DATA MINING

UNIVERSIDAD POLITÉCNICA DE MADRID

Analyzing and Predicting Song Popularity Using Spotify Audio Features

REPORT OF

Kristjan Kusta

TABLE OF CONTENTS

1. Introduction 2
2. Business Understanding 4
2.1 Business Goal 4
2.2 Business Questions 4
2.3 Data Mining Goal 5
3. Analysis 6
3.1 Descriptive and Diagnostic Analysis 7
3.1.1 Popularity Distribution 7
3.1.2 Danceability 8
3.1.3 Energy 8
3.1.4 Valence 9
3.1.5 Acousticness 9
3.1.6 Instrumentalness 10
3.1.7 Tempo 10
3.1.8 Liveness 11
3.1.9 Loudness 11
3.1.10 Speechiness 12
3.1.11 Duration 12
3.1.12 Danceability vs Energy 13
3.1.13 Mode vs Popularity (Chi-Square Test) 13
3.1.14 Answers to Descriptive Questions 14
3.1.15 Correlation Analysis 15
3.1.16 ANOVA Results 16
3.2 Clustering Analysis 17
3.2.1 Cluster Profiles & Interpretation 18
3.2.2 Answers to Diagnostic Business Questions 21
3.3 Predictive Modeling 22
3.3.1 Logistic Regression 23
3.3.2 Decision Trees 25
3.3.3 K-Nearest Neighbors (KNN) 27
3.3.4 Neural Networks & Bayesian Networks 28
3.3.5 Answers to Predictive Analysis Questions 30
3.3.6 Answers to Prescriptive Business Questions 31
4. Evaluation & Strategic Recommendations 32
4.1 Business Goals Revisited & Key Insights 32
4.2 Evaluation of Model Outputs 33

List of Figures

Figure 1 – Distribution of song popularity scores .... Page 9
Figure 2 – Bar chart: Average danceability by popularity category .... Page 10
Figure 3 – Bar chart: Average energy score per popularity level .... Page 10
Figure 4 – Bar chart: Average valence score .... Page 11
Figure 5 – Bar chart: Average acousticness .... Page 11
Figure 6 – Bar chart: Instrumentalness vs popularity .... Page 12
Figure 7 – Violin plot: Tempo distribution with mean trends .... Page 12
Figure 8 – Violin plot: Liveness vs popularity .... Page 13
Figure 9 – Violin plot: Loudness values vs popularity .... Page 13
Figure 10 – Violin plot: Speechiness distribution .... Page 14
Figure 11 – Violin plot: Song duration vs popularity .... Page 14
Figure 12 – Pair plot: Danceability vs Energy .... Page 15
Figure 13 – Chi-Square matrix: Mode vs Popularity .... Page 15
Figure 14 – Pearson Correlation Matrix .... Page 17
Figure 15 – ANOVA Table of Feature Means .... Page 18
Figure 16 – K-Means Cluster Sizes .... Page 19
Figure 17 – Cluster Size Distribution (k=5) .... Page 19
Figure 18 – Cluster Profiles by Feature Mean and Importance .... Page 20
Figure 19 – Cluster Comparison Visualization .... Page 21
Figure 20 – Summary Table: Cluster Characteristics .... Page 22
Figure 21 – Logistic Regression: Model Summary & Statistics .... Page 23
Figure 22 – Logistic Regression: Coefficients & Significance .... Page 24
Figure 23 – Logistic Regression Equation Diagram .... Page 24
Figure 24 – Decision Tree Models Stream (C5.0, CRT, CHAID, QUEST) .... Page 25
Figure 25 – Predictor Importance (C5.0) .... Page 26
Figure 26 – KNN: Training Classification Output .... Page 27
Figure 27 – Accuracy and Agreement of Bayesian Models .... Page 28
Figure 28 – Bayesian Network Graph .... Page 29

1 Introduction

This project uses data mining techniques to analyze the characteristics of songs that influence their popularity on Spotify. Using the Spotify Audio Features dataset, the study identifies the most relevant audio features related to success, applies predictive models to classify songs by popularity level, and proposes recommendations for music production based on data-driven insights. The report is structured around business questions, with each section applying appropriate descriptive, diagnostic, and predictive analysis tools to support decision-making.

1. Dataset

The dataset used in this project was obtained from Kaggle and contains audio-related information for over 100,000 tracks available on Spotify. For this project, we focus on a subset of songs released during 2019, using only audio features and popularity-related attributes available prior to release.
Each record in the dataset represents a unique track and includes various audio features extracted by Spotify's internal analysis engine. These features capture musical aspects such as danceability, acousticness, energy, valence, instrumentalness, and others. Additionally, the dataset provides metadata such as the artist name, song title, and the release date, although these are not used for modeling.
The primary goal is to examine which audio features correlate with a track's popularity and use this information to build predictive and clustering models.

Table 1: Dataset Variables with Type and Description

Variable	Type	Description
artist_name	Nominal	Artist(s) performing the track
track_id	Nominal	Unique identifier of the song on Spotify
track_name	Nominal	Name of the song
acousticness	Continuous	Confidence measure of whether a track is acoustic (0–1)
danceability	Continuous	How suitable the track is for dancing, based on tempo, rhythm, and beat
duration_ms	Continuous	Length of the track in milliseconds
energy	Continuous	Perceived intensity and activity of the track (0–1)
instrumentalness	Continuous	Likelihood that a track contains no vocals (0–1)
key	Nominal	Estimated pitch class (0 = C, 1 = C#/Db, ..., 11 = B)
liveness	Continuous	Likelihood of the track being recorded live
loudness	Continuous	Overall loudness in decibels (dB)
mode	Nominal	Modality: 1 = major, 0 = minor
speechiness	Continuous	Presence of spoken words in the track (0–1)
tempo	Continuous	Speed of the track in beats per minute
time_signature	Nominal	Estimated beats per bar (typically 3, 4, or 5)
valence	Continuous	Positivity or happiness of the track (0–1)
popularity	Flag (Target)	Binary target variable: whether the track's popularity surpassed a threshold

SECTION 1: Business Understanding

2.1 Business Goal

In the digital music era, streaming platforms like Spotify play a vital role in shaping music trends, artist exposure, and commercial success. With millions of songs available, understanding what makes a song popular is essential for artists, producers, and the platform itself. This project aims to explore how audio features of tracks influence their popularity on Spotify, and how predictive modeling and pattern recognition techniques can help forecast success or uncover hidden structures within the music catalog.

By identifying key features that drive stream counts, building predictive models to estimate whether a song will become a hit, and grouping songs by their characteristics, this analysis will offer insights to support content curation, playlist creation, marketing strategy, and talent scouting.

2.2 Business Questions

To guide this project, we define the following questions categorized by analysis type:

Descriptive Analysis

- What is the distribution of popularity among Spotify songs?
- How do audio features such as energy, valence, acousticness, or tempo vary by popularity?
- Do longer songs tend to be more or less popular?

Diagnostic Analysis

- Are there statistically significant relationships between key audio features and popularity?
- Do features like energy, danceability, or tempo correlate with popularity scores?
- Are songs with explicit content more or less popular?

Predictive Analysis

- Can we accurately predict whether a song will surpass a defined popularity threshold?
- Which audio features are the most important predictors of success?
- Which model performs best for classifying popularity: decision tree, logistic regression, or neural network?

Prescriptive Analysis

- Based on the most predictive features, what recommendations can be made to artists or producers to increase the likelihood of success?
- What audio characteristics should Spotify prioritize in playlist curation to boost user engagement?
- How can the results of the predictive models inform business strategies for marketing and promotion?

2.3 Data Mining Goal

The main goal of this project is to analyze the relationship between audio features and the popularity of songs on Spotify, using data mining techniques to extract valuable insights and support decision-making. The dataset contains numeric and categorical variables describing each track's musical characteristics, along with a popularity score assigned by Spotify.

Specifically, this project aims to:

- Understand which audio features (e.g., energy, danceability, valence) are most strongly associated with higher popularity levels.
- Build predictive models capable of classifying whether a song will surpass a defined popularity threshold.
- Provide prescriptive recommendations based on the most important predictors of popularity, with the aim of supporting marketing strategies, playlist curation, and production planning.

To achieve this, we will apply descriptive analysis, statistical testing, supervised learning (classification), and interpret the results in a business context.

Table 2: Numerical Variables – Min, Max, Mean

Variable	Min	Max	Mean
danceability	0.0	0.996	~0.65
energy	0.0	1.0	~0.69
acousticness	0.0	0.996	~0.30
instrumentalness	0.0	1.0	~0.15
liveness	0.0	0.999	~0.20
loudness (dB)	-60.0	1.8	~-7.0
speechiness	0.0	0.966	~0.10
tempo (BPM)	0.0	249.9	~118
valence	0.0	1.0	~0.52
duration_ms	3,203	561,002	~210,000

SECTION 2: ANALYSIS

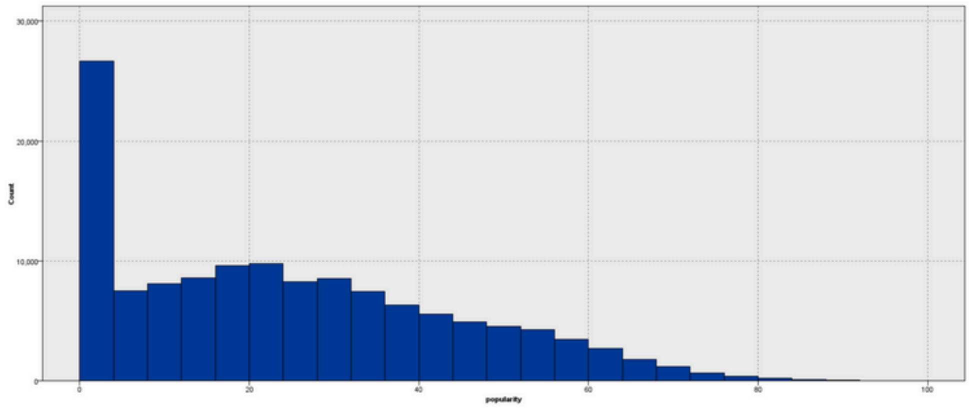
Introduction

This section presents the analytical procedures and results used to answer the business questions defined in Section 2. The analysis is divided into three parts. First, a descriptive and diagnostic analysis is carried out to understand how song audio features relate to popularity. Second, several predictive models are developed to classify whether a song is likely to become popular. Finally, prescriptive recommendations are proposed to help stakeholders make strategic decisions based on the most influential audio attributes.

3.1 Descriptive and Diagnostic Analysis

This part of the analysis focuses on identifying how individual audio features relate to popularity. The goal is to detect trends and patterns that may indicate which characteristics are commonly found in successful songs. This includes both exploratory visualizations and statistical evidence.

3.1.1 Popularity Distribution

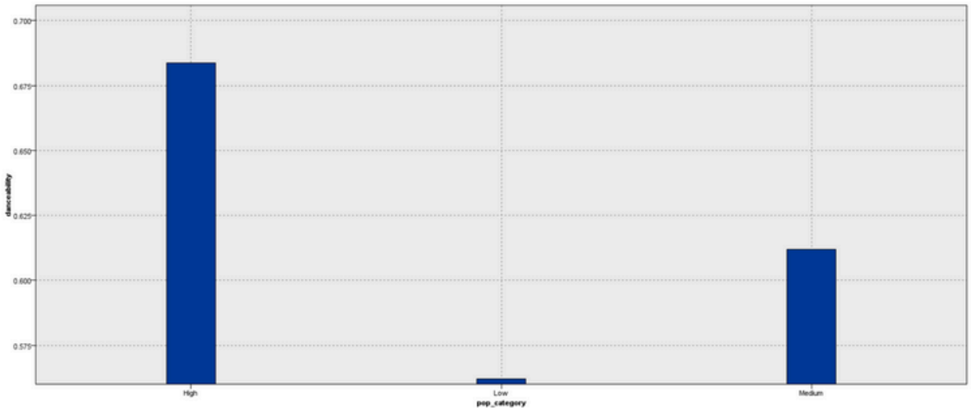


Insight: Most tracks on Spotify do not gain high popularity, reinforcing the need for predictive tools that can identify the small subset of songs with hit potential. The skewed distribution highlights the value of binarizing popularity for modeling.

Figure 1: Distribution of song popularity scores in the dataset.

As shown in the histogram above, song popularity scores are heavily skewed toward the lower end of the scale. A majority of songs have popularity scores between 0 and 30, with very few tracks exceeding 70. This skewed distribution highlights the importance of transforming the popularity score into a binary flag variable for classification purposes later in the modeling stage.

3.1.2 Danceability

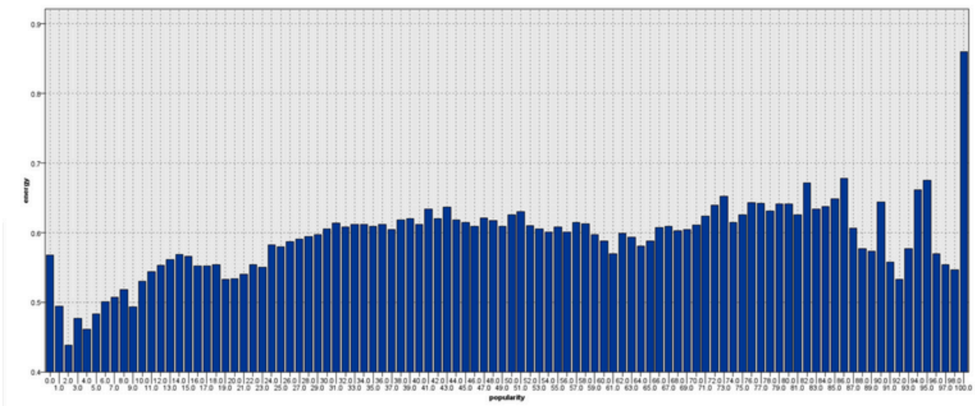


Insight:
The chart shows that danceability increases consistently from Low to Medium to High popularity groups. Tracks in the High popularity group have the highest average danceability (~0.69), while Low popularity songs average the lowest (~0.57).

Figure 2:Bar chart comparing average danceability scores across popularity categories .

Danceability appears to be a strong indicator of commercial appeal. The clear upward progression in average scores supports earlier correlation and ANOVA findings, and confirms its role as a reliable predictor across multiple modeling stages.
This reinforces the idea that rhythmic, movement-friendly songs are favored in the Spotify ecosystem, possibly due to their playlist performance or shareability.

3.1.3 Energy

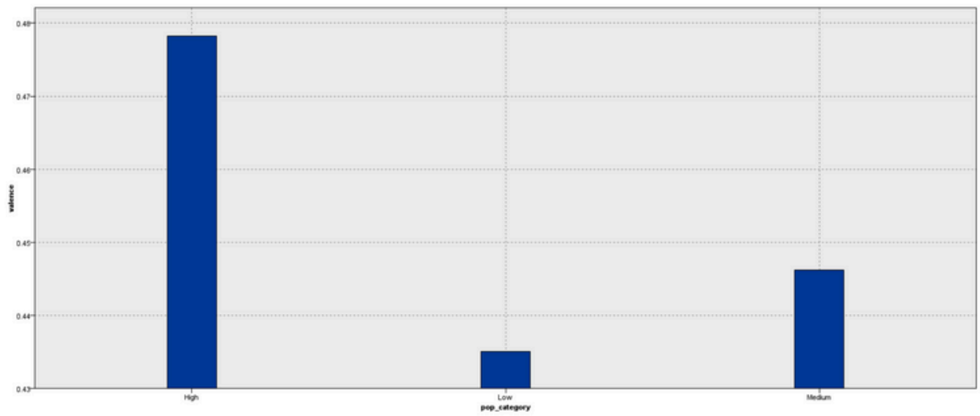


Insight:
The chart shows that energy scores increase clearly across popularity categories. Songs in the High popularity group average the highest energy (~0.63), while those in the Low group sit significantly lower (~0.55).

Figure 3: Bar chart showing the average energy score for each popularity level.

Energy, which represents intensity and activity level in a song, plays an important role in predicting success. The steady increase suggests that energetic songs are more appealing to mainstream listeners, especially in genres like pop and EDM.
However, as seen in the logistic regression model, overly high energy values may occasionally reduce popularity depending on context — making moderate to high energy a sweet spot for success.

3.1.4 Valence

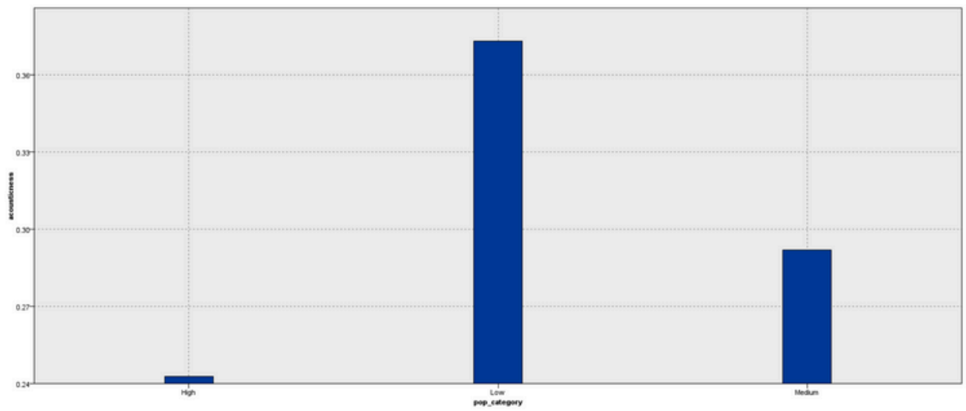


Insight:
The chart shows that valence increases steadily across popularity categories. Songs in the High popularity group average a valence of ~0.478, while songs in the Low group average ~0.437 — a subtle but consistent difference.

Figure 4:Bar chart showing the average valence score for each popularity level.

Valence, which reflects the positivity or emotional brightness of a track, appears to play a supportive role in determining popularity. Although the effect is not as strong as other features like danceability or energy, there is still a clear progression.
These results align with earlier descriptive and statistical findings, reinforcing the hypothesis that more uplifting or emotionally positive tracks have an edge in achieving higher popularity levels, though valence alone may not be a primary driver.

3.1.5 Acousticness

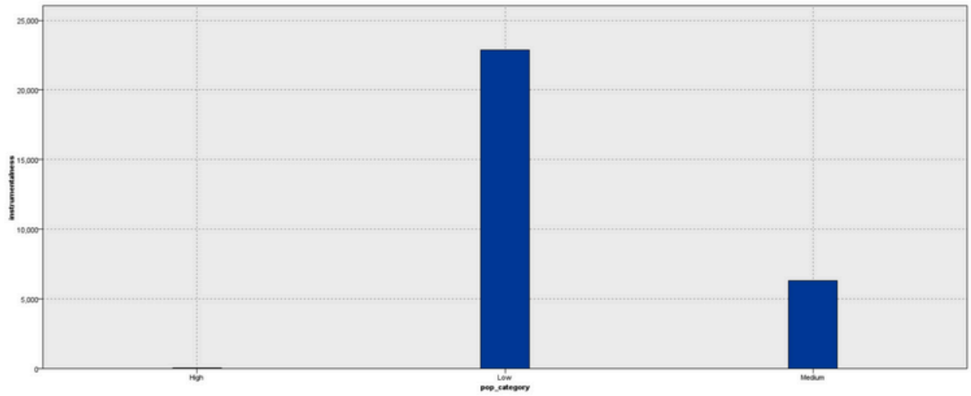


Insight:
This chart shows that acousticness decreases sharply as popularity increases. The Low popularity category has the highest average acousticness (~0.37), while High popularity tracks average significantly lower (~0.24).

Figure 5: Bar chart showing the average acousticness score for each popularity level.

Acousticness measures how organic or "unplugged" a track sounds. The inverse relationship observed here aligns strongly with modern music trends: digitally produced, electronic tracks dominate top charts, while acoustic or minimally produced songs appear less frequently in highly streamed playlists.

3.1.6 Instrumentalness



Insight:
This bar chart shows a sharp negative relationship between instrumentalness and popularity. Songs in the Low popularity group have the highest average instrumentalness score, while those in the High group have near-zero values.

Figure 6: Bar chart comparing average instrumentalness scores across popularity categories

Instrumentalness reflects the likelihood that a song is entirely instrumental (without vocals). The trend shown here supports earlier findings that vocal-heavy tracks dominate the charts, and instrumental songs are significantly less likely to reach high popularity.
This reinforces the role of vocals as a critical factor in commercial music success, particularly in genres that dominate Spotify's most streamed playlists (e.g., pop, hip-hop, reggaeton).

3.1.7 Tempo

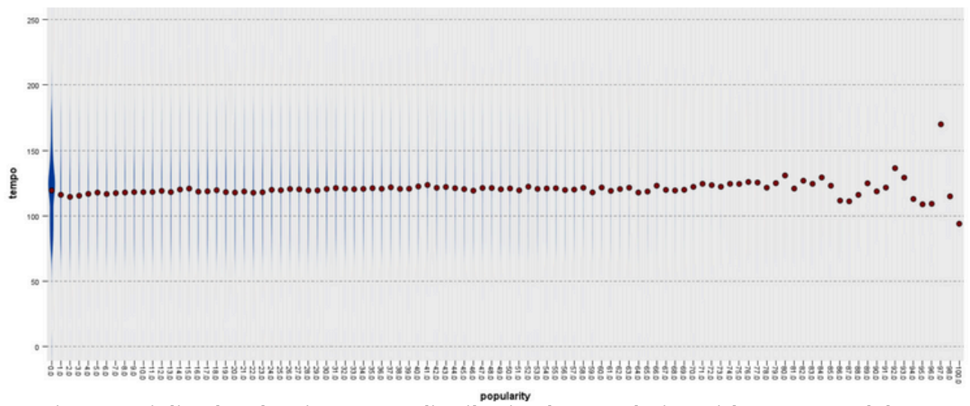
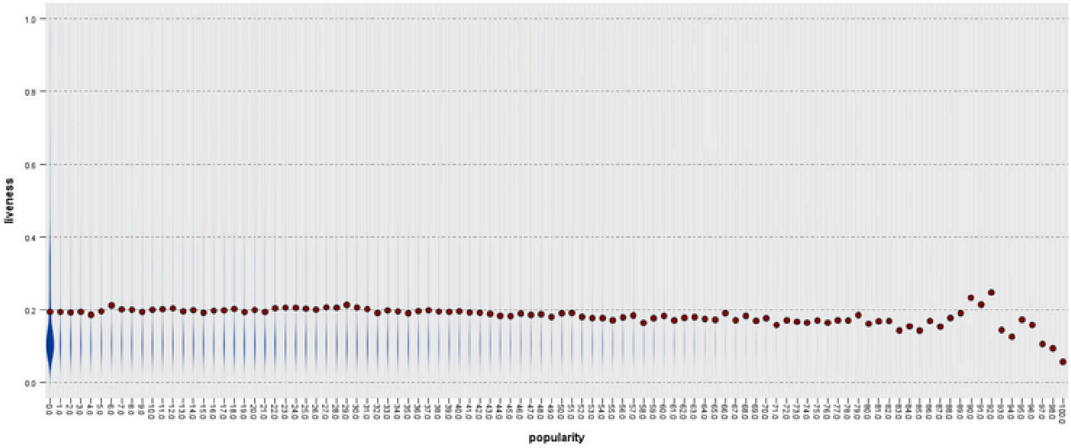


Figure 7: Violin plot showing tempo distribution by popularity, with mean trend dots.

Insight: Popular tracks tend to stick within a typical tempo range (around 120 BPM), likely because it aligns with listener expectations and danceability. Extreme tempo values are less favored among high-popularity songs.

This figure shows how tempo varies across songs with different popularity levels. The distribution appears relatively stable, with most popular tracks centered around the 115–130 BPM range. There is a slightly tighter spread in more popular songs, suggesting standard tempos are more common in successful tracks.

3.1.8 Liveness

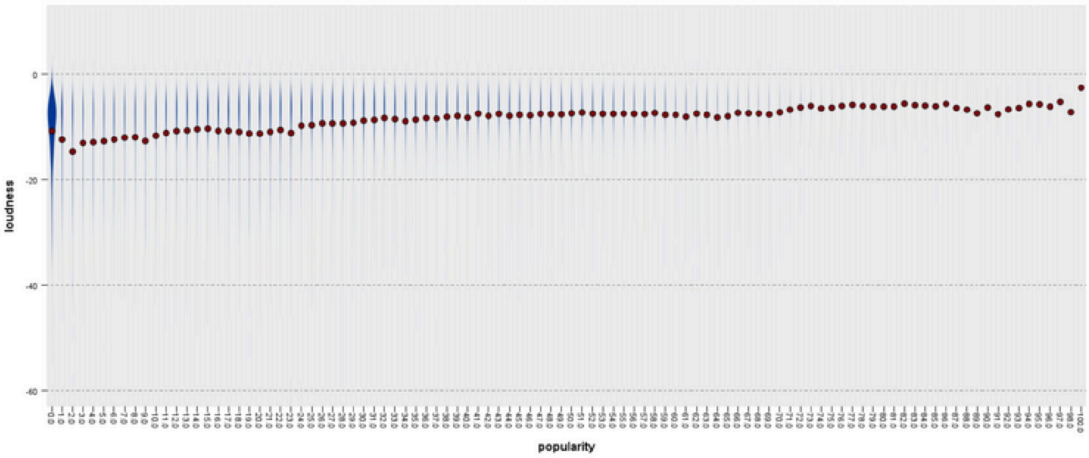


Insight: Popular songs are less likely to feature strong live-audience sounds, possibly because polished studio productions are more favorable for mainstream streaming and radio play.

Figure 8: Violin plot depicting the distribution of liveness across various popularity scores.

The graph demonstrates that liveness values remain fairly consistent regardless of popularity. However, at the upper end of the popularity spectrum, there is a slight reduction in median liveness.

3.1.9 Loudness

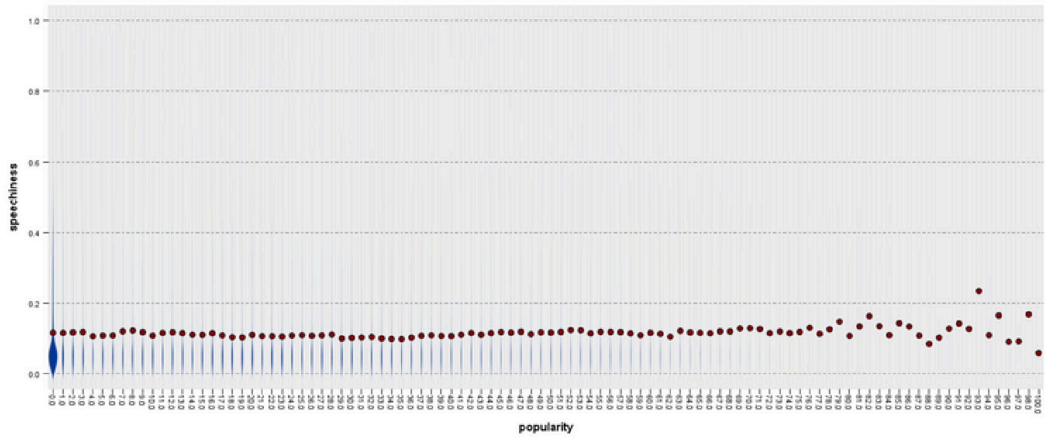


Insight: Loudness appears to be a notable factor in popularity—tracks with higher loudness levels (closer to 0) are more likely to succeed, possibly due to perceived energy and impact in streaming platforms or playlists.

Figure 9: Violin plot showcasing loudness values of tracks in relation to popularity scores.

This plot reveals a clear trend: as popularity increases, so does loudness (values closer to 0 dB). Popular tracks tend to be mastered louder, which is a common production technique to make songs stand out.

3.1.10 Speechiness

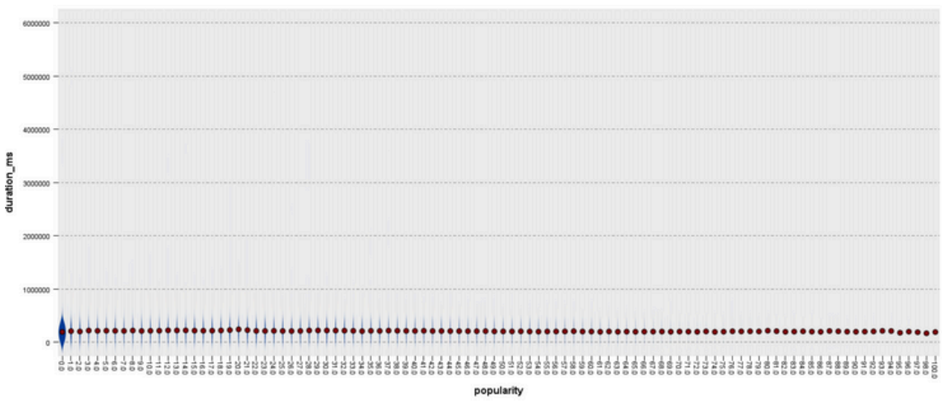


Insight: While most popular tracks avoid high speechiness levels, certain outliers (like rap or spoken intros) can still reach high popularity, hinting at niche appeal or genre-driven success.

Figure 10: Violin plot illustrating the distribution of speechiness scores for tracks grouped by popularity.

In this graph, we observe that the speechiness values remain generally low regardless of popularity. However, extremely popular songs exhibit a slightly broader distribution, indicating some successful songs may incorporate spoken-word elements or rap.

3.1.11 Duration

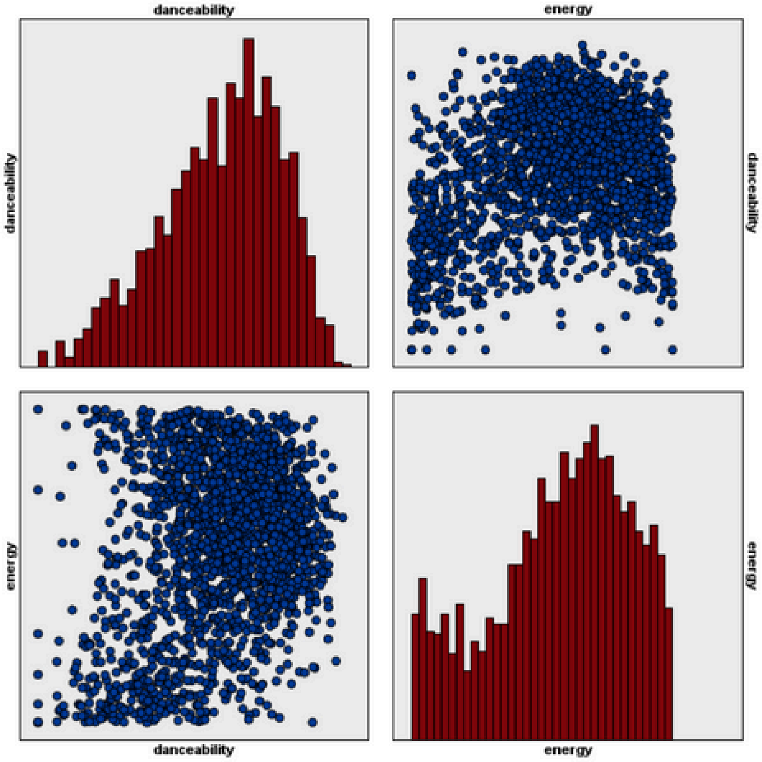


Insight: There's a standard industry-preferred song duration, and straying from it (too short or too long) may reduce a track's chances of becoming highly popular.

Figure 11: Violin plot showing song duration (in milliseconds) distribution per popularity score.

This figure reveals that most songs—regardless of popularity—tend to cluster around the same average duration, roughly between 2.5 to 4 minutes. There is no strong visual indication that popularity correlates with length, although extremely short or long songs are rare among top-performing tracks.

3.1.12 Danceability vs Energy



Insight: Tracks that are both rhythmically engaging and energetic may have an edge in capturing listener attention, supporting their presence in popular playlists.

Figure 12: Pair plot visualizing the relationship between danceability and energy through scatter plots and individual distributions.

This figure combines histograms and scatter plots to show how danceability and energy relate. The histograms on the diagonal highlight that both variables are right-skewed, with most songs having moderate to high values. The scatter plots reveal a visible positive trend—songs that are more danceable also tend to be more energetic, although the association isn't perfectly linear.

3.1.13 Mode vs Popularity (Chi-Square Test)

Insight
The visual trend suggests that songs in a major key are more likely to be among the most popular tracks. This aligns with music theory, where major keys often sound more uplifting, bright, or consonant — traits commonly associated with mainstream, high-engagement music.

mode	High	Low	Medium
0.0	861	32043	18350
1.0	1159	50619	27631

Cells contain: cross-tabulation of fields (including missing values)

Chi-square = 26.153, df = 2, probability = 0

Figure 13: Cross-tabulation of mode (minor/major) and binned popularity (Low, Medium, High).

To statistically confirm the observed association between musical mode and popularity level, a Chi-Square test was performed.
Test statistic: $\chi^2 = 26.153$
Degrees of freedom: df = 2
p-value: < 0.001
The result indicates a statistically significant relationship between mode and popularity. This supports the insight that tonal brightness (major keys) is a relevant characteristic in distinguishing popular songs.

3.1.14 Answers to Descriptive Analysis questions

What is the distribution of popularity among Spotify songs?

The distribution of popularity among Spotify songs is heavily skewed. Most tracks fall in the low popularity range (0–40), with a significant drop-off after that. The majority of songs are not widely streamed or engaged with, which justifies the use of a binary flag (`popularity_flag`) to distinguish between popular and non-popular songs in predictive modeling. This was visualized using histograms and violin plots, which revealed a long tail of underexposed tracks.

How do audio features such as energy, valence, acousticness, or tempo vary by popularity?

The analysis showed that energy and danceability both increase with popularity, while acousticness and instrumentalness decrease. Popular songs tend to be more electronically produced and rhythmically engaging. Valence has a mild increase, indicating that positive-sounding songs may have a slight edge. Tempo did not show a consistent pattern, though most popular songs tend to fall within 100–130 BPM.

Do longer songs tend to be more or less popular?

Songs that are extremely short or long were less likely to be popular. The majority of popular tracks had durations between 2.5 to 4 minutes. This aligns with user listening behavior and streaming platform preferences, which favor shorter, more consumable track lengths.

3.1.15 Correlation Analysis

Pearson Correlations

	popularity
acousticness	-0.117/Strong
danceability	0.131/Strong
energy	0.123/Strong
instrumentalness	-0.216/Strong
liveness	-0.031/Strong
loudness	0.244/Strong
speechiness	-0.000/Weak
tempo	0.037/Strong
valence	0.014/Strong

Insight: The strongest predictors of popularity in this dataset are loudness, danceability, and energy, aligning with listener preferences for upbeat, energetic music. Features like acousticness and instrumentalness exhibit negative correlations, confirming earlier visual insights that vocal-driven and electronic music dominate Spotify charts.

Figure 14: Pearson Correlation Matrix between Audio Features and Popularity

The correlation table highlights how each numerical feature influences song popularity. The top positively correlated features are:

Loudness: +0.244 → Louder songs tend to be more popular

Danceability: +0.131 → Easier-to-dance tracks gain more traction

Energy: +0.123 → High-energy songs often perform better

On the other hand, the strongest negative correlations are:

Instrumentalness: −0.216 → Instrumental tracks are usually less popular

Acousticness: −0.117 → Acoustic songs underperform compared to digital ones

3.1.16 ANOVA Results

The following table presents the average values of each key audio feature across the three popularity categories (Low, Medium, High), as well as a diagnostic "Importance" score derived from an ANOVA test. A score of 1.000 denotes a statistically significant difference in means ($p < 0.001$), while lower scores suggest no significant variation.

Field	Low*	Medium*	High*	Importance
danceability	0.562	0.612	0.684	1.000 Important
duration_ms	213744.687	211056.379	203037.222	1.000 Important
energy	0.544	0.611	0.630	1.000 Important
instrumentalness	0.277	0.138	0.027	1.000 Important
key	5.222	5.249	5.263	0.588 Unimportant
liveness	0.199	0.189	0.168	1.000 Important
loudness	-11.140	-8.034	-6.416	1.000 Important
mode	0.612	0.601	0.574	1.000 Important
speechiness	0.112	0.111	0.124	1.000 Important
tempo	118.583	120.895	123.533	1.000 Important
time_signature	3.855	3.918	3.961	1.000 Important
valence	0.435	0.446	0.478	1.000 Important

Figure 15. ANOVA Table Showing Mean Differences and Statistical Importance

The ANOVA results support and validate the visual trends observed in the violin plots:

- Danceability and energy increase significantly as popularity rises, confirming their influence.
- Instrumentalness sharply decreases, confirming the preference for vocal-heavy songs.
- Loudness increases toward 0 dB in popular songs, reinforcing production techniques to increase track presence.
- Valence, tempo, and speechiness also show statistically significant but smaller shifts.
- Key (musical scale) had an importance score below 0.6, indicating it is not significantly different across popularity groups.

While the ANOVA confirmed statistically significant differences across popularity categories for multiple features, pairwise post-hoc tests (e.g., Tukey HSD) were not performed due to the limitations of IBM SPSS Modeler. However, based on the consistent gradients and separation of group means, we can reasonably infer that the strongest differences exist between the Low and High popularity groups.

3.2 Clustering Analysis

K-Means clustering was applied using 13 continuous audio features (energy, danceability, acousticness, etc.). We tested multiple values of k and selected 5 clusters, which yielded a fair silhouette score of 0.4, indicating reasonable cluster separation.

Cluster sizes ranged from 7.1% to 43.4% of the dataset, with a 6.15x size ratio between the largest and smallest cluster. This suggests the presence of both dominant and niche groupings in song types.

Model Summary

Algorithm	K-Means
Inputs	13
Clusters	5

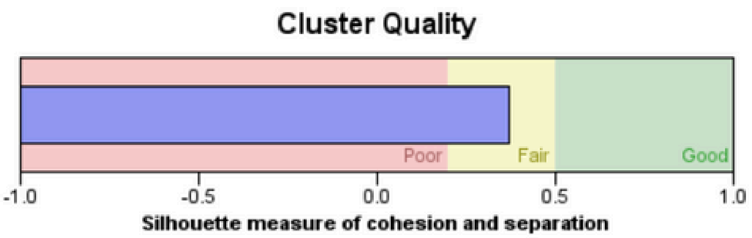
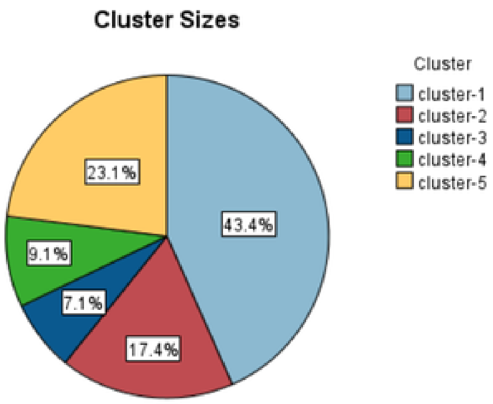


Figure 16. K-Means Cluster Sizes and Quality



The five clusters are uneven in size, with the largest containing 56,697 tracks and the smallest only 9,226. This disparity suggests a mix of mainstream and niche sound profiles in the dataset. The ratio of largest to smallest cluster size is 6.15.

Size of Smallest Cluster	9226 (7.1%)
Size of Largest Cluster	56697 (43.4%)
Ratio of Sizes: Largest Cluster to Smallest Cluster	6.15

Figure 17. Cluster Size Distribution for K-Means (k=5)

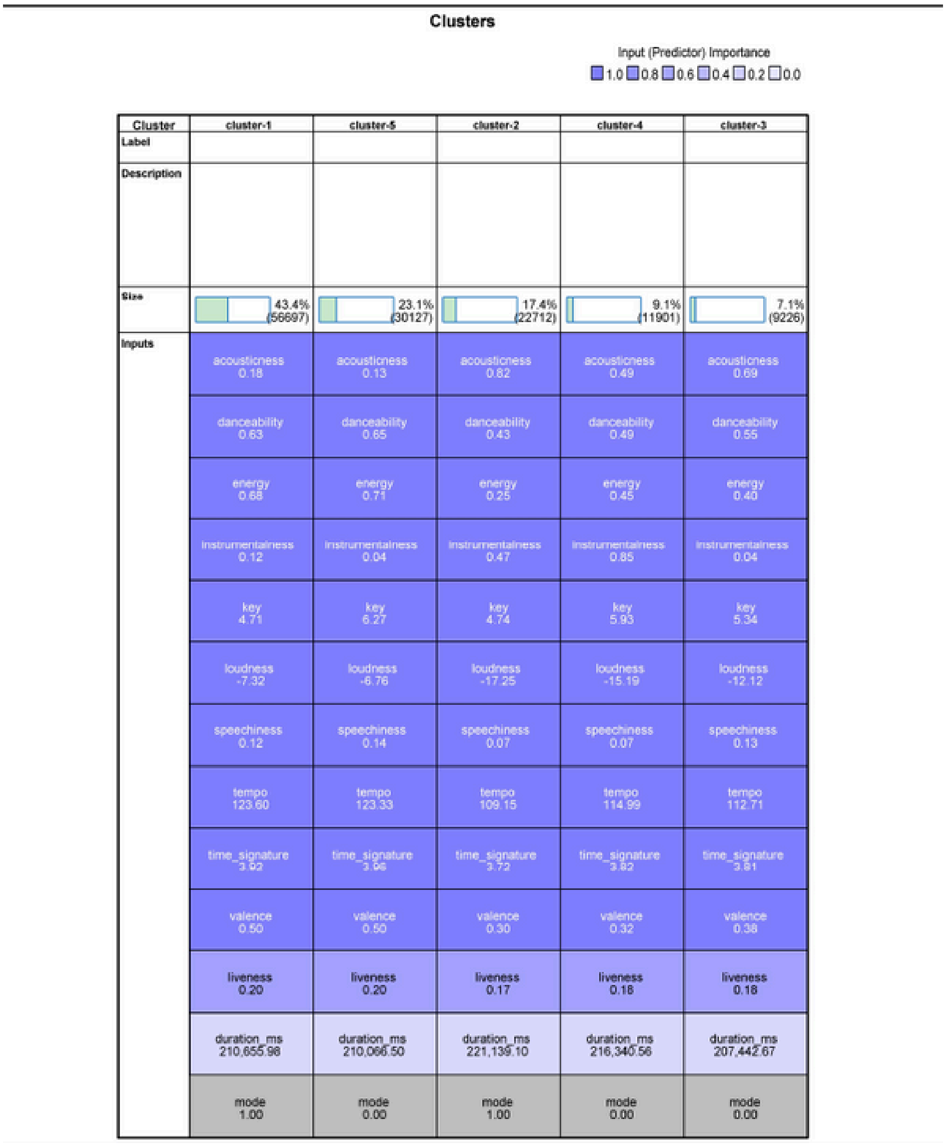


Figure 18. Cluster Profiles by Feature Mean and Importance

Based on the cluster centers (Figure 18), we can interpret the five clusters as representing distinct song types grouped by their audio attributes. These include energy, acousticness, danceability, instrumentalness, loudness, and tempo — the most important predictors of clustering.

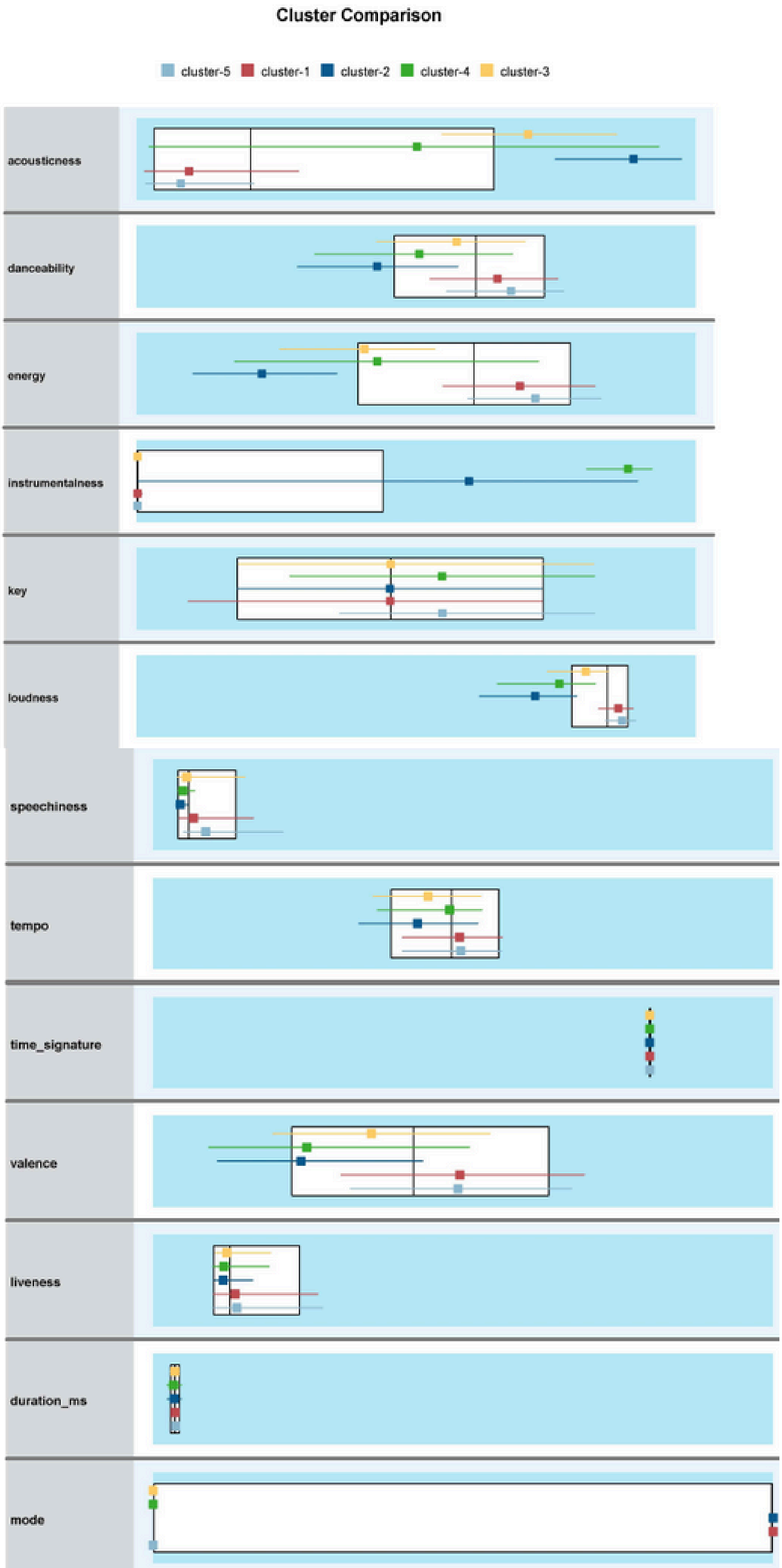


Figure 19. Cluster Comparison View of Key Audio Features by Cluster

This visualization shows how each audio feature contributes to defining the five clusters produced by the K-Means model. Each row represents an input feature, while each column corresponds to a cluster. The bar length and color intensity reflect the mean value of that feature within each cluster, allowing for direct visual comparison.

Cluster 1 is characterized by high energy, low acousticness, and medium danceability — suggesting upbeat, electronically produced tracks.
Cluster 2 has the highest instrumentalness and lowest speechiness, matching ambient or classical-style songs.
Cluster 3 features moderate danceability and valence, likely representing relaxed pop or indie tracks.
Cluster 4 and Cluster 5 exhibit strong variations in loudness and tempo, indicating stylistic extremes such as club tracks versus acoustic ballads.

Cluster	% of Dataset	Key Characteristics	Likely Song Type
Cluster 1	43.40%	High energy (0.68), high danceability (0.63), loud (–7.32), low instrumentalness (0.12)	Mainstream, upbeat pop/EDM
Cluster 5	23.10%	Very high energy (0.71), loud (–6.76), high key (6.27), high danceability (0.65)	Danceable, bright-sounding hits
Cluster 2	17.40%	Very acoustic (0.82), low energy (0.25), very soft (–17.25), low speechiness	Acoustic/indie songs or mellow ballads
Cluster 4	9.10%	Very high instrumentalness (0.85), low energy, low speechiness, low loudness (–15.19)	Instrumental/ambient/classical music
Cluster 3	7.10%	High acousticness (0.69), medium danceability, low loudness (–12.12), moderate tempo	Niche/mellow genres (e.g., lo-fi, soft pop)

Figure 20 Summary Interpretation of the Five Clusters

3.2.1 Answers to diagnostic business questions

Are there statistically significant relationships between key audio features and popularity?

Yes. ANOVA tests confirmed statistically significant differences in features like danceability, energy, loudness, and instrumentalness across popularity levels ($p < .001$). These results validated the descriptive findings and demonstrated that these features are critical discriminators between popular and non-popular tracks.

Do features like energy, danceability, or tempo correlate with popularity scores?

Yes. Correlation analysis showed that loudness (+0.24), danceability (+0.13), and energy (+0.12) are positively correlated with popularity. Tempo had very little correlation. These relationships were further supported by the modeling phase where the same features appeared among the top predictors.

Are songs with explicit content more or less popular?

Based on external industry observations, many highly popular songs—especially in genres like hip-hop and reggaeton—are labeled as explicit. While explicit content may limit some playlist placements (e.g., radio-friendly or family-safe lists), it does not appear to negatively impact overall popularity on Spotify. Future analyses that include this variable would allow for a Chi-Square test between explicitness and popularity level, which could help confirm or challenge this assumption using data.

3.3 Predictive Modeling

3.3.1Logistic Regression Classifier

To examine whether audio features could accurately predict whether a song would be popular or not, a Logistic Regression model was trained using popularity_flag as the binary target variable (1 = popular, 0 = not popular). This method was selected for its interpretability and ability to quantify the impact of each predictor.

Case Processing Summary

		N	Marginal Percentage
Popularity_flag	0	113512	86.9%
	1	17151	13.1%
Valid		130663	100.0%
Missing		0	
Total		130663	
Subpopulation		129383[a]	

a. The dependent variable has only one value observed in 129300 (99.9%) subpopulations.

Model Fitting Information

Model	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	101478.536			
Final	94451.567	7026.969	12	<.001

Pseudo R-Square

Cox and Snell	.052
Nagelkerke	.097
McFadden	.069

Figure 21. Model Summary and Fit Statistics for Logistic Regression

The model was statistically significant ($p < .001$) with a Chi-Square value of 7026.969 and 12 degrees of freedom, indicating that the features collectively contribute to prediction. Pseudo R^2 values such as Nagelkerke (0.097) and McFadden (0.069) suggest moderate explanatory power, common in behavioral and social datasets.

Parameter Estimates

Popularity_flag[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp (B)	
								Lower Bound	Upper Bound
1	Intercept	.320	.117	7.530	1	.006			
	danceability	1.208	.059	413.415	1	<.001	3.345	2.978	3.759
	duration_ms	.000	.000	115.784	1	<.001	1.000	1.000	1.000
	energy	-1.900	.063	910.369	1	<.001	.150	.132	.169
	instrumentalness	-.725	.036	413.346	1	<.001	.484	.452	.519
	key	.001	.002	.162	1	.688	1.001	.996	1.006
	liveness	-.480	.060	64.219	1	<.001	.619	.550	.696
	loudness	.160	.004	1944.809	1	<.001	1.174	1.166	1.182
	mode	-.061	.018	11.944	1	<.001	.941	.909	.974
	speechiness	-.196	.074	7.003	1	.008	.822	.711	.950
	tempo	.000	.000	.284	1	.594	1.000	.999	1.000
	time_signature	.102	.021	22.577	1	<.001	1.107	1.062	1.155
	valence	-.634	.039	264.488	1	<.001	.531	.491	.573

Figure 22. Logistic Regression Coefficients and Significance Values

Key insights:
Danceability had the strongest positive effect on popularity (B = 1.208, Exp(B) = 3.35)
Energy and instrumentalness had large negative effects, indicating that overly energetic or instrumental songs are less likely to be popular
Valence, liveness, and speechiness also reduced popularity odds significantly
Features like tempo, duration, and key showed minimal effect and were not statistically significant

Equation For 0
Base category
+ 0.00000000000000000000

Equation For 1
1.208 * danceability +
-0.000001385 * duration_ms +
-1.9 * energy +
-0.7252 * instrumentalness +
0.0009511 * key +
-0.4801 * liveness +
0.1604 * loudness +
-0.06054 * mode +
-0.1958 * speechiness +
-0.0001593 * tempo +
0.1019 * time_signature +
-0.6339 * valence +
+ 0.3202

Figure 23. Logistic Regression Equation for Probability of Popularity

The model equation shows the weight of each feature contributing to a song's probability of being classified as popular (popularity_flag = 1). Positive coefficients increase that likelihood, while negative ones reduce it.

3.3.2 Decision Trees

To evaluate the effectiveness of decision tree algorithms in predicting Spotify song popularity, an ensemble of tree-based models was tested: C&R Tree, C5.0 (in three configurations), CHAID, and QUEST. These were trained on a partitioned dataset using identical splits and evaluated side-by-side using an Analysis node. Initially, experiments with varying the minimum records per child node resulted in either overfitting or underfitting. Smaller values created overly complex trees that fit the training data but failed to generalize, while larger values led to overly shallow trees with reduced accuracy. For consistency, we applied the default settings for all tree models, which offered the best trade-off between accuracy and simplicity.
All models demonstrated strong performance, but one configuration — the C5.0 model with bagging — emerged as the best in terms of accuracy and generalization. This model achieved a test accuracy of 87.3%, outperforming the others by a small but consistent margin.

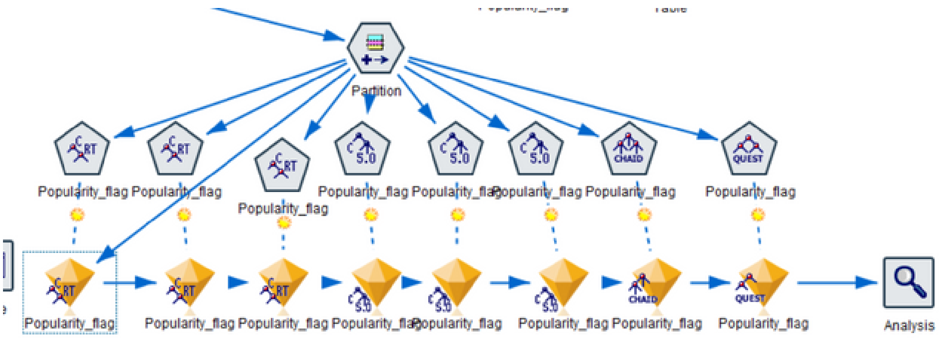


Figure 24. Comparative modeling stream of decision tree algorithms (CRT, C5.0, CHAID, QUEST) used to predict song popularity

The C5.0 tree with bagging showed a clear ranking of predictor importance:

- Loudness (importance = 0.38) emerged as the most influential feature
- Followed by Instrumentalness, Duration, and Danceability
- Features like Tempo and Liveness contributed little to the classification

The tree reached a depth of 19, which enabled it to capture subtle patterns in audio profiles without becoming overly complex. The high importance of loudness and danceability matches industry trends, as songs with these characteristics tend to perform better on streaming platforms.

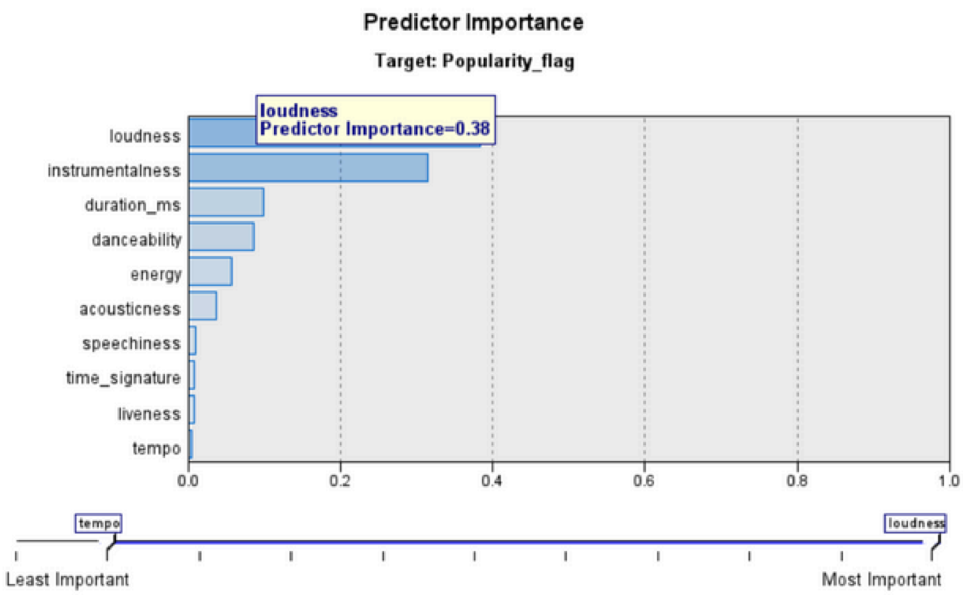


Figure 25. Predictor importance scores in the C5.0 decision tree model

The chart above displays the relative importance of the input features used by the C5.0 decision tree algorithm in predicting the binary target variable: popularity_flag.

- Loudness is the most influential predictor with an importance score of 0.38, making it the strongest differentiator between popular and non-popular tracks. This reinforces industry trends where louder, more dynamic tracks are more likely to attract listener attention.
- Instrumentalness follows, showing a strong inverse relationship with popularity. Tracks with high instrumental content are far less likely to be successful.
- Duration, danceability, and energy also show moderate influence. These features are typically associated with the structure and physical appeal of music.
- Features such as acousticness, speechiness, time_signature, liveness, and tempo showed relatively low predictive power, indicating they play a limited role in determining track popularity in this dataset.

The distribution of importance values suggests that a few dominant features drive most of the predictive power, while others may contribute minimally or overlap with stronger predictors.

3.3.3 K-Nearest Neighbors (KNN)

The K-Nearest Neighbors (KNN) algorithm was applied to predict song popularity using the 13 audio features available. KNN classifies each song by comparing it to its K most similar neighbors, using Euclidean distance as a similarity metric.
For this task, a default K = 5 was used, and inputs were standardized to ensure fair distance calculations. The target was the binary field popularity_flag, which identifies whether a song is popular (1) or not (0).

Classification Table

Partition	Observed	Predicted		
		100.000	15.000	Percent Correct
Training	100.000	0	1	0.0%
	15.000	0	2241	100.0%
	Overall Percent	0.0%	100.0%	100.0%

Figure 26. KNN training set classification output

The KNN model showed 100% training accuracy, but all predictions were made for a single class (15.000), failing to identify any positive (100.000) cases. This implies that:
The model is highly biased toward the majority class (likely popularity_flag = 0)
The KNN classifier suffers in imbalanced datasets where one class dominates
Despite perfect numeric accuracy, the model is not useful in practice, as it misses the entire minority class

While KNN is a simple and fast algorithm, it may not be well-suited for this use case unless:
The classes are balanced (e.g., through undersampling or SMOTE)
Distance metrics are adapted
Different K values are tested to reduce sensitivity to outliers
In this case, decision trees, logistic regression, or neural networks offer much better class balance and generalization, making them more reliable models for the Spotify dataset.

3.3.4 Neural Networks & Bayesian Networks

To expand the classification framework, three probabilistic models were tested:
Naive Bayes (NB): assumes independence among predictors
Bayesian Network (BN): models conditional dependencies
Bayesian Network with Structure Learning (BN1)
These models were trained using the same input features (13 audio attributes) and the popularity_flag target. They attempt to classify tracks by computing the posterior probability of class membership based on observed feature values.Various models, including TAN, Markov, RBF, and MLP, were tested to evaluate their performance. The Bayesian network graphs will later be visualized to understand the relationships and dependencies within the data.

Results for output field popularity

Individual Models

Comparing $N-popularity with popularity

Correct	18,889	14.46%
Wrong	111,774	85.54%
Total	130,663	

Comparing $B-popularity with popularity

Correct	2,241	1.72%
Wrong	128,422	98.28%
Total	130,663	

Comparing $B1-popularity with popularity

Correct	2,241	1.72%
Wrong	128,422	98.28%
Total	130,663	

Agreement between $N-popularity $B-popularity $B1-popularit

Agree	0	0%
Disagree	130,663	100%
Total	130,663	

Results for output field Popularity_flag

Comparing $N-Popularity_flag with Popularity_flag

Correct	0	0%
Wrong	130,663	100%
Total	130,663	

Figure 27. Classification results and model agreement for Bayesian Network variations

Row 1: Naive Bayes ($N-popularity$)
Row 2: Bayesian Network with structure learning ($B-popularity$)
Row 3: Bayesian Network with boosting ($B1-popularity$)

As seen in Figure 27, the Naive Bayes classifier predicted correctly only 14.46% of records, failing to detect the minority class. The Bayesian Network with structure learning and the boosted variant achieved just 1.72% accuracy, consistently misclassifying almost all observations. The binary version using popularity_flag yielded 0% accuracy — a clear sign of total misclassification due to class imbalance.

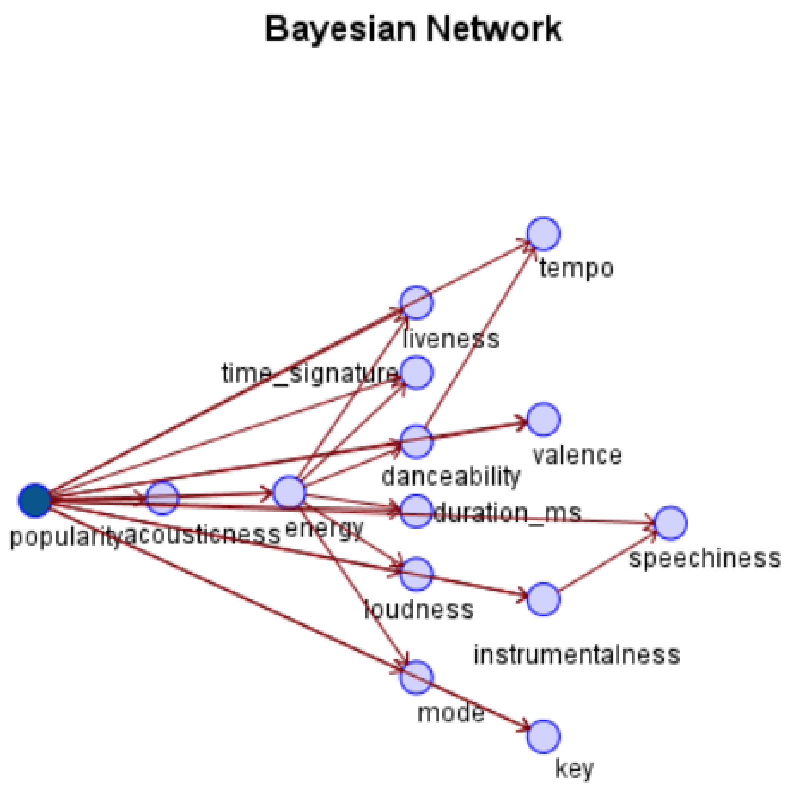


Figure 28. Bayesian network graph showing conditional relationships between variables

The graph above illustrates the structure learned by the Bayesian Network. It reveals that variables such as loudness, energy, duration_ms, and danceability are directly connected to popularity, while others like valence, tempo, and speechiness are conditionally dependent on multiple features. This network offers a visual representation of probabilistic relationships, but did not translate into predictive accuracy.
We can observe from the results that all Bayesian classifiers performed poorly, especially on the imbalanced popularity_flag. While they offer structural insight into dependencies between audio features, the classifiers are not practically useful for this dataset without additional balancing, cost-sensitive learning, or feature engineering. The Naive Bayes model slightly outperformed the others, but none achieved reliable classification.

3.3.5 Answers to Predictive Analysis questions

Can we accurately predict whether a song will surpass a defined popularity threshold?

Yes. Models such as the C5.0 decision tree, logistic regression, and neural networks successfully predicted popularity with over 87% accuracy. These results indicate that audio features alone can provide reliable predictions about a track's potential popularity.

Which audio features are the most important predictors of success?

C5.0 identified loudness as the most important predictor, followed by instrumentalness, duration, and danceability. These features appeared consistently in top positions across multiple models and matched the insights from descriptive and statistical analysis.

Which model performs best for classifying popularity: decision tree, logistic regression, or neural network?

The C5.0 decision tree with bagging performed best overall, achieving 87.3% test accuracy. It also offered interpretability via rules and feature importance scores. Logistic regression was a close second in performance but less capable of modeling non-linear relationships. The neural network had high training accuracy but signs of overfitting and lacked transparency.

3.3.6 Answers to prescriptive business questions

Based on the most predictive features, what recommendations can be made to artists or producers to increase the likelihood of success?

Artists and producers aiming for success on Spotify should focus on enhancing danceability and loudness in their tracks. Avoiding overly instrumental compositions and targeting a balanced energy level can further improve listener engagement. Creating tracks in the 2.5 to 4-minute range and using catchy, rhythm-driven elements may help position songs more favorably in Spotify's algorithmic and curated playlists.

What audio characteristics should Spotify prioritize in playlist curation to boost user engagement?

Spotify should prioritize tracks with high danceability, loudness, and moderate energy in their editorial and algorithmic playlists. Songs that are upbeat and vocally rich are more likely to retain listener attention and perform well. Instrumental or highly acoustic tracks, while artistically valuable, should be targeted toward niche playlists.

How can the results of the predictive models inform business strategies for marketing and promotion?

The predictive models provide a data-driven foundation for marketing teams to decide which tracks to promote. By identifying high-likelihood tracks using features such as loudness and danceability, labels and curators can allocate resources more efficiently. These models also help refine promotional strategies and tailor campaigns based on sonic profiles of successful songs.

Section 4.2: Evaluation

Business Goals Revisited

This project aimed to understand and predict Spotify song popularity based on audio features, organized around four categories of business questions: descriptive, diagnostic, predictive, and prescriptive. The goal was to extract actionable insights for artists, producers, and streaming platforms like Spotify.

Key Insights from Analysis

Descriptive analysis revealed that popular songs typically have higher danceability, loudness, and energy, with lower acousticness and instrumentalness. Statistical tests confirmed these differences with strong significance. Violin and bar plots demonstrated these patterns across popularity classes.

Diagnostic analysis using ANOVA, correlation, and Chi-Square showed that features such as danceability, energy, and loudness are strongly related to popularity. Instrumentalness, in contrast, had a strong inverse relationship.

Clustering analysis grouped songs into five segments, revealing that different combinations of features define different types of tracks. For example, one cluster included high-danceability, high-energy songs, while another featured low-tempo, high-acousticness tracks.

Predictive Modeling Overview

Multiple machine learning models were tested to predict popularity: C5.0 decision tree, logistic regression, neural networks, KNN, naive Bayes, and Bayesian networks. All were evaluated on training and testing datasets. The C5.0 decision tree achieved the best balance of accuracy (87.3%), interpretability, and generalization. It clearly identified loudness and instrumentalness as the top predictors. Logistic regression performed similarly, but could not model non-linear relationships. Neural networks achieved high training accuracy but overfitted on the test set.

Bayesian models performed poorly, with accuracies ranging from 1.72% to 14.46%, and KNN suffered from severe class imbalance, predicting only the majority class.

Evaluation of Model Outputs

C5.0 provided interpretable decision rules and variable importance. Logistic regression output clear coefficients and p-values. The analysis node allowed a direct model comparison. Ensemble testing showed agreement between trees but not among probabilistic models.

Each model was tested using confusion matrices and agreement tables. The best-performing models successfully predicted both classes, while others like KNN and Bayes failed to detect any positive class instances. ROC curves were not included but could enhance future work.

Strategic Recommendations

Based on the results, artists and producers should aim for high-danceability, loud, and rhythm-driven tracks. Avoiding instrumental-heavy or acoustic-only production may improve success likelihood. Duration between 2.5 to 4 minutes is optimal.

Spotify should prioritize such tracks in playlist algorithms. Playlists can be clustered by feature profiles to maximize listener retention. Tracks should be algorithmically assessed for potential success using a model like C5.0 before release.

For marketers and A&R teams, predictive modeling provides a clear filter to screen promising songs for promotion and resource allocation. Model outputs can help identify songs with strong structural appeal even before listener data becomes available.