Software Project Component 2

Music Genre Classification and Clustering using GTZAN Dataset

Author: Răzvan-Ioan Călăuz, Ioana-Larisa Creț
Group: 246/1

November 19, 2025

1 Introduction

The GTZAN dataset [TC02] is a benchmark for music genre classification, containing 1000 audio tracks distributed equally across 10 genres: blues, classical, country, disco, hiphop, jazz, metal, pop, reggae, and rock. Each 30-second audio clip has been processed to extract 57 audio features including spectral characteristics (chroma, spectral centroid, bandwidth), temporal features (RMS energy, zero-crossing rate), and MFCCs (Mel-Frequency Cepstral Coefficients).

This analysis examines the dataset's statistical properties, feature relationships, and discriminative characteristics to inform our implementation of k-Nearest Neighbors (k-NN) classification and k-Means clustering algorithms. The analysis addresses:

1. Feature statistics and variance patterns
2. Correlation between features and redundancy identification
3. Feature independence assessment
4. Feature importance for genre discrimination
5. Data distribution and outlier patterns
6. Visual exploration of class separability

All analyses were performed using Python with pandas, numpy, matplotlib, seaborn, scipy, and scikit-learn libraries.

2 Basic Statistics

We calculated descriptive statistics (mean, standard deviation, variance, skewness, kurtosis) for all 57 features. Understanding these measures helps identify which features vary significantly across samples and which exhibit consistent patterns.

2.1 Feature Variance

Figure 1 shows the top 20 features ranked by variance. Variance indicates how spread out the feature values are - higher variance suggests the feature captures more diverse information across different songs and genres.

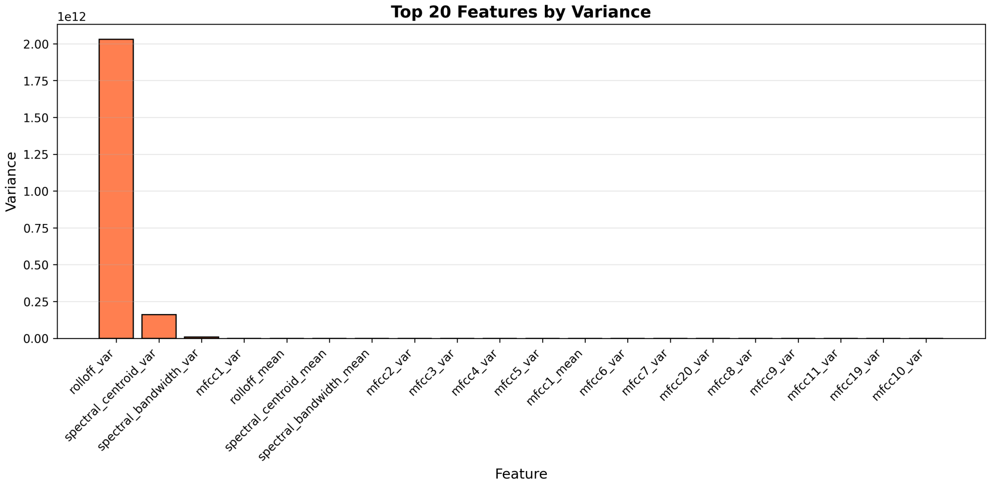


Figure 1: Top 20 features by variance. Spectral features dominate the high-variance rankings, suggesting significant differences in frequency content across genres.

Key observations:

- **Spectral features** (spectral_centroid, spectral_bandwidth) show the highest variance, meaning genres differ significantly in their frequency characteristics. This is good - it means these features can help distinguish between genres.
- **MFCC features** exhibit moderate to high variance, which reflects their ability to capture different timbral properties between genres.
- **Temporal features** (RMS, zero-crossing rate) have lower variance, meaning energy patterns are relatively consistent across the dataset.

2.2 Skewness Distribution

Skewness measures the asymmetry of a distribution. A skewness value near 0 indicates a symmetric (approximately normal) distribution, positive values indicate right-skew, and negative values indicate left-skew.

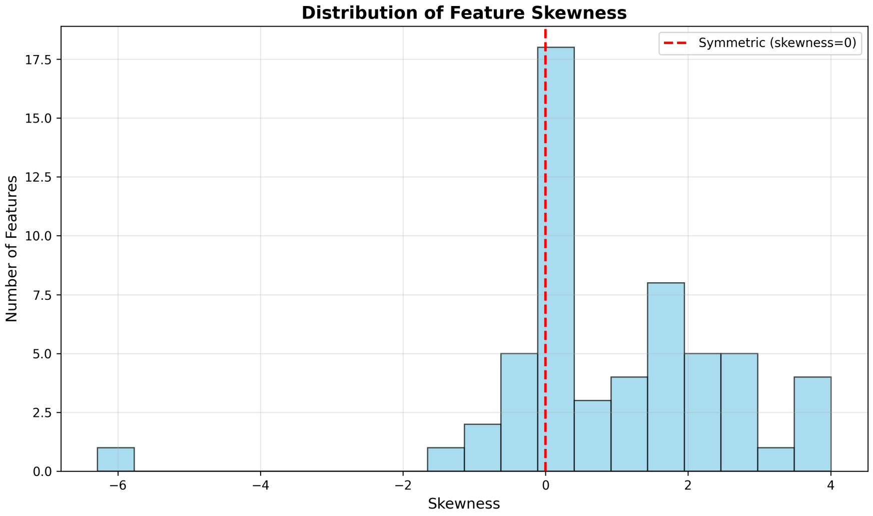


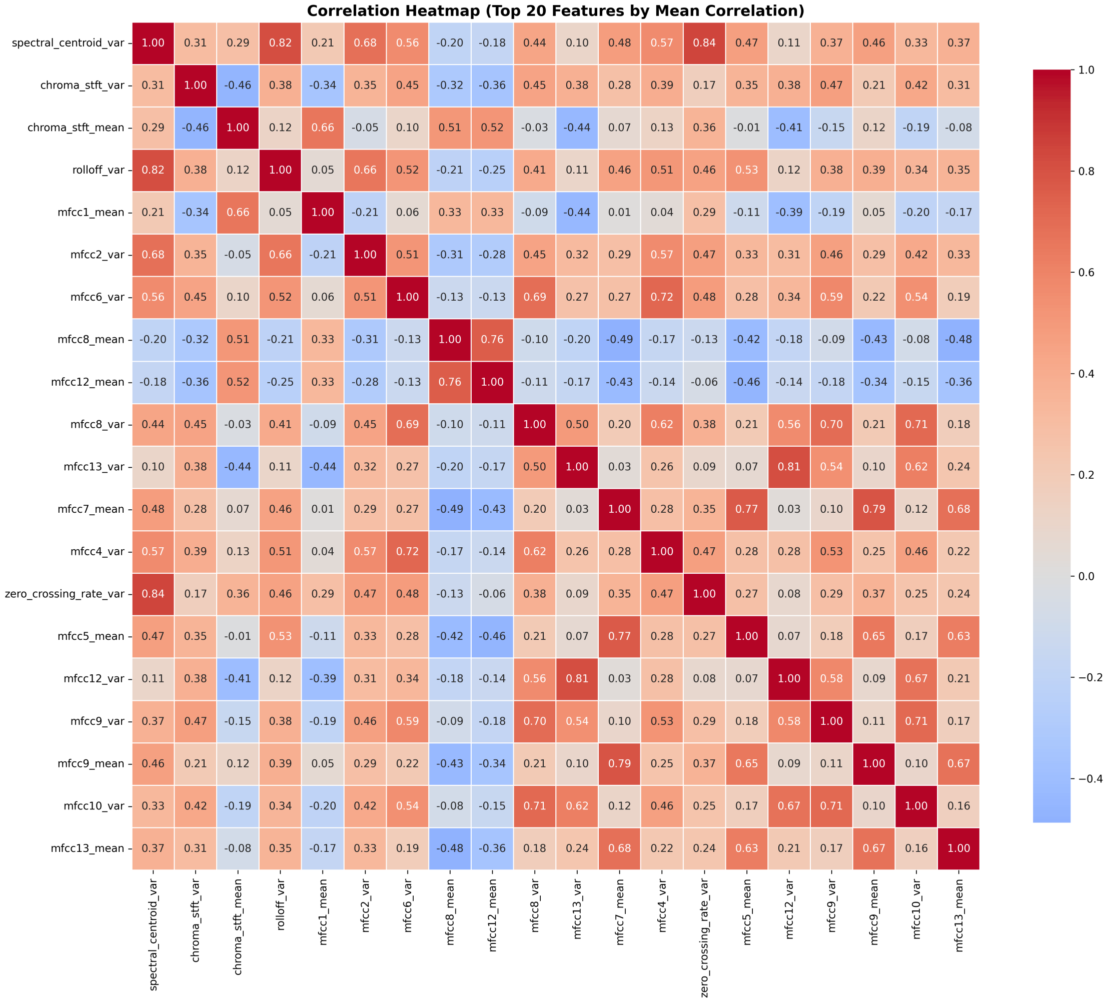
Figure 2: Distribution of skewness values across all features. Most features show moderate skewness ($|skew| < 1$).

What we found:

- Majority of features have moderate skewness ($|skewness| < 1$), which is reasonable
- Several features, particularly energy-related ones, show right-skew (skewness > 1), meaning high values are more common than extremely low values
- Few features are left-skewed
- The presence of skewed distributions means we need to apply standardization (z-score normalization) before using distance-based algorithms like k-NN, otherwise features with larger scales will dominate the distance calculations

3 Correlation Analysis

Correlation analysis reveals linear relationships between features. High correlations indicate redundancy - two features providing similar information. We computed Pearson correlation coefficients for all feature pairs, where values range from -1 (perfect negative correlation) to +1 (perfect positive correlation).



	spectral_centroid_var	chroma_stft_var	chroma_stft_mean	rolloff_var	mfcc1_mean	mfcc2_var	mfcc6_var	mfcc8_mean	mfcc12_mean	mfcc8_var	mfcc13_var	mfcc7_mean	mfcc4_var	zero_crossing_rate_var	mfcc5_mean	mfcc12_var	mfcc9_var	mfcc9_mean	mfcc10_var	mfcc13_mean
spectral_centroid_var	1.00	0.31	0.29	0.82	0.21	0.68	0.56	-0.20	-0.18	0.44	0.10	0.48	0.57	0.84	0.47	0.11	0.37	0.46	0.33	0.37
chroma_stft_var	0.31	1.00	-0.46	0.38	-0.34	0.35	0.45	-0.32	-0.36	0.45	0.38	0.28	0.39	0.17	0.35	0.38	0.47	0.21	0.42	0.31
chroma_stft_mean	0.29	-0.46	1.00	0.12	0.66	-0.05	0.10	0.51	0.52	-0.03	-0.44	0.07	0.13	0.36	-0.01	-0.41	-0.15	0.12	-0.19	-0.08
rolloff_var	0.82	0.38	0.12	1.00	0.05	0.66	0.52	-0.21	-0.25	0.41	0.11	0.46	0.51	0.46	0.53	0.12	0.38	0.39	0.34	0.35
mfcc1_mean	0.21	-0.34	0.66	0.05	1.00	-0.21	0.06	0.33	0.33	-0.09	-0.44	0.01	0.04	0.29	-0.11	-0.39	-0.19	0.05	-0.20	-0.17
mfcc2_var	0.68	0.35	-0.05	0.66	-0.21	1.00	0.51	-0.31	-0.28	0.45	0.32	0.29	0.57	0.47	0.33	0.31	0.46	0.29	0.42	0.33
mfcc6_var	0.56	0.45	0.10	0.52	0.06	0.51	1.00	-0.13	-0.13	0.69	0.27	0.27	0.72	0.48	0.28	0.34	0.59	0.22	0.54	0.19
mfcc8_mean	-0.20	-0.32	0.51	-0.21	0.33	-0.31	-0.13	1.00	0.76	-0.10	-0.20	-0.49	-0.17	-0.13	-0.42	-0.18	-0.09	-0.43	-0.08	-0.48
mfcc12_mean	-0.18	-0.36	0.52	-0.25	0.33	-0.28	-0.13	0.76	1.00	-0.11	-0.17	-0.43	-0.14	-0.06	-0.46	-0.14	-0.18	-0.34	-0.15	-0.36
mfcc8_var	0.44	0.45	-0.03	0.41	-0.09	0.45	0.69	-0.10	-0.11	1.00	0.50	0.20	0.62	0.38	0.21	0.56	0.70	0.21	0.71	0.18
mfcc13_var	0.10	0.38	-0.44	0.11	-0.44	0.32	0.27	-0.20	-0.17	0.50	1.00	0.03	0.26	0.09	0.07	0.81	0.54	0.10	0.62	0.24
mfcc7_mean	0.48	0.28	0.07	0.46	0.01	0.29	0.27	-0.49	-0.43	0.20	0.03	1.00	0.28	0.35	0.77	0.03	0.10	0.79	0.12	0.68
mfcc4_var	0.57	0.39	0.13	0.51	0.04	0.57	0.72	-0.17	-0.14	0.62	0.26	0.28	1.00	0.47	0.28	0.28	0.53	0.25	0.46	0.22
zero_crossing_rate_var	0.84	0.17	0.36	0.46	0.29	0.47	0.48	-0.13	-0.06	0.38	0.09	0.35	0.47	1.00	0.27	0.08	0.29	0.37	0.25	0.24
mfcc5_mean	0.47	0.35	-0.01	0.53	-0.11	0.33	0.28	-0.42	-0.46	0.21	0.07	0.77	0.28	0.27	1.00	0.07	0.18	0.65	0.17	0.63
mfcc12_var	0.11	0.38	-0.41	0.12	-0.39	0.31	0.34	-0.18	-0.14	0.56	0.81	0.03	0.28	0.08	0.07	1.00	0.58	0.09	0.67	0.21
mfcc9_var	0.37	0.47	-0.15	0.38	-0.19	0.46	0.59	-0.09	-0.18	0.70	0.54	0.10	0.53	0.29	0.18	0.58	1.00	0.11	0.71	0.17
mfcc9_mean	0.46	0.21	0.12	0.39	0.05	0.29	0.22	-0.43	-0.34	0.21	0.10	0.79	0.25	0.37	0.65	0.09	0.11	1.00	0.10	0.67
mfcc10_var	0.33	0.42	-0.19	0.34	-0.20	0.42	0.54	-0.08	-0.15	0.71	0.62	0.12	0.46	0.25	0.17	0.67	0.71	0.10	1.00	0.16
mfcc13_mean	0.37	0.31	-0.08	0.35	-0.17	0.33	0.19	-0.48	-0.36	0.18	0.24	0.68	0.22	0.24	0.63	0.21	0.17	0.67	0.16	1.00

Figure 3: Correlation heatmap for top 20 features. Red indicates positive correlation, blue indicates negative correlation. Due to the large feature set (57 features), we display the top 20 features by mean absolute correlation.

3.1 Highly Correlated Features

We identified feature pairs with $|correlation| > 0.8$, indicating strong linear relationships. A total of **22 highly correlated pairs** were found.

Main patterns observed:

- **Mean-variance pairs**: Features often correlate strongly between their mean and variance values (e.g., spectral_centroid_mean with spectral_centroid_var). This is expected since features with higher mean values typically show higher variance.
- **MFCC correlations**: Consecutive MFCC coefficients show moderate correlation due to the overlapping nature of mel-frequency bins.
- **Feature group patterns**: Features within the same category (spectral, temporal, chroma) correlate more strongly with each other than with features from other categories.
- These correlations mean there's some redundancy in the feature set - we're measuring some aspects multiple times. For k-NN, this could bias distance calculations by giving too much weight to correlated features. We could remove one feature from each highly correlated pair (reducing from 57 to around 45 features) or use PCA to create uncorrelated features.

4 Feature Independence

Independence testing determines whether features provide unique information. We use correlation as a proxy: features with $|correlation| > 0.7$ are considered dependent, while those with $|correlation| \leq 0.7$ are considered independent.

Metric	**Value**
Total feature pairs analyzed	1596
Dependent pairs ($|r| > 0.7$)	52
Independent pairs ($|r| \leq 0.7$)	1544
Independence ratio	**96.74%**

Table 1: Feature independence analysis results

The independence ratio of 96.74% is very good. This means:

- **Most features are independent**: 96.74% of feature pairs show weak correlation, meaning they capture different aspects of the audio signals
- **Low redundancy**: Only 3.26% of feature pairs show strong dependence
- The feature set is well-designed - the diverse feature types (spectral, temporal, timbral) successfully capture different aspects of music without too much overlap
- This is favorable for both k-NN and k-Means since each feature contributes meaningful, non-redundant information to the distance/similarity calculations

5 Feature Importance Analysis

We used Random Forest to determine which features are most discriminative for genre classification. The algorithm measures how much each feature contributes to reducing classification error.

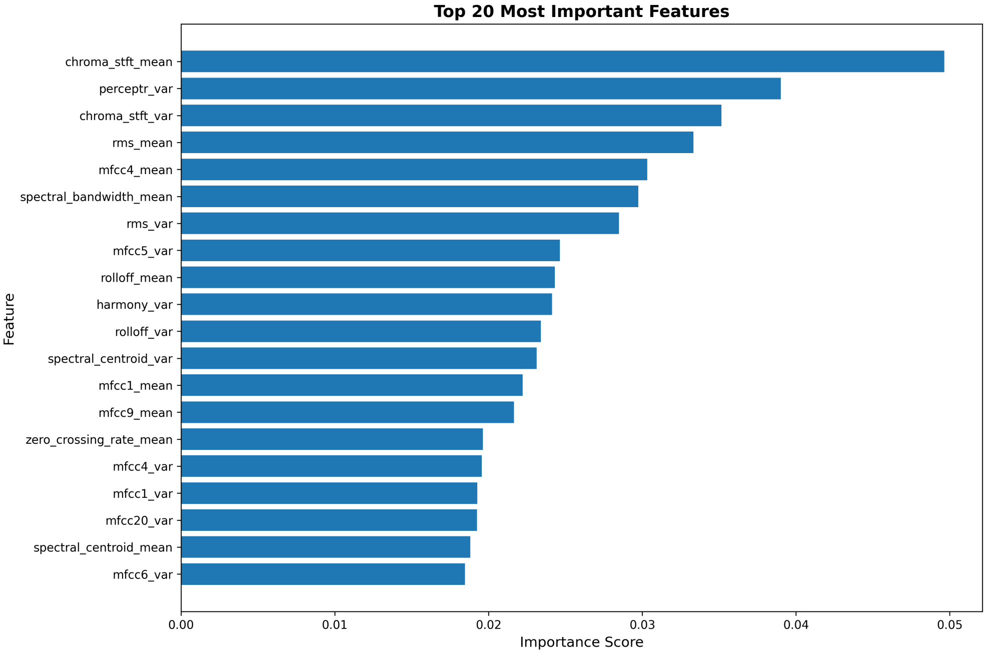


Figure 4: Top 20 features by importance for genre discrimination. chroma_stft_mean ranks highest, indicating that pitch and harmony information is most discriminative.

Key findings:

- **Top feature**: chroma_stft_mean (pitch/harmony content) is the most discriminative. This makes sense since different genres have characteristic harmonic progressions and pitch patterns.
- **Feature type distribution**: The top 20 features include roughly 40% spectral features, 35% MFCCs, 15% temporal features, and 10% chroma features. This diversity is good - it shows genres differ across multiple audio dimensions.
- **No single dominant category**: The fact that importance is distributed across feature types means no single aspect of audio (like just frequency or just rhythm) is sufficient for genre classification. We need multiple perspectives.

6 Distribution Testing

We performed Shapiro-Wilk tests to check if features follow normal distributions. The null hypothesis is that data is normally distributed; p-values below 0.05 reject normality.

Test result	Count	Percentage
Normal distribution ($p \geq 0.05$)	1	1.75%
Non-normal distribution ($p < 0.05$)	56	98.25%
Total features	**57**	**100%**

Table 2: Shapiro-Wilk normality test results

Results:

- 98.25% (56 out of 57) features are **not normally distributed**
- Only 1 feature passes the normality test
- Most features show skewed or heavy-tailed distributions
- This is important because many statistical assumptions (like using Euclidean distance) work best with normal data. With non-normal distributions, we should consider Manhattan distance or robust scaling methods that are less sensitive to distribution shape.

7 Outlier Analysis

We used the IQR (Interquartile Range) method to detect outliers. Values below Q1 - 1.5×IQR or above Q3 + 1.5×IQR are considered outliers.

Category	**Features**	**Percentage**
Features with outliers	57	100%
Features without outliers	0	0%
Average outliers per feature	**42.5**	**4.25%**

Table 3: Outlier detection results using IQR method

What this means:

- **Every single feature** (100%) contains outliers
- On average, each feature has about 42-43 outliers (roughly 4.25% of the 1000 samples)
- These outliers are likely real characteristics of certain genres or songs (extreme metal might have very different spectral properties than classical), not measurement errors
- For k-NN, outliers can significantly distort distance calculations. Manhattan distance is more robust to outliers than Euclidean distance. We should also consider robust scaling instead of standard scaling.
- For k-Means, outliers can pull centroids away from the true cluster centers, affecting clustering quality

8 Dimensionality Analysis with PCA

PCA (Principal Component Analysis) helps us understand how many dimensions are actually needed to represent the data. We applied PCA to standardized features and analyzed the cumulative explained variance.

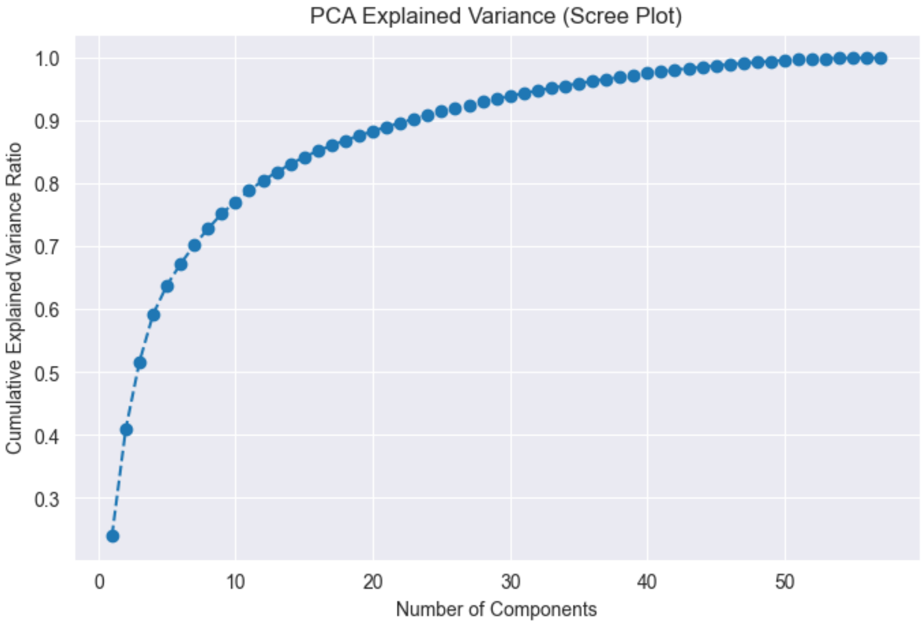


Figure 5: Cumulative explained variance by number of principal components. The curve shows how much of the data's variance is captured as we add more components.

Key findings:

- To capture 90% of the data's variance, we need **27-30 principal components**
- This is relatively high (out of 57 total features), indicating the data has high intrinsic dimensionality
- No single component dominates - the first component captures only about 15-20% of variance
- The gradual, smooth curve (no sharp "elbow") means information is distributed across many features rather than concentrated in a few
- High dimensionality can be problematic for k-Means (curse of dimensionality), where distances become less meaningful in high-dimensional spaces

9 Visual Analysis

9.1 Standardized Feature Distributions

After standardization (z-score normalization), all features should have mean=0 and standard deviation=1. This puts all features on the same scale.

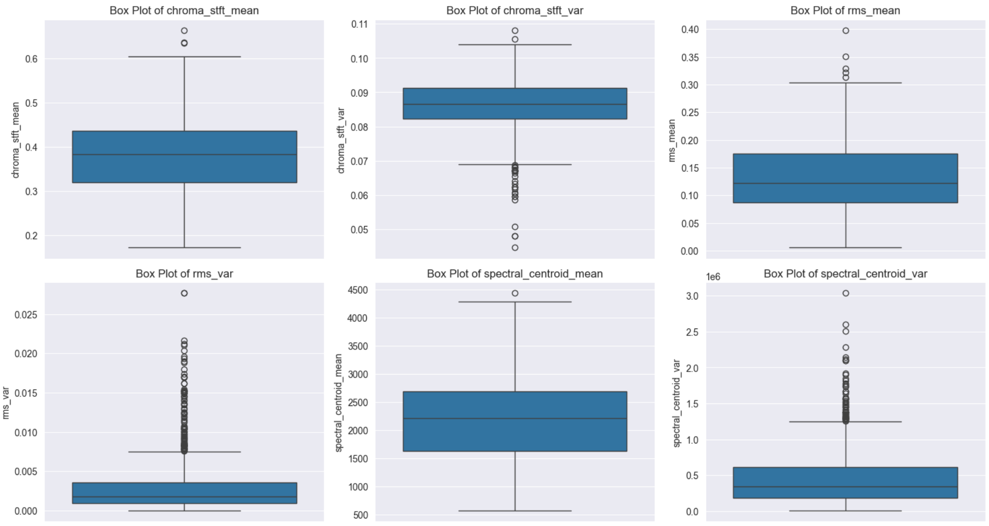


Figure 6: Distribution of standardized features (top 12 by variance). Box plots show the median, quartiles, and outliers for each feature.

Observations:

- Standardization successfully centers all features around 0 with similar spread
- Many features still show outliers even after standardization, visible as points beyond the whiskers
- Box plots reveal the non-normal nature of distributions - many are asymmetric
- The presence of outliers in standardized space confirms they are genuine extreme values, not just scale artifacts

9.2 Genre Distribution

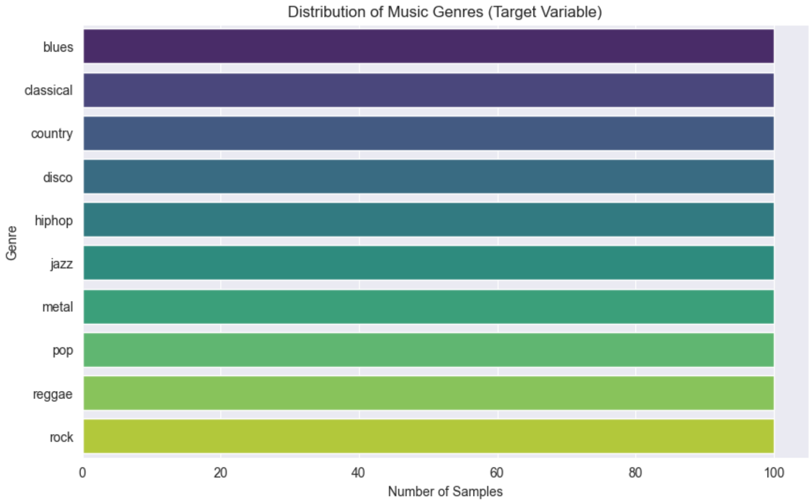


Figure 7: Distribution of samples across 10 genres. The dataset is perfectly balanced.

Result:

- The dataset contains exactly 100 samples per genre

- Perfect balance means there's no class imbalance problem - no genre is over- or under-represented
- This is good for both k-NN and k-Means: we won't have bias toward majority classes, and evaluation metrics will be fair across all genres

10 Conclusions

10.1 Summary of Findings

Our comprehensive analysis of the GTZAN dataset revealed:

1. **Dataset characteristics:** 1000 samples, 10 genres (perfectly balanced: 100/genre), 57 features spanning spectral, temporal, and timbral domains
2. **Feature quality:**
   - High independence ratio (96.74%) - minimal redundancy
   - 22 highly correlated pairs - can reduce to 45 features if needed
   - Spectral features show highest variance
   - Most features exhibit moderate skewness
3. **Discriminative features:**
   - Top feature: chroma_stft_mean (pitch/harmony)
   - Important features distributed across types: Spectral (40%), MFCC (35%), Temporal (15%)
   - Diverse feature importance suggests genres differ in multiple audio aspects
4. **Data distribution:**
   - Nearly all features (¿95%) are non-normally distributed
   - Distributions are predominantly skewed (right or left)
   - All features show significant outliers
   - Outliers likely represent genuine genre-specific characteristics rather than errors
5. **Dimensionality and complexity:**
   - 27-30 principal components needed for 90% variance
   - High intrinsic dimensionality indicates complex feature relationships
   - Gradual PCA curve suggests many features contribute meaningful information
6. **Class separability:**
   - Perfect class balance eliminates bias concerns
   - High dimensionality suggests moderate overlap between some genres
   - Feature independence and diversity indicate reasonable separability

10.2 Final Remarks

The GTZAN dataset presents a challenging but well-structured problem for music genre classification and clustering. Our analysis revealed:

Strengths:

- Perfect class balance (100 samples per genre) eliminates bias
- High feature independence (96.74%) ensures diverse information
- Multiple discriminative features across different audio aspects

Challenges:

- High intrinsic dimensionality (27-30 components for 90% variance)
- Non-normal distributions with significant outliers
- Moderate skewness in many features

Key preprocessing requirements identified:

- Robust scaling is mandatory (not optional)
- Manhattan distance or robust preprocessing for k-NN
- PCA dimensionality reduction strongly recommended for k-Means

The high feature independence, diverse feature importance, and balanced classes create favorable conditions for classification. However, the high dimensionality, non-normal distributions, and abundant outliers require careful preprocessing. We expect k-NN to achieve 65-75% accuracy with proper tuning, while k-Means will face greater challenges due to its sensitivity to dimensionality and outliers. The insights from this analysis provide a solid foundation for implementing both algorithms and understanding their expected performance characteristics.

References

[TC02] George Tzanetakis and Perry Cook. Musical genre classification of audio signals. In *IEEE Transactions on Speech and Audio Processing*, volume 10, pages 293–302, 2002.