Lyricalytics - Milestone 1

Dataset

For this project, we will be using three Kaggle datasets to analyze the relationship between musical success, song genre, and lyrics (under an emotion analysis lens):

1. Spotify Tracks Dataset – Contains detailed metadata on 89'741 songs available on Spotify, including track features such as tempo, danceability, energy, as well as the associated genre (such as pop, acoustic, rock, etc.). Available at https://www.kaggle.com/datasets/maharshipandya/-spotify-tracks-dataset.
2. 960K Spotify Songs With Lyrics Data – Includes lyrics for 955'320 of Spotify songs, providing the textual data necessary for sentiment and emotion analysis. Available at https://www.kaggle.com/datasets/bwandowando/spotify-songs-with-attributes-and-lyrics
3. Top Hits Spotify from 2000-2019 – Lists the 2000 top-charting songs over a two-decade period, helping us link sentiment and emotions to commercial success. Available at https://www.kaggle.com/datasets/paradisejoy/top-hits-spotify-from-20002019/data

When picking the datasets, we took care that all are related to Spotify data. The Spotify Tracks Dataset and 960K Spotify Songs With Lyrics Data both contain track ids that can be directly matched. Top Hits Spotify from 2000-2019 did not have track ids, but was directly matched to Spotify Tracks Dataset by guaranteeing cleaned versions of both the artist names and song title, as well as track features, were identical.

The final merged dataset was a total of 80274 songs with matching lyrics, 1868 of which are top hits. We have a total of 114 different genres, and there are no null or missing values. Please refer to Appendix 1 for the full list of dataset features and their descriptions.

Problematic

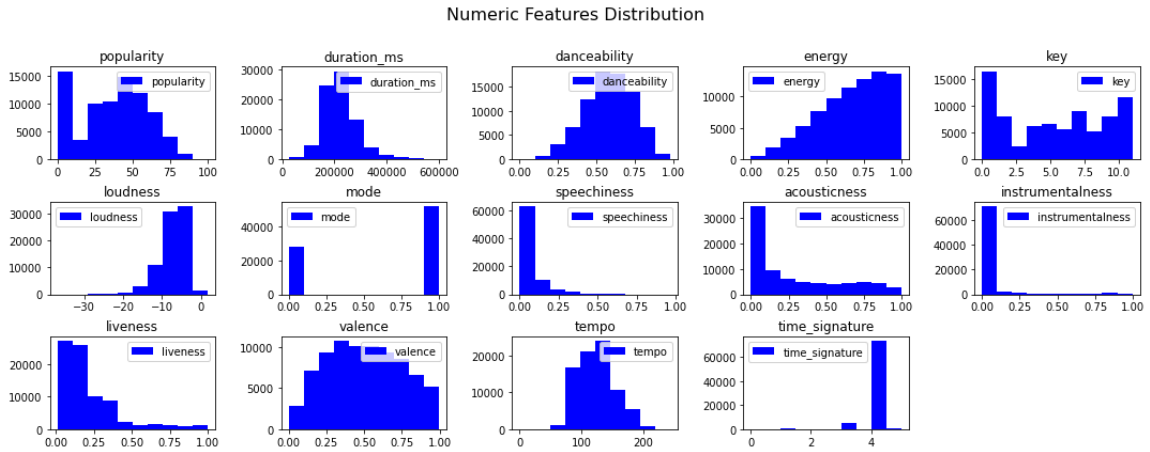
We seek to explore the relationship between a song lyrics' emotions, song features, and song popularity. In particular, with this project, we seek to understand the impact of songs' lyrical sentiments. We will investigate the correlation between the emotions present in lyrics and intrinsic song features – such as tempo, mode, liveness, etc. – and extrinsic features – namely, whether the song was a hit song or not.

We will answer questions such as:

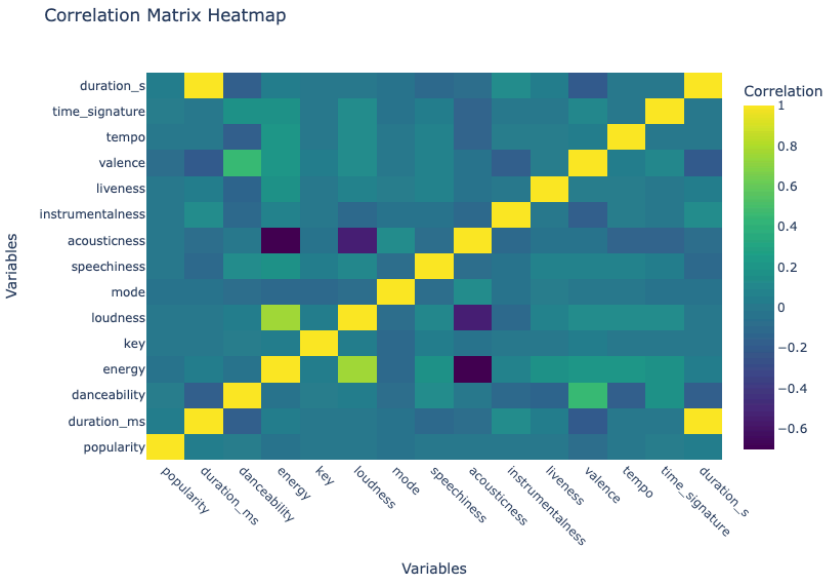
- Are happier or sadder songs more likely to become hits?
- In terms of features, do more energetic or danceable songs have a higher probability of being popular? And are these types of features correlated with positive emotions in lyrics?
- Is there a relationship between genre and popularity, and between genre and the lyrical sentiments?

Exploratory Data Analysis

We explore each feature's distribution and the correlation between them. Numeric features follow different kinds of distribution. For instance, track duration and danceability seem to follow a normal distribution. Energy follows a left skewed distribution. Mode seems to follow a binary distribution.



To understand the feature's correlation, we map them in the matrix below, where yellow represents the max positive correlation, and dark purple the max negative correlation in our data.

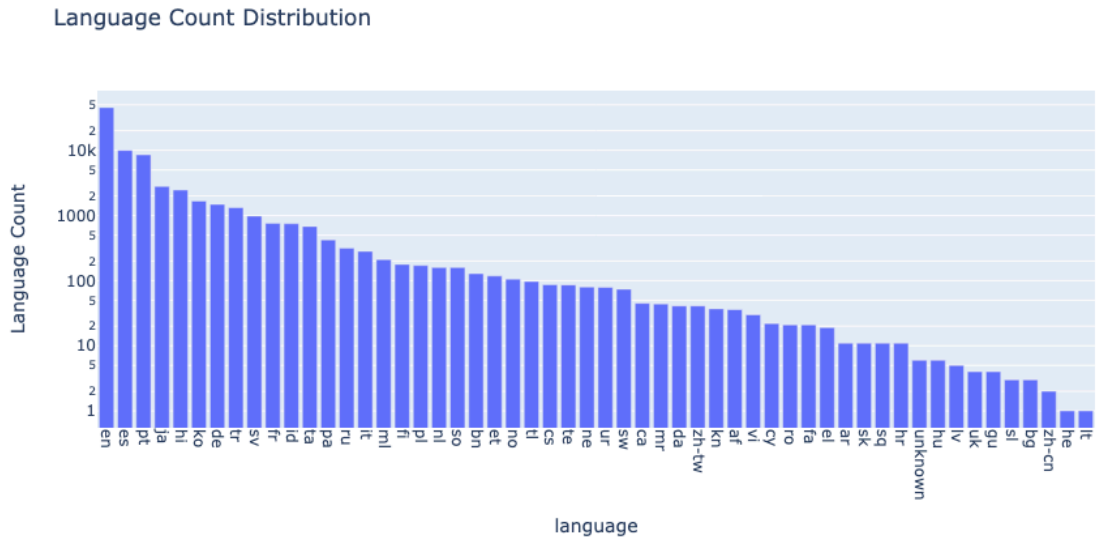


We can make several observations:

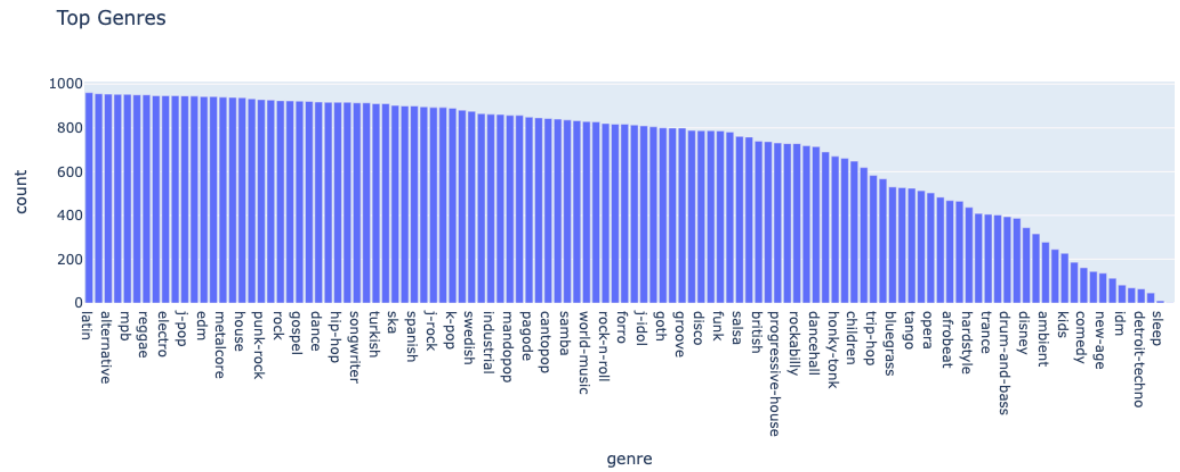
- Loudness and Energy are positively correlated; meaning that songs with high energy are louder (possibly due to heavy beats, louder vocals, and more intense instrumental sections).

- Valence and Danceability are also positively correlated. Upbeat songs tend to be more danceable because they naturally evoke more positive emotions (high valence), making people want to move and dance.
- Acousticness and Energy are negatively correlated. Songs with high acousticness tend to be softer, more laid-back, and less electronically enhanced, which can result in lower energy levels.
- Acousticness and loudness are also negatively correlated. Acoustic songs typically have lower loudness because they are not heavily compressed or amplified, therefore tending to be quieter.

After applying Google's *language-detection* (using langdetect), we identify a total of 53 different languages in our data, as well as several songs with unknown language. English has the most songs (45k+), followed by Spanish (9.9k+). We will discard the songs for which we could not identify the language, as they do not represent a significant amount of data (see the "unknown" column in the image below; please note the log y-axis).



The song genres are fairly evenly distributed, as opposed to the language's long-tailed distribution. The two least common genres will be discarded, resulting in a total of 112 unique genres.



Related Work

Sentiment and Emotion Analysis in Lyrics

- **Bebbington et al. (2023)**, *Cultural evolution of emotional expression in 50 years of song lyrics*: Despite lacking temporal data, we will similarly analyze sentiment and emotion in lyrics as well.
- **Kwon, L. et al. (2021)**, *Trends in Positive, Negative, and Neutral Themes of Popular Music From 1998 to 2018*: Again, while we can't track trends over time, we'll also examine whether certain genres show higher negative sentiment.

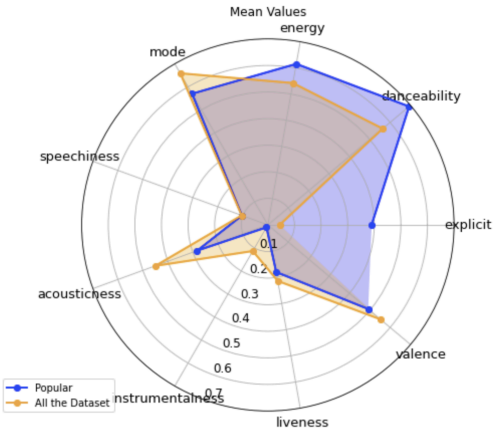
Lyrics Differentiation and Popularity

- **Berger & Packard (2020)**, *Are Atypical Songs More Popular? Lyrical differentiation and song popularity*: We'll also test if emotional differentiation in lyrics correlates with popularity.
- **Seufitelli, D. B. et al. (2023)**, *Hit song science: a comprehensive survey and research directions*: Partially supports our focus on how intrinsic lyrical features relate to popularity.

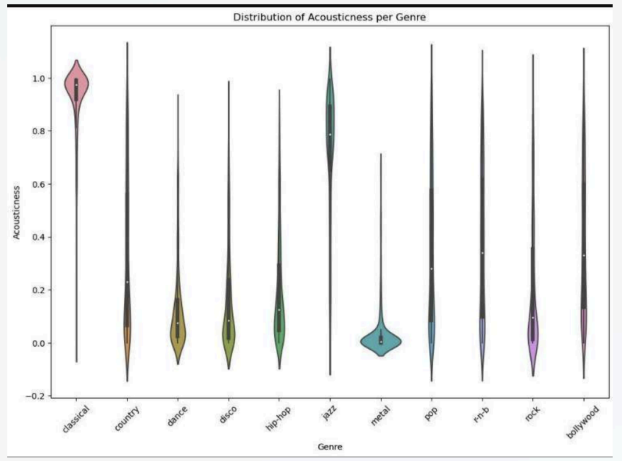
Please refer to Appendix 2 for more details on each work.

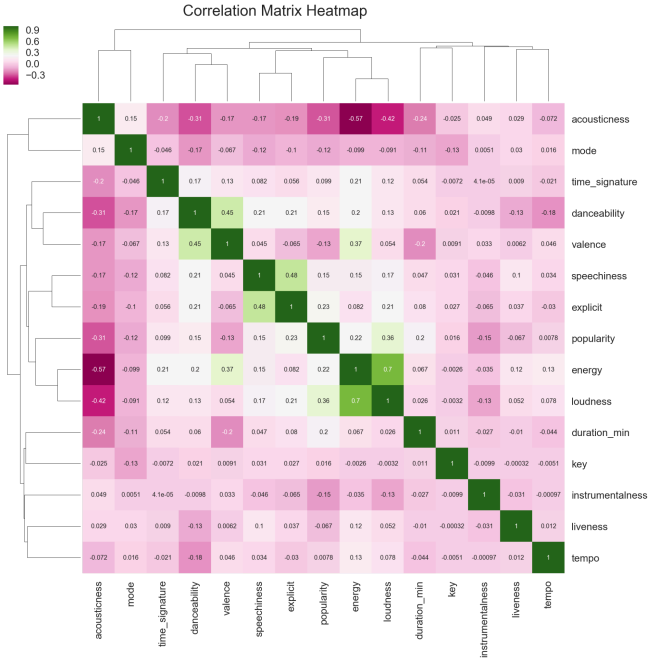
Data Visualization References

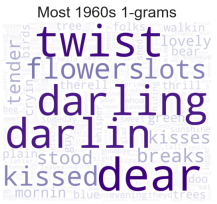
- **Spotify Dataset Visualization**
    - Informs our approach and warns of common pitfalls.
    - **General EDA**
    - **Lyric Analysis**
    - **ML Classification on Spotify Data**
    - **Basic Examples:**

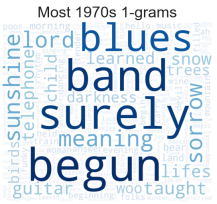


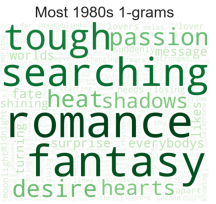
Radar chart of mean values of features of top 100 songs and the rest of the dataset

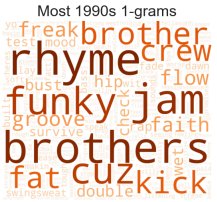

Distribution of Acousticness per Genre
Acousticness
Genre
classical
country
dance
disco
hip-hop
jazz
metal
pop
r+b
rock
bollywood


Correlation Matrix Heatmap
acousticness
mode
time_signature
danceability
valence
speechiness
explicit
popularity
energy
loudness
duration_min
key
instrumentalness
liveness
tempo

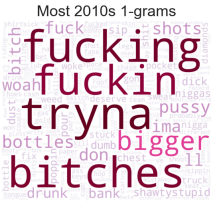

Most 1960s 1-grams
twist
flowers
darling
darlin
kisses
kissed
dear
breaks


Most 1970s 1-grams
blues
band
surely
begun
meaning
sunshine
sorrow

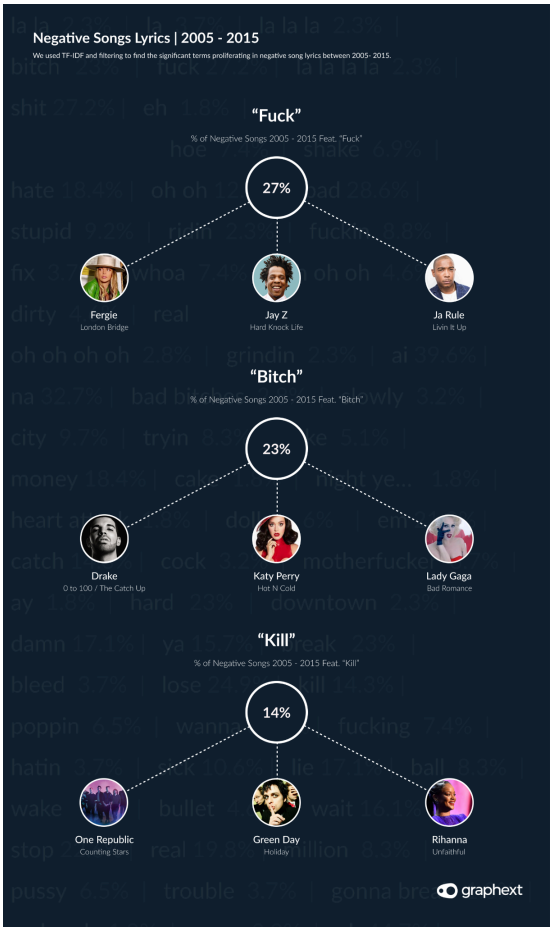

Most 1980s 1-grams
tough
passion
searching
romance
fantasy
desire
hearts


Most 1990s 1-grams
brother
rhyme
crew
funky
jam
brothers
fat
cuz
kick

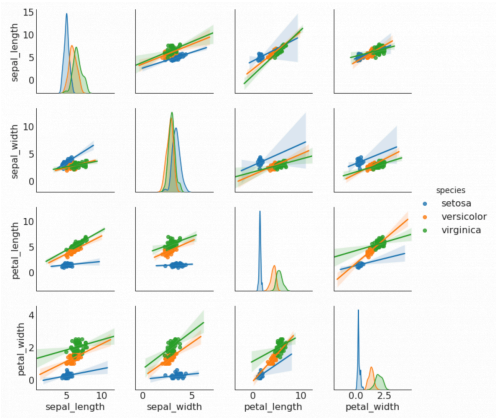

Most 2000s 1-grams
thugwit
shorty
holla
chicks
shawty


Most 2010s 1-grams
fucking
fuckin
tryna
bitches
bigger

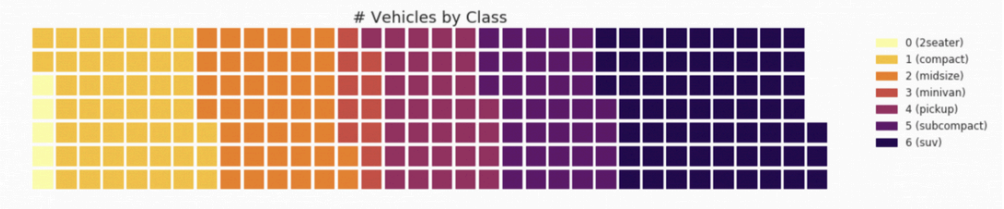
- **Sentiment Trends in Music** : Maps emotional tone across artists—an approach we could adopt.



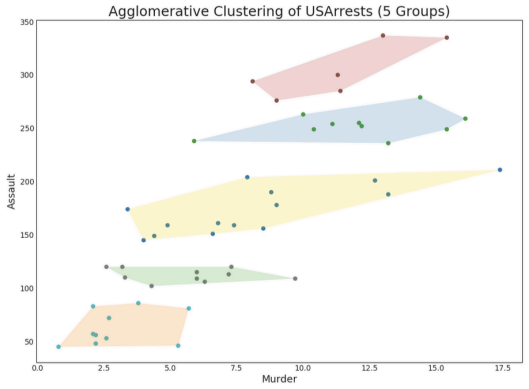
- **Ideas with Matplotlib Examples**
  - *Correlation Analysis*: Explore relationships among song features.



- *Waffle Plot*: Visualize lyric language distribution.



- *Cluster Plot (PCA/UMAP)*: Show emotional/lyrical similarity between songs.



Interactive Visualization References

- **How Music Taste Evolved** : Inspires adding audio playback.
- **Are Pop Lyrics Getting More Repetitive?** : Scroll-based interactions to explore lyrics.
- **Building Music Galaxy** : This visualization presents an embedding space for artists with an interactive Spotify connection, allowing users to explore musical relationships. We can try creating an interactive “emotion space” for songs as well.

Appendix 1 - Dataset Feature List

Our dataset contains a numerous amount of features. We describe the base features below:

1. **track_id:** spotify song id (string)
2. **artists:** name of the song artists (string)
3. **album_name:** name of the album the song comes from (string)
4. **track_name:** name of the song (string)
5. **popularity:** integer between 0 and 100 denoting the song's recent popularity[1] (integer)
6. **duration_ms:** duration of the song in milliseconds (integer)
7. **explicit:** whether the lyrics of a song or music video include one or more elements that may be deemed inappropriate or offensive for children (boolean)
8. **danceability:** how appropriate a track is for dancing (from 0 to 1, where 1 is the most danceable), determined by factors like tempo, rhythm consistency, beat strength, and overall regularity (float)
9. **energy:** a perceptual measure of intensity and activity in the song, from 0 to 1 where 1 is the most energetic (float)
10. **key:** an integer representation of the track's key, using standard Pitch Class notation (e.g., 0 = C, 1 = C♯/D♭, 2 = D, etc.; if no key is detected, the value is -1) (integer)
11. **loudness:** the total loudness of a track measured in decibels (dB)[2] (integer)
12. **mode:** indicates the musical scale (major (1) or minor (0)) of a track, reflecting the type of scale used for its melodic structure (integer)
13. **speechiness:** measures the extent of spoken words in a track; the more speech-like the content (e.g., talk shows, audiobooks, poetry), the closer the value is to 1[3] (float)
14. **acousticness:** a confidence measure from 0 to 1 of whether the track is acoustic; 1 represents high confidence the track is acoustic (float)
15. **instrumentalness:** whether a track is entirely instrumental, with sounds like 'ooh' and 'aah' considered as instrumental; the closer the instrumentalness value is to 1, the more likely the track has no vocal content[4] (float)
16. **liveness:** whether an audience is present in the recording; higher liveness values indicate a greater likelihood that the track was recorded live[5] (float)
17. **valence:** indicates the level of musical positivity in a track, ranging from 0 to 1; tracks with high valence sound more upbeat (e.g., happy, cheerful, euphoric), while tracks with low valence tend to sound more negative (e.g., sad, depressed, angry) (float)
18. **tempo:** the overall estimated tempo of a track, measured in beats per minute (BPM) (float)
19. **time_signature:** an estimated time signature, which is a notation that defines how many beats are in each measure (or bar); the values range from 3 to 7, representing time signatures such as 3/4 through 7/4 (integer)
20. **track_genre:** genre of the song (e.g., acoustic, pop, rock, etc.) (string)

---

[1] Popularity is determined by an algorithm, primarily based on the total number of plays a track has received and how recent those plays are. In general, tracks that are currently being played more frequently will have a higher popularity than those that were played a lot in the past. Duplicate tracks (e.g., the same track from both a single and an album) are rated separately. The popularity of an artist and album is calculated based on the popularity of their tracks

[2] Loudness values are averaged over the entire track and are helpful for comparing the relative loudness of different tracks. Loudness refers to the perception of sound intensity, which correlates with physical amplitude. Values generally range from -60 to 0 dB.

[3] Values above 0.66 indicate tracks mostly consisting of spoken words. Values between 0.33 and 0.66 suggest tracks that may contain a mix of music and speech, such as rap music. Values below 0.33 generally correspond to music and tracks with little or no speech.

[4] Rap or spoken word tracks are categorized as 'vocal'. Values above 0.5 suggest the track is instrumental, with higher confidence as the value approaches 1.0.

[5] A value above 0.8 strongly suggests the track was performed live.

We augmented the dataset to also include the following features:

21. **first_artist:** first artist if the song has a featuring artist (string)
22. **top_hit:** whether the song is a top commercial hit (part of the 2000 top-charting songs on Spotify) (boolean)
23. **lyrics:** the full text of the song's lyrics (string)

There are two important observations:

1. We have top seemingly overlapping features: ***popularity*** (from the base dataset) and ***top_hit*** (from the augmented dataset after merging our different data sources). As the ***popularity*** feature is significantly biased toward recent songs, we do not expect it to reference it in our analysis, opting instead to focus on the (temporally-invariant) top commercial hits (the ***top_hit*** variable).
2. We will remove outliers observed during our preliminary data exploration. In particular, we will **remove songs more than 10 minutes long** to ensure our analysis focuses on songs and not on other media (e.g., audiobooks, podcasts, etc.) We will also **remove the two least common genres** (sleep, with 10 occurrences; study, with 1 occurrence) and their associated entries, for the same reasons. Finally, we will **remove songs from unknown (unidentified) languages**.

Appendix 2 - Related Work Summary

Academic Papers

Sentiment and Emotion Analysis in Lyrics

- **Bebbington et al. (2023)**: *Cultural evolution of emotional expression in 50 years of song lyrics* (Paper)
    - **Takeaway:** The study finds an overall increase in emotionally negative lyrics and a decline in positive ones, providing evidence of content bias, prestige bias, and success bias in this trend. It suggests that negative emotional expression in music serves as a means for individuals to process and express their emotions while also identifying with prestigious figures who share similar negative experiences.
    - **Relevance:** While we do not have temporal data to conduct a longitudinal study, we will do sentiment and emotion analysis as well.
- **Kwon, L. et al. (2021)**: *Trends in Positive, Negative, and Neutral Themes of Popular Music From 1998 to 2018* (Paper)
    - **Takeaway:** Similar to Bebbington et al., this paper finds that negative themes have increased in frequency across all genres over two decades, with hip-hop/R&B exhibiting the highest frequency.
    - **Relevance:** Again, our project lacks a temporal component, but we will also explore whether certain genres have systematically higher negative sentiments.

Lyrics Differentiation and Popularity

- **Berger & Packard (2020)**: *Are Atypical Songs More Popular? Lyrical differentiation and song popularity* (PubMed)
    - **Takeaway:** Songs with lyrics that are more distinct from their genre norms tend to be more popular. However, this effect varies by genre (e.g., weaker for dance music).
    - **Relevance:** We will explore whether emotional differentiation in lyrics influences popularity rather than focusing purely on thematic differentiation.
- **Seufitelli, D. B. et al. (2023)**: *Hit song science: a comprehensive survey and research directions* (Journal of New Music Research)
    - **Takeaway:** Both intrinsic song features (lyrics, acoustic properties) and extrinsic factors (artist popularity, album strategy, streaming influence) contribute to song success.
    - **Relevance:** This motivates our analysis of the relationship between intrinsic lyrical features and song popularity, expecting at least partial correlations.