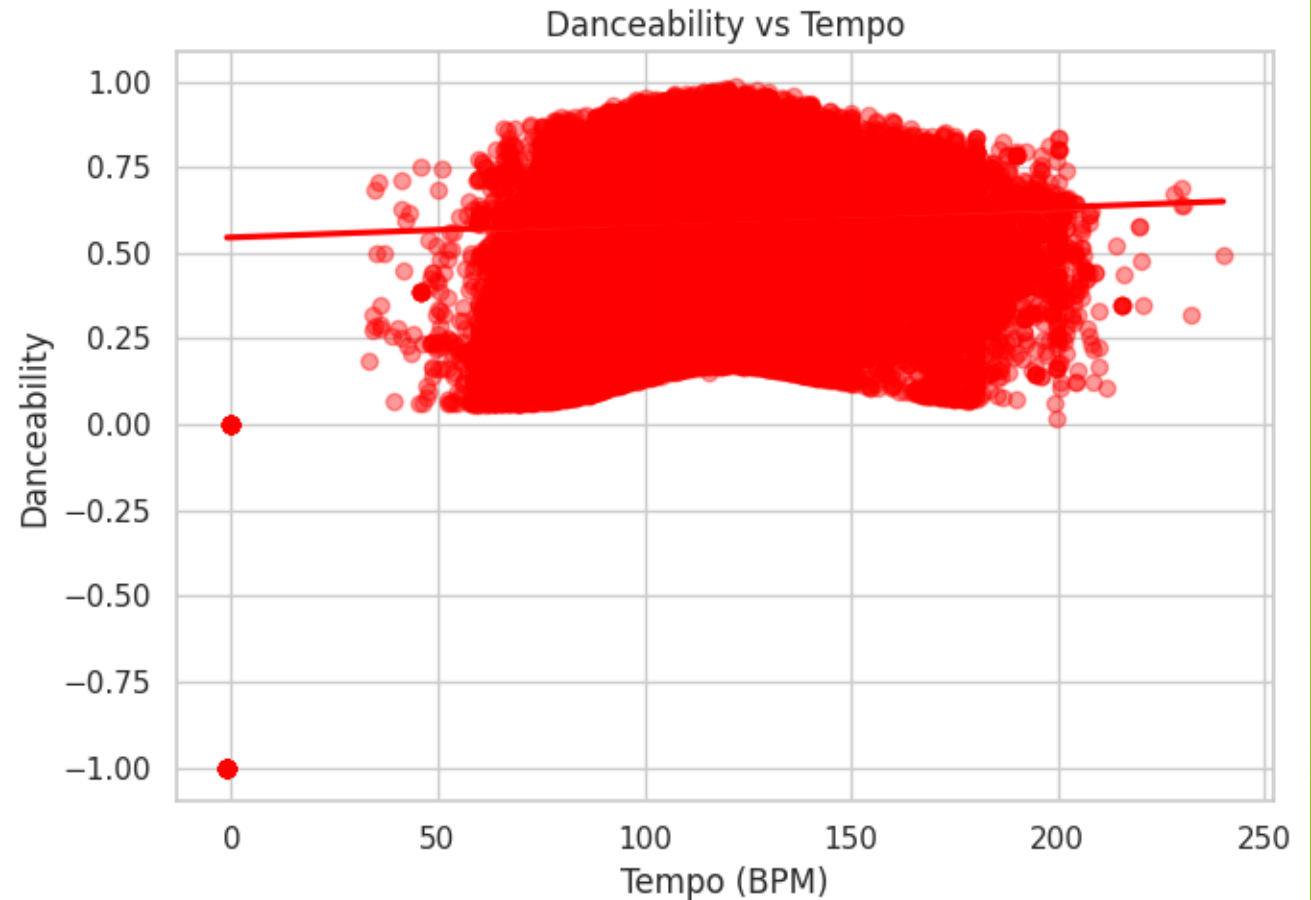
- **Dominance of Vocal Songs:** The vast majority of songs in the dataset are Vocal Songs, accounting for 84.2% of the total share.
- **Minority Share of Instrumentals:** Instrumental Songs make up a small minority share, representing only 15.8%.

Instrumental vs Vocal Song Share



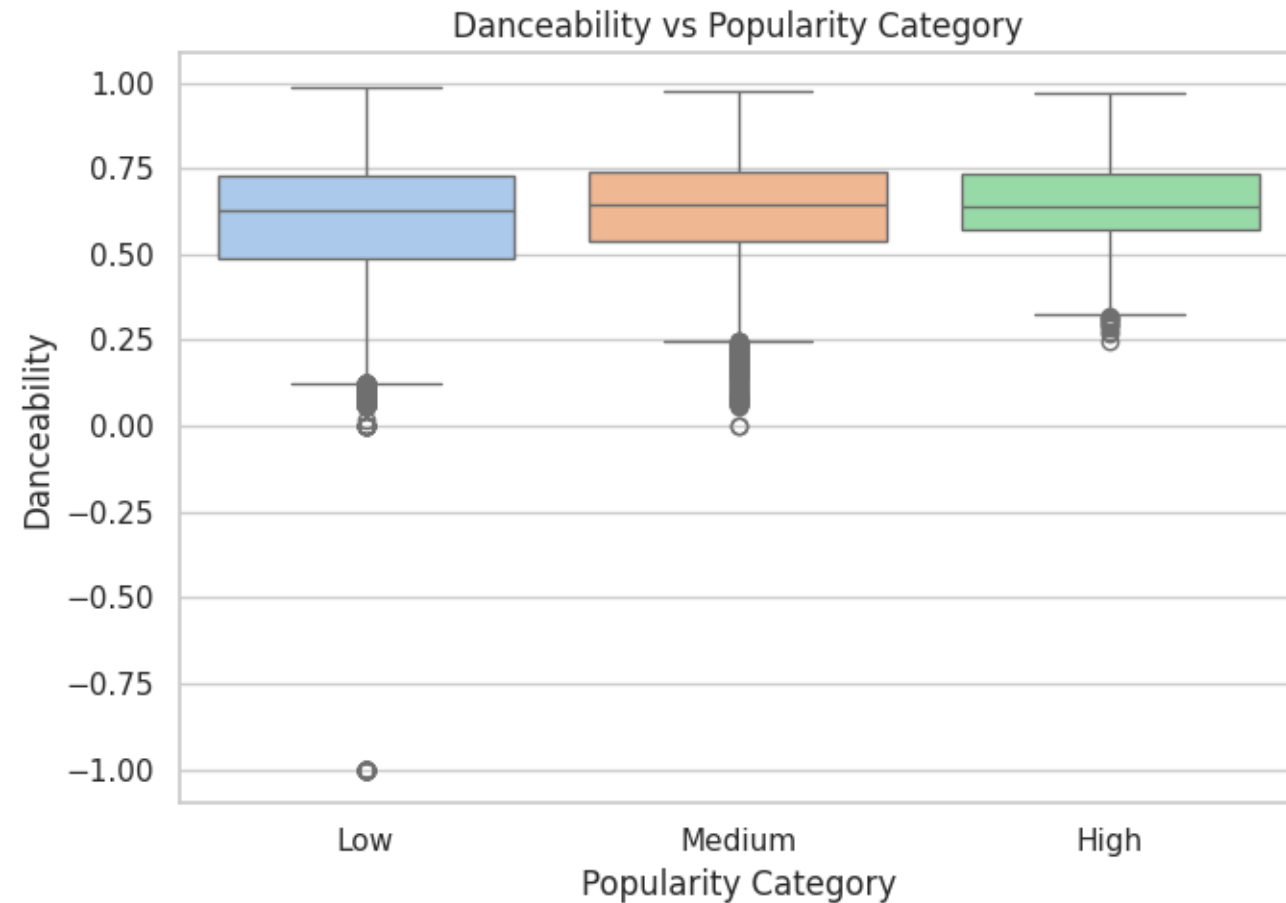
DANCEABILITY VS TEMPO

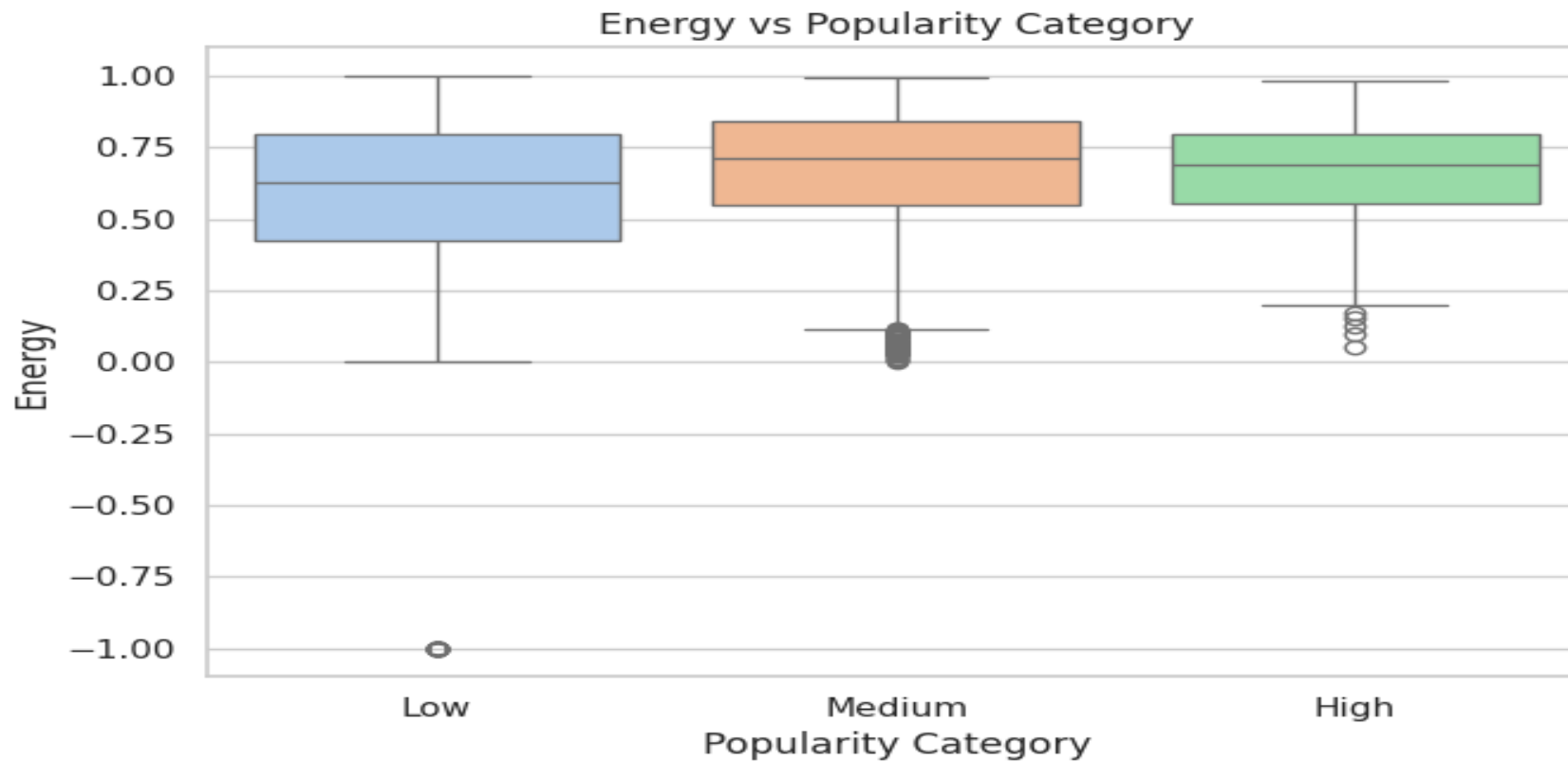
- **Speed doesn't matter.** A song's **speed (Tempo)** has little **correlation** with its **Danceability** score.
- **Danceable music is everywhere.** Highly danceable songs exist **across all common Tempos** (e.g., fast or slow).
- **Not a predictor:** Tempo will **not be a good predictor** for a song's Danceability.



ENERGY VS POPULARITY CATEGORY

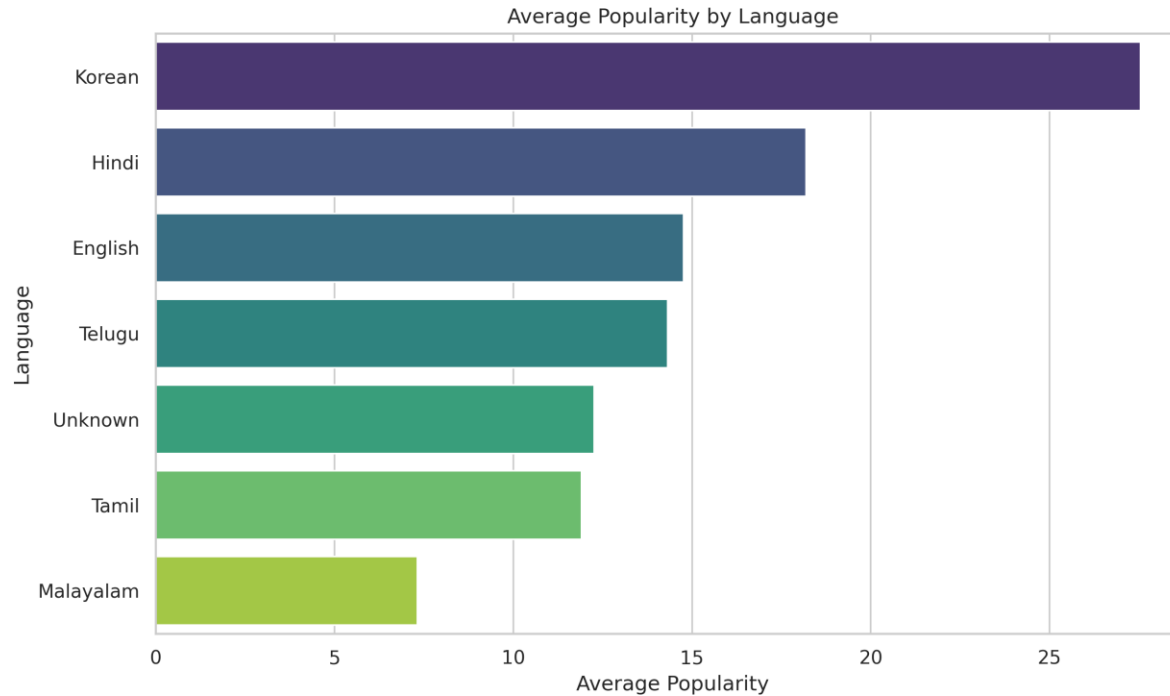
- **More Popular, More Energy:** High and Medium popular songs have a **higher median Energy** (≈ 0.7) than Low popular songs (≈ 0.6).
- **Low Energy is a Miss:** Songs with **low Energy** are concentrated in the 'Low' Popularity category; high-Energy songs dominate the hits.
- **Energy Range is Narrow:** The range of Energy is **tight** for Medium and High hits (≈ 0.5 to 0.8), suggesting a **sweet spot for Energy** that drives success.





DANCEABILITY VS POPULARITY CATEGORY

- **Danceability is Crucial:** High and Medium popular songs have a **higher median Danceability** (≈ 0.65) than Low popular songs (≈ 0.60).
- **Hits are More Consistent:** The **interquartile range (the box)** for High Popularity is slightly narrower and higher, meaning successful songs are consistently more danceable.
- **Low-Score Filtering:** Songs with very low Danceability are mostly in the 'Low' category, reinforcing it as a **good initial filter** for potential hits.

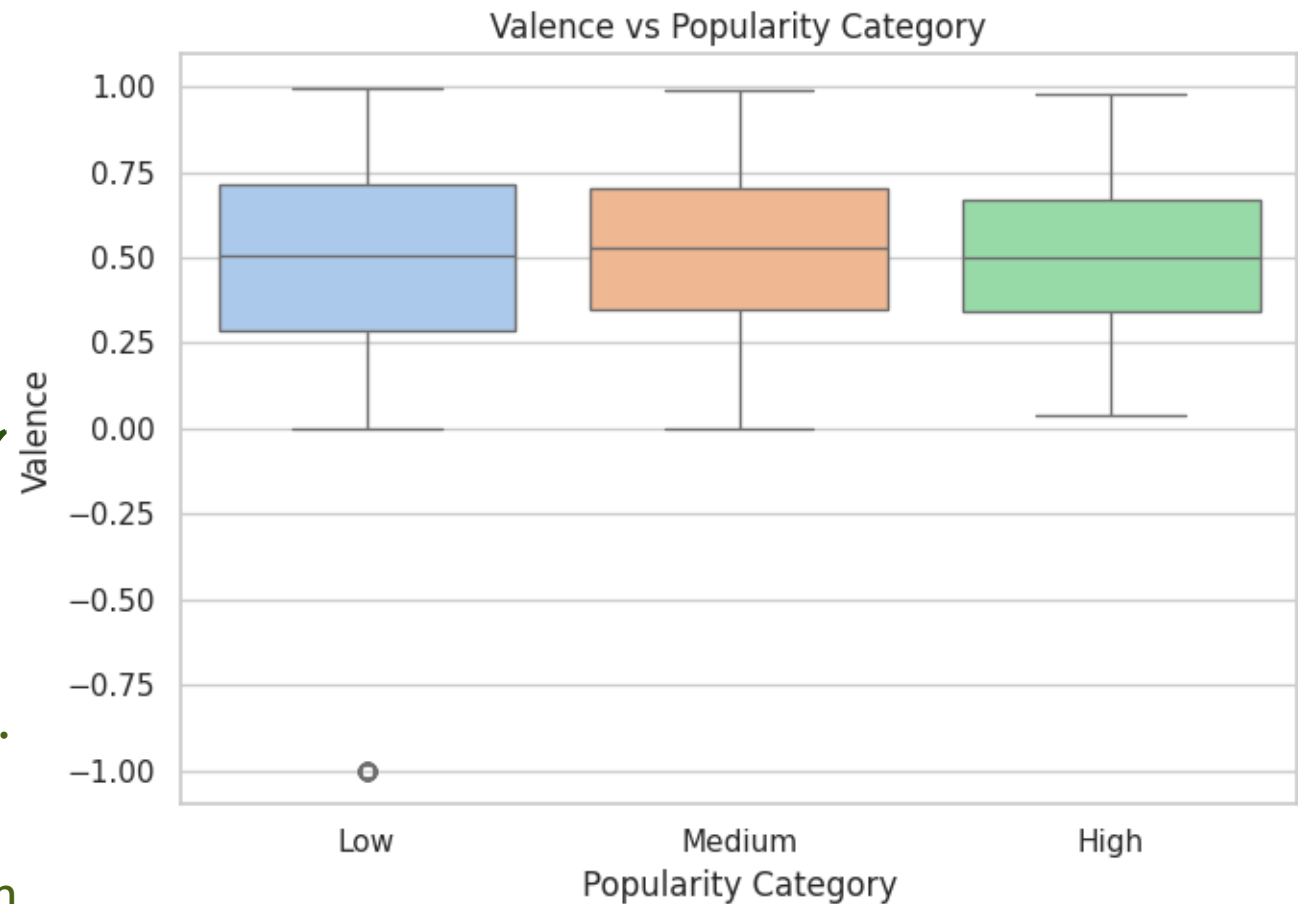


AVERAGE POPULARITY VS LANGUAGE

- **Korean Dominance:** Songs in Korean have the highest average popularity, significantly outperforming all other languages (≈ 27.5).
- **Lowest Popularity:** Malayalam has the lowest average popularity among the listed languages (≈ 7.0).
- **High Popularity Tier:** Hindi follows as the second most popular language (≈ 18.5), establishing a high-popularity tier with Korean.

VALENCE VS POPULARITY CATEGORY

- **Mood Doesn't Predict Category:** The median Valence is almost identical (≈ 0.5) across all three Popularity categories (Low, Medium, High).
- **Full Emotional Range:** All categories have a similarly wide range of emotions (Valence) from very sad to very happy.
- **Modeling Confirmation:** The box plots confirm the scatter plot finding: Valence has minimal independent power to predict if a song will be a hit.



CONCLUSION

- **SHIFT GOAL TO CLASSIFICATION:** Stop predicting scores; start classifying songs as "Hit" or "Miss."
- **PURGE COMPOSER TRACKS:** Remove all film score entries to focus your model only on standard popular music.
- **MANDATORY THRESHOLD FILTERS:** Discard any song that is too low in Energy or Danceability before modeling.
- **COMBINE EMOTIONAL FEATURES:** Use the interaction of Energy and Valence (mood) together, as mood alone is weak.
- **SIMPLIFY REDUNDANT FEATURES:** Drop one feature from highly correlated pairs (like Loudness or Valence) to streamline the final model.

The background features abstract, overlapping geometric shapes in various shades of green, primarily on the left and right sides, creating a modern, layered effect. The central area is white.

THANK YOU