Comparing Song Audio Features to Rankings on The Billboard Hot 100

Kevin Carr 501150122

Supervisor: Ceni Babaoglu, PhD

Date: October, 2022

Table of Contents

[Abstract 3](#_Toc117242108)

[Introduction 4](#_Toc117242109)

[Literature Review 4](#_Toc117242110)

[Data Description 7](#_Toc117242111)

[Data Sources 7](#_Toc117242112)

[Data Pre-Processing and Organisation 7](#_Toc117242113)

[Quality Assurance 10](#_Toc117242114)

[File and Calculation Locations 10](#_Toc117242115)

[Descriptive Statistics 11](#_Toc117242116)

[Audio Feature Descriptions 13](#_Toc117242117)

[Histograms 14](#_Toc117242118)

[Historical Changes in Audio Features 19](#_Toc117242119)

[Correlation Analysis 23](#_Toc117242120)

[Analysis of Genres 27](#_Toc117242121)

[Project Approach 28](#_Toc117242122)

[Data Mining 28](#_Toc117242123)

[Predictive Analytics 29](#_Toc117242124)

[Outline of Methodology 29](#_Toc117242125)

[References 30](#_Toc117242126)

[Attachments 31](#_Toc117242127)

# Abstract

The central problem in this project is to utilise publicly available audio feature data to predict whether or not a song is likely to appear on the Billboard Hot 100 charts. Music streaming services employ data models to characterise audio features for songs. This data is provided and publicly available for multiple streaming services, notably Spotify. The Billboard Hot 100 has been a music industry standard for approximately 70 years. The Billboard Hot 100 contains weekly rankings for songs which are based on sales, plays, and surveys.

Audio feature data and Billboard Hot 100 chart data for this project have been obtained from multiple sources and combined. Datasets were found using the Google dataset search engine. In cases where audio feature data for Billboard Hot 100 songs were not in available datasets, missing audio features were obtained using the Spotify API . Audio features were found for approximately 75% of songs in the Billboard Hot 100 list, as well as an approximately 10 million additional songs.

For this project, two main techniques will be employed, namely data mining and predictive analytics. First, data mining and knowledge discovery will be used to explore the data, cluster audio features, and determine correlations between audio features. Second, predictive analytics will be used to attempt to build a predictive model using the data. By utilising knowledge discovered during the data mining phase of the project, predictive analysis will be broken into sets of clustered genres or songs with similar audio features. It is anticipated that this segmentation will improve predictive accuracy and lead to further insights.

# Introduction

Music streaming services employ data models to characterise audio features for songs, and use this data to recommend songs and playlist to their listeners. This data is provided and publicly available for multiple streaming services, notably Spotify (Spotify, n.d).

The Billboard Hot 100 has been a music industry standard for approximately 70 years (Wikipedia, 2022). The Billboard Hot 100 contains weekly rankings for songs which are based on sales, plays, and surveys.

For this project, two main techniques will be employed, namely data mining and predictive analytics. First, data mining and knowledge discovery will be used to explore the data, cluster audio features, and determine correlations between audio features. Secondly, predictive analytics will be used to attempt to build a predictive model. This analysis has the potential to predict future trends in music or the performance of an individual song. These predictions could be useful to musicians or producers attempting to optimise success, or listeners looking for something new.

# Literature Review

The primary focus of this literature review was to gather understanding on previous research related to music clustering techniques and the prediction of popularity for songs, especially in cases where song audio features were used for clustering or prediction. Google Scholar (Google Scholar, n.d.), PaperDigest.org (Paper Digest, n.d.), and Elicit.org (Elicit, n.d.) were used to investigate and gather research materials.

Popularity can be defined in numerous ways. In this study, popularity will simply be considered to be an appearance on the Billboard Hot 100 charts. The magnitude of this popularity may be further be defined using total weeks on the chart, or top rank on the chart (Lee et. al., 2018). Using appearance on the Billboard Hot 100 as a metric for popularity has been used in other similar studies (Reiman et.al., 2018). Another commonly used popularity metric is ‘popularity’ as defined in the Spotify API (Kim, 2021; Gao, 2021). Since this definition of popularity does not include historical data prior to widespread Spotify use, this study focusses on the Billboard Hot 100 charts.

Audio features used in music classification have evolved through various stages. MIDI (Musical Instrument Digital Interface) format musical notation has been used to cluster music into categories (Cilibrasi et. al., 2004; Cataltepe et. al., 2007a). Low-level audio features such as mel-frequency cepstral coefficients, spectral flatness, and number of zero crossings have been used to predict steaming popularity (Yang et. al., 2017; Lee et. al., 2018; Araujo et. al., 2019), improve music recommendation systems (Li et. al., 2007; Schedl, 2013), and to classify emotion in music (Jia, 2022). High-level audio features such as danceability, instrumentalness, and speechiness are included in track information available from the Spotify API. These high-level audio features have been used to identify song attributes (Febirautami et. al., 2018), predict popularity (Reiman et.al., 2018; Martín-Gutiérrez et. al., 2020; Kim, 2021; Gao, 2021), to classify music into genres (Setiadi et. al., 2020), and to classify music into moods (Chen et. al., 2021).

Clustering music has been used for recommender systems (Li et. al., 2004; Li et. al., 2007; Huo, 2021), as well as to categorise music (Honingh et. al., 2011). In this study, overall trends, genres, and audio feature clusters will be considered to attempt to improve predictive analytics. It was hypothesised that the accuracy of predictions of song popularity using audio features could be improved by separating predictions by genre (Reiman et.al., 2018). Music genres have been used in combination with high-level audio features to predict popularity (Kim, 2021). It was noted that audio features within the Billboard Hot 100 within genres are relatively consistent over time (O'Toole et. al., 2022).

In this study, clustering and classification will be used. Similar studies have had success with a variety of techniques and models. Neural networks have been used to predict popularity (Yang et. al., 2017; Gao, 2021), improve music recommendation systems (Li, 2021; Shi, 2021), or classify music (Jia, 2022; Li et. al., 2022). K-Means Clustering has been used in a number of studies to cluster data (Li et. al., 2007; Xu et. al., 2021; Kim et. al., 2021). A variety of classification models have been used to classify music, notably Support Vector Machines (Laurier et. al., 2009; Lee et. al., 2018; Reiman et.al., 2018; Araujo et. al., 2019; Setiadi et. al., 2020; Wilkes et. al., 2021), K-Nearest Neighbours (Cataltepe et. al., 2007a; Reiman et.al., 2018; Kim, 2021), Decision Trees / Random Forests / Boosted Trees (West, 2008; Febirautami et. al., 2018; Chen et. al., 2021; Gao, 2021), and Logistic Regression (Reiman et.al., 2018; Chen et. al., 2021; Gao, 2021). In addition, Principle Component Analysis has been used to reduce dimensionality and improve predictive results (Gao, 2021).

Predicting popularity on the Billboard Hot 100 has been investigated in other studies (Lee et. al., 2018; Reiman et.al., 2018). However, no study has successfully used high-level audio features to predict popularity as defined above. Although high-level audio features were used, Reiman et.al. were not able to accurately predict song popularity. This was potentially due to an overly diverse dataset for non-hit songs. It was hypothesised that the accuracy of predictions of song popularity using audio features could be improved by separating predictions by genre (Reiman et.al., 2018). Additionally, although it was Lee et. al. were able to demonstrate that popularity could be predicted using audio feature alone, their study focussed on low-level audio features and different statistical descriptions for popularity. This study aims to accurately predict song popularity, defined as appearance on the Billboard Hot 100 charts, using high-level audio features and genre data available from the Spotify API.

# Data Description

## Data Sources

The data gathered for this project have been taken from multiple sources and combined. Data was found using the Google dataset search engine (Google Dataset Search, n.d.).

Three of the relevant sources were found on Kaggle.com, a popular online data science community where users can share datasets (Dhruvil Dave, 2021; Malte Grosse, 2022; Rodolfo Figueroa, 2020). Audio features from the large datasets were matched with the list of songs from the Billboard Hot 100.

## Data Pre-Processing and Organisation

Data Importing and preprocessing was completed in 2 stages, before and after the literature review. The two phases of dataset preprocessing are included as **Attachment 1** and **Attachment 2** respectively.

The “Billboard ‘The Hot 100’ Songs” dataset (Dhruvil Dave, 2021) was available in CSV format, and includes date, rank, song title, artist, last-week, peak-rank, and weeks-on-board. This data did not include Spotify song ids or audio features, so the Spotify API was used to gather this data in cases where the audio feature data was unavailable from the other sources. This CSV was imported into Python as a Pandas dataframe.

The SQLite dataset, “8+ M. Spotify Tracks, Genre, Audio Features” (Malte Grosse, 2022) included 9 tables totalling 44 columns. The database was queried to combine song title, artist, Spotify id, release date, and audio features. The queried data were exported to CSV and imported into Python as a Pandas dataframe.

The “Spotify 1.2M+ Songs” dataset (Rodolfo Figueroa, 2020) was available in CSV format. It included combine song title, artist, Spotify id, release date, and audio features. This CSV was imported into Python as a Pandas dataframe.

Based on findings from the literature review portion of this study, genre data was gathered for easily available songs. Since genre data was included in the SQLite database for many of the songs (Malte Grosse, 2022), this data was queried, exported as CSV, and imported into Python. Since multiple genres were often available for a given artist, all genres were populated, then the results were sorted by most common genre, then duplicate song entries were dropped. This results in songs being categorised as a singular genre. It should be noted that this is the most common genre, and not necessarily the most applicable genre. It should also be noted that not all songs have genre data associated with them. Approximately 69% of songs included genre data (approximately 6.6M entries).

Songs from the Billboard Hot 100 were found in the SQLite query data, were available. Songs still missing genre data, which included Spotify id were queried using the Spotify API to gather genre data were available. Since multiple genres were often available, all genres were obtained from the API, then the result corresponding to the most common genres from the SQLite query data were populated as the songs genre. Similar to above, it should be noted that this is the most common genre, and not necessarily the most applicable genre, and not all songs have genre data associated with them.

Since Get Requests from the Spotify API were time consuming, and the utilised workflow results in timeouts after 1 hour, undefined genres in the “Spotify 1.2M+ Songs” dataset (Rodolfo Figueroa, 2020) were left undefined due to time constraints.

The data from each dataset was combined to form 4 primary, non-distinct working datasets:

* All songs including audio features (approximately 10M entries)
* The Billboard Hot 100 historical charts (approximately 300k entries)
* All Songs from The Billboard Hot 100 historical charts (approximately 30k entries)
* Songs from The Billboard Hot 100 that include audio features and genre (approximately 20k entries)

Since the dataframes are not distinct, in cases where mutually exclusive groups were necessary, Pandas dataframe query functions and vectorized formula have been implemented.

Once working datasets were completed, they were exported as Parquet format. Parquet has a number of advantages over CSV format, most notably file size, the retention of data type formats, and the amount of time required to re-load the file into Python Pandas dataframe.

## Quality Assurance

In order to check the accuracy of the data and consistency between datasets, a Quality Assurance (QA) check was performed on the final dataset. Audio features from 100 songs were gathered from the Spotify API and compared to the datasets listed above. There were 3 non-trivial inconsistencies noted in 2 of the 100 songs:

* Madonna – “Live To Tell”
  + A significant difference in loudness (~7 dB)
  + Approximately 1 second difference in length
  + Both of these changes appear to result from re-mastering and re-uploading the track
* Lil Wayne – “Let It All Work Out”
  + The key signature was not consistent between the sources
  + This inconsistency is likely due to the characteristics of the song:
    - Atonal (most notably the singing)
    - Detuned (bass pitch automation, low-fi detuning effects)

All other inconsistencies were extremely small and appeared to be standard rounding errors. Overall, there is a large degree of consistency between datasets. Furthermore, inconsistencies are all explainable with reasonable suppositions. The data was therefore assumed to be consistent and accurate for the purposes of this study.

## File and Calculation Locations

Datasets and calculations used in this study can be found a the following URL:

<https://github.com/KevinCarr42/Billboard-100-Audio-Feature-Analytics>

Files too large to upload to GitHub been uploaded to a shared Google Drive folder (shared with Toronto Metropolitan University accounts):

<https://drive.google.com/drive/folders/10wpORzZURV11VAUPKmCDDKxHFCvwjloR>

## Descriptive Statistics

Descriptive statistics for the datasets are included in the following tables. More detailed descriptive calculations are included in **Attachment 3**.

Table . Descriptive Statistics - All Songs With Audio Features

| **Audio Feature** | **Mean** | **Standard Deviation** | **Minimum** | **Quartiles** | | | **Maximum** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **25%** | **50%** | **75%** |
| acousticness | 0.42 | 0.37 | 0.00 | 0.03 | 0.34 | 0.82 | 1.00 |
| danceability | 0.53 | 0.19 | 0.00 | 0.40 | 0.55 | 0.68 | 1.00 |
| duration\_ms | 238,210 | 159,342 | 0 | 169,600 | 216,933 | 275,080 | 19,672,058 |
| energy | 0.55 | 0.28 | 0.00 | 0.31 | 0.57 | 0.79 | 1.00 |
| instrumentalness | 0.26 | 0.37 | 0.00 | 0.00 | 0.00 | 0.65 | 1.00 |
| key | 5.2 | 3.5 | 0 | 2 | 5 | 8 | 11 |
| liveness | 0.21 | 0.18 | 0.00 | 0.10 | 0.13 | 0.26 | 1.00 |
| loudness | -11.0 | 6.3 | -60.0 | -13.7 | -9.2 | -6.4 | 7.2 |
| mode | 0.66 | 0.47 | 0.00 | 0.00 | 1.00 | 1.00 | 1.00 |
| speechiness | 0.10 | 0.14 | 0.00 | 0.04 | 0.05 | 0.08 | 0.97 |
| tempo | 119 | 31 | 0 | 95 | 119 | 137 | 250 |
| time\_signature | 3.84 | 0.57 | 0.00 | 4.00 | 4.00 | 4.00 | 5.00 |
| valence | 0.47 | 0.28 | 0.00 | 0.23 | 0.47 | 0.71 | 1.00 |

Table . Descriptive Statistics - Billboard Hot 100

| **Audio Feature** | **Mean** | **Standard Deviation** | **Minimum** | **Quartiles** | | | **Maximum** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **25%** | **50%** | **75%** |
| rank | 51 | 29 | 1 | 26 | 51 | 76 | 100 |
| last-week | 48 | 28 | 1 | 23 | 47 | 72 | 100 |
| peak-rank | 41 | 29 | 1 | 13 | 38 | 65 | 100 |
| weeks-on-board | 9 | 8 | 1 | 4 | 7 | 13 | 90 |
| acousticness | 0.28 | 0.28 | 0.00 | 0.04 | 0.18 | 0.47 | 1.00 |
| danceability | 0.60 | 0.15 | 0.00 | 0.51 | 0.61 | 0.71 | 0.99 |
| duration\_ms | 226,880 | 66,552 | 30,213 | 183,360 | 221,306 | 258,399 | 1,561,133 |
| energy | 0.63 | 0.20 | 0.01 | 0.48 | 0.64 | 0.79 | 1.00 |
| instrumentalness | 0.03 | 0.14 | 0.00 | 0.00 | 0.00 | 0.00 | 0.99 |
| key | 5.2 | 3.6 | 0 | 2 | 5 | 8 | 11 |
| liveness | 0.19 | 0.16 | 0.01 | 0.09 | 0.13 | 0.24 | 1.00 |
| loudness | -8.6 | 3.6 | -30.3 | -11.0 | -8.2 | -5.8 | 2.3 |
| mode | 0.73 | 0.44 | 0.00 | 0.00 | 1.00 | 1.00 | 1.00 |
| speechiness | 0.06 | 0.07 | 0.00 | 0.03 | 0.04 | 0.06 | 0.95 |
| tempo | 120 | 28 | 0 | 100 | 119 | 136 | 241 |
| time\_signature | 3.94 | 0.30 | 0.00 | 4.00 | 4.00 | 4.00 | 5.00 |
| valence | 0.61 | 0.24 | 0.00 | 0.42 | 0.63 | 0.81 | 0.99 |

Table . Descriptive Statistics - All Songs From Billboard Hot 100 With Audio Features

| **Audio Feature** | **Mean** | **Standard Deviation** | **Minimum** | **Quartiles** | | | **Maximum** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **25%** | **50%** | **75%** |
| acousticness | 0.32 | 0.29 | 0.00 | 0.05 | 0.22 | 0.56 | 1.00 |
| danceability | 0.59 | 0.15 | 0.00 | 0.49 | 0.60 | 0.70 | 0.99 |
| duration\_ms | 217,638 | 68,403 | 30,213 | 169,533 | 210,426 | 251,333 | 1,561,133 |
| energy | 0.61 | 0.20 | 0.01 | 0.46 | 0.62 | 0.78 | 1.00 |
| instrumentalness | 0.04 | 0.15 | 0.00 | 0.00 | 0.00 | 0.00 | 0.99 |
| key | 5.2 | 3.6 | 0 | 2 | 5 | 8 | 11 |
| liveness | 0.20 | 0.17 | 0.01 | 0.09 | 0.13 | 0.26 | 1.00 |
| loudness | -8.9 | 3.6 | -30.3 | -11.3 | -8.6 | -6.1 | 2.3 |
| mode | 0.74 | 0.44 | 0.00 | 0.00 | 1.00 | 1.00 | 1.00 |
| speechiness | 0.07 | 0.08 | 0.00 | 0.03 | 0.04 | 0.06 | 0.95 |
| tempo | 120 | 28 | 0 | 100 | 119 | 136 | 241 |
| time\_signature | 3.93 | 0.33 | 0.00 | 4.00 | 4.00 | 4.00 | 5.00 |
| valence | 0.61 | 0.24 | 0.00 | 0.43 | 0.64 | 0.81 | 0.99 |

Table . Descriptive Statistics - Songs From Billboard Hot 100 With Audio Features And Genre

| **Audio Feature** | **Mean** | **Standard Deviation** | **Minimum** | **Quartiles** | | | **Maximum** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **25%** | **50%** | **75%** |
| acousticness | 0.32 | 0.29 | 0.00 | 0.05 | 0.22 | 0.56 | 1.00 |
| danceability | 0.59 | 0.15 | 0.00 | 0.49 | 0.60 | 0.70 | 0.99 |
| duration\_ms | 217,638 | 68,403 | 30,213 | 169,533 | 210,426 | 251,333 | 1,561,133 |
| energy | 0.61 | 0.20 | 0.01 | 0.46 | 0.62 | 0.78 | 1.00 |
| instrumentalness | 0.04 | 0.15 | 0.00 | 0.00 | 0.00 | 0.00 | 0.99 |
| key | 5.2 | 3.6 | 0 | 2 | 5 | 8 | 11 |
| liveness | 0.20 | 0.17 | 0.01 | 0.09 | 0.13 | 0.26 | 1.00 |
| loudness | -8.9 | 3.6 | -30.3 | -11.3 | -8.6 | -6.1 | 2.3 |
| mode | 0.74 | 0.44 | 0.00 | 0.00 | 1.00 | 1.00 | 1.00 |
| speechiness | 0.07 | 0.08 | 0.00 | 0.03 | 0.04 | 0.06 | 0.95 |
| tempo | 120 | 28 | 0 | 100 | 119 | 136 | 241 |
| time\_signature | 3.93 | 0.33 | 0.00 | 4.00 | 4.00 | 4.00 | 5.00 |
| valence | 0.61 | 0.24 | 0.00 | 0.43 | 0.64 | 0.81 | 0.99 |

### Audio Feature Descriptions

Audio features available from the Spotify API used in this study are described in detail in the below table (Spotify, n.d.).

Table . Description of Audio Features From Spotify API

| **Audio Feature** | **Type** | **Description** | **Min** | **Max** |
| --- | --- | --- | --- | --- |
| acousticness | number  <float> | A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic. | 0 | 1 |
| danceability | number  <float> | Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable. | 0 | 1 |
| duration\_ms | integer | The duration of the track in milliseconds. | 0 | N/A |
| energy | number  <float> | Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. For example, death metal has high energy, while a Bach prelude scores low on the scale. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy. | 0 | 1 |
| instrumentalness | number  <float> | Predicts whether a track contains no vocals. "Ooh" and "aah" sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly "vocal". The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0. | 0 | 1 |
| key | integer | The key the track is in. Integers map to pitches using standard Pitch Class notation. E.g. 0 = C, 1 = C♯/D♭, 2 = D, and so on. If no key was detected, the value is -1. | -1 | 11 |
| liveness | number  <float> | Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live. | 0 | 1 |
| loudness | number  <float> | The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks. Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). Values typically range between -60 and 0 db. | -60 | 0 |
| mode | integer | Mode indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived. Major is represented by 1 and minor is 0. | 0 | 1 |
| speechiness | number  <float> | Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks. | 0 | 1 |
| tempo | number  <float> | The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration. | 0 | N/A |
| time\_signature | integer | An estimated time signature. The time signature (meter) is a notational convention to specify how many beats are in each bar (or measure). The time signature ranges from 3 to 7 indicating time signatures of "3/4", to "7/4". | 3 | 7 |
| valence | number  <float> | A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry). | 0 | 1 |

### Histograms

Histograms for each of the audio features are shown in the below **Figures**, comparing songs on the Billboard Hot 100 with all songs in this study.

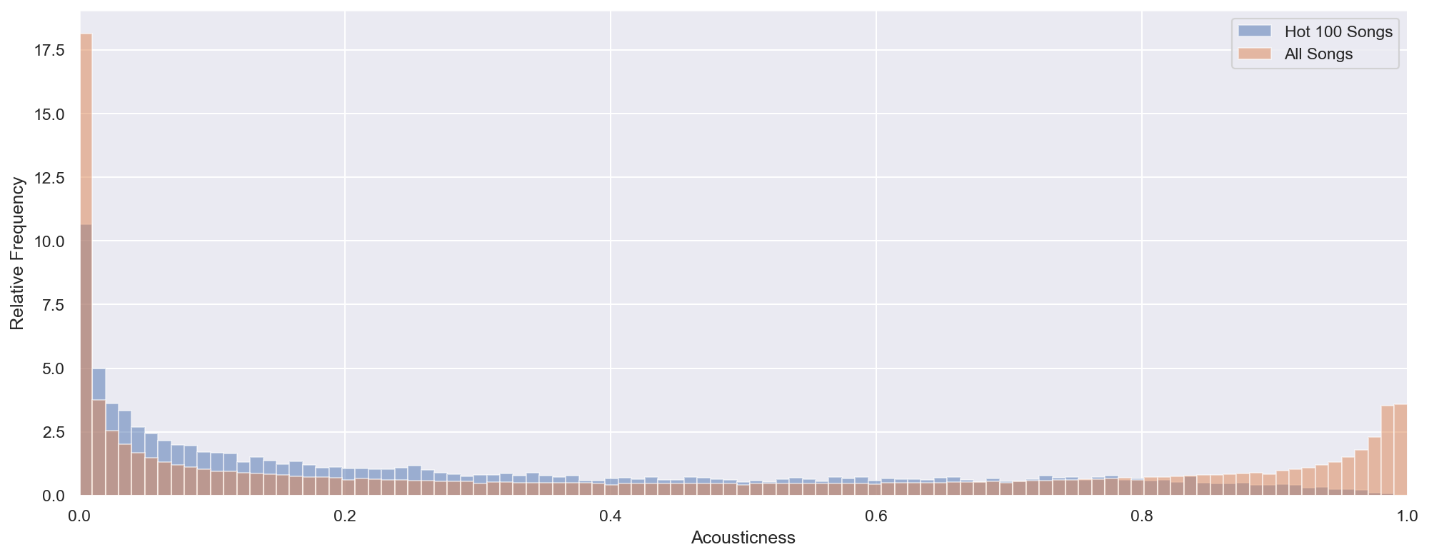


Figure . Acousticness Histogram

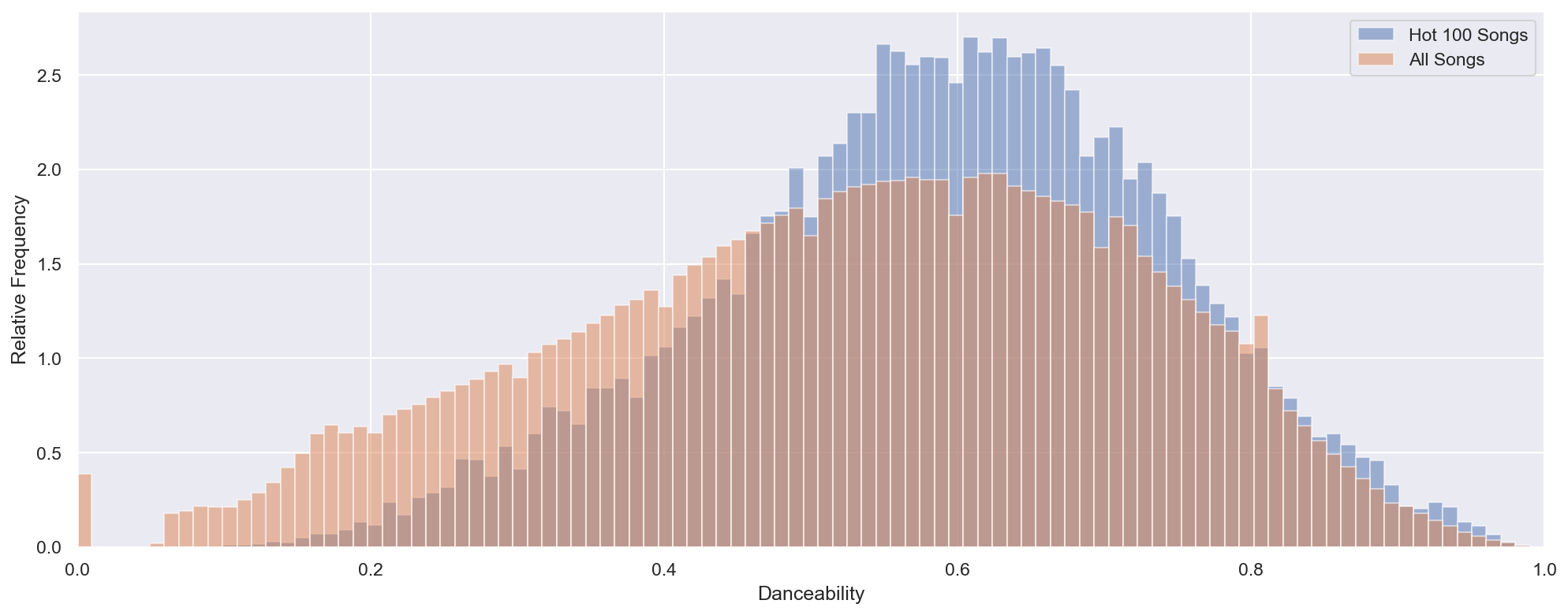


Figure . Danceability Histogram

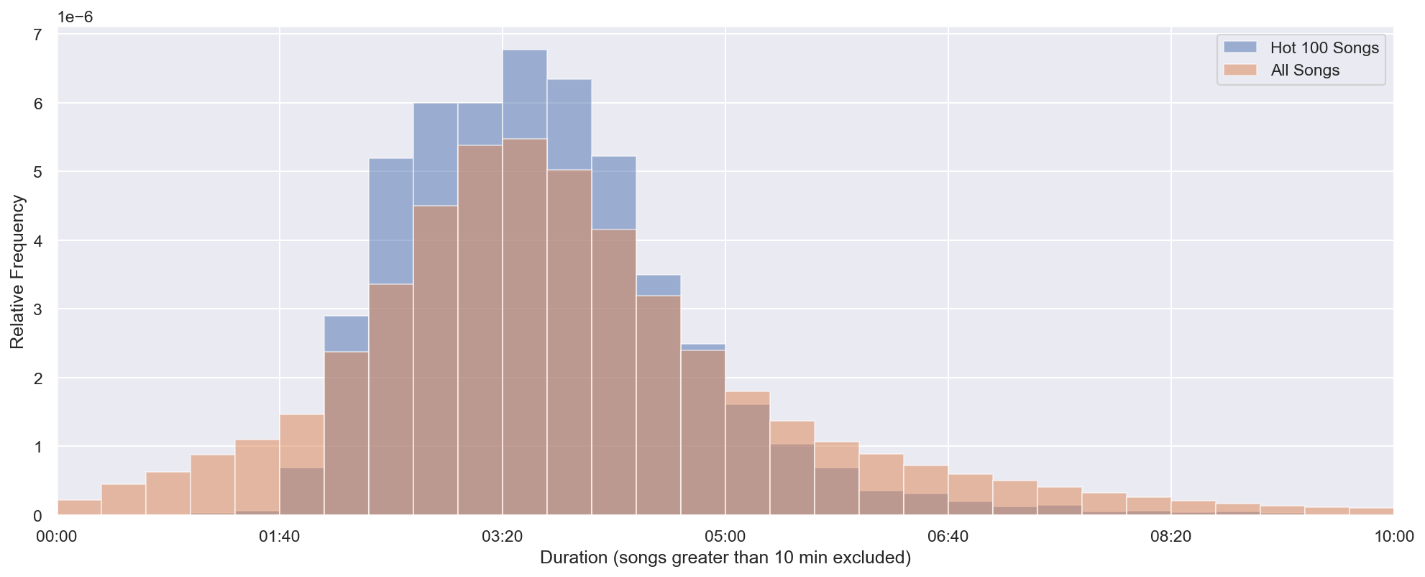


Figure . Duration Histogram

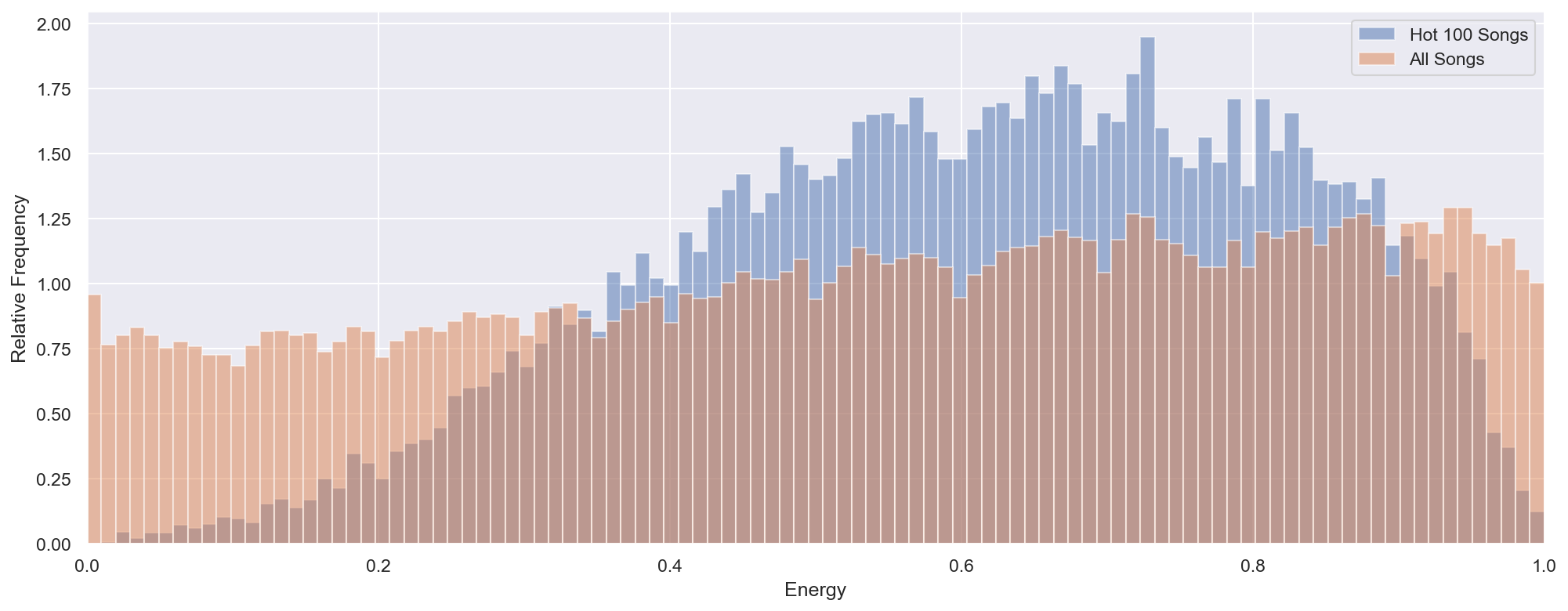


Figure . Energy Histogram

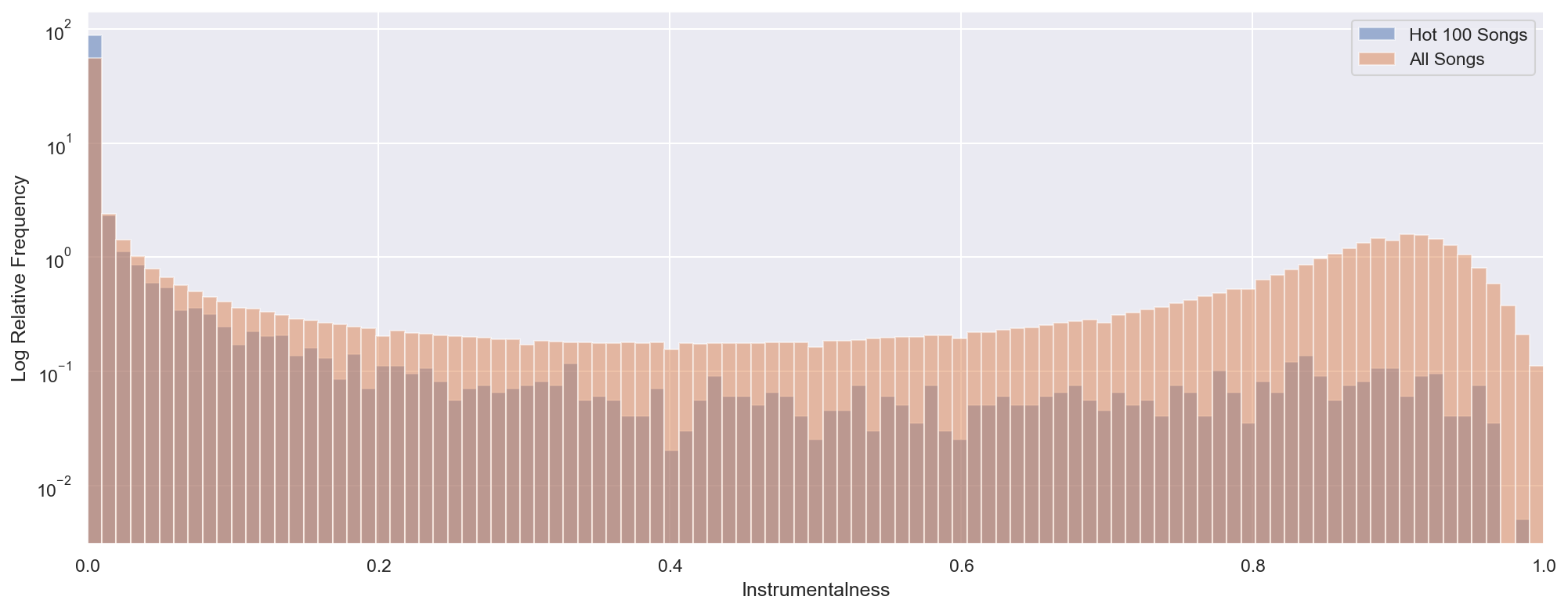


Figure . Instrumentalness Histogram

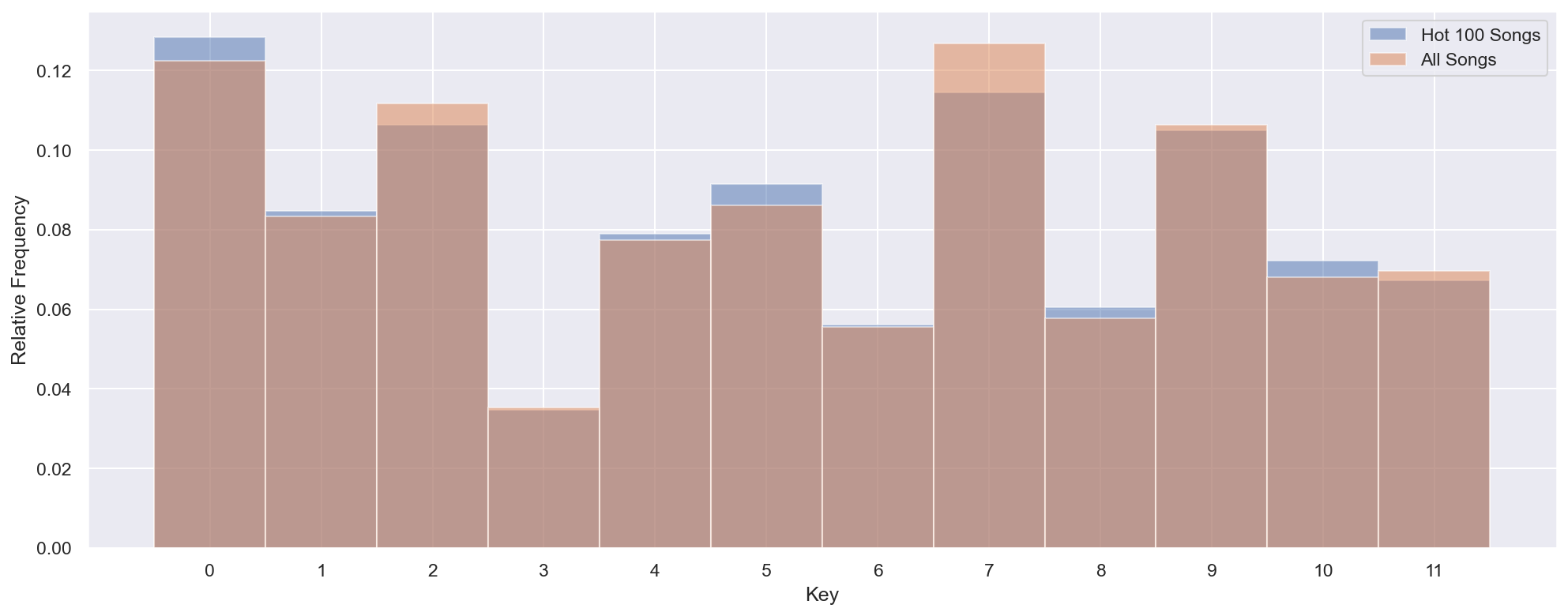


Figure . Key Histogram

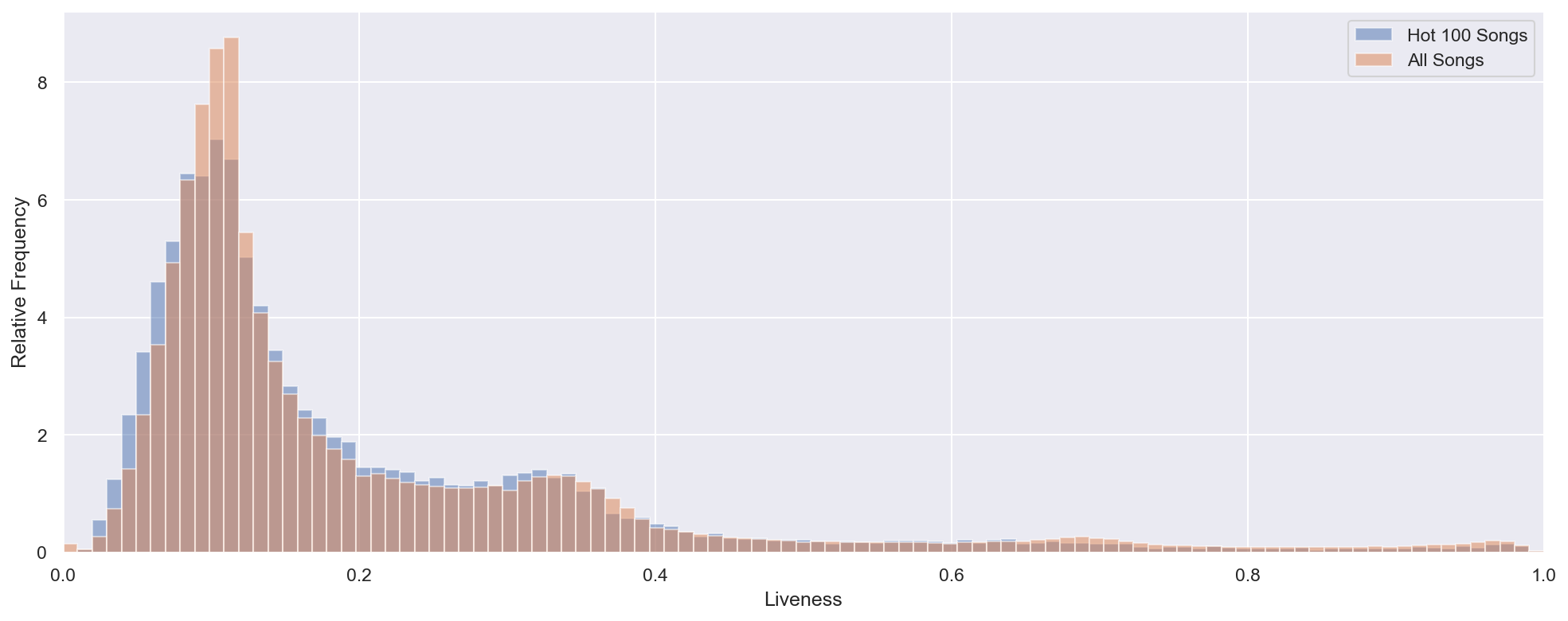


Figure . Liveness Histogram

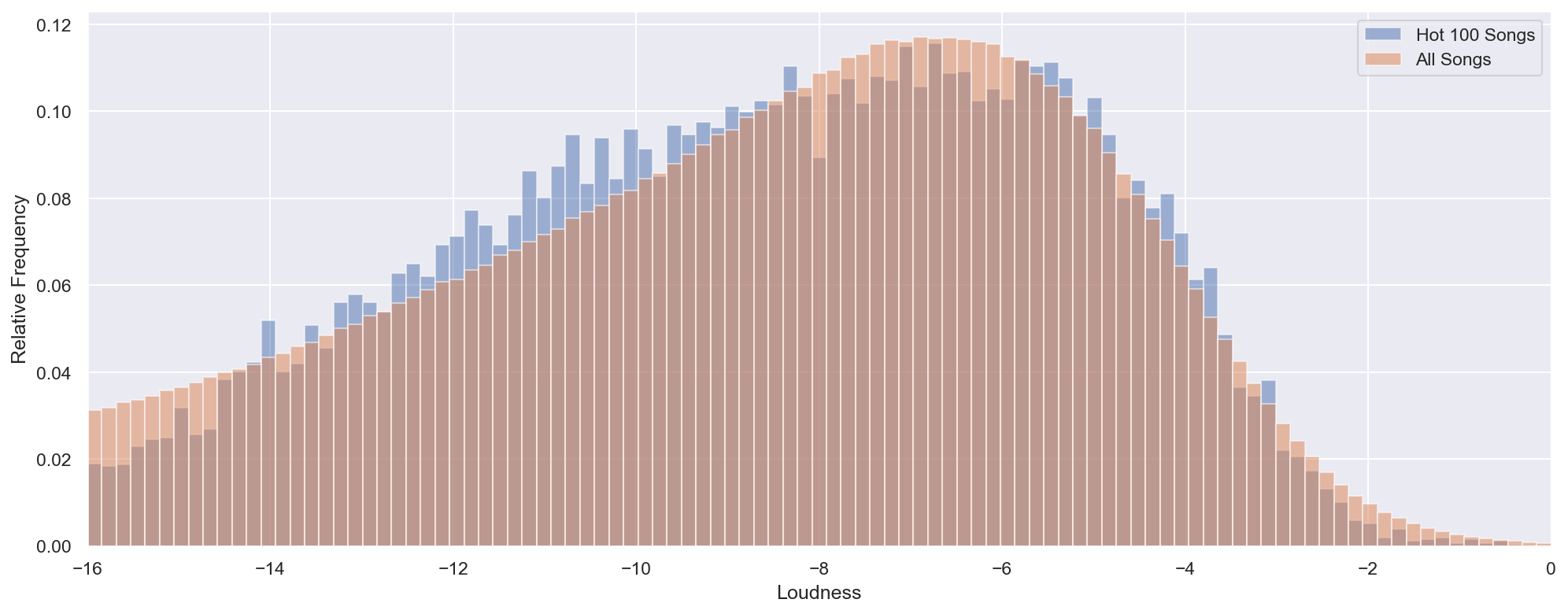


Figure . Loudness Histogram

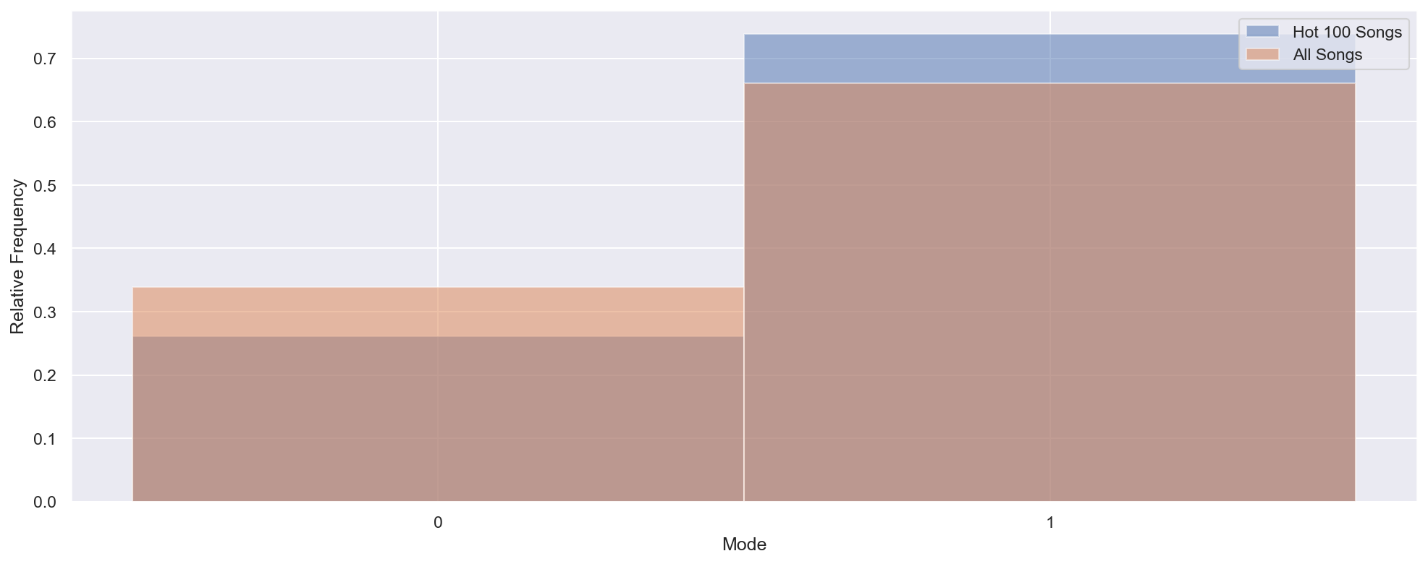


Figure . Mode Histogram



Figure . Speechiness Histogram

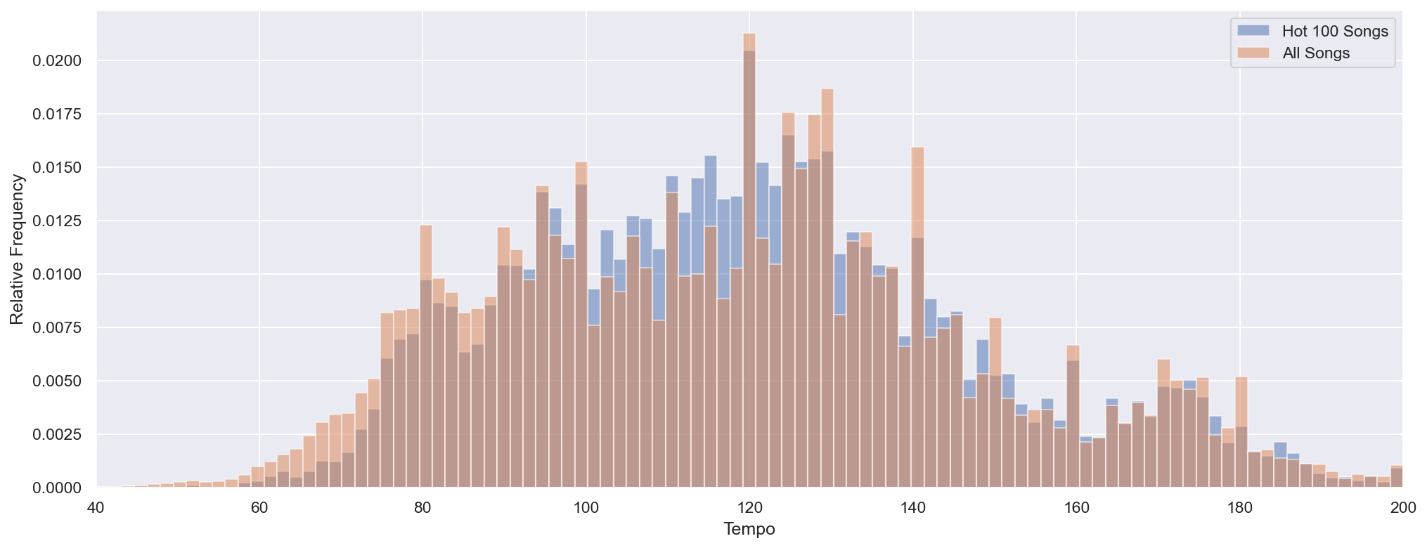


Figure . Tempo Histogram

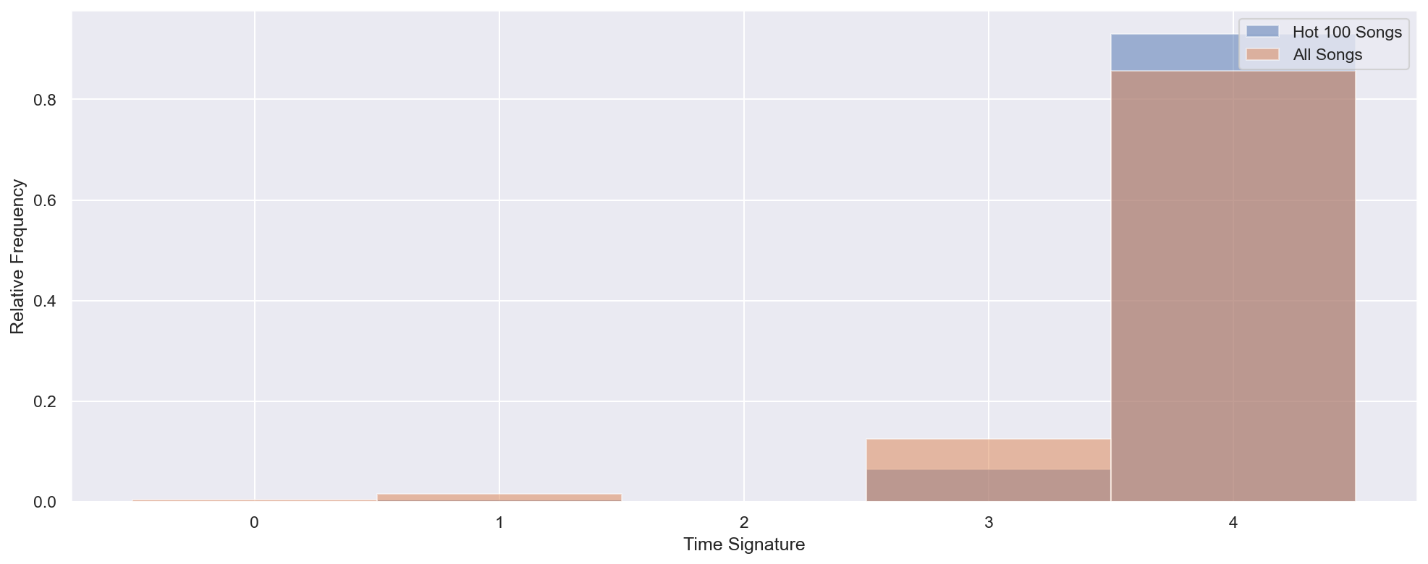


Figure . Time Signature Histogram

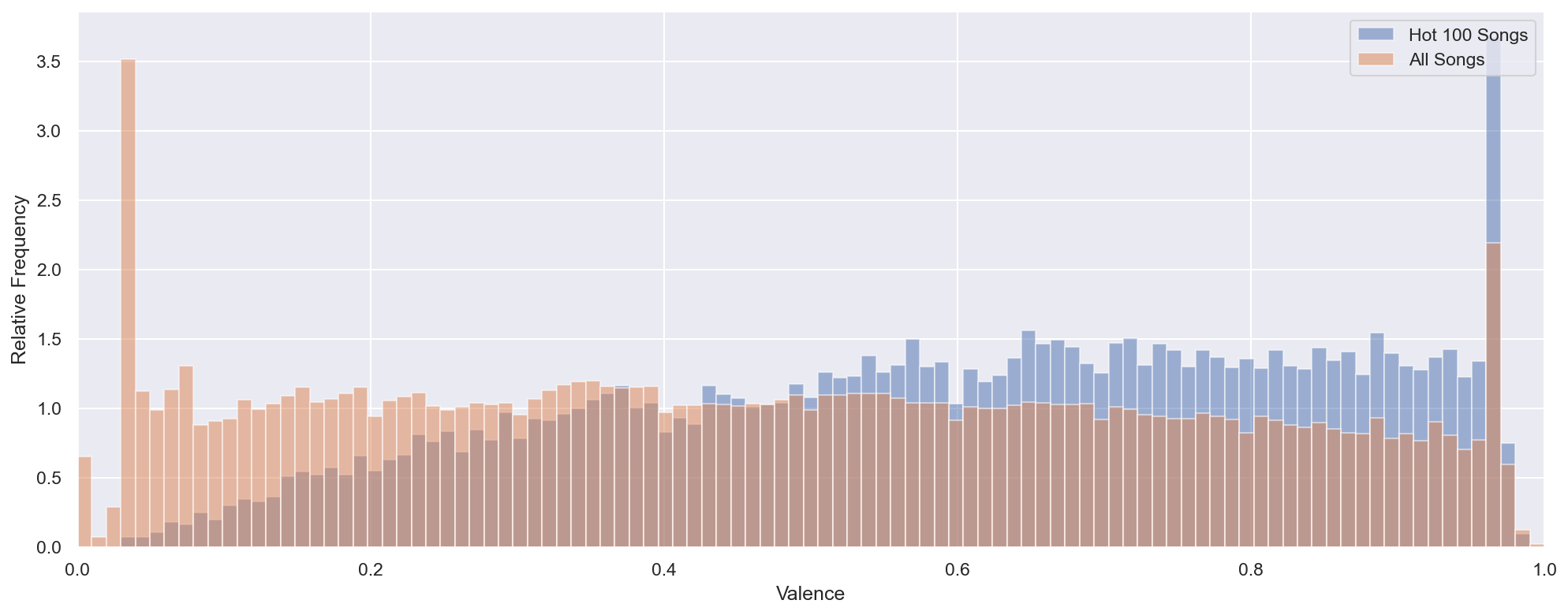


Figure . Valence Histogram

## Historical Changes in Audio Features

Historical line plots for each of the audio features are shown in the below Figures, comparing audio feature averaged by release year for songs on the Billboard Hot 100 compared with all songs in this study.

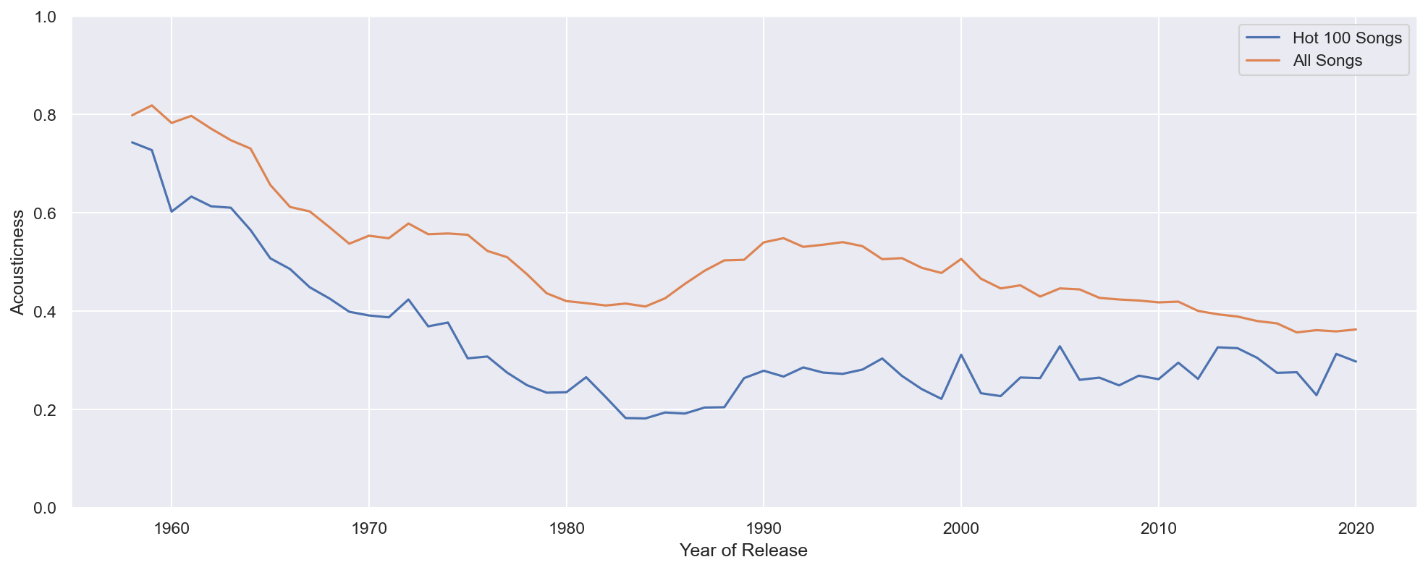


Figure . Acousticness History Comparing the Billboard Hot 100 with All Songs

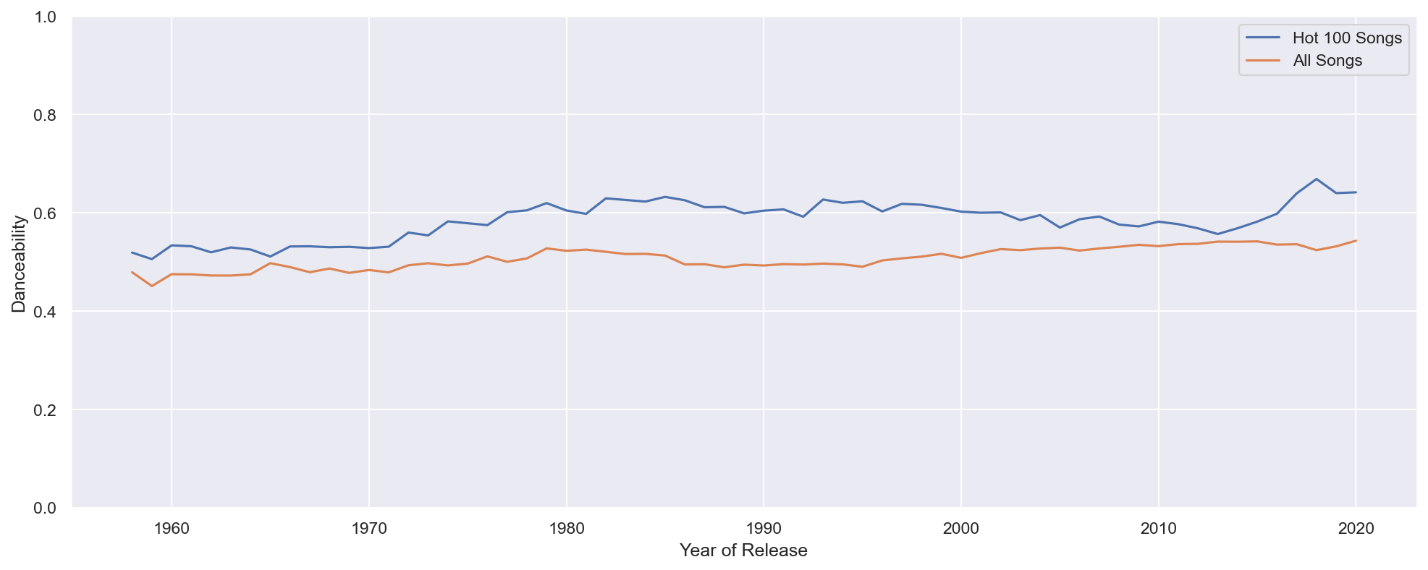


Figure . Danceability History Comparing the Billboard Hot 100 with All Songs

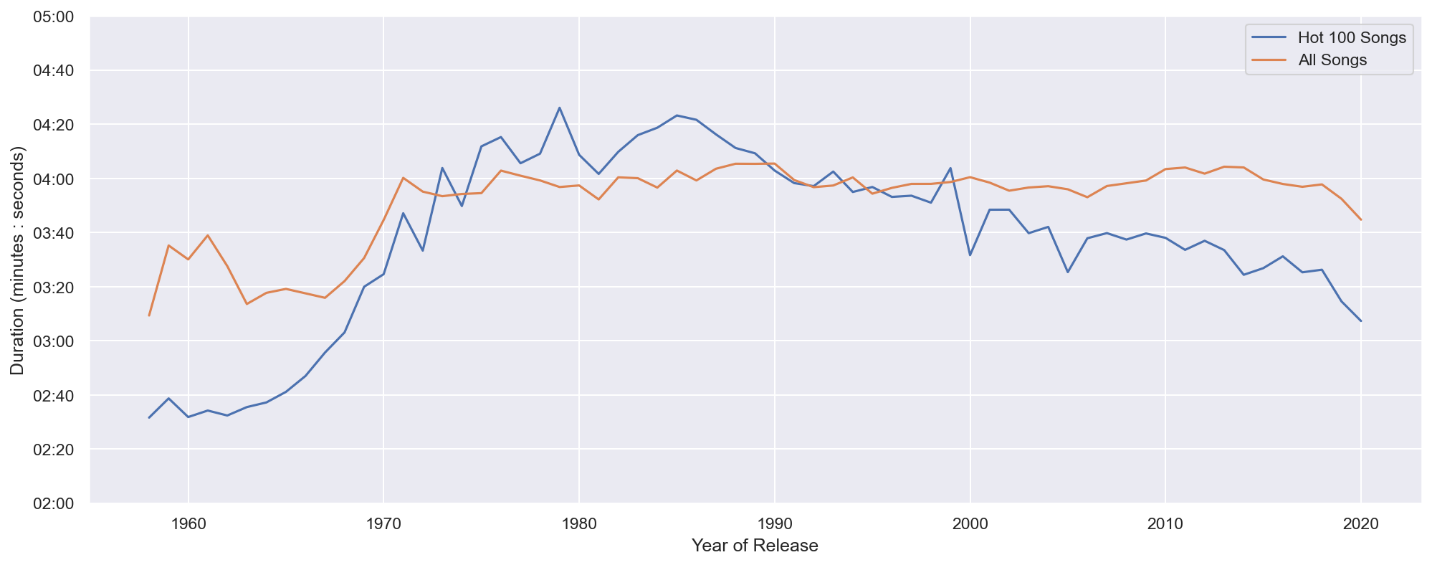


Figure . Duration History Comparing the Billboard Hot 100 with All Songs

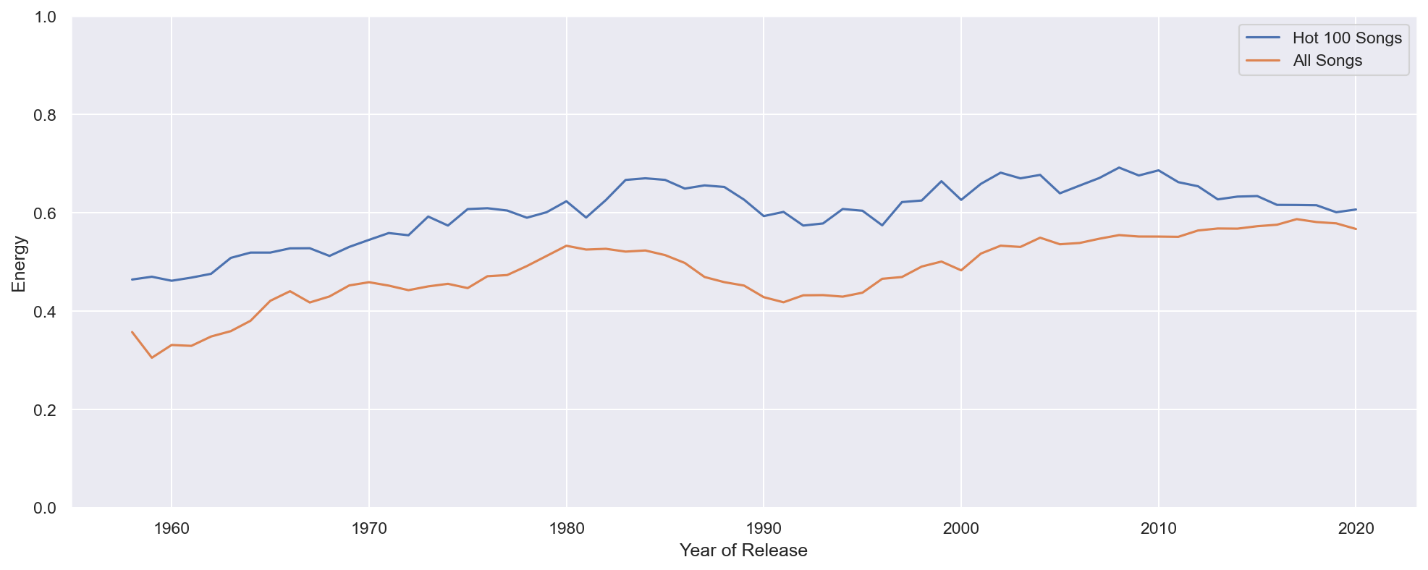


Figure . Energy History Comparing the Billboard Hot 100 with All Songs

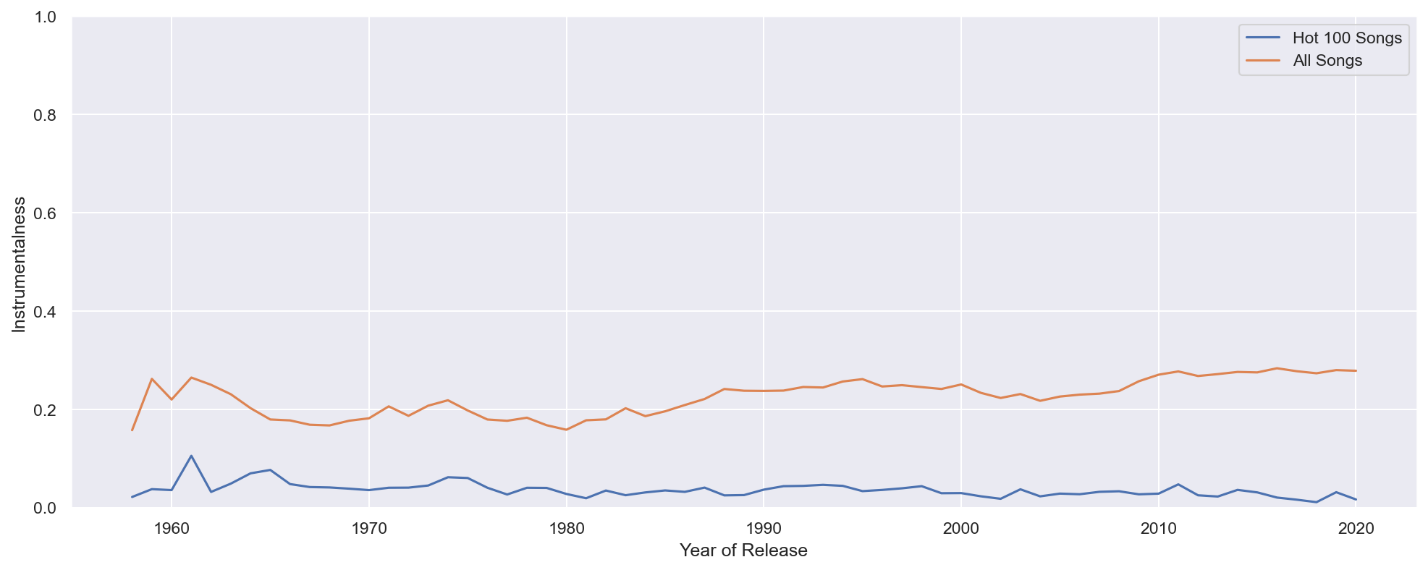


Figure . Instrumentalness History Comparing the Billboard Hot 100 with All Songs

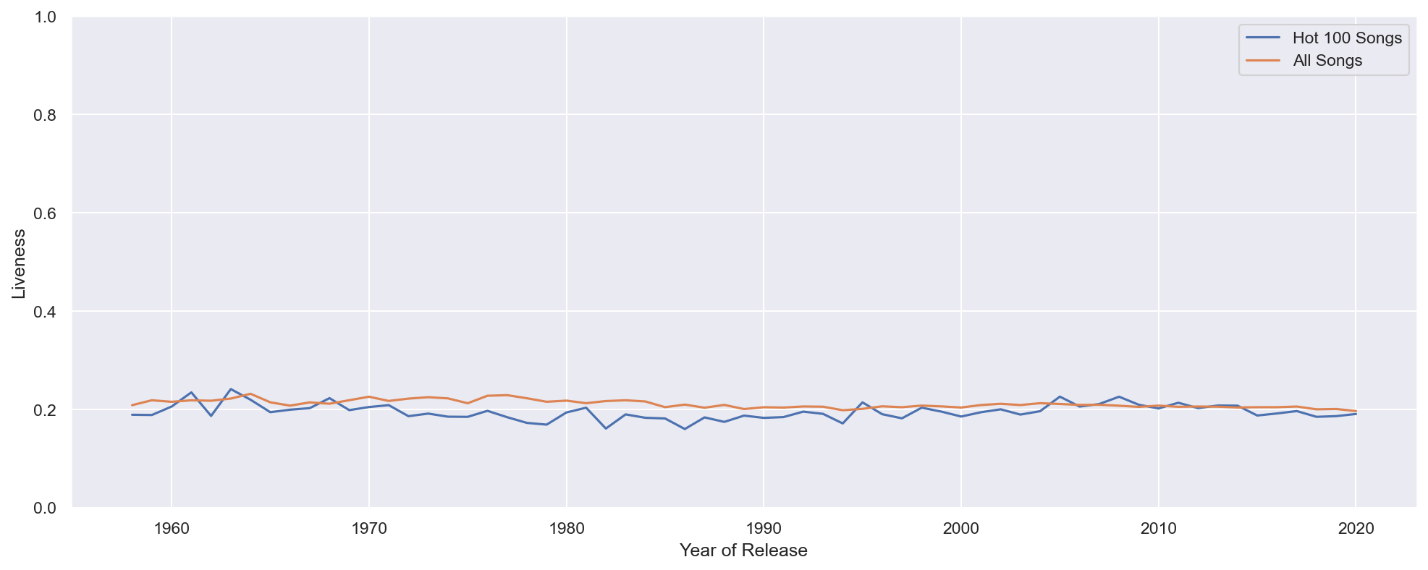


Figure . Liveness History Comparing the Billboard Hot 100 with All Songs

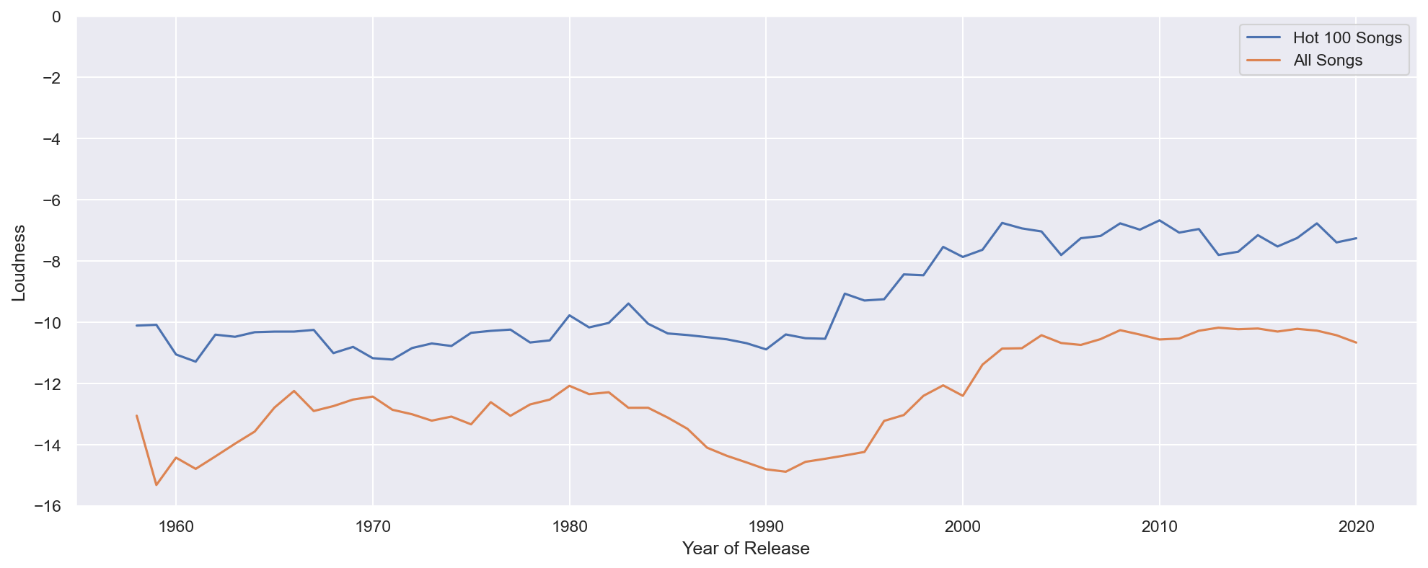


Figure . Loudness History Comparing the Billboard Hot 100 with All Songs

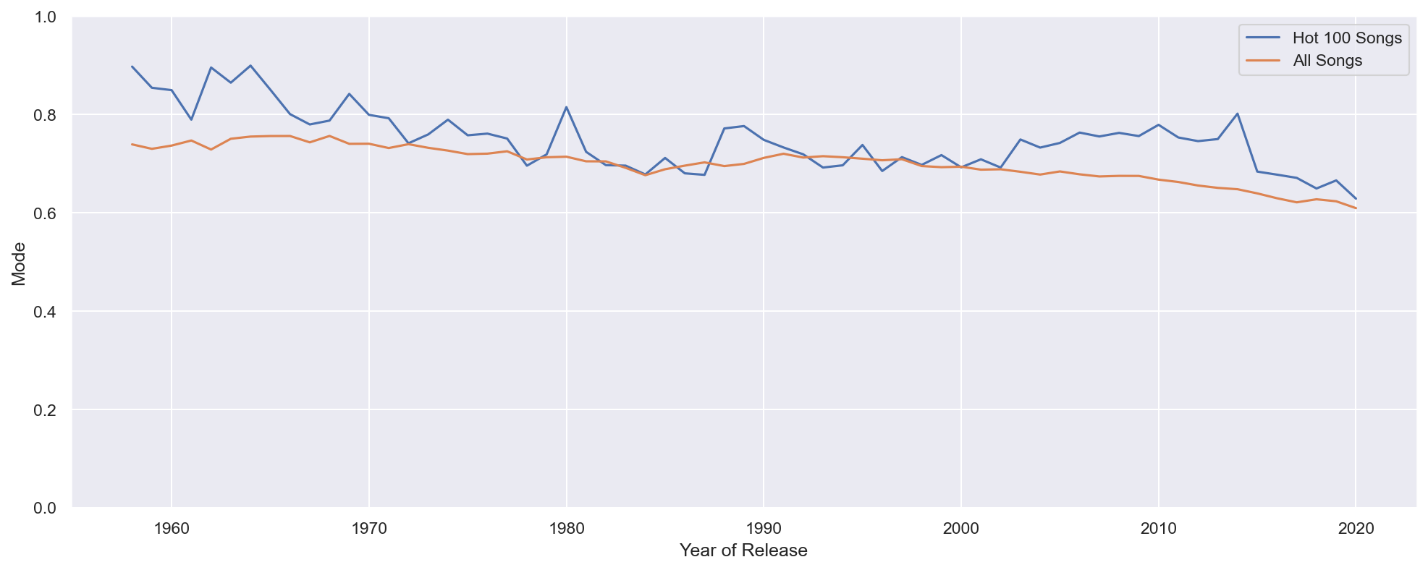


Figure . Mode History Comparing the Billboard Hot 100 with All Songs

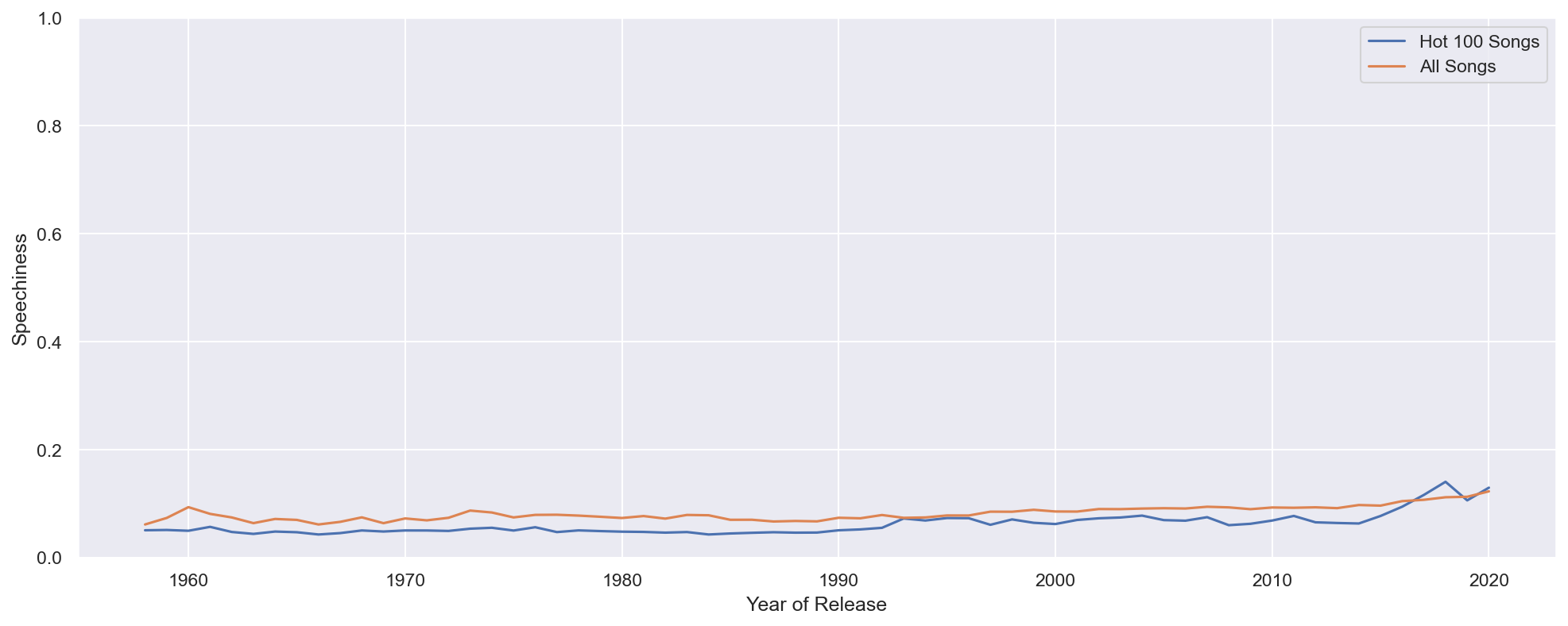


Figure . Speechiness History Comparing the Billboard Hot 100 with All Songs

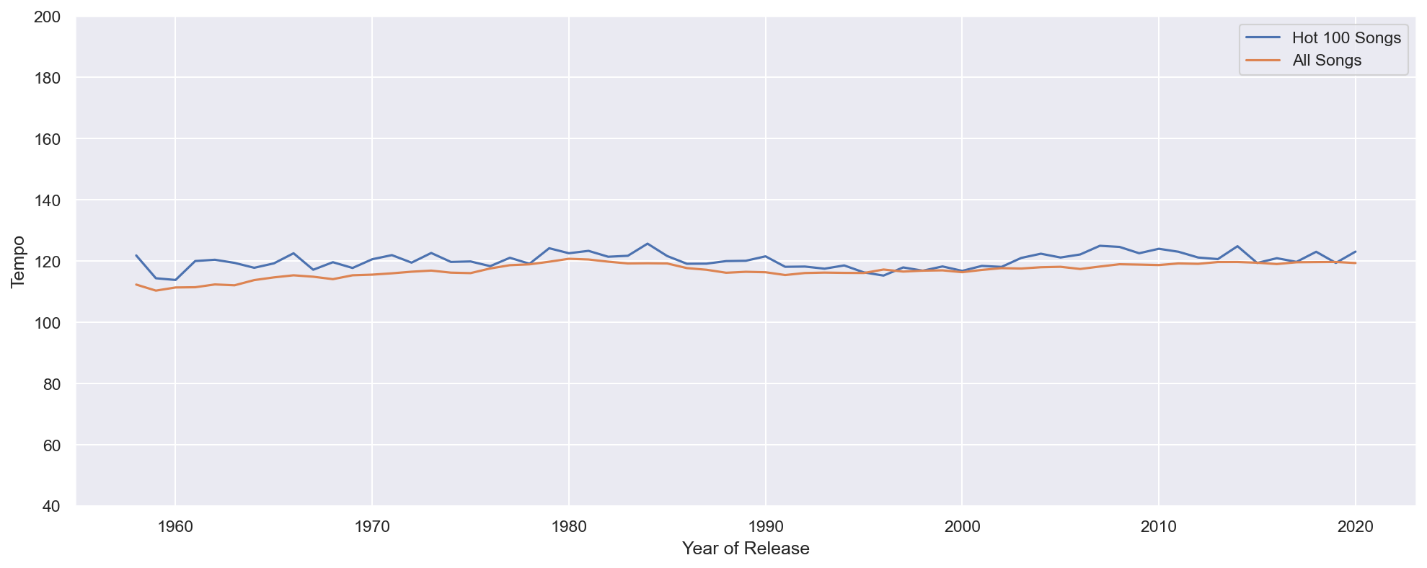


Figure . Tempo History Comparing the Billboard Hot 100 with All Songs

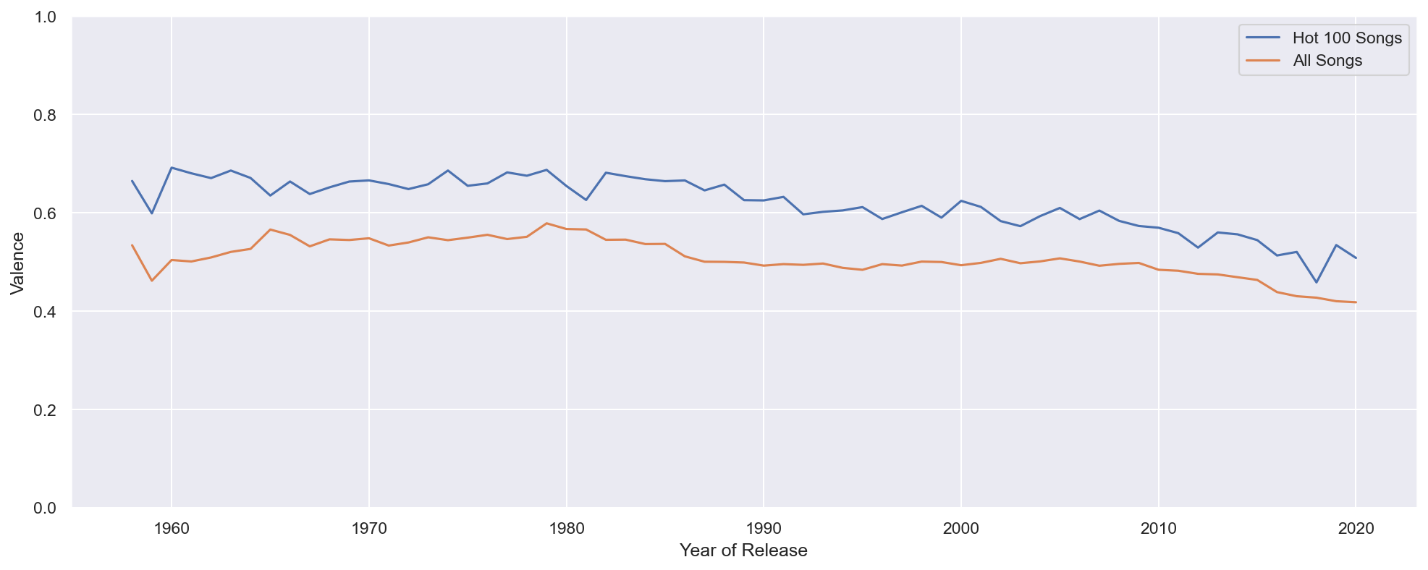


Figure . Valence History Comparing the Billboard Hot 100 with All Songs

The following Figure shows the Billboard Hot 100 chart yearly audio feature averages over time. In contrast to the above figures, this figure is grouped by song on the charts in a given year, and does not necessarily correspond to the release year for any given song. Also of note, all audio features were normalised to range from 0 to 1 in order to create a consistent plot.

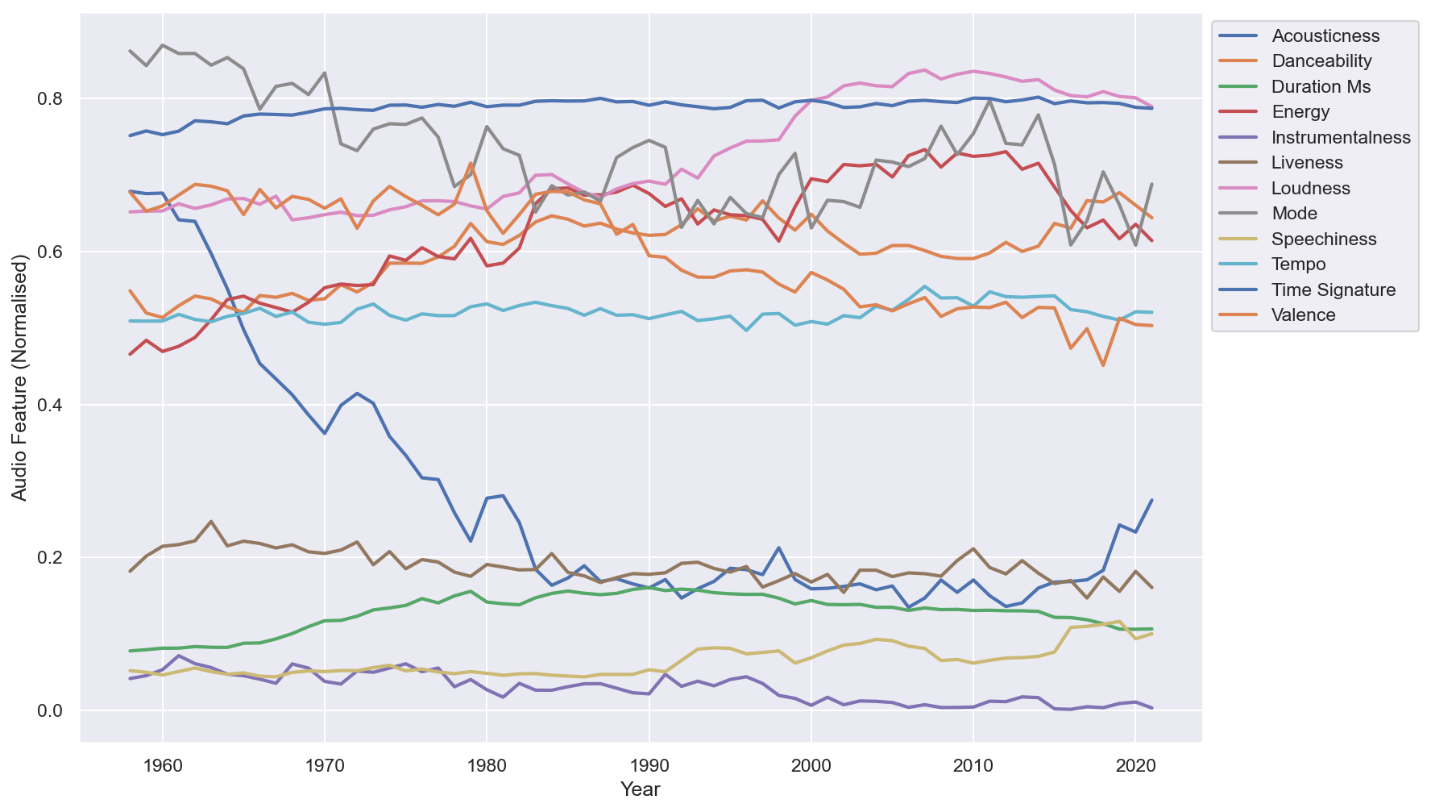


Figure . Billboard Hot 100 Historical Charts

A more detailed analysis of each audio feature over time is included in **Attachment 3**.

## Correlation Analysis

Correlation analysis is outlined in the below figure. A more detailed correlation analysis in included in **Attachment 3**. The analysis investigated which audio features correlate with each other, differences between Billboard Hot 100 songs and songs not on the list, as well as which features correlate with Popularity (as defined above).

Chart

Description automatically generated

Figure . Correlation Analysis Summary

In the above figure, the “POPULAR” feature denotes whether or not the song appeared on the Billboard Hot 100 charts. With a few exceptions, the analysis showed a relatively weak correlation between most audio features. This is unsurprising because audio features have been created to represent music simply, accurately, and without redundancies.

Also of note, every audio feature correlated very weakly with popularity. As noted in the Literature Review section, this is likely due to large varieties of musical styles all being included in the calculations. Separating analysis into genres yielded stronger correlations for some of the audio features. Genres are discussed in the next section, and detailed analysis is included in **Attachment 3**.

Correlation analysis was also conducted strictly on the Billboard Hot 100 dataset. The below figure shows the correlation between audio features and Billboard Hot 100 performance, both peak-rank and weeks-on-board.

Chart

Description automatically generated

Figure . Billboard Hot 100 Correlation Analysis Summary

As shown in the above figure, a number of audio features have a small correlation with performance on the Billboard Hot 100. The following figures show the sorted correlation coefficients corresponding to weeks-on-board and peak-rank, respectively. Note that the order of the charts is reversed because lower peak rank is optimal, whereas higher weeks-on-board is optimal. More detailed analysis is included in **Attachment 3**.

Chart

Description automatically generated

Figure . Ranked Audio Feature Correlations With Weeks On Billboard Hot 100 Charts

Chart, scatter chart

Description automatically generated

Figure . Ranked Audio Feature Correlations With Peak Rank On Billboard Hot 100

## Analysis of Genres

In order to improve correlations and future predictions, genre information has been investigated. Correlation analysis showed stronger correlation between audio features and popularity when restricting analysis to specific genres. Detailed genre analysis is included in **Attachment 3**.

Even after excluding all but the most popular genre for each artist, there were still over 5,000 genres in the dataset. Many of genres are rather obscure, and most read more like sub-genres. For this reason, bins of genres were grouped together to approximate more general genres. Regular Expressions were used to group songs into categories by matching patterns within genre info from the Spotify API. These genre groups were able to reduce variation in audio features to an extent, as shown in the below figure.

Chart, bar chart

Description automatically generated

Figure . Comparison of Audio Feature Variations Between Genres

It is important to note that these categories are based on assumptions about which genres belong together, and some of these grouping decisions have been made arbitrarily. Future phases on this study will investigate whether clustering analysis will create more useful clusters of songs by avoiding the need to label songs using genre.

# Project Approach

The central problem in this project is to utilise the datasets to predict the popularity of a song using audio features. For this projects, two main techniques will be employed, namely data mining and predictive analytics.

## Data Mining

First, data mining and knowledge discovery will be used to explore the data, cluster audio features, and determine correlations between audio features. Genres and audio feature correlation have been explored in detail; details are included in the Data Description section above, and detailed calculations are included in **Attachment 3**. In future phases of this study, this investigation will go more in depth and compare segmentation by genre to clustering performed using machine learning techniques. Ultimately, these clusters will be used in the predictive analytics described below.

For this analysis unsupervised machine learning techniques will employed. K-Means Clustering will most likely be employed, as it was the most common clustering algorithm noted in the Literature Review. Other machine learning models will be considered as the study progresses.

## Predictive Analytics

Secondly, predictive analytics will be used to attempt to build a predictive model using the data. We will attempt to predict whether or not a song will be make it onto the Billboard Hot 100 charts using audio features for that song. This prediction will likely incorporated the clusters from the previous phase of the study, as the correlation between audio features and popularity are very weak, as described above in the correlation analysis.

This analysis will investigate a variety of supervised classification algorithms, as well as ensemble methods. Based on the literature review, the most common algorithms for predicting popularity in music are Neural Networks, Support Vector Machines, and K-Nearest Neighbours, as well as some simpler models such as Naïve Bayes or Logistic Regression. Statistical significance and predictive power will be compared between the utilised algorithms.

## Outline of Methodology

The overall project methodology and timeline is outlined in the following chart.



Figure . Project Methodology Timeframe

# References

Spotify for Developers. (n.d.). <https://developer.spotify.com/documentation/web-api/>

Billboard Hot 100. (2022, September 6). Wikipedia. <https://en.wikipedia.org/w/index.php?title=Billboard_Hot_100&oldid=1108834581>

Google Scholar. (n.d. a). <https://scholar.google.com/>

Paper Digest. (n.d.). <https://www.paperdigest.org/>

Elicit. (n.d.). <https://elicit.org/>

Google Dataset Search. (n.d. b). <https://datasetsearch.research.google.com/>

Dhruvil Dave. (2021, November 9). Billboard "The Hot 100" Songs [Data set]. Kaggle. <https://doi.org/10.34740/KAGGLE/DS/1211465>

Malte Grosse. (2022, March 23). 8+ M. Spotify Tracks, Genre, Audio Features [Data set]. Kaggle. <https://www.kaggle.com/datasets/maltegrosse/8-m-spotify-tracks-genre-audio-features/>

Rodolfo Figueroa. (2020, December 22). Spotify 1.2M+ Songs [Data set]. Kaggle. <https://www.kaggle.com/datasets/rodolfofigueroa/spotify-12m-songs>

# Attachments