Semester Project Update 1: Spotify Audio Feature Time Series Analysis

Isaac Fry

October 24, 2022

1 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 my proposal, 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 [2]. This was possible due to a separate (now defunct) open source project that translated between different URIs for music data. The specific origin of each feature is listed in the data dictionary.

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

2 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 the hassle of learning the REST infrastructure and Spotify's APIs seems like an unnecessary step right now.

In order to take a time series approach and maintain a generalized approach, I decided to only keep the instance of a track 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.1, the distribution of tracks over time is fairly evenly distributed. This means a time series analysis would be in the valid realm of research.

From the initial data exploration, there were some interesting findings:

- Disappointingly, Popularity (Fig. 13) increases in a nearly $1:1$ relationship with time. This is expected, because the current popularity of older tracks on Spotify is lower than newer tracks (i.e. Justin Bieber is now more popular than Michael Jackson). This means that popularity **cannot** isn't necessarily an insightful feature.
- The Duration (Fig. 2) of tracks steadily increased from the 1960s to the 1990s, but then began decreasing again. There's probably some influence from the attention economy at play, especially after the introduction of music-backed social media beginning in the late 2010s.
- Songs have generally gotten louder over time (Fig. 6). After peaking in the mid 2000s, it stays consistent with current industry standard of -8 to -6 db.

- Speechiness (Fig. 8) increases after hip-hop became midstream in the late 1980s to early 1990s. There was a dramatic uptick in speechiness in the mid 2010s, presumably from the prominent rise of iconic hip-hop artists like Kendrick Lamar, Drake, Kanye West, and others.
- Perhaps most concerning for society, Valence (the measure of positivity) decreases consistently over time (Fig. 11).

At this point in the process, it seems unnecessary to do any dimensionality reduction as each of these features vary independently from one another.

As things stand, trying to predict a track's week of peak popularity based on audio features would be an interesting challenge. However, there are classification questions that can be paired with genres, though that would require extra data manipulation. Finally, doing feature prediction based on other features would be an question to fall back on (i.e. can a track's tempo be used to predict its key?).

3 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.

References

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

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

4 Appendix

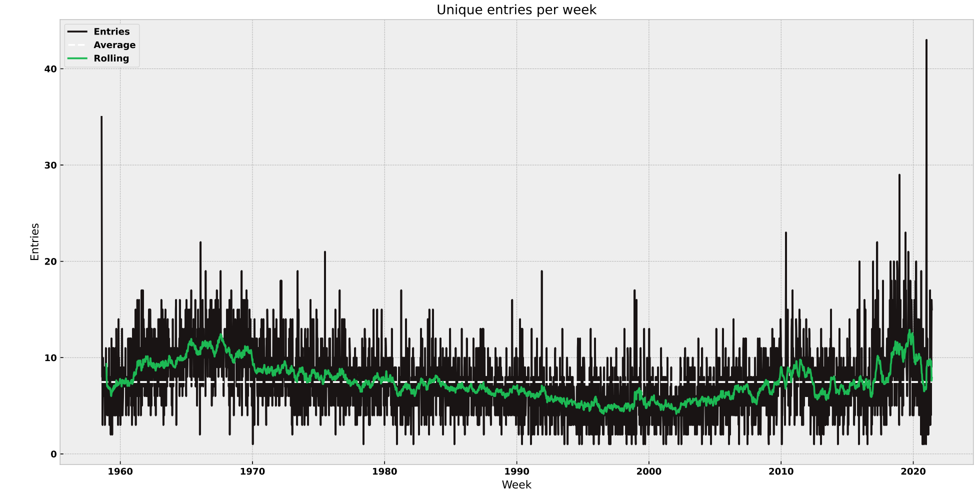

Figure 1: Analysis of Duration (ms) averages of the tracks that peaked on the Hot 100 Billboard Chart during a given week. See the data dictionary section for a more detailed description of this feature.

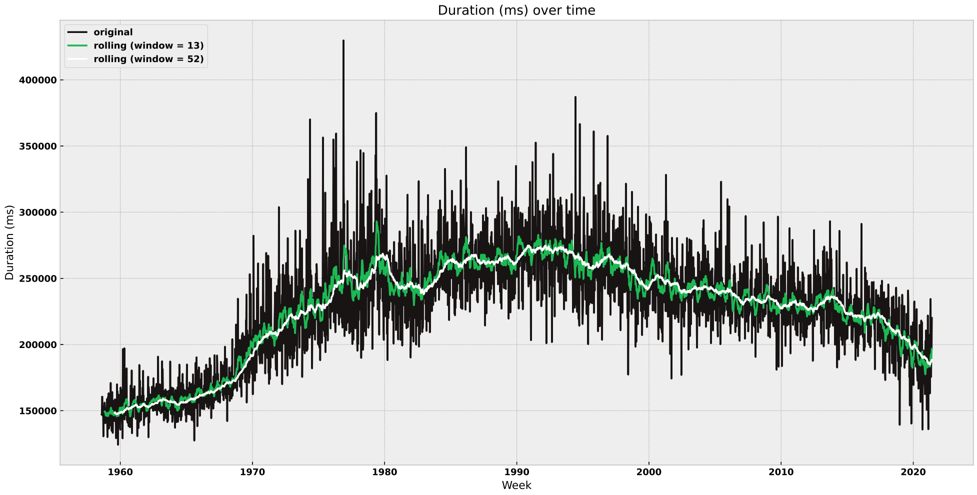

Figure 2: Analysis of Duration (ms) averages of the tracks that peaked on the Hot 100 Billboard Chart during a given week. See the data dictionary section for a more detailed description of this feature.

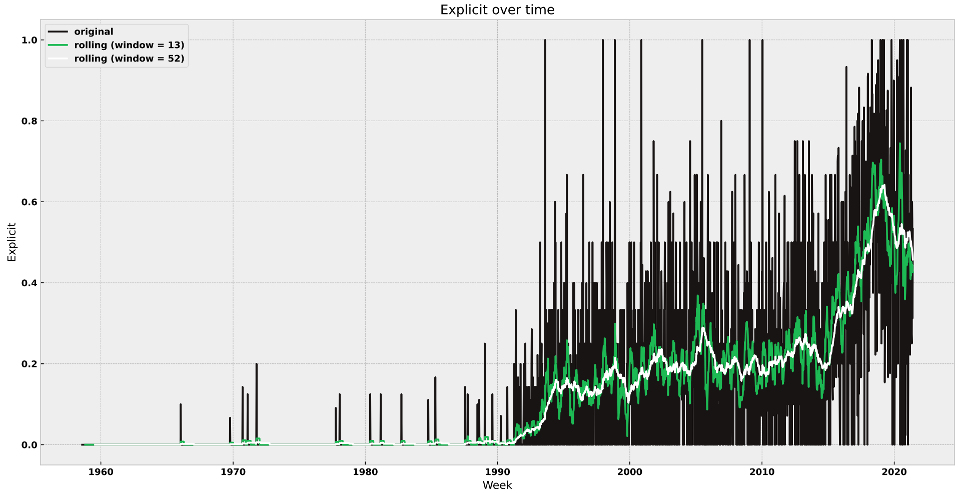

Figure 3: Analysis of Explicit averages of the tracks that peaked on the Hot 100 Billboard Chart during a given week. See the data dictionary section for a more detailed description of this feature.

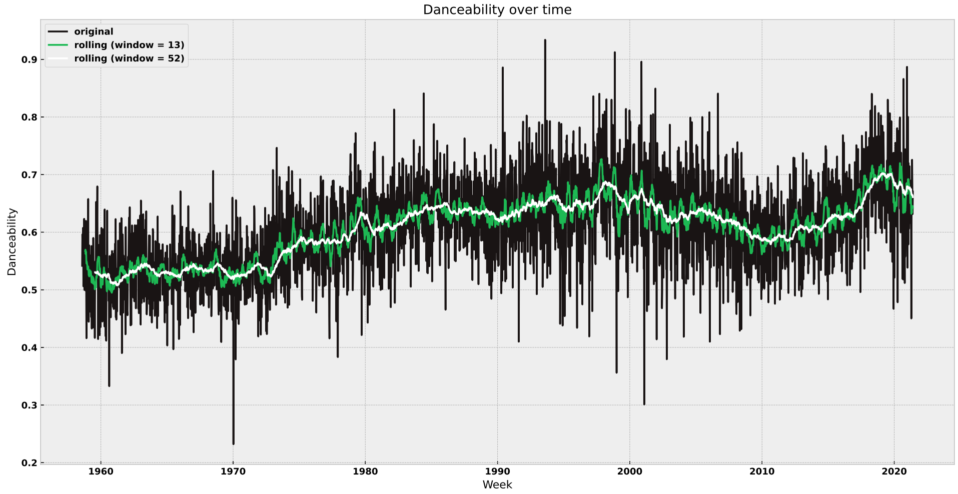

Figure 4: Analysis of Danceability averages of the tracks that peaked on the Hot 100 Billboard Chart during a given week. See the data dictionary section for a more detailed description of this feature.

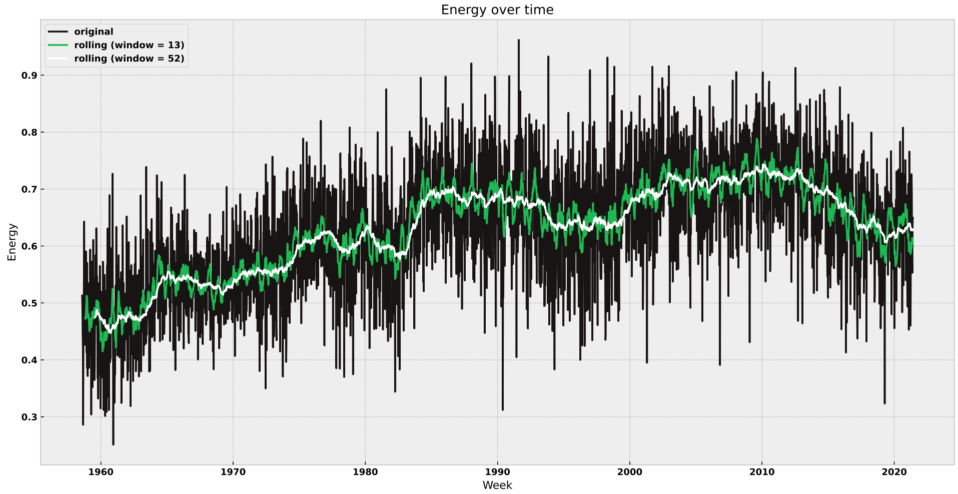

Figure 5: Analysis of Energy averages of the tracks that peaked on the Hot 100 Billboard Chart during a given week. See the data dictionary section for a more detailed description of this feature.

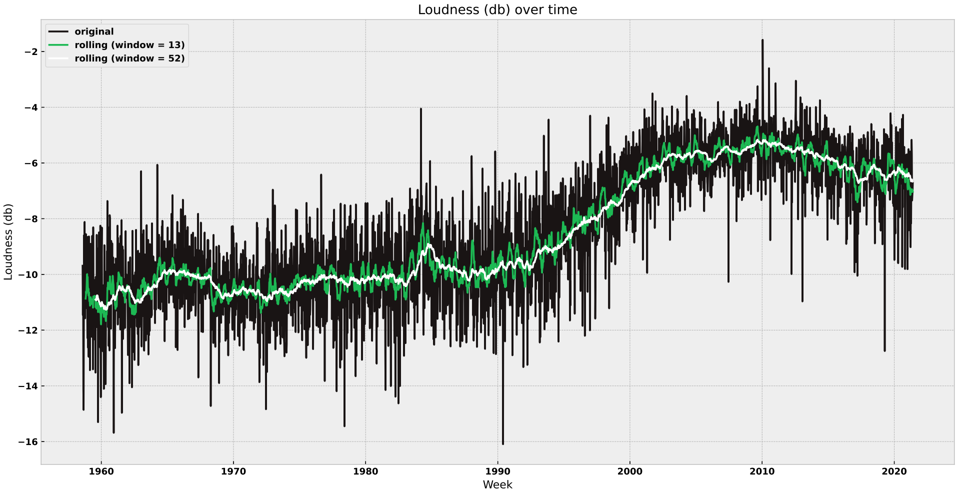

Figure 6: Analysis of Loudness (db) averages of the tracks that peaked on the Hot 100 Billboard Chart during a given week. See the data dictionary section for a more detailed description of this feature.

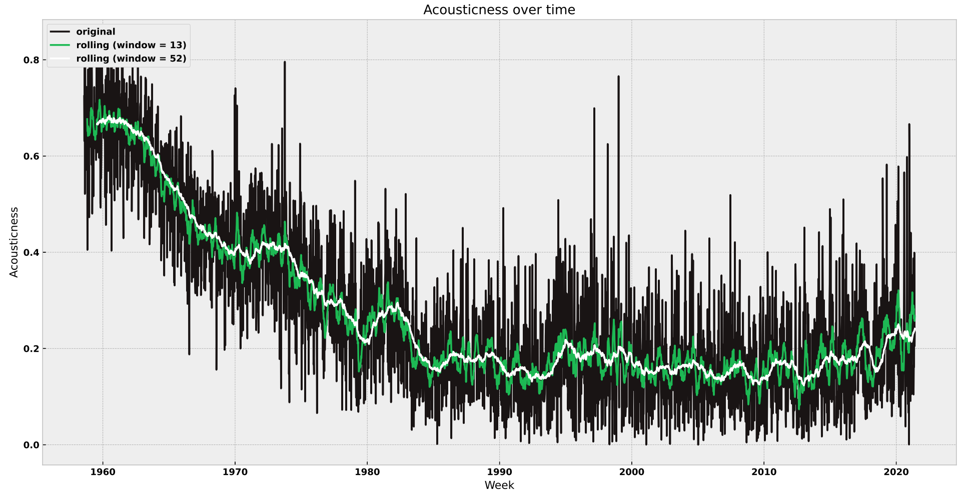

Figure 7: Analysis of Acousticness averages of the tracks that peaked on the Hot 100 Billboard Chart during a given week. See the data dictionary section for a more detailed description of this feature.

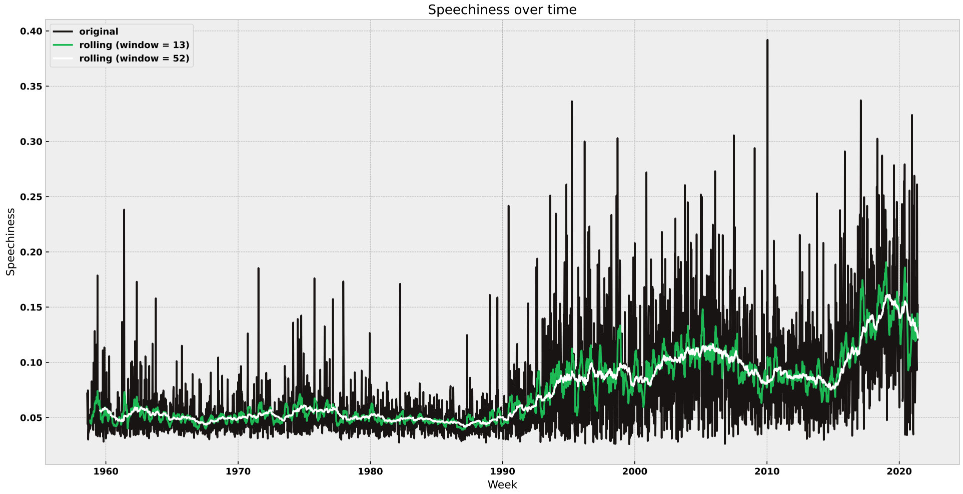

Figure 8: Analysis of Speechiness averages of the tracks that peaked on the Hot 100 Billboard Chart during a given week. See the data dictionary section for a more detailed description of this feature.

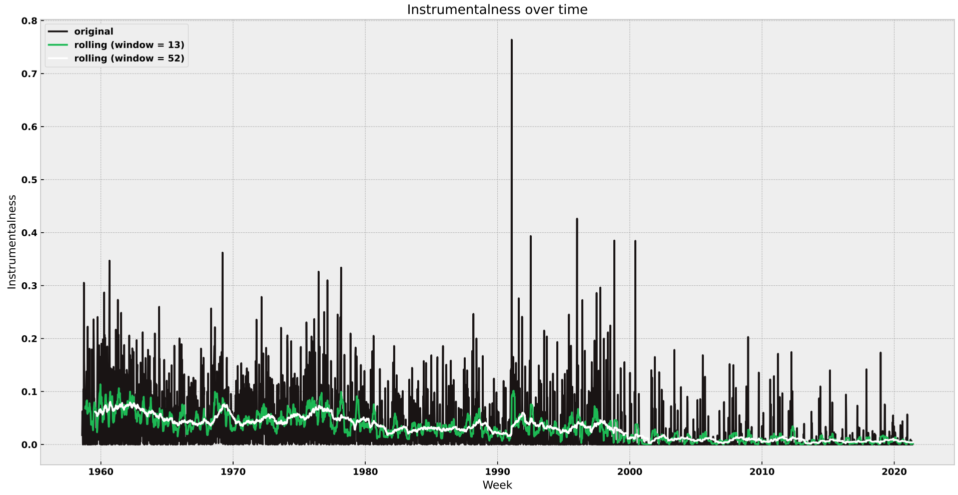

Figure 9: Analysis of Instrumentalness averages of the tracks that peaked on the Hot 100 Billboard Chart during a given week. See the data dictionary section for a more detailed description of this feature.

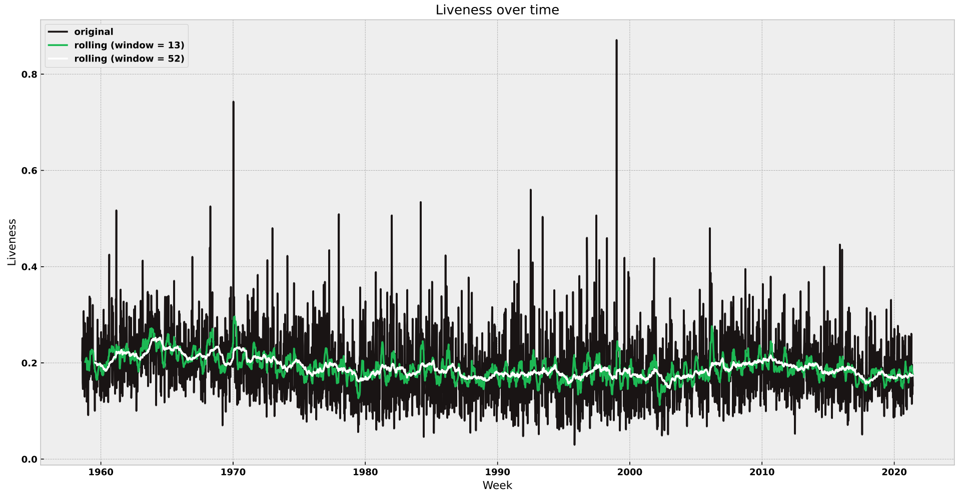

Figure 10: Analysis of Liveness averages of the tracks that peaked on the Hot 100 Billboard Chart during a given week. See the data dictionary section for a more detailed description of this feature.

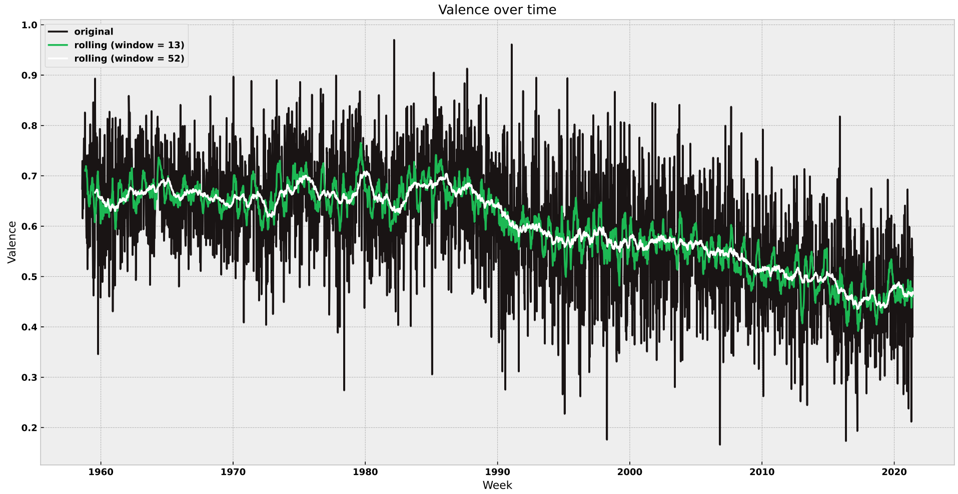

Figure 11: Analysis of Valence averages of the tracks that peaked on the Hot 100 Billboard Chart during a given week. See the data dictionary section for a more detailed description of this feature.

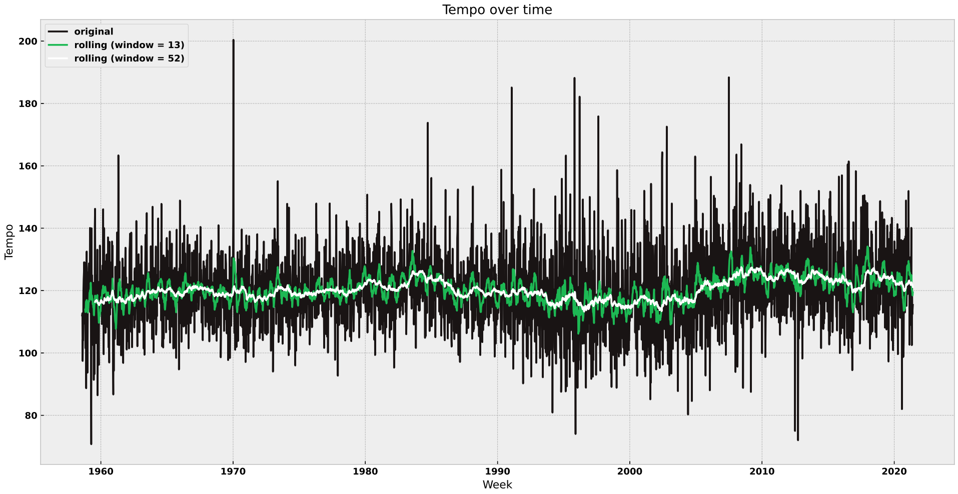

Figure 12: Analysis of Tempo averages of the tracks that peaked on the Hot 100 Billboard Chart during a given week. See the data dictionary section for a more detailed description of this feature.

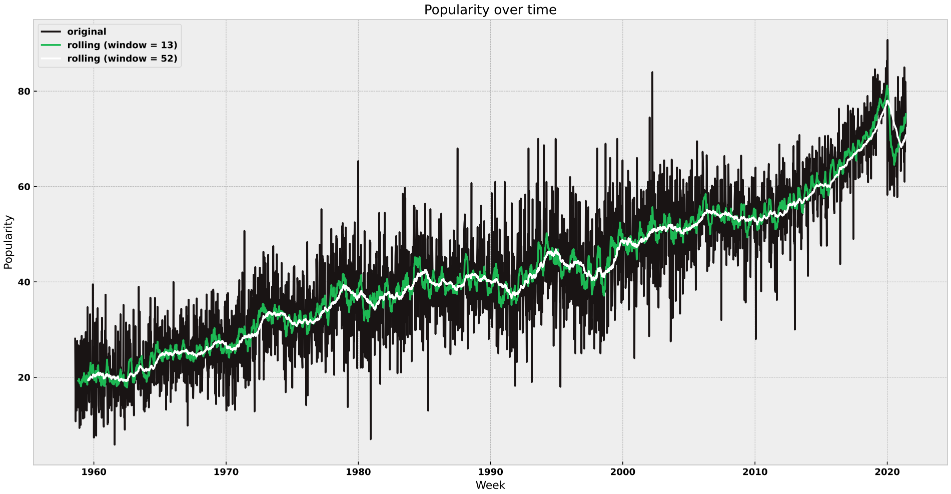

Figure 13: Analysis of Popularity averages of the tracks that peaked on the Hot 100 Billboard Chart during a given week. See the data dictionary section for a more detailed description of this feature.

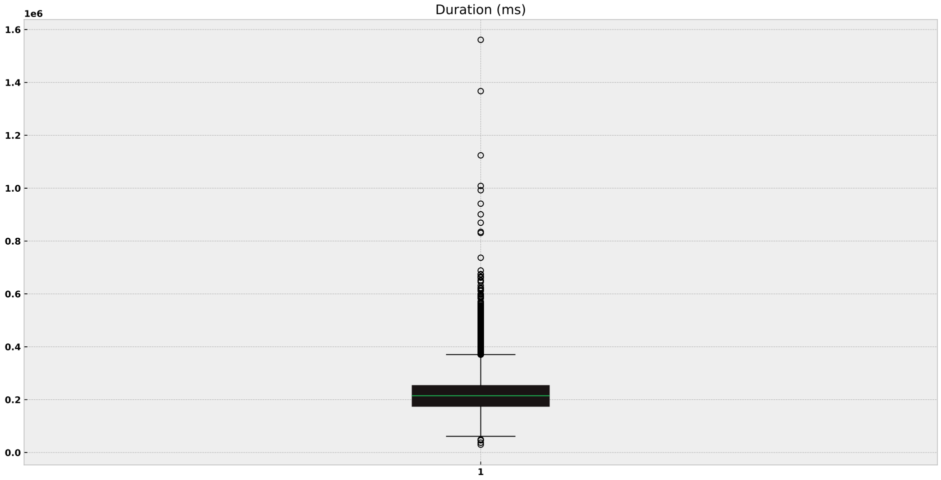

Figure 14: Analysis of Duration (ms) distributions of the tracks in the Billboard Hot 100 charts. See the data dictionary section for a more detailed description of this feature.

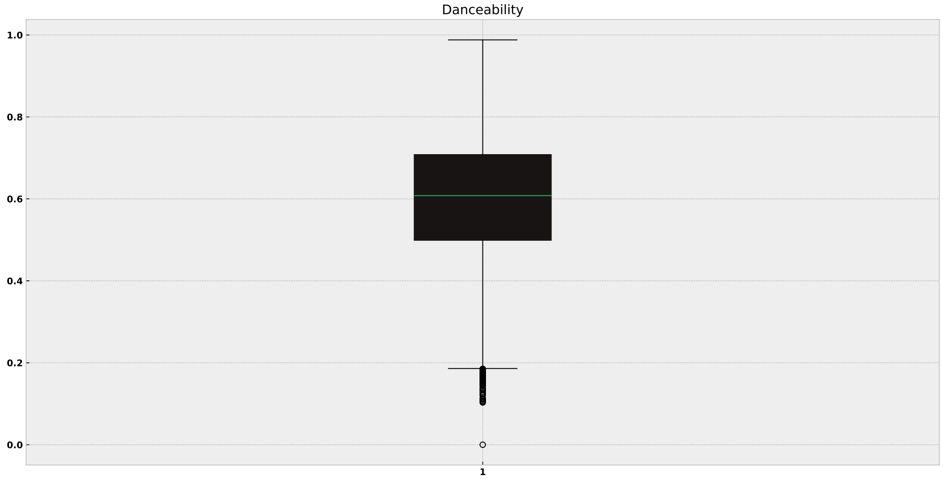

Figure 15: Analysis of Danceability distributions of the tracks in the Billboard Hot 100 charts. See the data dictionary section for a more detailed description of this feature.

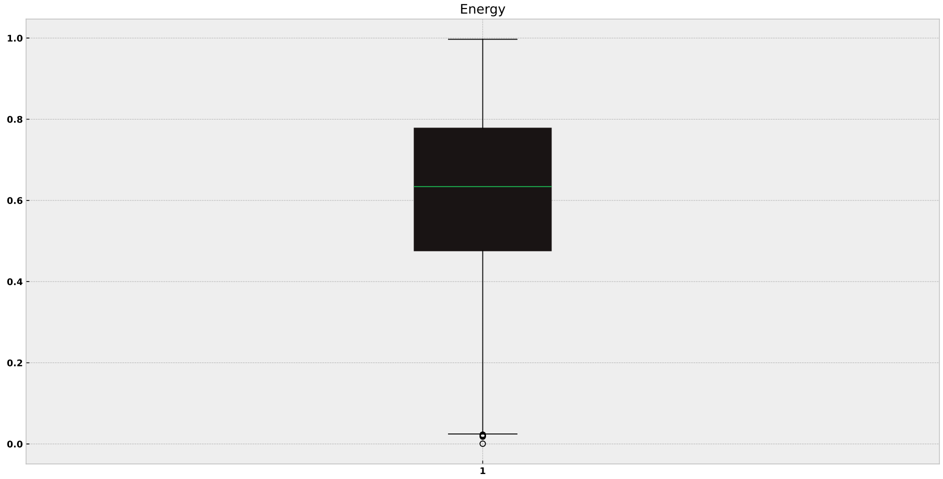

Figure 16: Analysis of Energy distributions of the tracks in the Billboard Hot 100 charts. See the data dictionary section for a more detailed description of this feature.

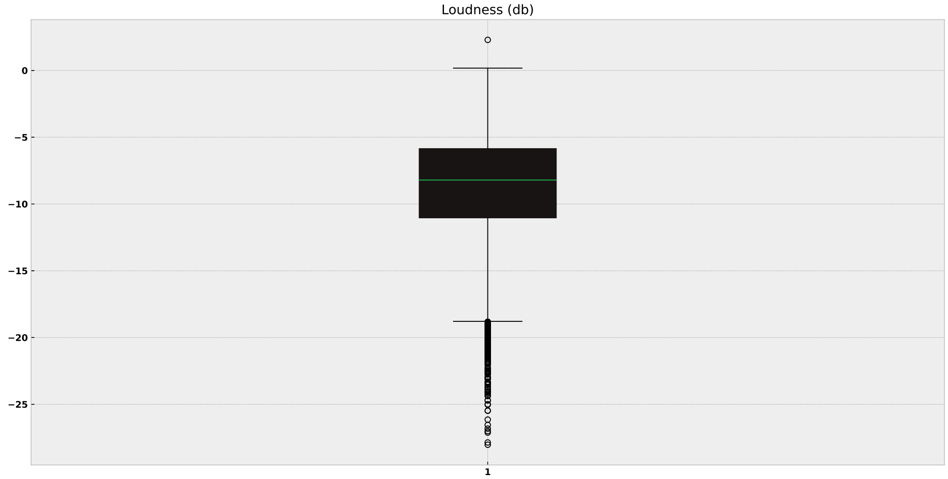

Figure 17: Analysis of Loudness (db) distributions of the tracks in the Billboard Hot 100 charts. See the data dictionary section for a more detailed description of this feature.

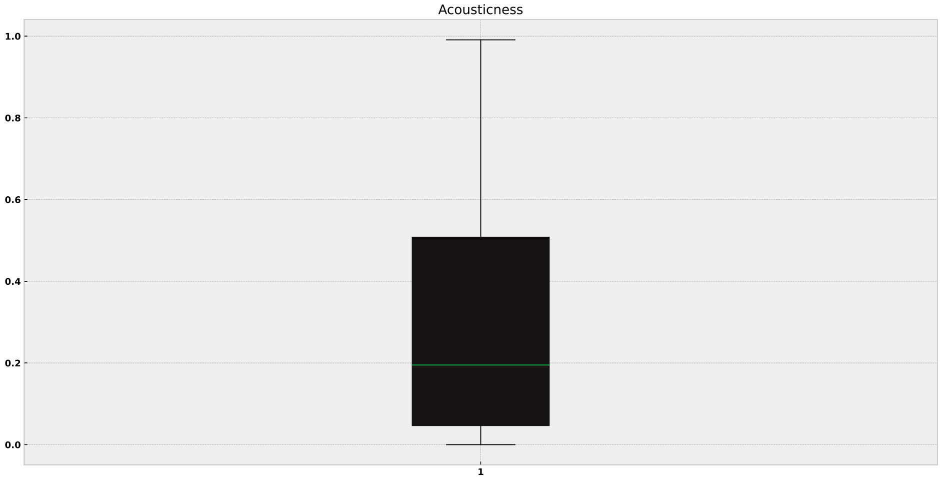

Figure 18: Analysis of Acousticness distributions of the tracks in the Billboard Hot 100 charts. See the data dictionary section for a more detailed description of this feature.

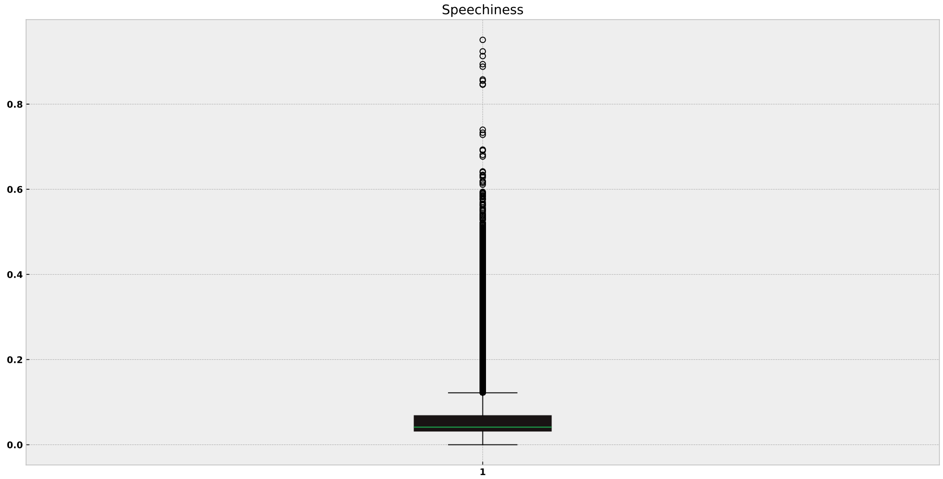

Figure 19: Analysis of Speechiness distributions of the tracks in the Billboard Hot 100 charts. See the data dictionary section for a more detailed description of this feature.

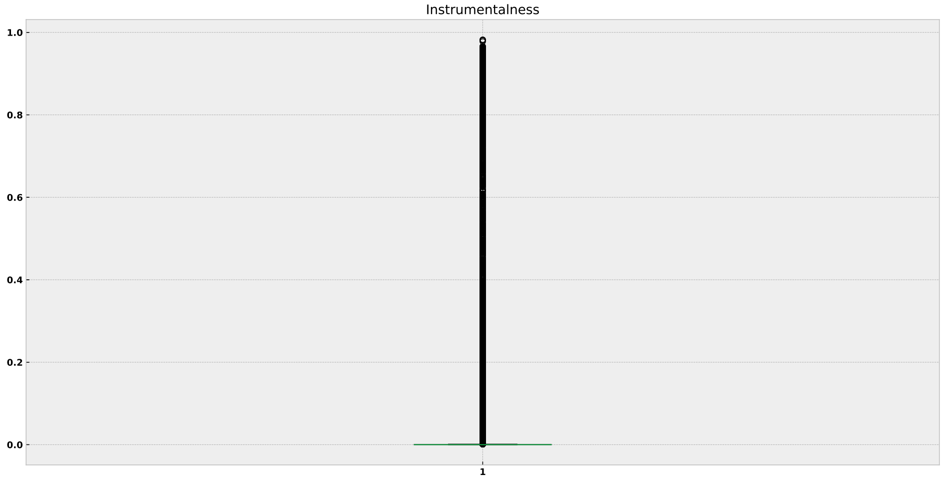

Figure 20: Analysis of Instrumentalness distributions of the tracks in the Billboard Hot 100 charts. See the data dictionary section for a more detailed description of this feature.

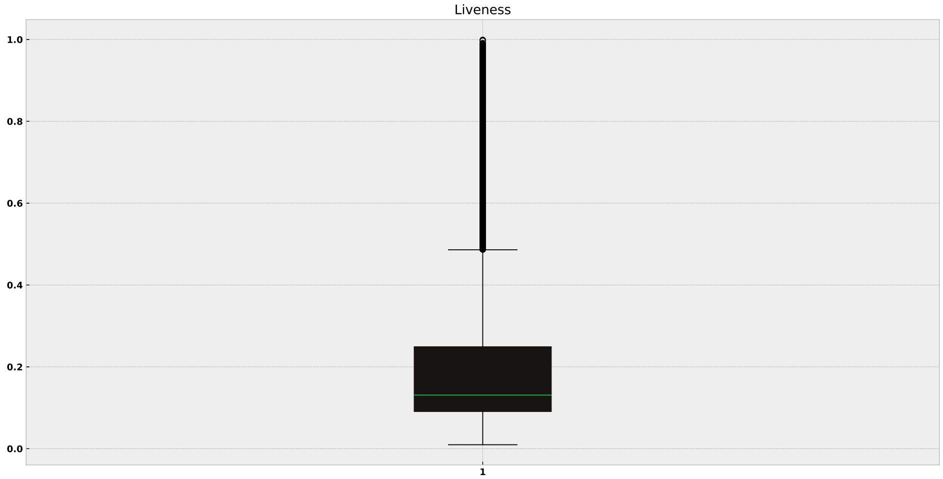

Figure 21: Analysis of Liveness distributions of the tracks in the Billboard Hot 100 charts. See the data dictionary section for a more detailed description of this feature.

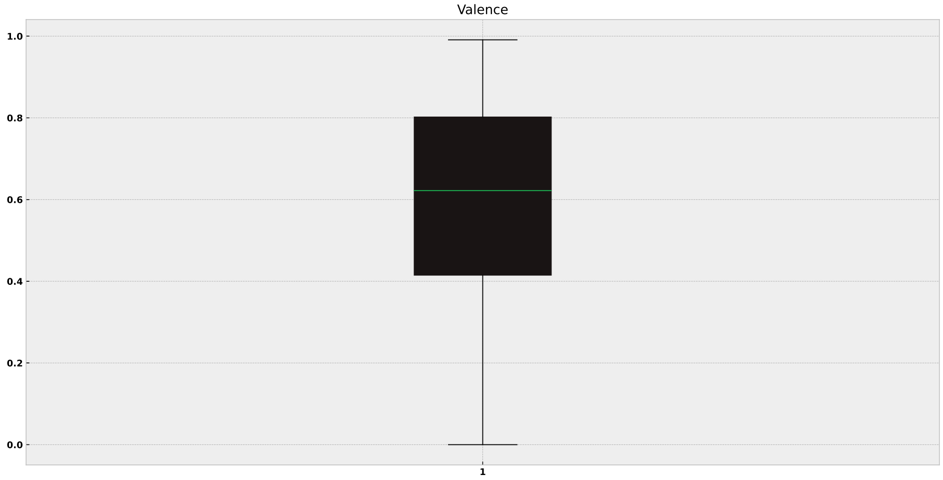

Figure 22: Analysis of Valence distributions of the tracks in the Billboard Hot 100 charts. See the data dictionary section for a more detailed description of this feature.

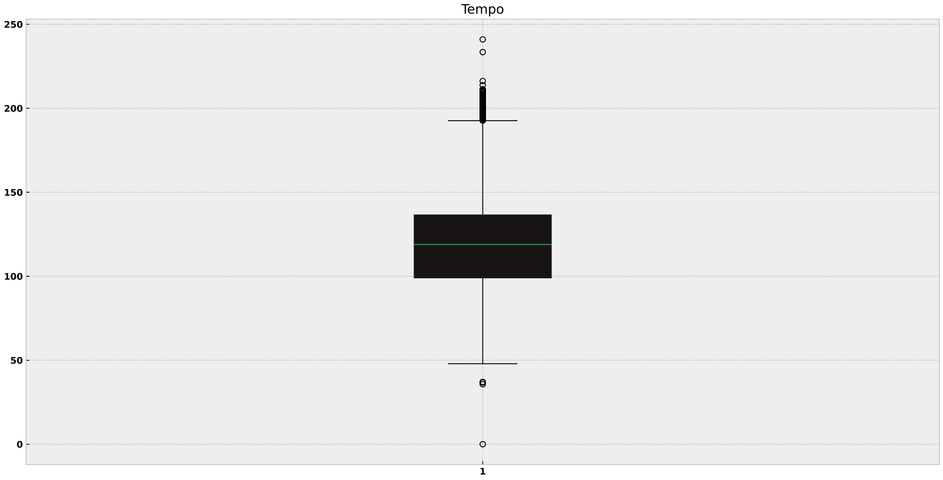

Figure 23: Analysis of Tempo distributions of the tracks in the Billboard Hot 100 charts. See the data dictionary section for a more detailed description of this feature.

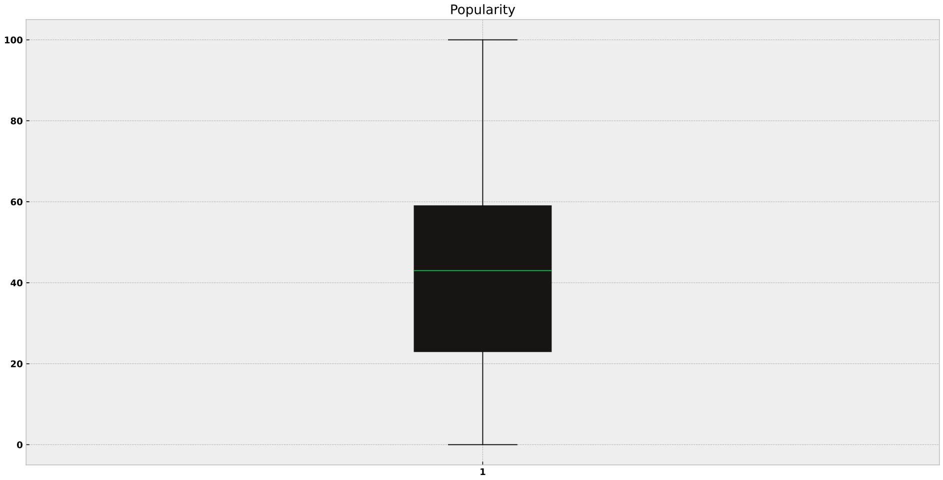

Figure 24: Analysis of Popularity distributions of the tracks in the Billboard Hot 100 charts. See the data dictionary section for a more detailed description of this feature.