Hit Songs Through the Decades

A Machine Learning Approach to Predicting Release Periods and Analyzing Trends in Popular Music

Andrea Lovato   Elwin Freudiger   Maya Wahart

2025-05-17

Table of contents

```
Rows: 10000 Columns: 35
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (16): Track URI, Track Name, Artist URI(s), Artist Name(s), Album URI, ...
dbl  (16): Disc Number, Track Number, Track Duration (ms), Popularity, Dance...
lgl   (2): Explicit, Album Genres
dttm  (1): Added At

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

1 Introduction

The primary objective of this project is to apply machine learning techniques to a dataset of 10,000 top-charting songs from the ARIA and Billboard rankings, spanning from the 1950s to 2024. Using audio and metadata features provided by Spotify, the project aims to explore the evolution of musical trends and understand the key characteristics that define popular music across different decades.

From a machine learning perspective, the project focuses on both supervised and unsupervised learning tasks. This includes building predictive models to estimate a song's release period based on its features, and using clustering algorithms to uncover latent patterns in the data. The goal is not only to analyze historical music trends but also to assess the feasibility and accuracy of predictive modeling in the context of music analytics.

2 Literature review

2.1 Music Analytics and Machine Learning

The intersection of music and data science has increasingly attracted scholarly attention, particularly due to the rise of streaming platforms that provide large-scale, structured musical data. Several studies have explored the potential of audio and metadata features—such as those provided by Spotify—for understanding musical trends, classifying genres, and predicting popularity (Schedl et al., 2015). These features typically include danceability, energy, tempo, and valence, which encode perceptual and structural dimensions of sound and have been shown to correlate with human musical preferences (Friberg et al., 2011).

Machine learning techniques have been applied extensively to music data for tasks such as mood classification, genre detection (Kim et al., 2010), and hit song prediction (Herremans et al., 2019). Supervised models, particularly random forests and support vector machines, have demonstrated robust performance in tasks involving feature-based classification and regression. More recent work has applied deep learning architectures to raw audio input, though these approaches often demand substantial computational resources and training data (Choi et al., 2017).

Unsupervised learning, including clustering and dimensionality reduction techniques such as PCA, has also been used to explore latent patterns in musical corpora. For example, Jan Van Balen et al. (2015) used clustering to discover prototypical musical structures, while Serrà et al. (2012) applied PCA to uncover dominant stylistic trends over time.

2.2 Temporal Trends and Musical Evolution

Temporal analyses have shown that the characteristics of popular music evolve in response to technological, cultural, and economic forces (Mauch et al., 2015). For instance, changes in average song duration have been linked to shifts in radio programming, physical formats (e.g., vinyl, CDs), and digital streaming incentives. The increasing prevalence of explicit lyrics has been attributed to looser content restrictions and changing cultural norms (Pachet, 2008).

Empirical work by Interiano et al. (2018) used Spotify audio features to track the emotional content of popular music, observing a long-term trend toward increased sadness and decreased acousticness in top-charting songs. Similarly, it has been demonstrated that temporal audio descriptors could be used to model the release decade of songs with reasonable accuracy.

2.3 Challenges in Modeling Popularity and Period

While popularity is a central variable in music analytics, its measurement remains opaque. Spotify's proprietary popularity score is influenced by recent streaming activity, skips, and playlist placements, making it a dynamic and platform-specific metric (Spotify, 2023). Consequently, its use in predictive models must be interpreted as a temporal snapshot rather than a static attribute.

Predicting a song's release year or period based on its acoustic profile poses challenges due to the high variability of musical styles within the same time frame and the enduring popularity of certain older songs. Nevertheless, several studies have demonstrated that audio features carry enough temporal signal to support classification tasks by decade or era, particularly when modeling broader stylistic shifts (Müller et al., 2010).

2.4 Contribution of the Present Study

Building on this body of literature, the present study contributes a comprehensive analysis of 10,000 top-charting songs from the ARIA and Billboard rankings, spanning more than seven decades. While prior work has focused either on genre classification or mood prediction, this study uniquely combines exploratory data analysis, unsupervised learning (e.g., PCA, clustering), and supervised modeling to assess both the temporal evolution of music and the predictive power of audio features.

Moreover, by investigating the relationships between explicitness, duration, and popularity in a temporal context, this work highlights how platform incentives and listener behaviors have shaped recent musical trends. It also addresses gaps in the literature regarding the feasibility of using readily available Spotify features for temporal classification, and evaluates the limitations of clustering methods in capturing stylistic boundaries.

Finally, this study serves as a practical application of machine learning techniques in the music domain, offering insights for musicologists, data scientists, and digital media analysts interested in the computational modeling of cultural data.

3 Data

3.1 Sources

The dataset employed in this study, titled *"Top 10,000 Spotify Songs – ARIA and Billboard Charts"*, was obtained from Kaggle, a widely used platform for sharing datasets and data science resources. It comprises a curated collection of 10,000 tracks that have achieved significant popularity, based on historical rankings from both the ARIA (Australian Recording Industry Association) and Billboard charts. This dual-source approach ensures a broad and balanced representation of commercially successful music across English-speaking markets.

The dataset spans a temporal range from the 1950s to 2024, capturing the dynamic evolution of popular music over more than seven decades. It includes metadata and audio-based features extracted via the Spotify API, such as tempo, energy, danceability, valence, and instrumentalness, which allow for detailed computational analysis.

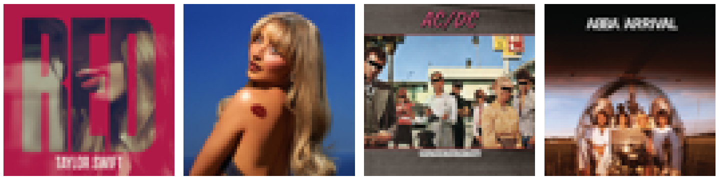
In addition to representing a variety of genres, artists, and time periods, the dataset reflects shifting cultural and musical preferences. As such, it provides a robust foundation for both exploratory data analysis and machine learning applications, particularly those aimed at uncovering temporal trends, predicting historical context (e.g., release period), and identifying latent patterns in music characteristics.

3.2 Image Data

In order to enhance the predictions, album covers were selected to be added as input in the neural network model. The orginal dataset uses a link provided by spotify to retrieve each image.

All images are retrieved from their respective links, resized to be 64x64 pixels. The size was chosen to balance concerns of space and usefulness. Indeed, a 16x16 pixel image would provide hardly any useful information. On the other hand, a 600x600 image would provide plenty of information, but would be difficult to store. All images are then saved locally, with their Spotify Song ID as a filename. This process was performed in parallel using multi-treading. This enables a much faster processing time.

Image examples can be seen here: ::: {#fig-album layout-ncol=4}



Examples of album covers, resized to 64×64 pixels and represented with RGB color channels
:::

3.3 Description

```
[1] 10000    35
```

The dataset comprises 10,000 entries and 35 variables, encompassing information related to song popularity, artist identity, release date, and various musical attributes.

Variable	Description	Category	Example
track_uri	Unique identifier for the track	character	spotify:track:123…
track_name	Name of the track	character	Bohemian Rhapsody
artist_uris	URIs of artists performing the track	character	spotify:artist:abc…
artist_names	Name(s) of the performing artist(s)	character	Queen
album_uri	Unique identifier for the album	character	spotify:album:def…
album_name	Title of the album	character	A Night at the Opera
album_artist_uris	URIs of the album's main artist(s)	character	spotify:artist:abc…
album_artist_names	Name(s) of the album's main artist(s)	character	Queen
release_date	Date the album was released	date	1975-11-21
album_image_url	Link to album cover image	character	https://i.scdn.co/image/…
disc_number	Disc number of the track in multi-disc sets	numeric	1
track_number	Track's position on the disc	numeric	11
duration_ms	Length of the track in milliseconds	numeric	354000

Variable	Description	Category	Example
preview_url	URL to 30-second preview of the track	character	https://p.scdn.co/mp3-preview/…
is_explicit	Indicates if track has explicit content	logical	TRUE
popularity	Spotify popularity score (0-100)	integer	85
isrc	International Standard Recording Code	character	GBUM71029604
added_by	User who added the track to playlist	character	user_id_123
added_at	Timestamp when track was added	datetime	2022-07-15T12:00:00Z
artist_genres	Genres associated with the artist(s)	character	rock, classic rock
danceability	How suitable a track is for dancing	numeric	0.6
energy	Intensity and activity level of the track	numeric	0.85
key	Musical key of the track (0=C, 1=C , …)	integer	5
loudness	Overall loudness in decibels	numeric	-5.3
mode	Modality: major (1) or minor (0)	integer	1
speechiness	Presence of spoken words in the track	numeric	0.05
acousticness	Confidence that track is acoustic	numeric	0.02
instrumentalness	Likelihood that track is instrumental	numeric	0.001
liveness	Likelihood of live audience presence	numeric	0.09
valence	Musical positiveness conveyed	numeric	0.7
tempo	Beats per minute (BPM)	numeric	120.5
time_signature	Estimated time signature	integer	4
album_genres	Genres associated with the album	character	rock, progressive rock
label	Record label	character	EMI

Variable	Description	Category	Example
copyrights	Copyright info for the album or track	character	℗ 1975 Queen Productions Ltd.

Track URI	Track Name	
spotify:track:0vNPJrUrBnMFdCs8b2MTNG	Fader	spoti
spotify:track:0NpvdCO506uO58D4AbKzki	Sherry	spot
spotify:track:1MtUq6Wp1eQ8PC6BbPCj8P	I Took A Pill In Ibiza - Seeb Remix	spotify:artist:2KsP6tYLJl
spotify:track:59lq75uFIqzUZcgZ4CbqFG	Let Go for Tonight	spo
spotify:track:7KdcZQ3GJeGdserhK61kfv	The Way I Want To Touch You	spo
spotify:track:000xQL6tZNLJzIrtIgxqSl	Still Got Time (feat. PARTYNEXTDOOR)	spotify:artist:5ZsFI1h6l

3.4 Wrangling / Cleaning

In preparation for analysis, the original dataset was cleaned and standardized by renaming variables to follow consistent, machine-readable naming conventions. Redundant or non-essential columns—such as URIs, preview links, and metadata unrelated to audio features—were subsequently removed. The resulting dataset *spotify_vr* retains only the relevant musical, temporal, and popularity-related attributes needed for the subsequent exploratory and predictive modeling tasks.

```
[1] 23
```

The resulting dataset consists of 23 columns, containing only the variables relevant for the subsequent analysis after the removal of redundant and non-informative features.

We only keep the ID from the Spotify track URL by extracting the final component of the *track_uri* string.

track_uri	track_name	artist_names	
0vNPJrUrBnMFdCs8b2MTNG	Fader	The Temper Trap	
0NpvdCO506uO58D4AbKzki	Sherry	Frankie Valli & The Four Seasons	Th
1MtUq6Wp1eQ8PC6BbPCj8P	I Took A Pill In Ibiza - Seeb Remix	Mike Posner, Seeb	
59lq75uFIqzUZcgZ4CbqFG	Let Go for Tonight	Foxes	
7KdcZQ3GJeGdserhK61kfv	The Way I Want To Touch You	Captain & Tennille	
000xQL6tZNLJzIrtIgxqSl	Still Got Time (feat. PARTYNEXTDOOR)	ZAYN, PARTYNEXTDOOR	S

3.5 Spotting Mistakes and Missing Data

The procedure involves identifying and counting missing values in the dataset, detecting rows containing incomplete information, and removing those with missing or empty loudness values. Then, the release year is extracted from the release date and converted to a numeric format, producing a cleaned dataset ready for further analysis.

The dataset have a total of 625 missing values, including 551 for the 'artist_genres'.

We are going to drop the instance with missing feature about the music.

```
[1] 625
```

```
       track_uri        track_name      artist_names        album_name
               0                 2                 2                 2
    release_date   album_image_url      track_number       duration_ms
               2                 4                 0                 0
     is_explicit        popularity     artist_genres      danceability
               0                 0               551                 5
          energy          loudness              mode       speechiness
               5                 5                 5                 5
    acousticness  instrumentalness          liveness           valence
               5                 5                 5                 5
           tempo    time_signature             label
               5                 5                 7
```

```
# A tibble: 557 x 23
   track_uri      track_name artist_names album_name release_date album_image_url
   <chr>          <chr>      <chr>        <chr>      <chr>        <chr>
 1 6TBUBqz21JmF~ Montego B~ Bobby Bloom  The Bobby~ 1970-01-01   https://i.scdn~
 2 5LenBcq9xeHL~ I Am the ~ The Look     The Look   1981         https://i.scdn~
 3 1A4a167Yf7bX~ Twistin' ~ Jimmy Soul   I You Wan~ 2013-12-02   https://i.scdn~
 4 1ezeinbhS1Hh~ Michelle   The Overlan~ Original ~ 2015-05-04   https://i.scdn~
 5 3tFECpGckFpp~ Forever Y~ Youth Group  Casino Tw~ 2007-01-30   https://i.scdn~
 6 46LTPfvrxlgN~ Teenage C~ Adrian Lux   Adrian Lux 2012-01-01   https://i.scdn~
 7 3LLeJjCRRkTL~ Feelin' A~ E.Y.C.       Express Y~ 1993-01-01   https://i.scdn~
 8 3kg05HyeoiSf~ All I Ask  Chase Martin All I Ask  2015-12-06   https://i.scdn~
 9 31UBcyZkwJRy~ Lady Bump  Penny McLean Lady Bump  1975         https://i.scdn~
10 4pCSXZwffXEm~ Shoop Sho~ Monte Video  Monte Vid~ 1983-12-01   https://i.scdn~
# i 547 more rows
# i 17 more variables: track_number <dbl>, duration_ms <dbl>,
#   is_explicit <lgl>, popularity <dbl>, artist_genres <chr>,
#   danceability <dbl>, energy <dbl>, loudness <dbl>, mode <dbl>,
```

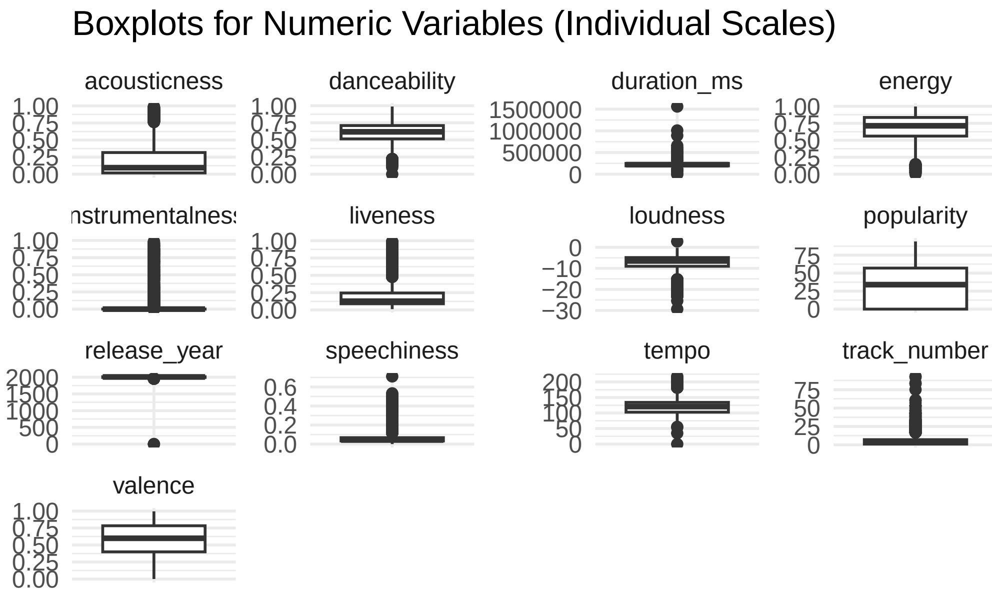
```
#   speechiness <dbl>, acousticness <dbl>, instrumentalness <dbl>,
#   liveness <dbl>, valence <dbl>, tempo <dbl>, time_signature <dbl>,
#   label <chr>
```

```
     track_uri       track_name     artist_names       album_name
             0                2                2                2
  release_date  album_image_url     track_number      duration_ms
             0                2                0                0
   is_explicit       popularity    artist_genres     danceability
             0                0              549                0
        energy         loudness             mode      speechiness
             0                0                0                0
  acousticness instrumentalness         liveness          valence
             0                0                0                0
         tempo   time_signature            label     release_year
             0                0                5                0
```

The dataset contains a total of 625 missing values, of which 551 correspond to the *artist_genres* variable. Instances with missing values in variables related to musical characteristics are excluded from the analysis to ensure data quality and consistency.

3.6 Listing Anomalies and Outliers

A subset of the dataset containing only numeric variables was selected, excluding *mode* and *time_signature*. The data was then transformed into long format to generate faceted boxplots, allowing for the visualization of the distribution and scale of each numeric variable individually.



3.6.1 Interpretation

The faceted boxplots illustrate the distribution of key numeric features. Variables such as acousticness, energy, danceability, instrumentalness, liveness, valence, and speechiness are bounded between 0 and 1. Most exhibit distributions concentrated near zero, particularly instrumentalness and speechiness, which show long right tails and outliers close to 1, suggesting that while the average track lacks strong instrumental or spoken elements, some are highly characterized by them.

Duration (in milliseconds) is slightly right-skewed, with most tracks clustered around the median and a few outliers representing exceptionally long songs. Loudness is centered around negative values, consistent with its decibel scale relative to silence, and shows a compact distribution with occasional extreme lows, likely corresponding to quiet or highly dynamic tracks.

Popularity displays a broad distribution with outliers at both ends, indicating notable variability in audience reception. Tempo shows high variance and several extreme values, potentially due to anomalous entries or experimental compositions. Lastly, track number is typically low—reflecting songs positioned early in albums—though outliers suggest the presence of long compilations or inconsistent metadata.

3.7 Summary statistics

Summary statistics are generated for all numeric variables in the cleaned dataset to provide an overview of central tendencies, dispersion, and the presence of potential outliers.

track_number	duration_ms	popularity	danceability	energy	loudness
Min. : 1.000	Min. : 0	Min. : 0.00	Min. :0.0000	Min. :0.0000203	Min. :-29.368
1st Qu.: 1.000	1st Qu.: 192084	1st Qu.: 0.00	1st Qu.:0.5155	1st Qu.:0.5610000	1st Qu.: -9.019
Median : 3.000	Median : 219413	Median :34.00	Median :0.6180	Median :0.7130000	Median : -6.49
Mean : 4.939	Mean : 224200	Mean :32.55	Mean :0.6084	Mean :0.6840504	Mean : -7.252
3rd Qu.: 7.000	3rd Qu.: 249820	3rd Qu.:57.00	3rd Qu.:0.7100	3rd Qu.:0.8360000	3rd Qu.: -4.878
Max. :93.000	Max. :1561133	Max. :94.00	Max. :0.9880	Max. :0.9970000	Max. : 2.769

The summary statistics table provides an overview of the central tendencies and dispersion of key variables in the dataset.

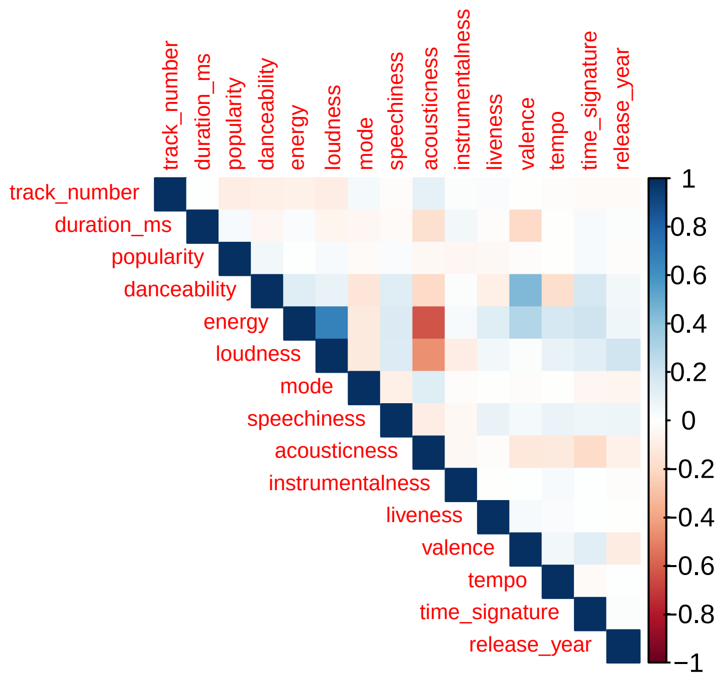
As we can see, regarding the release date, most songs have been released in more recent years. this may lead to unbalanced data. Ways to solve this unbalance will be discussed later in this paper. Regarding the track number in the song, the mean being at *4.9*, this tells us that most songs may part of an album or compilation. interestingly, the max of song number is *93* for a song named *Soul Revival* by *Johnny Diesel & The Injectors* part of a *Complete Eighties* compilation of 100 songs from the 80s.

The track duration is expressed in milliseconds, with a mean of *3 minutes and 44 seconds.* The longest song in record is a whopping 26 minutes. for *Tubular Bells - Pt. I* by *Mike Oldfield.* While this may not ring a bell (pun intended) for most readers, Amateurs of Horror may recognize this as the opening soundtrack for *The Exorcist (1973).*

The is_explicit variable is highly imbalanced, with approximately 95% of the songs labeled as non-explicit. The distribution of the popularity variable is not normally distributed, with a mean value of 33, indicating that most songs fall within a lower popularity range. Variables such as danceability, energy, key, loudness, mode, speechiness, acousticness, instrumentalness, liveness, valence, tempo, and time signature are audio features provided by Spotify that describe various musical characteristics of each track.

3.8 Correlation Matrix

A correlation matrix is computed and visualized using a color-coded upper-triangle plot to identify linear relationships among the numeric variables in the cleaned dataset. This helps reveal patterns of association and potential multicollinearity between audio features.



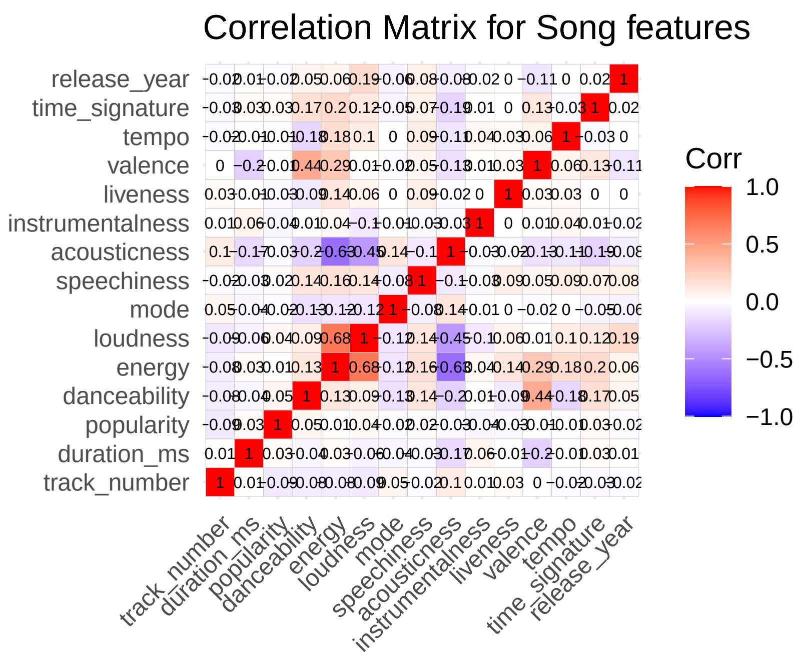
The correlation matrix reveals a moderate positive association between energy and loudness, suggesting that more energetic tracks tend to be louder. A notable negative correlation is observed between acousticness and energy, indicating that acoustic songs generally exhibit lower energy levels. Overall, the absence of strong correlations among most variables suggests low multicollinearity, supporting their joint inclusion in multivariate analyses.

4 Exploratory Data Analysis

To gain a comprehensive understanding of the dataset and uncover patterns relevant to the modeling phase, an exploratory data analysis (EDA) is conducted. This phase involves the systematic examination of the dataset's structure, distributions, and relationships among variables. We begin with summary statistics and visual inspections of key numeric features, followed by the analysis of correlations and potential multicollinearity. Subsequently, we investigate temporal trends, track characteristics, and the distribution of categorical variables such as explicit content and genre. This stepwise exploration helps identify data quality issues, potential outliers, and underlying trends that may influence or inform the subsequent application of machine learning techniques.

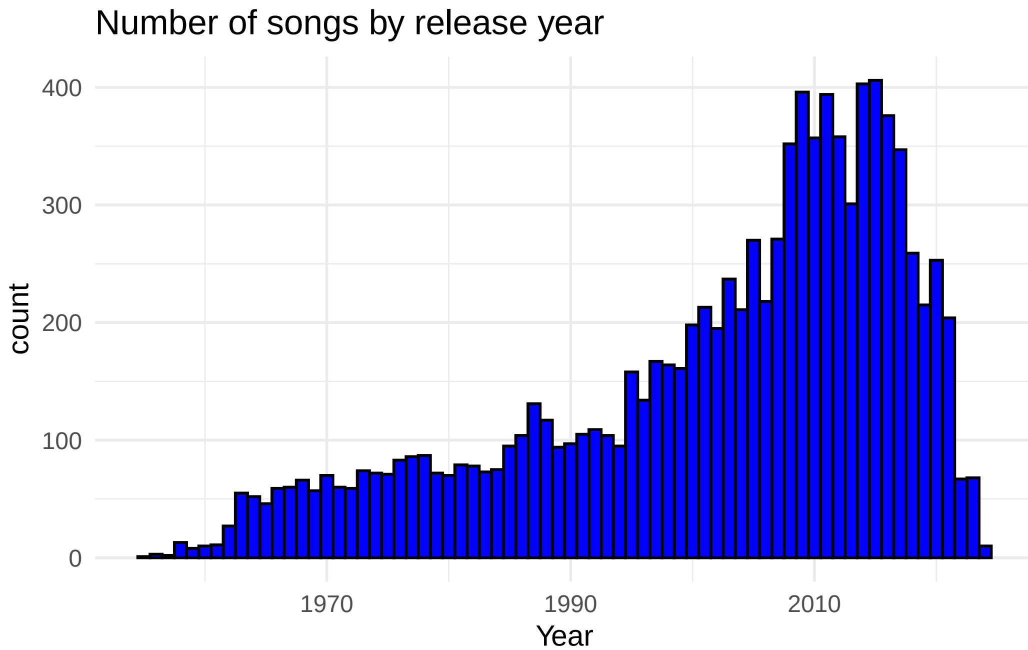
4.1 Correlation matrix

A correlation matrix was computed and visualized to examine linear relationships between all numeric features in the dataset, highlighting potential associations and redundancies among audio variables.



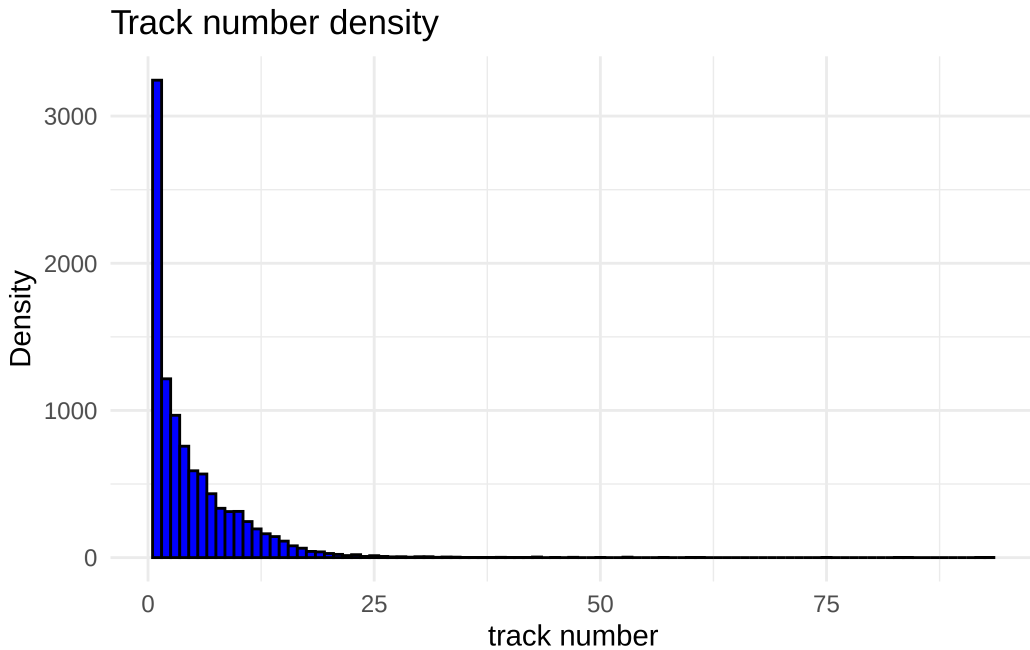
4.2 Variables analysis

4.2.1 Release year



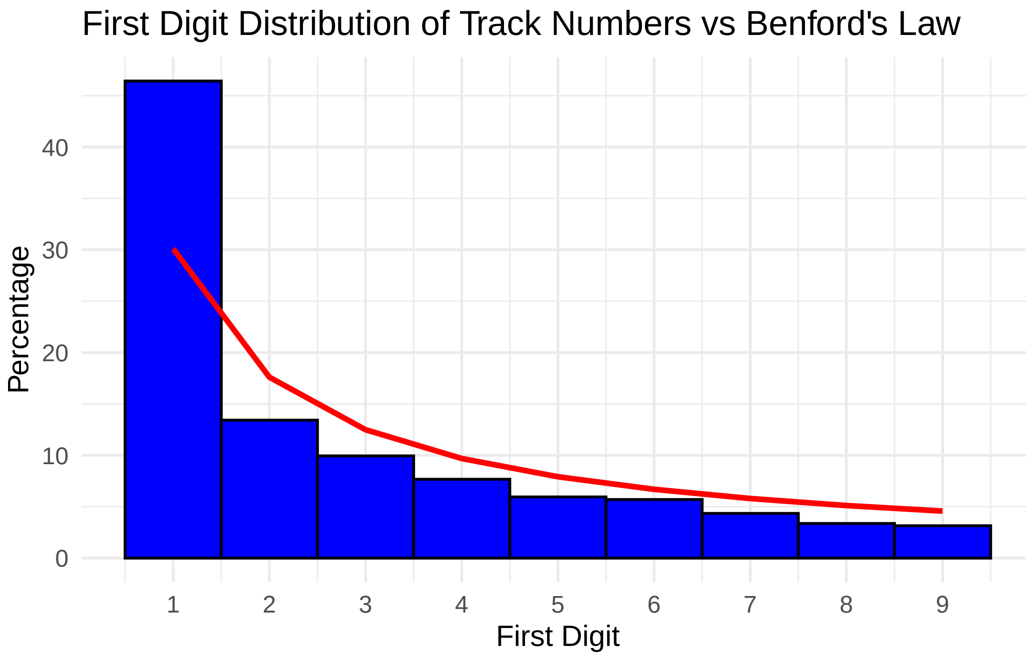
Release year is the dependent variable in this analysis, and it is considered at the level of yearly granularity. The histogram shows a strong concentration of songs in more recent years, indicating a temporal imbalance that may affect the representativeness of earlier decades in the modeling phase.

4.2.2 Track numbers



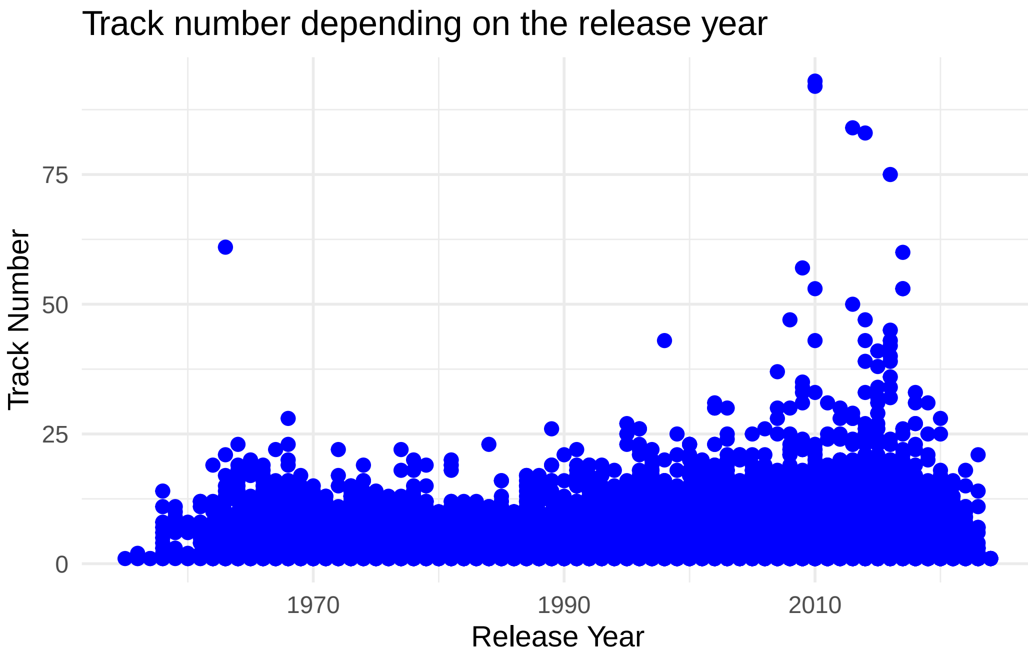
The distribution of track numbers is highly skewed toward lower values, indicating that most songs appear early in albums, while higher values likely reflect compilations or large tracklists.

It may be interesting to assess whether Benford's Law applies to the distribution of track numbers.



The distribution of the first digit in track numbers clearly deviates from the expected pattern described by Benford's Law. This discrepancy is likely due to structural constraints in how music albums are organized. Most releases—such as singles, EPs, and standard albums—contain a relatively small and fixed number of tracks, typically ranging from 1 to 15. As a result, lower digits, particularly 1, dominate the distribution, not because of a naturally logarithmic phenomenon but due to intentional sequencing and formatting practices in album production. This illustrates how domain-specific conventions can override general statistical laws in structured datasets.

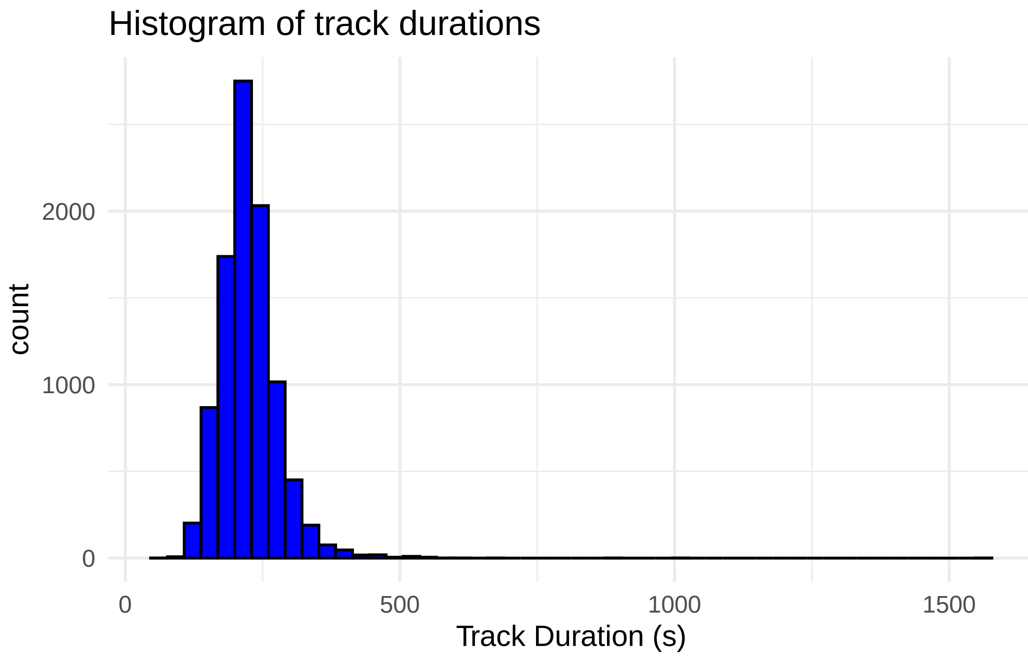
Lastly, the evolution of track numbers over time can be examined to assess whether album structure or track positioning has changed across decades, potentially reflecting shifts in music consumption formats or production practices.



The scatter plot shows no clear trend in the evolution of track numbers over time. While the majority of songs consistently appear in early album positions across decades, some outliers—particularly in more recent years—exceed typical album lengths, likely reflecting special editions, compilations, or digital releases with extended tracklists.

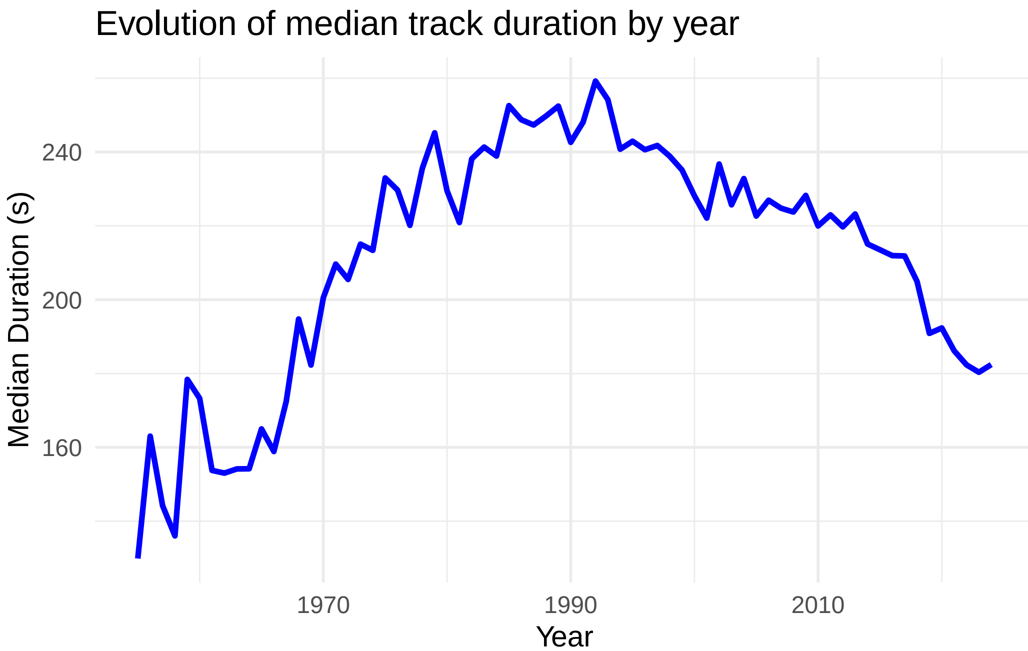
4.2.3 Track Duration

The distribution of track duration, expressed in milliseconds, is examined to determine typical song lengths and to identify potential outliers, including exceptionally short or long tracks, which may affect subsequent analyses.

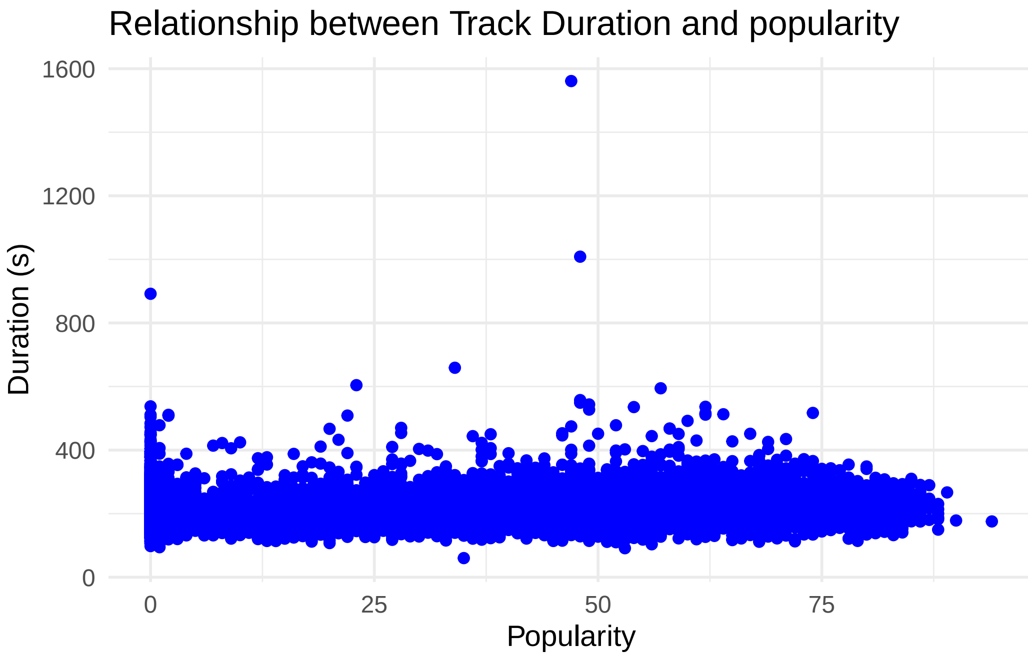


The histogram shows that the distribution of track durations is right-skewed, with most songs clustered around typical lengths and a limited number of extreme values representing unusually long tracks.

A line plot is used to visualize the evolution of median track duration over time. This allows identification of long-term trends, such as shifts in typical song lengths across decades, and highlights notable changes in production or consumption patterns.



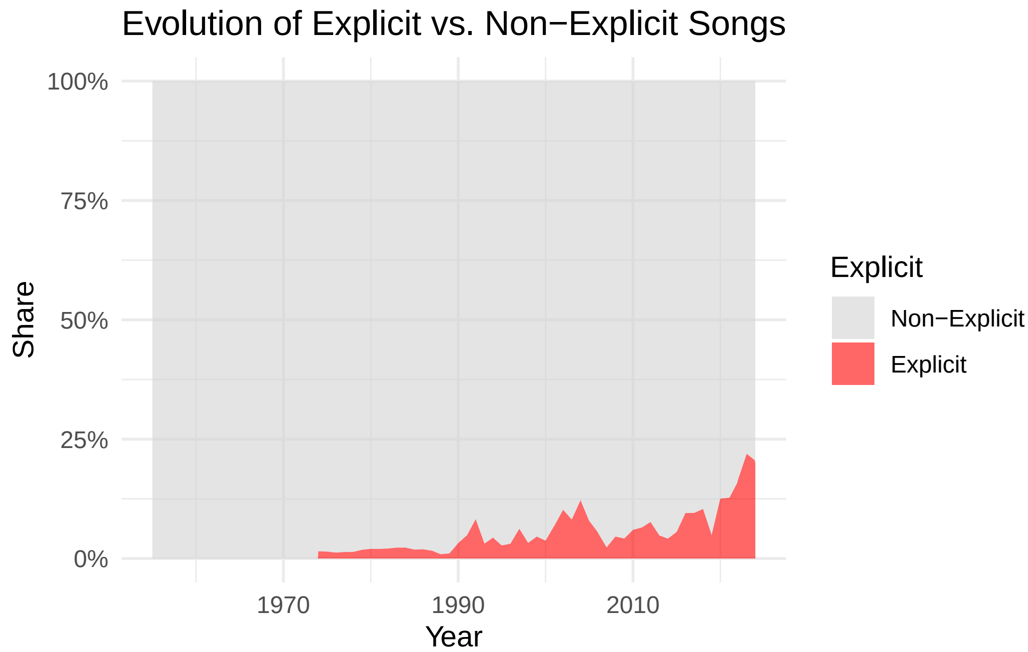
The plot shows that median track duration increased steadily throughout the second half of the 20th century, reaching a peak in 1992 at approximately 260 seconds (4 minutes and 20 seconds). A gradual decline follows, particularly from the 2010s onward. This recent downward trend is likely influenced by changes in digital consumption patterns, where streaming platforms incentivize shorter songs due to skip behavior and payout structures based on play counts.



The scatter plot shows no strong correlation between duration and popularity, though popular songs tend to have standard lengths, while very long tracks are generally less popular.

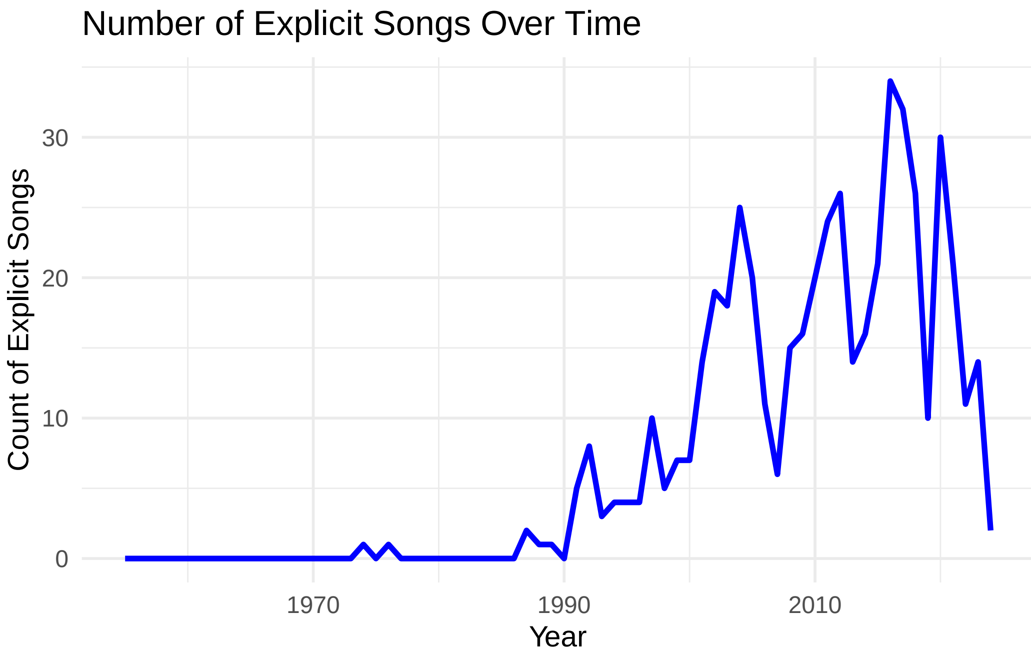
4.2.4 Explicit

The proportion of explicit versus non-explicit songs is analyzed over time to observe how the prevalence of explicit content has evolved across different release years.



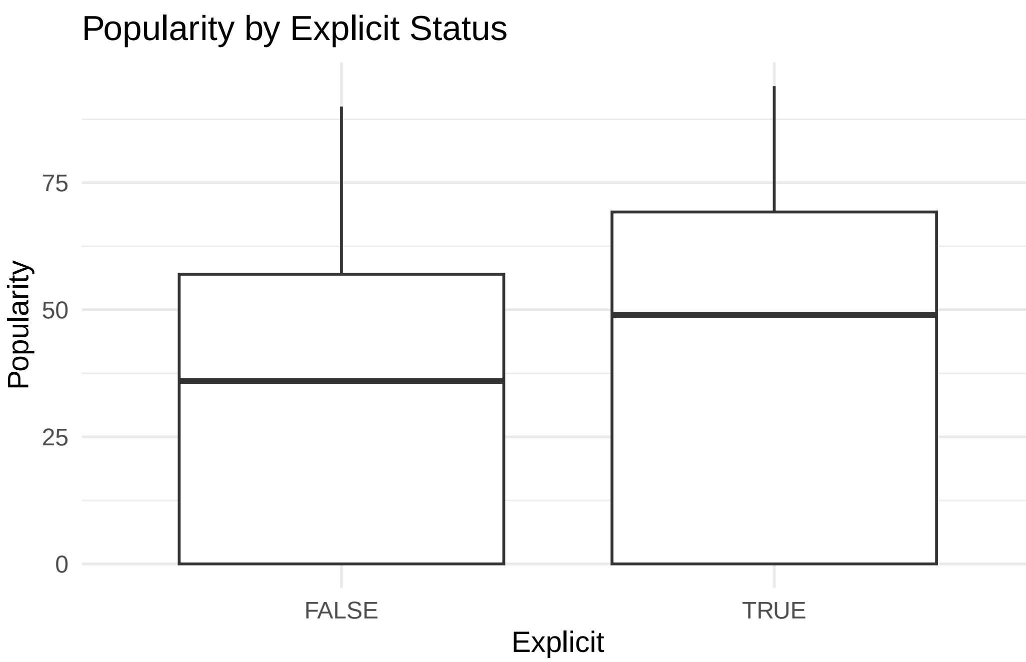
The analysis reveals a clear upward trend in the share of explicit songs over time, with a marked increase beginning in the early 2000s. This suggests a shift in lyrical content or labeling practices in the streaming era, where explicit content has become more common in popular music releases.

A line plot is now used to visualize the annual count of explicit songs over time, offering a more detailed view of their increasing presence in recent decades.

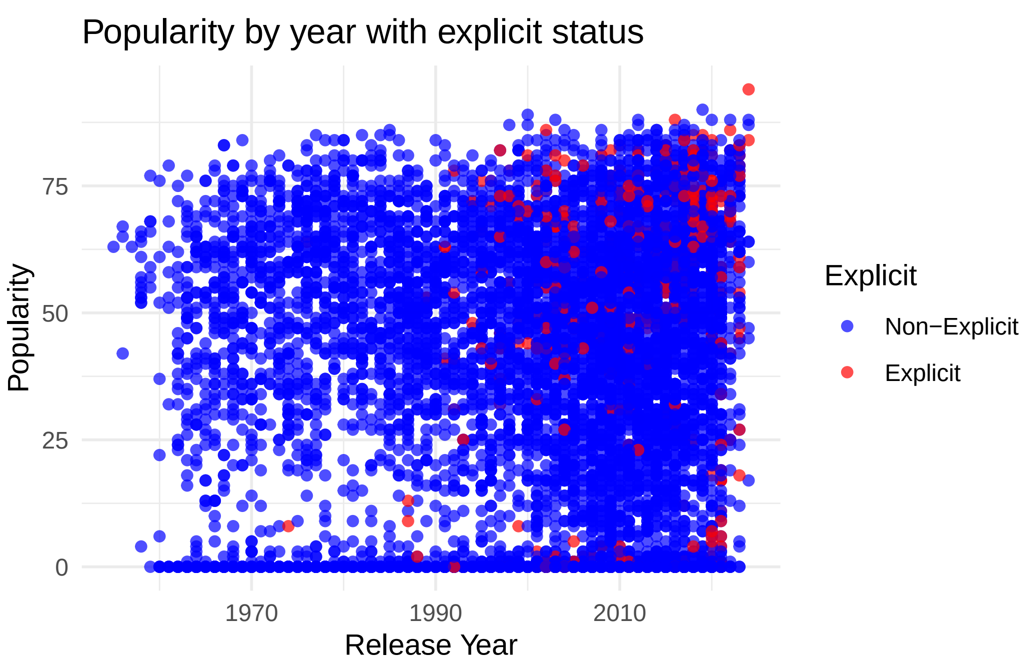


The line plot confirms a rising trend in the number of explicit songs over time, particularly from the early 2000s onward. However, this pattern should be interpreted cautiously, as the explicit label is assigned by content uploaders and may not consistently reflect the presence of explicit material. In some cases, tracks with potentially explicit content are also released in censored versions to ensure broader distribution, such as radio play.

A boxplot is used to compare the distribution of song popularity between explicit and non-explicit tracks, allowing for the assessment of whether explicit content is associated with higher or lower popularity levels.



To further explore the relationship between explicit content and popularity, a scatter plot is used to visualize popularity scores over time, differentiated by explicit status. This approach helps assess whether the observed higher popularity of explicit songs is inherently tied to their content or instead influenced by temporal release patterns.



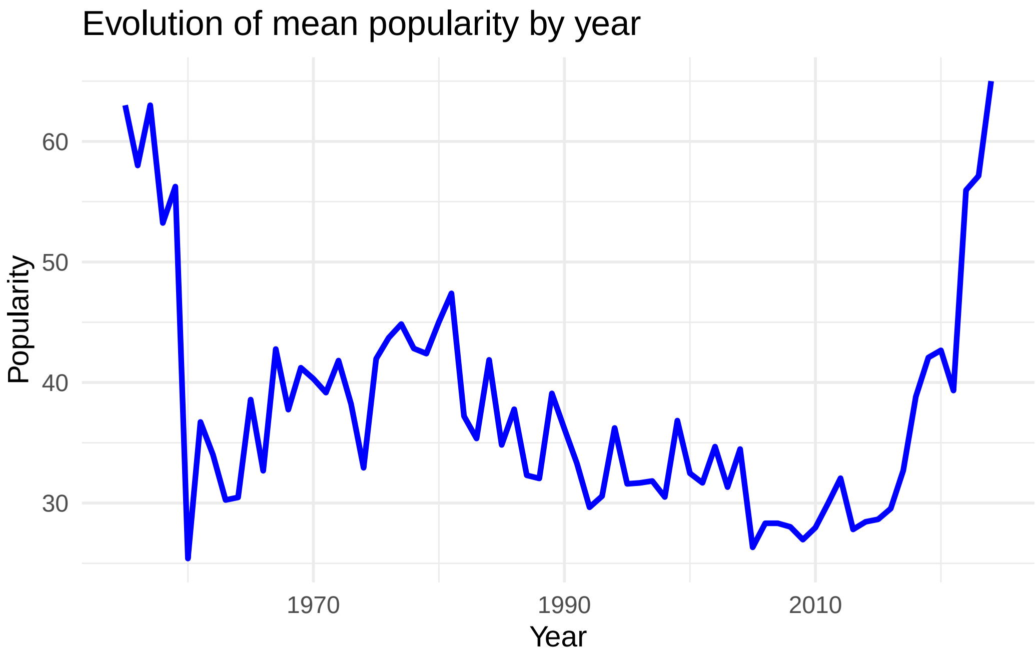
The scatter plot reveals that explicit songs (in red) are predominantly concentrated in more recent years and often exhibit high popularity scores. This supports the hypothesis that the observed popularity advantage of explicit tracks may be confounded by release date, as newer songs are both more likely to be explicit and to benefit from recency effects in popularity metrics.

4.2.5 Popularity

Popularity serves as a valuable proxy for a song's commercial success and audience reach. While the exact algorithm used to calculate Spotify's popularity score is not publicly disclosed, it is reasonable to assume that it reflects streaming frequency over a defined recent time window.

Track Name	Artist Names	Popularity
Espresso	Sabrina Carpenter	94
Cruel Summer	Taylor Swift	90
Yellow	Coldplay	89
Lose Control	Teddy Swims	89
Starboy	The Weeknd, Daft Punk	88
When I Was Your Man	Bruno Mars	88

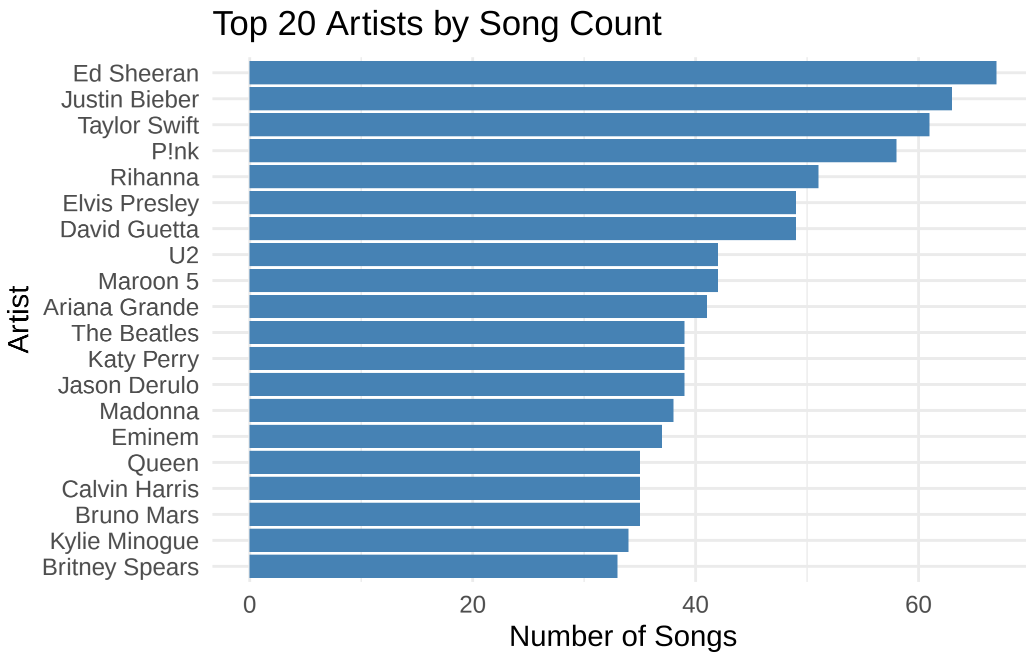
It is important to note that the dataset was last updated in October 2024, meaning that the popularity scores reflect a specific point in time and may have since changed. However, this temporal limitation does not affect the validity of the present analysis, as the values still offer a reliable snapshot for exploring general trends and patterns in music popularity.



The chart shows peaks in popularity for early and recent songs. While the high popularity of recent songs is to be expected, the elevated scores for tracks from the 1960s are more surprising. A possible explanation is that artists featured in rankings and playlists remain culturally significant. Artists such as Frank Sinatra, Elvis Presley or Johnny Cash have become timeless icons that continue to be cultural relevant and popular across generations.

4.3 Textual analysis

This section performs a basic text mining task by extracting and counting individual artist names from multi-artist tracks. The goal is to identify the top 20 most frequently appearing artists in the dataset.



This step identifies the top 10 artists with the highest average popularity scores, offering insights into which artists consistently produce well-received songs within the dataset.

Artist Name	Average Popularity
Benson Boone	87.0
a-ha	86.0
Teddy Swims	85.5
Alphaville	85.0
Nayer	85.0
Jung Kook	84.0
Nate Ruess	84.0
Tyla	84.0
D-Block Europe	83.0
cassö	83.0

The following analysis ranks artists by their average track duration, highlighting those associated with longer musical compositions. Such patterns may reflect stylistic tendencies common in certain genres, including progressive rock, ambient, or experimental music.

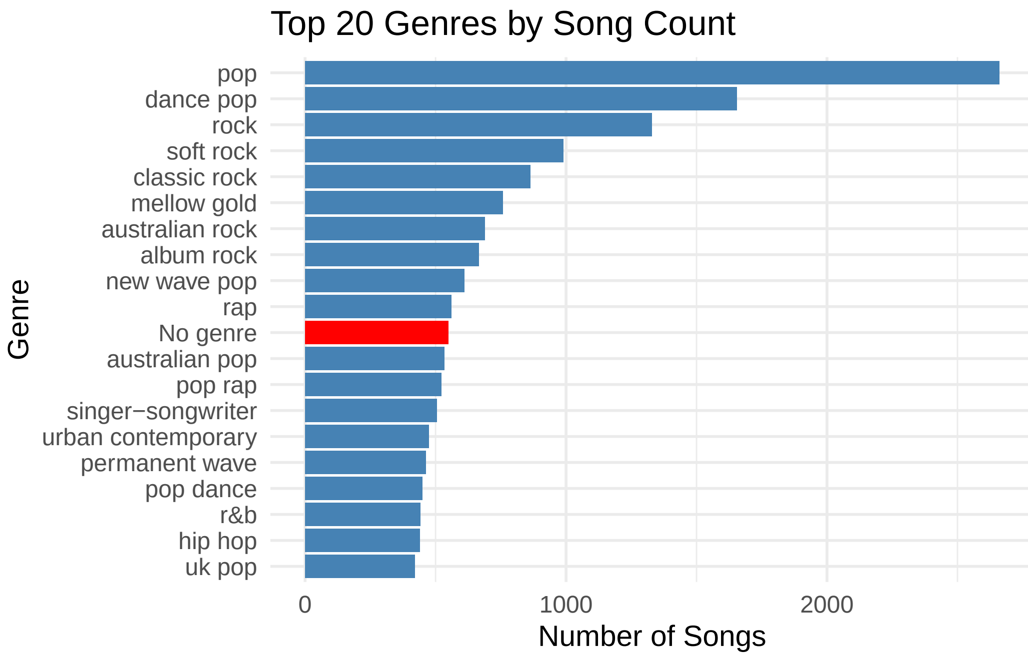
Artist Name	Avg. Song duration
The Bucketheads	891720

Mike Oldfield	633756
The Stone Roses	594453
The S.O.S Band	549000
Joe Walsh	536226
Curtis Mayfield	535333
Faithless	527240
Rollo Armstrong	527240
Sister Bliss	527240
Supercharge	523004

The following analysis focuses on the artists with the lowest average track durations, aiming to reveal trends related to brevity in musical production, which may be influenced by genre conventions or platform-driven listening behaviors.

Artist Name	Avg. Song duration
Liam Lynch	91226
Maurice Williams & The Zodiacs	97506
Fakebitcheshero	103026
The Swinging Blue Jeans	106066
Jerry Lee Lewis	111535
Hank Ballard & The Midnighters	111826
The Clovers	112200
Paul Russell	114233
Clyde McPhatter	116240
The String-A-Longs	116360

A similar type of analysis is now performed based on musical genre, in order to examine how average song characteristics vary across different stylistic categories.



Genres	Average Popularity
deep dance pop	85
melodic drill	83
indie rock italiano	82
italian pop	82
birmingham grime	80
indie r&b	79
bubblegrunge	78
sad lo-fi	78
sad rap	78
float house	77

_________________________ to be continued:

5 Machine Learning Methods

5.1 Unsupervised Learning

Several methods can be used for unsupervised learning, among them, K-means clustering, Hierarchical Clustering and PCA were performed, due to somewhat disappointing results for

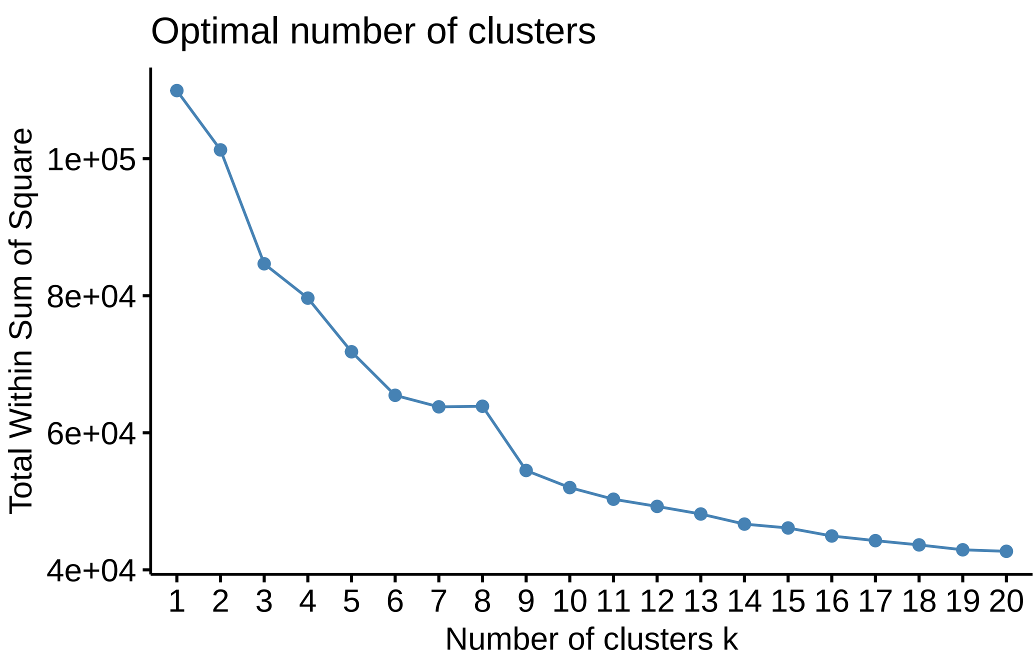
Hierarchical Clustering and PCA, only K-means clustering was kept. Nevertheless, the code for other methods can be found in the Appendix, presented as is, without result analysis.

5.1.1 Clustering

5.1.2 K-means clustering

We'll start by fitting a k-means clustering, using the R package *factoextra*. Only numeric variables were chosen to conduct this clustering experiment, and more specifically only track features were selected, these include: **danceability, energy, loudness, mode, speechiness, acousticness, instrumentalness, liveness, valence, tempo, time_signature**. The decision to only take up an interest in these variables is because these were created by spotify, who does not release any information regarding these features and the calculation behind. One lead may be that these features are caclulated through some sort of clustering to help understand different features and music genres trough it. A clustering analysis of tracks is therefore quite fitting.

The first step in clustering is to determine the optimal number of clusters. The metric used to choose the number of clusters is Total within-cluster sum of squares (TWSS). An elbow must be found, which will determine the choice of clusters.



Based on the graph, it was decided to chose 9 as the number of clusters. An analysis of these clusters can be found in result analysis section.

5.2 Supervised Learning Methodology

5.2.1 Random Forest Balanced

5.2.2 Regression

5.2.3 Random Forest with undersampling

```
# A tibble: 9 x 2
  decade     n
   <dbl> <int>
1   2010  3416
2   2000  2561
3   1990  1294
4   1980   916
5   1970   734
6   2020   602
7   1960   443
8   1950    27
9      0     2
```

```
# A tibble: 7 x 2
  decade_group     n
  <chr>        <int>
1 2010          3416
2 2000          2561
3 1990          1294
4 1980           916
5 1970           734
6 2020           602
7 1950-1960      470
```

```
# A tibble: 7 x 2
  decade_group     n
  <chr>        <int>
1 1950-1960      470
2 1970           470
3 1980           470
4 1990           470
5 2000           470
6 2010           470
7 2020           470
```

```
[1] 0.4346505

[1] 0.4356636

Confusion Matrix and Statistics

           Reference
Prediction  1950-1960 1970 1980 1990 2000 2010 2020
  1950-1960        89   32   14   20   11   16   14
  1970             20   45   16   17   11    3    5
  1980              9   33   68   30   15    4    6
  1990              4   16   20   40   13    9    6
  2000              4    7   11   23   47   28    8
  2010              0    8    5   11   22   53   26
  2020              7    4    4    6    9   30   88

Overall Statistics

               Accuracy : 0.4357
                 95% CI : (0.4045, 0.4673)
    No Information Rate : 0.155
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.3421

 Mcnemar's Test P-Value : 0.0001625

Statistics by Class:

                     Class: 1950-1960 Class: 1970 Class: 1980 Class: 1990
Sensitivity                   0.66917     0.31034      0.4928     0.27211
Specificity                   0.87471     0.91449      0.8857     0.91905
Pos Pred Value                0.45408     0.38462      0.4121     0.37037
Neg Pred Value                0.94437     0.88506      0.9148     0.87827
Prevalence                    0.13475     0.14691      0.1398     0.14894
Detection Rate                0.09017     0.04559      0.0689     0.04053
Detection Prevalence          0.19858     0.11854      0.1672     0.10942
Balanced Accuracy             0.77194     0.61242      0.6893     0.59558
                     Class: 2000 Class: 2010 Class: 2020
Sensitivity              0.36719      0.3706     0.57516
Specificity              0.90570      0.9147     0.92806
Pos Pred Value           0.36719      0.4240     0.59459
```

```
Neg Pred Value           0.90570      0.8956     0.92253
Prevalence               0.12969      0.1449     0.15502
Detection Rate           0.04762      0.0537     0.08916
Detection Prevalence     0.12969      0.1266     0.14995
Balanced Accuracy        0.63645      0.6427     0.75161
```

Accuracy (0.410): correctly predict the decade 41% of the time

Kappa (0.312): Fair agreement beyond chance, barely acceptable

P-value (< 2e-16): Statistically better than random guessing

No Info Rate (14.3%): Baseline accuracy if you predicted the majority

5.2.4 Random Forest on median year

```
[1] 2007
```

The median song was release in 2007. We are going to perform a random forest on whether the song was release before or after this year.

```
[1] 0.7638819
```

```
Confusion Matrix and Statistics

          Reference
Prediction after before
    after    683    211
    before   261    844

               Accuracy : 0.7639
                 95% CI : (0.7446, 0.7824)
    No Information Rate : 0.5278
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.525

 Mcnemar's Test P-Value : 0.02411

            Sensitivity : 0.7235
            Specificity : 0.8000
         Pos Pred Value : 0.7640
```

```
         Neg Pred Value : 0.7638
             Prevalence : 0.4722
         Detection Rate : 0.3417
   Detection Prevalence : 0.4472
      Balanced Accuracy : 0.7618

       'Positive' Class : after
```

Accuracy (0.764): correctly classify ~76% of the songs.

Kappa (0.525): moderate agreement beyond chance, pretty good.

Balanced Acc. (0.762): handles class imbalance, very healthy.

Sensitivity (0.724): correctly detect 72.4% of songs after 2007.

Specificity (0.800): correctly detect 80% of songs before 2007.

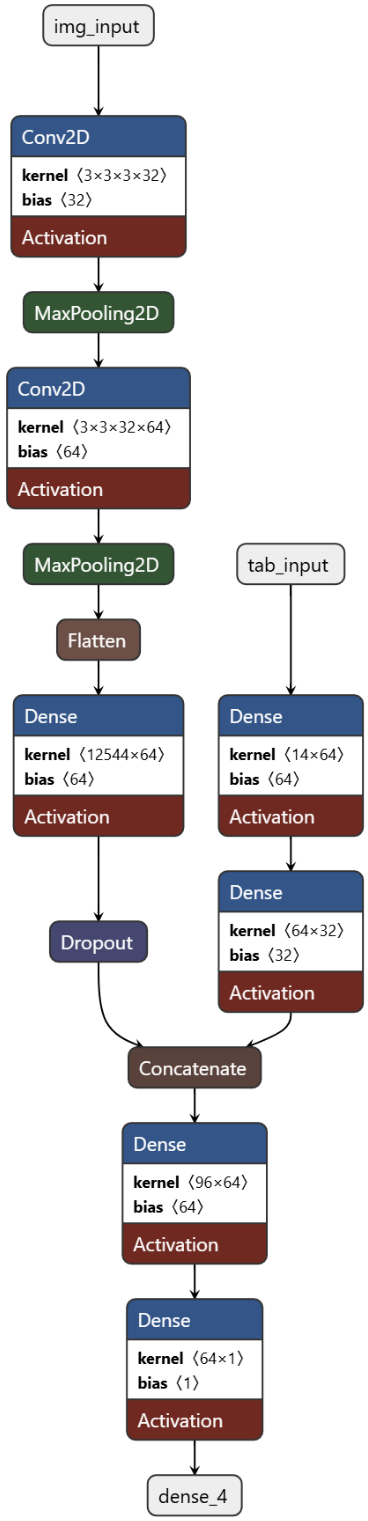
5.2.5 Neural Networks

While Random Forest models showed satisfying results when it came to classification of track decades, attempts involving regression tasks to predict the year saw issues. It was therefore decided to attempt to fit a neural network to predict the release year using tabular data as well as album covers.

5.2.5.1 Model Building

The model was built and fitted using python and the **Keras** library, allowing to easily build neural networks.

The decision was made to train both inputs, the image data in a CNN and the tabular data in a DNN. the layers were then merged. into one hidden layer thereby taking in all the information from both layers and finally making a prediction.



The model summary can be seen below:

To build the model, the release year was normalized, Therefore enabling predictions bounded between 1955 and 2024, the minimum and maximum release years. The final output layer uses a sigmoid function, making sure that forecasts are always bounded and that the model will not predict impossible years such as 2026 for example.

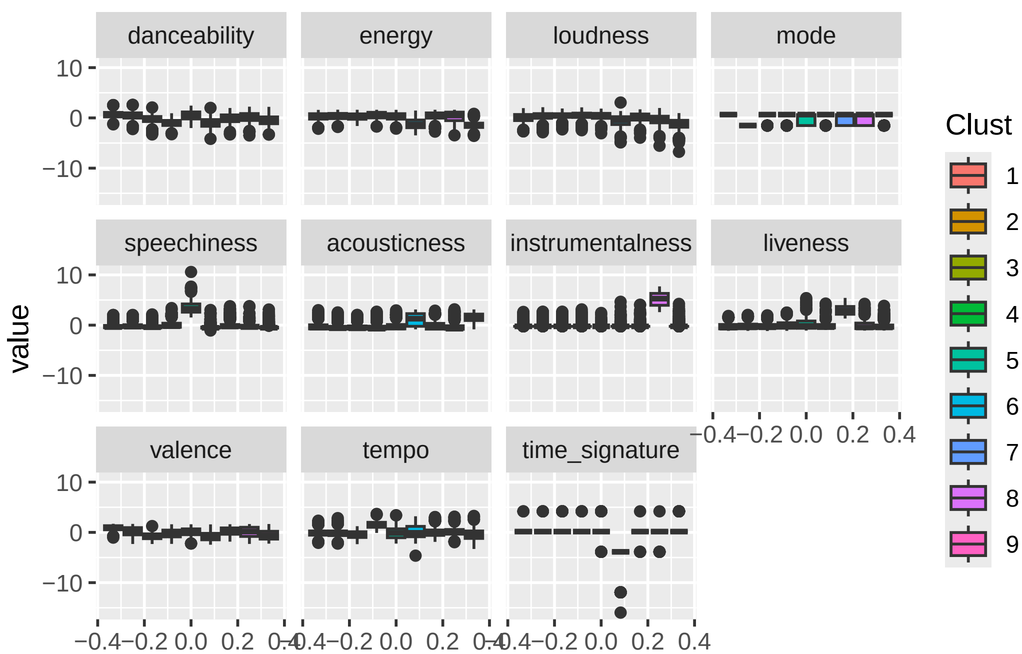
6 Results Discussion

6.1 Unsupervised Learning

6.1.1 Clustering

Based on the methodology discussed above, k-means clustering was run with 9 centers.

Boxplots are created to inspect these clusters.



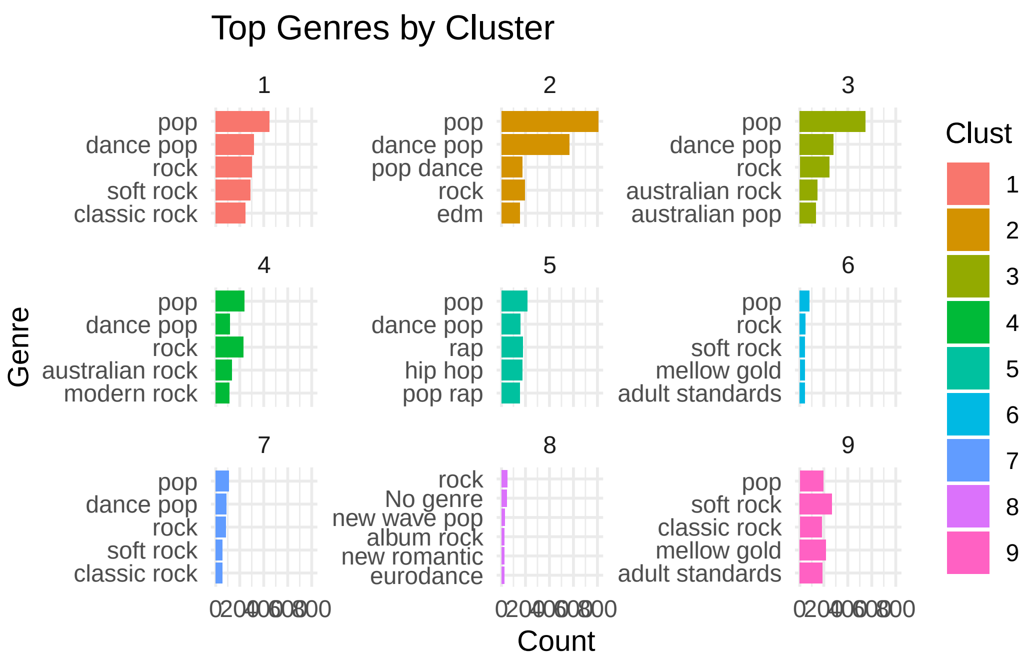
Some patterns seem to emerge from this clustering. Cluster 9 appears to have lower values for mode compared to others. Cluster 4 has higher speechiness, while clusters 5 and 6 show slightly more acousticness. Cluster 5 also has much higher instrumentalness, and cluster 7 has a significantly lower time signature than the rest.

Cluster	count	Mean Year	Median Year
1	2183	1999.061	2003

2	2055	2005.479	2009
3	1624	2006.442	2010
4	1038	2002.685	2007
5	488	2008.469	2010
6	397	1999.111	2003
7	489	2000.470	2006
8	303	1999.696	2002
9	1416	1994.894	1997

Looking at the release year, some patterns emerge. In terms of count, clusters **1** and **9** contain the most observations. Cluster 9 stands out for having the lowest median release year overall. It is also important to note that differences between the mean and median values—as well as the overall shift toward recent years—are likely due to an imbalance in the dataset, with more recent tracks represented.

The same analysis can be applied to identify genres within each cluster.

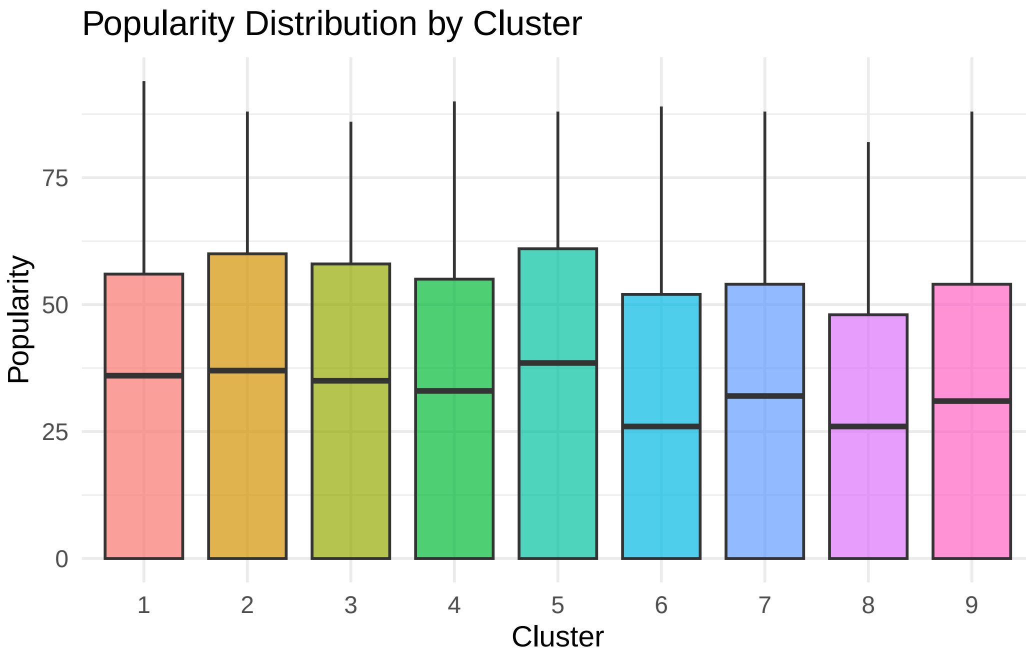


While this graph is interesting, it does not account for the overall distribution of genres. For instance, *Pop* and *Dance Pop* are by far the most present genres in the dataset with respectively *2660* and *1655* instances. To address this, we compute the relative proportion of each genre $rel_i n_c luster = cluster_n / total_n$, for example, pop appears 599 times in Cluster 1 out of 2660 total appearances. Only genres with at least 500 appearances are considered to focus only on main genres instead of niche ones.

Clust	artist_genres	rel_in_cluster	total_n
1	new wave pop	0.3371522	611
1	album rock	0.2987988	666
2	dance pop	0.3413897	1655
2	pop	0.3033835	2660
3	australian pop	0.2546816	534
3	australian rock	0.2122093	688
4	australian rock	0.1976744	688
4	rock	0.1753198	1329
5	rap	0.3172906	561
5	pop rap	0.2925430	523
6	No genre	0.0601093	549
6	singer-songwriter	0.0554455	505
7	singer-songwriter	0.0673267	505
7	classic rock	0.0660487	863
8	No genre	0.0819672	549
8	new wave pop	0.0458265	611
9	singer-songwriter	0.2910891	505
9	mellow gold	0.2892999	757

We identify the top two genres for each cluster.
- **Cluster 1**: Dominated by Pop songs.
- **Cluster 2**: Rock is the most frequent genre.
- **Cluster 3**: Features Singer-Songwriter and Mellow Gold, often from artists like Eric Clapton, George Harrison, or Bob Dylan—suggesting influences from folk, country, and soft rock.
- **Cluster 4**: Primarily contains rap music.
- **Cluster 5**: Largely composed of tracks labeled "No genre" and New Wave, likely reflecting more niche or lower-popularity songs.
- **Cluster 6**: Again features Soft Rock and Mellow Gold, indicating some similarity with Cluster 3.
- **Cluster 7**: Includes many songs with no genre and some Singer-Songwriter.
- **Cluster 8**: Appears to be focused on Australian music.
- **Cluster 9**: Dominated by Pop and Dance Pop.

This clustering also reveals trends related to popularity. Some clusters may represent more popular songs than others. This can be visualized easily using boxplots.



Surprisingly, the differences in popularity between clusters are not as pronounced as one might expect. Nevertheless, Cluster 4—dominated by rap—appears to be the most popular. Clusters 5 and 7, which include many songs with "No genre" labels, seem to be the least popular.

6.1.2 Conclusion

To conclude, one key takeaway from this clustering analysis is that the large volume of tracks likely affects the clarity of the results. Still, meaningful insights can be drawn—especially regarding genres. The analysis suggests that genres are, at least partially, distinguishable using Spotify's audio features.

Release year and popularity offer additional context. The main insight is that our dataset contains significantly more recent songs, and that these tend to be more popular. Currently, genres such as *Pop*, *Dance Pop*, *Rap*, and *Hip-Hop* dominate popular music.

6.2 Supervised Learning

6.2.1 Neural Network

The model was trained with 20 epochs and a batch size of 32 returned the following metrics:

Training MSE: 0.0274 Training MAE: 0.1133

Test MSE: 0.0270 Test MAE; 0.1139

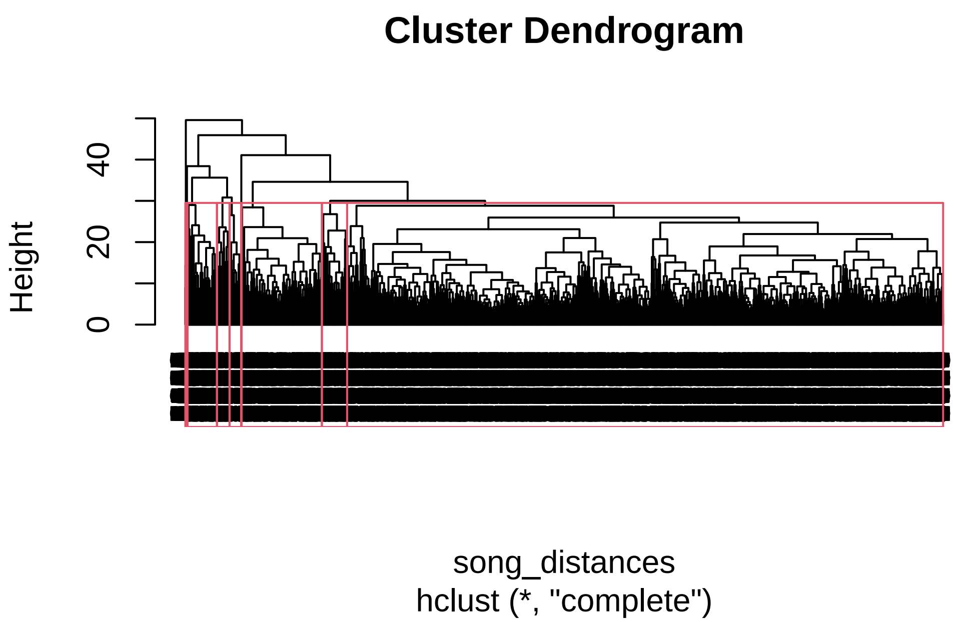
Test R^2 score: 0.4539

These results show no sign of overfitting, regarding the metrics themselves, one should keep in mind that these metrics are for normalized years. Therefore, a mean absolute error of 0.1139 over the 69(2024-1955) years of study means that the model is on average off by 7.8 years. The R-squared score shows that while the model may capture some of the variance, a big part is due to other factors.

7 Conclusive discussion

8 Appendix

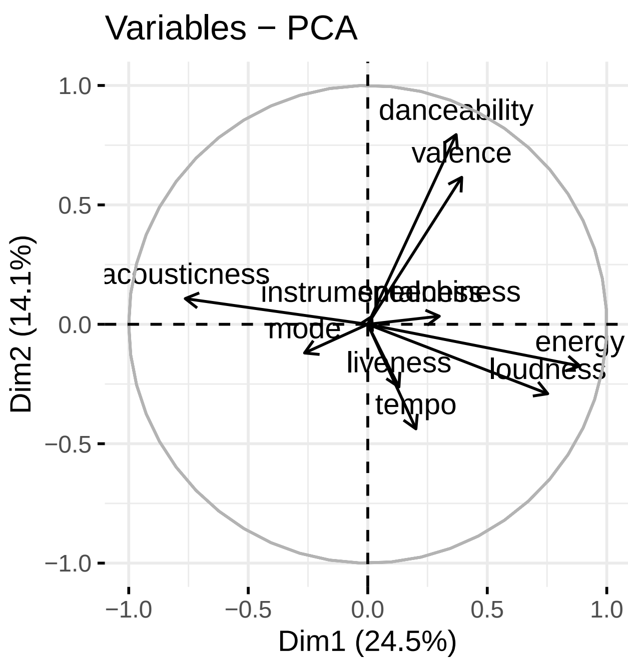
8.1 Hierarchical clustering

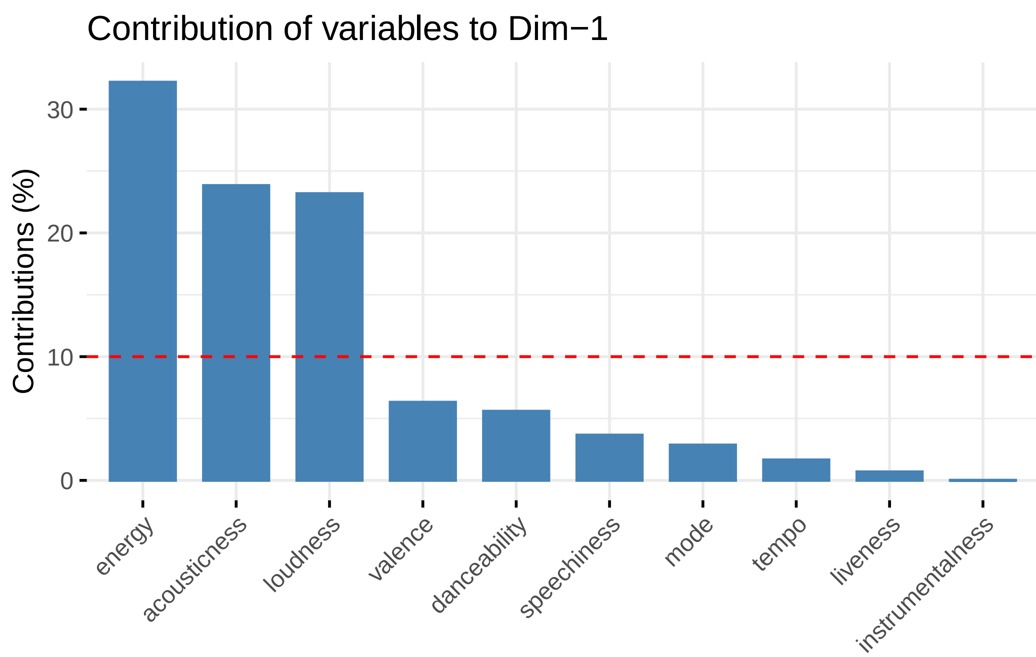


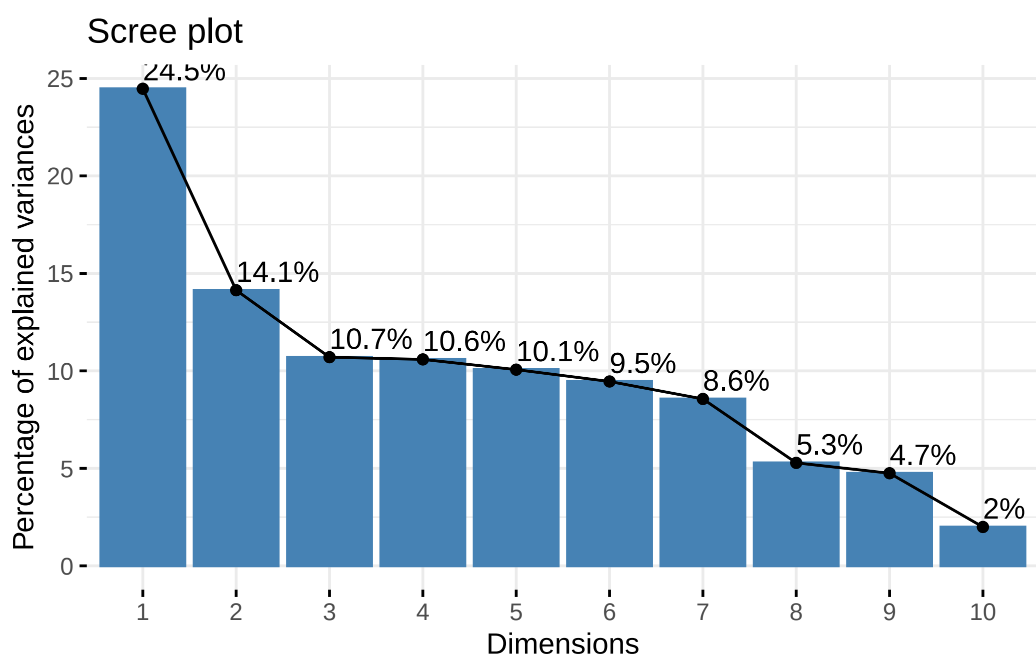
Due to poor performance from the hierarchical clustering, the decision was made to only focus on k-means clustering. In subsequent studies, a deeper focus on this clustering technique could be conducted.

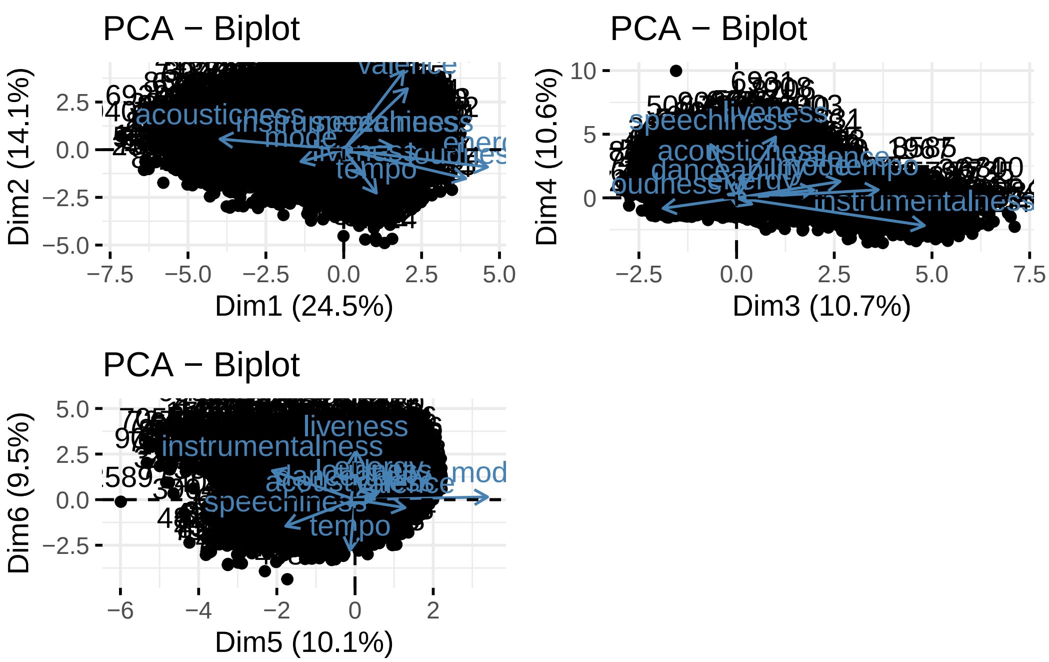
8.2 Principal component analysis

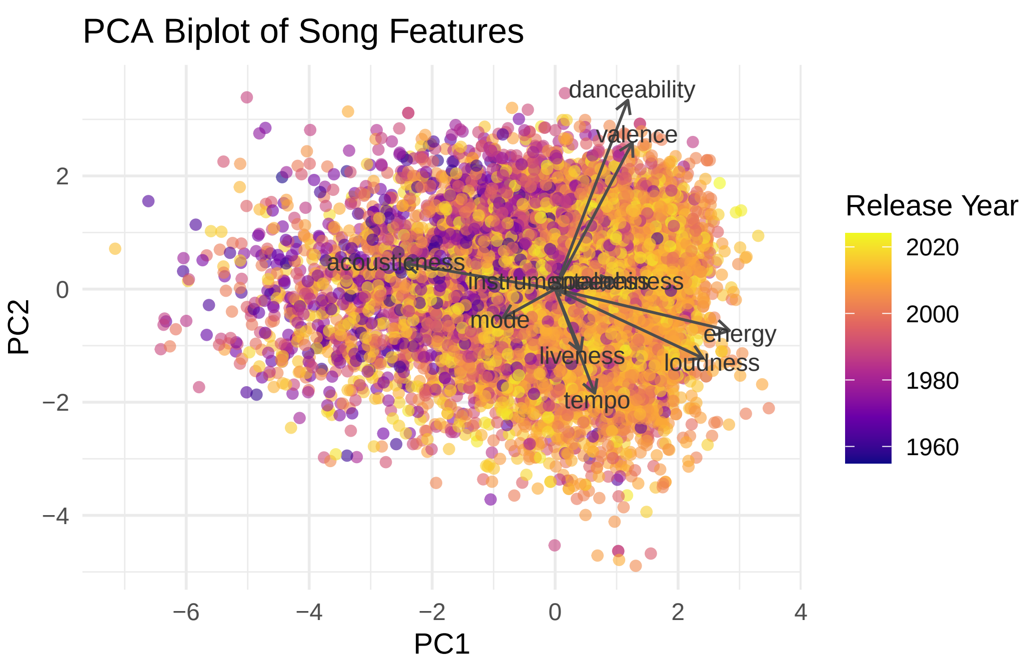
```
**Results for the Principal Component Analysis (PCA)**
The analysis was performed on 9443 individuals, described by 10 variables
*The results are available in the following objects:

   name               description
1  "$eig"             "eigenvalues"
2  "$var"             "results for the variables"
3  "$var$coord"       "coord. for the variables"
4  "$var$cor"         "correlations variables - dimensions"
5  "$var$cos2"        "cos2 for the variables"
6  "$var$contrib"     "contributions of the variables"
7  "$ind"             "results for the individuals"
8  "$ind$coord"       "coord. for the individuals"
9  "$ind$cos2"        "cos2 for the individuals"
10 "$ind$contrib"     "contributions of the individuals"
11 "$call"            "summary statistics"
12 "$call$centre"     "mean of the variables"
13 "$call$ecart.type" "standard error of the variables"
14 "$call$row.w"      "weights for the individuals"
15 "$call$col.w"      "weights for the variables"
```




Contribution of variables to Dim−1
Contributions (%)
0
10
20
30
energy
acousticness
loudness
valence
danceability
speechiness
mode
tempo
liveness
instrumentalness


Scree plot
Percentage of explained variances
0
5
10
15
20
25
24.5%
14.1%
10.7%
10.6%
10.1%
9.5%
8.6%
5.3%
4.7%
2%
1
2
3
4
5
6
7
8
9
10
Dimensions


PCA − Biplot
Dim1 (24.5%)
Dim2 (14.1%)
acousticness
mode
tempo
PCA − Biplot
Dim3 (10.7%)
Dim4 (10.6%)
speechiness
liveness
instrumentalness
PCA − Biplot
Dim5 (10.1%)
Dim6 (9.5%)
liveness
instrumentalness
speechiness
tempo


PCA Biplot of Song Features
danceability
valence
acousticness
mode
energy
loudness
liveness
tempo
PC1
PC2
Release Year
2020
2000
1980
1960

9 References

Beesa, P., Naregavi, V., Imandar, J., & Thatte, S (2023). *Songs popularity analysis using Spotify data: An exploratory study.* In Vidhyayana – An International Multidisciplinary Peer-Reviewed E-Journal, Volume 8, Special Issue 7, 4th National Student Research Conference on "Innovative Ideas and Invention in Computer Science & IT with its Sustainability" (pp. 211–223). https://www.researchgate.net/publication/384286217

Choi, K., Fazekas, G., Sandler, M., & Cho, K. (2017). *Convolutional recurrent neural networks for music classification.* 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2392–2396. https://doi.org/10.1109/ICASSP.2017.7952585

Friberg, A., Erkut, C., & Bresin, R. (2011). *Perceptual ratings of musical parameters.* Proceedings of the International Conference on Music Perception and Cognition (ICMPC), 193–198.

Gómez-Cañón, J. S., Cano, E., Eerola, T., Herrera, P., Hu, X., Yang, Y.-H., & Gómez, E. (2021). Music emotion recognition: Toward new, robust standards in personalized and context-sensitive applications. *EEE Signal Processing Magazine, 38*(6), 106–114.

Herremans, D., Martens, D., & Sörensen, K. (2019). *Dance hit song prediction.* University of Antwerp Operations Research Group Applied Data Mining Research Group, University of Antwerp.

Interiano, M., Kazemi, K., Wang, L., Yang, J., & Komarova, N. L. (2018). *Musical trends and predictability of success in contemporary songs in and out of the top charts.* Royal Society Open Science, 5(5), 171274. https://doi.org/10.1098/rsos.171274

Jan Van Balen, L., Burgoyne, J. A., Bountouridis, D., Müllensiefen, D., & Honingh, A. (2015). *Corpus analysis tools for computational hook discovery.* Proceedings of the 17th ISMIR Conference, 227–233.

Kim, Y. E., Schmidt, E. M., Migneco, R., Morton, B. G., Richardson, P., Scott, J., … & Turnbull, D. (2010). *Music emotion recognition: A state of the art review.* Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR), 255–266.

Mauch, M., MacCallum, R. M., Levy, M., & Leroi, A. M. (2015). *The evolution of popular music: USA 1960–2010.* Royal Society Open Science, 2(5), 150081. https://doi.org/10.1098/rsos.150081

Müller, M., Ewert, S., & Grosche, P. (2010). *Musical structure analysis using audio features.* In Müller, M. (Ed.), Fundamentals of Music Processing (pp. 135–154). Springer. https://doi.org/10.1007/978-3-319-21945-5_6

Pachet, F. (2008). *Hit song science is not yet a science.* Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR), 355–360.

Schedl, M., Gómez, E., & Urbano, J. (2015). *Music information retrieval: Recent developments and applications.* Foundations and Trends in Information Retrieval, 8(2-3), 127–261. https://doi.org/10.1561/1500000042

Serrà, J., Corral, Á., Boguñá, M., Haro, M., & Arcos, J. L. (2012). *Measuring the evolution of contemporary Western popular music.* Scientific Reports, 2(1), 521. https://doi.org/10.1038/srep00521

Spotify. (2023). *How Spotify's popularity score works.* Spotify Developer Documentation. Retrieved from https://developer.spotify.com/