Temporal Trends in Popular Music Genres: A Data-Driven Analysis

Elizabeth Hall - Data Analysis Project

Elizabeth Hall

2024-05-03

# 1. Abstract

This study investigates the temporal trends of popular music genres and the evolution of audio features in tracks from the Billboard Hot 100s charts between 1985 and 2023. Utilizing extensive datasets from Billboard and enriched with genre and audio feature data from the Spotify Web API, this study attempts to provide a comprehensive analysis of shifts in musical tastes and preferences among American listeners. Key findings include the sustained popularity of certain genres like pop and rock, alongside the emergent dominance of genres such as hip-hop and electronic music. Additionally the study examines changes in audio features over time, highlighting an increase in attributes such as danceability and loudness, which correlate with contemporary music production trends.

# 2. Introduction

## 2.1 Background

The landscape of popular music has undergone significant transformations since the mid-20th century, mirroring shifts in cultural, technological, and social dynamics. Studies such as those by F. Terroso-Saenz et al. [1] and Mauch, Matthias, et al. [2] have highlighted a growing scholarly and commercial interest in understanding global music trends.

The Billboard Hot 100 charts [3], established in 1985, have served as a barometer for musical popularity in the United States, chronicling the most popular genres, artists, and styles over the decades [4]. This project leverages extensive data from the Billboard charts and genre information from Spotify to examine genre trends over time.

The primary dataset for this project comprises records from the Billboard Hot 100 charts, spanning from 1958-2023. This dataset was sourced from a publicly available GitHub repository maintained by user HipsterVizNinja [5]. It encapsulates weekly rankings of songs in the United States, reflecting their commercial success and popularity. The raw dataset is structured as a CSV file titled “Hot\_100s.csv” located in the “raw-data” folder within the project repository.

To enrich this dataset with genre information, a cross-reference with the SpotifyAPI was performed. This process involved querying the SpotifyAPI for each unique artist to retrieve their associated musical genres [6]. The SpotifyAPI provides a comprehensive database of songs, artists, and their genres, making it an invaluable recourse for this analysis.

The data also underwent several cleaning and preprocessing steps to ensure its suitability for analysis.These steps included parsing and formatting dates, condensing the dataset for more efficient processing, and categorizing genres into broader classifications to simplify the analysis of genre trends.

### 2.2 The Question

This project is guided by two primary research questions:

1. **Genre Popularity**: How has the popularity of different music genres evolved over time?

This question seeks to chart the trajectory of musical genres over time, as represented in the Billboard Hot 100 charts from 1085-2023. By mapping these changes, this project aims to illustrate the shifts in musical tastes and preferences among American listeners.

1. **Audio Feature Evolution**: How have trends in audio features evolved over time?

Beyond genre preferences, this question examines the evolution of audio features within popular tracks. By analyzing aspects such as tempo, energy, danceability, and instrumentalness this project aims to reveal underlying patterns within the musical elements of popular tracks.

The ultimate goal of this project is to reveal deeper insights about the evolution of popular music, identifying temporal patterns and anomalies within the preferences of American listeners.

# 3. Methods

## 3.1 Data Acquisition

The primary dataset for this project was sourced from the Billboard Hot 100 charts, covering the period from 1958 to 2023. This dataset, which includes weekly ranking of songs based on their commercial success in the United States, was accessed via a publicly available GitHub repository managed by HipsterVizNinja. The data was stored in the file Hot\_100.csv located in the “raw-data” folder for this project’s repository.

To compliment the Billboard data with genre and audio feature information, the Spotify Web API was utilized. This API provides access to a comprehensive database of songs, artists, and associated musical genres and audio features which allowed the dataset to be enriched with additional information specific to each entry in the Billboard dataset.

## 3.2 Data Import

The raw dataset was imported into R using the read.csv function in R. This initial import involved loading the raw data from the Hot\_100.csv file, which contained several columns including artist names, song titles, chart positions, and other relevant metadata. To facilitate easier manipulation of the data, irrelevant columns such as chart\_debut, chart\_url, and song\_id were excluded during the import process.

The following R code snippet demonstrates the data import process:

data <- read.csv("./data/raw-data/Hot\_100.csv") %>%  
 select(-chart\_debut, -chart\_url, -song\_id) %>%  
 rename(artist = performer)

## 3.3 Data Cleaning

Data cleaning was a multi-step process aimed at preparing the dataset for analysis. This involved several key tasks:

1. **Date Parsing:** The chart\_date column was converted from a string format to a date object using the as.Date function. This conversion facilitated time-based analysis and grouping of data.

data$chart\_date <- as.Date(data$chart\_date)

1. **Adding Time Variables:** Additional columns for year, month, and week were created using the lubridate package to assist in more granular analyses of trends over time.

data$year <- year(data$chart\_date)  
data$month <- month(data$chart\_date)  
data$week <- week(data$chart\_date)

1. **Handling Duplicates and Aggregation:** To focus on the most significant entries, the dataset was condensed to include only the top song per week, per month, per year. This was achieved by grouping the data by these time variables and then arranging by chart position, followed by slicing the top entries.

top\_songs\_per\_week\_month\_year <- data %>%  
 group\_by(year, month, week) %>%  
 arrange(year, month, week, peak\_position, desc(time\_on\_chart)) %>%  
 slice(1) %>%  
 ungroup() %>%  
 distinct(song, .keep\_all = TRUE)

1. **Genre Enrichment:** Using the Spotify Web API, each track and artist was queried to fetch their associated genres. This step was crucial to map each song to one or more musical genres, facilitating a genre-based analysis. Since the API was being used to fetch the data, the dataset was condensed to exclude duplicate song entries, this API enhanced dataset was then used to enhance the full dataset without overloading the API.
2. **Genre Simplification:** To analyze trends more effectively the multitude of specific genres retrieved from the Spotify API were simplified into fewer, broader genre categories. This was achieved through a combination of automated processing and manual refinement. The first pass of categorization was done using unique genre labels (e.g. pop, rock, jazz, etc.) and guided by the presence of these key terms.

# Example of categorizing sub-genres into main genres  
genres <- unique\_songs$genres %>% unlist() %>% unique()  
pop\_genres <- genres[grepl("pop", genres)]  
rock\_genres <- genres[grepl("rock", genres)]  
# Additional categories were created similarly

For genres that were not clearly assignable through automated keyword matching, manual research was conducted to determine the most appropriate main genre category, ensuring that each genre label was as accurately represented as possible.

1. **Adding Missing Information:** Despite the extensive available through Spotify, some tracks/artists lacked genre information. To address this, a manual process was implemented to research and assign genres to these entries.

## 3.4 Exploratory Analysis

The exploratory data analysis aimed to uncover underlying patterns and trends in the genre distribution of the Billboard Hot 100 songs from 1985-2023. Various graphical and statistical techniques were employed in order to visualize and summarize the data. This provided initial insights into the evolution of musical genres.

**Summary Statistics**

Initially, summary statistics were calculated to provide a broad overview of the dataset attributes. This step helped to identify any outliers or anomalies that may affect future analyses and also served to provide a quick snapshot of the data.

**Genre Distribution Visualization**

To better understand the distribution of music genres over the dataset, a bar graph was created to show the total instances of each main genre. This visualization helped highlight the most dominant genres within the Billboard Hot 100 over the years.

|  |
| --- |
| Figure 3.4.1: Genre Distribution Visualization |

**Trend Analysis Over Time**

An area plot was used to visualize the percentage of each main genre over time, offering a clear view of how musical preferences have shifted. This initial analysis was crucial for getting a rough look at the dynamic nature of music trends present in the data.

|  |
| --- |
| Figure 3.4.2: Trend Analysis Over Time |

This exploratory analysis provided foundation insights into the data, setting the stage for more detailed analyses.

# 4. Results

## 4.1 Univariate Analysis

**Methodological Overview**

To analyze the trends and distribution of different music genres over time, binary flags were leveraged for each main genre, converting categorical genre data into a format suitable for quantitative analysis. This transformation enabled the structured tracking of genre prevalence within the dataset, spanning several decades. The tidyverse package was used to perform calculations to quantify the yearly count of songs per genre, facilitating a detailed exploration of temporal shifts in musical preferences.

**Results and Observations**

The evolution of music genres was visualized using area plots, which illustrate the total count of songs for each genre across years. These plots reveal historical popularity of genres but also highlight trends and deviations over time. The following insights were drawn from the visual and statistical analysis:

* **Pop** has shown consistent prevalence, with spikes indicating eras of heightened popularity.
* **Rock** experienced significant popularity in the late 20th century, with a gradual decline in recent years.
* **Country** music shows periodic fluctuations, without a clear long-term trend.
* **Jazz and Blues** have diminished in representation, suggesting a shift in popular taste away from these genres.
* **Soul and Funk** peaked around the 1970s and have since experienced a decline.
* **Hip-Hop and Rap** have seen a notable rise, reflecting their growing influence in mainstream music.
* **Electronic and Dance** genres exhibit an increasing trend, likely due to the rise of digital production methods.
* **R&B and Reggae** have maintained a steady presence with slight fluctuations over time.
* **Classical and Folk** music had a period of popularity in the mid-20th century but have since seen a decline.
* **World and Regional** music shows intermittent peaks, which may be associated with world events and global trends.
* **Indie and Alternative** genres appear to have gained a foothold in more recent years, indicating a shift towards more diverse and non-mainstream music.

|  |
| --- |
| Figure 4.1.1: Genre Distribution Over Time |

**Visual Analysis**

The area plots for each genre were enhanced by smoothing lines to better visualize trends, using the geom\_smooth function with a LOESS method. This approach provided a clearer picture of the overall direction of trends over time.

**Comparison with Existing Research**

The analysis of genre trends over time in this study aligns with the findings from other studies, notably the analysis presented by Pettibon and Hitchcock [7], which examined observed proportions and model-predicted probabilities for various music genres from 1975 to 2018 in Figure 2 of their study. Similar to the graph generated here, their findings indicate shifts in genre popularity over the decades which they represented through observed proportions and model-predicted probabilities. These studies support the idea that genre popularity shows clear trends which evolve over time.

**Analysis of Audio Features**

To explore how audio features have evolved in popular music, a General Additive Model (GAM) was employed to model each feature as a function of time and genre. This approach allowed for the capture of non-linear trends and the influence of genre on the auditory properties of music.

The analysis was conducted using the R package mgcv which supports the fitting of GAMs with smooth splines. Each audio feature was modeled as a function of time and musical genre. Cross validation was used to optimize the model fit, and the model’s performance for each feature was evaluated based on the RMSE. Trends were visualized using smoothed lines with confidence intervals to highlight the uncertainty around the estimates.

The results are as follows:

* **Danceability** shows an increasing trend, suggesting a preference for more rhythmically driven tracks in recent decades.
* **Energy** follows a similar upward trend, indicating that songs with higher intensity and activity have become more prevalent.
* **Loudness** has also increased notably, which could reflect both a change in musical style and advancements in recording technology.
* **Speechiness** has seen fluctuating prevalence, potentially mirroring the rise and fall of genres like rap and spoken word.
* **Acousticness** has declined overall, possibly due to the electronic and digital production shift.
* **Instrumentalness** demonstrates a decreasing trend, suggesting that vocal tracks have gained dominance.
* **Liveness** has remained relatively stable, with slight fluctuations that do not indicate a clear trend.
* **Valence** shows variation over time with no long term trend, which reflects the complex nature of mood in music.
* **Tempo** has experienced slight increases, indicating a slight trend towards faster-paced songs.

|  |
| --- |
| Figure 4.1.2: GAMs for Audio Features Over Time |

## 4.2 Multivariate Analysis

The objective of the multivariate analysis was to explore the relationships between various audio features and time, utilizing these audio features as predictors for the year of release.

**Modeling and Prediction**

Models were developed to predict release year based on audio features (instrumentalness, energy, tempo, danceability, speechiness, valence, liveness, acousticness, and loudness) from the Hot\_100\_Processed dataset. Data was split into 80% training and 20% testing sets.

Several models were evaluated and compared:

1. **Random Forest (RF):**

* The RF model was chosen because it is robust, handles complex datasets well, and provides insight into feature importance which was of particular interest given the topic of this project. The RF model was developed using the caret package. The model underwent hyper-parameter tuning performed via grid search to optimize parameters.

1. **Decision Tree:**

* The Decision Tree model was chosen because it is simple, easily interpretable, and is good at selecting relevant features which was important given the amount of audio features. The Decision Tree model was developed using the rpart and caret libraries. The model underwent 10-fold cross-validation and a tuning grid was utilized to optimize the complexity parameter.

1. **Gradient Boosting Machine (GBM):**

* The GBM model was chosen because it is flexible, handles various types of data well, and tends to be more accurate than other algorithms. The GBM model was developed using the gbm and caret libraries. The model underwent 10-fold cross validation and a tuning grid was utilized to optimize parameters such as number of trees and interaction depth.

**Comparison and Insights**

The performance of each model was quantitatively assessed using the following metrics:

|  |
| --- |
| Table 4.2.1: Accuracy, Kappa, RMSE, & R-squared Values for ML Models |

The Decision Tree model demonstrated the highest accuracy and Kappa, indicating better classification capabilities but with a trade-off in model complexity, which was shown in its lower R-square and higher RMSE values.

The GBM model performed better than the Random Forest in terms of Kappa and accuracy, and slightly better than the Decision Tree model in terms of R-squared. While it performed similarly to the Random Forest model, it showed less efficiency in error reduction which is reflected in the RMSE value.

|  |
| --- |
| Figure 4.2.1: Visual Metric Comparison of ML Models |

The RF model, despite having a lower accuracy and Kappa than the Decision Tree, showed the best performance in terms of RMSE and R-squared. These results indicated that the Random Forest model had a higher capacity to explain the variance in the data, so it was chosen as the final model for analysis.

**Random Forest Analysis and Implications**

The scatter-plot in figure 4.2.2 indicates that the RF model performs adequately in predicting the release year of songs, albeit with some variability. While the model’s primary purpose in this project was to assess the importance of different audio features, it is worth noting the aspects of it’s predictive performance.

|  |
| --- |
| Figure 4.2.2: RF Observed vs Predicted Values |

The model appears to predict older songs with greater success than newer ones. This trend might suggest that audio features from earlier decades are more distinct or consistent. In contrast the model’s prediction for newer songs shows greater deviation from the line of prediction. This could be due to a variety of factors, but could indicate a greater diversity in modern music production or trends that are not as well captured by the model.

Despite these limitations, the predictive accuracy of the model surpasses random chance. While this indicates room for improvement, also shows that the model can provide valuable insights, especially in regard to feature importance. The RF model provides a tool for evaluating the importance of various features. By leveraging this capability, insight was able to be gained on the specific contribution of different audio features to the predictive accuracy of the model.

|  |
| --- |
| Figure 4.2.3: Importance of Audio Features for RF Model |

Figure 4.2.3 illustrates feature importance as determined by a RF model, specifically showing the mean decrease in Gini Impurity for various audio features. Each point represents a different audio feature from the dataset. The vertical axis lists the features while the horizontal quantifies their importance. A higher value on the horizontal axis indicates a greater impact on the model’s decision making process, highlighting which characteristics are most influential.

The most influential features were loudness, acousticness, and liveness, while the least influential feature was instrumentalness.

# 5. Discussion

This study demonstrates that there are temporal trends present for popular music genres and their audio features.

## 5.1 Summary and Interpretation

This study set out to chart the temporal trends of popular music genres and the evolution of audio features in tracks on the Billboard Hot 100 charts form 1985-2023. Temporal were clearly present for both genre and audio features, and those trends were able to be visualized using the tools available in R Studio. The Random Forest model’s capacity to accurately predict the year of the song release based on audio features implies that certain audio features play a key role in a song’s popularity.

## 5.2 Strengths and Limitations

One of the key strengths of this study is the comprehensive dataset which provided plenty of temporal data to be able to observe trends. Additionally the utilization of the Spotify API allowed for that data to be enriched, allowing for even more temporal analysis. Additionally, this study lent itself well to visualization.

However, this study did not lend itself as well to quantifiable results. Additionally the genre classifications, while informed by Spotify’s comprehensive database, could still subject some songs to misclassification due to the subjective nature of music genres.

## 5.3 Conclusions

The insights from this study establish that popular music is not static; there are definite trends present both in genre and in audio features suggesting an ever changing music preference from listeners.

# 6. References

[1] Terroso-Saenz F, Soto J, Muñoz A. Evolution of global music trends: An exploratory and predictive approach based on spotify data. Entertainment Computing [Internet]. 2023;44:100536. Available from: <https://www.sciencedirect.com/science/article/pii/S1875952122000593>.

[2] Mauch M, MacCallum RM, Levy M, et al. The evolution of popular music: USA 1960–2010. Royal Society Open Science [Internet]. 2015;2:150081. Available from: <https://royalsocietypublishing.org/doi/abs/10.1098/rsos.150081>.

[3] Tolsen. Billboard hot 100 [Internet]. Billboard; 2024. Available from: <https://www.billboard.com/charts/hot-100/>.

[4] Molanphy C. How the hot 100 became america’s hit barometer [Internet]. NPR; 2013. Available from: <https://www.npr.org/sections/therecord/2013/08/16/207879695/how-the-hot-100-became-americas-hit-barometer>.

[5] HipsterVizNinja. Hot 100 raw data [Internet]. GitHub. Available from: <https://github.com/HipsterVizNinja/random-data/tree/main/Music/hot-100>.

[6] Web API | spotify for developers [Internet]. Available from: <https://developer.spotify.com/documentation/web-api/>.

[7] Petitbon AM, Hitchcock DB. [What kind of music do you like? A statistical analysis of music genre popularity over time](https://doi.org/10.6339/22-JDS1040). Journal of Data Science. 2022;20:168–187.