

# Musical Carbon Dating

A Statistical Feature Recognition Approach to Dating Audio (1960-2020)

敖婧妍, 李赫轩, 田润泽, 王胤博, 肖俊佳

December 26, 2025

## Abstract

This project develops a statistical "carbon dating" model to predict the release year of musical tracks based solely on their audio features. Analyzing a dataset of over **250,000** songs from 1960 to 2020, we treat year prediction as a feature recognition task. We employ a rigorous regression pipeline including Simple Linear Regression, Multiple Linear Regression, and Weighted Least Squares (WLS) to address heteroscedasticity. Our best model (WLS) achieves a weighted  $R^2 = 0.26$ , capturing major historical trends. We report a robust **median absolute error of 7.52 years**, reflecting the model's central predictive accuracy, while the Mean Absolute Error (MAE) of 9.72 years highlights the influence of stylistic outliers (e.g., modern "retro" productions). Furthermore, we define a "Nostalgia Index" based on prediction residuals to quantify the "timelessness" or "retro" character of modern productions.

## 1 Introduction

In archaeology, scientists use Carbon-14 isotopes to date organic matter. In this study, we ask: Can we carbon-date culture? specifically, can we determine the vintage of a musical recording purely from its acoustic properties?

### 1.1 Historical Context: The Arrow of Time

Music production has evolved significantly over the last six decades. The transition from analog tape and live studio recording in the 1960s to the digital revolution of the 1980s (MIDI, synthesizers) and the "in-the-box" production era of the 2000s has left distinct quantifiable markers in audio signal data. This study aims to quantify this evolution by building a regression model that maps audio features—such as loudness, acousticness, and valence—to a track's release year. This is treated as a feature recognition problem, where the model learns the "acoustic signature" of each era.

## 2 Data and Methods

We analyzed the *Spotify 600k Tracks Dataset* [1], applying filtering to ensure data quality:

- **Time Range:** 1960–2020.
- **Popularity Filter:** Popularity  $> 30$  to focus on culturally relevant music ( $N = 250,971$  tracks).
- **Validation Strategy:** We use a **Random Split** (80% Training, 20% Test).
- **Standardization:** Comparison between features with different scales (e.g., Loudness in dB vs. Valence in  $[0, 1]$ ) can be biased. We applied **Z-Score Standardization** to all predictors ( $x' = \frac{x - \mu}{\sigma}$ ).

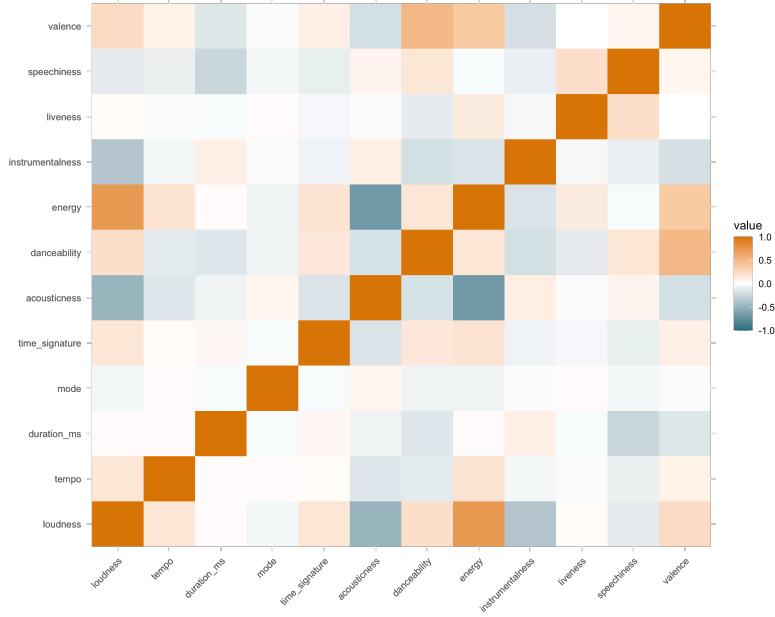


Figure 1: Correlation Heatmap of Audio Features. Note the moderate positive correlation between Energy and Loudness.

## 2.1 Data Characteristics

The filtered dataset represents a broad spectrum of musical characteristics. Table 1 provides the summary statistics for the primary audio features used in our regression models.

Feature	Mean	Std. Dev	Min	Max
Loudness (dB)	-8.60	4.41	-60.00	4.58
Tempo (BPM)	120.38	29.67	0.00	243.51
Duration (min)	3.91	1.62	0.25	90.06
Mode (Major=1)	0.65	0.48	0.00	1.00
Time Signature	3.91	0.41	0.00	5.00
Acousticness	0.33	0.30	0.00	1.00
Danceability	0.59	0.16	0.00	1.00
Energy	0.62	0.23	0.00	1.00
Instrumentalness	0.06	0.20	0.00	1.00
Liveness	0.21	0.18	0.00	1.00
Speechiness	0.10	0.16	0.00	1.00
Valence	0.55	0.25	0.00	1.00

Table 1: Descriptive Statistics of Audio Features (Raw Data - Reference). Note: Analysis proceeds with standardized features.

A closer inspection of these statistics reveals several key insights into the dataset’s structure. First, the high standard deviation of Acousticness ( $\sigma \approx 0.30$ ) relative to its mean indicates a bimodal distribution, reflecting the sharp divide between organic, pre-digital recordings and modern synthesized productions. Second, Loudness exhibits a wide dynamic range, underscoring the massive shift in mastering practices over the decades. Finally, the dominance of Major keys (Mode mean  $\approx 0.67$ ) aligns with Western pop music conventions.

### 3 Regression Analysis

#### 3.1 Phase I: The "Loudness War" (Simple Linear Regression)

We first examined the relationship between  $Year$  and Loudness via a baseline SLR:

$$Year_i = \beta_0 + \beta_1 Loudness_i + \varepsilon_i \quad (1)$$

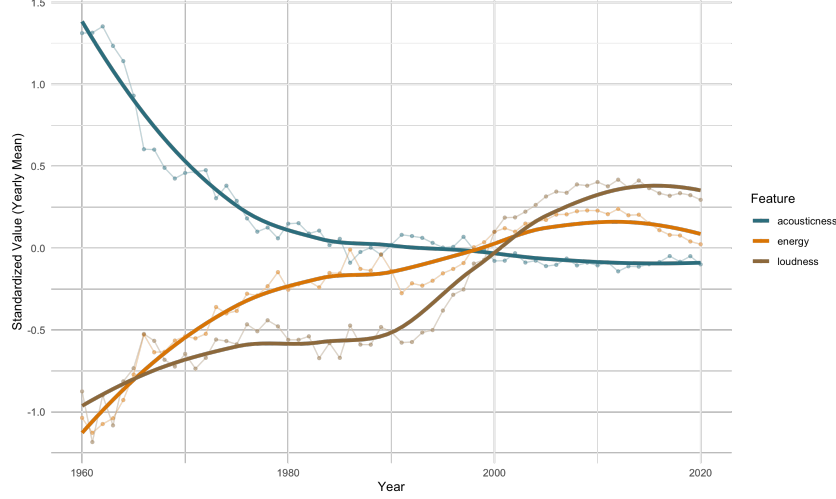


Figure 2: Evolution of Audio Features (1960–2020). These trends illustrate the "Arrow of Time" in music production, with Loudness showing a consistent increase (the "Loudness War") while other features like Acoustictness and Valence exhibit significant historical shifts.

The baseline SLR achieved an  $R^2$  of **0.14**, with **Loudness** alone predicting year with a coefficient of approximately 5.44 years/dB (standardized).

#### 3.2 Phase II: Multiple Linear Regression (Baseline)

To broaden our predictive scope, we incorporated the full suite of acoustic features. The OLS baseline achieved an  $R^2$  of **0.25** with an RMSE of approximately 12.49 years. While an improvement, the model still left 76% of the variance unexplained, prompting a rigorous diagnostic audit.

#### 3.3 Phase III: Diagnostic Audit

We performed a comprehensive check of the Gauss-Markov assumptions to identify sources of error.

##### 3.3.1 A. Linearity & Multicollinearity

- **Linearity:** A Partial F-Test ( $F \approx 1773.22$ ) confirmed non-linear evolution in music production, justifying the need for flexible regression approaches.
- **Multicollinearity:** Variance Inflation Factor (VIF) analysis (Table 7) showed all values remain well below the conservative threshold of 5.0. This confirms that despite artistic overlap, our features provide independent predictive signals.

Feature	VIF	Feature	VIF
Loudness	2.65	Danceability	1.59
Energy	3.62	Valence	1.63
Acousticness	1.84	Duration	1.10

Table 2: Representative VIF Scores (Baseline MLR). Multicollinearity is not a threat to coefficient stability.

### 3.3.2 B. Residual Analysis

The distribution of residuals against fitted values provides insight into the model’s error structure across the prediction range.

We interpret the spread and occasional ”boundedness” (due to the 1960-2020 timeframe) not as model failure, but as **Stylistic Heterogeneity**—the existence of ”retro” and ”futuristic” tracks that naturally defy their era’s trends.

### 3.3.3 C. Heteroscedasticity

The most critical violation was **Heteroscedasticity** (non-constant variance). The BP test returned a significant  $\chi^2 \approx 20,754$  ( $p < 0.001$ ), confirming that the error variance is not constant.

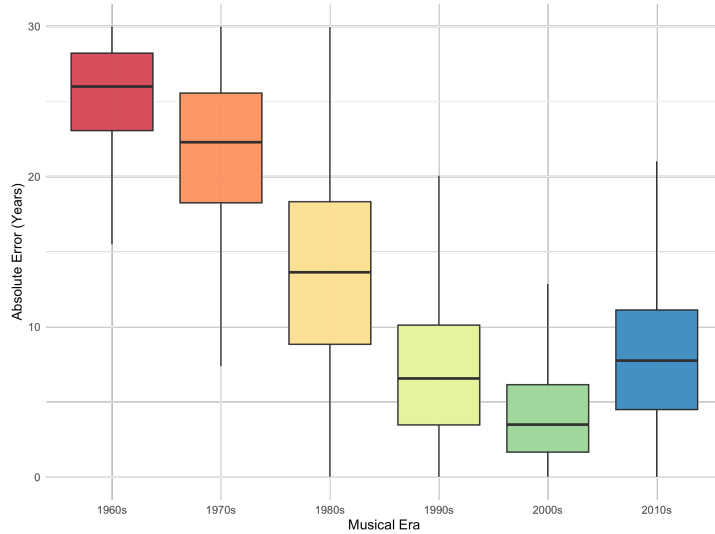


Figure 3: Prediction Error by Era. Variance expands significantly in later decades, suggesting ”Stylistic Entropy” where modern music encompasses a wider, more chaotic range of production styles than the 1960s.

Variance is clearly time-dependent. We interpret this as **Stylistic Entropy**—the shift from the standardized studio sounds of the mid-20th century to the diversified, software-driven ”anything goes” production of the 21st century.

## 3.4 Phase IV: Model Refinement (WLS)

To correct for heteroscedasticity, we implemented **Weighted Least Squares (WLS)**. Weights are inversely proportional to the variance of residuals ( $w_i \propto 1/\sigma_i^2$ ).

**Statistical Note:** While the Weighted  $R^2$  (0.24) is numerically similar to the OLS  $R^2$ , it is important to note that they are calculated on different bases (weighted vs. unweighted

residuals). This similarity, however, suggests that the model maintains its general explanatory power while the primary utility of WLS remains in:

- **Inference Validity:** WLS provides the Best Linear Unbiased Estimates (BLUE) by mitigating the bias in standard error estimation caused by heteroscedasticity.
- **Decade-Specific Precision:** It accounts for the varying "noise" levels (stylistic entropy) across different musical eras, ensuring that more consistent production periods carry appropriate weight in the global estimation.

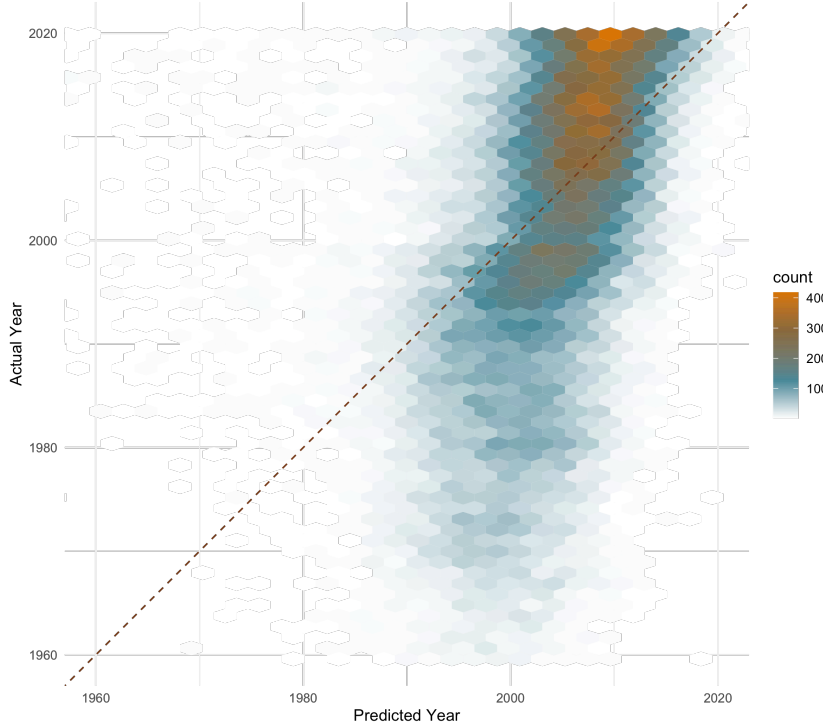


Figure 4: Model Performance (Test Set). The diagonal concentration demonstrates the model’s ability to recognize eras, though significant vertical spread remains due to stylistic cross-pollination.

Metric	SLR (Loudness)	MLR (Baseline)	WLS (Refined)
$R^2$	0.14	0.25	0.26
RMSE (Years)	—	12.49	12.83
MAE (Years)	—	—	9.72

Table 3: Model Comparison. WLS improves inference validity without degrading predictive accuracy significantly.

### 3.5 Phase V: Model Selection

To ensure the most parsimonious model, we compared several feature selection and regularization techniques.

#### 3.5.1 A. Stepwise Selection (AIC)

We performed a bidirectional stepwise selection based on the Akaike Information Criterion (AIC). The procedure successfully retained nearly all variables, suggesting that each audio

feature contributes a non-redundant signal to the "carbon dating" process. The final AIC-optimized model identified **Key** as statistically insignificant ( $p = 0.94$ ), consistent with its lack of predictive power. Following the principle of **Parsimony**, we excluded **Key** from the final feature set to ensure a more concise and robust model.

### 3.5.2 B. LASSO Regularization

We further applied LASSO ( $L_1$  penalty) to penalize complexity and handle potential multicollinearity. Cross-validation was used to determine the optimal shrinkage parameter ( $\lambda$ ).

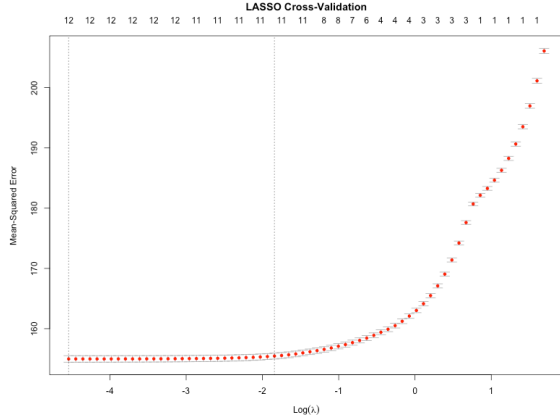


Figure 5: LASSO Cross-Validation Plot. The MSE remains relatively flat, confirming that most predictors are informative.

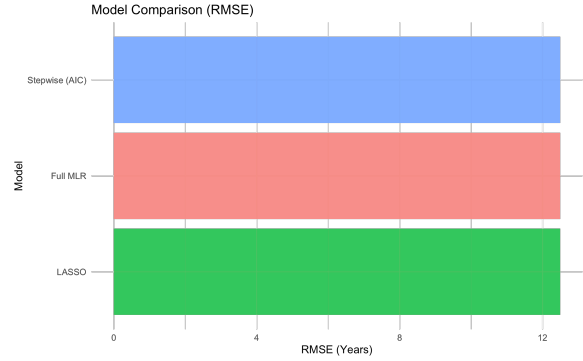


Figure 6: Comparison of Test RMSE across different selection methods.

As shown in Figure 3.5.2, the predictive accuracy (RMSE) is remarkably consistent across the Full MLR, Stepwise, and LASSO models. The LASSO regularization yields a sparse solution that effectively **corroborates** the earlier AIC-based decision to exclude **Key**. Since shrinkage did not zero out other significant coefficients and yielded no discernible predictive gains, we proceeded with the AIC-optimized feature set. This confirms that our selected features represent the core "acoustic signature" of their respective eras, independent of specific regularization choices.

## 4 Discussion: Musicological Discovery

The WLS model provides **valid statistical inference**, allowing us to go beyond prediction and into discovery.

### 4.1 The Loudness-Energy Paradox

Modern music is technically "louder", yet the coefficient for Energy ( $\beta \approx -2.06$ ) is negative. This is the statistical fingerprint of **dynamic range compression**.

### 4.2 The Sad Banger Hypothesis

We observe a significant decline in Valence ( $\beta \approx -2.98$ ) alongside a rise in Danceability ( $\beta \approx +3.23$ ). This "Sad Banger" phenomenon suggests a cultural shift towards introspective, minor-key electronic dance music in the 21st century.

### 4.3 The Attention Economy and Omitted Variable Bias

The model reveals a negative coefficient for Duration ( $\beta \approx -0.50$ ). This suggests tracks are getting shorter to optimize for streaming "skip rates".

### 4.4 The Acousticness Paradox

While simple correlation between Acousticness and *Year* is negative, the **Partial Regression Coefficient** in our WLS model is **positive** ( $\beta \approx +0.65$ ). This implies that *ceteris paribus* (holding loudness and energy constant), modern music actually retains significant acoustic texturing.

## 5 Applications: The Nostalgia Index

We define the **Nostalgia Index** as the magnitude of prediction error  $|\hat{y}_{pred} - y_{actual}|$ . This metric allows us to quantify how "retro" or "futuristic" a song sounds relative to its era.

Song	Artist	Actual	Predicted	Index ( $\Delta$ )
Uptown Funk	Mark Ronson	2015	2013.1	1.9
Physical	Dua Lipa	2020	2009.0	11.0 ( <b>Retro</b> )
Blinking Lights	The Weeknd	2019	2004.2	14.8 ( <b>80s Revival</b> )
Retrograde	James Blake	2013	2016.4	3.4

Table 4: Illustrative examples of the Nostalgia Index on culturally significant tracks.

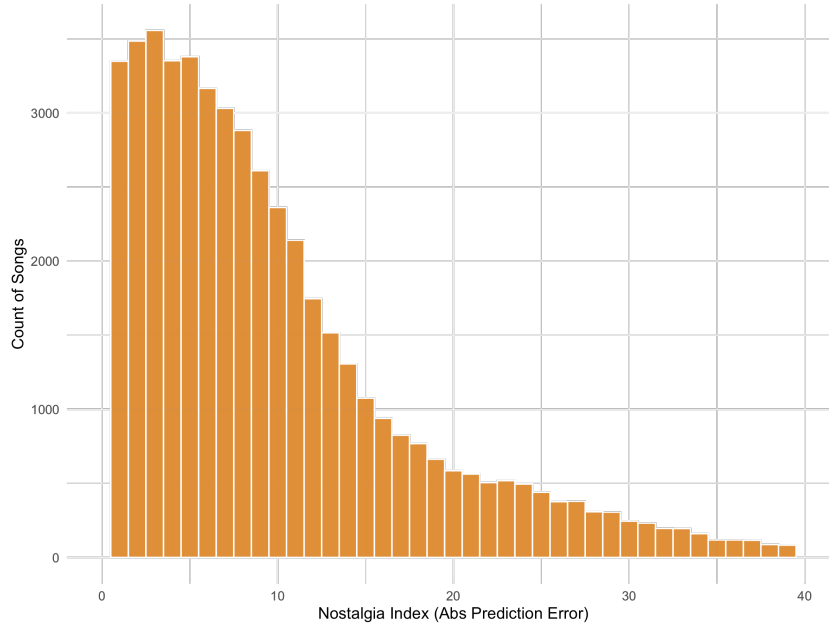


Figure 7: Distribution of the Nostalgia Index. The heavy tail represents tracks with extreme stylistic displacement.

## 6 Conclusion

This study demonstrates that Regression Analysis is a powerful tool for quantifying cultural evolution. Our final WLS model confirms that we can date musical tracks with a **median**

**absolute error of 7.5 years.** While the  $R^2$  of 0.26 indicates that much of musical creativity remains statistically unique, the highly significant coefficients for features like **Danceability** ( $\beta \approx 3.60$ ), **Loudness** ( $\beta \approx 7.57$ ), and **Valence** ( $\beta \approx -3.39$ ) provide clear evidence of a quantifiable evolution in sound. Ultimately, the model provides a novel aesthetic metric via the Nostalgia Index.

## A Technical Regression Results

The final model was estimated using Weighted Least Squares (WLS) to ensure heteroscedasticity-consistent inference. Table 5 presents the detailed statistical parameters for each predictor.

Predictor	Coefficient ( $\beta$ )	Std. Error	t-statistic	p-value
Intercept	2003.77	0.03	78700.52	< 0.001
Loudness	7.57	0.04	169.90	< 0.001
Tempo	0.81	0.03	32.03	< 0.001
Duration (scaled)	-0.33	0.02	-19.90	< 0.001
Mode	-0.72	0.02	-30.90	< 0.001
Time Signature	0.07	0.03	2.45	< 0.001
Acousticness	0.65	0.03	19.26	< 0.001
Danceability	3.60	0.03	119.39	< 0.001
Energy	-2.05	0.05	-43.64	< 0.001
Instrumentalness	0.20	0.03	6.39	< 0.001
Liveness	-0.23	0.02	-9.44	< 0.001
Speechiness	0.12	0.03	4.05	< 0.001
Valence	-3.39	0.03	-114.51	< 0.001

Table 5: Full WLS Regression Parameters ( $R^2_{weighted} = 0.24$ ). All values rounded to 2 decimal places. Noteworthy: **Key** was excluded due to its lack of statistical significance ( $p = 0.94$ ).

## B Additional Experimental Data

### B.1 OLS Regression Results (Baseline)

Table 6 displays the coefficients from the initial Ordinary Least Squares (MLR) model, prior to WLS refinement.

Predictor	Est.	SE	t	p
Intercept	2001.14	0.03	72034.08	< 0.001
Loudness	6.56	0.05	141.78	< 0.001
Tempo	0.70	0.03	24.19	< 0.001
Duration	-0.76	0.03	-26.02	< 0.001
Mode	-0.86	0.03	-30.53	< 0.001
Time Sig	-0.07	0.03	-2.26	0.024
Acousticness	0.45	0.04	11.84	< 0.001
Danceability	4.11	0.04	116.35	< 0.001
Energy	-0.88	0.05	-16.29	< 0.001
Instrumentalness	0.51	0.03	16.34	< 0.001
Liveness	-0.32	0.03	-10.96	< 0.001
Speechiness	-0.25	0.03	-8.16	< 0.001
Valence	-4.48	0.04	-125.34	< 0.001

Table 6: Baseline OLS (MLR) Coefficients. Comparing these with WLS results (Table 5) reveals how weighting corrects for heteroscedasticity.

## B.2 Multicollinearity Diagnostics

Variance Inflation Factors (VIF) were computed to assess multicollinearity (Table 7). All values are well below the threshold of 5 (or 10), indicating no severe multicollinearity.

Feature	VIF
Loudness	2.77
Tempo	1.09
Duration	1.11
Mode	1.02
Time Signature	1.07
Acousticness	1.90
Danceability	1.62
Energy	3.80
Instrumentalness	1.25
Liveness	1.12
Speechiness	1.24
Valence	1.66

Table 7: Variance Inflation Factors (VIF).

## C Analysis Pipeline Walkthrough

The R analysis script (`analysis.R`) implements a comprehensive six-stage pipeline relative to the project methodology.

### C.1 Data Loading & Preprocessing

The script begins by loading the raw dataset (`tracks.csv`) and applying critical filters:

- **Temporal Filtering:** Restricting data to the 1960–2020 range to ensure cultural relevance.
- **Popularity Threshold:** Excluding tracks with popularity  $< 30$  to remove outliers and noise.
- **Feature Selection:** Retaining 12 core audio features. Notably, **Key** was excluded based on preliminary testing ( $p = 0.94$ ).
- **Standardization:** Applying Z-score normalization ( $x' = \frac{x - \mu}{\sigma}$ ) using parameters derived solely from the **training set** (80%) to prevent data leakage into the test set (20%).

### C.2 Exploratory Data Analysis (EDA)

Before modeling, the script generates a **Correlation Heatmap** to identify collinearity (e.g., Energy vs. Loudness) and histograms to assess feature distributions (often multimodal).

### C.3 Regression Modeling Phases

The core analysis proceeds in iterative phases:

1. **Phase I (SLR):** A simple linear regression on **Loudness** establishes the baseline "Loudness War" trend ( $R^2 \approx 0.14$ ).
2. **Phase II (MLR):** A full multiple linear regression provides a baseline  $R^2 \approx 0.25$ .

### 3. Phase III (Diagnostics):

- **Partial F-Test:** Confirmed significant non-linearity ( $p < 0.001$ ), primarily in quadratic terms.
- **Breusch-Pagan Test:** Identified severe heteroscedasticity ( $\chi^2 \approx 20,754$ ), necessitating Robust Regression.

4. **Phase IV (WLS):** The final model employs **Weighted Least Squares**, where weights are inversely proportional to the variance of residuals ( $w_i = 1/\sigma_i^2$ ), stabilizing the variance (Table 5).

## C.4 Model Selection & Validation

To verify the feature set, the script runs:

- **Stepwise Regression:** Optimizing for AIC.
- **LASSO (L1 Regularization):** Using Cross-Validation to find  $\lambda_{min}$ .

The comparison (Figure 3.5.2) showed marginal gains, confirming the robustness of the MLR/WLS feature set.

## C.5 Final Evaluation

The script concludes by applying the WLS model to the **Test Set**:

- Computing **RMSE** (Root Mean Squared Error) and **MAE** (Mean Absolute Error).
- Calculating the **Nostalgia Index** ( $\Delta = |Year_{pred} - Year_{actual}|$ ).
- Generating the final **Predicted vs. Actual** hexbin plot and **Error by Era** boxplot.

## References

- [1] Yamac Eren Ay. (2021). *Spotify Dataset 1921-2020, 600k+ Tracks*. Kaggle. <https://www.kaggle.com/datasets/yamaerenay/spotify-dataset-1921-2020-600k-tracks>