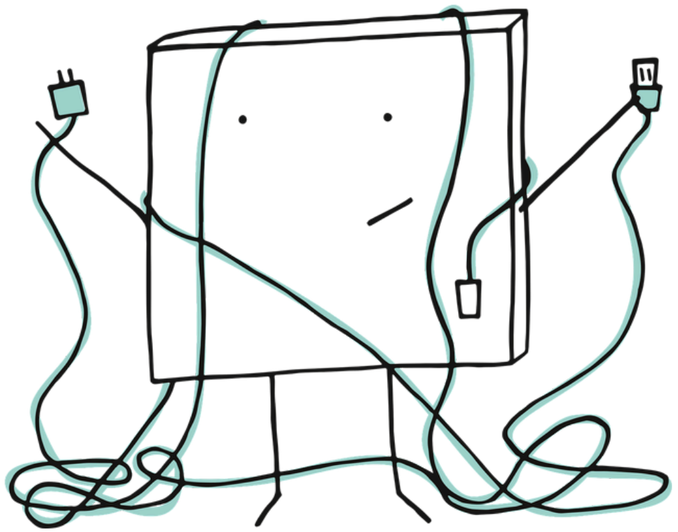
MULTIVARIATE ANALYSIS OF SPOTIFY TRACK POPULARITY

Ismet Eren - 2561249
Ahmet Doğan - 2426401
Ismet Selman Loğoğlu - 2507515

PROBLEM
MAIN RESEARCH QUESTIONS

- Can **audio features** (e.g., Danceability, Energy, Valence) effectively discriminate between **Popular** and **Non-Popular** tracks?
- How are the **physical properties** of sound (Loudness, Tempo) related to the **emotional qualities** (Valence, Danceability)?



DATASET DESCRIPTION

The original dataset was obtained from the Spotify Web API and contains 114,000 tracks across 114 different genres. For computational efficiency and statistical manageability, a random sample of 10,000 tracks was drawn using stratified sampling

TABLE I: Complete Variable Inventory

Variable Type	Count	Variables
Continuous Numeric	10	danceability, energy, loudness, speechiness, acousticness, instrumentalness, liveness, valence, tempo, duration_ms
Categorical (Nominal)	5	track_id, artists, album_name, track_name, track_genre
Categorical (Ordinal)	3	key (0-11), mode (0-1), time_signature (3-7)
Binary	2	explicit (True/False), popularity (derived: 0/1)

TIDYING DATA

- Missing Values: **No missing** values were detected across all 20 cont. variables
- Duplicates: **Duplicate** tracks (based on track_id) were identified and **removed**
- Outliers: **Duration** outliers were **filtered** using the **1st-99th percentile** range, removing extremely short clips (**<30 seconds**) and unusually long tracks (**>10 minutes**)
- Data Types: All variables were verified to have appropriate data types for analysis

- **0 (Not Popular):** popularity score ≤ 50
- **1 (Popular):** popularity score > 50

EXPLORATORY DATA ANALYSIS

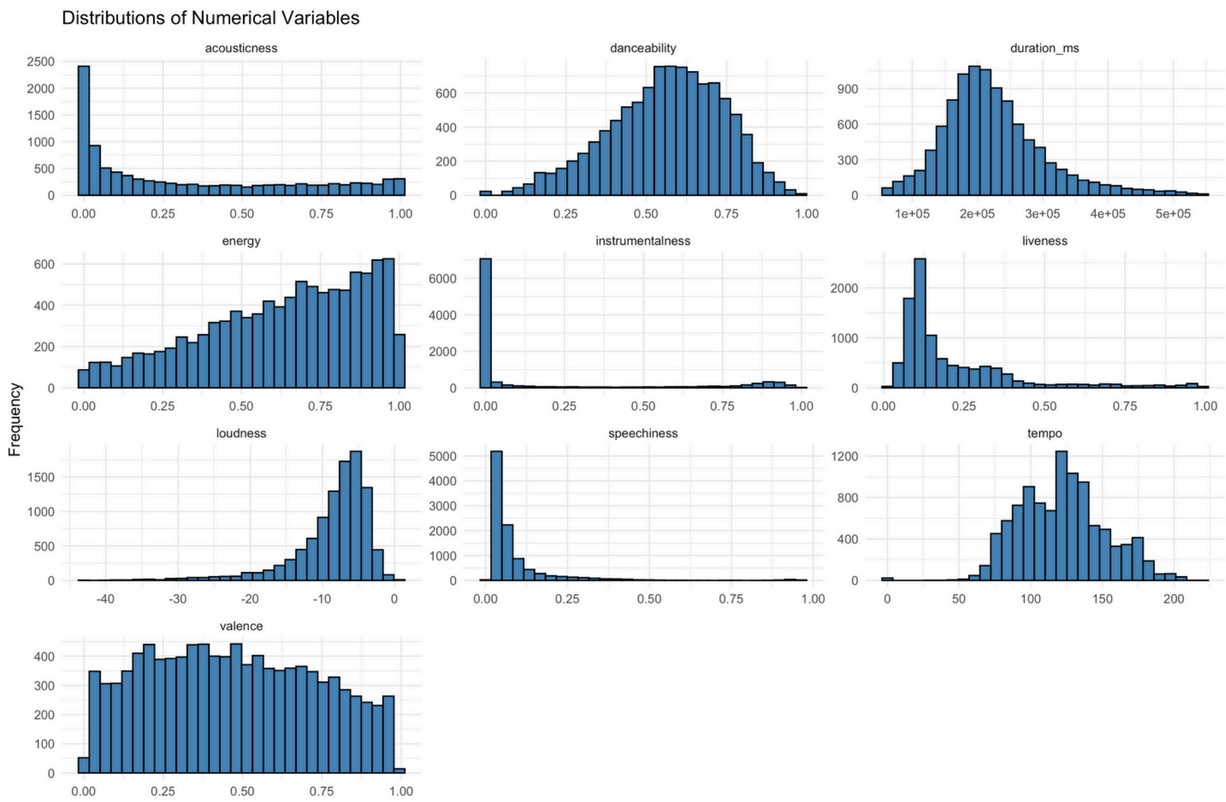


***Fig. 1**. Distribution histograms for all 10 continuous audio features*

- **Univariate Analysis:** Histograms for all 10 continuous variables to assess distributional shapes, identify skewness, and detect potential outliers

The diverse distributional shapes, ranging from Gaussian (tempo) to strictly positive/skewed (speechiness), highlight the necessity of variable-specific transformations independent of simple linear scaling.

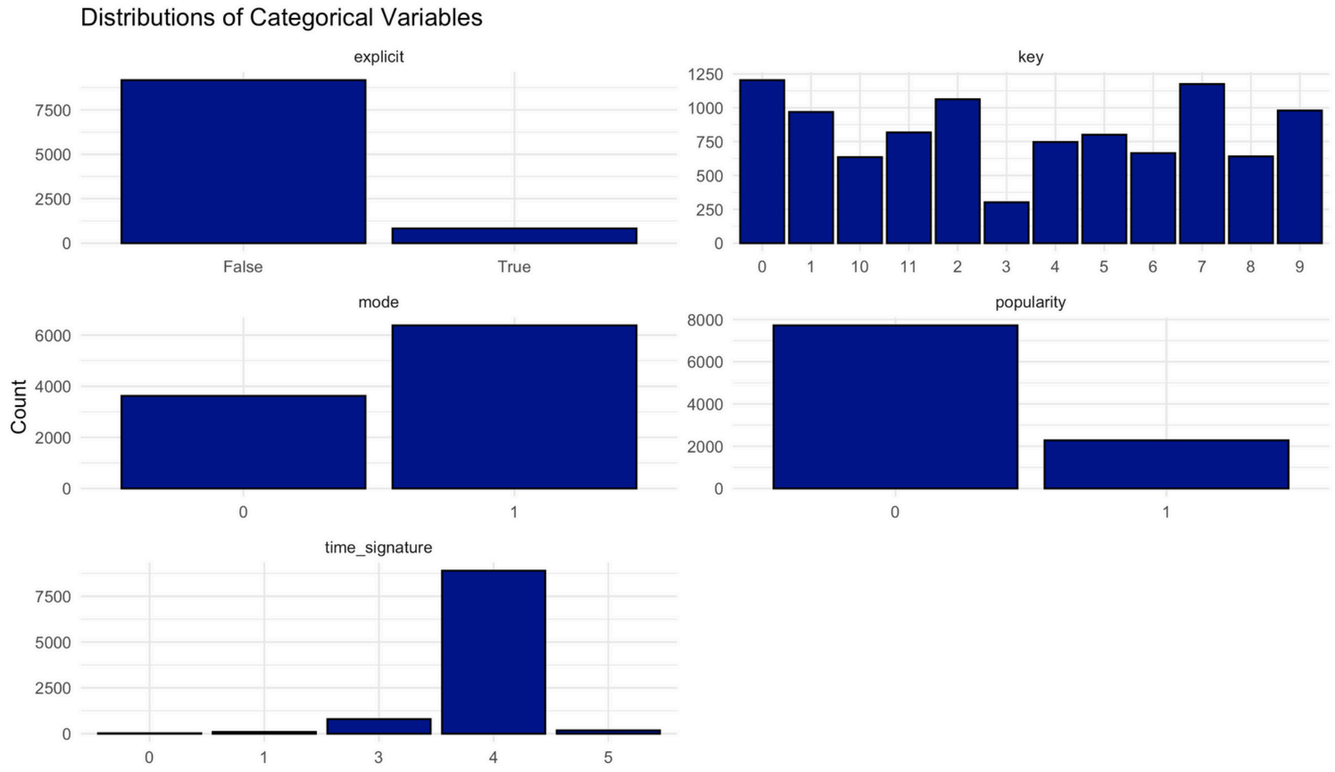


Fig. 2. *Frequency distributions of categorical variables.*

- **Bivariate Analysis:** Bar plots for categorical variables (explicit, key, mode, time_signature, popularity) to understand frequency distributions

The class imbalance (40% popular) is moderate. The dominance of Major keys and 4/4 time signatures reflects the standardized structure characteristic of Western commercial music.

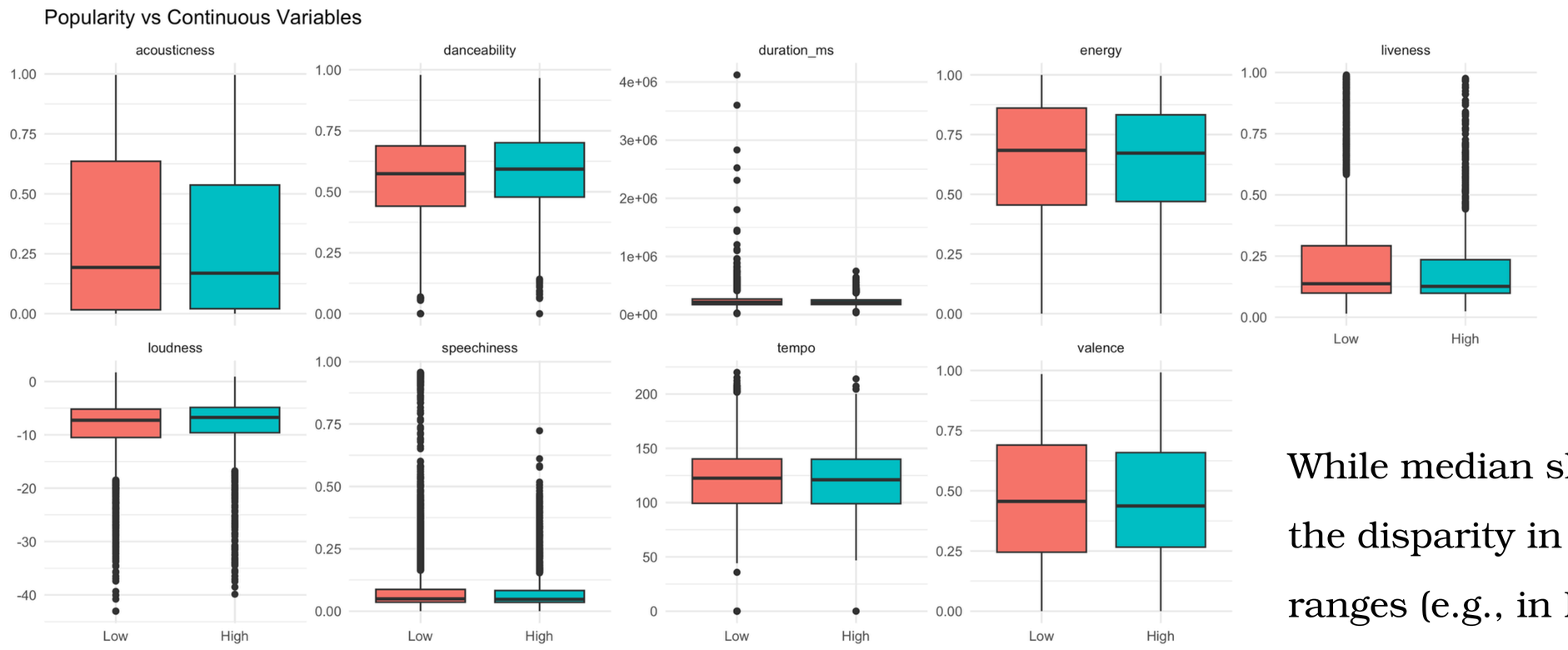


Fig. 3. *Bivariate boxplots displaying the distributions of continuous audio features across popularity groups.*

While median shifts are subtle, the disparity in interquartile ranges (e.g., in Loudness) suggests that popular tracks exhibit less variance (are more formulaic) than non-popular ones.

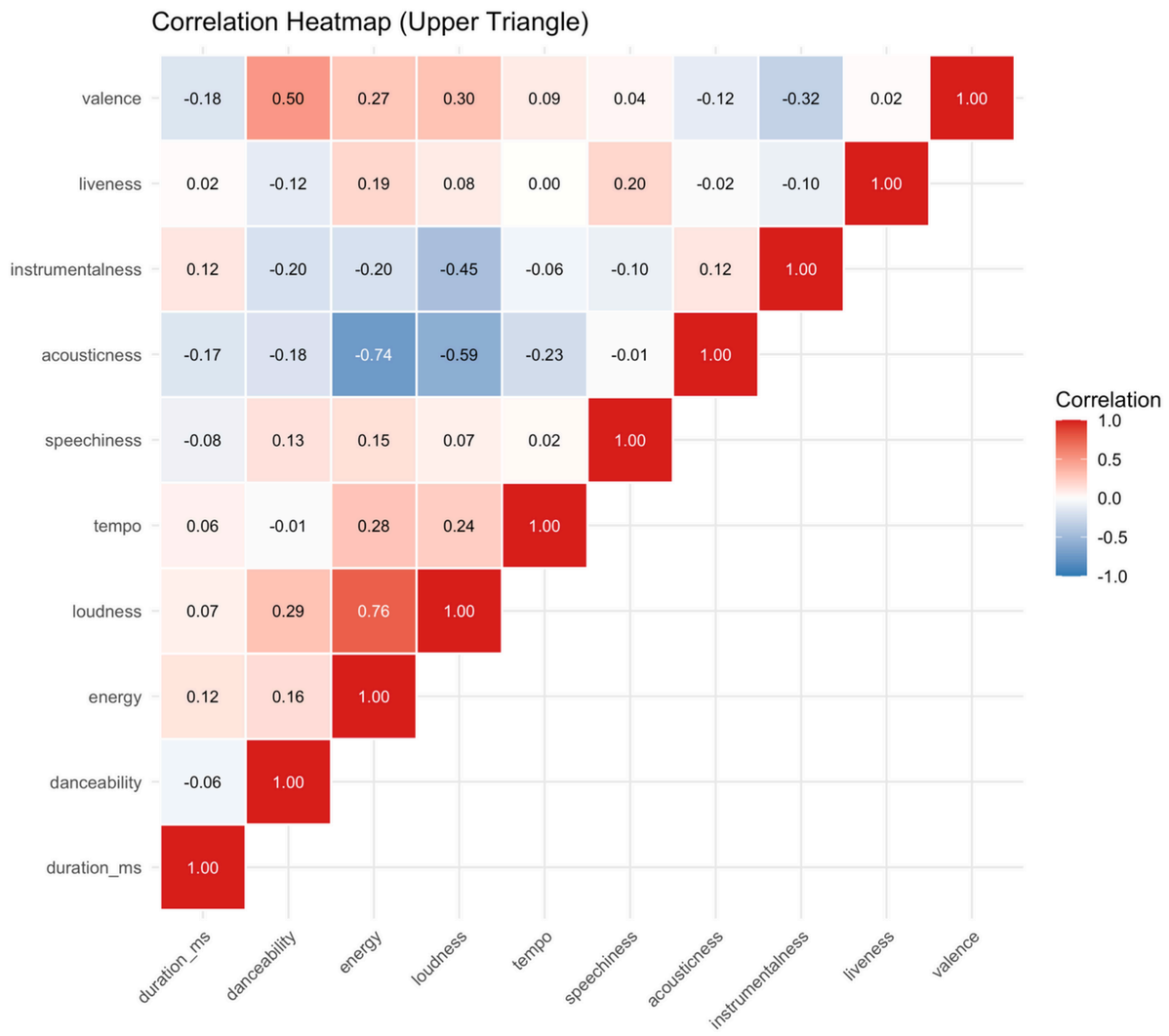


Fig. 4. *Pearson correlation heatmap of continuous audio features*

- **Multivariate Analysis:** Correlation heatmap to identify linear relationships among continuous features and potential multicollinearity issues

The Pearson correlation matrix (Fig. 4) revealed several statistically and substantively significant relationships among the audio features. Structurally, 'Energy' acts as a latent factor driving both 'Loudness' (+) and 'Acousticness' (-)

INFERENCES ABOUT A MEAN VECTOR

Assumption	Test Method	Test Statistic	Result	Implication
Univariate Normality	Q-Q Plots (Visual)	-	✓ Satisfied	All 9 continuous variables normalized
Multivariate Normality	Chi-square Q-Q Plot (D^2)	$D^2 \sim \chi^2(10)$	✓ Satisfied	After outlier removal, distribution confirmed
No Multivariate Outliers	Mahalanobis Distance	Cutoff: $\chi^2(10, 0.001) = 29.59$	✓ Satisfied	198 outliers (2%) removed
Independence	Study Design	-	✓ Assumed	Random sample from Spotify database
Homogeneity of Covariances	Box's M Test	$M = 892.4$, $F \approx 16.2$	✗ Violated ($p < 0.001$)	Groups have different covariance structures

To validate the Multivariate Normality assumption and draw valid inferences about the mean vector, we analyzed the distributional properties and applied necessary transformations.

NORMALIZATION

TABLE IV: bestNormalize Transformation Selection for Each Variable

Variable	Transformation Method	Original Skewness	Post-Transform Skewness
duration_ms	Box-Cox	+5.82 (Severely Right)	~0.0
tempo	Box-Cox	+0.25 (Approx. Normal)	~0.0
loudness	Box-Cox	-1.15 (Left)	~0.0
danceability	Yeo-Johnson	-0.48 (Approx. Normal)	~0.0
energy	Yeo-Johnson	-0.72 (Moderately Left)	~0.0
valence	Yeo-Johnson	+0.08 (Approx. Normal)	~0.0
speechiness	Ordered Quantile (ORQ)	+2.85 (Severely Right)	~0.0
liveness	Ordered Quantile (ORQ)	+1.58 (Right)	~0.0
acousticness	Ordered Quantile (ORQ)	+1.12 (Right)	~0.0

The **bestNormalize** algorithm from the R library selected the following transformations for each variable:

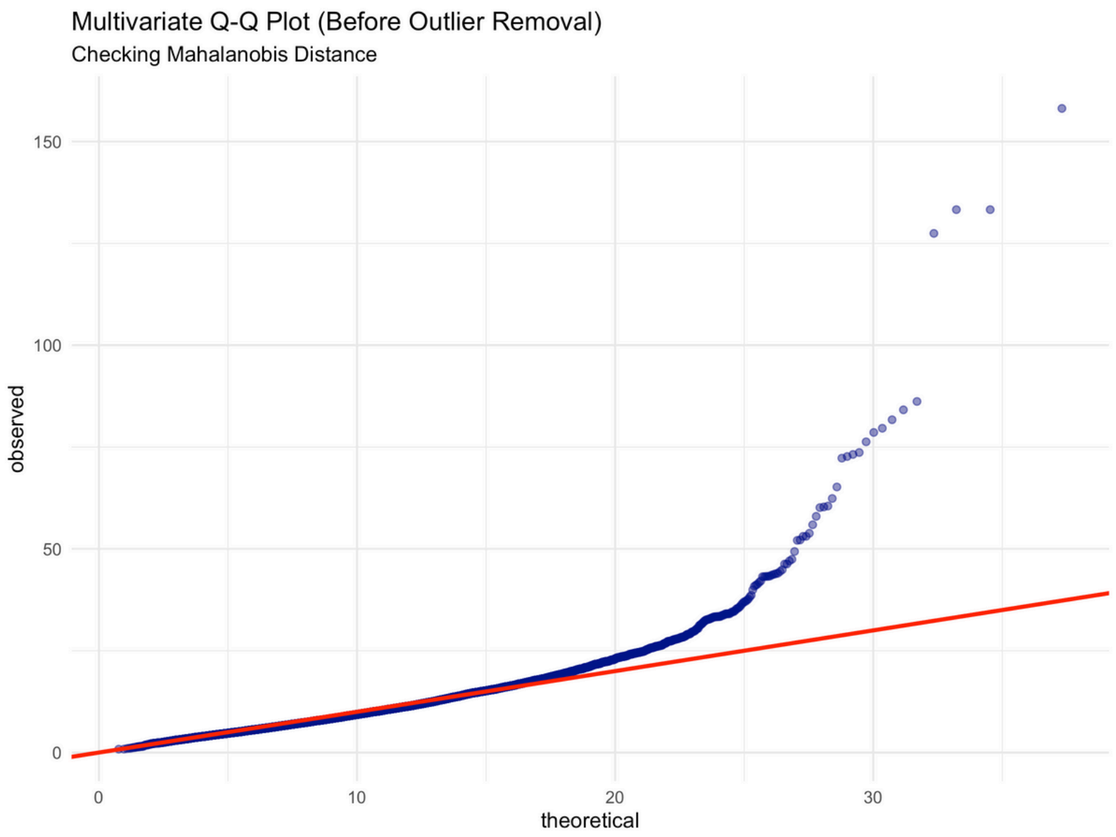
TABLE V: Mahalanobis Distance Outlier Detection Summary

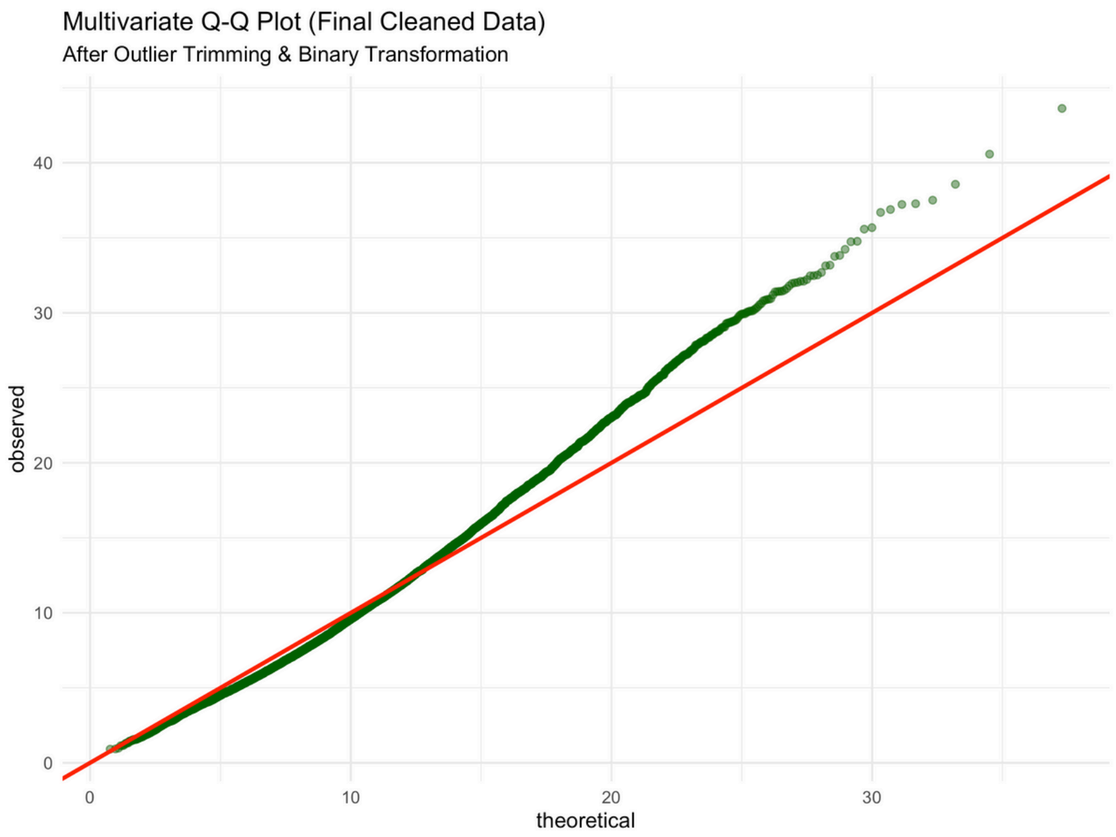
Statistic	Value
Original Sample Size	$n = 10{,}000$
Number of Variables (p)	10
Chi-Square Cutoff ($\alpha = 0.001$)	$\chi^2(10) = 29.59$
Maximum Observed D^2	~52.4
Outliers Identified	198 (1.98%)
Final Sample Size	$n = 9{,}802$

Mahalanobis distances were computed for all observations using the normalized data matrix.

- Good adherence to the reference line for ~95% of observations
- Severe upper-tail deviation indicating multivariate outliers
- Maximum D^2 values exceeding 50 (compared to $\chi^2(0.999, 10) = 29.59$ cutoff)

MULTIVARIATE NORMALITY ASSESSMENT





The near-perfect fit ($R^2 \approx 1$) validates the assumption of multivariate normality for the cleaned dataset, legitimizing the subsequent use of **Mahalanobis**-based inference.

COMPARISONS OF SEVERAL MULTIVARIATE MEANS

The core objective was to test if the mean vector of audio features differs between "Popular" and "Not Popular" tracks ($H_0 : \boldsymbol{\mu}_{pop} = \boldsymbol{\mu}_{non-pop}$).

Assumption Checking: Box's M Test

We first tested the equality of covariance matrices ($H_0 : \boldsymbol{\Sigma}_{pop} = \boldsymbol{\Sigma}_{non-pop}$) using Box's M test.

Box's M Result:
Chi-Sq = 892.4, p < 0.001
Conclusion: Reject H_0. Covariance matrices are NOT equal.

The violation of homogeneity of covariance matrices ($p < 0.001$) invalidates standard MANOVA (Wilks' Lambda). Therefore, we proceeded with a robust alternative.

ROBUST TEST: PERMANOVA

Source	Df	Sum of Squares	R2	F	Pr(>F)
Model (Popularity)	1	36.5	0.00199	3.9913	0.005 **
Residuals	1998	18267.5	0.99801		
Total	1999	18304.0	1.00000		

To overcome the assumption violation, we employed Permutational Multivariate Analysis of Variance (PERMANOVA) using **Euclidean distances** and **199 permutations.**

The test yielded a statistically significant result (***F* = 3.99, *p* = 0.005**), rejecting the null hypothesis of equal centroids. However, the effect size ($R^2 \approx 0.002$) is practically negligible, indicating that popularity status explains only 0.2% of the total variance in audio features.

PRINCIPAL COMPONENTS ANALYSIS AND PRINCIPAL COMPONENTS REGRESSION

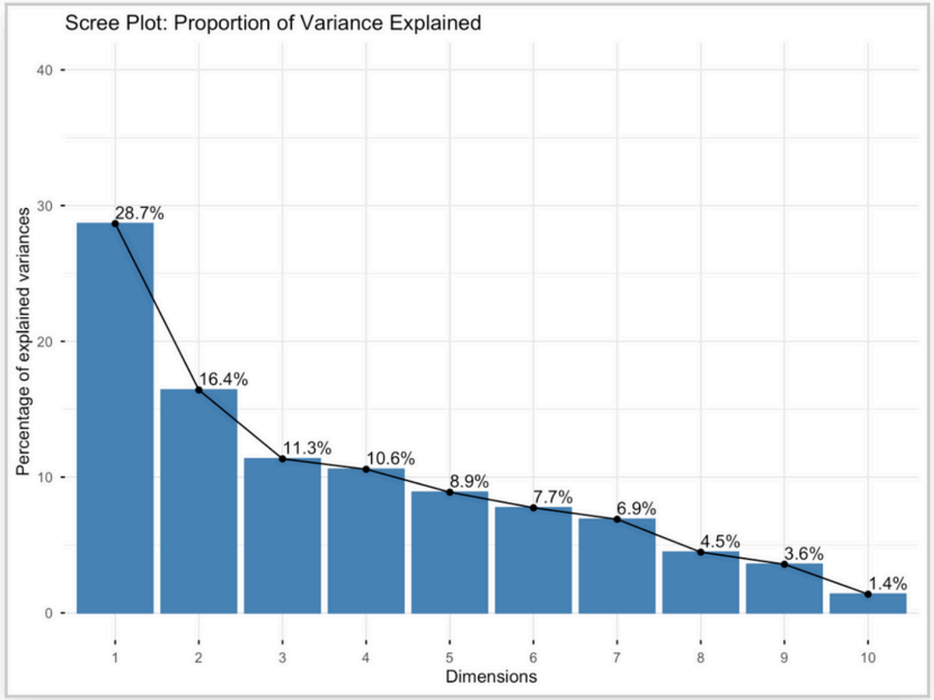


Fig. 4a. Scree Plot showing variance explained by each PC.

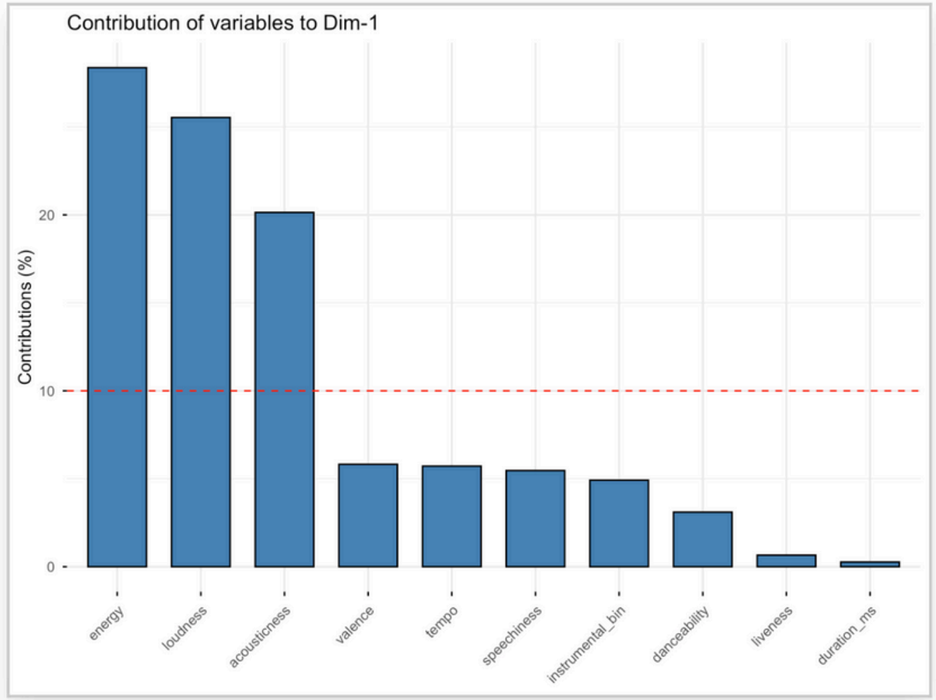
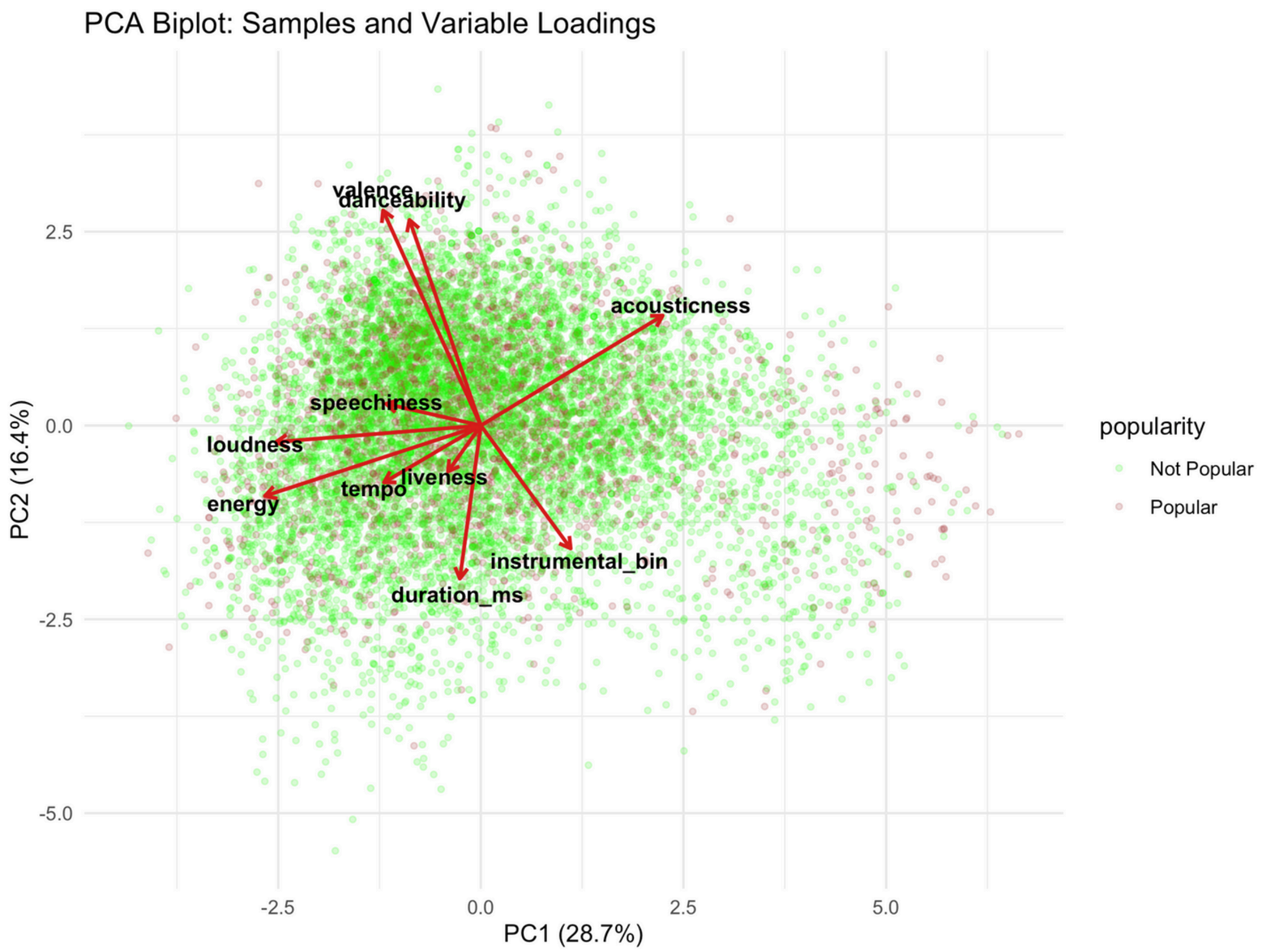
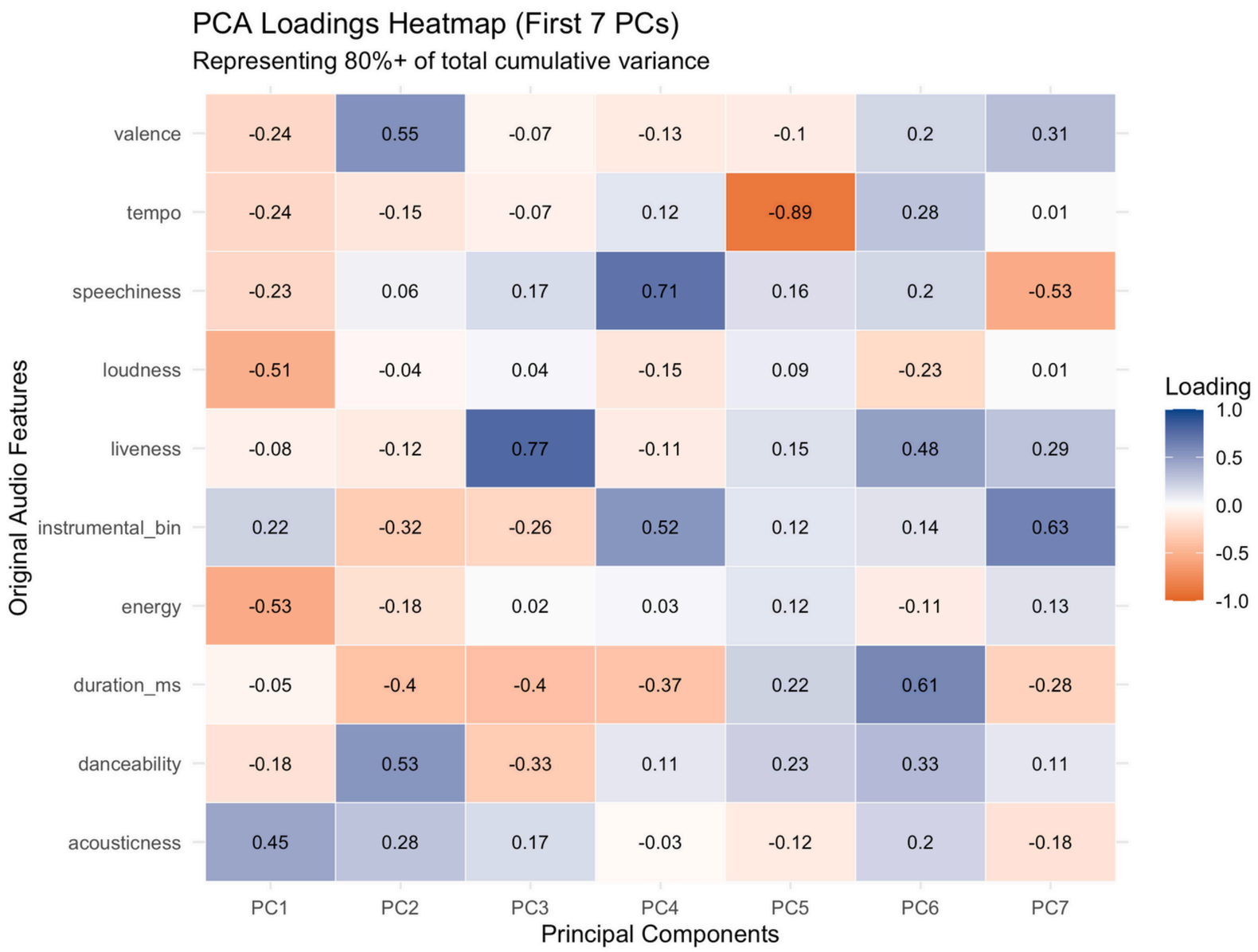


Fig. 4b. Variable Contributions to the First Principal Component.

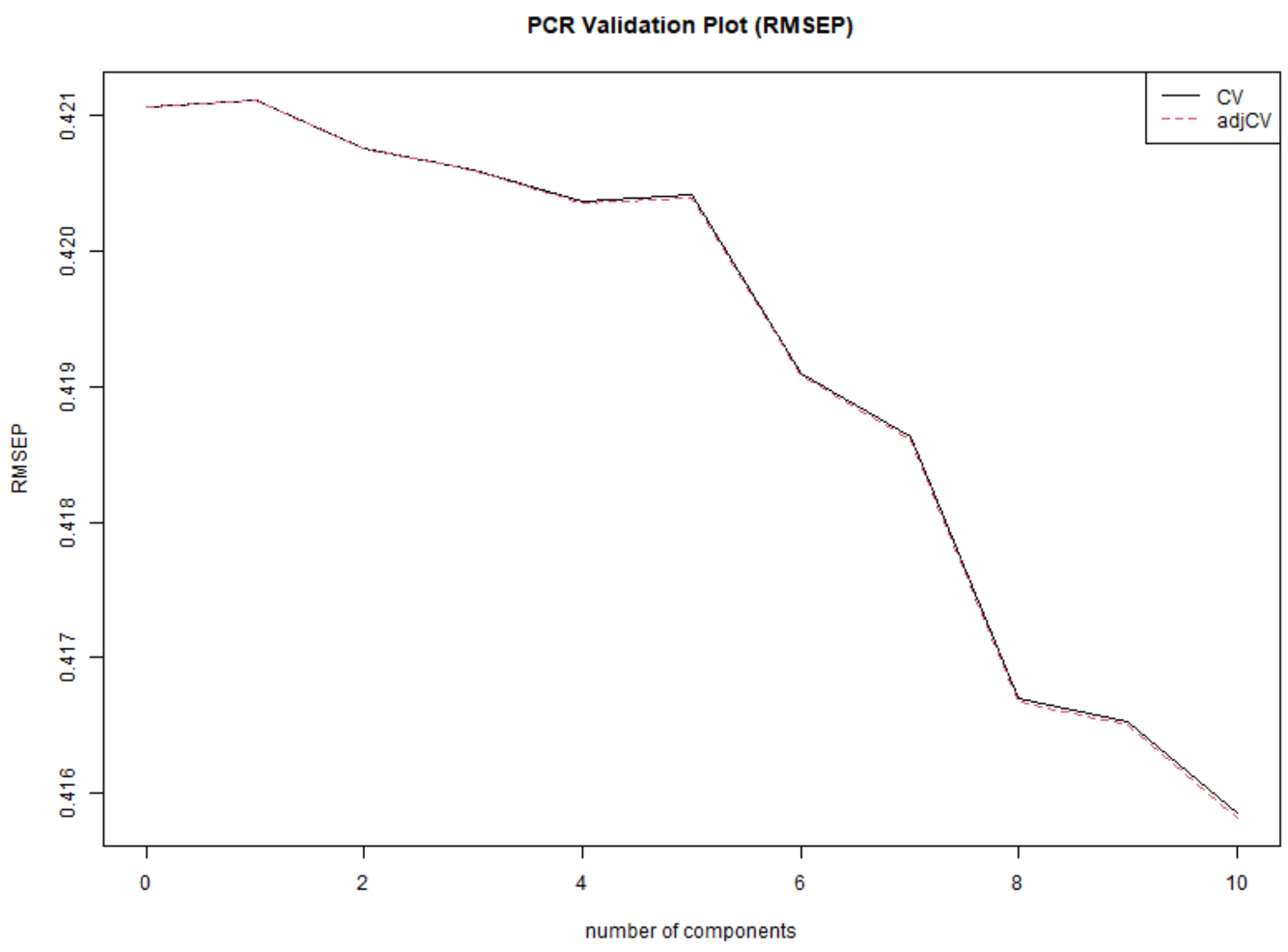
The scree plot (Fig. 4a) suggests that the **first 4 components** capture the majority of the information (**Cumulative Variance ≈ *67%***). The first principal component (PC1) is dominated by Loudness, Energy, and Acousticness (Fig. 4b), representing a "Loudness/Energy" axis.



The Biplot visualizes the samples in the PC1-PC2 space. A crucial observation is the **absence of distinct clusters** corresponding to popularity. Both "Popular" and "Not Popular" tracks are dispersed throughout the entire feature space, heavily overlapping.



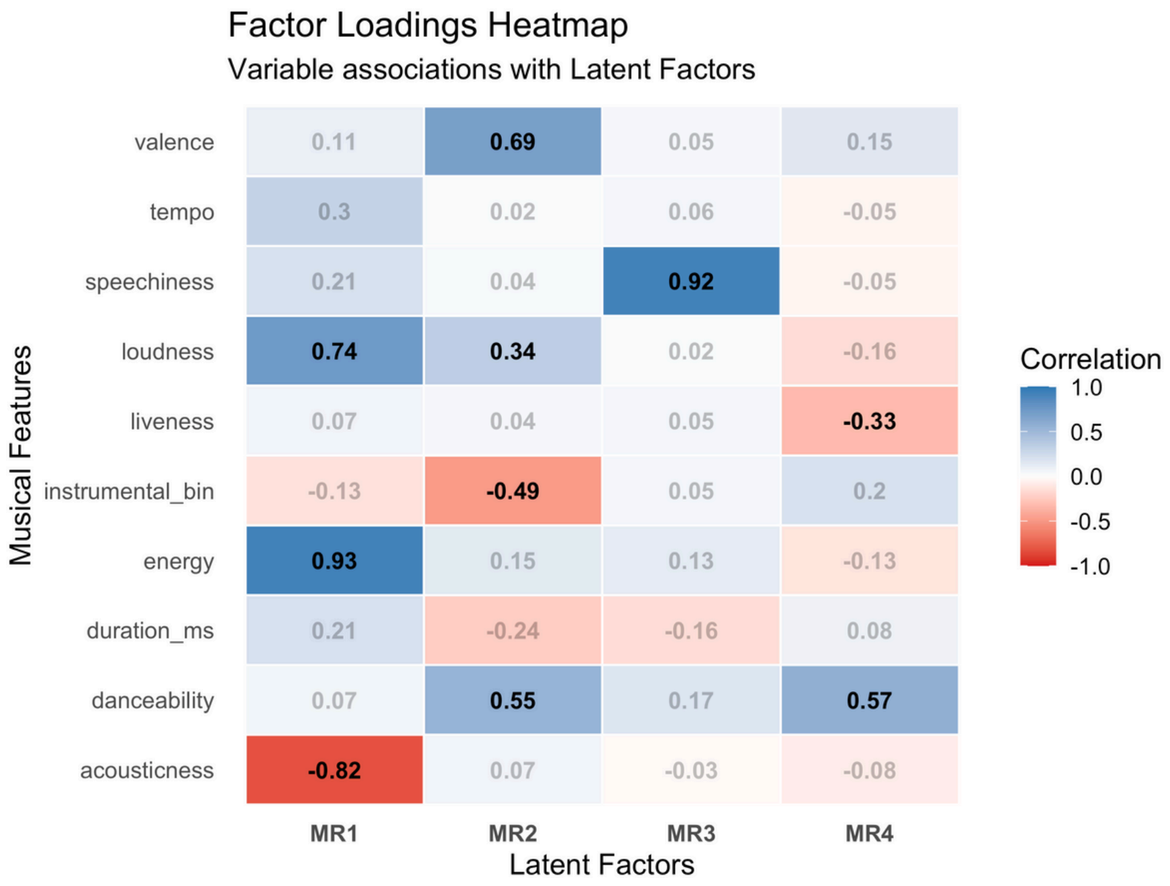
The Heatmap provides a detailed explanation of how original variables map to principal components. PC1 is clearly an "Intensity" component (High Energy/Loudness, Low Acousticness), while PC2 is a "Rhythm" component (High Valence/Danceability). Despite this clear structural definition of musical features, they do not map to popularity success.



While the model successfully captured 100% of the variance in the audio features (X) using 10 components, it failed to translate this information into popularity prediction. As seen in Table V, the explained variance for Popularity (Y) saturated at a mere 2.69%, and the error rate (RMSEP) showed no significant improvement (0.421 to 0.416).

Components	X Variance Explained	Y Variance Explained (Popularity)
1 Comp	2.868%	2%
5 Comps	7.592%	43%
10 Comps	**10.000%**	**269%**

FACTOR ANALYSIS AND FACTOR ROTATION



This heatmap reveals the latent structure of the audio features.

The Factor Analysis yielded a 4-factor solution that is substantively more interpretable than PCA:

Factor	Dominant Variables	Label
Factor 1	Energy (+), Loudness (+), Acousticness (-)	**Physical Intensity**
Factor 2	Danceability (+), Valence (+), Speechiness (+)	**Mood & Rhythm**
Factor 3	Instrumentalness (+)	**Vocal Presence**
Factor 4	Duration_ms (+)	**Song Length**

This structure confirms that Spotify audio features reliably measure distinct musical attributes (Intensity, Mood, Instrumentality, Length)

DISCRIMINATION AND CLASSIFICATION

Linear Discriminant Analysis (LDA)

LDA assumes multivariate normality and equal covariance matrices (which we know are violated, but LDA is often robust). The results on the unbalanced test set were:

TABLE V(a): LDA Performance (Test Set)

Balanced Accuracy	55.6%
Sensitivity (Popular Class)	57.4%
Specificity (Not Popular)	53.9%

6.2 Random Forest Classifier

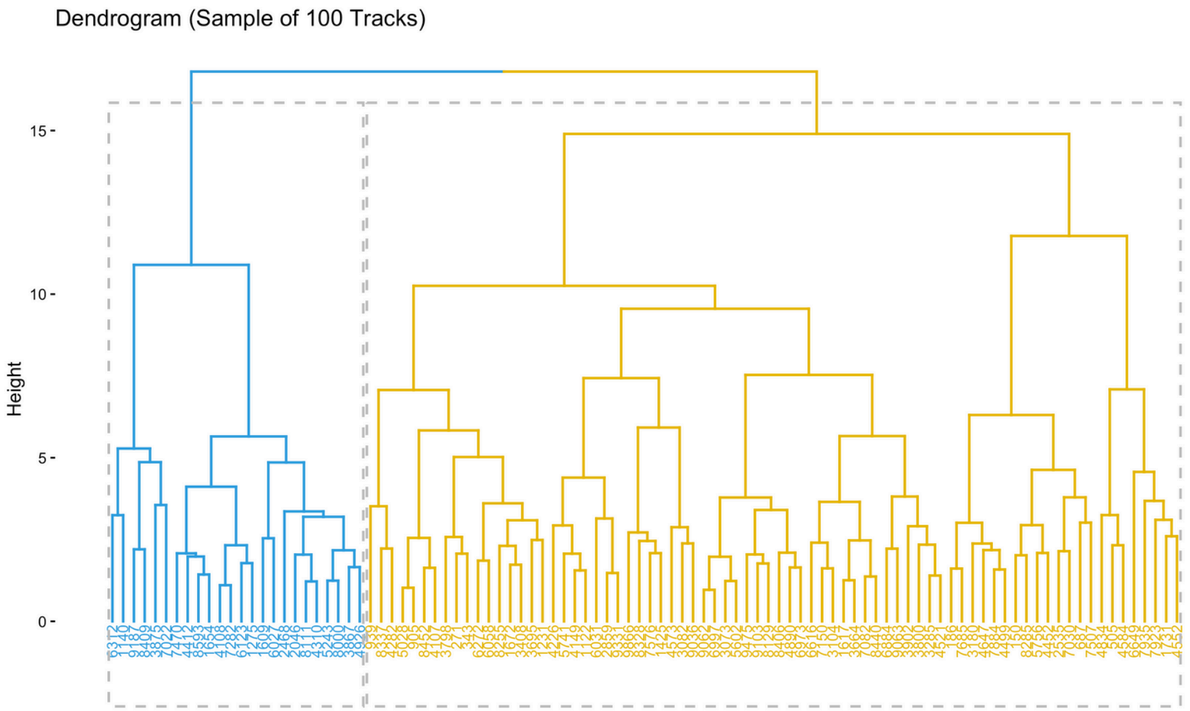
We trained a Random Forest (ntree = 500) on the same balanced training data to capture potential non-linear interactions.

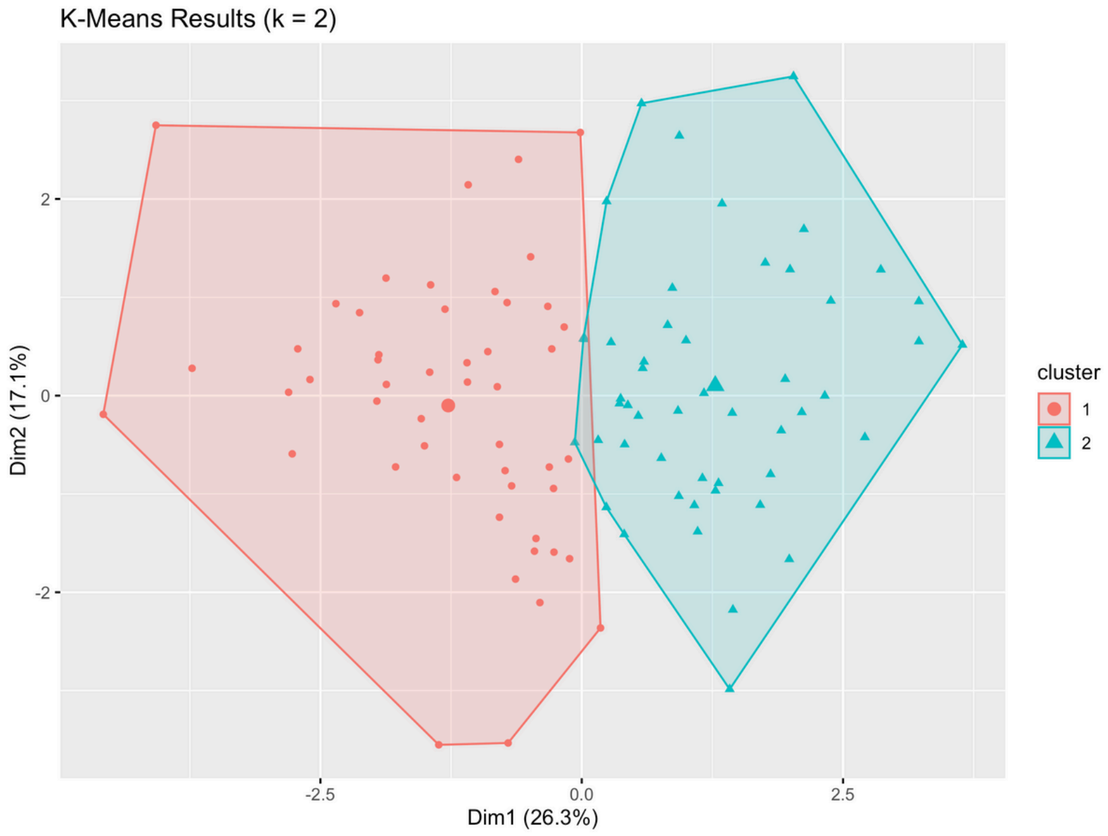
TABLE V(b): Random Forest Performance (Test Set)

Balanced Accuracy	61.4%
Sensitivity (Popular Class)	63.7%
Specificity (Not Popular)	59.1%

CLUSTERING

We switched from supervised to unsupervised learning to checking for natural group structures. We applied K-Means (k = 2) and **Hierarchical Agglomerative Clustering** (HAC, Ward's method) on a random subsample (N = 100) to visualize the dendrogram.





CLUSTERING

We evaluated the agreement between the **generated clusters** and the **actual Popularity** labels using the Adjusted Rand Index (ARI).

TABLE VI: Clustering Validation Results (N=100)

Algorithm	Mean Silhouette Width	Adjusted Rand Index (ARI)
K-Means (k=2)	0.154	**0.0004**
Hierarchical (Ward)	0.143	**0.016**

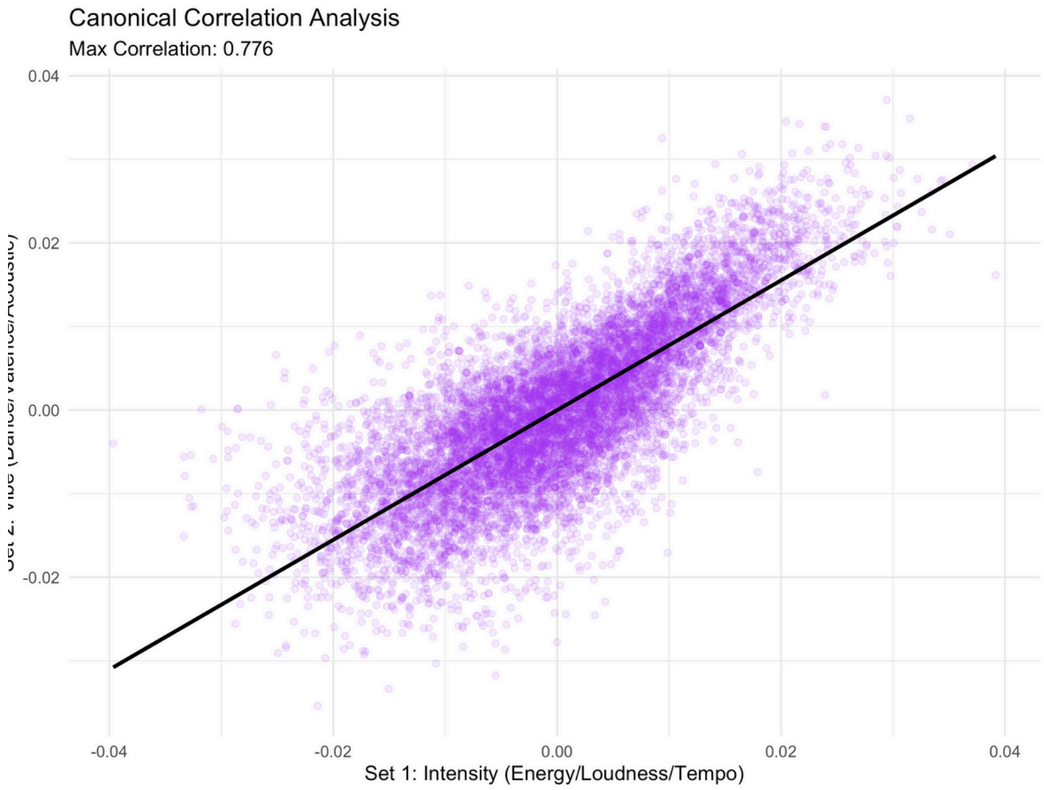
An ARI of 0 implies that the clustering is effectively random with respect to popularity. The low **Silhouette widths (< 0.2)** indicate that the data does not contain tight, **well-separated natural clusters.**

CANONICAL CORRELATION ANALYSIS

As a final exploratory step, we investigated the internal relationships between different types of audio features. We defined two sets of variables:

- **Set 1 (Physical Intensity):** Energy, Loudness, Tempo
- **Set 2 (Musical Vibe):** Danceability, Valence, Acousticness

Max Canonical Correlation: $r = 0.776$



This high correlation demonstrates that "Intensity" and "Vibe" are intrinsically linked in music production. High-intensity tracks (Loud/Fast) tend to have a specific vibe (High Valence/Low Acousticness), while low-intensity tracks map to the opposite. This successful modeling of internal structure contrasts sharply with the failure to model popularity.

CONCLUSIONS

- **Prediction Failed:** Despite rigorous statistical modeling, audio features failed to predict popularity.

- **Strong Internal Structure:** While we cannot predict popularity, we successfully modeled the music itself. CCA revealed a strong correlation ($r = 0.78$) between a track's physical intensity and its emotional vibe, confirming that the audio features are meaningful descriptors of music.

- **Implication:** Track popularity is likely driven by external factors, artist reputation, marketing budget, playlist placement, and cultural trends, rather than the audio signal itself.

THANK *YOU*