Spotify Audio Feature Clustering Analysis: Generalizing Hyper-specific Genre Classifications

Isaac Fry

December 7, 2022

Abstract

As one of the world's top music streaming services, Spotify's data infrastructure is immense and thoroughly engineered. Spotify's data backend boasts the ability to classify songs with an extreme degree of precision, assigning anywhere from 3 to 10 different genres to a single track. However, this hyper-specificity can hinder the communication of a track's genre to an average music listener.

My research utilizes K-Means Clustering on an open-source dataset that cross-references the Billboard Hot 100 charts from 1958-2020 to Spotify's audio features. By leveraging Principal Component Analysis on a set of Spotify audio features, I reduce 1042 unique genres into $k = 3$ and $k = 10$ genre clusters. Each cluster is defined by a relatively distinct accumulation of genres, and the cluster centers indicate the general archetypal song for that genre cluster.

There is immense opportunity for future work in the field of genre classifcation. By analyzing the frequency of genres within a cluster, a given hyper-specific genre in Spotify's extensive catalog can be better understood in relation to other genres. This insight can enable musicians to better understand the potential reach of a given song to previously unreached audience groups. Furthermore, these same techniques can be used for classifying different types of media with similar results.

1 Overview

As one of the world's top music streaming services, Spotify's data infrastructure is immense and thoroughly engineered. A strong data backend enables some of Spotify's flagship features, including song discoverability, artist suggestion, playlist generation, radio, marketing, and more. Furthermore, Spotify leverages this data to classify songs with an extreme degree of precision, assigning anywhere from 3 to 10 different genres to a single track.

Unlike other streaming services, like Apple Music, Pandora, or Amazon Music, Spotify maintains a widespread public access to their data backend through API calls. These APIs allow developers to access the same data that power Spotify's impressive array of music-oriented algorithms. Each track is described by a set of audio features as well as a list of wide-ranging and hyper-specific genres (see **Data Dictionary**). However, this hyper-specificity can hinder the communication of a track's genre to an average music listener. For instance, genres like "dirty south rap," "Hi-nrg," "skramz," "ottawa rap," and "rap conscient" are vague and undescriptive for the average music purveyor. This poses a problem in discerning how different genres interact with one another, and, more generally, how Spotify uses genre classification in its backend.

It should be noted that Spotify's API calls do not perform well for dataset generation (seeing as its API calls are designed for data retrieval). Any song must be called by a specific `Spotify ID`, and Spotify does not publicly list a dataset of songs that are each correlated with a `Spotify ID`. Thus, finding a dataset of songs with sufficiently interesting features is difficult.

Fortunately, an open source developer created a large dataset of 330000 entries with each *BillBoard Hot 100* chart from 1958-2020 and a corresponding dataset with information from a Spotify API call for each song [1]. This is one of the only large, publicly available datasets that is modern, includes a large number of songs, and links to a descriptive set of audio feature analysis.

The features included in the base dataset are described in the **Data Dictionary**. Furthermore, each song is linked to a list of Spotify-deigned genres, with 1042 unique genres in the dataset. All of these were previously called from one of Spotify's APIs [5, 4].

Other datasets exist besides the one cited above, and they may have a higher quality of information. In the audio data analysis community, the Million Song Dataset (MSD) has been used regularly and remains a staple of ongoing open source research [3, 7, 2]. The proprietor of the dataset, EchoNest, published MSD with data generated from Spotify between 2010-2017. At the time, another project, known as Rosetta Stone, allowed a researcher to translate an `EchoNest ID` into a `Spotify ID`, but after Spotify acquired EchoNest in 2016, Spotify absorbed many of the translation features into its private backend [8]. After that, MSD's insight for modern music was reduced, especially as Spotify continued to advance the data available through its privately-owned APIs.

Thus, a strong motivation exists for using the *Billboard Hot 100* due to its high availability of data. Since the *Billboard Hot 100* is a time series data set (with a new chart each week), an extra temporal component can be added to the analysis. Furthermore, using songs only from the *Billboard Hot 100* gives a highly representative sample of songs. While not every chart-topping song provides accurate insight into the musical feelings of the public, the collective sum of the top songs provides a generally healthy sample of the cultural trends of the time.

A few critical questions motivate this analysis:

- How can Spotify's extensive genre classification be reduced to more comprehensible genre groups?
- How do musical genres and audio features generally change over time?
- What audio features are most indicative of a track's genre?

2 Data Acquisition

This data includes a wide variety of features, all of which are detailed extensively in the **Data Dictionary**. Furthermore, these features are represented graphically as boxplots and as a changes over time in the **Appendix**.

As mentioned in the **Overview**, this data comes from an open source dataset [1] that cross-references the Billboard Hot 100 charts (circa 1960 - 2020) with Spotify's web API for audio feature analysis [6].

There are no limitations on sharing this data due to its open source nature.

3 Pre-processing

As a whole, this data set is remarkably clean. It would probably be beneficial to re-call each track's audio features from the Spotify API, but that requires a significant time investment that was outisde the scope of this project.

In order to maintain a generalized approach, the instance of a track is only kept when it reached its highest point on the Billboard Hot 100 charts. While the chart is not an objective measure of popularity, it does indicate a general societal appreciation for that music. Thus, the date when a track is most popular roughly indicates the peak of societal appreciation for that track.

Via this approach, the Billboard Hot 100 dataset was reduced from 327895 entries to 24280 unique songs after joining it with the valid Spotify Audio Features dataset. This still encapsulates the 3252 weeks available in the dataset. An unfortunate downside is that a given week may have 1 or 0 songs (i.e. no tracks peaked during that week), but the overwhelming scale of the dataset mitigates the influence those weeks may have.

As seen in Fig. 6, the distribution of tracks over time is fairly evenly distributed, indicating that a temporal component would be a valid feature to include in the model. Furthermore, the boxplots generated from the data describe a well-distributed and usable dataset (see **Boxplot Distributions**).

While the data from temporal analysis did not factor into the clustering analysis, the visualizations from that exploration are still included in this report for general amusement and curiosity (see **Temporal Analysis**).

4 Model Selection: `K-Means Clustering`

This analysis uses `K-Means Clustering` for my model, leveraging a standard implementation using Principal Component Analysis (PCA) with 2 components. The Spotify feature set is fairly extensive, which may lead to erroneous results. Furthermore, PCA with 2 components will enable the generation of comprehensible two-dimensional graphs.

The motivation for using `K-Means Clustering` stems from a few main advantages::

- K-Means performs well with large numbers of samples and small numbers of clusters, which fits the descriptions of my objectives.
- K-Means is optimized for geometric distances between points.
- The act of clustering will with K-Means will generate some type of nodal center. Seeing which tracks are closest to the center (the archetype for that center) can reveal an unexpected "average" for a given genre.
- Most of the songs in the feature set have a plain-text list of Spotify-assigned genres. Spotify's genre's are incredibly wide-ranging and hyper-specific, indicating that their backend algorithm for genre classification is advanced. However, the hyper-specific genres aren't necessarily approachable for an average music listener. Clustering songs and then cross-referencing Spotify's listed genres helps reveal the implicit similarities in Spotify's genre classification.

For this analysis, the feature set includes: Duration (ms), Explicit, Danceability, Energy, Key, Loudness (db), Mode, Speechiness, Acousticness, Instrumentalness, Liveness, Valence, Tempo, and Time Signature (see **Data Dictionary** for more details).

5 Results and Evaluation

The reader is encouraged to look at the figures prior to reading this analysis. All figures are included in the **Appendix**. It might be worthwhile to have two copies of the document available to cross-reference the Genre Clusters and Genre Groups.

Date was excluded to eliminate the influence of non-audio information.

Below are the steps used to arrive at this technique:

- Using the dataset from **Pre-processing**, each feature in the feature space was scaled using the Standard Scaler from `scikit-learn`. This helped improve the intuitiveness of visualizations despite a reduction in silhouette scores.
- A series of PCA component tests was performed to determine both the optimal number of components and optimal number of clusters. I found PCA $= 2$, $k = 3$ provided the highest silhouette score (see Fig. 1).
- With the modified data from the above processes, I ran K-Means clustering for $k = 3$ (optimal) and $k = 10$ (arbitrary). Each trial was then visualized as a scatter plot.
- Each group was then parsed to find the relative frequency of genres within a that group. This helped highlight the genres that were abnormally frequent compared to the baseline frequency of that genre in the entire dataset.
- The 15 highest relative frequency genres in each cluster were visualized as bar graphs.
- The cluster centers (i.e. the "average" song) for each genre space was determined.
- The prominence of a genre space over time was visualized as a time series (not included in this analysis).

The above was completed for the entire dataset (1960s - 2021) and the 2010s (2010 - 2021).

5.1 Validation Metrics

As an unsupervised learning model, K-Means Clustering is difficult to validate. However, there were a few ways I verified the efficacy of my techniques:

- Silhouette scores were used to determine that a PCA with 2 components and $k = 3$ are optimal hyperparameters (see Fig. 1). A silhouette score of 0.5 is not exceptional, but it performs better than an analysis without any dimensionality reduction.
- Each feature passed into the PCA was analyzed for explained variance. Features that varied over time (see **Temporal Analysis**) and that describe certain genres had the highest explained variance (see Fig. 2).
- Musical intuition was employed to generally verify the likelihood that a cluster center would describe a genre center. Of course, the "average" song for a cluster may not be the genre-defining song per a music critic, but each cluster center generally fits its assigned genre space.

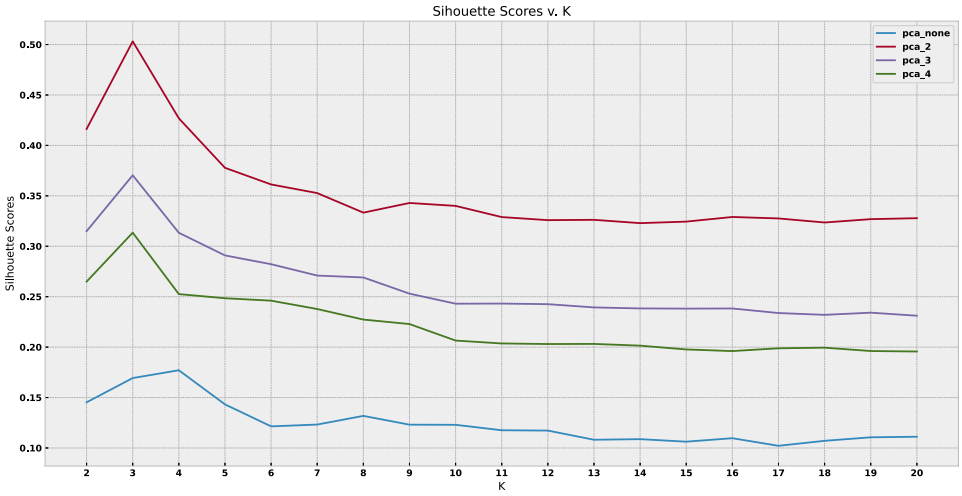

Figure 1: Silhouette scores for PCAs with various numbers of components.

	Duration (ms)	Explicit	Danceability	Energy	Key	\
pca1	-0.145162	-0.271096	-0.332995	-0.471340	-0.039513	
pca2	0.086259	0.568112	0.066847	-0.335607	0.060582	

	Loudness (db)	Mode	Speechiness	Acousticness	Instrumentalness	\
pca1	-0.409061	0.175897	-0.275937	0.445833	0.049230	
pca2	-0.094595	-0.190802	0.488759	0.112881	-0.128162	

	Liveness	Valence	Tempo	Time Signature
pca1	-0.016532	-0.183947	-0.076230	-0.235165
pca2	-0.029132	-0.448811	-0.150941	-0.102794

Figure 2: Explained variance for each feature in a PCA with 2 components.

5.2 General Results

Using the entire dataset and $k = 3$ (Fig. 3), I found that there were 3 fairly distinct genre groups. Group 1 (Fig. 35) seemed to primarily consist of rap/hip-hop tracks, while Group 0 (Fig. 34) contained pop and rock from the 60s and 70s, and Group 2 (Fig. 36) held more of the older rock tracks (Fig. 35).

This general grouping held true for the $k = 10$ trial (Fig. 4). For instance, the upper left groups (Groups 2, 5 and 6) in the $k = 10$ trial are all rap/hip-hop genres (Fig. 39, Fig. 42, Fig. 43). Some more specific genres emerge, such as country in Groups 0 and 4 (Fig. 37, Fig. 41).

Some of cluster centers do not seem to accurately represent their genre space, but that's most likely due to coincidence. For instance, the center for Group 0 in the $k = 10$ trial is a synthpop song despite it being in a generally country genre space (Fig. 37).

However, there are plenty of instances where the cluster center accurately describes its cluster. For instance, Group 6's center is *Lose You* by Drake, and some of Group 6's descriptive genres are Toronto Rap and Canadian Pop (Fig. 56). Or, for instance, *Redneck Crazy* by Tyler Farr is the song center for the Group 4 country genre space (Fig. 54).

There are also some instances where some audio features may have been incorrectly gathered. For instance, in the $k = 10$ trial for the 2010s (Fig. 5), there are 10 songs that are greater than 6 for the `pca1` feature (light green far right, Fig. 55). Some of these songs are very clearly outliers, like *Forward* by Beyonce and James Blake, or *I Dreamed a Dream* from the *Les Miserables* movie musical. However, there's also *close* by J. Cole, which would very easily fit in the Group 2, 5, or 6 genre space. Thus, either the predictions can have extreme mis-classifications, or there's some incorrectly wrangled data.

Overall, this analysis helps consolidate the Spotify's sprawling genre list to comprehensible genre spaces. Ideally, this would be represented as an interactive platform, but static graphs and plain text will hopefully communicate the essence of the idea.

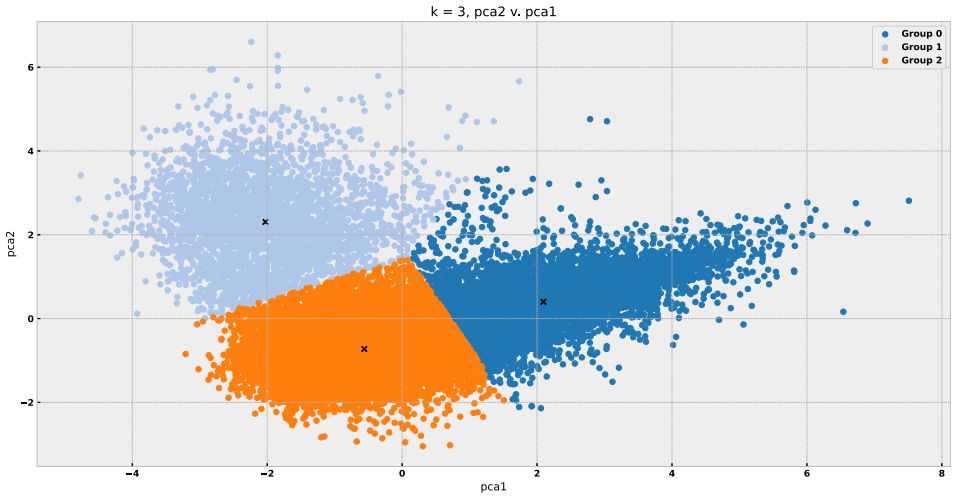

Figure 3: Clustering scatter plot with $k = 3$, full dataset

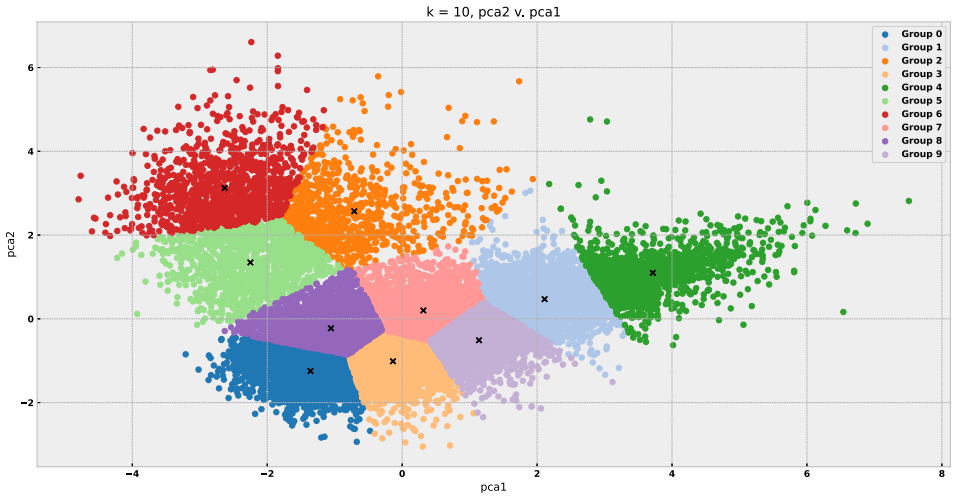

Figure 4: Clustering scatter plot with $k = 10$, full dataset

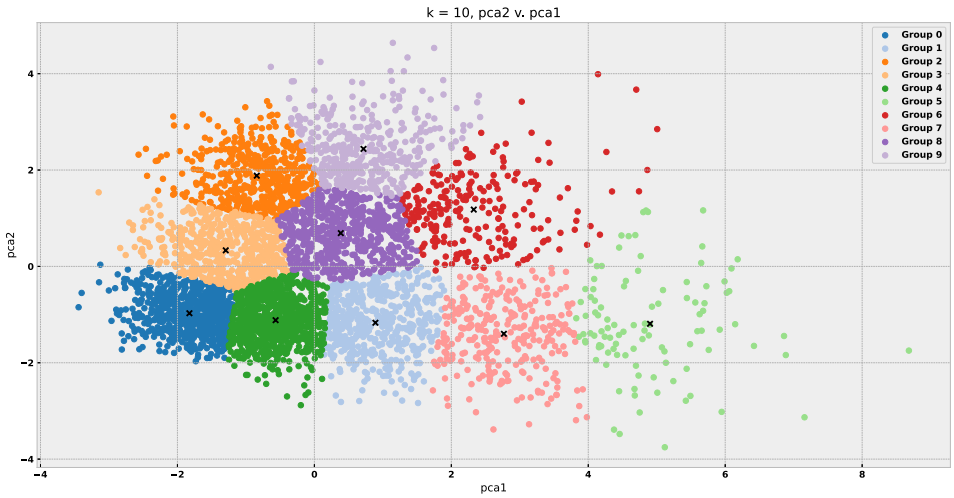

Figure 5: Clustering scatter plot with $k = 10$, 2010s

5.3 Specific and Interesting Results

- From the entire dataset, the most average song is *The Same Love That Made Me Laugh* by Bill Withers. It's interesting that a soul/funk song would be the most average song from the 1960s to the 2021s.
- From just the 2010s, the most average song is *Sex Room* by Ludacris. This also helps highlight how vague "Dirth South Rap" is as a genre for the layman. (After some research, Dirty South Rap is a rap region style, akin to West Coast and East Coast.)
- Each artist seems to be linked to a specific geographic area regardless if that track is representative of that style. For instance, the cluster center for 2010s $k = 10$ Group 9 (Fig. 59) is *Fr Fr* by Wiz Khalifa and Lil Skies, both from the Pittsburgh scene. *Fr Fr* is marked as Pittsburgh Rap *and* Southern Hip Hop, which are clearly incongruent.

- Similarly, all artists seem to have genre tags they're assigned to regardless of the audio qualities of the track. For instance, the center for 2010s $k = 10$ Group 5 (Fig. 55) is *Nothing Like Us* by Justin Bieber. *Nothing Like Us* is an original piano ballad from an acoustic album, but it's marked as Dance Pop.
- Billie Eilish is fascinating outlier. In the 2010s $k = 10$ trial (Fig. 5), Group 5 is a conglomerate of multiple fringe genres (Fig. 55, Fig. 33). However, out of the 10 tracks greater than `pca1` $= 6$, 4 of those are Billie Eilish: *Xanny*, *Listen Before I Go*, *I Love You*, and *everything i wanted*. The first three are from *WHEN WE ALL FALL ASLEEP, WHERE DO WE GO?*, and it's likely that they were dragged up to the charts from the streaming popularity of the entire album rather than their own merit. However, *everything i wanted* was released as a single and presumably climbed the charts on its own merit.
- Lil Wayne is absurdly average. In the both the 2010s $k = 3$ **and** and 2010s $k = 10$, he appears as an artist in the cluster center. In the entire dataset $k = 10$, he appears as the center for two different genre spaces. Perhaps this is just a product of Lil Wayne having a "prolific" rate of output, so it's more likely that he'll appear in the dataset more often. Or, it might be because Lil Wayne is an extremely average rapper.

6 Appendix

6.1 Data Dictionary

hot 100 url, string

Origin: Billboard Hot 100

Billboard Hot 100 URL used to scrape data

WeekID, datetime

Origin: Billboard Hot 100

Day that the weekly chart was published, as a datetime object

---

Week Positon, integer

Origin: Billboard Hot 100

Current position corresponding with the WeekID

Song, string

Origin: Billboard Hot 100

Title of the track

Performer, string

Origin: Billboard Hot 100

Name of the performer/artist listed on the Billboard Hot 100

SongID, string

Origin: Billboard Hot 100

Concatenation of Perfomer and Song to create a unique ID

Instance, integer

Origin: Billboard Hot 100

Indicates how many times a song has appeared on the Hot 100 Billboard chart (i.e. if a song was on the chart, fell off the chart, and then returned, it would have a value of 2)

Previous Week Position, integer

Origin: Billboard Hot 100

Position of the song on the previous Hot 100 Billboard chart

Peak Position, integer

Origin: Billboard Hot 100

Highest position attained by the song as of the corresponding week

---

Weeks on Chart, integer

Origin: Billboard Hot 100

Weeks on the chart as of the corresponding week

Genres, string

Origin: Spotify

A list of the genres the artist is associated with. If not yet classified, the array is empty

---

spotify track id, string

Origin: Spotify

The Spotify ID for the track

---

spotify track preview url, string

Origin: Spotify

The preview URL for the song on Spotify

Duration (ms), integer

Origin: Spotify

The length of the track in ms

Explicit, boolean

Origin: Spotify

Whether or not the track has explicit lyrics (true = yes it does; false = no it does not OR unknown)

Album, string

Origin: Spotify

The album on which the track appears

Danceability, float

Origin: Spotify

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

---

Energy, float

Origin: Spotify

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.

Key, float

Origin: Spotify

The key the track is in. Integers map to pitches using standard Pitch Class notation. E.g. 0 = C, 1 = C#/Db, 2 = D, and so on. If no key was detected, the value is -1.

Loudness (db), float

Origin: Spotify

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.

Mode, float

Origin: Spotify

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.

Acousticness, float

Origin: Spotify

A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic.

Speechiness, float

Origin: Spotify

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.

Instrumentalness, float

Origin: Spotify

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.

Liveness, float

Origin: Spotify

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.

Valence, float

Origin: Spotify

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).

Tempo, float

Origin: Spotify

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.

Time Signature, integer

Origin: Spotify

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".

Popularity, integer

Origin: Spotify

The popularity of a track is a value between 0 and 100, with 100 being the most popular. The popularity is calculated by algorithm and is based, in the most part, on the total number of plays the track has had and how recent those plays are. Generally speaking, songs that are being played a lot now will have a higher popularity than songs that were played a lot in the past. Duplicate tracks (e.g. the same track from a single and an album) are rated independently. Artist and album popularity is derived mathematically from track popularity. Note: the popularity value may lag actual popularity by a few days: the value is not updated in real time.

6.2 Temporal Analysis

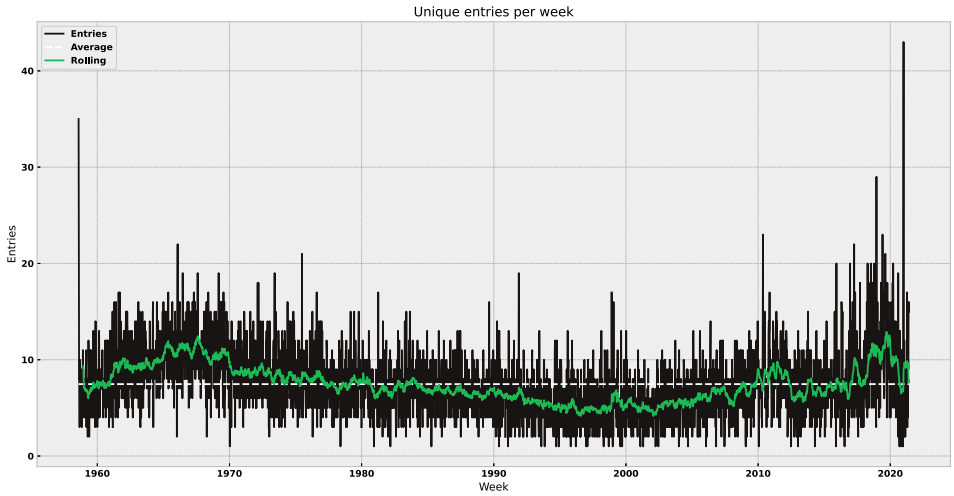

Figure 6: Distribution of unique entries from the Hot 100 Billboard chart during over time.

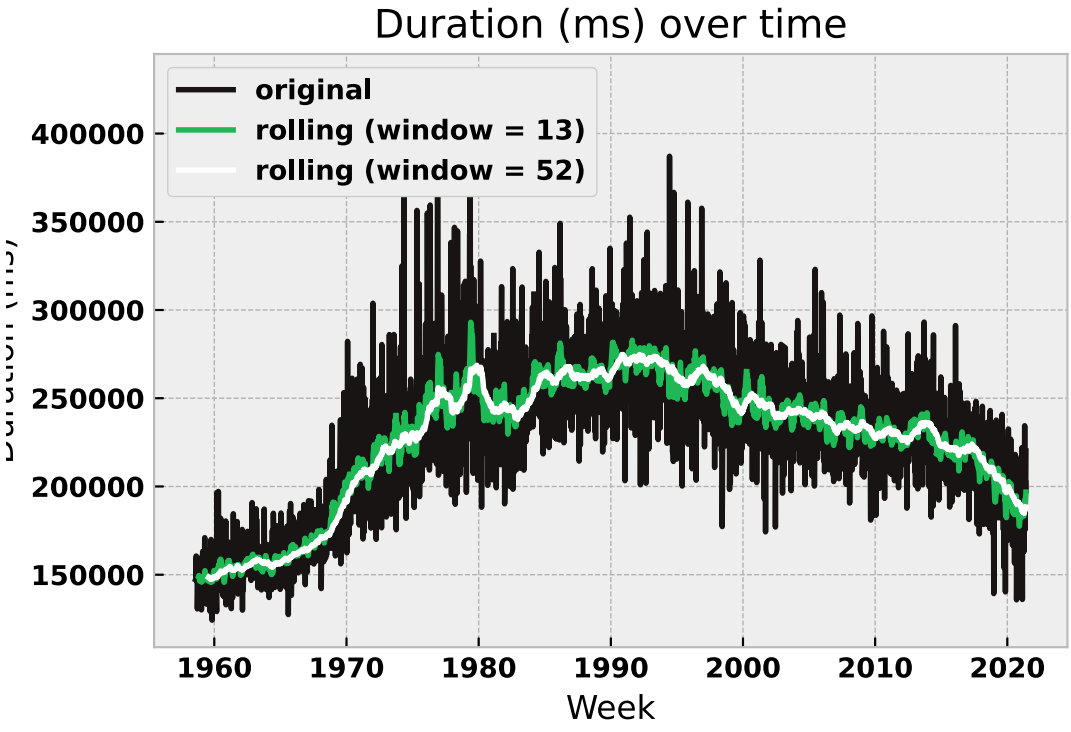

Figure 7: Analysis of Duration (ms) averages of the tracks that peaked on the Hot 100 Billboard Chart during a given week. See the data dictionary section for a more detailed description of this feature.

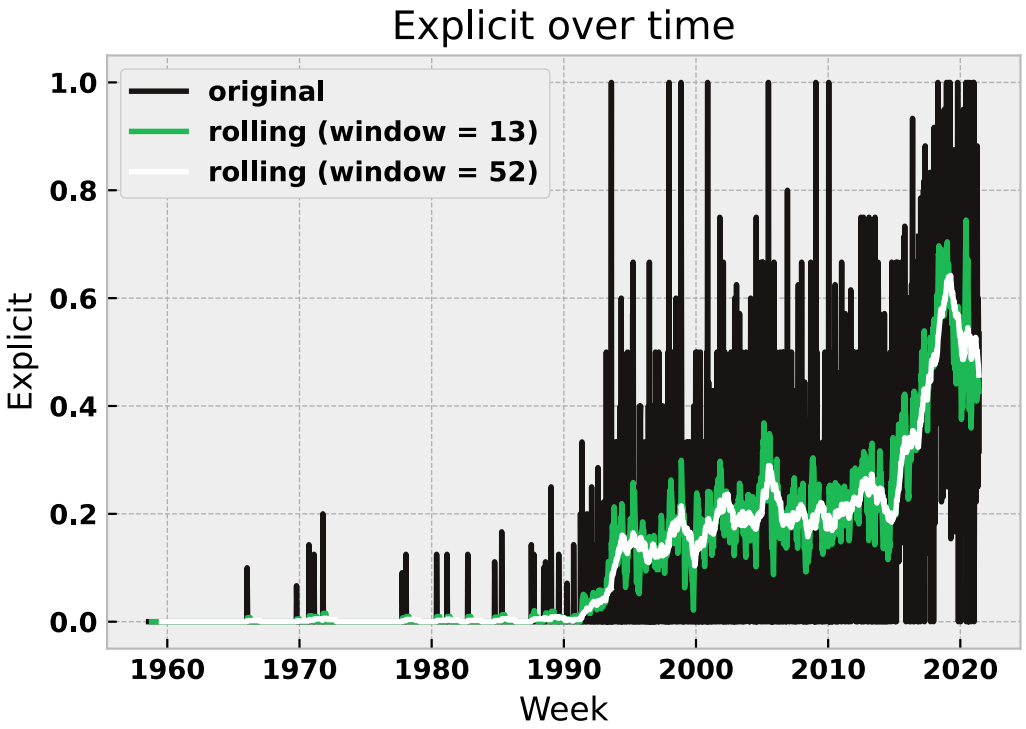

Figure 8: Analysis of Explicit averages of the tracks that peaked on the Hot 100 Billboard Chart during a given week. See the data dictionary section for a more detailed description of this feature.

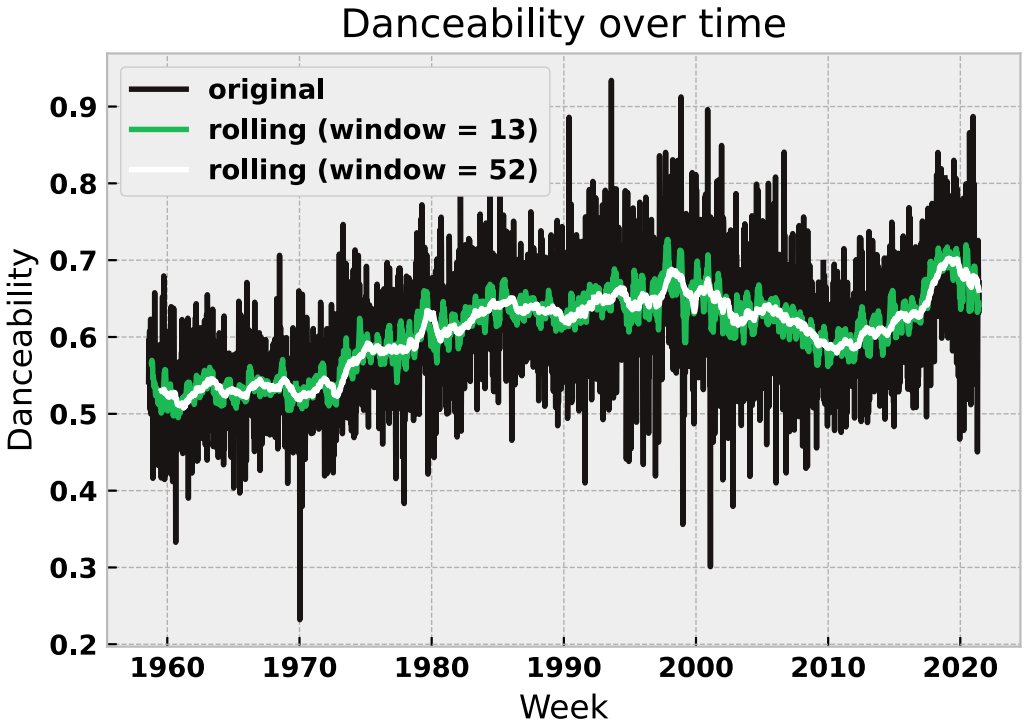

Figure 9: Analysis of Danceability averages of the tracks that peaked on the Hot 100 Billboard Chart during a given week. See the data dictionary section for a more detailed description of this feature.

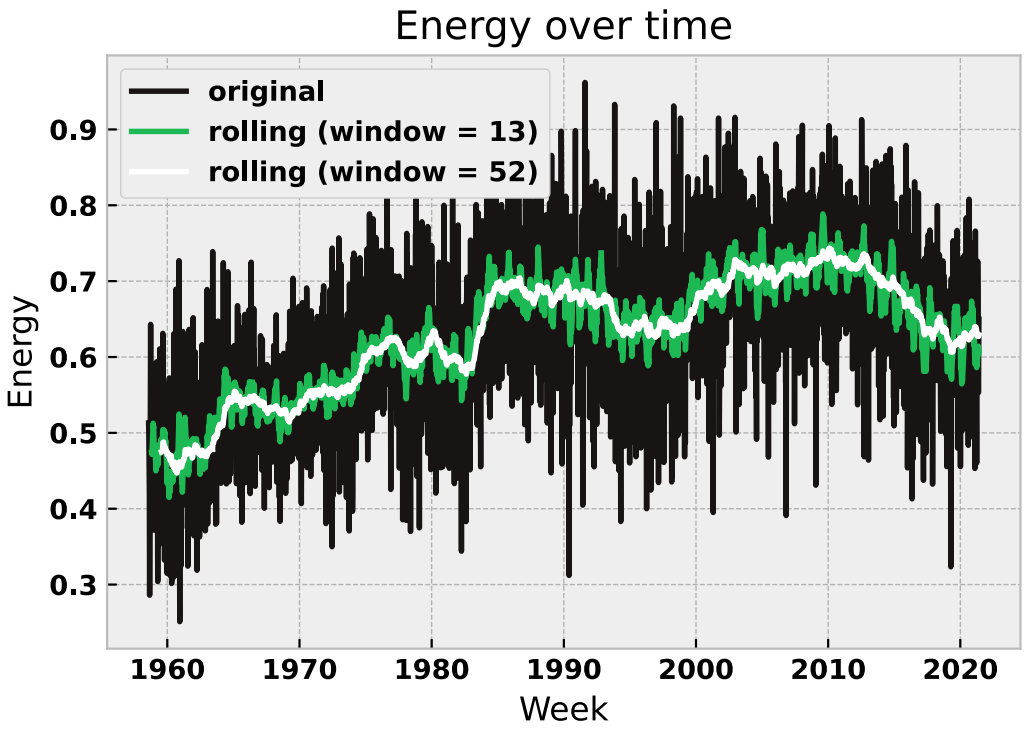

Figure 10: Analysis of Energy averages of the tracks that peaked on the Hot 100 Billboard Chart during a given week. See the data dictionary section for a more detailed description of this feature.

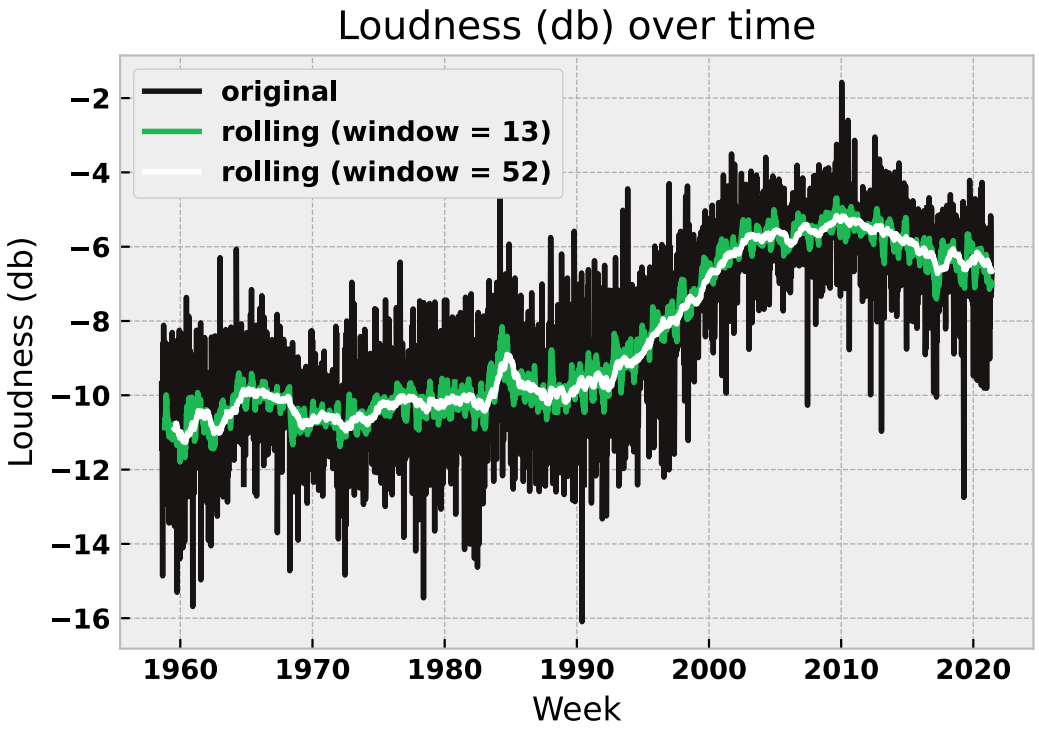

Figure 11: Analysis of Loudness (db) averages of the tracks that peaked on the Hot 100 Billboard Chart during a given week. See the data dictionary section for a more detailed description of this feature.

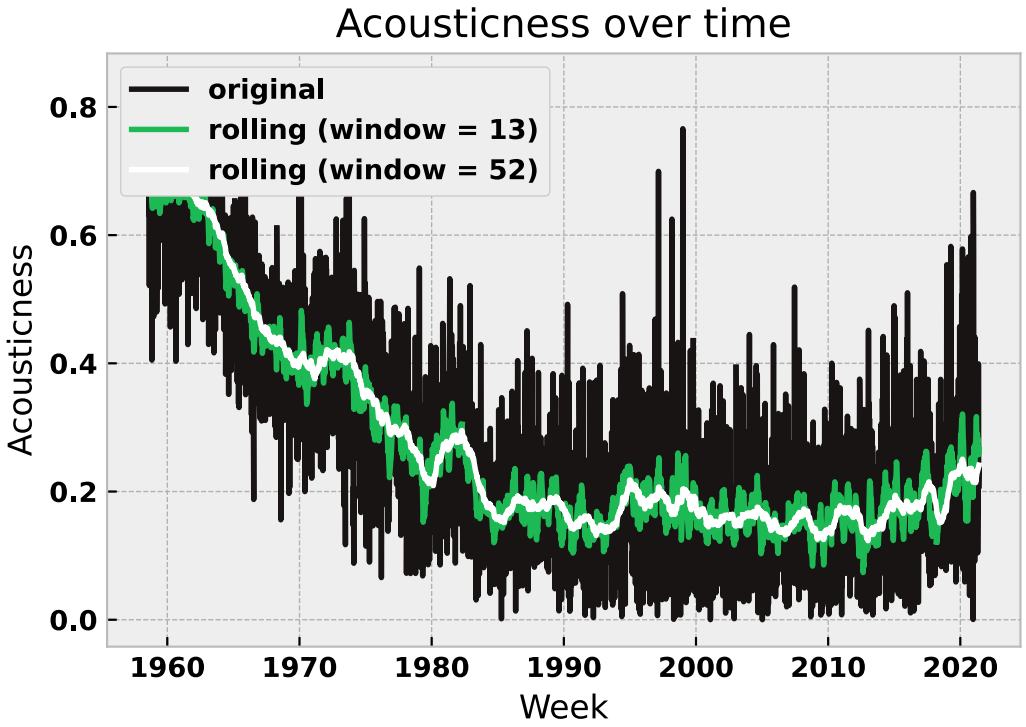

Figure 12: Analysis of Acousticness averages of the tracks that peaked on the Hot 100 Billboard Chart during a given week. See the data dictionary section for a more detailed description of this feature.

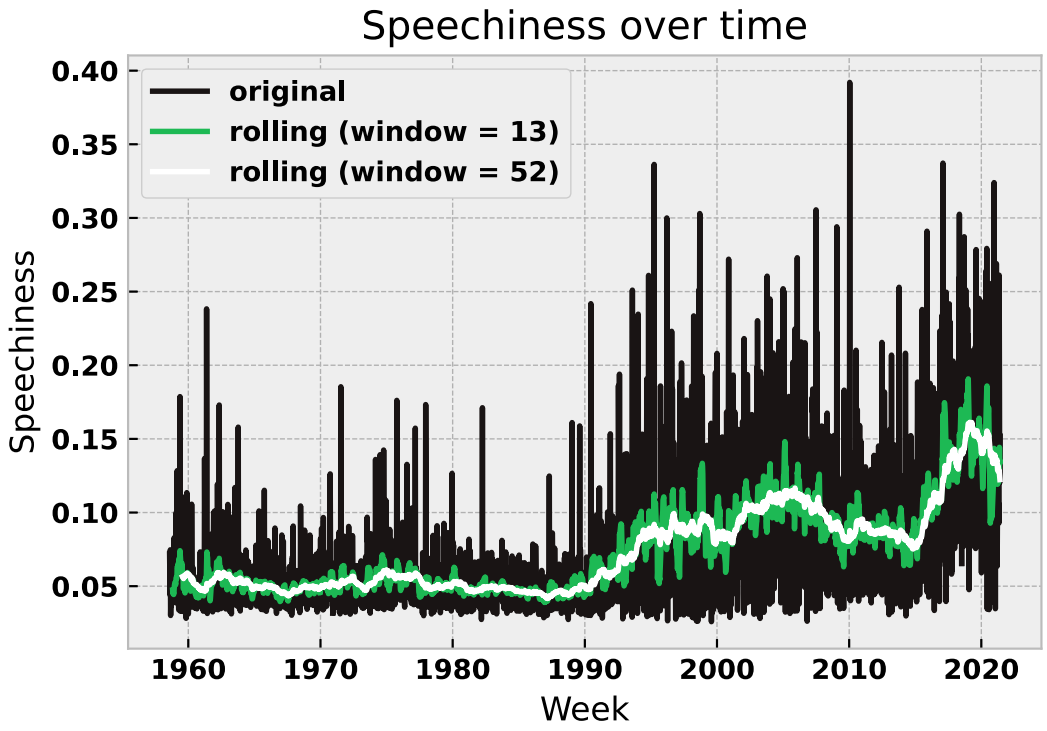

Figure 13: Analysis of Speechiness averages of the tracks that peaked on the Hot 100 Billboard Chart during a given week. See the data dictionary section for a more detailed description of this feature.

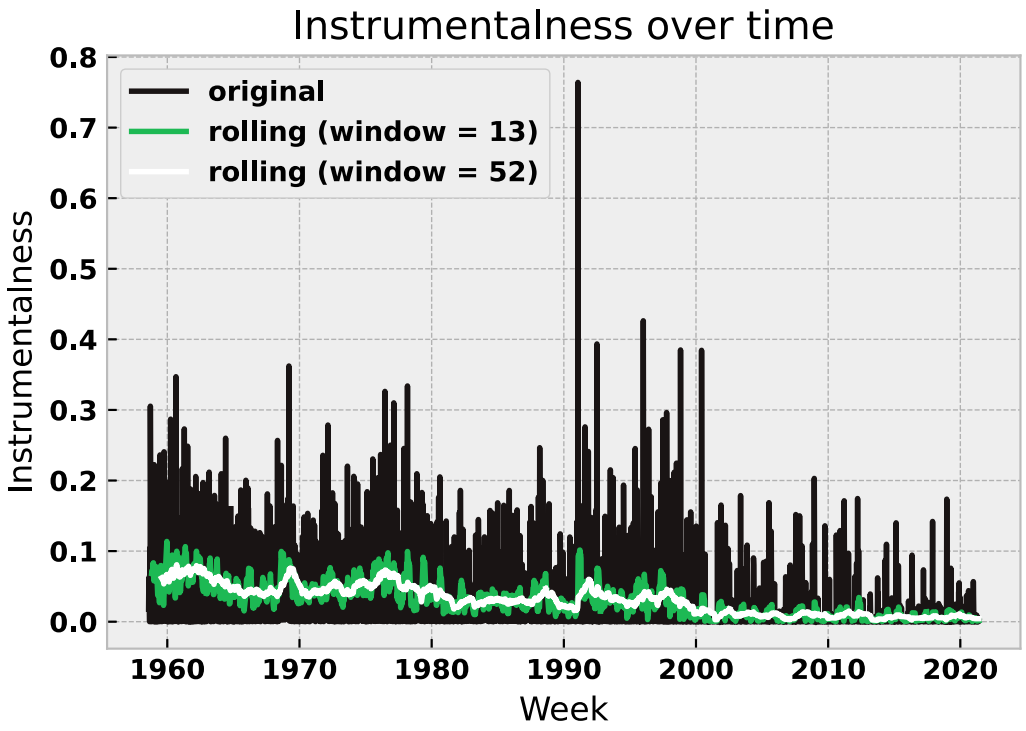

Figure 14: Analysis of Instrumentalness averages of the tracks that peaked on the Hot 100 Billboard Chart during a given week. See the data dictionary section for a more detailed description of this feature.

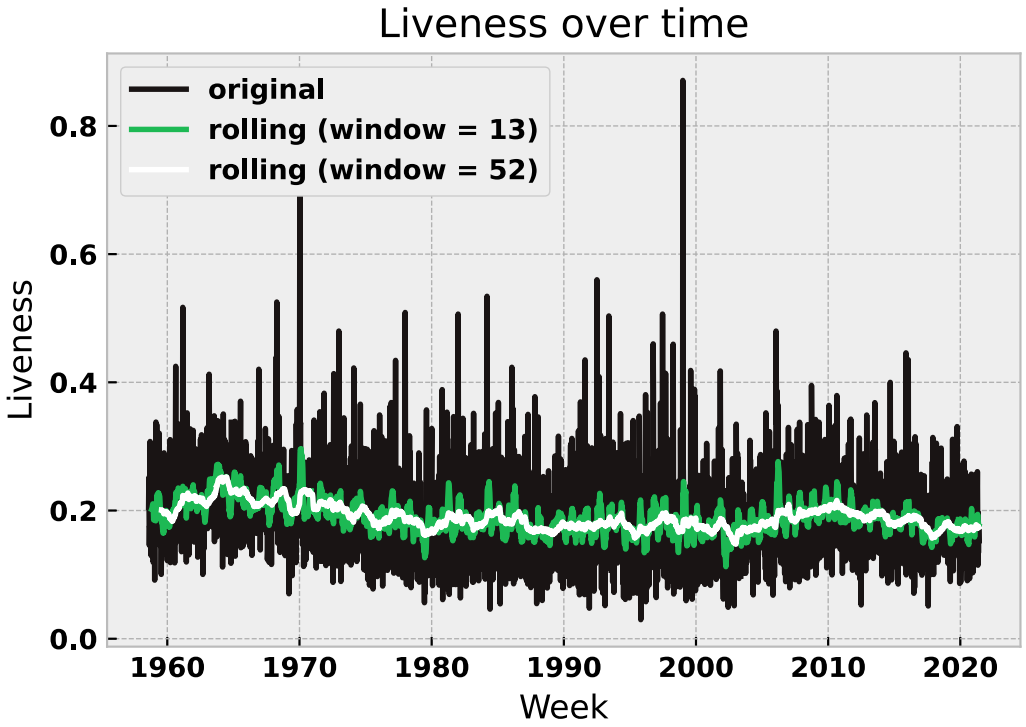

Figure 15: Analysis of Liveness averages of the tracks that peaked on the Hot 100 Billboard Chart during a given week. See the data dictionary section for a more detailed description of this feature.

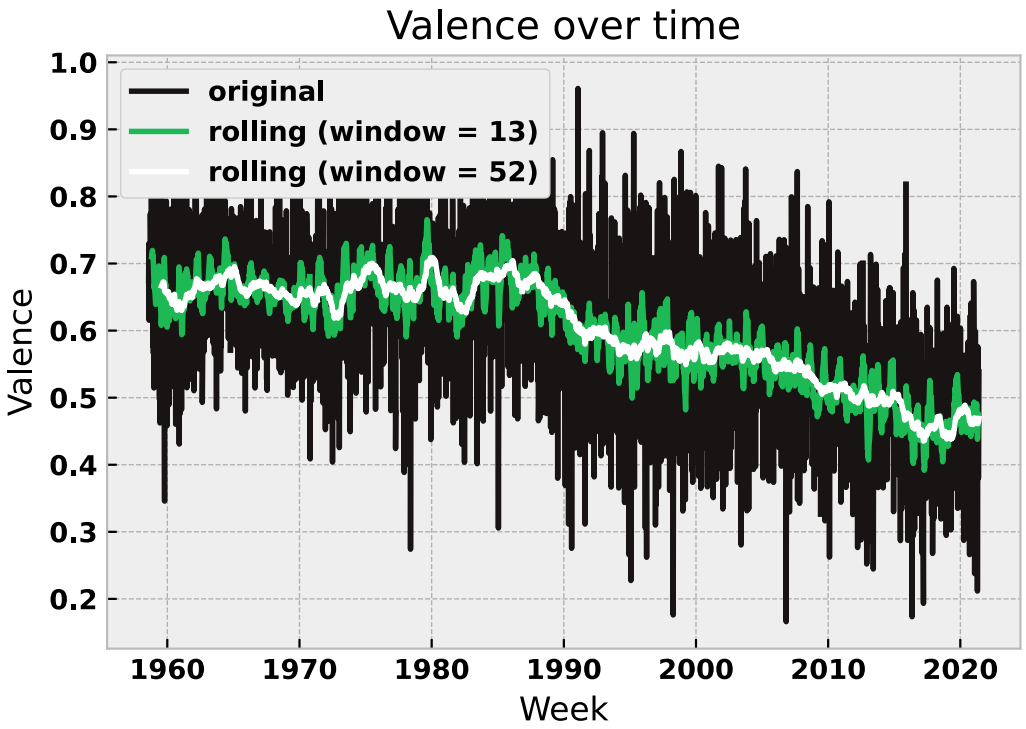

Figure 16: Analysis of Valence averages of the tracks that peaked on the Hot 100 Billboard Chart during a given week. See the data dictionary section for a more detailed description of this feature.

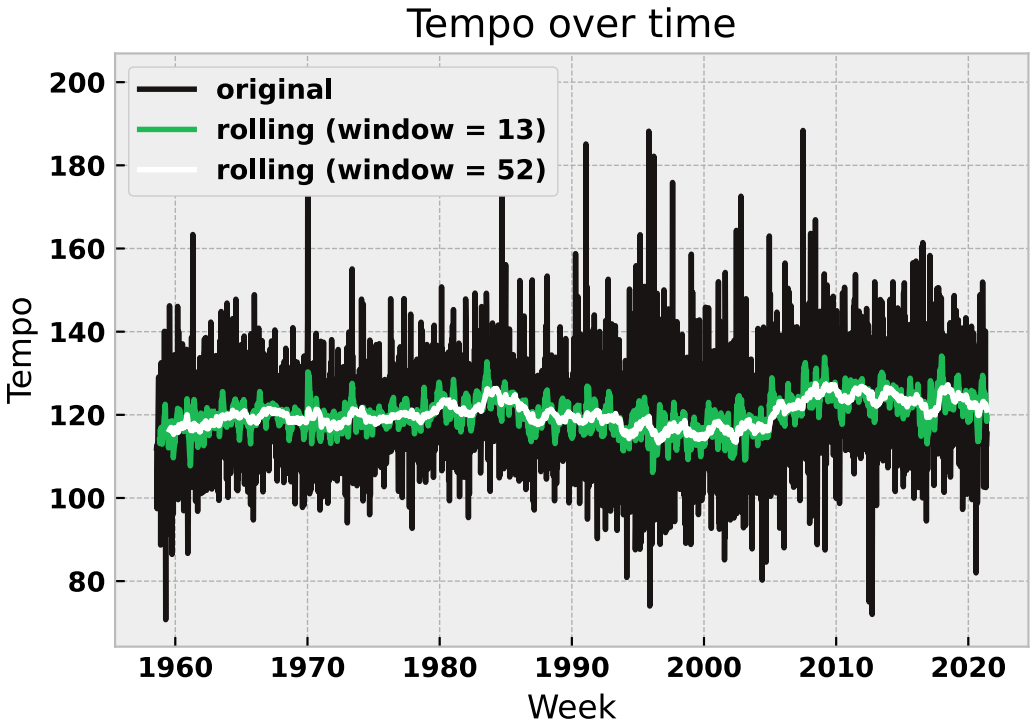

Figure 17: Analysis of Tempo averages of the tracks that peaked on the Hot 100 Billboard Chart during a given week. See the data dictionary section for a more detailed description of this feature.

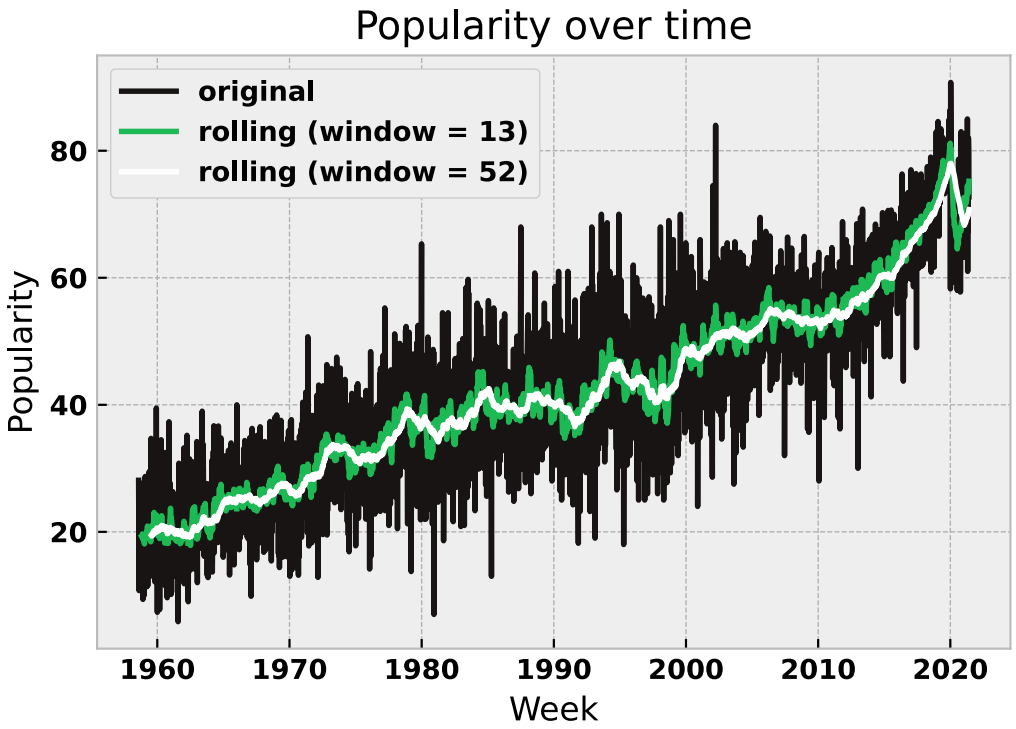

Figure 18: Analysis of Popularity averages of the tracks that peaked on the Hot 100 Billboard Chart during a given week. See the data dictionary section for a more detailed description of this feature.

6.3 Boxplot Distributions

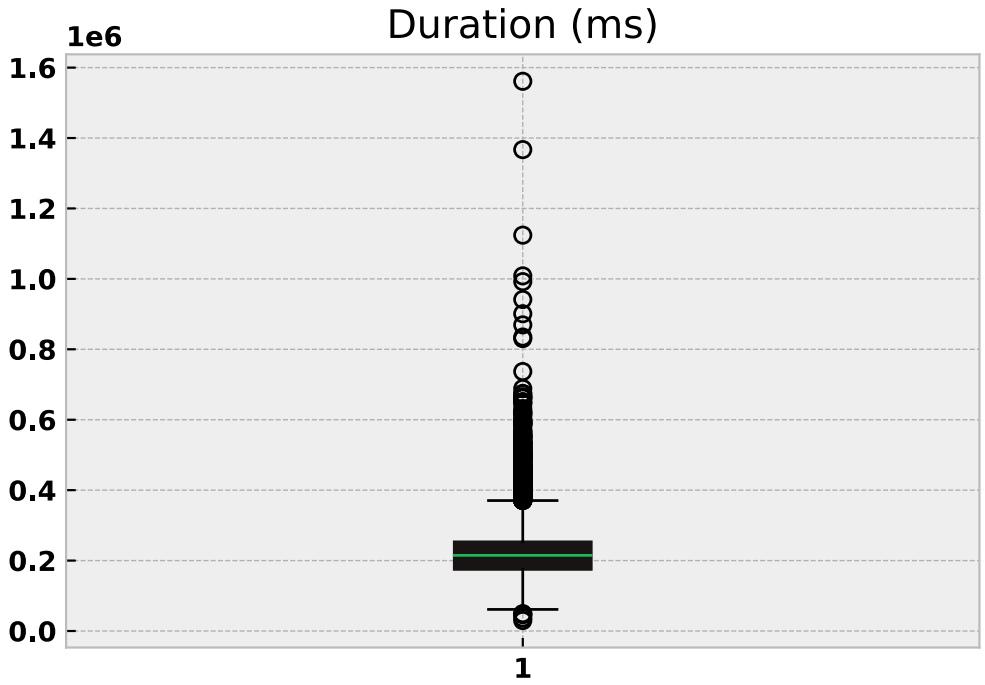

Figure 19: Analysis of Duration (ms) distributions of the tracks in the Billboard Hot 100 charts. See the data dictionary section for a more detailed description of this feature.

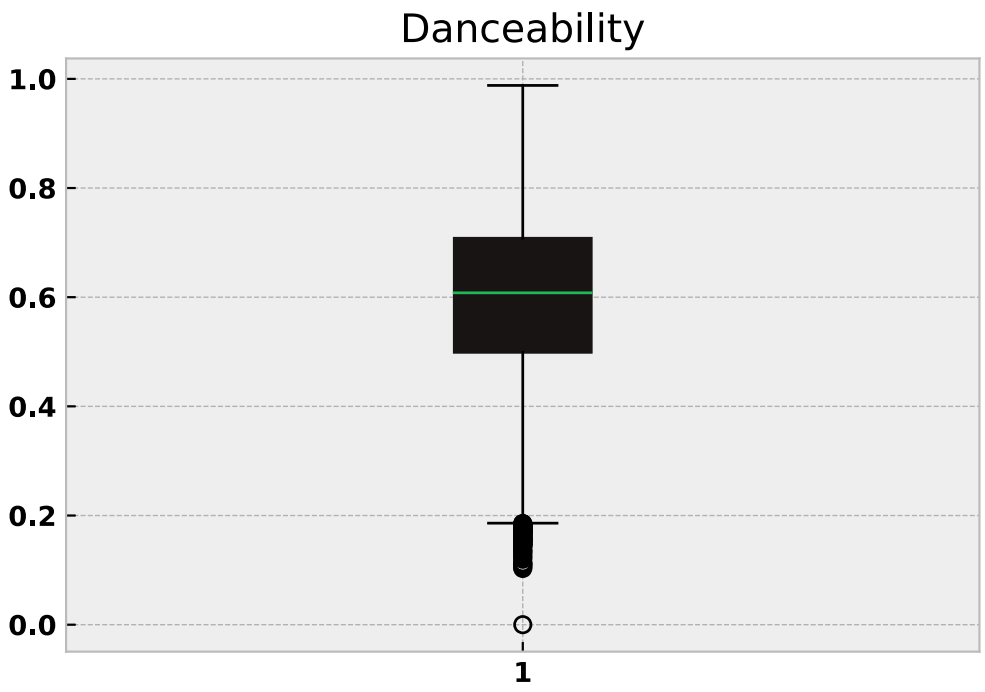

Figure 20: Analysis of Danceability distributions of the tracks in the Billboard Hot 100 charts. See the data dictionary section for a more detailed description of this feature.

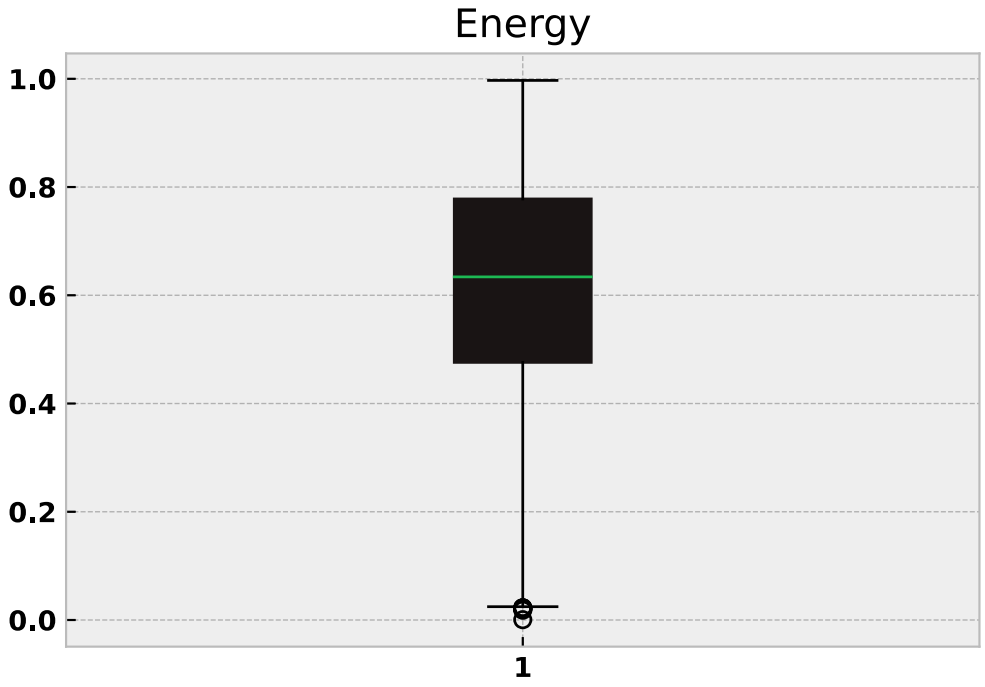

Figure 21: Analysis of Energy distributions of the tracks in the Billboard Hot 100 charts. See the data dictionary section for a more detailed description of this feature.

Figure 22: Analysis of Loudness (db) distributions of the tracks in the Billboard Hot 100 charts. See the data dictionary section for a more detailed description of this feature.

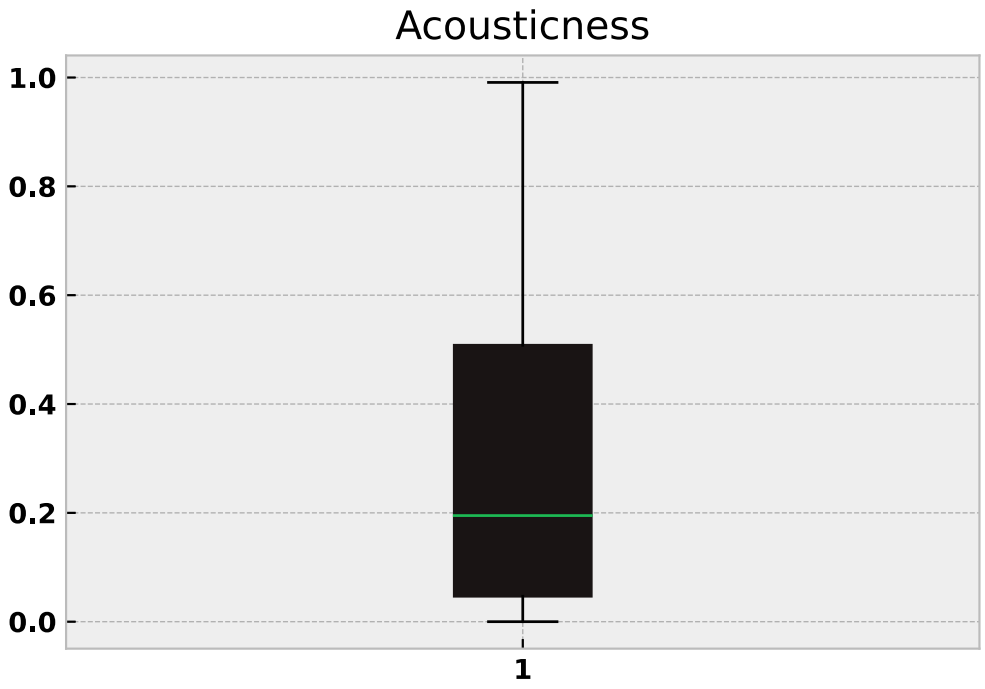

Figure 23: Analysis of Acousticness distributions of the tracks in the Billboard Hot 100 charts. See the data dictionary section for a more detailed description of this feature.

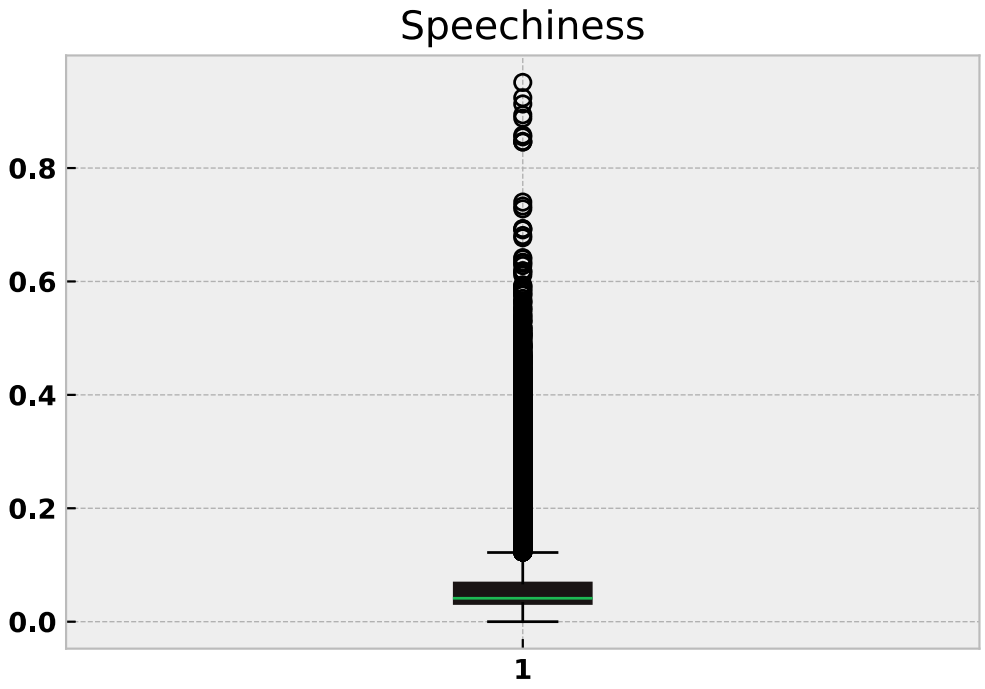

Figure 24: Analysis of Speechiness distributions of the tracks in the Billboard Hot 100 charts. See the data dictionary section for a more detailed description of this feature.

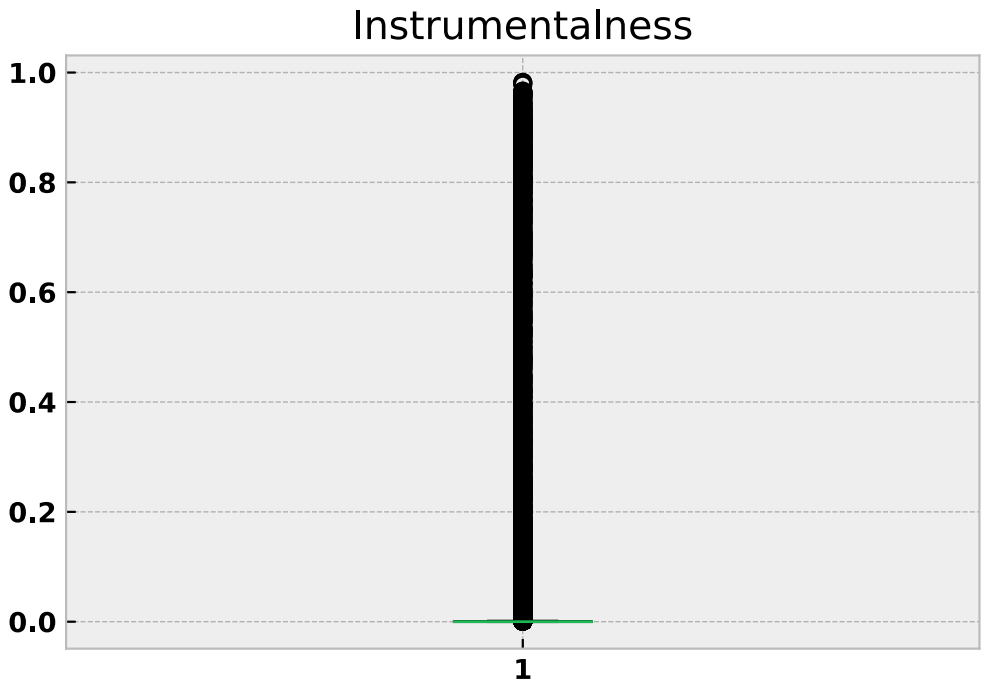

Figure 25: Analysis of Instrumentalness distributions of the tracks in the Billboard Hot 100 charts. See the data dictionary section for a more detailed description of this feature.

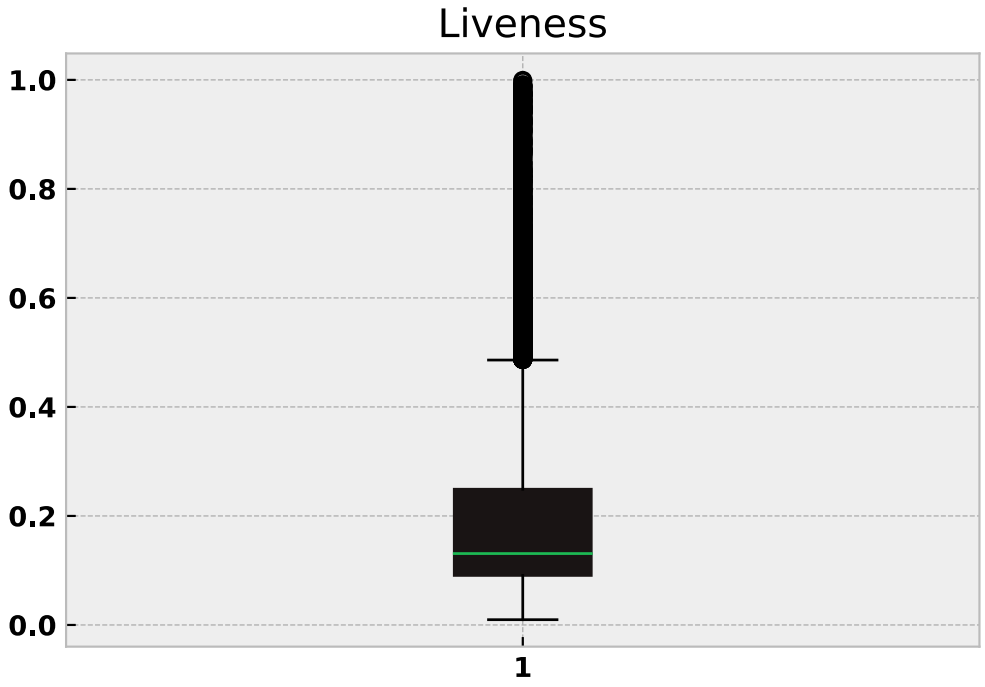

Figure 26: Analysis of Liveness distributions of the tracks in the Billboard Hot 100 charts. See the data dictionary section for a more detailed description of this feature.

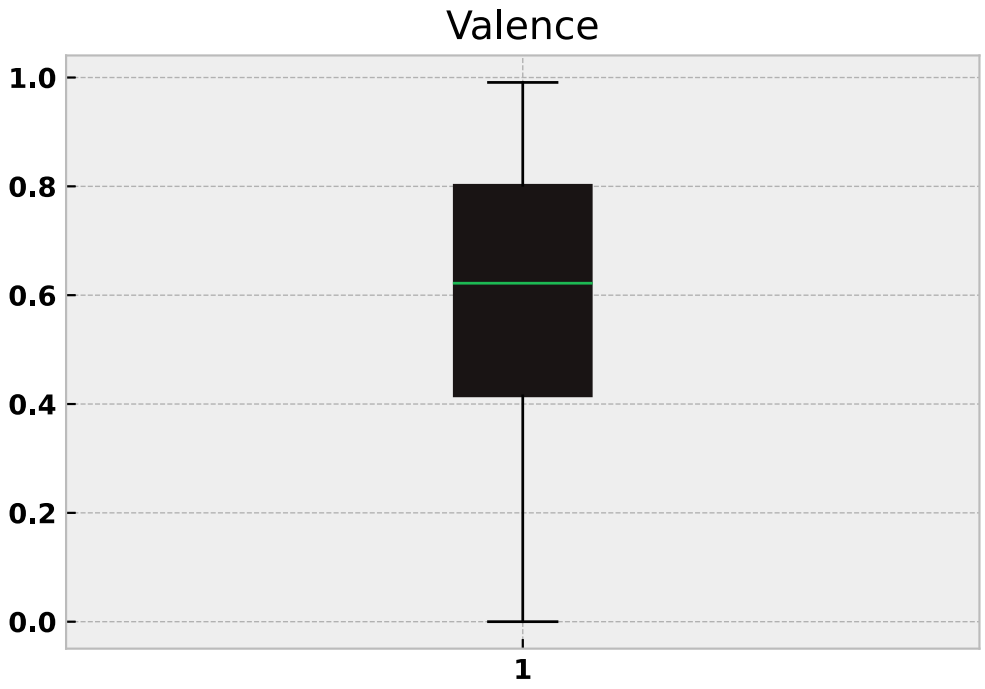

Figure 27: Analysis of Valence distributions of the tracks in the Billboard Hot 100 charts. See the data dictionary section for a more detailed description of this feature.

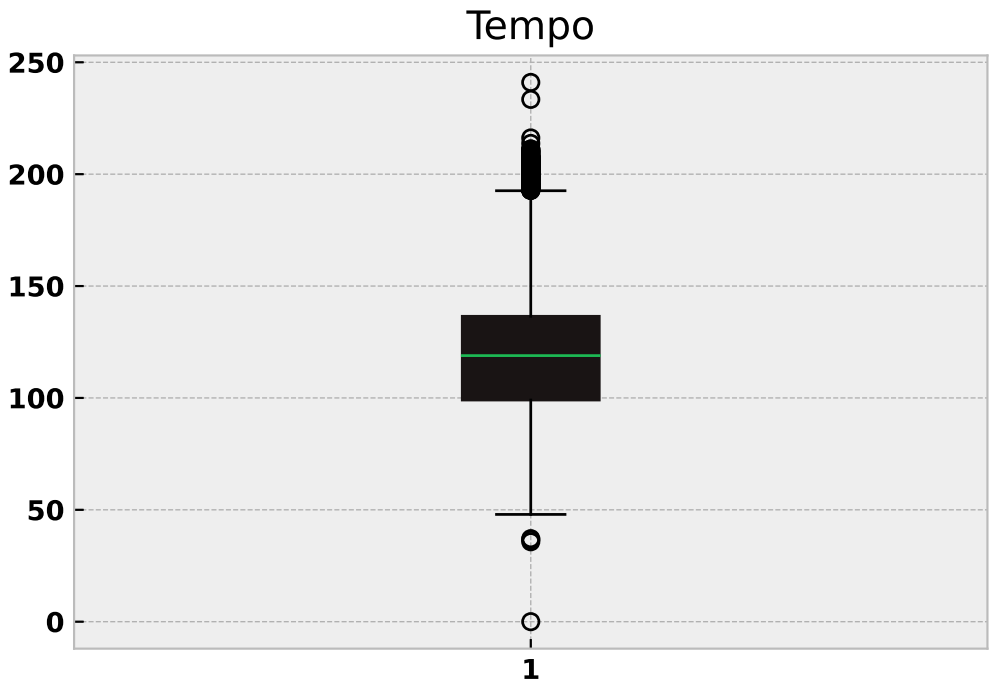

Figure 28: Analysis of Tempo distributions of the tracks in the Billboard Hot 100 charts. See the data dictionary section for a more detailed description of this feature.

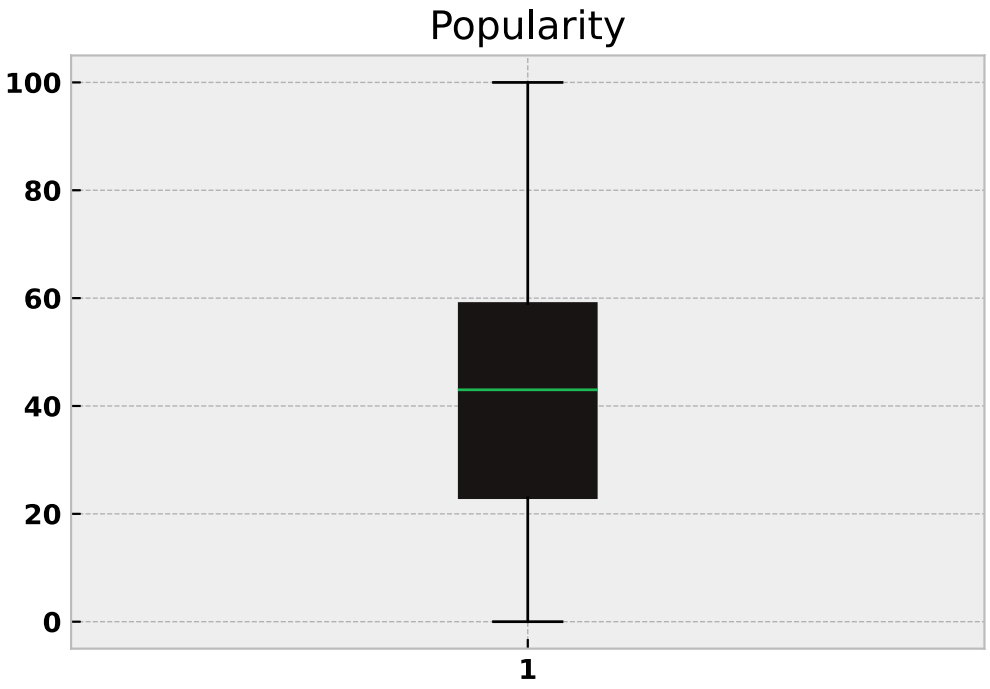

Figure 29: Analysis of Popularity distributions of the tracks in the Billboard Hot 100 charts. See the data dictionary section for a more detailed description of this feature.

6.4 Genre Centers

6.4.1 k = 1, full dataset

==========

Group 0: The Same Love That Made Me Laugh, by Bill Withers

Genre list for cluster center: ['classic soul', 'funk', 'motown', 'quiet storm', 'soul']

Coordinates: (-8.336121292975984e-16, -4.979578279536245e-17)

6.4.2 k = 3, full dataset

==========

Group 0 (Fig. 34): See You In September, by The Tempos

Genre list for cluster center: []

Coordinates: (2.0947419166334798, 0.40120042496312497)

==========

Group 1 (Fig. 35): Umma Do Me, by Rocko

Genre list for cluster center: ['atl hip hop', 'atl trap', 'deep southern trap', 'dirty south rap', 'gangster rap', 'pop rap', 'rap', 'southern hip hop', 'trap']

Coordinates: (-2.0268926619062886, 2.3069156404945494)

==========

Group 2 (Fig. 36): Nothin' At All, by Heart

Genre list for cluster center: ['folk-pop', 'indie folk', 'indie pop', 'neo mellow', 'new americana', 'seattle indie', 'stomp and holler']

Coordinates: (-0.5612625052394398, -0.7262582040259067)

6.4.3 k = 10, full dataset

==========

Group 0 (Fig. 37): Another Minute, by Cause And Effect

Genre list for cluster center: ['neo-synthpop', 'synthpop']

Coordinates: (-1.3586792126736627, -1.2465001887967935)

==========

Group 1 (Fig. 38): Tonight I Wanna Be Your Man, by Andy Griggs

Genre list for cluster center: ['country', 'country road', 'modern country rock']

Coordinates: (2.112331506526272, 0.4701100314096748)

==========

Group 2 (Fig. 39): Cudi Montage, by KIDS SEE GHOSTS

Genre list for cluster center: ['hip hop', 'rap']

Coordinates: (-0.7099272696299451, 2.569104124593737)

==========

Group 3 (Fig. 40): Cars, by Gary Numan

Genre list for cluster center: ['art pop', 'art rock', 'dance rock', 'industrial rock', 'new romantic', 'new wave', 'new wave pop', 'permanent wave', 'post-punk', 'synthpop']

Coordinates: (-0.13570878637768932, -1.0129694035993908)

==========

Group 4 (Fig. 41): Like Strangers, by The Everly Brothers

Genre list for cluster center: ['adult standards', 'brill building pop', 'bubblegum pop', 'folk rock', 'lounge', 'mellow gold', 'rock-and-roll', 'rockabilly', 'sunshine pop']

Coordinates: (3.7157936975812467, 1.0975600939592274)

==========

Group 5 (Fig. 42): Knockout, by Lil Wayne Featuring Nicki Minaj

Genre list for cluster center: ['hip hop', 'new orleans rap', 'pop rap', 'rap', 'trap']

Coordinates: (-2.2487897503870284, 1.3462054872013571)

==========

Group 6 (Fig. 43): Mr. Carter, by Lil Wayne Featuring Jay-Z

Genre list for cluster center: ['hip hop', 'new orleans rap', 'pop rap', 'rap', 'trap']

Coordinates: (-2.6316457013309913, 3.1248715960641937)

==========

Group 7 (Fig. 44): Behind Blue Eyes, by The Who

Genre list for cluster center: ['album rock', 'art rock', 'british invasion', 'classic rock', 'folk rock', 'hard rock', 'mellow gold', 'protopunk', 'rock', 'roots rock']

Coordinates: (0.31493723990760636, 0.20037562868408118)

==========

Group 8 (Fig. 45): Steam, by Peter Gabriel

Genre list for cluster center: ['album rock', 'art pop', 'art rock', 'dance rock', 'mellow gold', 'permanent wave', 'rock', 'soft rock', 'symphonic rock']

Coordinates: (-1.055106621243965, -0.2256033336750082)

==========

Group 9 (Fig. 46): Don't Forget About Me, by Dusty Springfield

Genre list for cluster center: ['adult standards', 'brill building pop', 'british invasion', 'bubblegum pop', 'classic uk pop', 'folk', 'folk rock', 'lounge', 'mellow gold', 'motown', 'rock', 'soul', 'vocal jazz']

Coordinates: (1.1392915648270854, -0.5105270395169049)

6.4.4 k = 1, 2010s

==========

Group 0: Sex Room, by Ludacris Featuring Trey Songz

Genre list for cluster center: ['atl hip hop', 'dance pop', 'dirty south rap', 'hip hop', 'pop rap', 'rap', 'southern hip hop', 'trap']

Coordinates: (1.3055400382166194e-16, -1.3971240924365399e-16)

6.4.5 k = 3, 2010s

==========

Group 0 (Fig. 47): 9 Piece, by Rick Ross Featuring Lil Wayne Or T.I.

Genre list for cluster center: ['dirty south rap', 'gangster rap', 'hip hop', 'pop rap', 'rap', 'southern hip hop', 'trap']

Coordinates: (0.14545397134417323, 1.5997790125192437)

==========

Group 1 (Fig. 48): Everybody Hates Me, by The Chainsmokers

Genre list for cluster center: ['electropop', 'pop', 'tropical house']

Coordinates: (-1.0013605538908235, -0.7865593792614018)

==========

Group 2 (Fig. 49): Into The Unknown, by Idina Menzel and AURORA

Genre list for cluster center: ['hollywood', 'show tunes']

Coordinates: (2.402917608246099, -0.9801049522379964)

6.4.6 k = 10, 2010s

==========

Group 0 (Fig. 50): Get Low, by Zedd Liam Payne

Genre list for cluster center: ['complextro', 'dance pop', 'edm', 'electro house', 'german techno', 'pop', 'post-teen pop', 'tropical house']

Coordinates: (-1.8251786781637704, -0.9723416725717746)

==========

Group 1 (Fig. 51): I Almost Do, by Taylor Swift

Genre list for cluster center: ['pop', 'post-teen pop']

Coordinates: (0.8911310274466424, -1.1693142051395569)

==========

Group 2 (Fig. 52): Its Every Night Sis, by RiceGum Featuring Alissa Violet

Genre list for cluster center: ['social media pop']

Coordinates: (-0.841387663185129, 1.8827461959034557)

==========

Group 3 (Fig. 53): Holy Key, by DJ Khaled Featuring Big Sean, Kendrick Lamar and Betty Wright

Genre list for cluster center: ['dance pop', 'hip hop', 'miami hip hop', 'pop', 'pop rap', 'rap', 'southern hip hop', 'trap']

Coordinates: (-1.2946607769911924, 0.33362154428167384)

==========

Group 4 (Fig. 54): Redneck Crazy, by Tyler Farr

Genre list for cluster center: ['contemporary country', 'country', 'country pop', 'country road', 'modern country rock', 'redneck']

Coordinates: (-0.5659302724051143, -1.1162711135981391)

==========

Group 5 (Fig. 55): Nothing Like Us, by Justin Bieber

Genre list for cluster center: ['canadian pop', 'dance pop', 'pop', 'post-teen pop']

Coordinates: (4.9021844918414725, -1.1907152924887112)

==========

Group 6 (Fig. 56): Lose You, by Drake

Genre list for cluster center: ['canadian hip hop', 'canadian pop', 'hip hop', 'pop rap', 'rap', 'toronto rap']

Coordinates: (2.327254274015786, 1.17908172684473)

==========

Group 7 (Fig. 57): Imma Be, by The Black Eyed Peas

Genre list for cluster center: ['dance pop', 'pop', 'pop rap']

Coordinates: (2.767543158647383, -1.3989570004440441)

==========

Group 8 (Fig. 58): Famous, by Lil Wayne Featuring Reginae Carter

Genre list for cluster center: ['hip hop', 'new orleans rap', 'pop rap', 'rap', 'trap']

Coordinates: (0.38549387292094106, 0.6908536023932265)

==========

Group 9 (Fig. 59): Fr Fr, by Wiz Khalifa Featuring Lil Skies

Genre list for cluster center: ['hip hop', 'pittsburgh rap', 'pop rap', 'rap', 'southern hip hop', 'trap']

Coordinates: (0.7172228880141721, 2.4388794029164913)

6.5 Clusters

6.5.1 1960 - 2021 (full dataset)

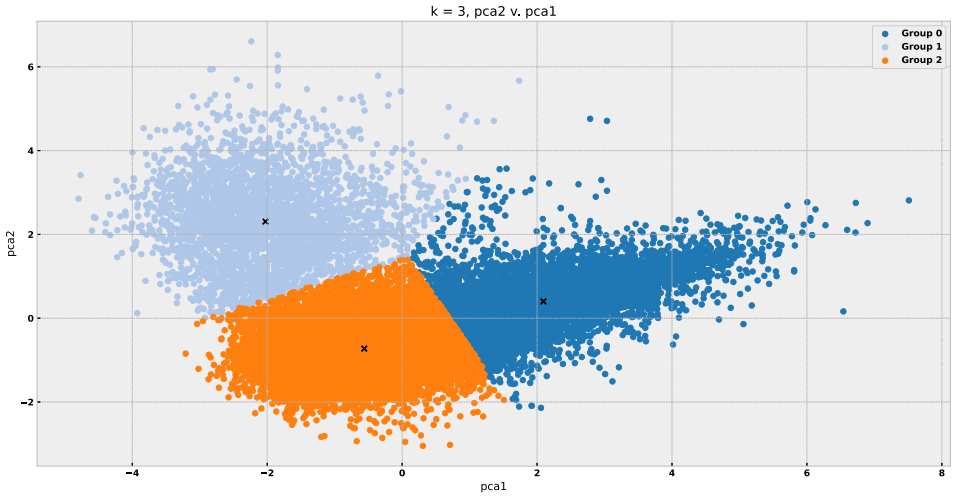

Figure 30: Clustering scatter plot with $k = 3$, full dataset

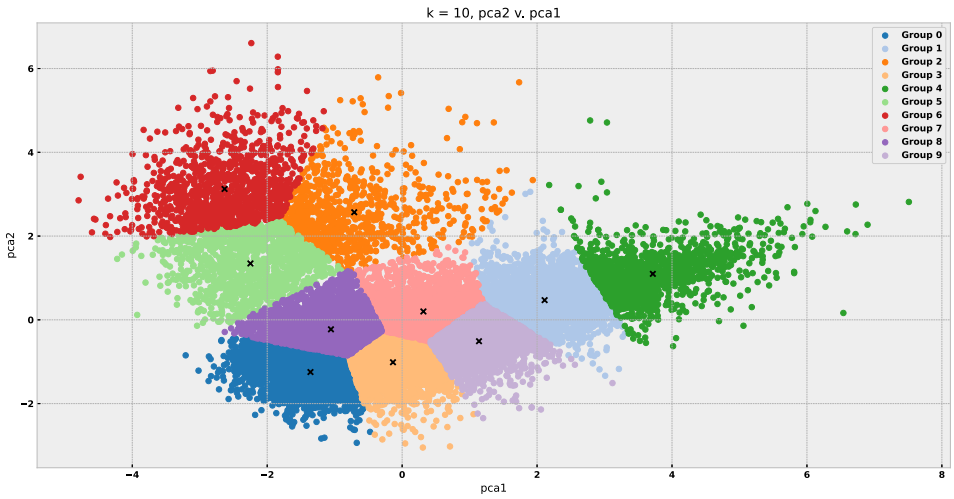

Figure 31: Clustering scatter plot with $k = 10$, full dataset

6.5.2 2010 - 2021 (the 2010s)

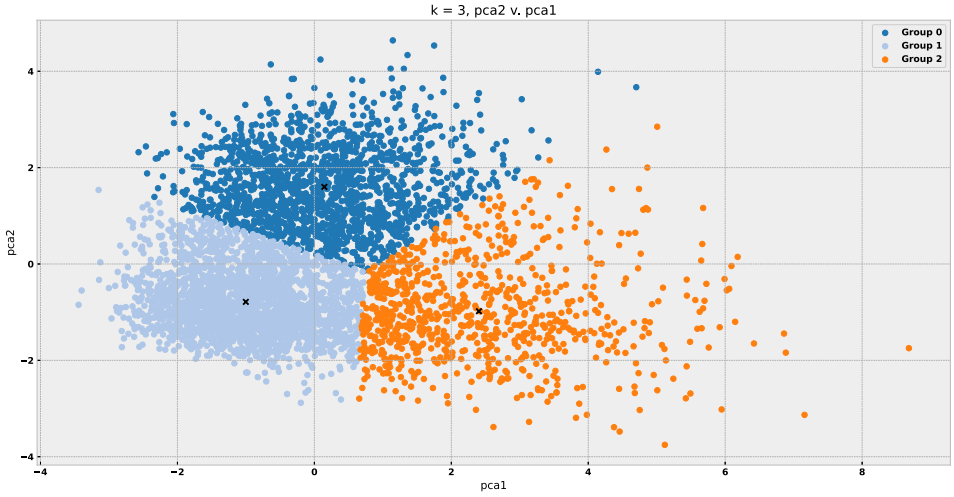

Figure 32: Clustering scatter plot with $k = 3$, 2010s

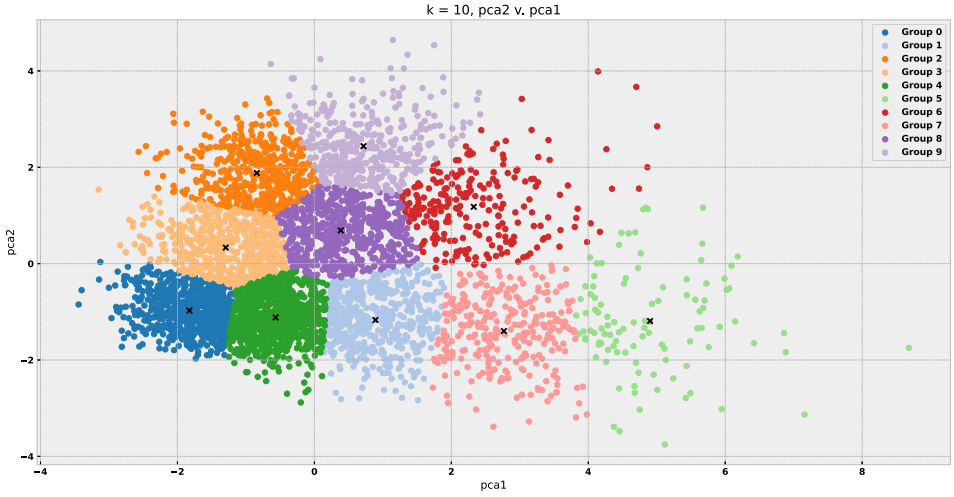

Figure 33: Clustering scatter plot with $k = 10$, 2010s

6.6 Genre Groups, 1960 - 2021 (full dataset)

6.6.1 k = 3

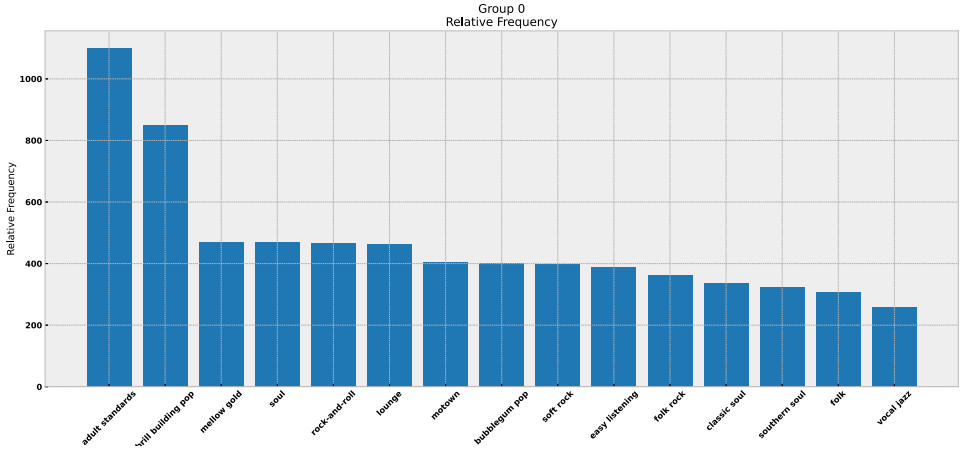

Figure 34: Genre relative frequency, k = 3, Group 0, full dataset

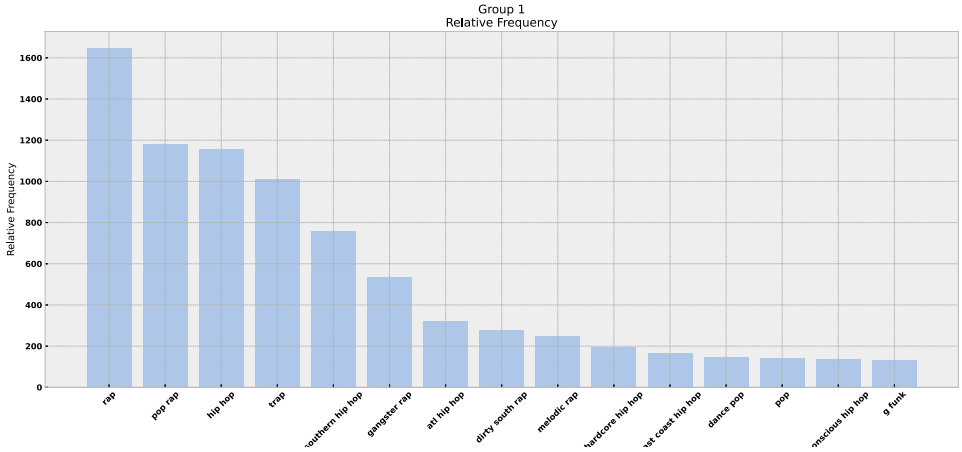

Figure 35: Genre relative frequency, k = 3, Group 1, full dataset

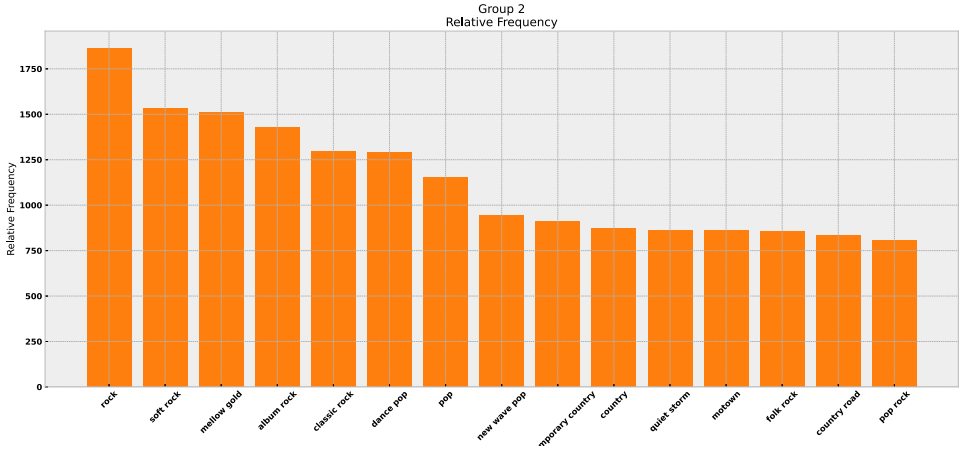

Figure 36: Genre relative frequency, k = 3, Group 2, full dataset

6.6.2 k = 10

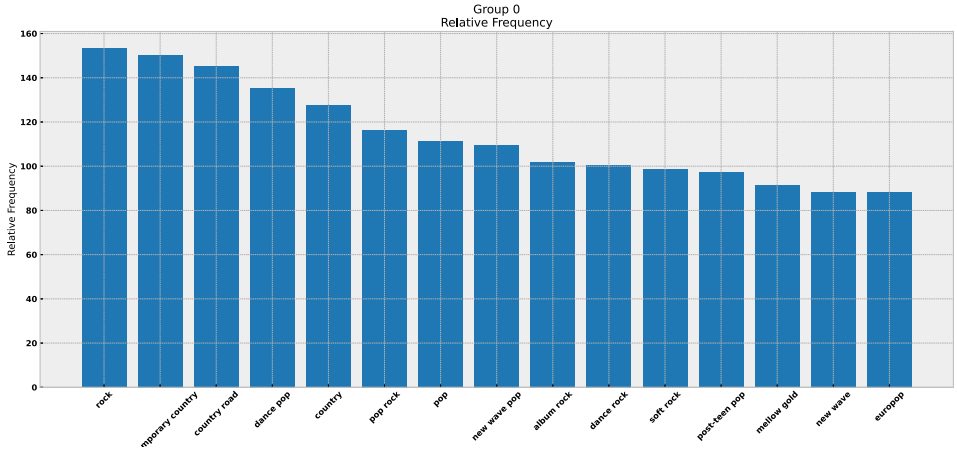

Figure 37: Genre relative frequency, k = 10, Group 0, full dataset

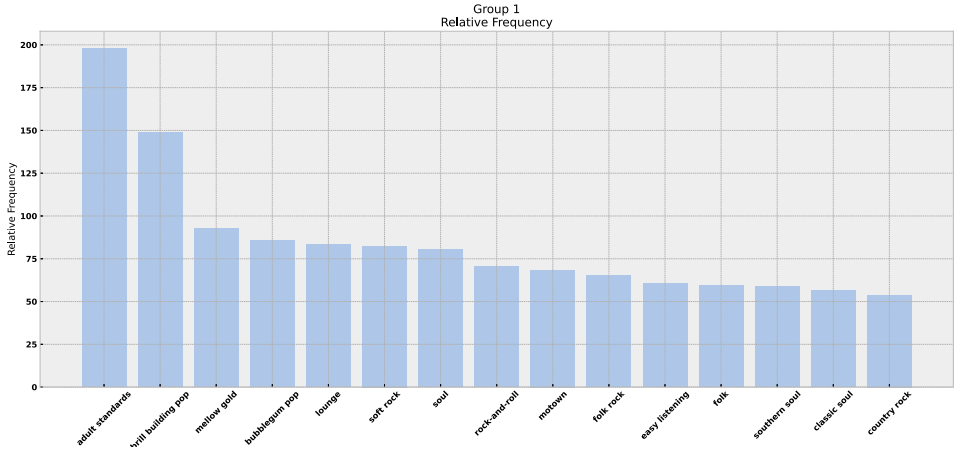

Figure 38: Genre relative frequency, k = 10, Group 1, full dataset

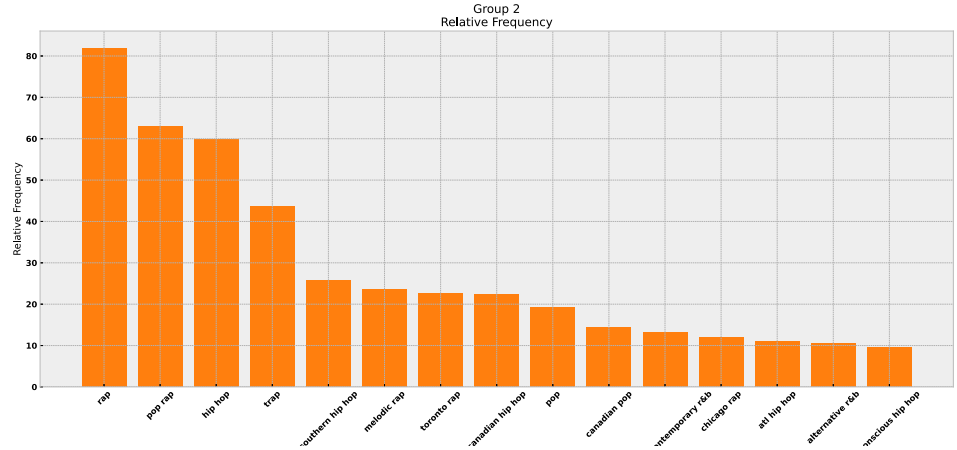

Figure 39: Genre relative frequency, k = 10, Group 2, full dataset

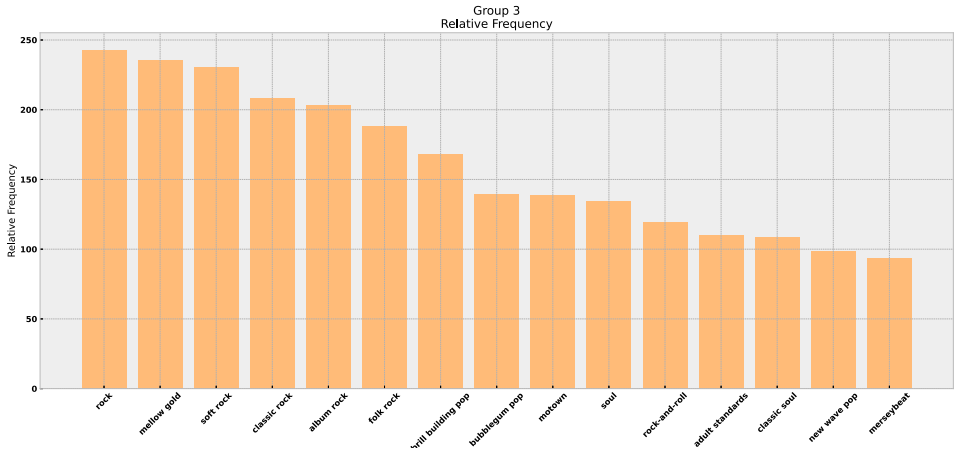

Figure 40: Genre relative frequency, k = 10, Group 3, full dataset

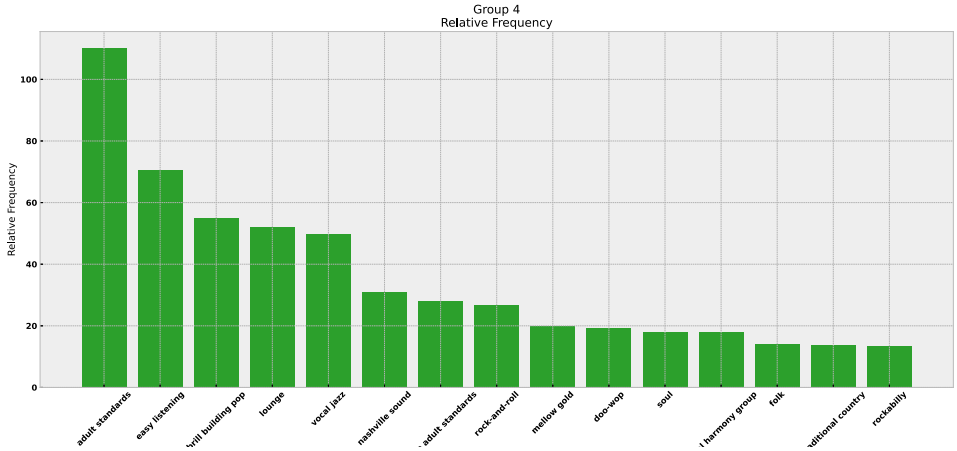

Figure 41: Genre relative frequency, k = 10, Group 4, full dataset

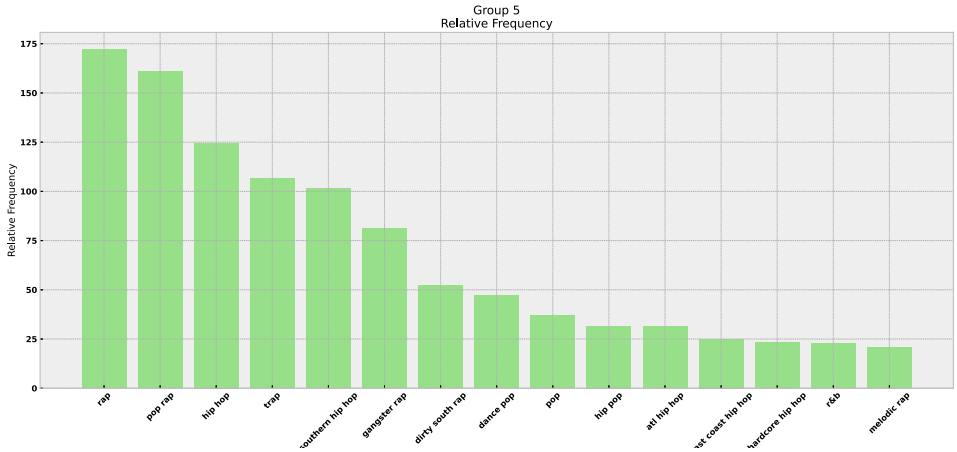

Figure 42: Genre relative frequency, k = 10, Group 5, full dataset

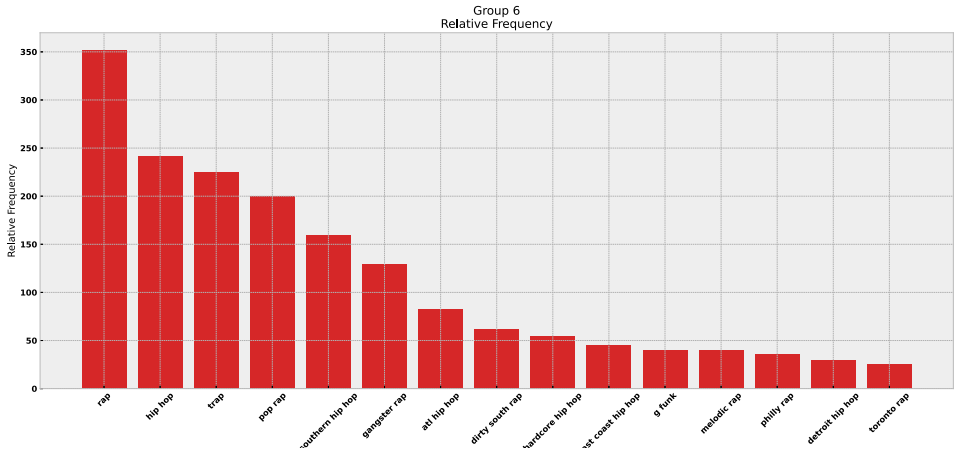

Figure 43: Genre relative frequency, k = 10, Group 6, full dataset

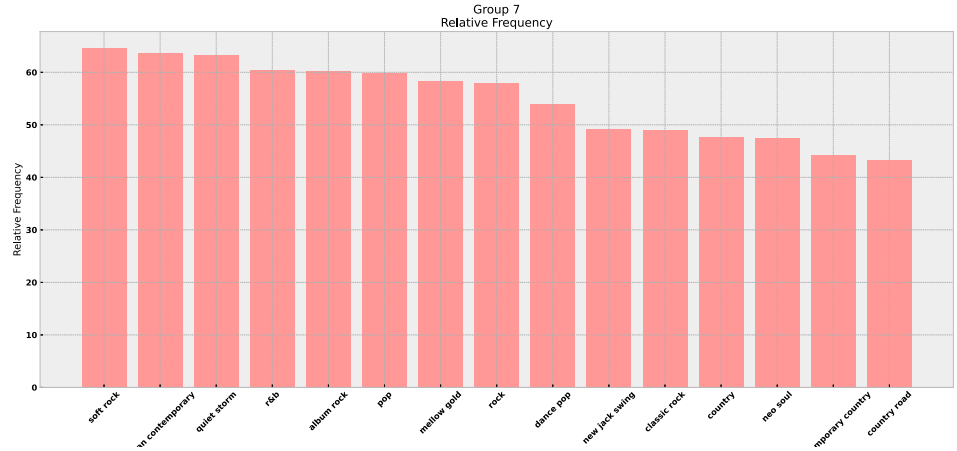

Figure 44: Genre relative frequency, k = 10, Group 7, full dataset

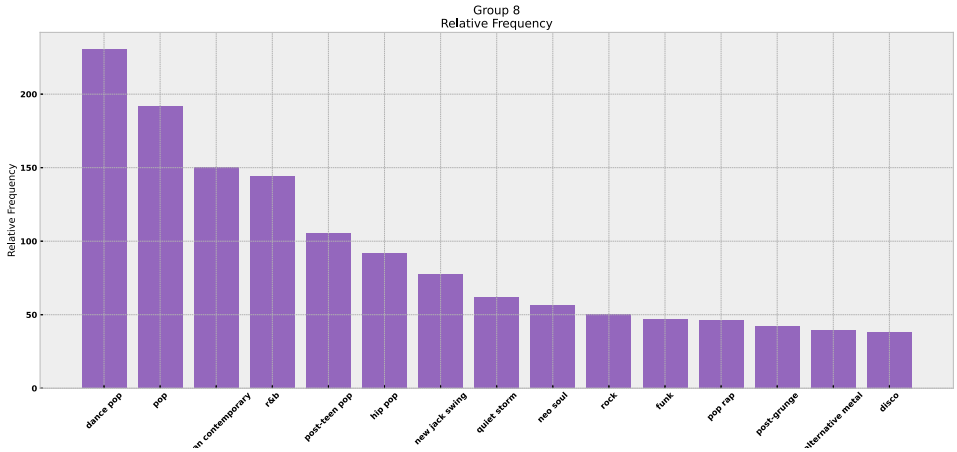

Figure 45: Genre relative frequency, k = 10, Group 8, full dataset

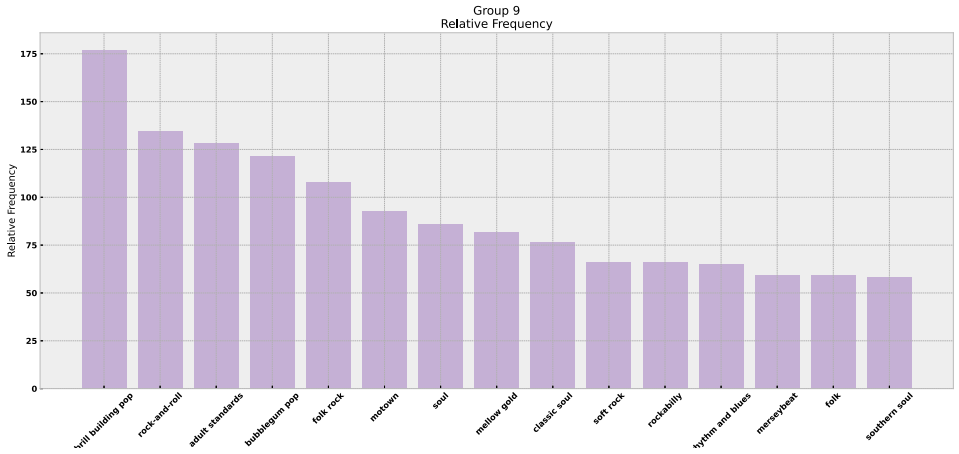

Figure 46: Genre relative frequency, k = 10, Group 9, full dataset

6.7 Genre Groups, 2010 - 2021 (the 2010s)

6.7.1 k = 3

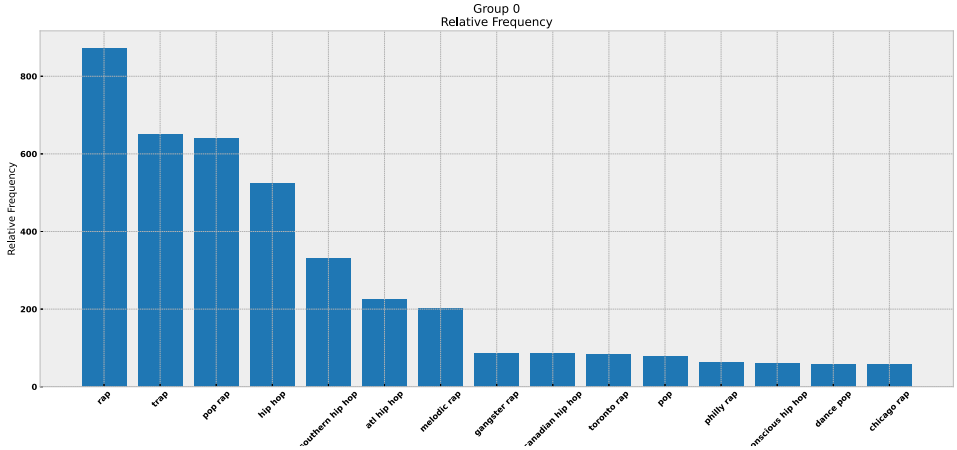

Figure 47: Genre relative frequency, k = 3, Group 0, 2010s

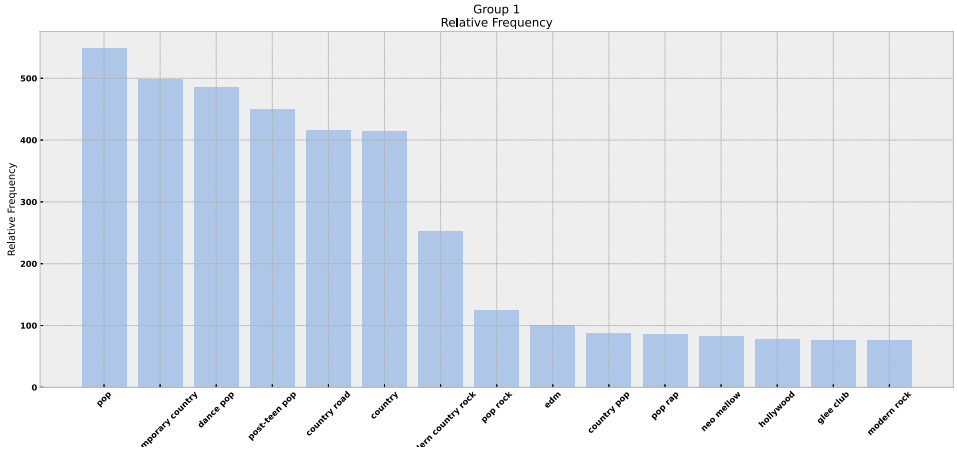

Figure 48: Genre relative frequency, k = 3, Group 1, 2010s

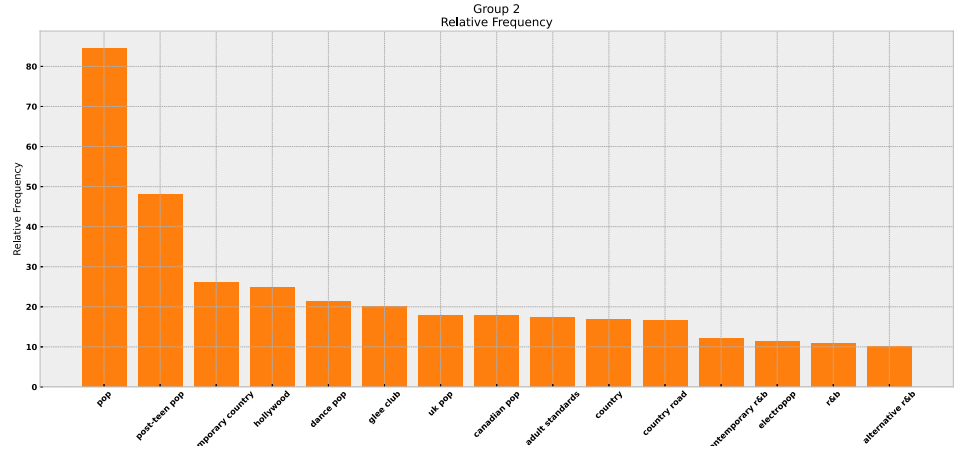

Figure 49: Genre relative frequency, k = 3, Group 2, 2010s

6.7.2 k = 10

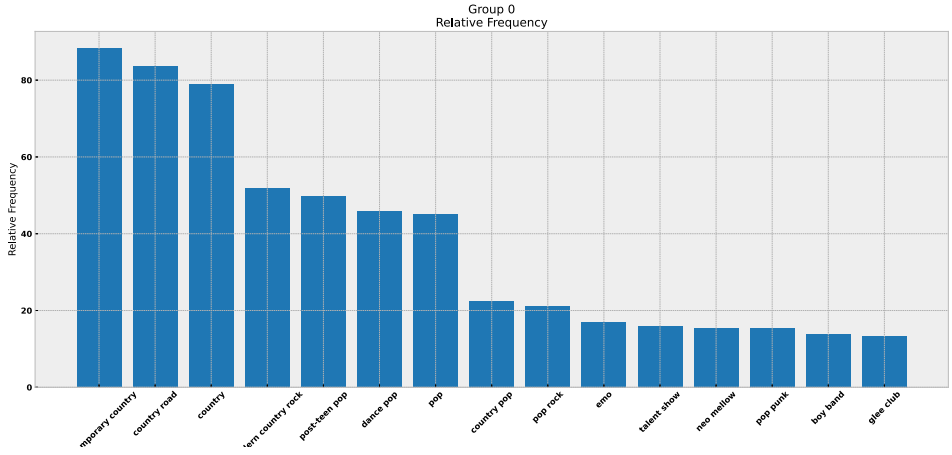

Figure 50: Genre relative frequency, k = 10, Group 0, 2010s

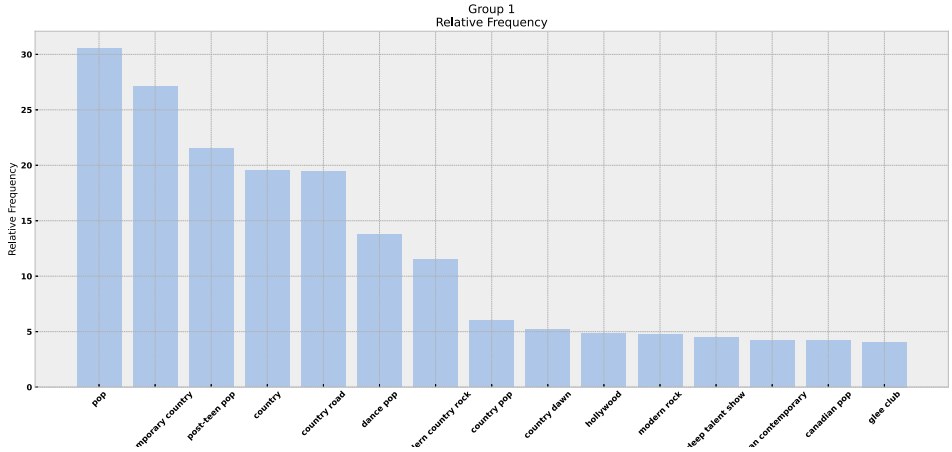

Figure 51: Genre relative frequency, k = 10, Group 1, 2010s

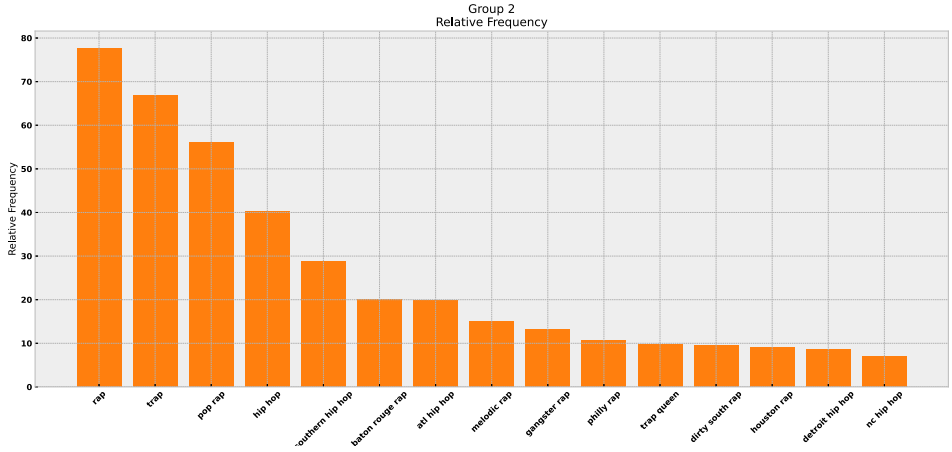

Figure 52: Genre relative frequency, k = 10, Group 2, 2010s

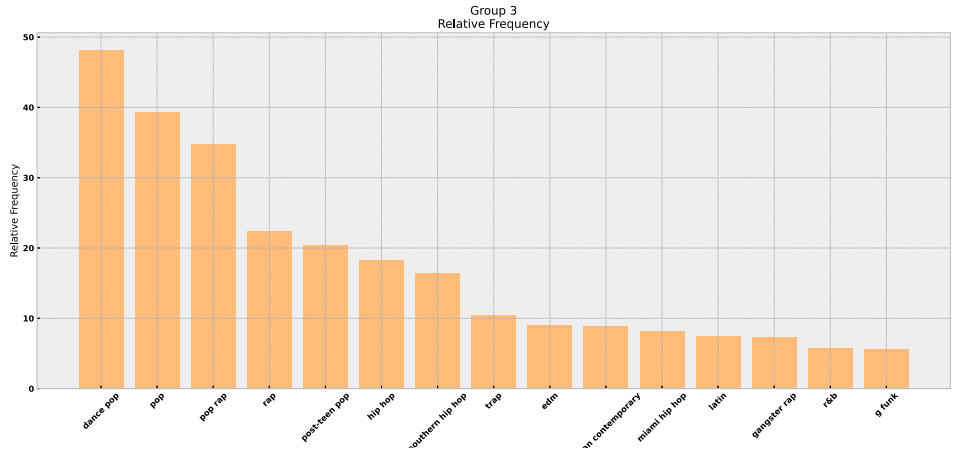

Figure 53: Genre relative frequency, k = 10, Group 3, 2010s

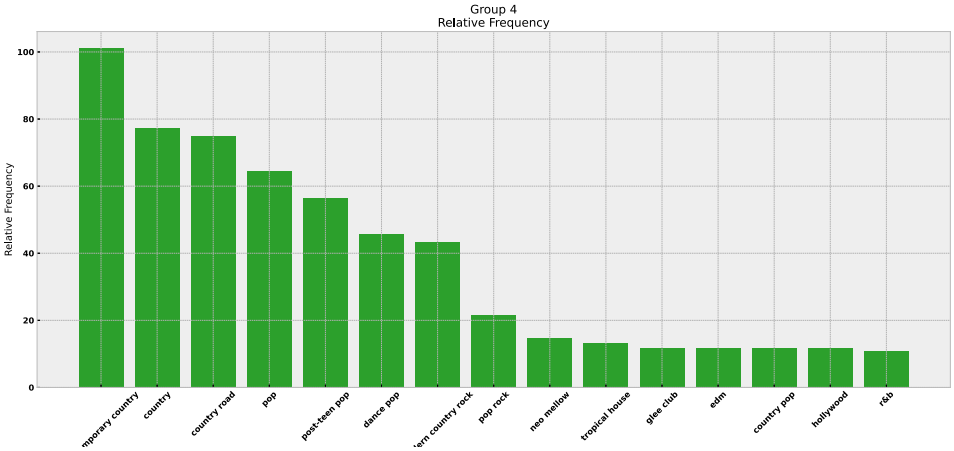

Figure 54: Genre relative frequency, k = 10, Group 4, 2010s

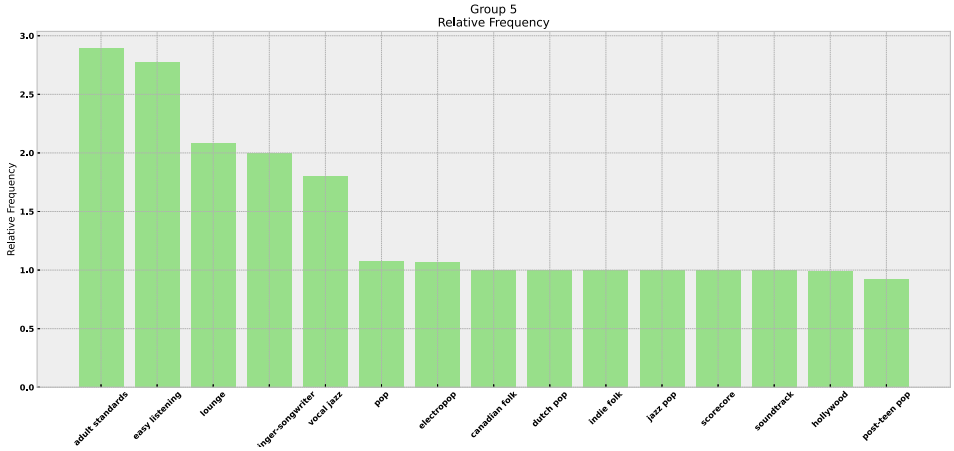

Figure 55: Genre relative frequency, k = 10, Group 5, 2010s

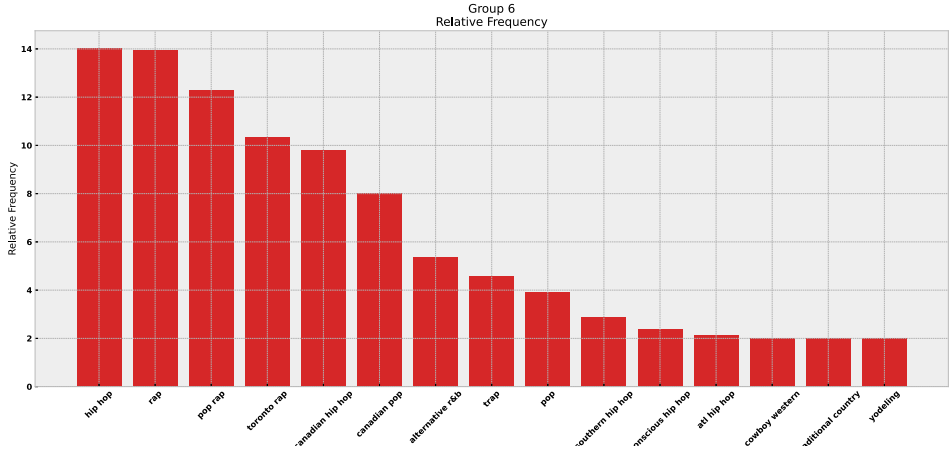

Figure 56: Genre relative frequency, k = 10, Group 6, 2010s

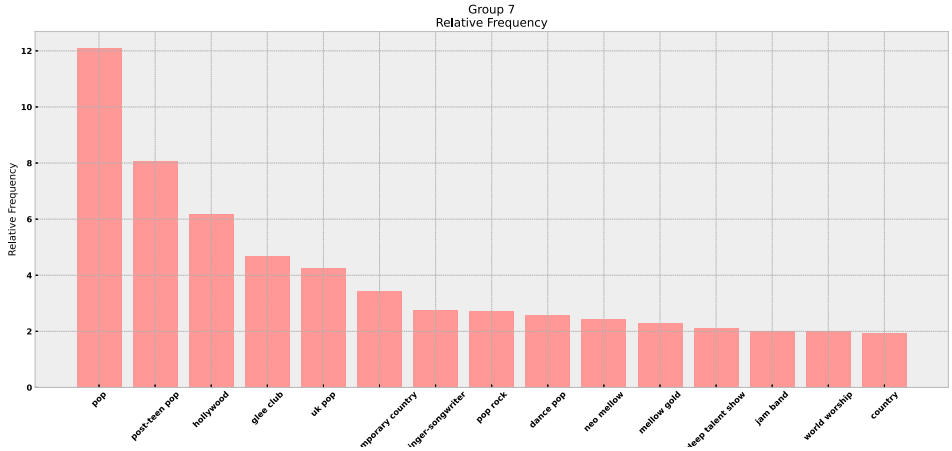

Figure 57: Genre relative frequency, k = 10, Group 7, 2010s

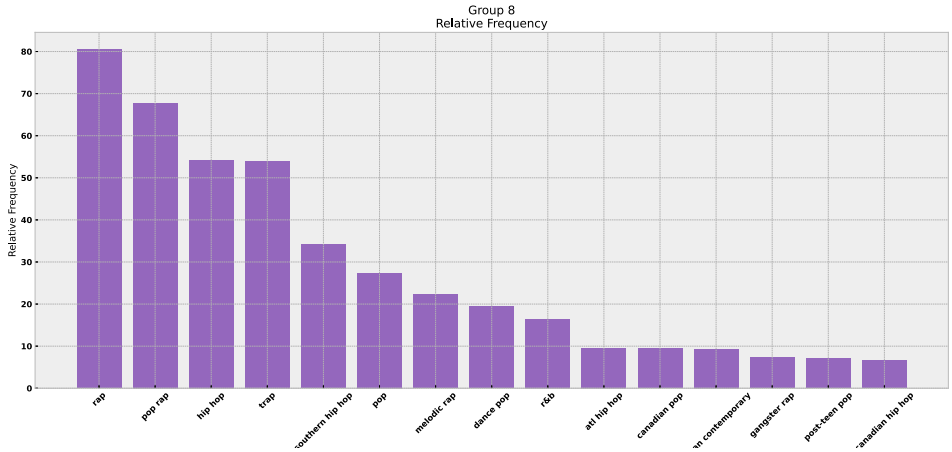

Figure 58: Genre relative frequency, k = 10, Group 8, 2010s

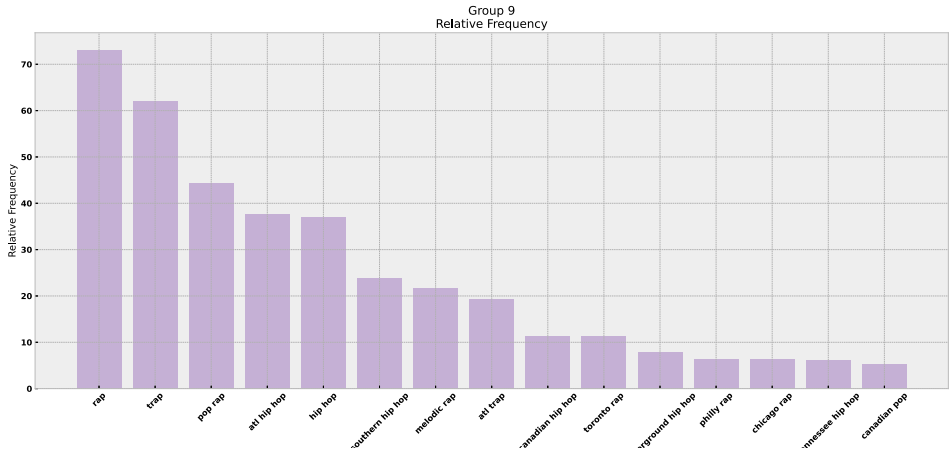

Figure 59: Genre relative frequency, k = 10, Group 9, 2010s

References

[1] @kcmillersean. *Billboard Hot weekly charts*. URL: https://data.world/kcmillersean/billboard-hot-100-1958-2017. (accessed 10.09.2022).

[2] Thierry Bertin-Mahieux et al. "The Million Song Dataset". In: *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)* (Jan. 2011), pp. 591–596.

[3] EchoNest. *Million Song Dataset*. URL: http://millionsongdataset.com/. (accessed 10.09.2022).

[4] Spotify Engineering. *Get Track Documentation*. URL: https://developer.spotify.com/documentation/web-api/reference/#/operations/get-track. (accessed 10.09.2022).

[5] Spotify Engineering. *Get Track's Audio Features Documentation*. URL: https://developer.spotify.com/documentation/web-api/reference/#/operations/get-audio-features. (accessed 10.09.2022).

[6] Spotify Engineering. *Spotify Web API*. URL: https://developer.spotify.com/documentation/web-api/reference/#/. (accessed 10.09.2022).

[7] James Pham, Edric Kyauk, and Edwin Park. "Predicting Song Popularity". In: *Stanford Project Repository* (2015). DOI: http://cs229.stanford.edu/proj2015/140_report.pdf.

[8] Peter Skiden. *API Improvements and U*. URL: https://developer.spotify.com/community/news/2016/03/29/api-improvements-update/. (accessed 10.09.2022).