Kai King

Professor Johnson

ENG 101

15 April 2025

Problem Statement and Background

The music industry has undergone dramatic transformations over the past few decades, driven by technological advancements, the rise of streaming platforms, and shifting cultural trends. Popular music serves as more than entertainment—it acts as a cultural snapshot, reflecting the emotions, values, and identities of the time. By analyzing which genres rise to prominence and how lyrical themes shift, we can gain insight into broader societal trends within listeners.

For example, the mainstream ascent of hip hop signals not just a musical preference, but the rising adoption of black culture. This project tries to understand how these important sonic and lyrical trends evolve. In an era where music data is more recorded than ever, analyzing the evolution of popular music is essential for listeners, artists, and industry professionals to better understand the path mainstream music and culture is heading.

Datasets used

The Billboard Hot 100 Dataset (1958–2024) contains weekly chart rankings of songs from Billboard Hot 100. The data was from the University of Texas at Austin's School of Journalism's GitHub. The data was obtained by web scraping Billboard's website. It contains only public, non-personal data (song titles, artists, rankings). However, there are ethical concerns as scraping may violate Billboard’s terms (non-commercial use restrictions). Additionally, the dataset may be biased as it does not represent non-mainstream music scenes. Billboard's ranking methodology has also changed over the years, incorporating radio, physical sales, and streaming in different ways at different times.

The Genius API Data provides song lyrics from Billboard Hot 100 (1958–2024). The source is Genius.com API, accessed via the lyrics genius Python wrapper. Lyrics were scraped using artist and song titles. The lyrics genius library requests the Genius API to obtain the song’s URL metadata to then web scrape using BeautifulSoup. Genius lyrics are publicly available, but scraping may violate terms of service. There is some bias, as due to web scraping errors some songs were not detected, and in rare cases some lyrics are scraped in a different language.

The Spotify API gives music genres for artists on the Billboard Hot 100 (1958–2024). The data comes from Spotify's artist metadata, accessed via Spotipy. Genres were fetched using artist names from the Billboard data through API calls. This uses public Spotify data and no personal user data was collected. However, there is bias—Spotify's genre detection is partially algorithm-based, and genre is subjective, so it is not always accurate. Spotify does not classify songs by genre, only the associated artist, and many artists don’t even have a genre assigned—so the data should be taken lightly.

The Chosic Genre Dataset contained mappings for Spotify's subgenres into bigger widely recognized categories. The data was collected from Chosic.com by scraping and parsing genre names from the site into a spreadsheet to help simplify Spotify's artist genre tags. There are no direct privacy concerns, as the site lists non-personal, publicly available genre information. Ethical concerns could arise around usage, as the site does not have a way to properly access its data and scraping may go against its terms of service. There could be bias as the categorization of subgenres is subjective and is not always accurate algorithmically.

Data Science Approaches

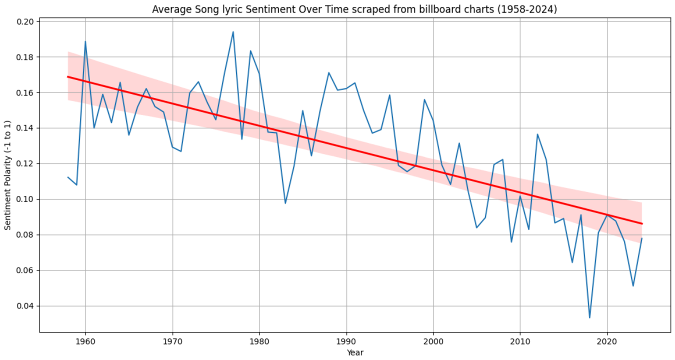
To assess the emotional tone of song lyrics, I used NLP algorithms such as sentiment analysis through the TextBlob library. This algorithm quantifies the polarity of each word on a scale from -1 (negative) to 1 (positive) to give a sentiment score for the whole song. Sentiment analysis shows some algorithmic patterns in lyrics, but it doesn't capture everything. To better understand how lyrics have changed over different eras, I also created word clouds. These highlight the most common words and phrases by displaying them in larger text, making trends easier to spot with human eyes.

Additionally, I used Seaborn to create a linear regression for my sentiment vs year and profanity vs year line graphs. This regression was developed using the Least Squares Method to algorithmically find the best-fitting line, minimizing the total squared differences between the actual y value and the predicted y value on the line.

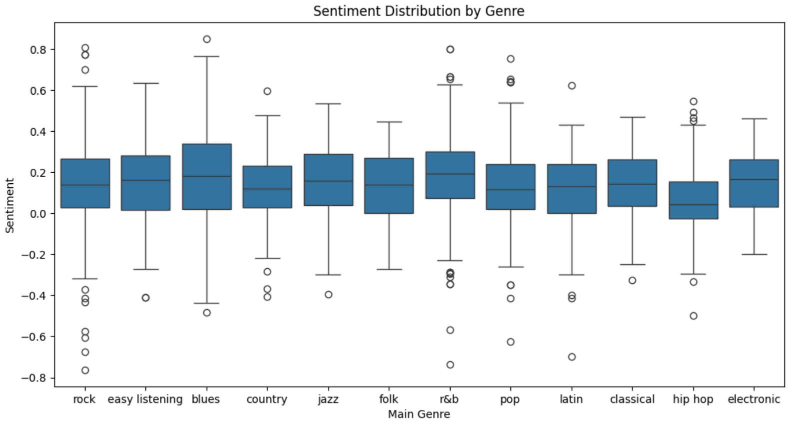
To collect all this data, I also had to run my data-retrieving algorithm overnight to ensure I didn't break API rate limits. I automatically saved instances of my data frames as CSVs on my computer to avoid having to repeat this process during the analysis section.

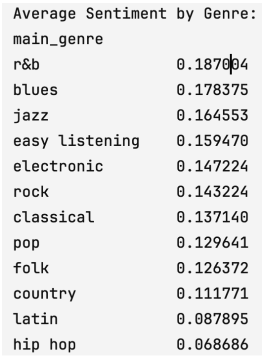
Results and Conclusions

The "Average Song Lyric Sentiment Over Time" graph indicates a gradual decline in positivity from the 1960s to the 2020s, with lyrics becoming more neutral or emotionally complex. Peaks in sentiment (e.g., late 1970s, 1990s) may correlate with the puritanical beliefs of the time, while the dips from the 2000s could reflect the adoption of profanity even within the charts.



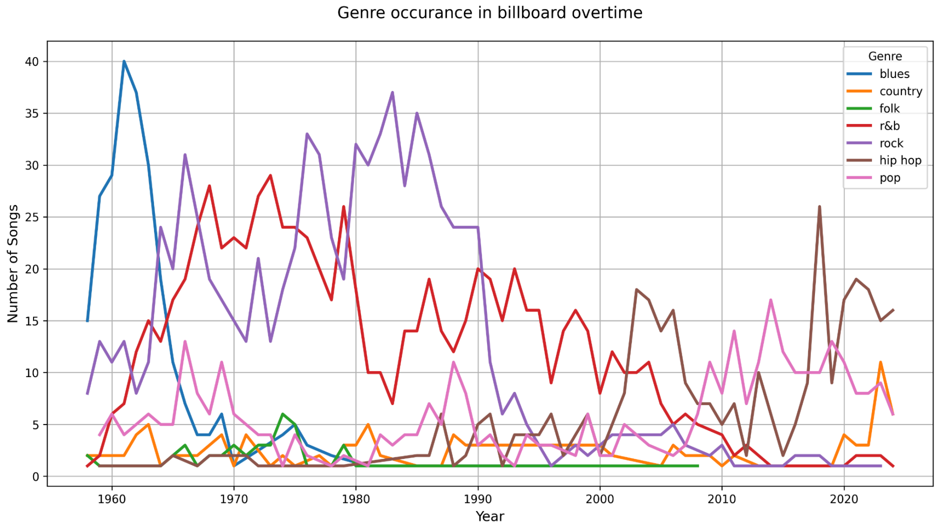
The "Sentiment Distribution by Genre" box plot reveals that R&B and blues songs have the most positive lyrics on average, with the highest sentiment around 18%. Meanwhile, hip-hop has the least positive lyrics on average, with a sentiment of 6%. Other genres like rock, pop, and electronic rank significantly higher than hip hop. These observations could reflect R&B's often love-filled lyrics and hip hop's rebellious profanity.



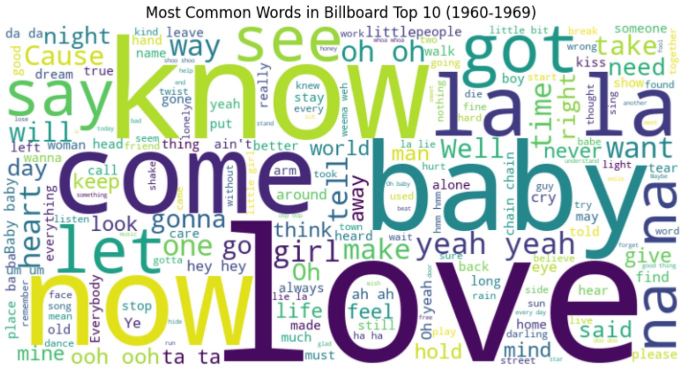


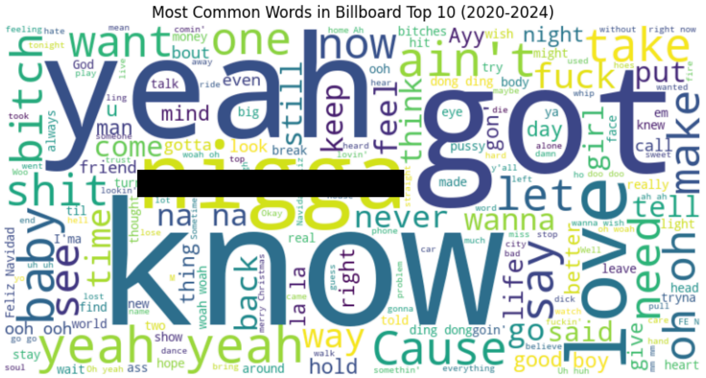
```
Average Sentiment by Genre:
main_genre
r&b                 0.187004
blues               0.178375
jazz                0.164553
easy listening      0.159470
electronic          0.147224
rock                0.143224
classical           0.137140
pop                 0.129641
folk                0.126372
country             0.111771
latin               0.087895
hip hop             0.068686
```

The "Top Genres Over Time" graph highlights changes in genre prevalence. Blues initially dominated the charts but quickly fell into irrelevance by the 1980s. Rock's and R&B’s dominance in the 1960s–1980s was then followed by the rise of hip hop, country, and pop from the 1990s onward. Now, hip-hop dominates the charts in the 2000s, with pop in close second. The rise of hip hop may be contributing to the decline in lyrical sentiment, given its lower average positivity.



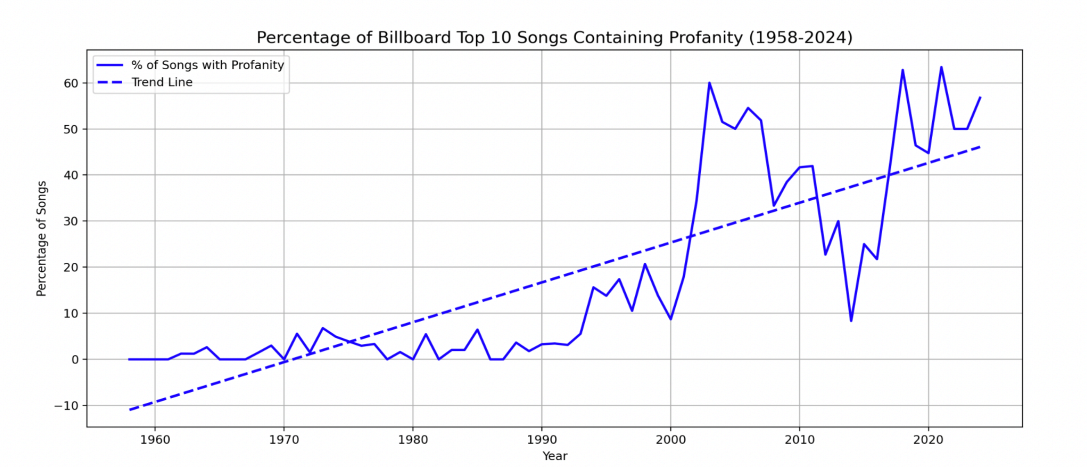
To better understand how lyrics have changed over time, word clouds for every decade were used to identify patterns visible to the human eye. Looking at these clouds, we see an increase in the prevalence of profanity in recent years. (All years: http://bit.ly/43WY03m, also attached)





Songs in the 1900s were mostly about love, with words like “love” and “baby” dominating the charts. By the 2000s, this shifted to more materialistic and colloquial language with words like “know” and “got.” Profanity and slang also became more common, with frequent use of words like “don’,” “gotta,” “bout,” and “ima.”

This increase in profanity is also visualized in the graph below, showing how the percentage of Billboard top 10 songs with swear words has increased significantly in the past few years. This further reinforces the idea that the rise of pop culture and genres like hip hop has changed what kind of lyrics are trendy today.



The decline in lyrical positivity, the rise of hip-hop, and the increasing prevalence of profanity all suggest that mainstream music has become more unfiltered and raw than before. This shows how the qualities that define trending songs shift with changes in public attitudes.

During my research, I also discovered how difficult it is to obtain key music industry statistics about music consumption like my analysis of song lyrics. Almost all of my data had to be sourced via web scraping, as there are few viable alternatives. This is not a coincidence. Today, music data is largely collected by massive streaming services such as Spotify. However, Just this year, Spotify abruptly shut down public access to many of its API’s most important features to maintain its industry power. This gives streaming services a monopoly on music data, posing a serious threat to independent research and transparency in the music industry—limiting the public’s understanding of music’s cultural impact.

Another major challenge I learned about was genre classification. After using Spotify’s API to find artist genres, I ended up with over 300 genres and subgenres that had to be simplified for visualization. Mapping out genre trees is incredibly complex for computers and would have been an entire project on its own (if I had not found a site that already did this).

Future Work

This project’s findings reveal how lyrics define what's on the charts but lyrics alone are only a small slice of the pie. The next step for this project would be to incorporate other musical characteristics like tempo, energy, and danceability - features that Spotify's API once provided before discontinuing public access. With these additional variables, we could employ K-means clustering to group songs by their sonic similarities. Such an analysis could uncover patterns within subgenres and map them into broader genre trees—which I realized I needed during my current analysis. Mapping out songs by their similarities can also provide artists with data driven insights on the sonic qualities that resonate with their audiences.

Citations:

Chosic. (2024, July 26). *Music Genres Explorer: 6000+ Music Genres & Types - Chosic*. https://www.chosic.com/list-of-music-genres/

Utdata. (n.d.). rwd-billboard-data/data-out/hot-100-current.csv at main · utdata/rwd-billboard-data [Dataset]. In *GitHub*. https://github.com/utdata/rwd-billboard-data/blob/main/data-out/hot-100-current.csv

Web API | Spotify for Developers. (n.d.). [Dataset]. https://developer.spotify.com/documentation/web-api

Genius API Documentation. (n.d.). [Dataset]. https://docs.genius.com/

Spotipy-Dev. (n.d.). GitHub - spotipy-dev/spotipy: A light weight Python library for the Spotify Web API [Dataset]. In *GitHub*. https://github.com/spotipy-dev/spotipy

Johnwmillr. (n.d.). GitHub - johnwmillr/LyricsGenius: Download song lyrics and metadata from Genius.com 🎶🎤 [Dataset]. In *GitHub*. https://github.com/johnwmillr/LyricsGenius