Statistics 208 Project Report

The Evolution and Popularity of Music Over Time

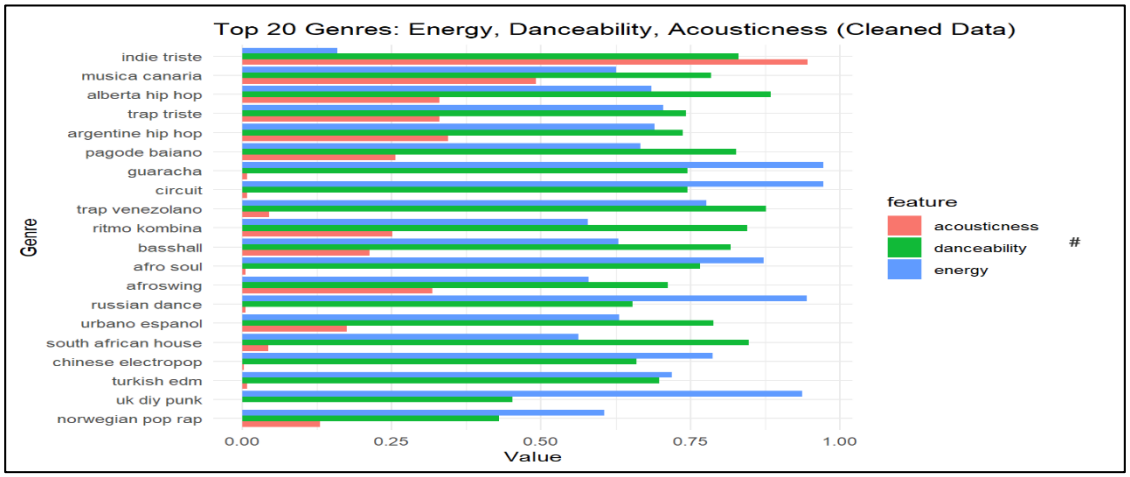
Topic Motivation: Our team was intrigued by how music styles have evolved and how certain audio characteristics contribute to a track's success. Streaming platforms and social media have transformed music discovery, and we were curious about identifying patterns behind popular tracks and understanding their changes over time. This analysis can help artists, producers, and marketers make informed decisions.

Research Objectives:

1) **Objective 1:** Identify key musical features linked to track popularity using multiple linear regression and feature importance from tree-based models.
2) **Objective 2:** Analyze how popularity and musical features have changed over time through time series plots, regression with year as a predictor, and PCA for feature evolution.

Exploratory Data Analysis

To gain initial insights into the dataset and support our research objectives, we performed a thorough exploratory data analysis using both time-based and feature-based visualizations. These graphs helped us understand trends, relationships, and potential predictors of track popularity.



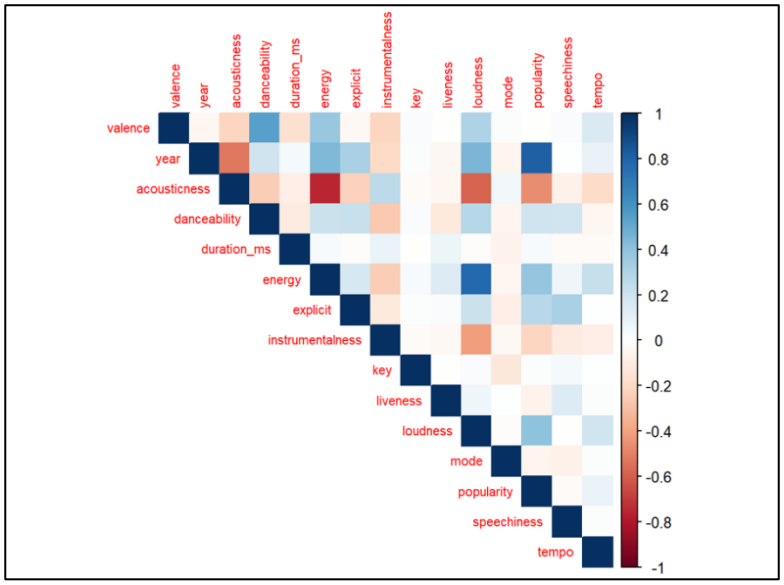
This grouped bar chart compares the top 20 genres across three features: energy, danceability, and acoustic Ness.

Genres like trap triste, Argentine hip hop, and urbano español are high in energy and danceability, suggesting strong commercial potential.

In contrast, musica canaria and indie triste show high acousticness, indicating a softer, more natural sound often associated with less mainstream styles.

This analysis helps visualize how musical traits vary by genre and identifies which genres cluster around features associated with popularity.

This genre-level comparison reinforces findings from regression and tree-based models, further validating Objective 1.

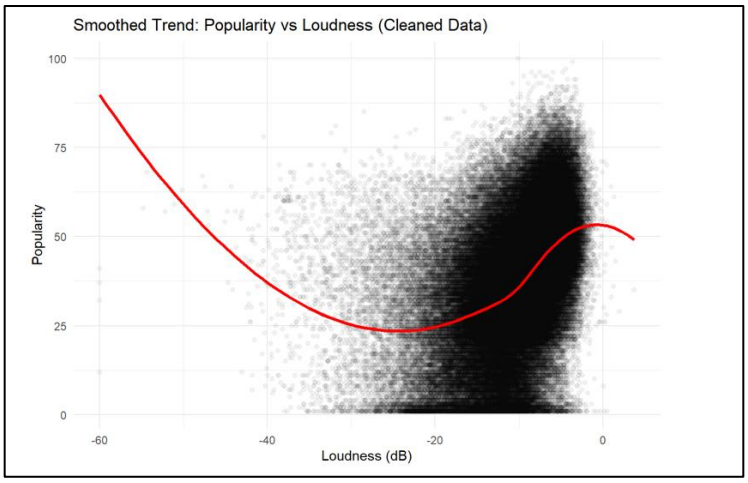


The heatmap shows pairwise correlations between all numerical features.

Danceability, energy, and loudness are positively correlated with popularity, supporting their inclusion in regression modeling.

Speechiness, acousticness, and instrumentalness show negative correlations with popularity, suggesting less “instrumental” or spoken-word content may be favored in popular tracks.

Other notable relationships include a strong positive correlation between energy and loudness, and a negative correlation between valence and speechiness.

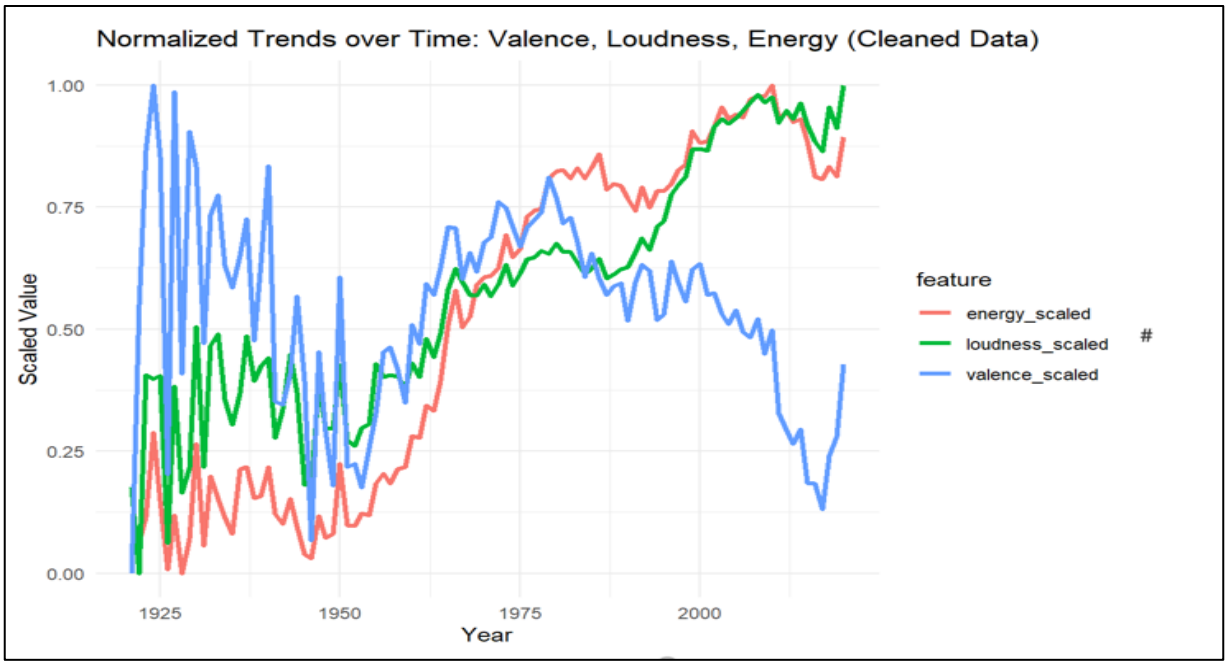


This scatterplot uses a smoothed red line to show the relationship between loudness and popularity.

The trend reveals a non-linear relationship: while extremely quiet and extremely loud tracks tend to be less popular, songs with moderately high loudness (around -5 to -2 dB) are more likely to be popular.

This supports the idea that modern mastering techniques (e.g., loudness normalization) may contribute to making tracks more appealing.

This insight connects with both Objectives 1 and 2, as it informs how loudness affects popularity and how this effect may have intensified in recent years.



This time-series line graph shows how the normalized values of valence, loudness, and energy have evolved from 1921 to 2020.

Energy and loudness have both increased steadily since the 1960s, peaking in the 2000s, reflecting a shift toward more powerful and high-intensity music.

Valence, on the other hand, shows a decline post-2000, suggesting that modern music may be less emotionally "happy" despite being more energetic and louder.

The graph indicates a historical trend where popular music has become louder and more energetic over time, possibly due to production trends and audience preferences.

This visualization aligns with Objective 2, highlighting how core musical attributes have shifted across decades.

Summary of Insights:

- Over the past century, music has become louder, more energetic, and less positive in valence.
- Popularity is most positively associated with danceability, loudness, and energy.
- Genres vary significantly in feature composition, helping to explain trends in audience preferences.
- These visual trends and correlations provided a strong foundation for building predictive models and validating assumptions through regression and random forest analysis.

Data Modeling

For Objective 1:

To determine which musical attributes most strongly influence a track's popularity, we conducted multiple predictive modeling techniques on the cleaned dataset.

- **Multiple Linear Regression** served as the baseline model. It achieved an adjusted R^2 of 0.3088 and an RMSE of 15.29 on the test set. Several predictors were found to be statistically significant at the 0.001 level, including:
    - **Danceability** ($\beta = 3.97$, $p < 2e\text{-}16$)
    - **Loudness** ($\beta = 3.30$, $p < 2e\text{-}16$)
    - **Energy** ($\beta = 1.18$, $p < 2e\text{-}16$)
    - **Valence** had a negative coefficient ($\beta = -5.23$, $p < 2e\text{-}16$), suggesting that higher valence (which corresponds to more positive, happy, or upbeat mood) was associated with lower predicted popularity in this linear context. This may seem counterintuitive but could reflect complex genre or audience preferences where popular tracks might have a more varied emotional tone rather than purely upbeat characteristics.
    - **Speechiness**, and **Liveness** also showed negative relationships with popularity.

  These results highlight that tracks with higher danceability, loudness, and energy are more likely to be popular, while features like speechiness and valence may reduce predicted popularity in a linear context.

  Although the model explains approximately 31% of the variance, it provides valuable interpretability and insight into the direction and statistical significance of individual predictors.

- To improve model generalization and address potential multicollinearity, **Lasso** and **Ridge Regression** were applied with cross-validated regularization:
    - The best lambda for Lasso was 0.0185, which shrank several coefficients toward zero while retaining key predictors.
    - The best lambda for Ridge was 0.8589, which retained all variables but reduced the influence of less relevant features. Both models achieved RMSE values on the test set very close to the linear model, at 15.29 (Lasso: 15.29, Ridge: 15.29), indicating comparable predictive performance with added benefits of regularization.
- To capture complex non-linear interactions and potentially improve prediction accuracy, a **Random Forest** model was implemented. It achieved a significantly lower test RMSE of 13.87, demonstrating better predictive performance than all linear-based models. Feature importance scores from the Random Forest identified loudness, energy, acousticness, and valence as the most influential predictors of popularity. Notably, valence's importance in the Random Forest suggests that while its relationship to popularity is not strictly linear, it plays a key role in combination with other features in influencing popularity outcomes.

Together, these modeling approaches provided complementary insights. The linear models offered interpretability by revealing the direction and strength of influence of each attribute, including the intriguing negative linear association of valence with popularity. The regularized models helped confirm key predictors while managing multicollinearity. Finally, the Random Forest model improved overall prediction accuracy and highlighted the importance of complex, potentially non-linear effects of features like valence.

The Random Forest's lower RMSE suggests it is the optimal model for predicting track popularity on this dataset, particularly when prediction accuracy is the primary goal. However, the interpretability of linear models remains valuable for understanding the underlying relationships, especially for exploring how emotional characteristics such as valence influence popularity.

Objective 2:

To examine how musical features and popularity have evolved over time, we employed time series visualization, linear regression, and Principal Component Analysis (PCA).

- **Time Series Visualization** was used to observe trends in audio features across decades. Faceted line plots showed distinct temporal patterns in key attributes such as:
    - *Danceability, Energy, and Valence*: Generally exhibited upward trends over time.
    - *Loudness*: Increased consistently, suggesting modern music is louder on average.
    - *Tempo and Popularity*: Revealed subtle yet meaningful long-term shifts.
- **Linear Regression** was applied to assess whether the *year* significantly predicts *popularity*:
    - The model yielded an adjusted R^2 of 0.9492 and a p-value < 2e-16, indicating a very strong linear relationship.
    - The regression coefficient for *year* was 0.6954, suggesting a consistent annual increase in track popularity over time.
    - These results support the hypothesis that newer music is systematically more popular in the dataset.
- **Principal Component Analysis (PCA)** was conducted to reduce dimensionality and detect dominant patterns of change in musical attributes over time:
    - The first two principal components explained a substantial proportion of the variance (PC1 = 54.6%, PC2 = 14.3%).
    - The PCA biplot showed a clear trajectory of evolution, with features such as *energy*, *danceability*, and *tempo* increasing in prominence in more recent years.
    - Earlier years clustered separately from recent years, affirming shifts in musical styles and production trends.

Conclusion:

By integrating both predictive modeling and temporal analysis, this study offers a holistic understanding of what drives music popularity and how these drivers have evolved over time. Regression techniques, including Multiple Linear, Lasso, and Ridge models, consistently identified danceability, energy, and loudness as strong positive predictors of popularity. Interestingly, valence and speechiness demonstrated a negative association, suggesting that tracks with a more emotionally complex or less "happy" tone may actually resonate more broadly in certain contexts — potentially reflecting trends in genres, listener preferences, or even streaming algorithms that favor emotionally nuanced tracks.

The Random Forest model outperformed all other approaches with the lowest RMSE of 13.87, underscoring the importance of capturing non-linear interactions between musical features when the goal is accurate prediction. The model also highlighted acousticness and valence as key variables, reaffirming that popularity is influenced by more than just tempo or volume — mood, texture, and production style matter, too.

Beyond static modeling, our time series and PCA analyses revealed that popular music has undergone substantial shifts over recent decades. Attributes such as energy, danceability, and loudness have increased, suggesting that modern music is becoming more upbeat, rhythm-driven, and production-heavy. These changes likely reflect broader technological advances, evolving listener tastes, and the impact of digital platforms in shaping musical trends.

Together, these insights offer both academic and practical value. They not only inform how we understand musical success in a data-driven era, but also provide valuable guidance for artists, producers, and music platforms seeking to align with evolving listener expectations. Future research could explore how these patterns vary across genres, cultures, or platforms, and whether similar dynamics apply to emerging music forms such as AI-generated compositions or short-form viral audio.

Ultimately, music popularity is not just a product of catchy melodies or fast beats — it’s shaped by a complex, evolving interplay of emotional, acoustic, and cultural signals that reflect the times we live in.

Recommendations:

Based on both objectives, we recommend the following:

- For musicians and producers: Focus on enhancing danceability, energy, and loudness to increase a track's popularity. However, be mindful that extremely high valence (overly happy/upbeat tone) may not always correlate with popularity—audiences may prefer more nuanced emotional tones.
- For streaming platforms and analysts: Improve algorithmic recommendations by prioritizing tracks with optimal combinations of energy, loudness, and moderate valence, as identified by both linear and Random Forest models.
- For future researchers: Extend this analysis to genre- or artist-level data to uncover deeper segment-specific trends, and use non-linear models (e.g., QDA) to capture complex interactions between musical features and popularity over time.