Harmonizing Data: A Decade in Music - Comprehensive Analysis of Musical Features and Lyrics

Madhvi Malhotra

Executive Summary

This report presents an in-depth analysis of popular music trends over the past decade, combining two key areas of study: the musical features of songs by Billboard's greatest artists and the lyrical content of top 100 Billboard songs over the last 10 years. By integrating these two components, the report offers a holistic understanding of the evolution and characteristics of popular music, both in terms of its composition and thematic elements.

Background and Objective

In the rapidly evolving landscape of the music industry, understanding shifts in musical styles and lyrical themes is vital. This study aims to dissect a decade's worth of popular music, revealing key trends and underlying patterns in both musical composition and lyrical content. It provides a window into how global events, technological advancements, and evolving societal norms have influenced these aspects of musical expression. This dual analysis involves separate evaluations of quantitative (musical features) and qualitative (lyrical content) elements, aiming to uncover the relationship between a song's musicality and its thematic expressions, providing insights into the changing dynamics of popular music.

Methodology and Insights

Part 1: Analysis of Musical Features

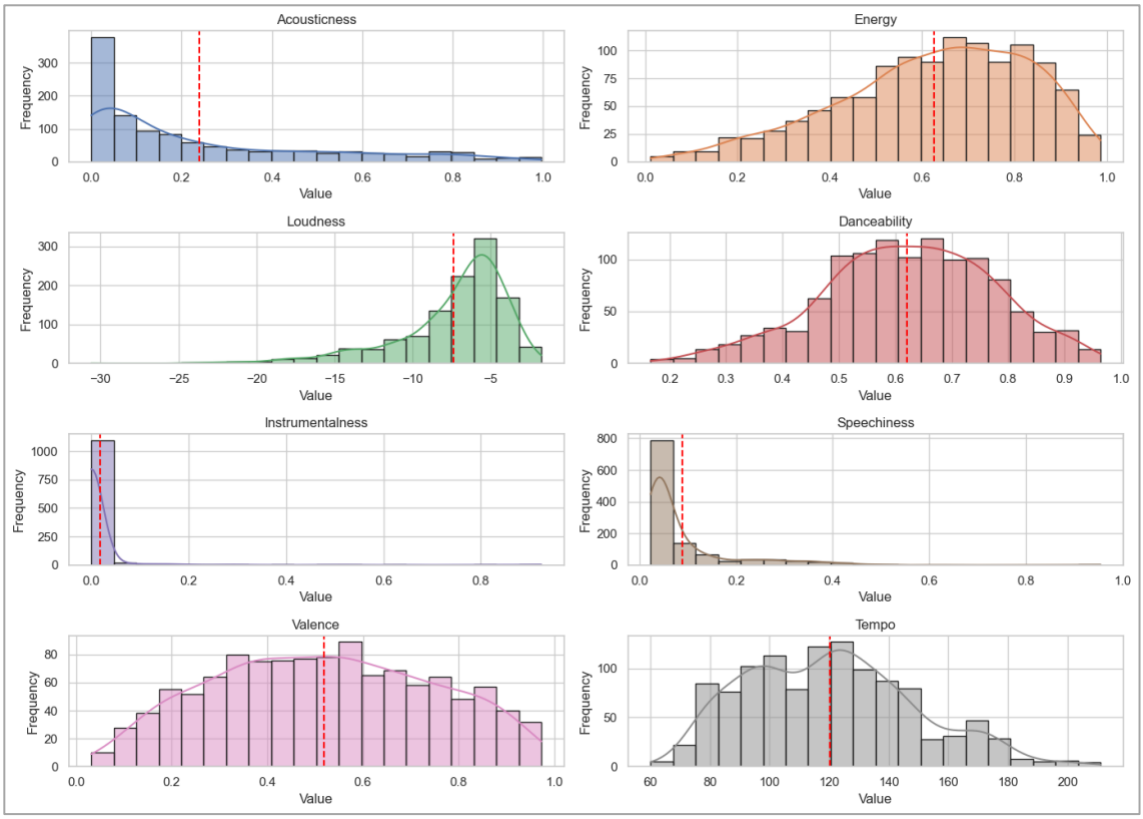
Data Sources: The project utilized Billboard charts for identifying "*125 Greatest of All Time Artists"* artists. Discogs was used to fetch Top Albums for these 125 Artists and the Genre for these Albums. Additionally, Spotify's API was leveraged to fetch detailed musical features for each track in these Albums. Then the dataset was reduced to include only top 10 popular songs based on the popularity index from Spotify.

Data Processing:

- **Cleaning and Standardization:** The data from different sources was cleaned and duplicates were removed to ensure consistency. This included standardizing the format of artist names and song titles and normalizing numerical features like tempo and loudness.
- **Feature Selection and Preparation:** The key audio features like Acousticness, Energy, Loudness, Danceability, Instrumentalness, Speechiness, Valence and Tempo for each song was selected for clustering analysis. These features were standardized to ensure comparability across different songs.

Figure 1: Here's a visual representation of the distribution of various audio features across all albums in our dataset. The histograms displayed here showcase the diversity of audio features found in a decade's worth of music. These features are key to understanding what makes each song unique. The density plots provide a smooth curve representing the distribution of each feature, while the red dashed lines mark the mean, serving as a quick reference to the typical values within each feature's range.

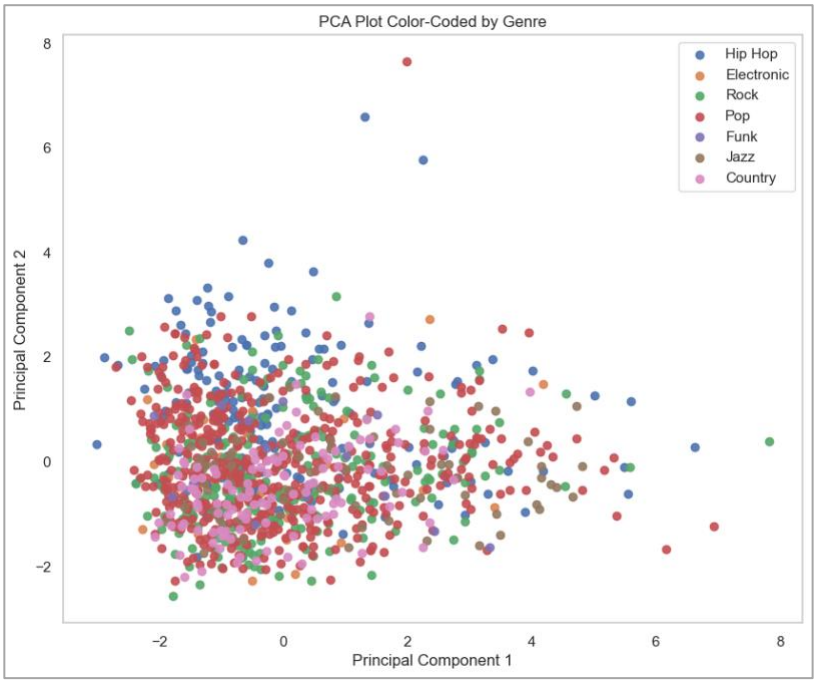
By examining these distributions, we gain insights into the common audio profiles of popular music tracks and how they contribute to the listener's experience



Analytical Techniques: To understand the evolving landscape of music over the past decade, K-means clustering was employed, a robust algorithm ideal for grouping data based on shared characteristics. This unsupervised machine learning technique iteratively assigns each album to the nearest cluster, based on the similarity of its features to the cluster's centroid. In this section, we explore how this technique has helped us identify unique patterns and groupings in music albums, based on a diverse range of audio features.

Dimensionality Reduction/ Principal Component Analysis (PCA) of Song Features: PCA was applied to reduce the complexity of the data, facilitating a clear visualization of clusters in a two-dimensional space

Figure 2: The PCA plot displayed here visually represents the multifaceted nature of musical tracks based on their audio features. By reducing the complexity of our dataset to two principal components, we can observe the variance and relationships between songs in a simplified two-dimensional space.



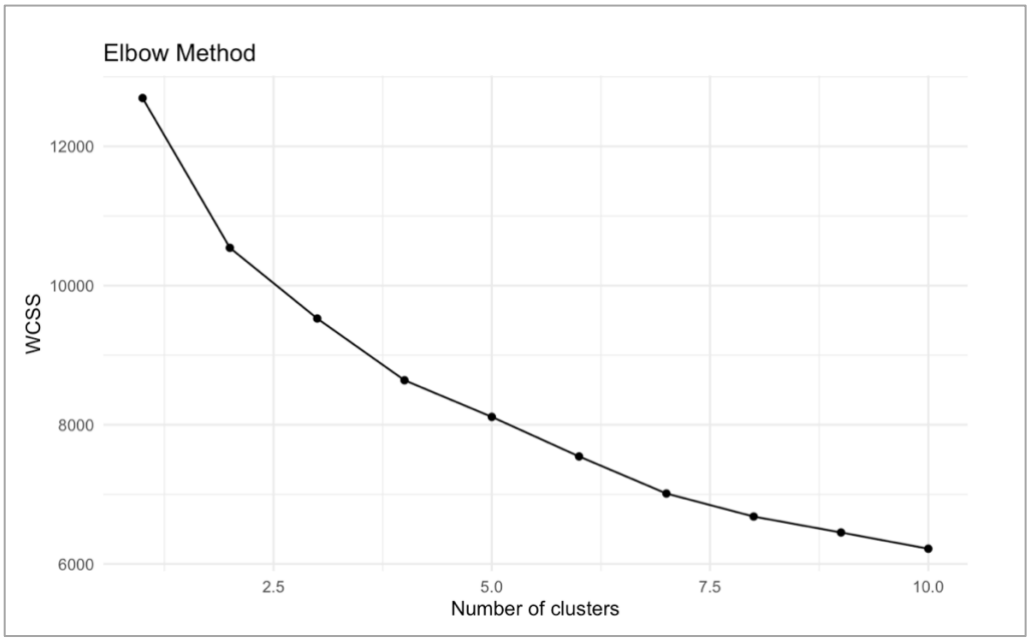
Each point on the plot corresponds to a song, color-coded by its genre, allowing us to discern how different genres of music are distributed and to identify any inherent clustering. For instance, we can see how certain genres like Rock (red dots) and Pop (pink dots) share a similar space, suggesting commonalities in their audio characteristics. The plot also reflects the diversity within genres, as evidenced by the wide spread of

points within each color, suggesting a range of sub-genres and stylistic variations. Overall, the PCA plot illustrates the fluidity and dynamic nature of modern music, with genres both merging and maintaining their unique identities, creating a complex and interconnected musical landscape. It's a testament to the diversity of music and its ability to transcend simple categorization, reflecting a complex tapestry of rhythm, melody, and emotion.

Optimizing Cluster Number: The Elbow Method and Silhouette Score were used to determine the optimal number of clusters.

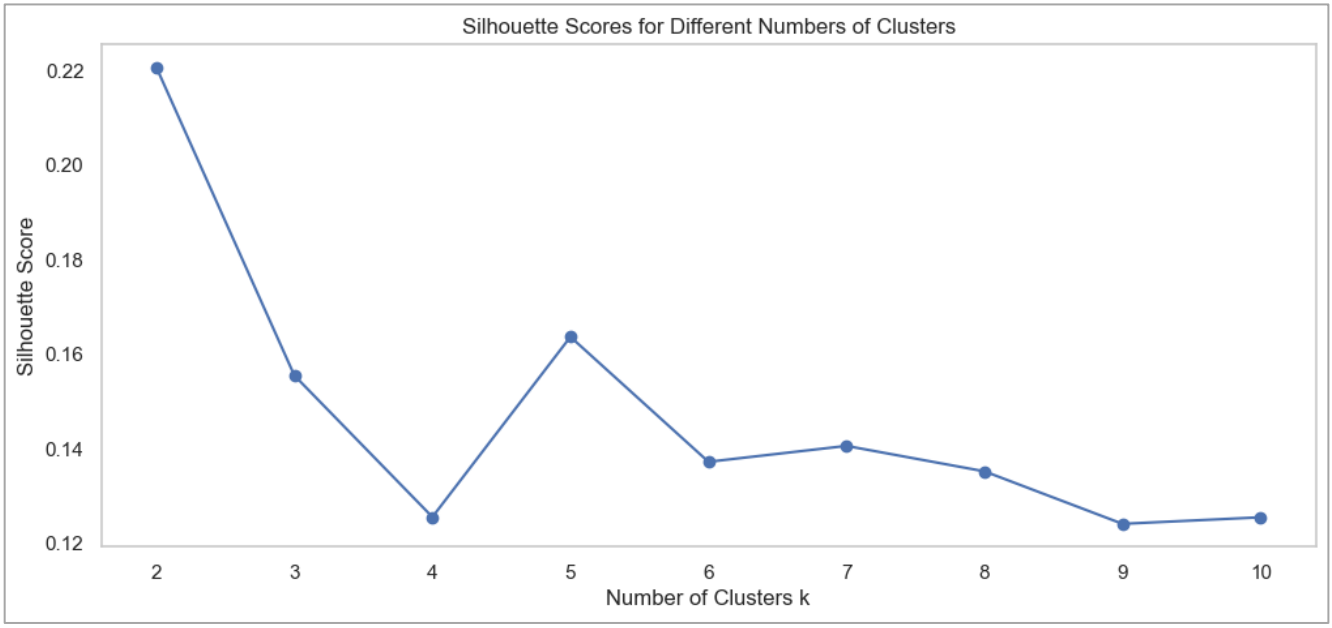
In our clustering analysis, determining the optimal number of clusters is crucial for meaningful segmentation. The elbow method is a visual tool used to estimate the number of clusters in which the variance Within the clusters (WCSS) starts to diminish. In the plot, we look for a point where the decrease in WCSS slows down significantly, indicating a balance between the number of clusters and the sum of squared distances to the cluster centroids.

Figure 3: As we increase the number of clusters from 1 to 10, the WCSS decreases sharply at first and then begins to level off. This 'elbow' point suggests an optimal number of clusters beyond which adding more clusters does not contribute to a significant decrease in WCSS, hence providing diminishing returns.



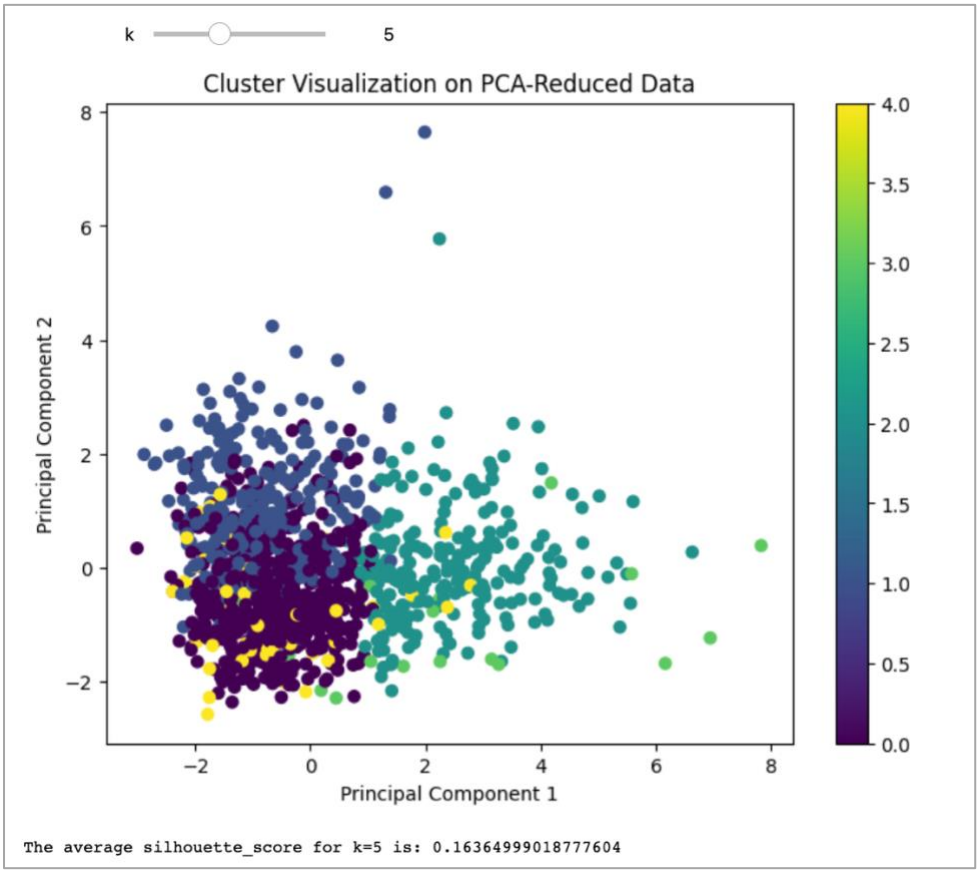
Silhouette scores are a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). High silhouette scores indicate clear, well-separated clusters, while lower scores suggest overlapping clusters.

Figure 4: Shows silhouette scores for different numbers of clusters (k), although the highest average silhouette score occurs at k = 2, the score experiences a notable increase at k = 5. The Elbow plot corroborates the silhouette analysis, together justifying the selection of k = 5 for clustering the music features.



The clustering result at k = 5, visualized in the PCA-reduced space (***Figure 5 below***), shows the data points grouped into five distinct clusters. This visualization, combined with a moderate silhouette score of 0.164, indicates a reasonable structure within the data, capturing meaningful patterns in music features that could correspond to different genres or styles.

Figure 5: Clusters at k = 5



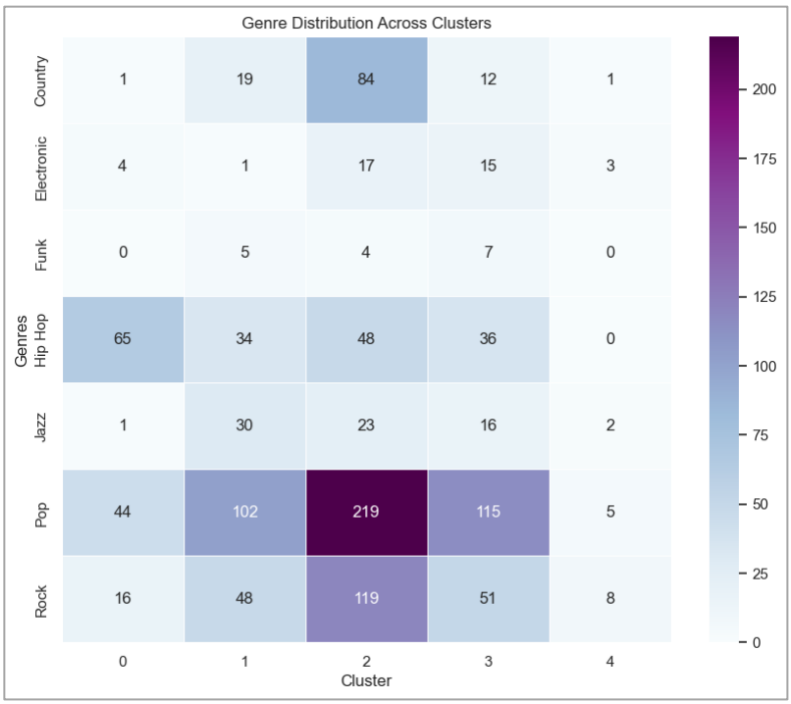
We can also observe from above how ***Figure 5 and Figure 2*** are quite similar, showing the features could provide information about the genre. Now to see this we can use crosstabulation of genres and clusters

Genre Distribution Across Clusters: Building on the PCA-reduced cluster visualization in Figure 5, we further analyzed the relationship between genres and the clusters by employing a crosstabulation approach. This method quantifies the number of songs from each genre within each cluster, providing a numerical basis for the patterns observed in the PCA plot.

Figure 6 presents a heatmap based on this crosstabulation, offering an immediate visual impression of the distribution and density of genres across the clusters. The heatmap uses a diverging color scheme that is colorblind-friendly, ensuring accessibility while effectively highlighting variations in song count across the clusters. Darker colors

represent higher concentrations of songs within a cluster for a given genre, while lighter colors indicate fewer songs.

Figure 6: Heatmap



The heatmap reveals several key insights into our dataset:

- Hip Hop shows a significant presence in Cluster 0, suggesting that the defining audio features of Hip Hop are captured within this cluster.
- The distribution of Country music is predominantly centered in Cluster 2, alongside Pop, indicating shared musical characteristics between these genres that are captured by the clustering algorithm
- Rock and Electronic genres are more dispersed across clusters, suggesting a wider variety in their audio features or a blend of styles within these genres.

Reflection on Data Spread and Recommendations for Further Research

While the heatmap in Figure 6 provides valuable insights into the distribution of genres across the clusters, it also highlights an important aspect of the dataset's composition. Notably, the spread of data across genres is not consistent, with certain genres like Pop and Hip Hop showing a higher concentration within specific clusters, whereas genres like Jazz and Funk are more sparsely represented. This inconsistency could potentially skew

the clustering analysis, as the algorithm may be disproportionately influenced by the genres with more data points.

The current dataset, derived from Billboard charts and Spotify's API, gives us a snapshot of popular music by popular artists but may not encompass the full diversity within each genre. For a more balanced and comprehensive understanding of the musical landscape, future analyses would benefit from expanding the dataset. Including a broader range of artists, albums, and tracks that cover more subgenres and less mainstream music could provide a richer and more nuanced clustering result. Moreover, the inclusion of additional data may also help in addressing any potential biases in the dataset. For instance, the popularity index from Spotify is a dynamic and platform-specific measure that might not fully represent the global or historical popularity and influence of certain songs or genres.

In light of these observations, I would recommend extending the dataset for future analyses to include a wider array of genre which might lead to more robust clustering outcomes.

Part 2: Lyrical Landscapes: Unveiling Trends Through Text Analysis

Data Sources: This analysis draws from a decade of musical history, specifically the Year-end Top 100 Songs for each year as listed by Billboard. Lyrics for these songs were meticulously extracted using the *"Genius"* API, resulting in a rich dataset ripe for exploring lyrical trends.

Data Processing:

- **Lyric Cleaning and Preprocessing:** Initial steps involved the removal of extraneous text and non-lyrical elements. The text was then standardized to ensure a clean dataset, allowing for more accurate analysis. To ensure uniformity and focus, songs with non-English lyrics were excluded.

- **Text Normalization:** The lyrics underwent further processing which included conversion to lowercase, tokenization, and the removal of stopwords, setting the stage for in-depth textual analysis

Analytical Techniques

Word Frequency Analysis:

A cornerstone of our analysis, word frequency techniques unveiled the most commonly used words and phrases across the dataset, providing insights into recurring lyrical themes.

Overview:

A summary statistics table was constructed to present a high-level view of the dataset's composition over the years. It included metrics such as the total number of songs analyzed, the number of unique words, the average word count per song, and the most common word for each year. The table below shows the summary statistics of song lyrics, including total songs analyzed, unique word counts, average word count per song, and the most common word

Figure 7: Summary Statistics

Year	Total Number of Songs	Total Number of Unique Words	Average Word Count per Song	Most Common Word
2013	100	3175	196	love
2014	100	3227	206	like
2015	100	3084	204	like
2016	99	3070	209	know
2017	99	3862	224	like
2018	97	3695	234	like
2019	99	3766	210	like
2020	100	3611	210	yeah
2021	97	3699	233	like
2022	92	3030	189	like
2023	92	3078	191	like

We can see from the Table (*Figure 7)* above that consistency in the number of songs per year provides a stable foundation for comparative analysis. Notably, there is a fluctuation in the total number of unique words used annually, with a peak in 2017, suggesting a diversification in lyrical content. The recurrence of certain words as the most common each year, such as "like," "love," "know," and "yeah," underscores their enduring presence in popular music themes. ***Figure 8*** below show the word cloud for most common words across these 10 years, these top words, tells the story of a decade's hopes, dreams, fears, and joys, all echoed in the lyrics we sing.

Figure 8: Word cloud

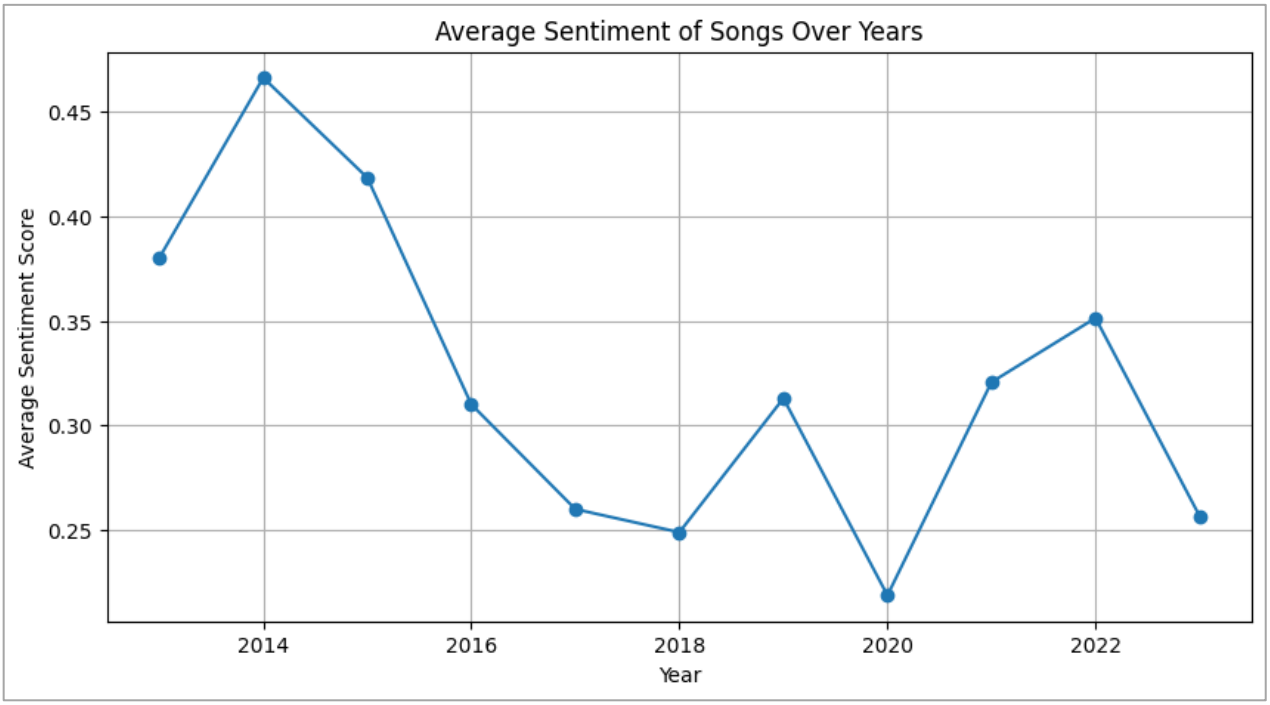
Sentiment Analysis

Natural language processing tools were used to analyze the sentiment of the lyrics, allowing for an examination of the emotional tone and shifts over time. Sentiment scores typically range from negative to positive, reflecting a spectrum of emotions that can be conveyed through lyrics. A higher average score suggests a prevalence of positive sentiments such as joy and happiness, while a lower score might indicate more negative emotions like sadness or anger.

Figure 9: The sentiment analysis graph presented here offers an insightful look into the emotional undertones of song lyrics over a span of years. By applying natural language

processing techniques, we've quantified the average sentiment score of songs for each year, revealing how the emotional content of music has fluctuated over time.

Figure 9: Sentiment Analysis



The song ***"Mirrors" by Justin Timberlake from the year 2013*** has the highest sentiment score, which suggests that its lyrics have a very positive tone according to the Sentiment Intensity Analyzer. Conversely, ***"All Time Low" by Jon Bellion from the year 2017*** has the lowest sentiment score, indicating that its lyrics are perceived as having a very negative tone.

The sentiment analysis chart tells an interesting story about the emotions in popular music over the years. For instance, in 2014, the songs had the happiest lyrics, suggesting that was a cheerful year in music. But by 2018, the mood had dipped, showing that songs had taken on a sadder tone. The chart also points out that 2020 had some of the lowest mood scores, which isn't too surprising given that it was a tough year worldwide because of the pandemic. This drop in mood could reflect the challenging times and how they influenced the themes and emotions expressed in the songs of that year. It's a clear example of how real-world events can directly affect the music we listen to and the overall vibe of the times.

Conclusion

To conclude it's clear that the last ten years have been a dynamic period in the music industry. The analysis, spanning from the intricacies of musical features to the nuances of lyrical content, has revealed trends and patterns that characterize this era. The insights gleaned from this study have several implications for the music industry. For artists and producers, understanding these trends can guide the creative process, helping them tap into current listener preferences or explore new directions. For music platforms and marketers, this analysis provides valuable data to tailor recommendations and marketing strategies, aligning with the listener tastes.

As we look to the future, it's evident that embracing diversity in musical expression and remaining attuned to societal shifts will be key to resonating with audiences.

Bibliography (References)

- Billboard “125 Greatest of All Time Artists”: https://www.billboard.com/charts/greatest-of-all-time-artists/
- Billboard “Year-end charts Hot 100 Songs”: https://www.billboard.com/charts/year-end/2023/hot-100-songs/
- Spotify API: https://developer.spotify.com/documentation/web-api
- Discogs API: https://www.discogs.com/developers
- Genius API: https://docs.genius.com/
- Billboard.py API: billboard.py is a Python API for accessing music charts from Billboard.com https://github.com/guoguo12/billboard-charts
- Lyrics Genius: a Python client for the Genius.com API used to extract Lyrics https://github.com/johnwmillr/LyricsGenius