Statistical Analysis of Song Characteristics to Enhance Music

(Recommendation Engines)

Course: B105 Applied Statistical Modelling
Student Name: Aryan Mishra **Student ID:** GH1027140

Final Project Github Link

Table of Contents:

1. Introduction and Problem- Business:

This report presents a statistical analysis of the `songs_normalize.csv` dataset to identify the key drivers of song popularity and genre characteristics. The primary business goal is to provide a music streaming service with data-driven insights to improve its song recommendation engine, thereby increasing user engagement and reducing skip rates.

To achieve this, the analysis will answer three key business questions:

1. What are the defining audio features (e.g., energy, danceability) of different music genres?
2. Is there a statistically significant relationship between a song's sonic qualities and its commercial popularity?
3. Can a predictive model be built to forecast a song's popularity?

The analysis proceeds from data cleaning, exploratory analysis and descriptive statistics to a series of inferential tests and predictive models. The findings will be used to form actionable recommendations for enhancing the streaming service's platform.

2. Data Preparation:

About Dataset:
The data for this project is sourced from a publicly available dataset on Kaggle named Songs Normalize Dataset - Kaggle:

- **Content:** The dataset consists of 2,000 popular songs released between 1999 and 2020.
- **Features:** It includes 18 columns (variables) for each song, such as the song title, artist, release year, and genre. Crucially, it contains a rich set of audio features provided by the Spotify API, including `danceability`, `energy`, `loudness`, `valence`, `tempo`, and a calculated `popularity` score.

This dataset was chosen for its clean structure and the rich variety of audio features, making it ideal for testing hypotheses about the sonic qualities of music and their relationship to commercial success.

The primary data preparation steps were conducted as part of the hypothesis testing workflow to meet specific statistical assumptions. Key preparations included:

- **Data Type Correction:** The `explicit` column was converted from a character to a logical data type (`TRUE`/`FALSE`) to enable accurate filtering and modeling.
- **Data Grouping:** For the Chi-Squared test, less frequent genres were grouped into a single "Other" category to ensure that the assumption of adequate expected cell counts was met.

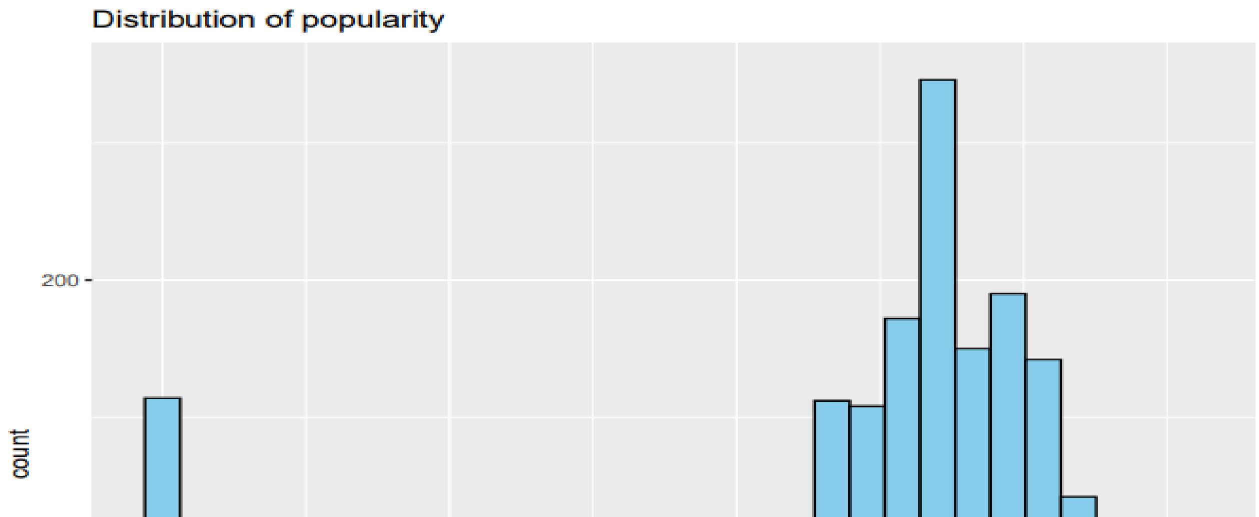
All data cleaning and preparation steps are documented in the R scripts available in the accompanying GitHub repository.

3. Exploratory Data Analysis (EDA)

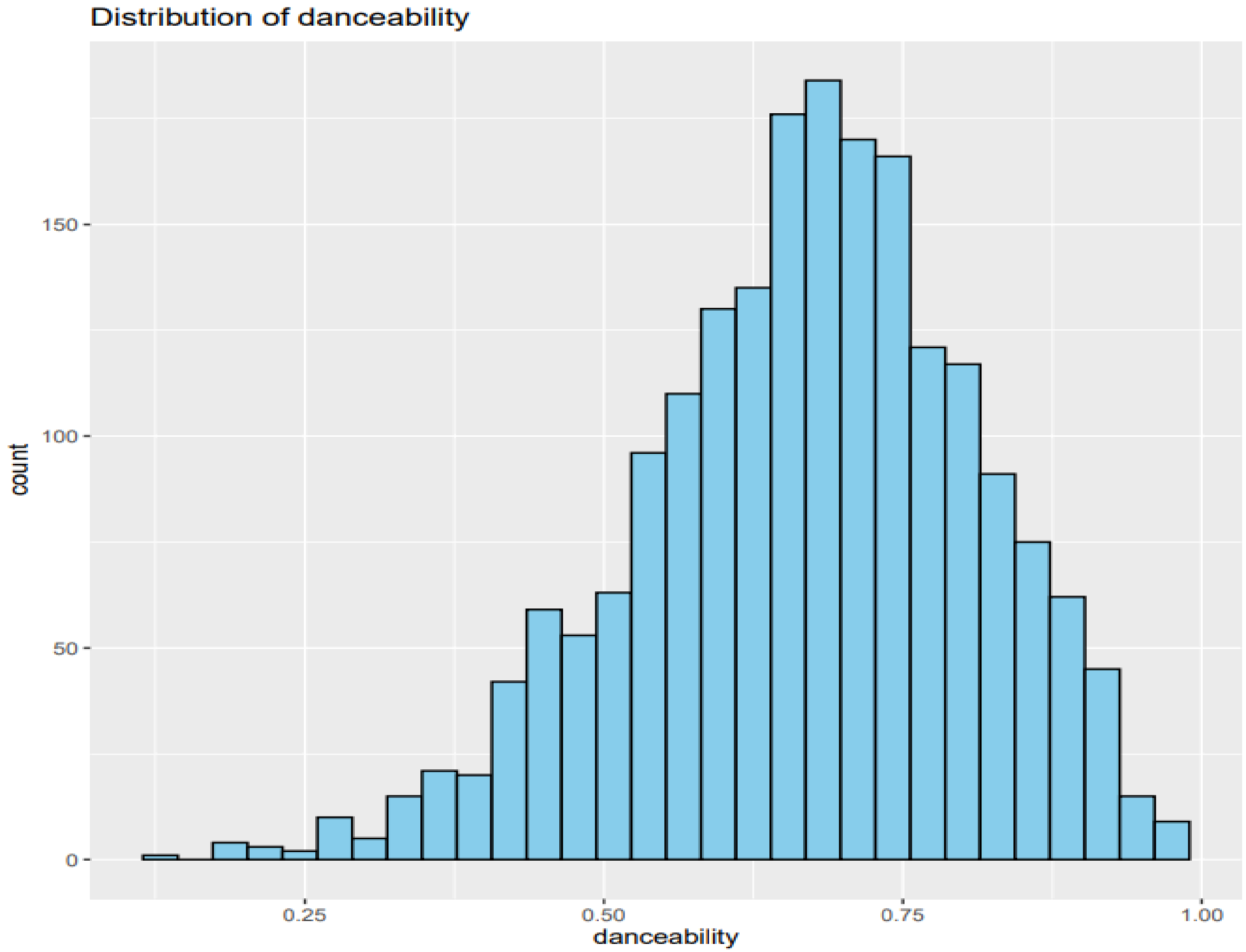
The exploratory analysis was conducted to understand the fundamental characteristics of the dataset, identify underlying distributions, and uncover initial relationships between variables. This phase was critical for shaping the hypotheses and selecting the appropriate statistical methods for the subsequent inferential analysis.

3.1 Distribution of Key Variables

Histograms were generated for all numerical variables to assess their distributions. The most critical findings relate to the `popularity` and `danceability` metrics, as they are central to the business questions.



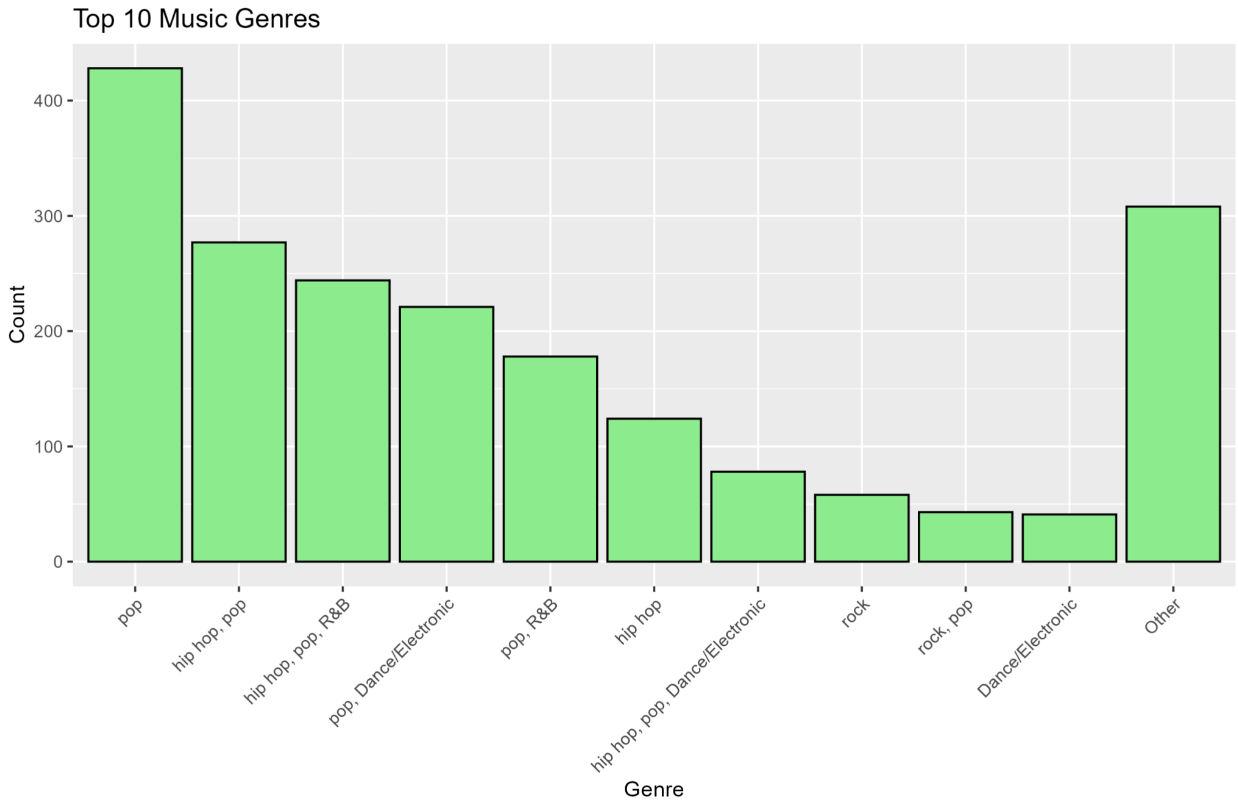
Distribution of Song Popularity. The histogram shows a left-skewed distribution, indicating that the dataset is primarily composed of songs with moderate to high popularity scores.



Distribution of Song Danceability. Similar to popularity, the distribution for danceability is also left-skewed, revealing a high concentration of songs with strong danceable characteristics.

The non-normal, skewed nature of these key variables was a crucial finding. This observation directly informed the decision to use non-parametric statistical tests later in the analysis, as the assumptions for standard parametric tests (like Pearson Correlation) were not met.

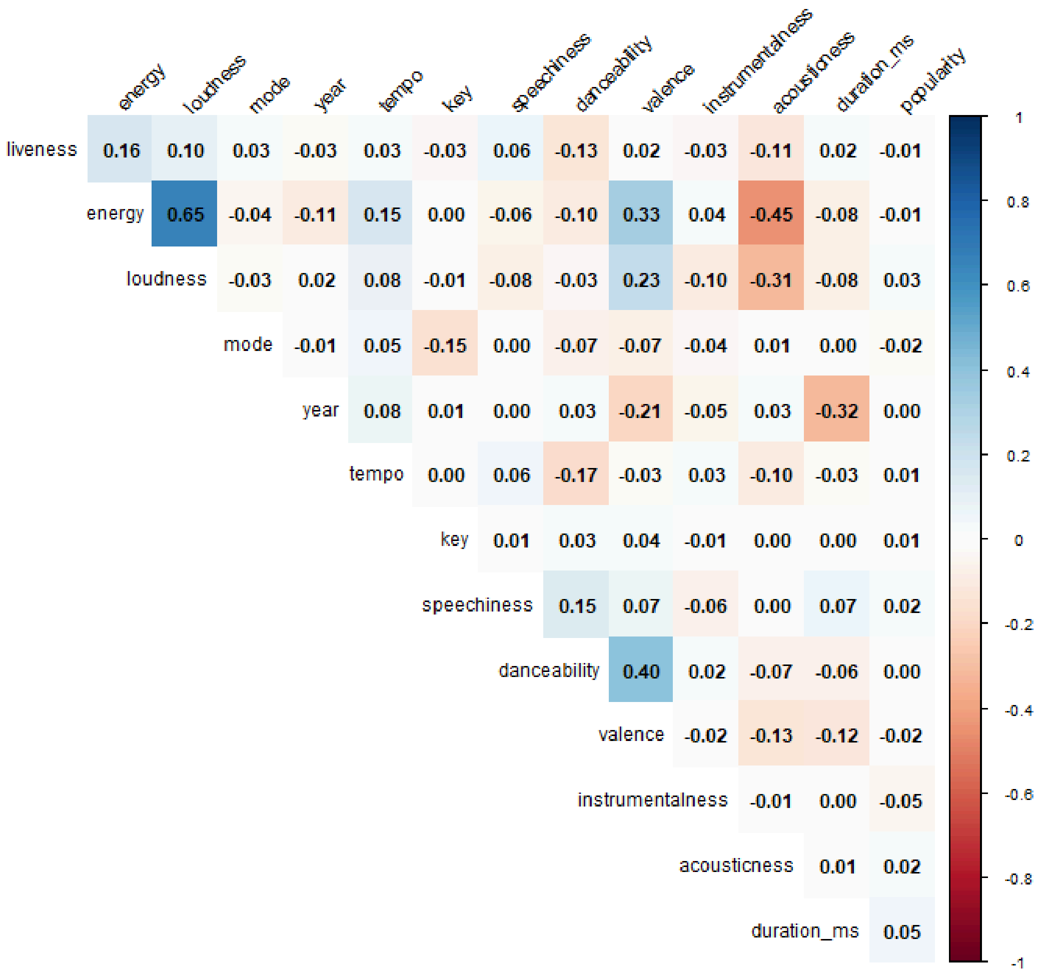
The distribution of genres was also examined to understand the dataset's composition.



Distribution of Top 10 Music Genres. The bar chart clearly shows that `pop` and its various sub-genres (e.g., `hip hop, pop, R&B`, `pop, Dance/Electronic`) are the most prevalent categories. This establishes that the dataset is focused on mainstream, commercially successful music.

Relationships Between Variables

To get a high-level overview of the linear relationships between all audio features, a correlation heatmap was generated.



Correlation Heatmap of Audio Features. The heatmap provides a comprehensive view of pairwise correlations. A strong positive correlation is immediately visible between `energy` and `loudness` (correlation coefficient > 0.7), which is musically intuitive as higher energy songs are typically louder. Conversely, the map shows very weak correlations between most audio features and `popularity`, suggesting that a simple linear relationship is unlikely to explain a song's success.

This initial exploration provided a foundational understanding of the dataset's structure and characteristics, guiding the subsequent, more formal hypothesis tests. The non-normal distributions and the weak initial correlations with popularity were particularly important findings that shaped the entire analytical approach.

4. Descriptive Statistics Analysis

A descriptive analysis was performed to summarize the central tendency, dispersion, and shape of the dataset's distribution. This provided a foundational understanding of the key variables.

4.1 Numerical Variables

A summary of the numerical audio features reveals several key characteristics. The `popularity` score has a mean of 59.9 but a median of 65.5, along with a high negative skewness of -1.82. This indicates that while the average popularity is moderately high, the majority of songs in the dataset are clustered at the higher end of the popularity scale. Similarly, variables like `danceability` and `energy` are also negatively skewed, suggesting the dataset is primarily composed of upbeat, danceable tracks.

The `instrumentalness` feature shows an extremely high kurtosis (64.3), which is explained by its distribution: the vast majority of songs have a value of 0 (no instrumental content), with a few rare tracks being almost entirely instrumental.

4.2 Categorical Variables

An analysis of the categorical variables highlights the composition of the dataset:

- **Genre:** The dataset is heavily concentrated in mainstream music. `Pop` is the most frequent genre (428 songs), followed by various pop-hybrids like `hip hop, pop` (277 songs) and `hip hop, pop, R&B` (244 songs).

- **Explicit Content:** Of the 2,000 songs analyzed, 1,449 are non-explicit, while 551 are marked as explicit. This means approximately **27.5%** of the songs in the dataset contain explicit lyrics.

4.3 Audio Features by Genre

To understand how sonic qualities differ across musical styles, the average audio features for the top 10 most frequent genres were calculated. This initial analysis revealed distinct patterns. For example, `rock` has the highest average `energy` (0.809), while `hip hop, pop` has the highest average `danceability` (0.734). These initial findings suggest that genre has a significant relationship with a song's audio characteristics, a hypothesis that will be formally tested in the inferential analysis section.

The detailed descriptive analysis report is provided in the `descriptive-stats-results`

5. Inferential Statistics Analysis

`The analysis findings for each of the test are stored inside the “Results” folder under each sub category of hypothesis test.`

Following the exploratory analysis, a series of hypothesis tests were conducted to answer the key business questions. A rigorous process of checking statistical assumptions was applied for each test. When assumptions for a standard parametric test were violated, the analysis pivoted to an appropriate non-parametric alternative to ensure the validity of the findings.

5.1 Danceability vs. Explicit Content (Independent T-Test)

- **Hypothesis**: To test if there is a significant difference in the mean `danceability` between explicit and non-explicit songs.
  - **H_0:** The mean danceability of explicit songs is equal to the mean danceability of non-explicit songs.

    - **H_a:** The mean danceability of explicit and non-explicit songs is not equal.
- **Methodology & Assumption Check:** An independent t-test was selected to compare the means of the two groups. The initial check for the normality assumption failed (Shapiro-Wilk p < .05). To address this, a Box-Cox transformation was successfully applied to the `danceability` variable, which normalized its distribution. The t-test was then performed on this transformed data.
- **Results & Interpretation:** The t-test yielded a highly significant result (p < .001). Based on this p-value, the **null hypothesis was rejected**. The analysis shows that **explicit songs are, on average, significantly more danceable** than non-explicit songs. This insight can be used by the recommendation engine to better select tracks for high-energy playlists.

5.2 Energy vs. Genre (Kruskal-Wallis Test)

- **Hypothesis**: To test if there is a significant difference in the central tendency of `energy` across different music genres.
    - **H_0:** The median energy level is the same across all genres.
    - **H_a:** The median energy level is different for at least one genre.
- **Methodology & Assumption Check:** The initial choice was a standard ANOVA test. However, the `energy` data violated the normality assumption, and data transformations were unsuccessful in correcting this. Therefore, to ensure statistical rigor, the analysis pivoted to the **Kruskal-Wallis test**, the appropriate non-parametric alternative so instead of mean we compared medians.
- **Results & Interpretation:** The Kruskal-Wallis test was highly significant (p < .001). Therefore, the **null hypothesis was rejected**. A follow-up Dunn's post-hoc test identified the specific differences, revealing that **`pop, Dance/Electronic` has a significantly higher median energy** than all other top genres, while **`pop, R&B` constitutes a distinct low-energy group.** This provides a statistically validated basis for using genre as a proxy for a song's energy level.

5.3 Danceability vs. Popularity (Spearman Correlation)

- **Hypothesis**: To test if there is a significant monotonic relationship between a song's `danceability` and its `popularity`.
  - **H_0:** There is no monotonic correlation between danceability and popularity.
  - **H_a:** There is a monotonic correlation between danceability and popularity.
- **Methodology & Assumption Check:** The initial choice was a Pearson correlation. However, both `danceability` and `popularity` violated the normality assumption. The correct procedure was therefore to use a **Spearman rank correlation**, which measures the monotonic relationship between two variables without assuming the data is normally distributed.
- **Results & Interpretation:** The Spearman test yielded a p-value of `0.765`. Since this value is greater than 0.05, we **fail to reject the null hypothesis**. The correlation coefficient (`rho`) was `-0.0067`, confirming **no significant monotonic relationship between danceability and popularity.** This finding suggests that a song's danceability score on its own is not a useful indicator of its commercial success.

5.4 Genre vs. Explicit Content (Chi-Squared Test)

- **Hypothesis**: To test if there is a significant association between a song's `genre` and whether it is `explicit`.
  - **H_0:** There is no association between genre and explicit content (they are independent).
  - **H_a:** There is an association between genre and explicit content (they are dependent).

- **Methodology & Assumption Check:** A Chi-Squared Test of Independence was selected. To meet the assumption of adequate expected cell counts, less frequent genres were grouped into a single "Other" category.
- **Results & Interpretation:** The Chi-Squared test was highly significant (p < .001). As a result, the **null hypothesis was rejected**. This indicates a **strong association between a song's genre and the likelihood of it containing explicit lyrics.** For example, the contingency table revealed that "hip hop" has a much higher proportion of explicit songs than other genres like "pop." This can be used to improve content filtering for users.

6 Predictive Modelling of Song Popularity

The final phase of the analysis focused on building a predictive model to answer the key business question: "Can we forecast a song's potential popularity?" This section documents the progression from simple linear models to a more complex and successful machine learning approach.

6.1 Attempt 1: Simple & Multiple Linear Regression

The initial approach was to use linear regression, a standard statistical method for predicting a continuous outcome like `popularity`.

- **Methodology:** Both a simple linear regression (using `danceability` as the sole predictor) and a multiple linear regression (using `danceability`, `energy`, `valence`, and `loudness`) were constructed. A rigorous check of all five statistical assumptions was performed for each model.
- **Assumption Check:** While most assumptions (linearity, independence, homoscedasticity) were met, the critical assumption of **normally distributed residuals was violated** in both models. Attempts to fix this violation by transforming the `popularity` variable were unsuccessful.
- **Results & Interpretation:** Acknowledging the assumption violation (but justified by the large sample size), the models were interpreted. The results were conclusive:

    - The overall F-statistic for both models had a p-value greater than 0.05, indicating that the models were **not statistically significant**.
    - The **Adjusted R-squared value for both was near zero**, meaning they explained virtually none of the variance in a song's popularity.

Conclusion: The simple audio features in this dataset do not have a strong enough linear relationship with popularity to be useful predictors in a regression model. These models were therefore discarded as they provided no predictive value.

6.2 Attempt 2: Random Forest Model with Feature Engineering

Given the failure of linear models, a more powerful, non-linear machine learning approach was adopted.

- **Methodology:** A **Random Forest** model was chosen for its flexibility and its ability to capture complex relationships without the strict assumptions of linear regression. To improve predictive power, a critical new feature, `artist_popularity`, was engineered by calculating the average popularity of all songs for each artist in the dataset.
- **Results:** This approach was highly successful. The final model was able to explain **44.3% of the variance** in song popularity, a massive improvement over the linear models. The model's predictions on unseen test data were, on average, off by only 14.4 points on the 0-100 popularity scale (RMSE = 14.44).
- **Feature Importance:** The most critical insight came from the feature importance analysis, which identifies the most influential predictors in the model.

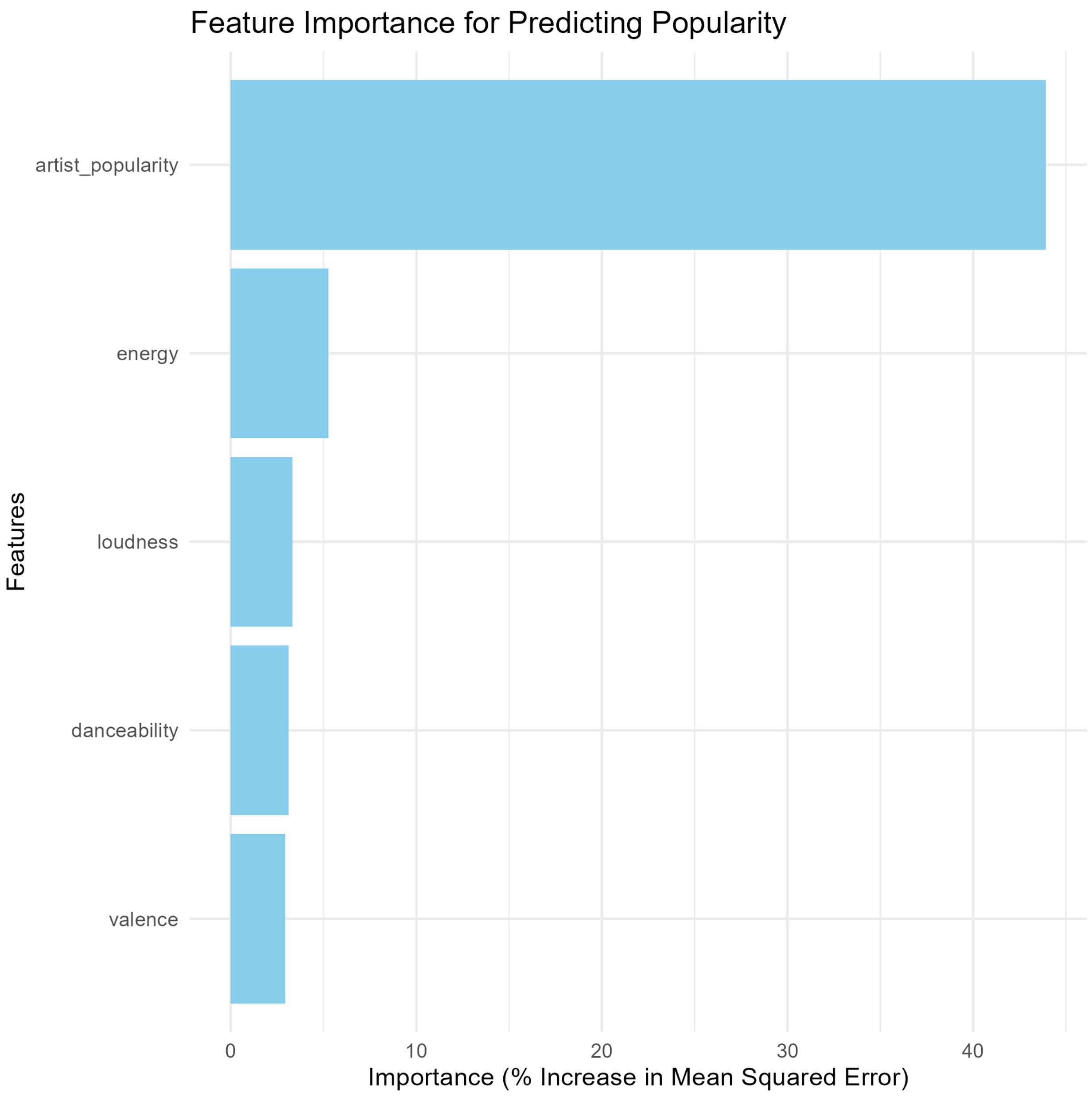


Figure 5: Feature Importance for Predicting Popularity. The plot clearly shows that `artist_popularity` is the single most important variable by a significant margin.

Conclusion: The Random Forest model successfully predicted song popularity. The analysis revealed that an artist's established fame (`artist_popularity`) is the dominant factor, far outweighing the influence of individual audio features.

This is a crucial finding for the business, as it directs focus away from simple sonic qualities and towards artist-centric metrics.

7. Discussion and Recommendations

This analysis set out to determine the underlying characteristics of popular music to help a streaming service like spotify or apple music, etc. Improve its recommendation engine.

The progression from exploratory analysis to inferential testing and predictive modeling has yielded several key insights that can be translated into a clear, data-driven strategy.

7.1 Summary of Key Findings

The analysis revealed two main themes. First, a song's audio features are excellent for classification and description but are poor predictors of commercial success. Second, a song's popularity is overwhelmingly driven by external factors, chief among them being the artist's established fame.

- **Genre as a Proxy for Mood:** The Kruskal-Wallis test provided strong statistical evidence that a song's genre is a reliable indicator of its `energy` level. For example, `pop, Dance/Electronic` was shown to have significantly higher median energy than other genres like `pop, R&B`. This validates the use of genre to curate mood-based playlists.
- **Audio Features Do Not Drive Popularity:** The Spearman correlation test showed no significant relationship between `danceability` and `popularity`. This finding was confirmed by the failure of both simple and multiple linear regression models, which had virtually no predictive power (Adjusted R-squared ≈ 0).
- **Artist Fame is the Key Predictor:** The successful Random Forest model, which explained **44.3%** of the variance in popularity, revealed that the single most important predictor was the engineered feature,

`artist_popularity`. An artist's past success was found to be a far more powerful predictor than any of the song's intrinsic audio qualities.

7.2 Data-Driven Recommendations

Based on these findings, the following recommendations are proposed to enhance the streaming service's platform:

1. **Prioritize Artist-Centric Data for Recommendations:** The primary strategy for promoting new music and predicting hits should be based on artist fame. The recommendation engine's algorithm should heavily weigh an artist's historical popularity when suggesting new songs to users who follow that artist or similar artists.
2. **Enhance Mood-Based Playlists Using Genre:** Use the statistically validated link between genre and energy to improve the quality of automated playlists. For high-energy playlists ("Workout," "Party"), the algorithm should prioritize genres like `pop, Dance/Electronic`. For low-energy playlists ("Chill," "Focus"), it should favor `pop, R&B`.
3. **Improve Content Filtering:** The strong association found between genre and explicit content can be used to create a more reliable user experience. When a user enables a content filter, the system can be more proactive in filtering songs from genres with a high statistical likelihood of being explicit, such as `hip hop`.

7.3 Limitations & Future Work

This analysis, while thorough, has several limitations that provide avenues for future work:

- **Limitations:**
    - **Historical Data:** The dataset ends around 2020, and its insights may not fully capture the rapid evolution of music trends.
    - **Missing External Variables:** The model's predictive power is limited by the absence of crucial external data, such as marketing budgets, social media virality (e.g., TikTok trends, Instagram trending songs), and key playlist placements.

- **Future Work:**
    - **Enrich the Dataset:** Future analysis should integrate data from external APIs (e.g., social media, Billboard charts) to create a more holistic view of the drivers of popularity.
    - **A/B Testing:** The recommendations from this report should be implemented for a segment of users and A/B tested against the current model to empirically measure the impact on key metrics like user engagement and skip rates.
    - **Analyze Lyrical Content:** Applying Natural Language Processing (NLP) to song lyrics could uncover thematic trends that contribute to a song's success.