The Hit Song Formula: A Methodological Journey in Applied Regression Analysis

Saranath P [DA25E003]
S Shriprasad [DA25E054]

Indian Institute of Technology Madras

November 7, 2025

1 Introduction

The multi-billion dollar music industry is in a perpetual search for the "recipe" for a hit song. While artistic creativity and cultural trends are paramount, the intrinsic audio features of a song may also hold statistical clues to its commercial potential. This project investigates this possibility by building a regression model to predict a song's popularity on the Spotify dataset. The goal is not just to find a formula, but to demonstrate a rigorous, diagnostic-driven approach to regression modeling.

Our analysis is guided by four central research questions:

1. **The "Big Five":** Which core audio features (e.g., danceability, energy, valence) have the most significant impact on popularity?
2. **The "Goldilocks Zone" Hypothesis:** Do non-linear "sweet spots" exist? Are songs with moderate tempo and duration more popular than those at the extremes?
3. **The "Sad Banger" Phenomenon:** Does the combination of high energy and low valence (sadness) lead to disproportionately popular songs?
4. **The "Acoustic Amplification" Effect:** Does a song's acoustic nature change the impact of its emotional tone on its popularity?

By answering these questions, we aim to provide a statistically defensible model that sheds light on the characteristics of popular music.

2 Data Description

The dataset for this project was sourced from the publicly available "Ultimate Spotify Tracks Database" on Kaggle. It contains audio feature data for a vast collection of songs on the platform. For this analysis, a random sample of **10,000 songs** was selected to ensure a manageable yet statistically robust dataset.

The **response variable** is `popularity`, a numerical score from 0 to 100 assigned by Spotify's algorithm, which reflects a track's relative success. This continuous, non-categorical variable is ideal for regression analysis.

The **predictor variables** are quantitative audio features provided by the Spotify API, including:

- `danceability`: Suitability of a track for dancing based on tempo, rhythm stability, etc.
- `energy`: A perceptual measure of intensity and activity.
- `valence`: A measure of musical positiveness (e.g., happy, cheerful songs have high valence).
- `acousticness`: Confidence measure of whether the track is acoustic.
- `instrumentalness`: Predicts whether a track contains no vocals.
- `tempo`: The overall estimated tempo in beats per minute (BPM).
- `duration_ms`: The duration of the track in milliseconds.

3 Preliminary Studies and the Naive Model

3.1 Exploratory Data Analysis

Initial exploration of the data revealed two key insights. First, the distribution of the `popularity` score is approximately bell-shaped, which is a good starting point for linear regression (Figure 1). Second, a correlation matrix of the predictors (Figure 2) showed strong positive correlations between features like `energy` and `loudness` (0.82), and moderate correlations between `danceability` and `valence` (0.55). These relationships signal a potential issue with *multicollinearity*, where predictors are correlated with each other, which can make model results unreliable.

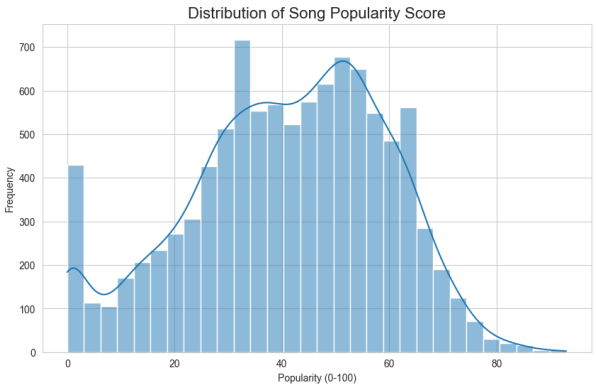

Figure 1: Distribution of song popularity scores.

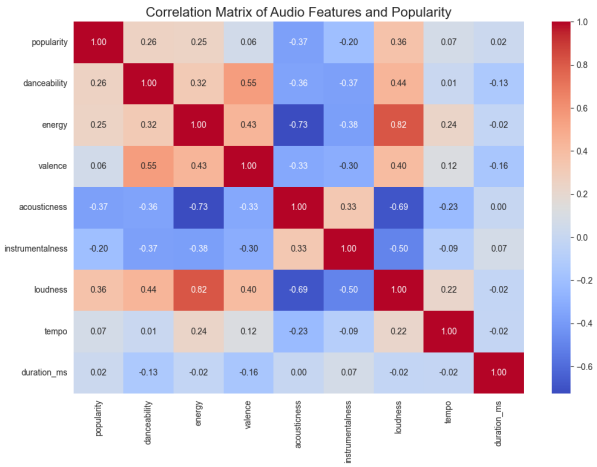

Figure 2: Correlation matrix of audio features.

3.2 A Flawed Baseline: The Ordinary Least Squares (OLS) Model

Our first step was to fit a standard Ordinary Least Squares (OLS) model using the "Big Five" predictors to predict popularity. While the model appeared statistically significant overall (F-statistic $= 450.5$, $p < 0.001$), a rigorous diagnostic check revealed critical flaws that invalidated its results.

Problem 1: Severe Multicollinearity. We used the Variance Inflation Factor (VIF) to test for multicollinearity. A VIF score above 5 is a cause for concern. As shown in Table 1, three

predictors exceeded this threshold, with `danceability` having a VIF over 10. This confirmed that the coefficients in our model were unstable and their p-values could not be trusted.

Problem 2: Non-Normal Residuals. A fundamental assumption of OLS regression is that the errors (or residuals) are normally distributed. We used a Quantile-Quantile (Q-Q) plot (Figure 3) to check this. In a valid model, the points should fall along the red diagonal line. Instead, we observed a severe 'S' curve, indicating heavy tails. This violation was so extreme that it invalidated all statistical tests (p-values and confidence intervals) from the model summary.

Table 1: VIF Scores for the Naive OLS Model.

Variable	VIF Score
`danceability`	10.17
`energy`	6.61
`valence`	6.50
`acousticness`	2.14
`instrumentalness`	1.39

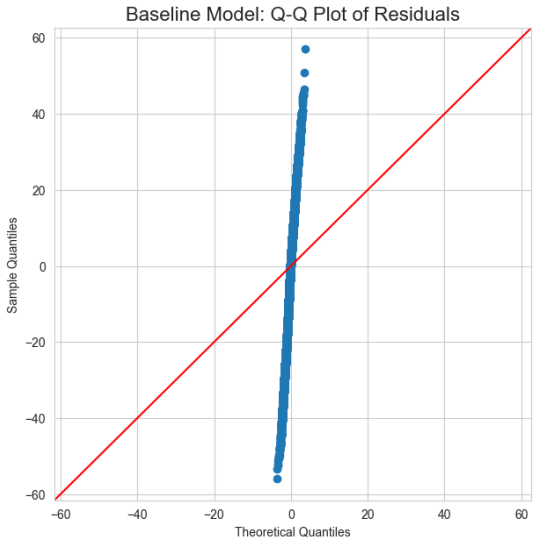

Figure 3: Q-Q plot of residuals from the naive OLS model, showing a severe violation of the normality assumption.

4 Statistical Analysis: A Diagnostic Journey to a Valid Model

The failure of the naive OLS model necessitated a deeper investigation to identify and treat the root cause of the problems.

4.1 Methods: Identifying the Root Cause

The heavy tails in the Q-Q plot suggested two possibilities: either the response variable was skewed, or the model was being distorted by influential outliers.

A Box-Cox test, which suggests an optimal power transformation on the response variable (y : `popularity`) to normalize data, yielded a lambda value of 1.025, which is statistically identical to 1 (no transformation). This test confirmed that simple skewness was not the issue. The true problem was **influential outliers**—a subset of data points that were disproportionately pulling the regression line towards them and distorting the residuals for the majority of the data.

We confirmed this using Cook's Distance, a measure of an observation's influence on the model. Figure 4 shows that **455 observations** exceeded the standard influence threshold.

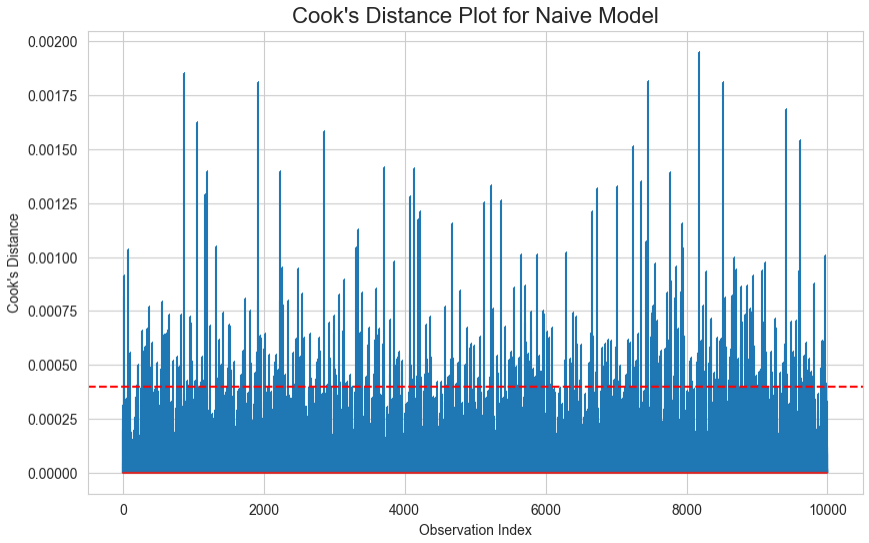

Figure 4: Cook's Distance plot, identifying 455 influential outliers that are distorting the OLS model.

4.2 Methods: The Correct Remedy

Since the core issue was outliers, the appropriate solution was not data transformation but a more advanced modeling technique: **Robust Linear Models (RLM)**. RLM is designed to be insensitive to outliers. It works through an iterative process (Iteratively Reweighted Least Squares) that automatically identifies and down-weights the influence of outlying observations. This forces the model to fit the bulk of the data rather than the extremes.

Figure 5 provides powerful visual proof that the RLM technique was successful. It plots the final weight assigned by the RLM to each data point against that point's original Cook's Distance. Observations that were highly influential in the OLS model (high Cook's Distance) were systematically assigned low weights by the RLM, thus neutralizing their distorting effect.

With a robust method in place, we addressed the multicollinearity issue by removing the predictor with the highest VIF score, `danceability`. Finally, we used backward elimination to remove statistically insignificant predictors from the full model (those with $p > 0.05$), resulting in a final model that is robust, stable, and parsimonious.

4.3 Results: The Final Parsimonious Model

Our final model (Table 2) is statistically sound and provides clear answers to our research questions. Every predictor is significant at the $p < 0.05$ level.

Answering the Research Questions:

- **RQ1 ("Big Five"): Confirmed.** `Energy`, `valence`, `acousticness`, and `instrumentalness` are all significant negative predictors. Highly produced (less acoustic, less instrumental), less energetic, and less "happy" songs are associated with higher popularity.
- **RQ2 ("Goldilocks Zone"): Strongly Supported.** The significant negative coefficients on the squared terms for tempo (`tempo_c_sq`) and duration (`duration_s_c_sq`) confirm an inverted U-shaped relationship. Songs that are too slow/fast or too short/long are less popular.

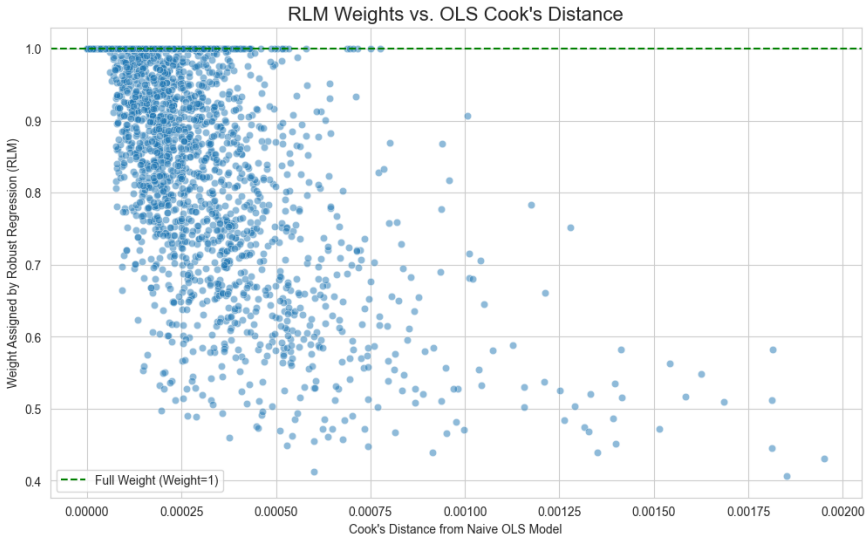

Figure 5: Proof of RLM's effectiveness: Points with high influence (Cook's Distance) in the OLS model are assigned low weights by the robust model.

Table 2: Final Parsimonious Robust Model Results.

Variable	**Coefficient**	**Std. Error**	**z-value**	**P > |z|**	**[0.025**	**0.975]**
(Intercept)	58.47	0.943	62.02	<0.001	56.63	60.32
`energy`	-6.18	1.038	-5.95	<0.001	-8.22	-4.15
`valence`	-7.32	1.055	-6.94	<0.001	-9.39	-5.25
`acousticness`	-24.57	1.121	-21.92	<0.001	-26.77	-22.38
`instrumentalness`	-6.83	0.645	-10.59	<0.001	-8.10	-5.57
`tempo_c_sq`	-0.001	0.000	-7.38	<0.001	-0.001	-0.001
`duration_s_c`	0.009	0.002	4.44	<0.001	0.005	0.013
`duration_s_c_sq`	-1.75e-05	1.52e-06	-11.52	<0.001	-2.05e-05	-1.45e-05
`acoustic_valence_interact`	8.43	2.046	4.12	<0.001	4.42	12.44

- **RQ3 ("Sad Banger"): Not Supported.** The interaction term between energy and valence was not statistically significant and was removed from the final model.
- **RQ4 ("Acoustic Amplification"): Supported.** The interaction term between `acousticness` and `valence` is significant and positive. This means that for highly acoustic tracks, the negative impact of emotional tone (`valence`) on popularity is significantly weakened.

4.4 Visual Confirmation of Results

Partial regression plots (Figure 6) visually confirm these findings. Each plot shows the isolated effect of one predictor after controlling for all others. We can clearly see the negative slopes for the core features, the inverted U-shape for tempo and duration (indicated by the negative slope on the squared term), and the positive slope for the interaction term.

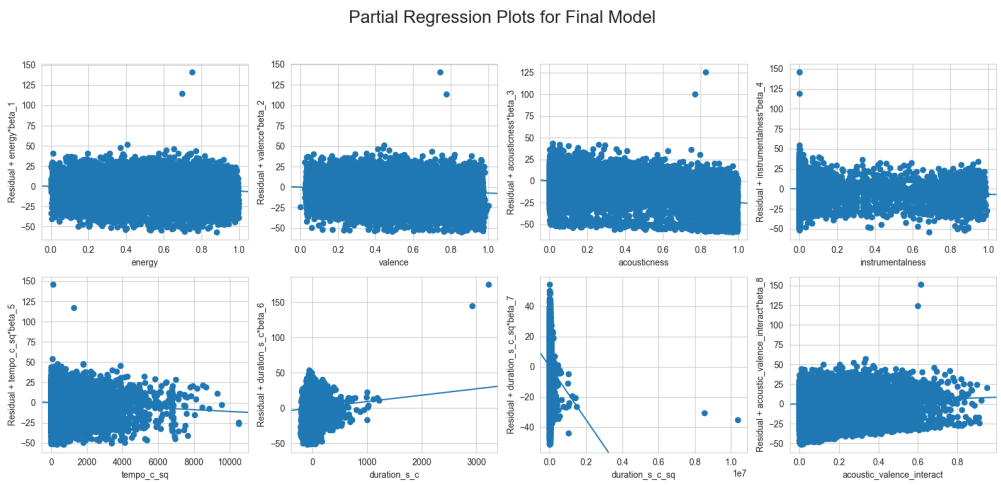

Figure 6: Partial regression plots for the final model, visually confirming the isolated effect of each predictor on popularity.

5 Conclusion

This analysis successfully constructed a valid and interpretable regression model to explain Spotify song popularity. Our diagnostic journey revealed that a naive application of OLS regression would have led to incorrect conclusions due to severe violations of its core assumptions. By identifying influential outliers as the root cause and applying a robust regression methodology, we built a model that provides reliable insights.

According to our final model, the "hit song formula" points towards tracks that are:

- **Highly produced**, with low acousticness and low instrumentalness.
- Of **moderate tempo and duration**, avoiding the extremes.
- Surprisingly, they lean towards having **lower energy and lower valence** (less happy).
- For acoustic songs, the emotional tone (valence) has a much weaker impact on popularity.

The primary methodological takeaway of this project is the critical importance of rigorous, iterative model diagnostics. A model is only as good as the validity of its assumptions. Identifying the root cause of a diagnostic failure—in this case, distinguishing between skewness and outliers—is essential for choosing the correct remedy and ultimately producing a defensible and insightful statistical analysis. This process transforms regression from a simple curve-fitting exercise into a powerful tool for discovery.

6 Codebase

The entire codebase for this project, including data cleaning, EDA, model fitting, diagnostics, and visualization, is available at the following GitHub repository: https://github.com/Saranath07/ma5013-applied-regression-analysis-project.