A Deep Dive into Spotify's Musical Landscape

s2580177 **s2614493** **s2465354** **s2520327**

Abstract

Throughout our lives, we all connect with music for a variety of reasons in different situations, be it for entertainment, a means of relaxation, a mood booster, and even an outlet for dancing. This analysis delves into the intricacies of the Spotify top 200 global dataset, showing trends, correlations, and unique characteristics of songs. The exploration covers audio features, genres, and the evolution of the musical landscape over a six-year period. Quantifying the influence of the COVID-19 pandemic, the prevalence of genres and the effects of collaboration, the analysis provides a comprehensive understanding of Spotify's musical ecosystem.

1 Introduction

Spotify characterises songs via numerical audio features, such as loudness, danceability, valence and others. The aim of this project is to extract interpretable insights that are useful for Spotify and help improve the user experience. Using correlation analysis we establish connections between audio features. By applying spectral clustering we aim to split the songs into genres as well as quantify the correlation between audio-features. Comparing our work to prior research [1] on different but similar data sets, we try to estimate how accurate it is to split the genres by audio features. Our temporal analysis of the data produces insights into users' preferences over time and may provide guidance on what songs to recommend in specific scenarios. Our analysis of collaborative works confirms Spotify's own research [2] as well as long-established global trends [3] and provides an insight into the positive effect of collaboration.

The dataset comprises the daily top 200 songs spanning from January 1, 2017, to May 29, 2023, featuring a total of 7,457 unique songs. Each entry includes seven distinct audio features, and additional details about the song, such as the date, rank, artist's name and nationality, and points are provided.

Currently, around 10% of the world's preferred music consumption is trouogh Spotify [4][5]. Hence the recommendation algorithm employed to present users with music to listen has a huge impact on our daily lives. We explore ways to promote music that will be more popular with users.

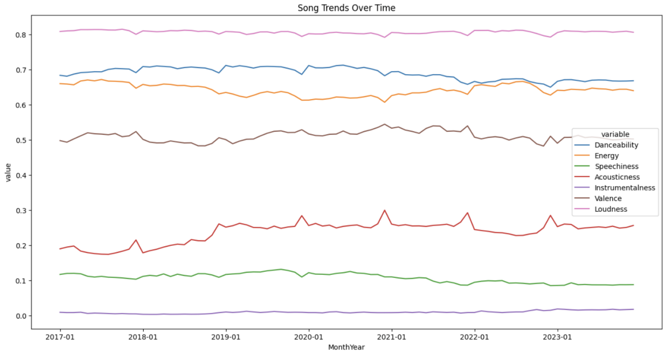
2 Data preparation

It was observed that all audio features were scaled between 0 and 1, except for the 'Loudness' column, which exhibited a significantly different range (-34500 to 1500). To ensure consistency and facilitate meaningful comparisons between features, a normalization process was applied specifically to the 'Loudness' column. We further use an additional genre column to represent genre of each of the songs. The 'Genre' column is added to categorize songs into genres such as EDM, Hip-Hop, Rap, Pop, R&B, and Latin, based on specific range of important audio features. To further enhance our analytical capabilities, we expanded the dataset to include information about the season and day of the week for each row. This addition lays the groundwork for seasonal analysis and the identification of periodic trends within the dataset. When exploring the impact of collaboration we use the sum of the involved artists' points to analyse the popularity of songs. Some tracks have the artists listed in a different order depending on the day that song was in the top 200 global playlist. We rearrange where necessary to avoid duplicate listings later.

3 Exploratory data analysis

3.1 Audio Feature Trends Over Time

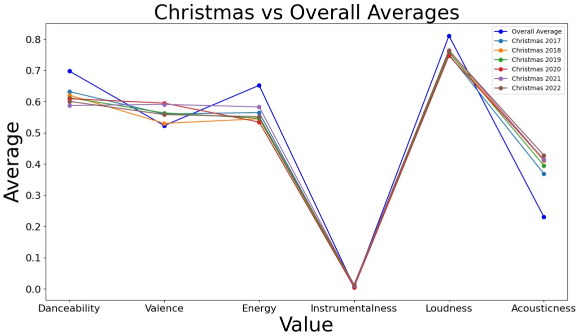
We plotted [6] multiple line charts on the same plot to find out any significant changes in the audio features [7]. The multiple line charts revealed trends in audio features over the six-year period, and specific attention was paid to potential changes associated with the COVID-19 pandemic. Key findings include:



Covid: During the COVID-19 pandemic we observe a slight dip in energy acompanied by a slight increase in acousticness and valence. Danceability declines overall in the entire observed timeframe. Decreasing danceability and energy during COVID may be due to restrictions on social gatherings or reflect a shift towards more subdued and introspective music during challenging times.
Speechiness remains stable around slightly more than 0.1, but there is a discernible increase by 0.01 during the COVID-19 period.
The rise in speechiness during the COVID-19 period may be indicative of changes in lyrical styles or a greater emphasis on spoken words in music during this time.

Seasons: Instrumentalness experiences a dip in the summer, aligning with a potential shift towards more vocal-centric party-music during warmer months. This is confirmed by higher values of danceability, energy, loudness and valence observed during the summer, possibly influenced by more frequent social gatherings. Small dips at the end of each year are observed, indicating a potential seasonal influence on danceability trends.
The seasonal preferences imply that listeners associate specific musical characteristics with different times of the year. These preferences may be influenced by changes in social behaviour, or they may indicate that users seek music that complements the mood and atmosphere of each season.
Notably, there are spikes in acousticness at the end of each year, suggesting a potential seasonal influence or a preference for acoustic tracks during certain periods.

Holidays: Mariah Carey's "All I Want for Christmas Is You" consistently ranks #1 during Christmas from 2017 to 2022 suggesting a strong preference for well-known classics. Analyzing the musical features of songs consistently ranked high during Christmas shows some regularities: firstly, a prevalence of higher instrumentalness; secondly, these songs tend to exhibit lower energy levels; thirdly, lower danceability and lastly, higher valence which all together make up the spirit of Christmas festivities.

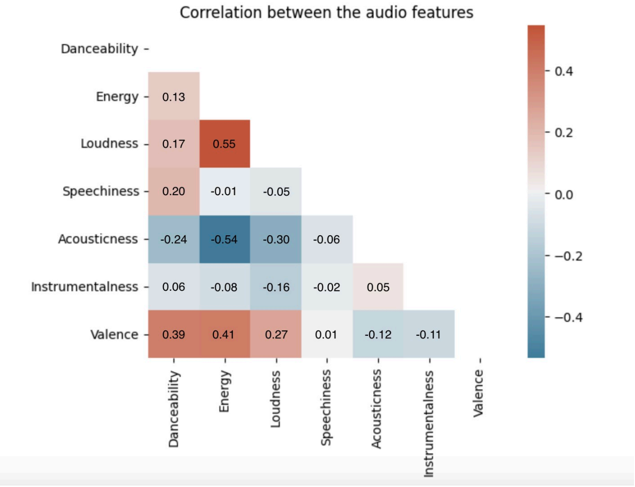


"All of Me" by John Legend maintains a consistent presence in the top 10 during Valentine's Day.

Days: The slight variation in instrumentalness and loudness across different days suggests that listeners have specific preferences throughout the week. Mondays and Fridays may see a preference for more instrumental music, while Fridays and Saturdays may attract listeners who enjoy louder tracks with higher danceability.

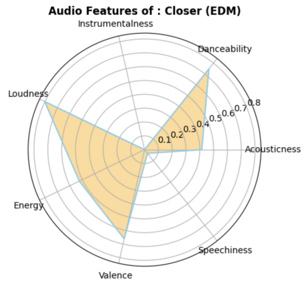
3.2 Correlation between audio features

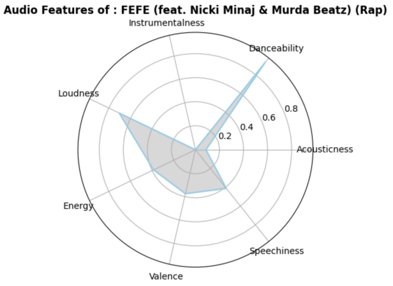
We check the correlation between audio features using a seaborn heatmap. Majorly, Loudness seems to be positively correlated with Energy, and Acousticness seems to be negatively correlated with Energy.
Since there are no variables with extremely high correlation, we can't drop any column based on this analysis.

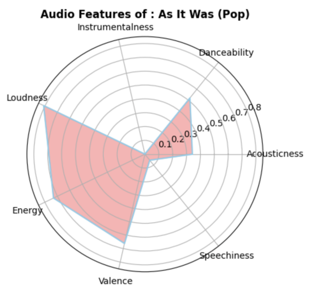


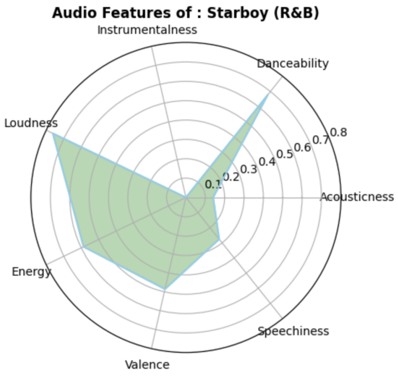
3.3 Audio features of Top-Ranking Genres

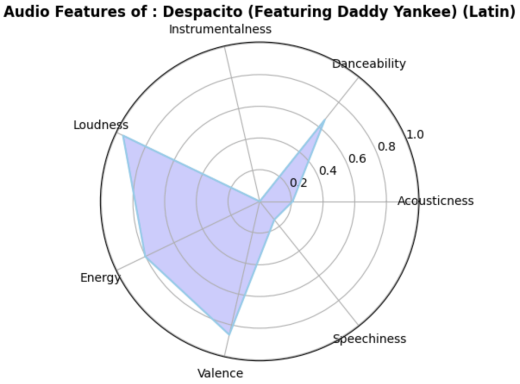
To find out the unique audio features of a particular song, we make a radar charts, which provided a detailed view of audio features [8]. As an example, we have taken one song per genre to compare the audio features. The use of radar charts for individual songs, enhances our ability to grasp the nuanced audio features that define a particular piece of music. This approach facilitates a more in-depth appreciation of the distinct musical characteristics that contribute to the song's identity.







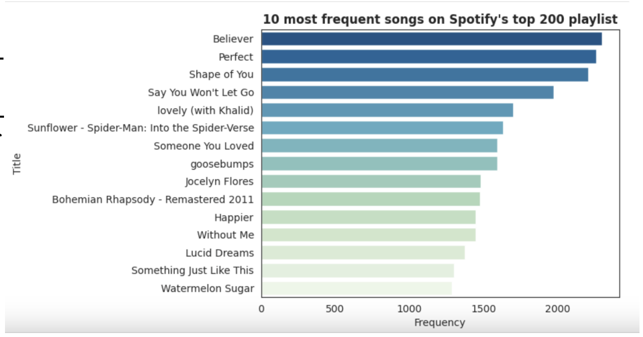




The audio charts highlight distinct patterns among the top-ranked songs; we can see from the audio charts, the top ranked songs are loud, energetic, and danceable. However, it changes from genre to genre. E.g., EDM stands out for its elevated levels of energy, loudness, and danceability. Hip-Hop, while still danceable and loud, exhibits slightly lower energy levels. In contrast, Rap songs are characterized by high speechiness, emphasizing the prominence of spoken words. Pop songs, although loud, tend to be less danceable compared to other genres.

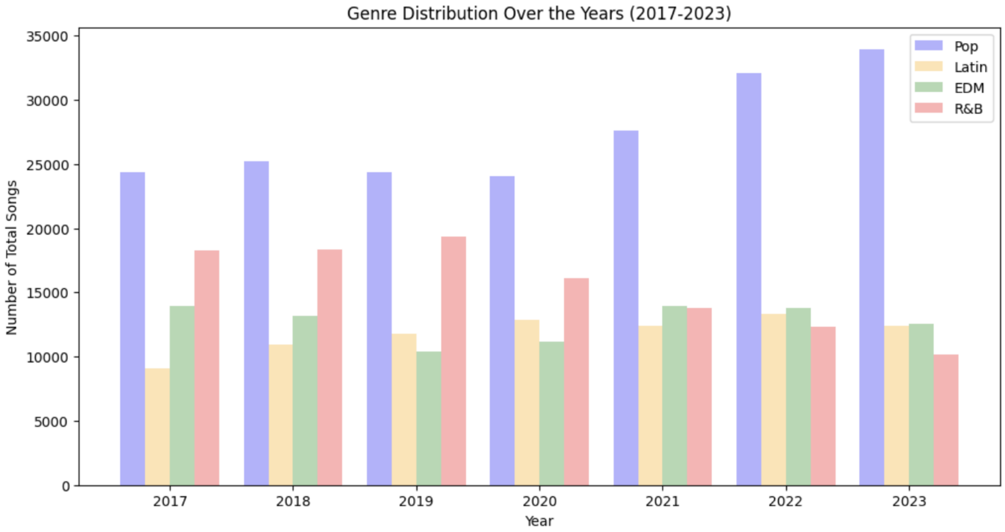
3.4 10 Most Popular songs

We showcase bar plots of the most popular songs based on their recurrence in the list of songs. To get the correct analysis, we dropped columns related to # of artist, their nationality, continent etc. so that one entry of a songs for one date is not repeated. Thus, we find out the top 10 most recurring songs during 2017 to 2023.



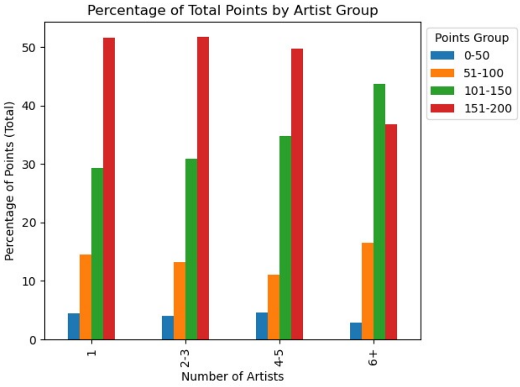
3.5 Genre Distribution over the Years

Examining the graph reveals a consistent upward trend in the number of Pop songs each year, notably surging from 2020 onward. Furthermore, the presence of Latin songs within the top 200 playlist increased after 2018. The popularity of EDM experienced a peak in 2017 and 2018 but exhibited a decline in 2019 and 2020, potentially influenced by the COVID-19 pandemic. However, a resurgence in EDM's popularity is evident post-2020. Additionally, the graph showcases a peak in R&B popularity from 2017 to 2019, followed by a successive yearly decline after 2019.



3.6 The impact of collaboration

Confirming long-established global trends [3] we see that songs with more than one artist are more popular on average. We see no statistically significant increase in popularity for subsequent additional artists though. This indicates that collaboration has a binary effect on popularity.



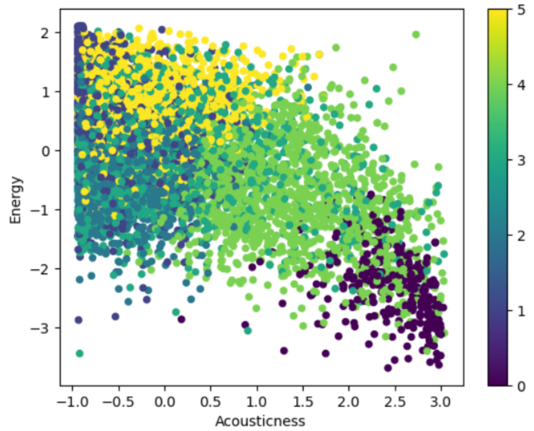
4 Learning methods

We wish to test the hypothesis that genres are split by numerical audio features. Under this hypothesis we need to understand how the border between genres is characterised to choose the appropriate clustering algorithm. The songs belonging to same genre are similar to each other in terms of some numerical features producing a connected point cloud [9]. But the numerical features themselves may vary. Then genres would not be grouped around a genre mean, but instead be spatially connected point clouds. To test this we employ spectral clustering due to its ability to recognise those connected shapes.
Logistic Regression and Random Forest fail due to the nature of the data points, each only achieving scores of around 4%.

Spectral clustering [10] uses the eigenvalues of the similarity matrix of the data to perform dimensionality reduction before clustering in a lower-dimensional space. The similarity matrix calculated as the matrix of relative similarity of each pair of points in the dataset. A lower bound on the complexity is O(n) because a sparse matrix is constructed using the nearest neighbours method and the number of eigenvectors is a lot smaller than the number of data points.

A way to intuitively motivate this choice is to think about the rock music genre. Tracks belonging to that genre may vary greatly in loudness and valence as well as energy. The different sub-genres of rock music (country rock, folk rock, blues rock, punk rock and others).
We use the genres identified in 3.5, discounting the latin-american songs because those are varied in style and split into many smaller categories. We want to compare the genre clustering provided by the spectral clustering method to a manual identification of genres for a sample of songs.

Now we check whether we have accurately identified music genres. Taking samples we see that although some of our clusters, in particular the clusters 0,2 and 3 are identified with the genres Pop, EDM and R&B our other clusters are not representative of specific genres. The clusters we identify with genres do not perfectly line up with the manually labeled genres, but they contain a majority of the genre associated with them.

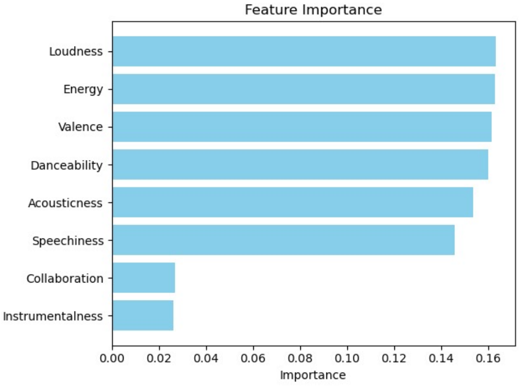


label	largest proportion of genre
0	pop
1	undefined
2	Hip-Hop, Rap
3	EDM
4	R&B
5	undefined

It is clearly visible that cluster 0 above is defined by high acousticness and low energy. The other clusters, in particular 1,3 and 5 are separated by other features. This reaffirms our choice of clustering method, since we want to see some cluster boundaries ignore subsets of features, as motivated in the example of rock music above.

We also want to predict a song's popularity from directly from its numerical features and from the number of collaborators.

Because some of the features, such as the size of the collaborative collective and instrumentalness, behave more like a categorical features, we use a random forest regression and achieve a mean squared error of 3449. Since the y-values are 1-200, this means that our prediction quality is poor, but we may still extract useful information from the feature importance graph.



5 Results

The above findings show that the genre, while sometimes characterised by numerical features, is reliant on human perception. This agrees with prior research [1] done on the subject using the Free Music Archive data set featuring a similar number of songs and similar numerical features.

The feature importance graph clearly shows that the numerical audio features, barring instrumentalness, are much more important than the collaborators in predicting a song's popularity. No feature stands out though, and the model itself does not perform well, so we cannot conclude that a specific feature predicts a song's success. This aligns with our understanding that music cannot be easily quantified.

Using our exploratory data analysis, Spotify can curate Christmas playlists to be most successful by combining timeless hits with potential new favorites that match the seasonal preference to create a festive mood, capturing the both nostalgia and novelty in holiday music.

6 Conclusions

A question remains whether Spotify itself influences the taste of listeners. It is reasonable to assume that the recommendation algorithms push some songs into popularity that would otherwise not have been as popular. With 10% of music consumption happening through Spotify it is highly likely that when we measure popularity on Spotify we also measure how favourably the algorithm looks upon that song.

Overall, this analysis provides meaningful insights into the popularity of songs on Spotify and attempts to create a way to label each song with a genre. This is particularly useful when creating recommended playlists for Spotify users. Genre-playlists in particular are an example where the classification of a song into a genre has to meet an estimation of its popularity.

Our analysis further confirms a general belief about human preferences, that we strongly associate well-known and well-liked pieces of music to certain celebrations and periodic events such as Christmas and Valentine's Day.

References

[1] V. Ramirez, “Discovering descriptive music genres using k-means clustering,” 2018, last accessed 22 November 2023. [Online]. Available: https://medium.com/latinxinai/discovering-descriptive-music-genres-using-k-means-clustering-d19bdea5e443

[2] unknown, “The crossover effect: Artist collaborations thrive on spotify,” 2023, last accessed 22 November 2023. [Online]. Available: https://newsroom.spotify.com/2023-05-24/crossover-collaborations-genre-collabs-streaming-data-spotify/

[3] P. Kaplan, “The rise of “feat.” in today’s music,” 2017, last accessed 22 November 2023. [Online]. Available: https://news.distrokid.com/the-rise-of-collaborations-in-todays-music-8a8bcd386ea

[4] A. Alexander, “Infographic: How does the world consume music?” 2023, last accessed 22 November 2023. [Online]. Available: https://www.weforum.org/agenda/2023/02/world-consume-music-infographic/

[5] F. Duarte, “Music streaming services stats (2023),” 2023, last accessed 22 November 2023. [Online]. Available: https://explodingtopics.com/blog/music-streaming-stats

[6] “Plotting multiple bar charts using matplotlib in python,” 2021, last accessed 22 November 2023. [Online]. Available: https://www.geeksforgeeks.org/plotting-multiple-bar-charts-using-matplotlib-in-python/

[7] A. Mahajan, “Exploring spotify dataset,” 2021, last accessed 22 November 2023. [Online]. Available: https://www.kaggle.com/code/alankarmahajan/exploring-spotify-dataset

[8] A. M. P. T. G. L. Deniz Duman, Pedro Neto, “Music we move to: Spotify audio features and reasons for listening,” *PLOS ONE*, 2022.

[9] C. P. R. R. G. R. M. R. S. Y. A. D. Michelangelo Harris, Brian Liu and J. Pender, “Analyzing the spotify top 200 through a point process lens,” *School of Operations Research Information Engineering*, 2019.

[10] U. von Luxburg, “A tutorial on spectral clustering,” *Statistics and Computing*, vol. 17, no. 4, 2007.

Statement of Contribution

In this collaborative project, each team member made significant contributions to various aspects of the analysis, ensuring a comprehensive exploration of the dataset. Our collective efforts were directed towards understanding different facets of the music dataset, ranging from audio feature trends to genre clustering and the impact of collaboration.

s2465354's Contribution: s2465354 took charge of investigating the trends of audio features over time, diving deep into the nuanced variations present in Christmas, seasonal, daily, and yearly patterns. Her analysis provided valuable insights into how audio features evolve and adapt to different temporal contexts, enriching our understanding of the dynamic nature of music.

s2614493's Contribution: s2614493 led the genre clustering efforts, applying sophisticated techniques to group similar genres together with unsupervised machine learning. His work laid the foundation for a more organized and coherent representation of the diverse music present in the dataset. By identifying commonalities among genres, Daniel's clustering analysis offered a deeper understanding of the inherent relationships between different music styles.

s2580177's Contribution: s2580177 focused on clustering using strict boundaries and contributed to all analyses related to genres. Her approach to defining boundaries and exploring the intricacies within genre categories brought depth to our understanding of the dataset. s2580177's work added granularity to the genre classification, ensuring a more nuanced representation of musical styles.

s2520327's Contribution: s2520327's primary focus was on exploring the impact of collaboration within the dataset. By examining collaborative relationships between artists and their influence on song popularity, Rain shed light on the social dynamics within the music industry. His analysis provided valuable insights into how collaborative efforts contribute to the overall success and popularity of songs.

Collective Contribution: As a team, we collaborated on the introduction and data preparation stages of the project. By collectively working on these foundational aspects, we ensured a cohesive and well-prepared dataset for subsequent analyses. Our combined efforts in these initial phases set the stage for the in-depth explorations conducted by individual team members.

Through our collaborative and individual contributions, we aimed to provide a comprehensive understanding of the dataset, uncovering patterns, relationships, and insights that contribute to the broader discourse on music trends and collaborations.

Appendix

A Description of Audio Features

- **Danceability:** 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:** 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.
- **Loudness:** The overall loudness of a track in some unit. Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks.
- **Speechiness:** 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.
- **Acousticness:** A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic.
- **Instrumentalness:** Predicts whether a track contains no vocals. The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content.
- **Valence:**Valence: A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive, while tracks with low valence sound more negative.