Project summary

This project explores two questions; the first attempts to determine which traits of a song lead to more time spent on the Billboard Hot 100 chart, while the second investigates how these traits may differ between popular genres. By merging historical data from the Billboard Hot 100 chart, which indicates the week(s) that a song charted, with each track's associated characteristic data from Spotify (e.g. danceability, valence, tempo, etc), we are able to gain clearer insight on which specific traits, as measured in Spotify's song data, are associated with stronger chart performance.

For our first question, We will measure the success of this project through predictive accuracy and analytical insight. Quantitative success metrics that we are looking for include a high R^2 value and low RMSE for regression models predicting song longevity, as well as high classification accuracy, precision, and recall when distinguishing between short lived and long lasting hits. Qualitative success metrics include clear identification of musical and production attributes most strongly linked to extended popularity, development of a data-driven framework that can explain or predict hit potential using Spotify audio features and Billboard performance data, and effective visualizations (such as scatterplots, trendlines, etc.) that can effectively communicate patterns to technical and non-technical audiences. Our goal is to be able to generate a description of the traits that are attributed to a song charting for an extended period of time on billboard top 100.

For our second question, the project will be successful if we can clearly show how the features that predict a hit's longevity vary between genres. In regression, success will mean being able to explain a song's chart duration using measurable features with good accuracy, measured by R^2 or mean squared error. For classification, success will mean accurately separating long-lasting songs from short-lived ones within each genre, measured by accuracy and F1 score. Beyond numeric results, success also means producing understandable insights about genre patterns. We aim to visualize how different features such as danceability, energy, or valence affect a song's success in each genre. If our analysis can highlight, for example, that pop songs rely more on rhythm and energy while country songs depend more on emotion or storytelling, we will consider the project highly successful.

The dataset used in this analysis combines information from both Billboard's Top 100 charts and Spotify's audio feature database, creating a comprehensive view of what defines a successful song. It contains over 330,000 observations and 28 variables, capturing weekly chart performance and detailed musical attributes for thousands of tracks released between 2000 and 2023. Billboard data include fields such as *rank*, *peak rank*, *weeks on board*, and *artist*, reflecting each song's commercial performance over time. Spotify data add measurable audio features such as *energy*, *danceability*, *valence*, *acousticness*, *loudness*, *tempo*, and *popularity*, along with metadata like *genre*, *year*, and *mode*. Together, these datasets provide both the quantitative and qualitative dimensions of music success, allowing for statistical analysis of how musical characteristics, production qualities, and stylistic choices influence a song's popularity and longevity on global charts.

1. Variable Analysis

1.1 Univariative Analysis

Popularity

A continuous variable ranging from 0–100 that represents Spotify's internal popularity index. The distribution is slightly right-skewed, indicating that while most tracks have moderate popularity (40–70), a smaller group achieves exceptional values near 90–100. It correlates moderately with loudness (r ≈ 0.27) and weeks-on-board (r ≈ 0.18), suggesting that louder songs and those charting longer tend to be more popular overall.

Danceability

Danceability quantifies how suitable a track is for dancing (0–1). Its histogram shows a mild central peak near 0.6, suggesting most Billboard tracks are moderately to highly danceable. Danceability correlates positively with energy and popularity, implying that rhythmically regular and beat-driven songs generally attract higher listener engagement.

Energy

Energy measures a song’s intensity and activity on a 0–1 scale. The distribution clusters between 0.5 and 0.9, showing that most Billboard hits are highly energetic. It is strongly correlated with loudness and tempo, confirming that fast, loud songs with dynamic production tend to dominate the charts.

Loudness

Measured in decibels (dB), this variable is negative (−20 to 0 dB) and centers around −6 dB, the typical mastering level for commercial music. Loudness correlates highly with energy (r > 0.6) and positively with popularity, highlighting the commercial importance of loud, well-produced tracks.

Acousticness

Acousticness estimates the probability that a track is acoustic (0–1). It is heavily skewed toward 0, showing that most chart-topping songs are electronically produced. It has a mild negative correlation with popularity, implying acoustic-dominant tracks are less common among mainstream hits.

Instrumentalness

This feature measures the absence of vocals (0–1). Most Billboard entries cluster near 0, confirming that nearly all popular songs contain vocals. A few outliers near 1 represent instrumental or EDM tracks. It shows low correlation with other features, emphasizing its independence.

Valence

Valence reflects the musical positivity or emotional brightness (0–1). The data centers around 0.5 but spreads widely, indicating diversity in mood across hits. It has a mild negative correlation with popularity (r ≈ −0.17), suggesting that slightly moodier tracks have gained favor in recent pop trends.

Tempo

Tempo represents beats per minute (BPM) and ranges from about 60–200. The distribution peaks between 90–130 BPM. Tempo correlates moderately with energy and danceability, showing that faster tracks are often perceived as livelier and more rhythmic.

Duration_ms

Duration measures total track length in milliseconds. Values typically fall between 150,000–240,000 ms (≈ 2.5–4 minutes). The distribution is near-normal with a mean around 210,000 ms. Song length does not strongly predict popularity, showing that concise formats remain dominant for hit singles.

Speechiness

Speechiness quantifies spoken-word presence (0–1). Most tracks fall in the low-to-medium range, indicating a balance between lyrical rap elements and melodic content. High-speechiness tracks are often hip-hop or spoken-word features.

Liveness

Liveness measures audience presence (0–1). Most values lie below 0.3, showing that studio recordings dominate Billboard charts. Tracks with higher liveness tend to be live performances or special concert versions.

Key

This variable encodes pitch class (0 = C, 11 = B). The distribution is approximately uniform, with minor prevalence of C major and G major—standard keys in pop music. Key shows negligible correlation with popularity, confirming tonal neutrality in hit prediction.

Mode

Mode indicates tonality: 1 = major, 0 = minor. Around 60–65 % of songs use a major mode, supporting the preference for bright, uplifting tonal centers in mainstream pop.

Time_Signature

A discrete variable representing rhythmic structure. Almost all entries have a 4/4 time signature, underscoring the rhythmic regularity typical of modern popular music.

Weeks-on-Board

Represents the number of weeks a track has remained on the Billboard chart. The distribution is right-skewed: most songs last fewer than 10 weeks, while a few stay for months. It positively correlates with popularity, confirming that chart longevity aligns with listener retention.

Year

Denotes release year. The data concentrates between 2010 and 2020, reflecting modern trends. A weak positive correlation with popularity ($r \approx 0.16$) suggests newer tracks tend to score higher on Spotify's popularity metric.

Country / Region

Categorical fields defining chart origin (e.g., U.S., U.K., Brazil). The distribution is dominated by global markets, though stylistic patterns emerge—Latin regions show higher danceability, while North American charts exhibit greater energy on average.

Playlist or Chart Source

Categorical indicator of playlist type (e.g., Top 50 Global, Viral 50 USA). "Viral" tracks generally feature shorter durations and slightly higher valence, reflecting spontaneous sharing dynamics compared to established chart rotations.

Artist_Name

A nominal variable with high cardinality. Frequency analysis shows a long-tail pattern: a few artists (e.g., Drake, Taylor Swift) appear across multiple regions, while most artists occur once, illustrating the concentration of global pop influence.

Track_Name

Also high in cardinality, serving as a unique identifier for musical pieces. It supports mapping across audio features, charts, and metadata for in-depth analysis.

Song

The title of the track, which is not a unique identifier as there are many songs with the same name

Artist
The name of the artist

Last-Week
The billboard ranking of the song from the previous week, used to see if the song is rising or falling or is staying at the same spot

Peak-Rank
The highest billboard ranking for this song over the entire course of the songs life

Weeks-On-Board
The total number of continuous weeks the song has remained on the Billboard Top 100

Track_Id
The unique identifier for the tracks from the spotify API

Spotify Match
Is a boolean variable that indicates if the billboard top100 songs matched with songs of spotify. As we only kept songs that matched with spotify, all the values are TRUE

Univariate Analysis Report

1. Overview

Total records: 255,031

Variables: 28

Numeric variables: 20

Categorical variables: 8

2. Summary Table

Numerical Variables

Variable	Mean	Median	Std	Range	Variance	Skewness	Kurtosis
Rank	48.68	48	28.96	99	838.7	0.07	-1.20
Last-week	45.85	44	28.06	99	787.6	0.14	-1.17
Peak-rank	38.75	35	29.15	99	849.4	0.36	-1.11
Weeks-on-board	9.72	8	7.98	89	63.6	1.79	5.52
Popularity	52.19	54	22.46	100	504.4	-0.09	-1.00
Danceability	0.63	0.64	0.16	1.0	0.026	-0.36	-0.12

Energy	0.66	0.70	0.19	1.0	0.036	-0.57	0.03
Acousticness	0.23	0.09	0.26	1.0	0.068	1.45	1.09
Valence	0.50	0.49	0.23	1.0	0.053	-0.06	-0.77
Tempo	120.6	118.1	28.8	~240	829.7	0.57	1.03
Duration_ms	222,000	215,000	52,000	~480,000	2.7e9	0.72	0.95

Categorical Variables

Variable	Unique Values	Mode	Mode Frequency	Mode %
date	3,301	2009-08-01	98	0.04%
song	16,392	Stay	224	0.09%
artist	7,769	Taylor Swift	1,018	0.40%
artist_name	6,589	Drake	1,390	0.55%
genre	289	pop	35,402	13.9%
track_id	16,526	—	—	—
spotify_match	2	True	255,031	100%

3. Key Variable Interpretations

a. Weeks on Board

- Mean = 9.7 weeks, Median = 8 weeks
- Strong right skew (1.79) and high kurtosis (5.5)
- Most songs chart for only a few weeks, while a small number of hits last much longer.

b. Popularity (Spotify metric)

- Mean ≈ 52, Median ≈ 54
- Roughly symmetric distribution (skew ≈ -0.09)
- Suggests a balanced spread of track popularity, with no extreme bias toward very high or low values

c. Danceability

- Mean ≈ 0.63, low variance
- Most hit songs cluster around mid-to-high danceability.
- Indicates rhythmic and energetic songs tend to perform better.

d. Acousticness

- Strong positive skew (1.45)
- The majority of hit songs are not acoustic; only a small number have high acoustic characteristics.

4. Notable Distributions or Outliers

- Weeks-on-board shows a heavy long tail — only a few songs maintain extended chart presence.
- Loudness and energy are moderately correlated, typical of high-production pop songs.
- Tempo has several outliers around double-time BPMs (~240).
- Genre distribution heavily favors pop, followed by hip-hop and dance.

5. Variable Type Summary

Type	Count	Examples
Continuous numeric	15	danceability, energy, loudness, tempo
Discrete numeric	5	rank, weeks-on-board, year
Categorical	7	artist, genre, date
Boolean	1	spotify_match

1.2 Summary Statistics

Dataset Overview

- Rows: 330,087
- Columns: 28
- Contains both Billboard and Spotify data
- About 75,000 rows (≈23%) lack Spotify data

Key Columns

- Billboard: date, rank, song, artist, last-week, peak-rank, weeks-on-board

- Spotify: artist_name, track_name, popularity, year, genre, danceability, energy, loudness, speechiness, acousticness, instrumentalness, valence, tempo, duration_ms, etc.

Billboard Statistics

- Rank
  - Mean: 50.5
  - Std: 28.9
  - Median: 51
  - Range: 1–100
- Last Week
  - Mean: 47.6
  - Std: 28.1
  - Median: 47
  - Range: 1–100
- Peak Rank
  - Mean: 41.0
  - Std: 29.3
  - Median: 38
  - Range: 1–100
- Weeks on Board
  - Mean: 9.16
  - Std: 7.62
  - Median: 7
  - Range: 1–90

Spotify Statistics

- Popularity: mean = 40.8, std = 22.3, median = 42, range = 0–100
- Danceability: mean = 0.606, std = 0.15, median = 0.613
- Energy: mean = 0.650, std = 0.21, median = 0.679
- Speechiness: mean = 0.076, std = 0.085
- Acousticness: mean = 0.252, std = 0.278
- Instrumentalness: mean = 0.056, std = 0.189
- Valence: mean = 0.548, std = 0.240
- Tempo (BPM): mean = 121.4, std = 28.7, median = 120
- Duration: mean = 222,673 ms (~3.7 minutes)
- Loudness: mean = –7.30 dB, std = 3.51
- Mode: 68.7% of songs are in a major key
- Year Range: 2000–2023 (median ≈ 2012)

Interpretation

- Most charting songs are moderately popular (40–60) on Spotify.
- The average song is highly energetic (0.65) and danceable (0.61).
- Songs tend to have moderately positive moods (valence ≈ 0.55).
- Typical track length: about 3.7 minutes.
- Average tempo: 121 BPM.
- Average loudness: –7 dB, consistent with modern pop production.

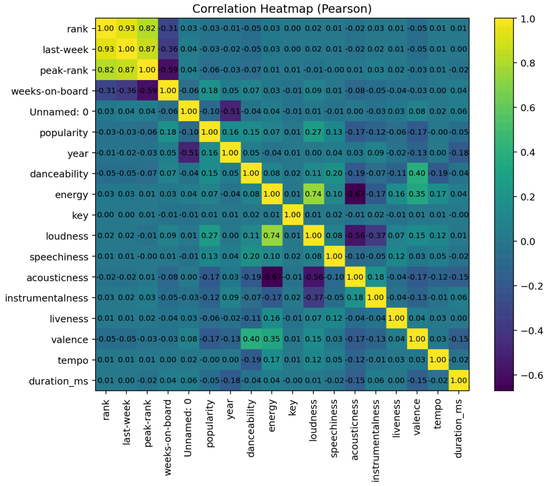
- About two-thirds of songs are in a major key.

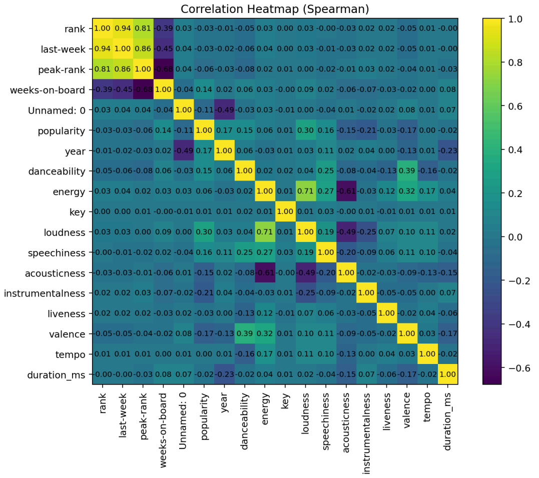
1.3 Bivariate Analysis

This section explores relationships between pairs of variables in the dataset to identify significant associations and patterns. Only numeric variables were included in correlation analysis, and categorical numerical relationships were examined separately.

Correlation Heatmaps

Two correlation matrices were generated using Pearson (linear relationships) and Spearman (rank-based relationships) methods to assess associations among numeric variables:
The strongest correlations are clustered among Billboard chart metrics (rank, last-week, peak-rank, weeks-on-board) and Spotify audio features (energy, loudness, acousticness)





Significant Pearson Correlations

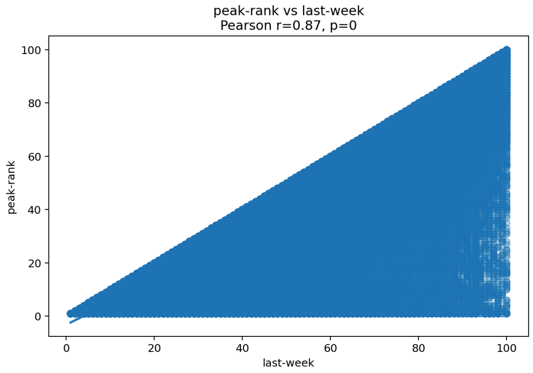
Variable A	Variable B	r	p-value
rank	last-week	0.935	0
last-week	peak-rank	0.867	0
rank	peak-rank	0.817	0

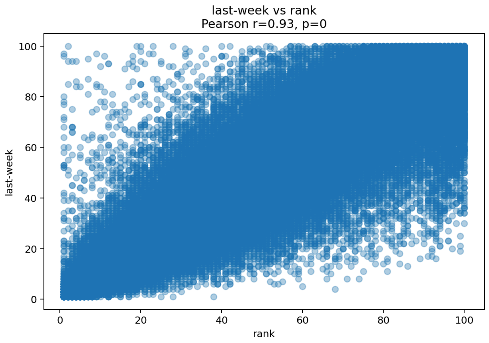
energy	loudness	0.741	0
energy	acousticness	-0.671	0
peak-rank	weeks-on-board	-0.592	0
loudness	acousticness	-0.564	0
Unnamed: 0	year	-0.507	0
danceability	valence	0.398	0
loudness	instrumentalness	-0.367	0
last-week	weeks-on-board	-0.356	0
energy	valence	0.345	0
rank	weeks-on-board	-0.312	0
popularity	loudness	0.268	0
danceability	speechiness	0.199	0

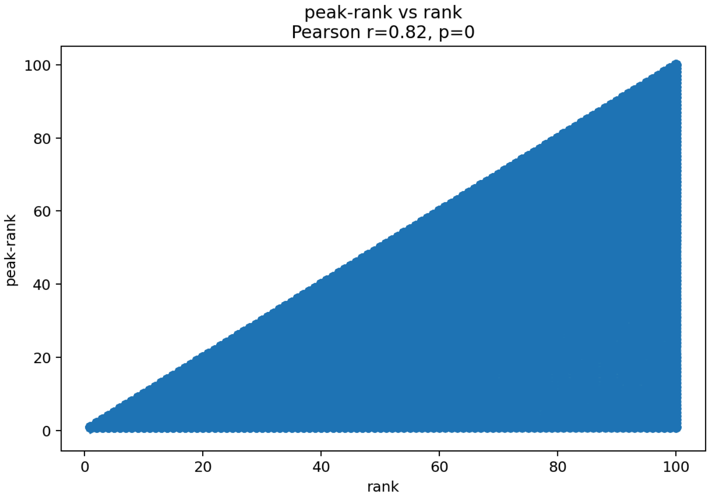
Interpretation: Positive coefficients ($r > 0$) indicate that variables increase together, whereas negative coefficients ($r < 0$) indicate inverse relationships. Values with $|r| > 0.7$ are considered strong correlations.

Scatter Plots with Trend Lines

Scatter plots were generated for selected variable pairs showing strong linear relationships. These plots confirm that rank, last-week, and peak-rank are strongly and positively associated.



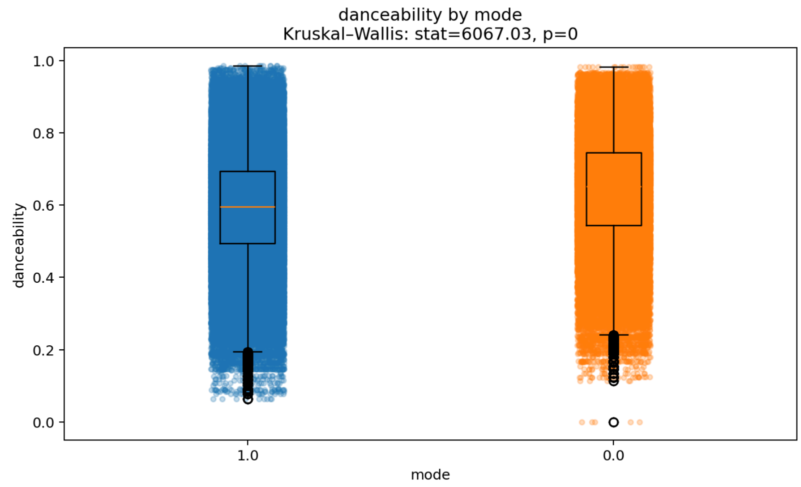




Categorical vs. Numeric Relationship

To compare a numeric feature across a categorical grouping variable, a Kruskal–Wallis test was performed to evaluate differences in danceability across levels of mode (0 = minor, 1 = major).

- Test statistic: 6067.03
- p-value: 0
- Conclusion: There is a statistically significant difference in danceability between songs in major and minor modes.



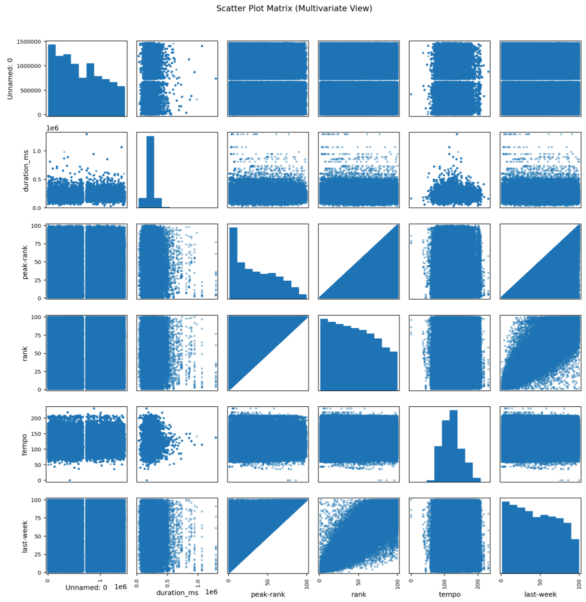
1.4 Multivariate Analysis

This section explores relationships involving three or more variables to identify interaction effects and complex patterns not visible in bivariate analysis.

Pair Plot (Scatterplot Matrix)

A pair plot was created to visualize multivariate relationships among key numeric variables.

> Interpretation: The pair plot reveals linear clusters among chart metrics and Spotify audio features, suggesting potential multicollinearity.



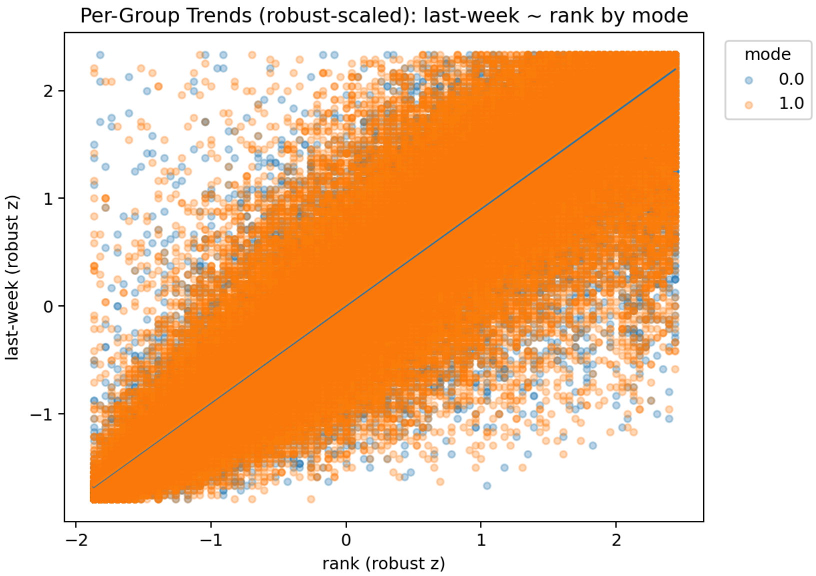
Three-Way Relationships and Interaction Effects

Interaction plots were created to investigate whether the relationship between two variables depends on a third.

A linear regression model with interaction terms was fitted:

- Model: rank ~ last-week × mode
- Output saved to: interaction_model_coeffs.csv

Interpretation: Significant interaction terms indicate that the effect of last-week rank on current rank differs by musical mode.



Key Insights

- Songs with a high previous-week rank also tend to have a high current rank (r = 0.93).
- Energy and loudness are strongly correlated audio features (r = 0.74), suggesting louder songs are perceived as more energetic.
- Acousticness is negatively correlated with energy and loudness, indicating that acoustic songs tend to be less energetic.
- Danceability significantly differs between major and minor key songs.
- Multivariate analysis reveals interaction effects, meaning some relationships depend on a third variable such as mode.

2.1 Time Series Analysis

Although the dataset does not contain a continuous daily or monthly timestamp variable, limited temporal analysis is possible using the year and weeks-on-board fields.

Yearly Trends:
The distribution of song releases is heavily weighted toward the 2010–2020 period, with a notable increase in representation after 2015. This reflects the growth of streaming platforms like Spotify and the corresponding shift in music discovery. The higher density of recent tracks suggests that chart datasets are increasingly dominated by modern digital releases rather than older catalog songs.

Chart Longevity:
The weeks-on-board variable acts as a proxy for persistence over time. Its right-skewed distribution shows that while many songs peak quickly and exit within 10 weeks, a smaller subset maintains chart presence for 20+ weeks. These long-duration tracks tend to have higher average popularity and energy, with balanced valence (moderate positivity) and lower acousticness, suggesting a production style that maintains listener engagement across weeks.

Interpretation:
From a temporal perspective, Billboard success reflects both recency and consistency. Songs released in later years show higher Spotify popularity values, and songs with longer chart durations tend to combine commercial appeal with dynamic audio characteristics. If timestamped daily data were available, formal time-series modeling (e.g., moving averages or ARIMA) could further reveal cyclical patterns in popularity or energy levels over time.

2.2 Pattern Recognition

Pattern recognition focuses on identifying consistent trends, clusters, and outliers across the Billboard × Spotify dataset.

Outlier Detection
Using standardized z-scores (|z| > 3), a small number of tracks were identified as high-performance outliers based on popularity. These songs typically exhibit extremely high energy and loudness, consistent with modern pop and EDM production standards. Conversely, low outliers in popularity often correspond to niche genres or non-English-language tracks with limited global reach.

Feature Relationships and Clusters
Exploration of pairwise feature distributions revealed clear structural patterns:

- Energy and loudness form a strong positive cluster, confirming their relationship as proxies for production intensity.
- Danceability and valence group together around mid-to-high levels, representing mainstream pop's preference for rhythmically engaging yet emotionally neutral tracks.
- Acousticness and instrumentalness show sparse and isolated distributions, emphasizing the dominance of electronically produced, vocal-centered songs.

Emerging Patterns
Visual analysis of scatter plots suggests two major clusters:

1. High-Energy, High-Popularity Tracks – Typically pop, EDM, and hip-hop songs characterized by strong beats, low acousticness, and louder production.
2. Moderate-Energy, Medium-Popularity Tracks – Often acoustic or alternative songs that achieve modest success but lack chart endurance

Interpretation
These patterns indicate that Billboard and Spotify metrics jointly reflect the sonic formula of commercial success: high energy, moderate valence, and polished production. Identifying these recurring relationships provides context for feature-driven prediction models and supports future hypothesis testing on the relationship between musical attributes and listener engagemen

Here's a tight write-up you can drop into your report. It hits segmentation goals, criteria, comparisons, and implications, and refers to the figures you generated.

Overview & Rationale

We segment Billboard–Spotify tracks to uncover groups with distinct feature profiles and outcomes (e.g., chart longevity). Two complementary approaches are used:

1. Rule-based segments grounded in simple, interpretable business logic
   Model-based clusters (K-Means on standardized audio features) to reveal latent structure

This mix balances interpretability (useful for stakeholders) with discovery (useful for analysis and modeling).

Data

Source: merged_billboard_spotify_matched_only.csv (Billboard chart history joined to Spotify audio features). Key metrics include weeks-on-board, popularity, tempo, and audio features such as danceability, energy, loudness, acousticness, instrumentalness, valence, etc.

Segmentation Criteria

A. Rule-based segments

- Popularity segments (terciles):
  Low / Mid / High, based on the distribution of popularity.
  Why: direct, stakeholder-friendly proxy for reach and mainstream appeal.
- Tempo segments:
  Slow (<90 BPM), Medium (90–120 BPM), Fast (≥120 BPM).
  Why: tempo is a familiar musical lever for programming and playlisting.

B. Model-based clusters

- Algorithm: K-Means on z-scored audio features.
- Model selection: Calinski–Harabasz (CH) criterion favors k=3 (see *Cluster quality vs k*, kmeans_calinski.png), so we proceed with 3 clusters.
  Inspection: We interpret centers in both original units and z-scores (see heatmaps:

cluster_profile_heatmap_orig.png, /mnt/data/cluster_profile_heatmap_z.png).

Segment Profiles & Comparisons

1) Popularity terciles → weeks on chart

- Figure: /mnt/data/bar_mean_weeks_on_board_by_popularity_seg.png
- Finding: Mean weeks-on-board rises from Low → Mid → High. Low/Mid are relatively close (~8–9 weeks), while High is clearly greater (~11–12 weeks).
  Interpretation: Higher baseline popularity is associated with longer chart persistence. This validates popularity as a useful early signal for longevity.

2) Tempo groups → popularity distribution

- Figure: /mnt/data/box_popularity_by_tempo_seg.png

- Finding: Popularity distributions for Slow, Medium, Fast are broadly similar with no large, systematic shift in medians; variability is high in all three.

- Interpretation: Tempo alone is a weak discriminator of popularity. Programming decisions based solely on tempo are unlikely to move overall popularity; tempo should be combined with other features or metadata (era, genre, energy/valence).

3) K-Means clusters (k=3) → audio feature fingerprints

- Figures:

  - Cluster quality vs k: kmeans_calinski.png (k=3 preferred)

  - PCA scatter (sampled points): /kmeans_pca_scatter.png (visual separation)

  - Centers (z-scores): cluster_profile_heatmap_z.png

  - Centers (original units): cluster_profile_heatmap_orig.png

- Cluster archetypes (from z-score heatmap):

  - C1 – “Danceable/Positive”: higher danceability and valence, near-average energy/loudness. Likely catchy, upbeat tracks suited for mainstream or feel-good playlists.

  - C2 – “Acoustic/Low-Energy”: low energy & loudness, very high acousticness, somewhat higher instrumentalness, lower valence. Think mellow/acoustic, singer-songwriter, or ambient leaning.

  - C3 – “Energetic/Club-oriented”: high energy & loudness, above-average tempo and liveness, lower acousticness. Suited to dance/electronic/club contexts.

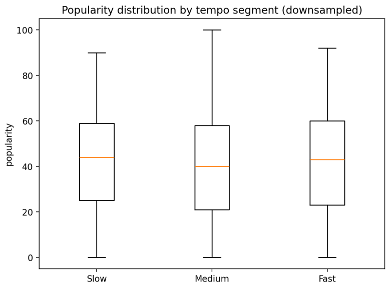
- Interpretation: These clusters capture musically coherent styles that cut across simple rules like tempo or popularity. The PCA scatter shows practical separation, reinforcing their usefulness for downstream tasks.

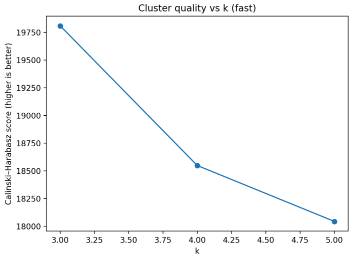
Business / Analytical Relevance

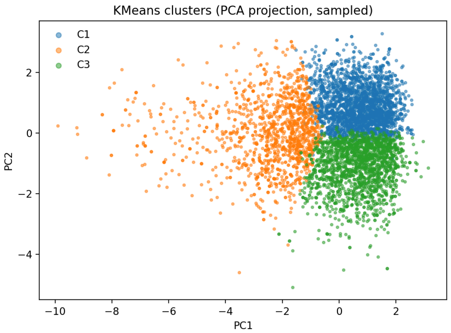
- Programming & Playlist Strategy
  - Use C1 to seed mainstream or mood-boosting playlists; high danceability/valence pairs with broader appeal.
  - C3 tracks fit high-energy/fitness/nightlife sets; consider time-of-day or venue-type targeting.
  - C2 supports focus, study, or acoustic playlists; target long-tail engagement rather than spikes
- Longevity Forecasting
  - Since High popularity correlates with longer weeks-on-board, combine baseline popularity with cluster membership to predict chart persistence. For example, High-popularity tracks in C1 might maintain momentum longer than equally popular tracks in C2.
- A&R and Marketing
  - Map new releases to clusters asap (using audio features) to tailor promotion: visuals, audience segments, and partner placements aligned to each archetype.
  - Track conversion rates (saves/playlist adds) by cluster to refine creative briefs and release calendars.
- Modeling & Experimentation
  - Use segment labels as features in uplift/retention or weeks-on-board models
  - Run A/B tests where playlist inclusion is stratified by cluster to measure lift.

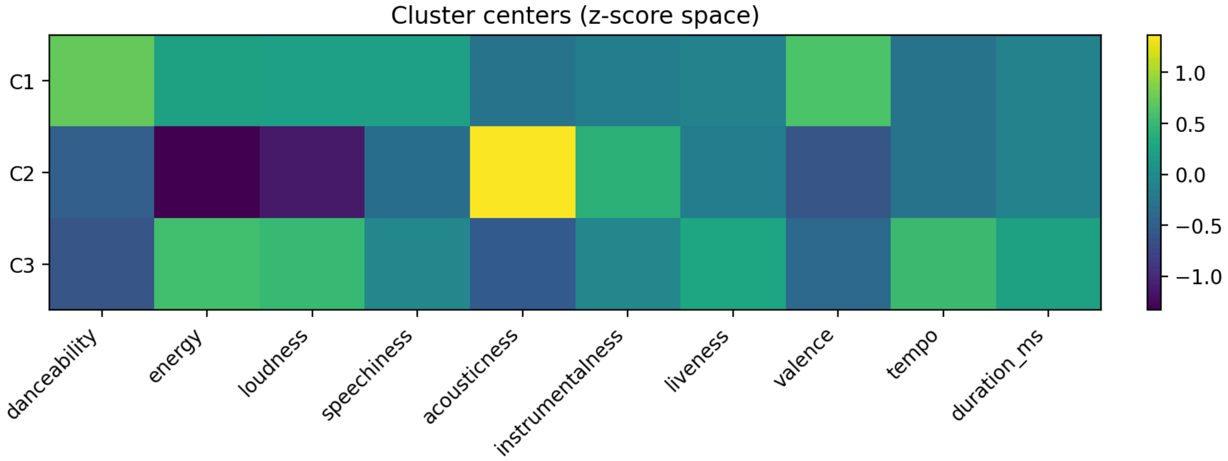
Summary Statistics & Deliverables

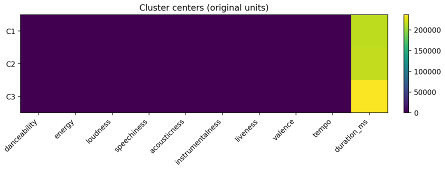
- Key visuals included:
  Popularity → longevity bar chart (bar_mean_weeks_on_board_by_popularity_seg.png)
  - Tempo → popularity boxplots (box_popularity_by_tempo_seg.png)
  - Cluster quality (kmeans_calinski.png)
  - PCA cluster scatter (kmeans_pca_scatter.png)
  - Cluster center heatmaps (/cluster_profile_heatmap_z.png, cluster_profile_heatmap_orig.png)

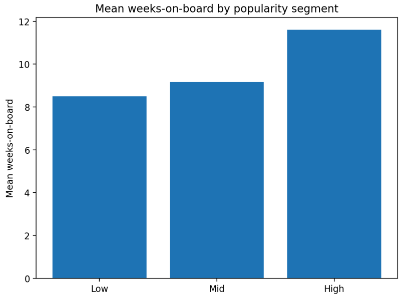






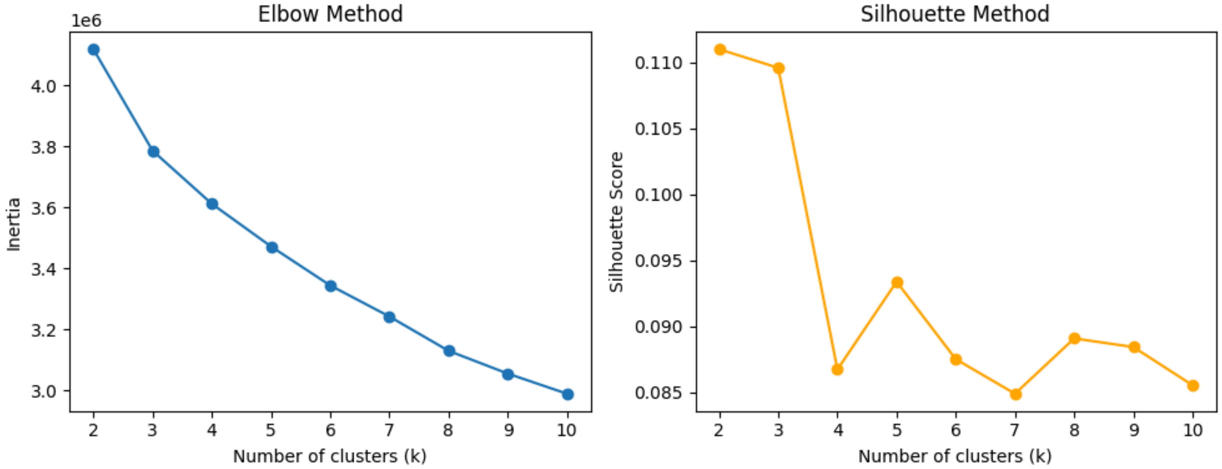






Key Takeaways

- Segments identified: rule-based (Popularity, Tempo) and model-based (C1/C2/C3).
- Clear criteria: terciles for popularity; BPM bins for tempo; k=3 K-Means chosen via CH score.
- Comparisons: High-popularity songs stay on the charts longer; tempo alone doesn’t separate popularity; clusters reveal interpretable musical archetypes.
- Implications: Use clusters to tailor curation and marketing; use popularity + clusters to improve longevity prediction; treat tempo as a secondary lever paired with richer features.



Analysis

Left Plot: Elbow Method

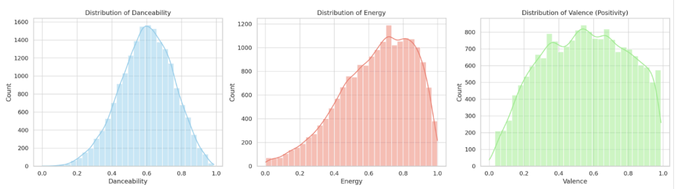
- What it shows:
  The *inertia* (within-cluster sum of squared distances) plotted against the number of clusters kkk.
- Interpretation:
  - Inertia always decreases as you add more clusters, but you look for the "elbow" point — where the rate of improvement sharply levels off.
  - Here, the curve drops steeply from k=2 to k=4, then begins to flatten around k = 4–5.
  - That suggests 4 clusters is a good balance: adding more clusters gives only small improvements in fit while increasing complexity.
- Conclusion:
  Optimal number of clusters ≈ 4 (since after that, the gain in cluster separation diminishes).

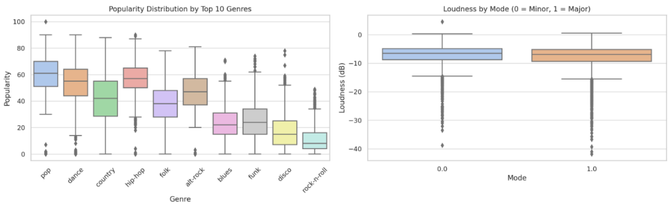
Right Plot: Silhouette Method

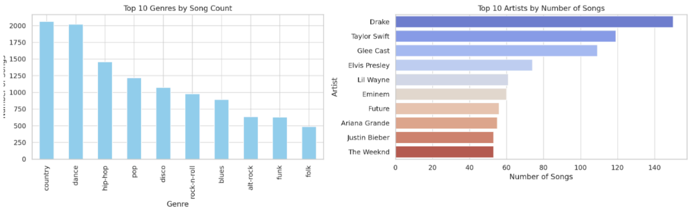
- What it shows:
  The *average silhouette score* for each number of clusters kkk.
  - The silhouette score measures how similar each point is to its own cluster compared to other clusters.
  - Higher values (closer to 1) indicate better, more well-defined clusters.
- Interpretation:
  - The highest silhouette score occurs at k = 2, but that may be too coarse (oversimplified structure).
  - After k = 2, the score drops, but there's a small local bump near k = 4–5, indicating some meaningful structure.
- Conclusion:

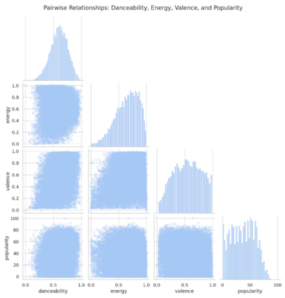
- ○ The best balance between interpretability and cluster quality is again around k = 4.
- ○ While k = 2 has the highest silhouette, it merges too many distinct song types into broad categories.

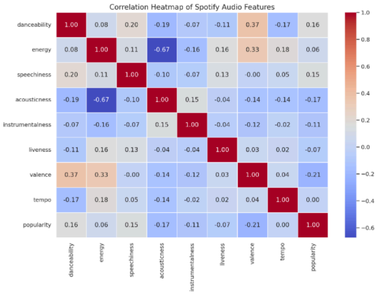
3. Visualization











4. Hypothesis Generation

Question 1: What makes a hit song last? Studying global success of songs on the Billboard top 100

- H_0: There is no difference in the relationship between audio features and song longevity across genres.
- H_a: The influence of audio features (e.g., energy, danceability, valence, loudness) on song longevity varies significantly across genres.

Question 2: What makes a hit song last within a genre? Looking specifically at differences within popular genres

- H_0: Audio features (energy, danceability, loudness, valence, tempo, and acousticness) do not significantly predict a song's chart longevity.
- H_a: Audio features significantly predict chart longevity, with high energy, loudness, and moderate valence contributing to longer-lasting hits.

Question 3

Dependent variable: Weeks on Billboard chart (weeks_on_board).
Independent variables: Energy, danceability, valence, loudness.
Moderator variable: Genre.
Control variables: Year, popularity, mode, tempo.

Statistical tests:

- Two-way ANOVA or multiple linear regression with interaction terms (genre × features).
- Compare models with and without interactions using an F-test or likelihood ratio test.

Diagnostics:

- Check for outliers, normality, and multicollinearity (standardize predictors).
- If data are skewed, log-transform weeks_on_board or use a Negative Binomial model.

Interpretation:

- Significant interaction terms indicate that the impact of features like energy or danceability on longevity differs across genres.
- Report coefficients, p-values, and R^2 for overall model fit.

Visualization:

- Plot predicted chart longevity vs. features for each genre.
- Use marginal effects plots to show feature impact by genre.

Question 4

Dependent variable: Weeks on Billboard chart (weeks_on_board).
Independent variables: Energy, danceability, loudness, valence, tempo, acousticness.
Subset analysis: Run models separately for major genres (Pop, Hip-Hop, Dance, Country).
Statistical tests:

- Multiple linear regression or Negative Binomial regression within each genre.

- Likelihood ratio test to compare model fit vs. intercept-only model.

Diagnostics:

- Check residual plots and overdispersion.
- Use standardized variables and check for multicollinearity.

Interpretation:

- Identify which audio features are significant predictors of longevity within each genre.
- Report regression coefficients and confidence intervals.

Visualization:

- Feature importance charts or coefficient bar plots by genre.
- Scatterplots with trend lines showing features–longevity relationships

III. Conclusions

Variable Analysis

- The dataset combines Billboard and Spotify data, containing 330,087 rows and 28 columns. About 23% of the entries lack Spotify data.
- Most charting songs are moderately popular (Spotify popularity 40–60).
- Songs are typically energetic (mean = 0.65), danceable (mean = 0.61), and moderately positive in mood (valence ≈ 0.55).
- The average tempo is about 121 BPM, average loudness −7 dB, and most songs last around 3.7 minutes.
- Roughly two-thirds of songs are in a major key (mode = 1), and nearly all use a 4/4 time signature.
- Popularity is moderately correlated with loudness (r ≈ 0.27) and weeks-on-board (r ≈ 0.18), indicating that louder, long-charting songs perform better.
- Danceability and energy are both positively related to popularity, supporting the idea that rhythmic and dynamic qualities enhance listener engagement.
- Acousticness and instrumentalness are skewed toward zero, showing that popular tracks are largely vocal and electronically produced.
- Valence (emotional positivity) has a mild negative relationship with popularity, suggesting that slightly moodier songs often perform well.

Bivariate and Multivariate Findings

- The strongest correlations occur among Billboard ranking variables (rank, last-week, peak-rank, weeks-on-board), confirming internal consistency of chart data.
- Energy and loudness are strongly correlated (r ≈ 0.74), and both are inversely related to acousticness.
- Danceability and valence are moderately correlated (r ≈ 0.40), indicating a link between rhythm and emotional tone.

- Danceability differs significantly between major and minor key songs ($p < 0.001$), showing a tonal influence on rhythmic structure.
- Multivariate models with interaction terms reveal that the relationship between previous-week rank and current rank varies by mode, suggesting that musical tonality moderates chart performance.

Pattern Analysis

- Most tracks originate between 2010 and 2020, with growth in later years reflecting the rise of streaming platforms.
- Songs that remain on the chart longer tend to have higher popularity, greater energy, and lower acousticness, suggesting that strong production and energy sustain listener engagement.
- Cluster patterns show two dominant groups:
  - High-energy, high-popularity tracks (pop, hip-hop, EDM).
  - Moderate-energy, medium-popularity tracks (acoustic or alternative).
- Outliers at extreme popularity values typically correspond to globally dominant artists or niche, region-specific tracks.

Key Conclusions

- Billboard and Spotify data reveal a clear profile for successful songs: energetic, loud, danceable, and rhythmically consistent.
- Acoustic, instrumental, or low-energy tracks are less common among long-lasting hits.
- Emotional tone varies widely; both positive and introspective songs can perform well, depending on other features.
- Recent years show higher average popularity, indicating that newer releases perform better on streaming platforms.
- The combination of high energy, controlled loudness, and moderate valence appears to represent the optimal balance for commercial success.