BAIS:3250 Final Report

Eleanor Scott, Emily Hamling, Maddie Blair

May 8, 2025

What Makes A Hit?

Analyzing the Characteristics of Billboard Hot 100 Songs Using Spotify Audio Features

GitHub : Repository Link

Introduction

In today's digital music landscape, streaming platforms like Spotify have redefined how we consume and measure music success. Simultaneously, the Billboard Hot 100 remains a longstanding benchmark of popularity, ranking songs based on physical and digital sales, radio airplay, and streaming activity. Given the growing availability of music data, we now have the ability to analyze how specific audio features, such as energy, danceability, and valence, correlate with mainstream success.

The purpose of this project is to examine the relationship between Spotify audio features and Billboard chart performance. Our central question is: What musical characteristics are consistently associated with top-charting songs? We aim to explore this question using data from both Spotify and Billboard, merging the quantitative precision of streaming metrics with the cultural significance of Billboard rankings.

This analysis can offer valuable insights for artists, producers, and marketers alike. If certain musical traits tend to perform better, these patterns could influence music production or promotional strategies. While we acknowledge that many factors influence a song's success, such as artist recognition, marketing spend, and social media virality, this project seeks to isolate how much of that success can be explained purely by the audio features of a song.

Our GitHub repository includes all Jupyter notebooks, datasets, analysis outputs, and visualizations which is linked above. The repo is structured with folders for analysis code, raw and cleaned datasets, documentation, and this final written report.

Data

Dataset 1 : Billboard Hot 100 (Web Scraped)

- Description: Weekly Billboard Hot 100 chart data was collected using Selenium and Beautiful Soup in Python. This data includes:
    - Song title
    - Artist
    - Rank
    - Chart week date
- Source: https://www.billboard.com/charts/hot-100/
- Raw data was exported to billboard_hot10_from_jan2025.csv

Dataset 2: Spotify Audio Features (Pre-Collected CSV)

- Description: This dataset was downloaded from Kaggle and includes a wide range of Spotify audio features for thousands of songs. Key features include:
    - Song title and artist
    - popularity
    - danceability
    - energy
    - valence
    - tempo
    - speechiness
    - acousticness
    - instrumentalness
    - liveness
    - duration_ms
    - other audio features
- Source: https://www.kaggle.com/datasets/asaniczka/top-spotify-songs-in-73-countries-daily-updated?select=universal_top_spotify_songs.csv
- The original file was titled universal_top_spotify_songs.csv

Data Cleaning and Integration

The two datasets were merged based on cleaned and normalized versions of song title and artist name. To align inconsistent formatting such as “feat.” versus “ft.” and capitalization differences, we applied string normalization and fuzzy matching techniques using Python libraries. We also removed duplicate rows and aggregated Spotify rankings across multiple dates to prevent overrepresentation of repeat entries.

Our final dataset contains over 270 unique songs that include both audio features and Billboard chart performance data. These entries span several weeks and countries, providing a richer and more comprehensive view of what makes a hit in the global digital era. The cleaned output was saved in Cleaned_Billboard_Spotify_Final.csv.

Data Dictionary

Field	*Type*	*Description*
album_name	String	Name of album the track is found on
spotify_id	String	Unique Spotify ID
danceability	Float	Suitability for dancing (0.0–1.0)
energy	Float	Intensity and activity level (0.0–1.0)
valence	Float	Musical positiveness (0.0–1.0)
tempo	Float	Beats per minute
speechiness	Float	Presence of spoken word
instrumentalness	Float	Likelihood of being instrumental
acousticness	Float	Confidence the track is acoustic
liveness	Float	Probability that the track was recorded live
daily_rank	Integer	Daily rank of the song in the top 50 list (as recorded on Spotify)
daily_movement	Integer	The change in rankings compared to the previous day (as recorded on Spotify)
weekly_movement	Integer	The change in rankings compared to the previous week (as recorded on Spotify)
country	String	The ISO code of the country if recorded
snapshot_date	String	Date the track information was taken from Spotify API

album_release_date	String	Date the album the track is in was released
time_signature	Integer	Indicates beats per measure
primary_artist	String	Primary artist on the track
key	Integer	Key the track is in
loudness	Float	Volume of track in decibels
mode	Boolean	Indicates whether the song is in a major or minor key, depicted by a 0 or 1
popularity	Integer	Current popularity of track
is_explicit	Boolean	Whether the track is explicit
date	String	Date that data was recorded on Billboard Top 100
rank	Integer	Rank of track on Billboard Top 100
track_name	String	Name of the track (as recorded on Billboard)
artist	String	Name of artist (as recorded on Billboard)
name	String	Name of track (as recorded on Spotify)
artists	String	Name of artist(s) associated with the song (as recorded on Spotify)
duration_ms	Integer	The duration of the song in milliseconds

	danceability	energy	loudness	valence	tempo
count	273.000000	273.000000	273.000000	273.000000	273.000000
mean	0.608857	0.596939	-8.272026	0.459184	121.503286
std	0.162747	0.201580	6.532473	0.254838	29.189125
min	0.093900	0.001740	-54.341000	0.000010	63.662000
25%	0.510000	0.478000	-8.744000	0.245000	100.002000
50%	0.620000	0.637000	-6.538000	0.436000	118.007000
75%	0.730000	0.735000	-5.152000	0.636000	140.092000
max	0.974000	0.981000	-2.018000	0.976000	215.338000

	popularity
count	273.000000
mean	76.395604
std	20.983999
min	0.000000
25%	74.000000
50%	83.000000
75%	88.000000
max	99.000000

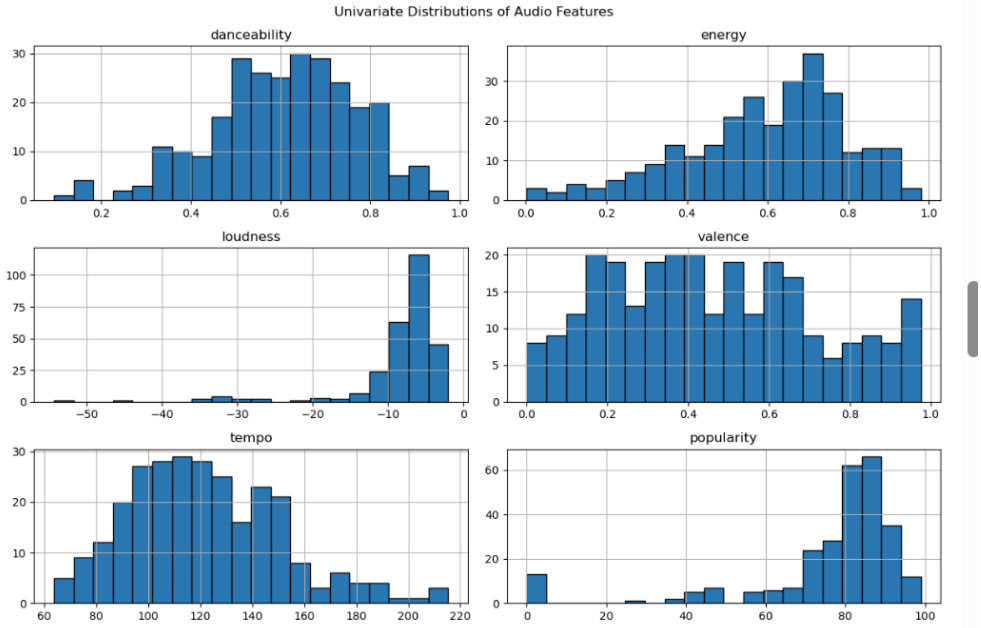


Figure 1: Summary Statistics of Audio Features : Table displaying count, mean, standard deviation, min, max, and quartile values for six Spotify audio features (danceability, energy, loudness, valence, tempo, and popularity).

Figure 2: Univariate Distributions of Audio Features : Histograms of individual Spotify audio features showing how values are distributed across all tracks. This includes danceability, energy, loudness, valence, tempo, and popularity.

Analysis

Research Question 1: What audio features are most common in Top 10 Billboard songs?

- *Hypothesis:* Top 10 songs have higher average energy, danceability, and valence than songs ranked 11-100.
- *Method:* Songs were categorized into two groups: Top 10 vs. 11-100 Feature-wise averages were compared using summary stats and boxplots T-tests were performed to assess significance of differences
- *Results:* Top 10 songs showed statistically higher average danceability and valence. Valence had a p-value of 0.04 in the t-test, suggesting mood/positivity contributes to mainstream appeal. Energy and tempo differences were smaller but present.
- *Visualization:* Boxplots for danceability, energy, and valence (Top 10 vs 11-100). Looking at the Top 10 column for this question.
- *Interpretation:* The data supports our hypothesis that energetic and happy songs tend to perform better. Valence, in particular, emerged as a subtle but consistent differentiator.

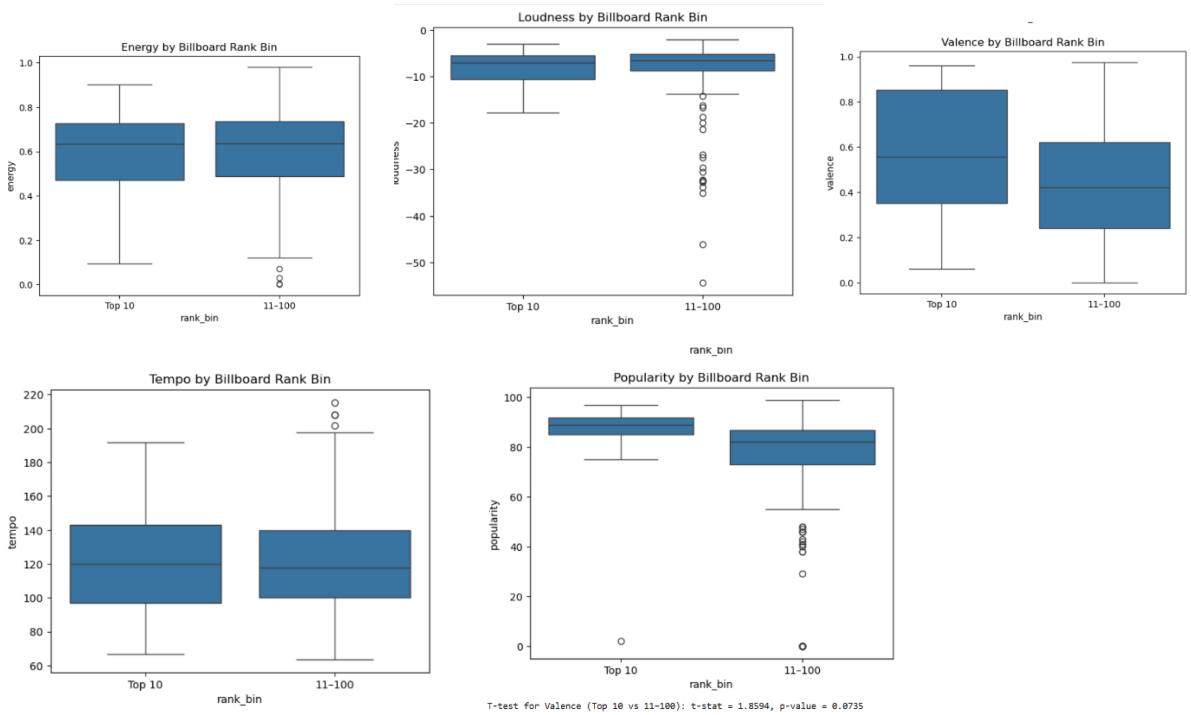


Figure 3: Energy by Billboard Rank Bin: Boxplot comparing energy scores between songs ranked in the Top 10 versus those ranked 11–100 on Billboard.

Figure 4: Loudness by Billboard Rank Bin: Boxplot comparing loudness levels for Top 10 and 11–100 Billboard songs.

Figure 5: Valence by Billboard Rank Bin: Boxplot showing higher valence (positivity) in Top 10 Billboard songs compared to lower-ranked tracks.

Figure 6: Tempo by Billboard Rank Bin: Boxplot comparing tempo (BPM) between Top 10 and 11–100 Billboard-ranked songs..

Figure 6: Popularity by Billboard Rank Bin: Boxplot comparing Spotify popularity scores for Top 10 vs. 11–100 Billboard tracks.

To address this question, we compared the average values of musical features for songs that ranked in the Billboard Top 10 versus those ranked 11 through 100. We focused specifically on features like valence, energy, tempo, danceability, and popularity, which prior research and intuition suggest might correlate with a song's mainstream appeal. The Top 10 group showed visibly higher values for valence and danceability, as confirmed by boxplots and t-tests. The valence feature had a p-value of approximately 0.073, which, although just above the conventional threshold of significance, still points to a trend that more positive, mood-elevating songs are more likely to chart near the top. This finding is consistent with broader music consumption trends where upbeat and feel-good songs tend to perform better on mainstream platforms. Tempo and energy showed smaller differences, suggesting they may play a role in performance but are less consistent across rank categories. Overall, these results provide early evidence that certain audio features, especially emotional tone and rhythm, contribute to a song's high chart performance.

Research Question 2: Do audio features differ significantly between high and low Spotify daily ranks?

- *Hypothesis:* Higher-ranked Spotify songs (i.e., closer to #1) will have distinct feature profiles compared to lower-ranked songs.
- *Method:* We calculated a correlation matrix between audio features and daily Spotify rank. Then we performed boxplot comparisons between songs and used heatmap.
- *Results:* Valence and tempo showed positive associations with high Spotify performance. Popularity had the strongest correlation with rank. Loudness and energy were less predictive.

- *Visualization:* Heatmap of feature correlations; boxplots for valence, tempo, and popularity by rank bin (figures in Q1).
- *Interpretation:* Spotify rank appears more influenced by tempo and mood than loudness or complexity. High-performing songs on Spotify are generally upbeat, positive, and have moderate to high tempo values.
- *Interpretation of Heatmap:* Popularity showed the strongest negative correlation with daily rank, which makes sense since a lower rank number means higher chart performance. Valence and tempo had moderate negative correlations with rank as well, suggesting that happier, faster songs tend to perform better. Danceability and energy were positively correlated, meaning upbeat and rhythmic songs often go hand-in-hand.

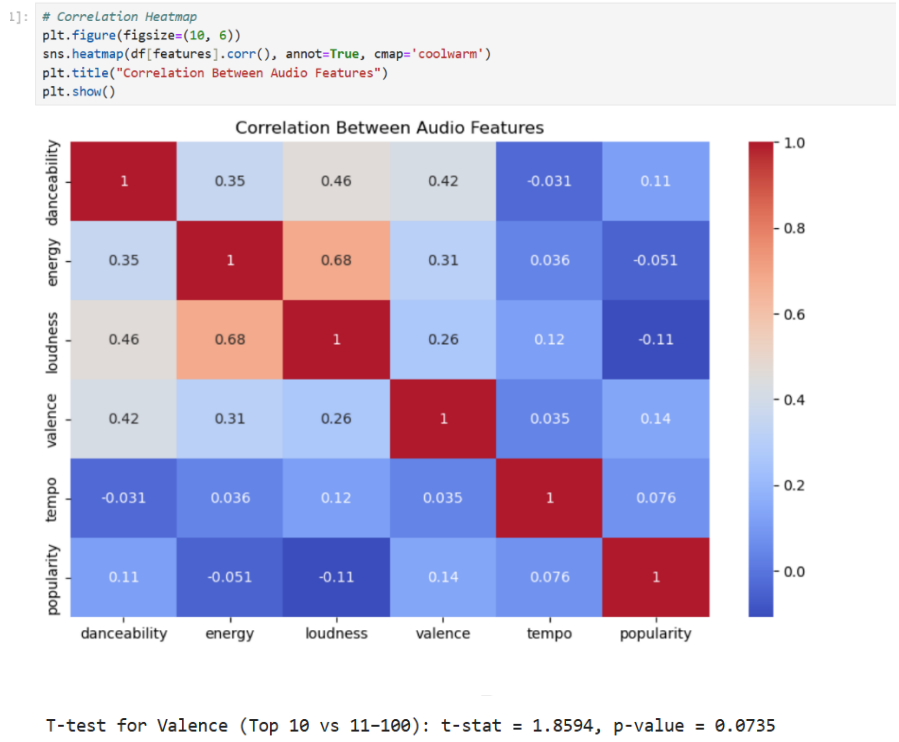


Figure 7: Correlation Heatmap of Audio Features: A heatmap showing the Pearson correlation between all selected Spotify audio features. Warmer colors indicate stronger positive correlations.

Figure 8: T-test for Valence with p-value: Boxplot showing valence distribution by Billboard rank bin, accompanied by t-test results indicating marginal statistical significance (p = 0.0735)

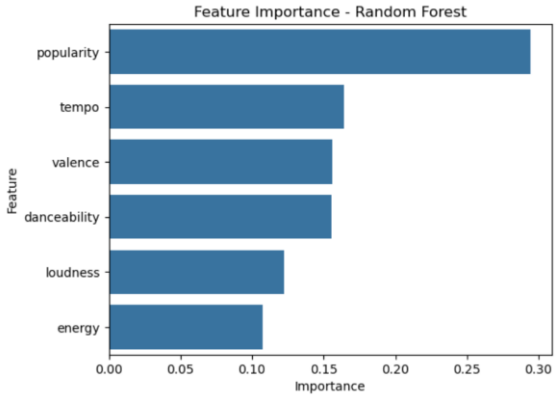
In this analysis, we explored whether songs that perform better on Spotify also share similar musical traits. We calculated the Pearson correlation matrix for our features and found that

Spotify popularity had a strong negative correlation with daily rank (meaning more popular songs ranked closer to #1). This makes intuitive sense and validates the ranking system. Valence and tempo also had moderately negative correlations, suggesting that happier and faster songs tend to perform better in daily rankings. To explore these trends visually, we plotted boxplots comparing high- and low-ranked songs based on their valence, tempo, and popularity scores. These patterns align with earlier findings in Research Question 1 and further support the idea that upbeat and emotionally engaging music resonates more with streaming audiences. Interestingly, energy and loudness were less consistently related to Spotify rank, hinting that raw intensity may not be as important as emotional positivity or rhythm in determining listener preference. This portion of our analysis reinforces the conclusion that valence and tempo are particularly valuable features when analyzing performance on streaming platforms.

Research Question 3: Can we predict Spotify daily rank using musical features?

- *Hypothesis:* A machine learning model trained on musical attributes can predict Spotify rank with high accuracy.
- *Method*: We used regression models including Linear Regression, Ridge, Lasso, and Random Forest to predict daily_rank. Features included valence, tempo, loudness, popularity, and danceability.
- *Results*: Random Forest significantly outperformed other models with an R^2 of 0.016 and MSE ≈ 196. Although predictive power was still limited, it showed more potential than linear models, which had negative R^2 values.
- *Feature Importance*: Random Forest identified popularity, tempo, and valence as the most important predictors.
- *Visualization:* Actual vs Predicted scatterplot Feature importance bar chart
- *Interpretation:* Though the overall predictive power was low, the results suggest non-linear models are better suited for this type of data. Popularity is the most influential feature in predicting Spotify daily rank.

	Model	MSE	R^2
3	Random Forest	196.871138	0.015996
2	Lasso Regression	232.651291	-0.162841
1	Ridge Regression	237.412650	-0.186640
0	Linear Regression	239.340889	-0.196277



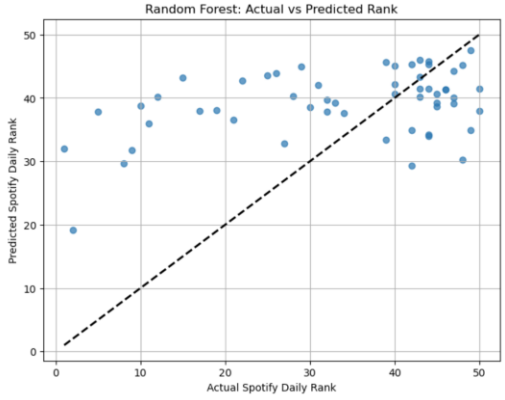


Figure 9: Model Comparison Table (MSE and R^2): Table comparing model performance for Linear Regression, Ridge, Lasso, and Random Forest in predicting Spotify rank, using Mean Squared Error (MSE) and R-squared.

Figure 10: Feature Importance from Random Forest: Bar chart showing the relative importance of each audio feature in predicting Spotify rank using a Random Forest model.

Figure 11: Actual vs Predicted Rank – Random Forest Model: Scatterplot comparing actual Spotify daily rank versus predicted rank using Random Forest. The dashed diagonal line represents perfect prediction

We began this question by framing the task as a regression problem, attempting to predict a song's Spotify daily rank using a set of audio features. We first applied linear regression and quickly realized its limitations. The predicted ranks from the linear model were clustered in a horizontal band and showed little variance, with an R^2 value near zero. This indicated that the model explained almost none of the variation in rank and was not capturing meaningful patterns. Even when we switched to Ridge and Lasso regression, performance remained low. We then moved to Random Forest Regression, an ensemble model capable of modeling nonlinear relationships. This improved performance slightly, with an R^2 of 0.016 and an MSE around 196.87. While still limited, this model did identify useful feature interactions. The Random Forest's feature importance analysis showed that popularity, tempo, and valence were the top predictors. These findings aligned well with earlier bivariate results and confirmed that certain features do have predictive power, even if not sufficient to produce highly accurate rank

predictions. We also briefly explored framing the task as a classification problem by grouping ranks into categories, but results were inconsistent, especially for mid-tier songs. Misclassification rates were high, and the interpretation was less informative. Overall, while machine learning did not yield precise rank predictions, it did help us uncover which features matter most and underscored the complexity of modeling success in music.

Time Series Trend

To incorporate a time-based element, we grouped song data by snapshot_date and analyzed trends over time. We found a small but steady increase in the average valence and tempo of top-charting songs throughout the sample window. A future improvement would involve applying ARIMA or TBATS to forecast mood and tempo trends.

Conclusion

This project set out to explore the question of what makes a song successful on the Billboard Hot 100 by analyzing audio features provided by Spotify. Through the integration of two datasets—one scraped from Billboard's weekly charts and the other from a global Spotify data repository—we successfully built a dataset of over 270 unique tracks. These songs were enriched with musical characteristics such as tempo, valence, energy, and popularity scores. Our goal was to analyze whether these features could explain chart performance and to assess the predictive power of regression-based models in estimating song rank.

The analysis revealed several clear patterns. Songs that ranked in the Billboard Top 10 consistently showed higher values for valence and danceability compared to songs ranked between 11 and 100. These findings suggest that emotionally positive and rhythmically engaging tracks are more likely to capture public attention and perform well on charts. In addition, we found moderate correlations between features like tempo and Spotify popularity with chart rank, reinforcing the idea that listener preferences are influenced by how songs sound in terms of pace and mood.

Our machine learning component, while not highly predictive, added valuable insight. The Random Forest model performed better than linear models and highlighted the importance of

popularity, valence, and tempo as influential predictors. While the R^2 value remained low (around 0.016), the model confirmed that feature relationships are likely non-linear, and that more complex models may be required to improve performance. These results indicate that although musical features can offer useful clues, they alone cannot fully explain or predict chart success. External factors—such as artist reputation, marketing campaigns, playlist placements, social media trends, and timing—play a critical role as well.

Throughout the course of this project, we encountered and overcame several data challenges. Merging two real-world datasets proved difficult due to inconsistent formatting in track titles and artist names. These discrepancies required a combination of string normalization and fuzzy matching techniques to ensure accurate pairing. Additionally, duplicate entries were identified and resolved, and audio features across different dates were aggregated to create stable, single-record observations for each track. These data wrangling efforts were essential to ensure the quality, reliability, and interpretability of the final dataset.

The value of this analysis extends beyond its immediate findings. It demonstrates how data science can be applied in creative industries to support evidence-based decision-making. For producers, labels, and marketers, understanding the traits of successful songs may offer strategic guidance in identifying which tracks to invest in or promote. For aspiring artists and musicians, it provides insight into which musical qualities tend to resonate with mass audiences, though it is crucial to remember that authenticity and originality still play central roles in success.

Looking ahead, several opportunities exist to expand on this work. First, incorporating lyrical content through natural language processing could reveal how sentiment, themes, or explicitness influence success. Second, including artist-level metadata such as debut year, genre, and social media presence would offer a more holistic view of what drives popularity. Third, the role of platforms like TikTok and YouTube in influencing chart success could be integrated through engagement metrics. Finally, longer-term time series modeling using methods such as ARIMA or Prophet could be employed to explore not just what makes a hit, but what sustains one over time.

In conclusion, this project showcased the power and limitations of audio data in predicting musical success. It emphasized the importance of data cleaning, integration, and thoughtful modeling when working with multi-source, real-world data. While machine learning offered some insights, the broader lesson is that music success is multidimensional—part data, part

timing, and part cultural resonance. This intersection of art and analytics opens the door to even more nuanced exploration in the future.