A Project Report on

THE POWER OF LYRICS

Analyzing the relationship between Lyrical Complexity and Song Popularity

Sarah Dias Barreto
Neha Anil Chede
Angel Reji George

Fall 2024 | Indiana University

Abstract

The proliferation of digital music platforms has revolutionized access to song lyrics and metadata, paving the way for innovative music analysis. This project investigates the relationship between lyrical complexity and song popularity, analyzing a dataset of 60,000 Spotify tracks. Using natural language processing techniques, we measure lyrical complexity through readability tools such as the Flesch-Kincaid Index and Gunning Fog Index. Through visualizations like scatter plots, heatmaps, and word clouds, we identify patterns in lyrical complexity across time periods, and artists and examine its connection to popularity metrics such as stream counts. The research seeks to determine whether audiences favor simpler or more complex lyrics and how these preferences differ by region or evolve over time. Additionally, word cloud visualizations reveal recurring themes in the lyrics of popular songs. This study provides fresh insights into the role of lyrical content in shaping modern music consumption and its impact on a song's success.

Keywords: Spotify, Lyrical complexity, Natural Language Processing, Flesch-Kincaid Index, Gunning Fog Index, Music Popularity, Genre Analysis, Word cloud, Song Popularity Correlation, Heatmap, Sentiment analysis

Table of Contents

1. Introduction

1.1 Motivation

The exponential growth of digital music platforms like Spotify has fundamentally transformed music consumption, offering unprecedented access to vast song libraries and listener data. With record labels investing over $2 billion annually to discover and promote talent, the demand for data-driven insights into what drives song popularity has never been higher. Within the fields of hit song science and Music Information Retrieval (MIR), researchers have explored various factors influencing song success, including acoustic features, metadata, and artist popularity.

For example, Herremans[1] used classifiers like Logistic Regression and Naive Bayes to predict whether dance tracks would enter the top 10 charts, analysing a combination of acoustic and temporal features with promising results. Similarly, Interiano[2] applied random forest models to over 500,000 tracks, incorporating a 'superstar' attribute to account for artists' prior chart success, yielding robust predictions. Research on Turkish music[3] further demonstrated that combining audio and lyrical features with SVMs could achieve classification accuracy exceeding 99%, though the limited diversity of performance metrics raised concerns about the reliability of such results.

While these studies emphasize the predictive power of acoustic features, the role of lyrics in determining song popularity remains underexplored. Cognitive research suggests that humans are drawn to complex linguistic structures, which can provide intellectual stimulation and emotional resonance[4]. This raises an important question: could lyrical complexity play a significant role in influencing a song's success?

This study aims to address this gap by investigating the relationship between lyrical complexity and song popularity on Spotify. Using readability metrics like the Flesch-Kincaid Grade Level and Gunning Fog Index, we will analyse how the semantic richness of lyrics correlates with popularity metrics, such as stream counts. By focusing on lyrical content, our research seeks to complement

existing work on acoustic and structural features, offering a more holistic understanding of the factors driving musical engagement. These insights have the potential to inform musicians and record labels about crafting lyrics that resonate deeply with audiences, ultimately enhancing a song's commercial appeal.

This project introduces a novel intersection of music analytics and cognitive linguistics, hypothesizing that emotionally engaging or intellectually stimulating lyrics may foster higher listener engagement and repeat plays. By filling this critical research gap, we aim to contribute to the growing body of knowledge on modern music consumption and the elements that shape a song's success.

1.2 Existing Visualizations

Two influential papers, *"Veritas AI Tuning into Trends: Machine Learning Models for Song Popularity Prediction on Spotify"[5]* and *"Music Popularity Prediction Through Data Analysis of Music's Characteristics"[6]* , provide diverse visual analyses that contribute to understanding song popularity metrics and modeling. These visualizations play a pivotal role in highlighting the challenges and insights derived from their respective datasets and methodologies.
In the *"Veritas AI"* paper, a correlation matrix was used to illustrate the relationships between various song attributes and their popularity scores. The visualization demonstrated low correlations across most features, indicating that no single feature could robustly predict popularity. For example, acoustic and temporal attributes such as Danceability, Acousticness, Energy, and Tempo showed negligible relationships with popularity, necessitating the exploration of more complex models and feature combinations to capture subtle patterns.

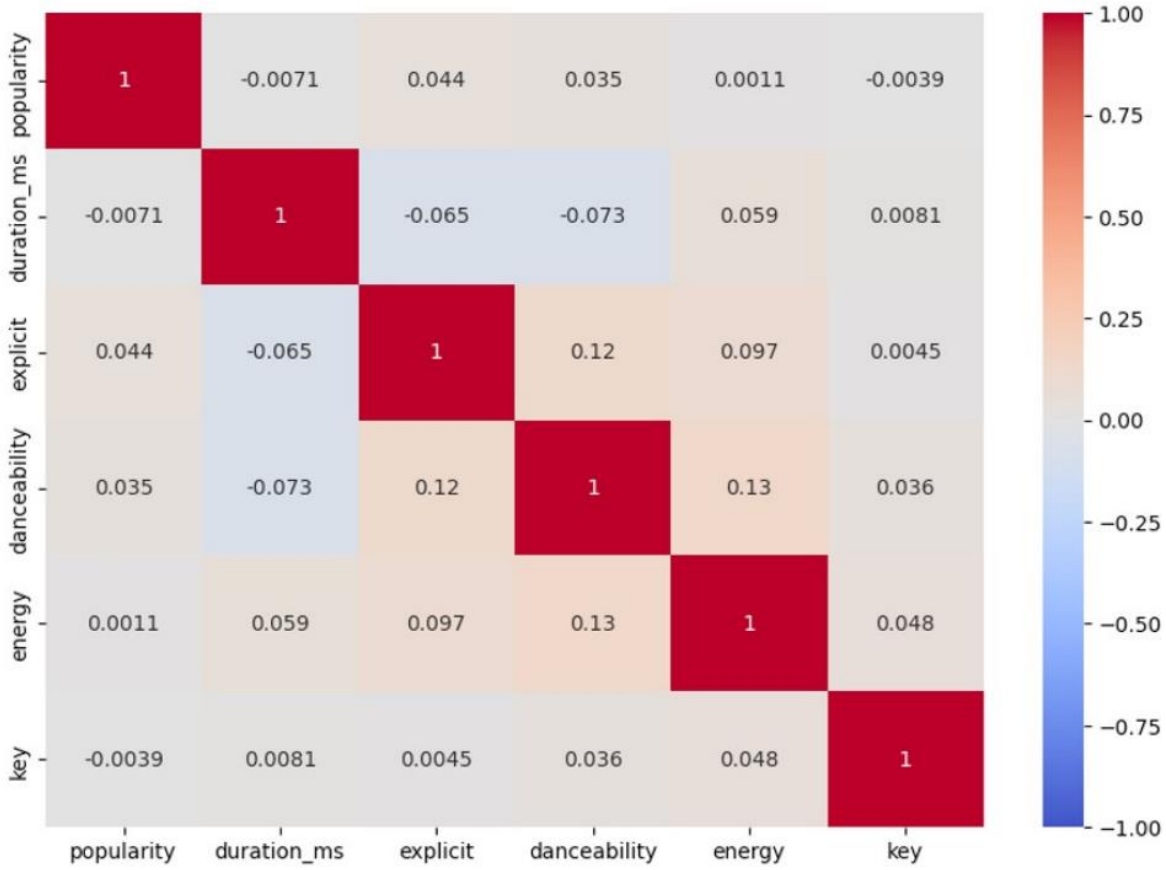

Fig 1.1 A correlation matrix displaying a very low correlation between any of our features and the target popularity. [5]

A scatter plot comparing predicted and actual popularity values showed that regression models, including MLP, failed to adequately capture the target variable's distribution. Predicted values were largely clustered within a narrow range (15–45), reflecting the limitations of the model and possibly the dataset. This visualization underscores the need for further refinement of features, model architecture, and possibly reframing the problem to a classification task, as discussed in their solution section.

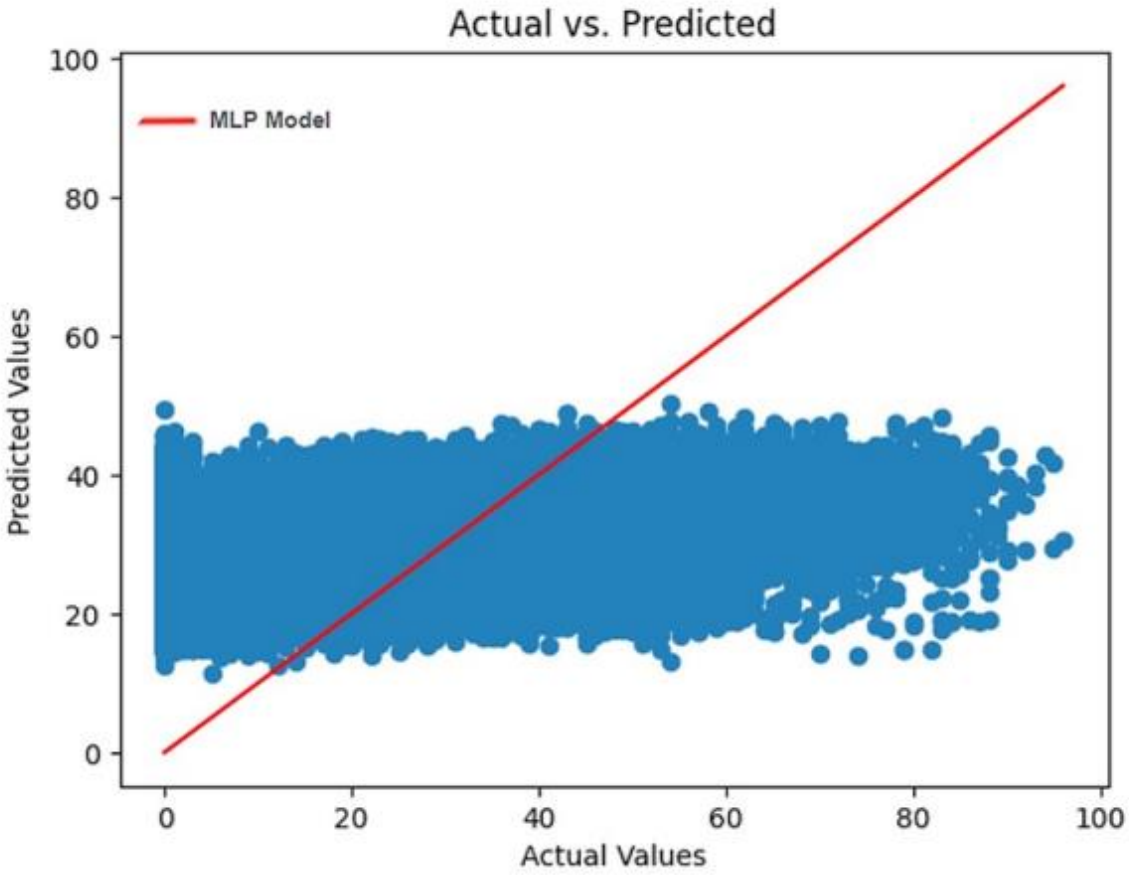

Fig. 1.2. Scatter plot for Predicted vs Actual values [5]

The heatmap in Fig. 1.2. is limited in its effectiveness as a visualization for several reasons. While it provides an overview of correlations between features and song popularity, it oversimplifies relationships by focusing only on linear correlations, ignoring potential nonlinear or interactive effects among features. Additionally, the visualization lacks actionable insights or contextual annotations to explain the implications of these correlations, such as why Loudness is strongly correlated with popularity or how features like Danceability and Valence interact. The dense presentation of numeric values without highlighting key takeaways can overwhelm the reader, making it difficult to discern meaningful patterns at a glance. Finally, the heatmap fails to engage with temporal dynamics or genre-specific trends, missing opportunities to add depth to the analysis.

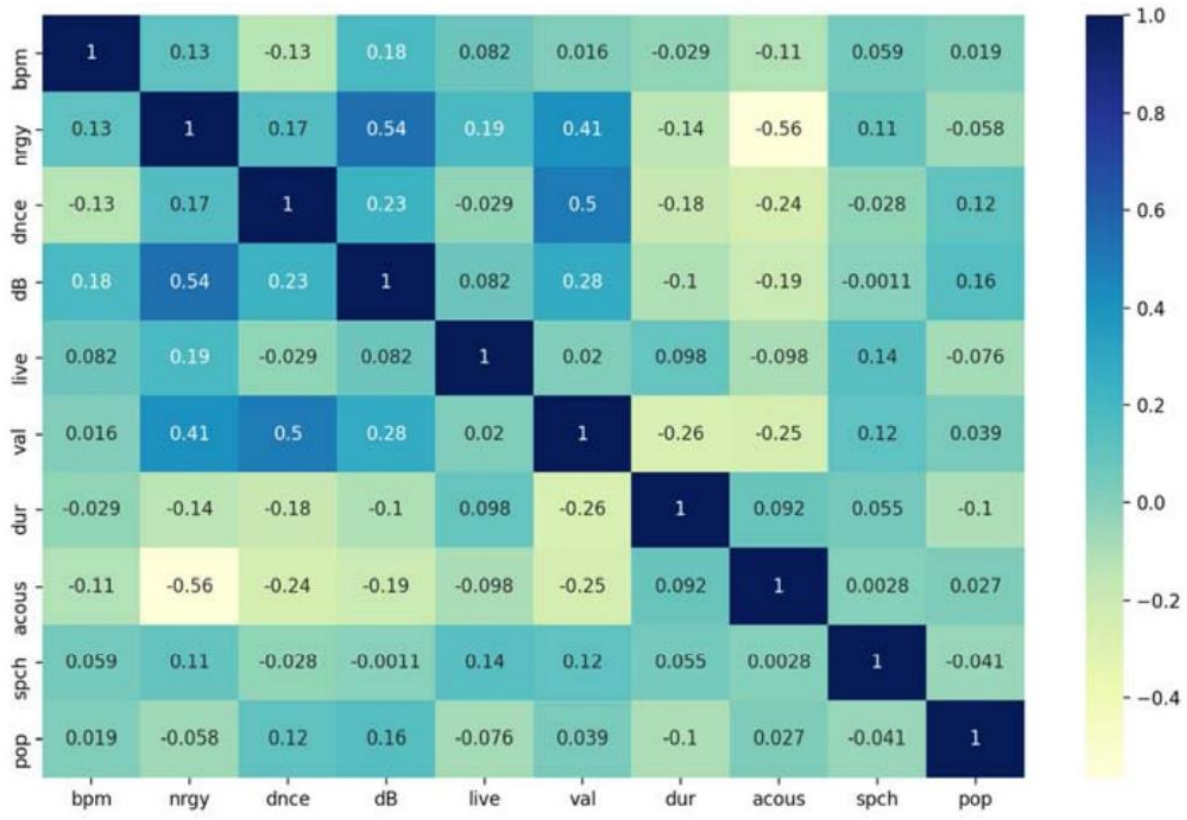

Fig. 1.3. Confusion matrix for SVM model with unbalanced class weights, trained on the tf-idf feature set[5]

Each visualization is accompanied by a brief explanation, which helps in understanding the context and significance of the results. However, more detailed annotations or callouts within the graphs could highlight key findings or anomalies, making the visualizations more informative. While static visualizations are straightforward and easy to interpret, incorporating interactive elements in digital versions of the report could enhance user engagement and allow for deeper exploration of the data.

Both studies could benefit from more engaging and user-friendly visualizations. For example, annotated scatter plots could better highlight anomalies or patterns, while interactive digital versions could allow deeper exploration of correlations or model predictions. Preprocessing

steps, such as transformations or class balancing, could also be visualized to enhance transparency and methodological rigor.

Additionally, the study's visualizations could benefit from a more dynamic representation of temporal trends, such as how lyrical complexity correlates with popularity across decades. Line or bar charts comparing these metrics over time could reveal valuable insights into shifting preferences. Lastly, visuals explaining preprocessing steps, such as dataset transformations or content filtering, would improve transparency and strengthen the methodological narrative Enhancing these areas could make the visualizations more comprehensive and engaging, allowing them to convey insights more effectively to varied audiences.

2. Data and Methods

2.1 Data Description

For this project, we will rely on two main data sources:

- Spotify's Worldwide Daily Song Ranking Dataset[7]:

This dataset contains the daily rankings of approximately 3.4 million songs across various countries, specifically focusing on the daily ranking of the 200 most listened to songs in 53 countries from 2017 and 2018 by Spotify users. It includes over 3 million rows, encompassing 6,629 artists and 19,923 songs, totaling an impressive 178 billion stream counts. The data spans from January 1, 2017, to January 9, 2018, collected from Spotify's regional chart data. This dataset provides crucial metadata, including song popularity metrics such as stream counts, user ratings, and listener engagement over time, allowing for a comprehensive analysis of patterns in song popularity across different geographical areas, offering insights into both local and international music trends. Both well-known and obscure songs are included, providing a diverse perspective on music consumption.

Column Name	Description	Data Type
Position	The daily ranking position of the song	Integer
Track Name	The title of the song	String
Artist	The performer or group of the song	String
Streams	The total number of streams the song received that day	Integer
URL	The Spotify URL link to the song	String (URL)
Date	The date of the ranking entry	Date
Region	The country or region where the ranking applies	String

Table 1.1 Schema for Spotify's Worldwide Daily Song Ranking Dataset

– Song Lyrics Dataset[8]:

This dataset contains around 60,000 song lyrics with fields for artist, song title, lyrics, and hyperlinks to the original text. This allows for in-depth text analysis in the ways in which lyrics impact a song's popularity or impact over time. An extensive summary of the corresponding columns and data types is provided below:

Column	Description	Data Type
Artist	The performer or group of the song	String
Song	The title of the song	String
Link	URL link to the original lyrics source	String (URL)
Text	The full lyrics of the song.	String (Text)

Table 1.2 Schema for Song Lyrics Dataset

We will ensure the dataset is complete by checking for any missing or incomplete data.

2.2 Data Analysis

Our research methodology focuses on three key stages:

1. Preprocessing the Lyrics:
   - Clean the song lyrics by removing stop words, punctuation, and unnecessary symbols
   - Apply tokenization and lemmatization to divide the lyrics into meaningful linguistic components
2. Lyrical Complexity Analysis:
   - Evaluate the readability and complexity of each song's lyrics using tools such as the Flesch-Kincaid Reading Ease and the Gunning Fog Index
   - Calculate lexical diversity, measured by the type-token ratio, to determine the variety of vocabulary employed in each song
3. Popularity Correlation:
   - Extract key popularity metrics from Spotify, including stream counts, likes, and listener retention rates for each track.

- Analyze the relationship between lyrical complexity and song popularity using Pearson's correlation coefficient
- Segment the analysis by region, artist, and release year to explore how complexity impacts popularity across different categories

4. Visualizations:

- Use scatter plots to illustrate the correlation between complexity metrics and popularity scores.
- Generate heatmaps to depict the relationship between lyrical diversity and engagement metrics across various genres and timeframes
- Create word clouds and sentiment maps to visually represent thematic patterns and their influence on a song's success

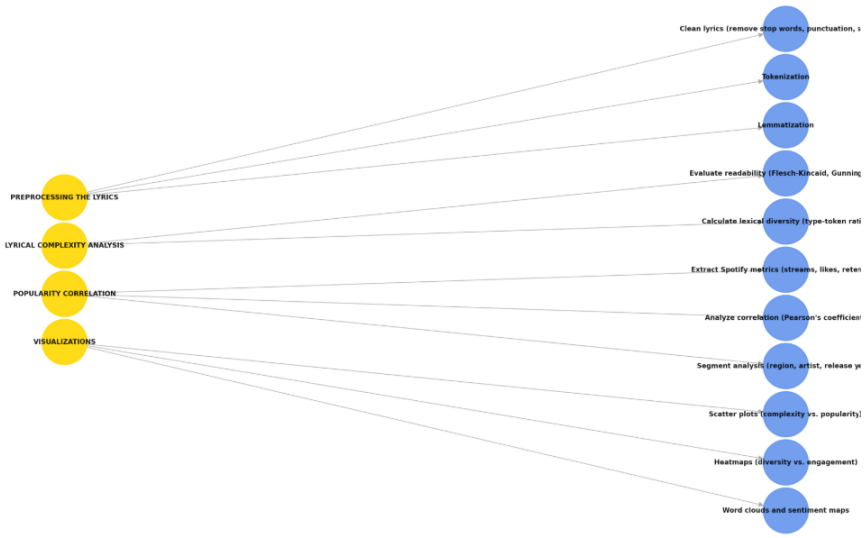

Fig 2.1 Flowchart of Methodology

2.3 Data Exploration

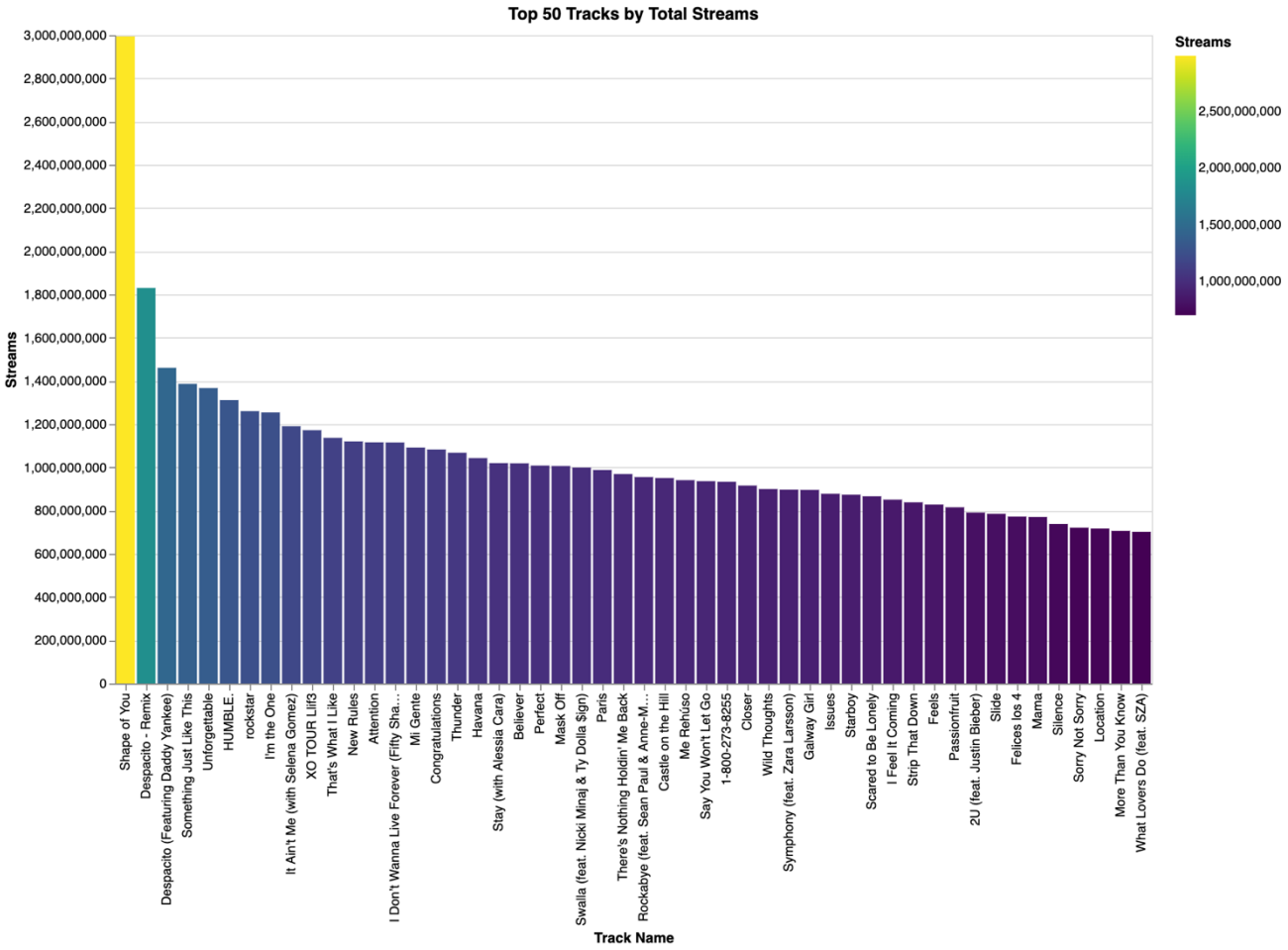

Fig 2.2 Top 50 Tracks by total streams globally

The visualizations presented explore the global popularity of music tracks and artists based on streaming data. *Figure 2.2* showcases the top 50 tracks by total streams, revealing a steep drop-off in streams after the most popular track, "Shape of You." This suggests a significant disparity in streaming numbers between the most and moderately popular tracks, highlighting the immense concentration of listening activity around a few global hits. The gradual decline in bar heights also indicates the "long tail" effect, where many tracks achieve substantial but less dominant popularity.

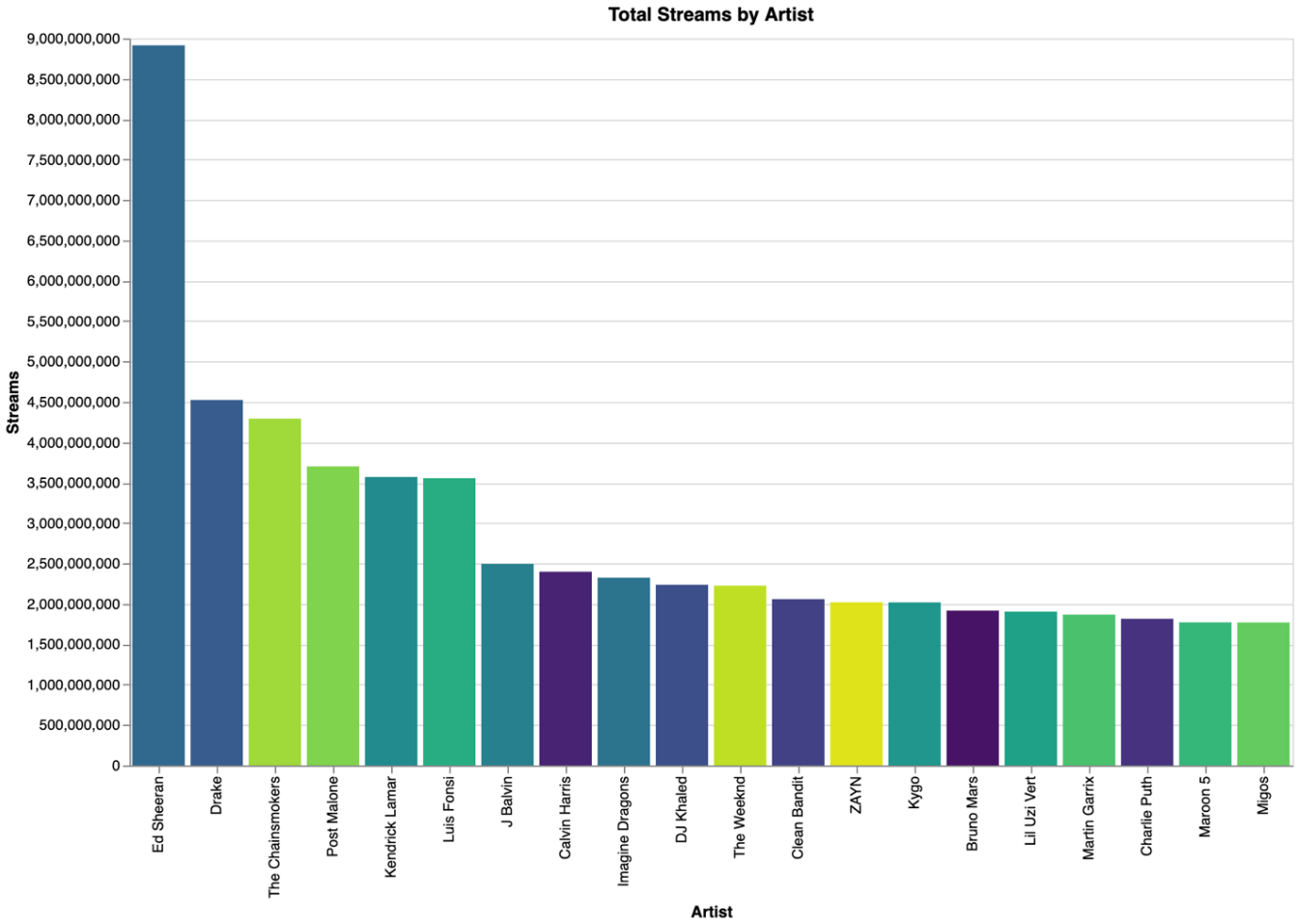

Fig 2.3 Top 20 Tracks by total streams globally

The visualizations depict streaming data at artist levels, providing insights into trends. *Figure 2.3*, which ranks the top 20 artists globally by total streams, highlights Ed Sheeran as a dominant figure, significantly outpacing other artists with over 8.5 billion streams. Other prominent artists include Drake, The Chainsmokers, and Post Malone, who also achieve substantial global presence.

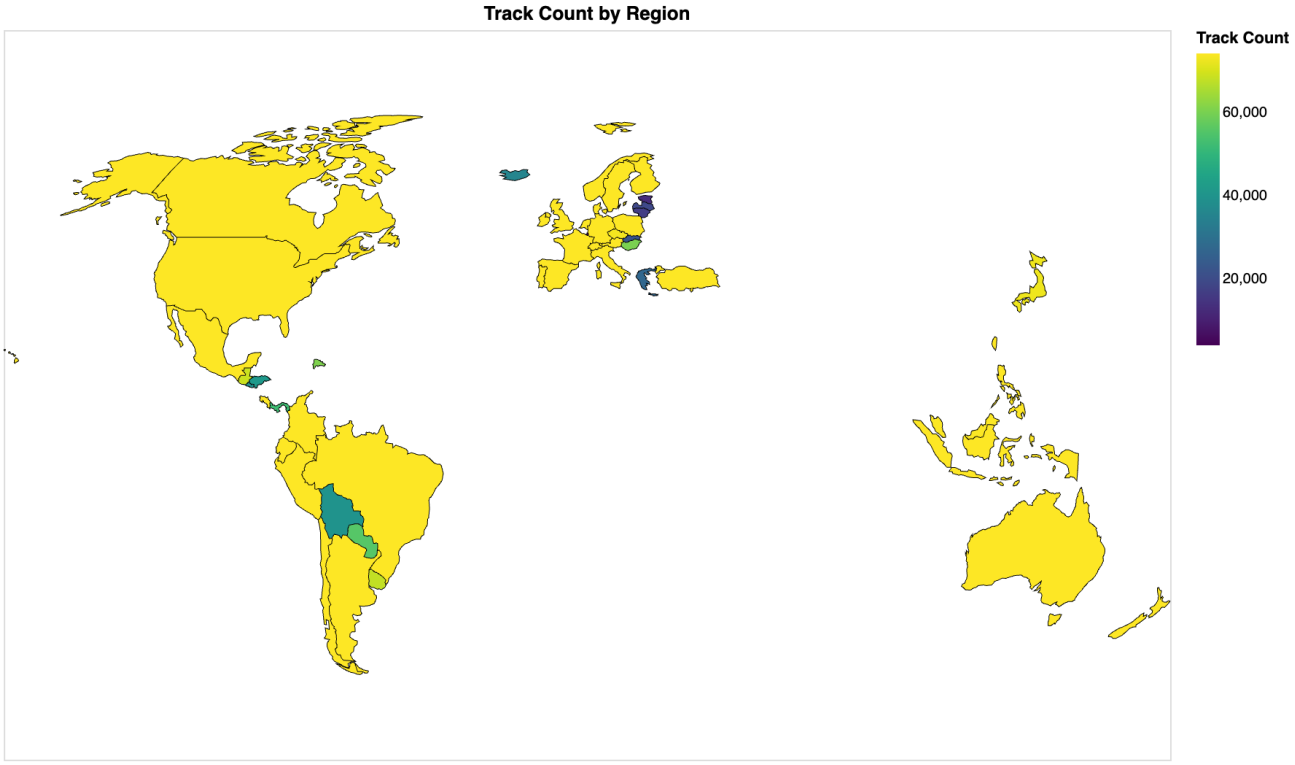

Fig 2.4 Geographical distribution of the streams

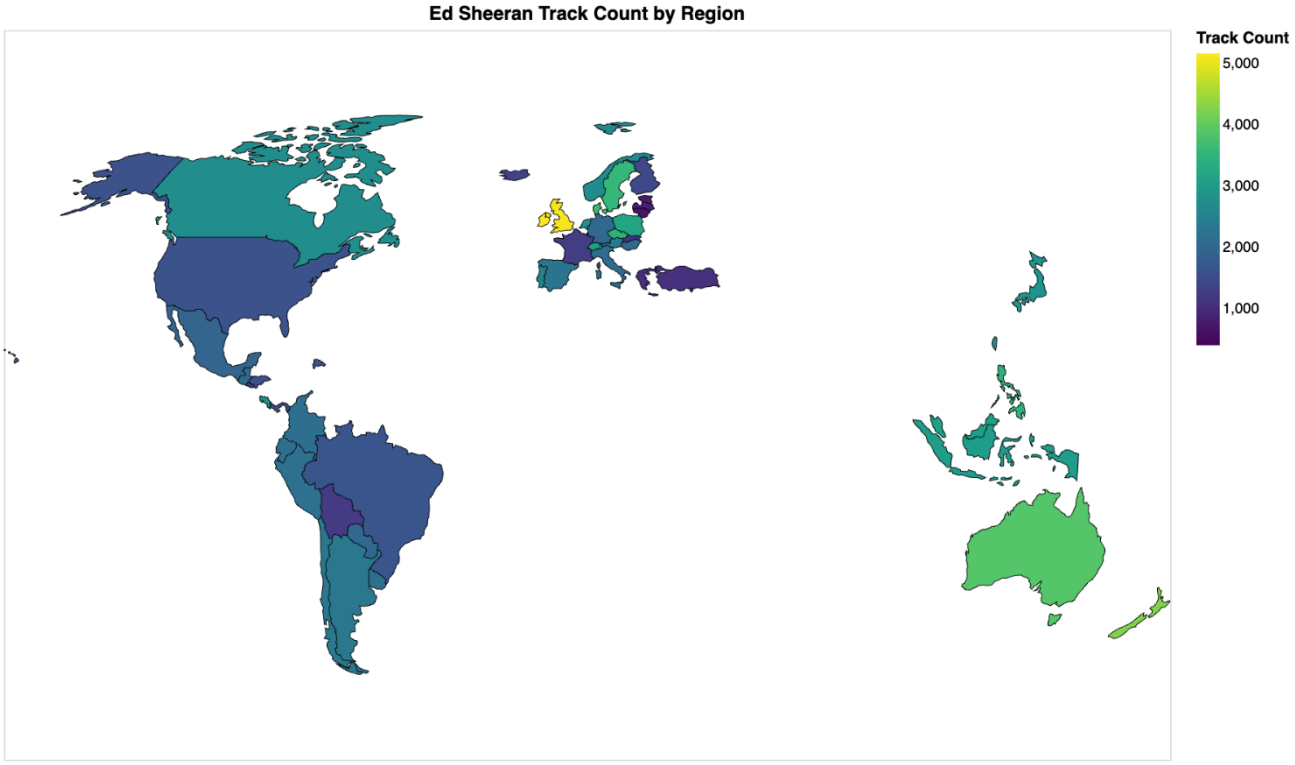

Fig 2.5 Geographical distribution of the for Ed Sheeran, the leading artist

Figures 2.4 and 2.5 delve into the geographical breakdown of streams. *Figure 2.4* maps the overall track count by region, showcasing higher activity in North America, Europe, Australia and parts of South America, suggesting these areas are major contributors to global streaming trends. Figure 2.5 narrows the focus to Ed Sheeran, visualizing his dominance across multiple regions, particularly in Europe and the Americas, where track counts are highest. This emphasizes his widespread popularity and influence.

The bar charts are straightforward, with clear axes and labels, ensuring that viewers can easily interpret the data without confusion. The gradient coloring effectively draws attention to the highest values in each chart, creating a visually appealing and intuitive representation of the data. The choice to represent both tracks and artists allows for a nuanced understanding of the industry, bridging the gap between individual hits and aggregate artist performance. The inclusion of track names and artist names ensures that viewers can connect the data to real-world examples, making the insights tangible and relatable. Using color-coded choropleth maps that provide an intuitive understanding of regional variations, helping to identify hotspots of streaming activity. The consistent use of color intensity to denote higher values enhances interpretability and ensures visual consistency across maps

2.4 Visualization Methods

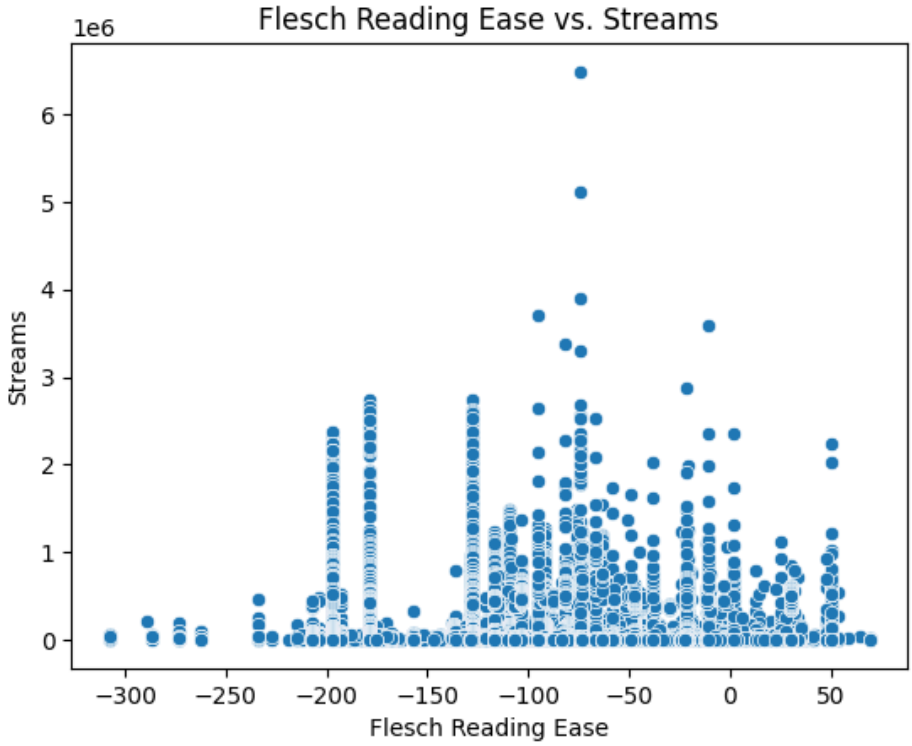

Fig 3.1 Scatter plot of Popularity vs Complexity

The visualization aims to explore the relationship between **Flesch Reading Ease** (a measure of textual complexity) and **Streams**. The data points are predominantly clustered between Flesch Reading Ease scores of -150 and -50, suggesting that moderately complex texts dominate the dataset. These texts likely appeal to a broader audience, striking a balance between readability and some level of depth or sophistication.

A few points stand out with significantly high streaming values, particularly around very low Flesch scores (high complexity). This indicates that even highly complex texts can achieve notable popularity, although they appear as exceptions rather than the norm. The dataset spans a wide range of Flesch scores, from approximately -300 to +50, indicating that the texts vary in

complexity from highly complex (negative scores) to very simple (positive scores), representing diverse readability levels.

No clear linear trend emerges between Flesch Reading Ease and Streams, suggesting that textual complexity alone may not be the primary determinant of a text's popularity. While this visualization provides useful insights, further enhancements could make it more visually appealing and easier to comprehend. An enhanced version of this visualization can be found in the Results section.

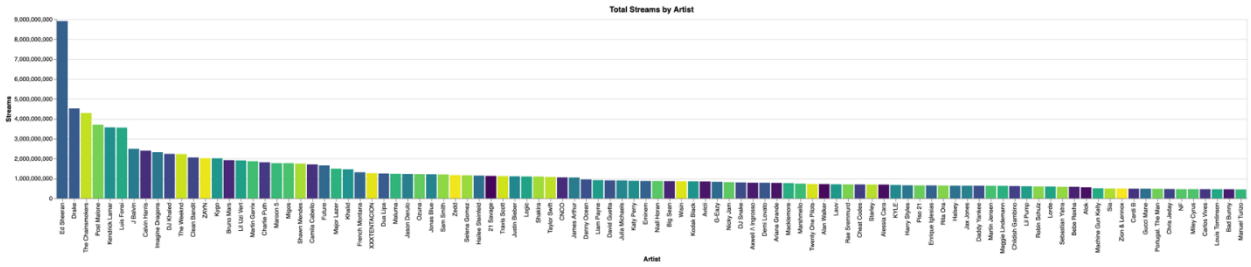

Fig 3.2 Top 100 artists by total streams globally

Figure 3.2 presents the top 100 artists by total streams, revealing a similar trend of concentration, with a small group of artists dominating the total streams. However, compared to individual tracks, the disparity between artists is less pronounced, likely due to the aggregation of streams across multiple songs per artist. The inclusion of a wide range of artists in the top 100 reflects the industry's breadth while also reinforcing the dominance of globally renowned artists at the top.

These findings highlight the dual dynamics of the music industry: a strong preference for a few megahits and artists, alongside a broad, diverse base of moderately popular tracks and performers. This understanding of listener behavior is valuable for marketing, playlist curation, and talent promotion in the music streaming ecosystem.

The consistent Y-axis scale facilitates comparison, and the visual representation of streaming numbers enhances clarity. However, visualization suffers from overcrowding, which makes reading individual artist names difficult, especially for those with fewer streams. The issue is exacerbated by the dense X-axis labeling, where names overlap and become nearly illegible.

Additionally, the use of color, while enhancing the aesthetic, appears arbitrary since no legend or explanation is provided. This lack of clarity can confuse viewers, reducing the effectiveness of the chart in conveying meaningful patterns. To improve the visualization, focusing on a subset of the data, such as the top 20 or 50 artists, could reduce overcrowding. Grouping less prominent artists into an "Other" category would also help simplify the chart. Adjusting the axis labels—by rotating, staggering, or implementing interactive visualization tools—could enhance readability. Moreover, using a single-color gradient to represent the magnitude of streams would create a more intuitive and cohesive visual experience. These adjustments would make the chart more accessible and impactful for viewers. An alternative to this visualization could be a scatter plot of Artist vs Popularity which is displayed in the next section.

3.Results

Fig 3.3 Enhanced Scatter plot of Popularity vs Complexity

Most streamed content correlates with moderately negative Flesch Reading Ease scores (-120 to -50), suggesting that content with a certain level of complexity might be more appealing or relatable to the audience.

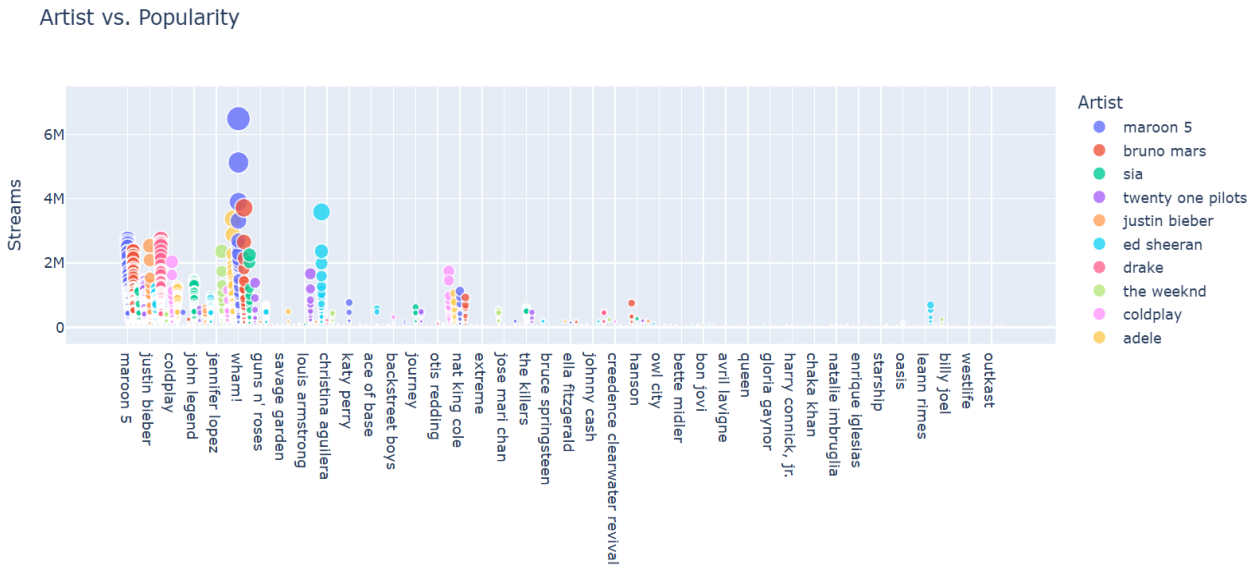

Fig 3.4 Relation between Artists, Streams and Popularity

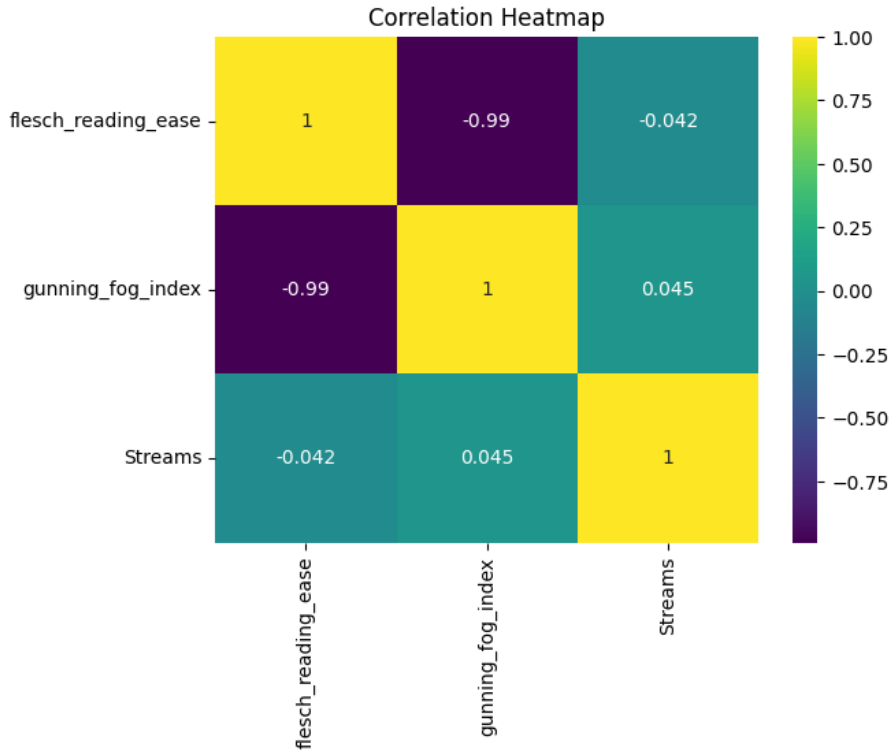

Fig 3.5 Correlation Matrix of Flesch Reading Ease, Gunning Fog Index, and Streams.

Figure 3.5 presents a correlation heatmap that shows the relationships between Flesch Reading Ease, Gunning Fog Index, and Streams. The heatmap reveals a strong negative correlation (-.99)

between the Flesch Reading Ease and the Gunning Fog Index, which is expected as these metrics measure textual complexity in inverse ways. A weak negative correlation (-0.042) is observed between Flesch Reading Ease and Streams, suggesting that simpler texts are only slightly associated with higher streaming numbers. Similarly, the Gunning Fog Index and Streams show a weak positive correlation (0.045), indicating that text complexity has minimal direct influence on streaming popularity. The heatmap effectively summarizes these relationships, though it is important to note that correlation does not imply causation, and other factors may influence the results.

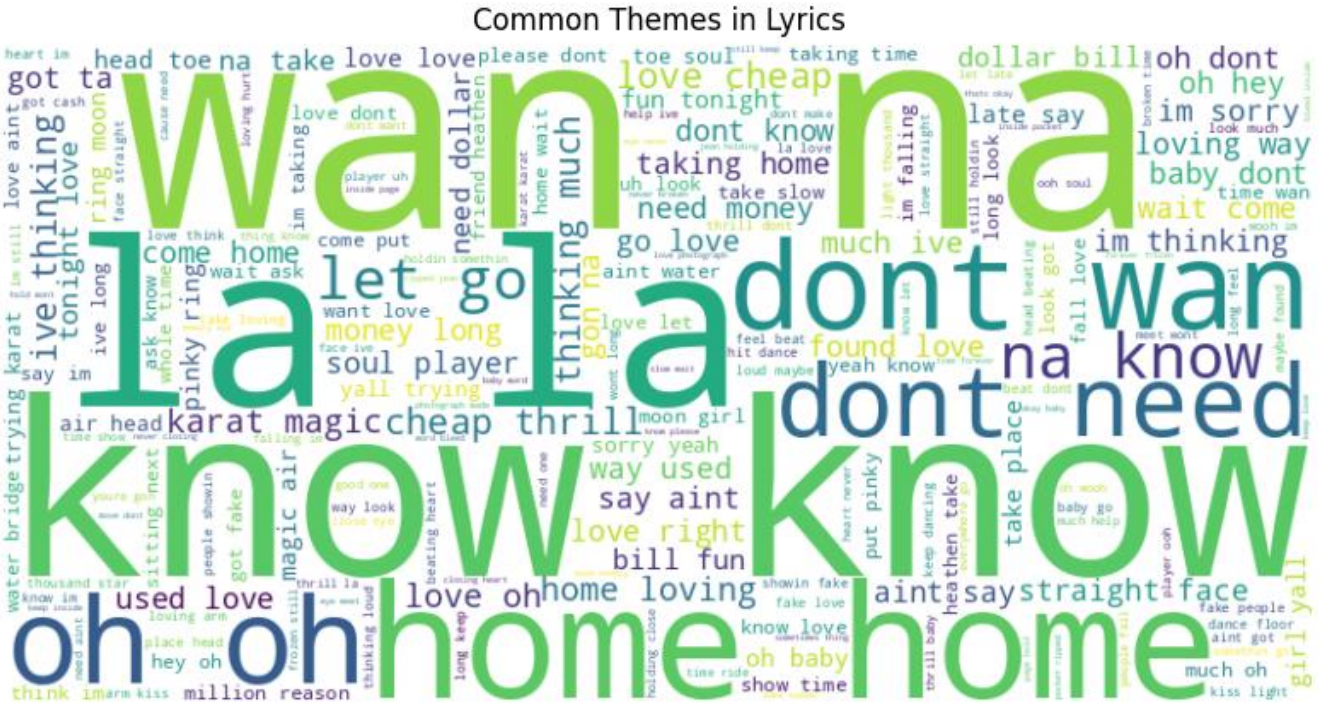

Fig 3.6 Word cloud for most frequent words in the dataset

Figure 3.6 provides a word cloud illustrating the most frequent words in the lyrics dataset. Larger words like “know,” “home,” “wanna,” and “don’t” suggest common themes of introspection, longing, and relational dynamics. This visualization is particularly effective for qualitative analysis, giving a quick overview of dominant themes and emotions conveyed in the lyrics. However, the lack of contextual information or frequency counts for the words limits

deeper interpretation. For instance, while “know” appears prominently, it is unclear whether it signifies questions, declarations, or emotional states without additional context.

Together, these visualizations offer complementary insights. The heatmap quantitatively establishes weak relationships between textual complexity and popularity, while the word cloud qualitatively highlights recurring themes that may resonate with audiences and contribute to the popularity of songs. These methods align well with the dataset’s goals, providing a balance of statistical analysis and thematic exploration.

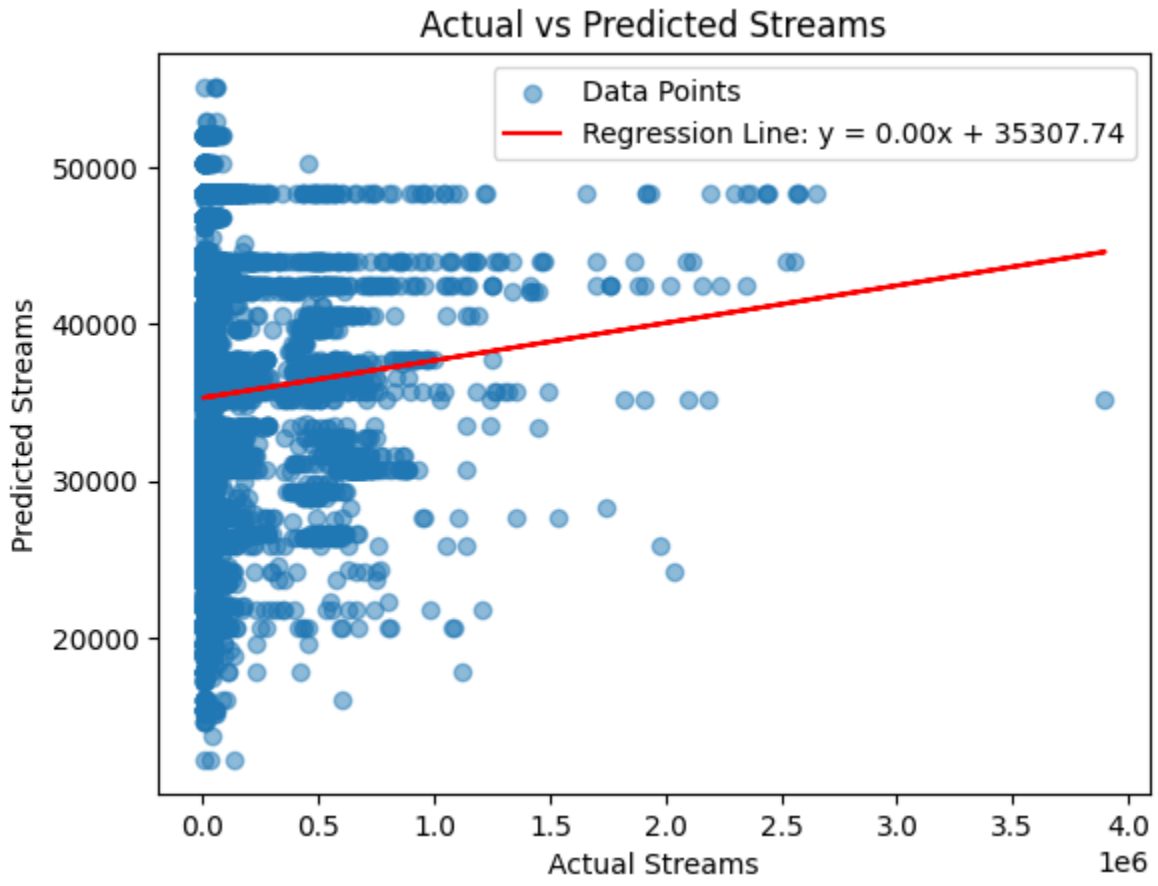

Fig 3.7 Actual Streams vs Predicted Streams

Figure 3.7 displays a scatter plot comparing **actual streams** (on the x-axis) with **predicted streams** (on the y-axis). The plot indicates a significant mismatch between predicted and actual

values, with most predictions clustering in a narrow range (below 50,000) while actual streams vary widely, reaching up to 4 million. This suggests that the prediction model may lack the capability to accurately capture the variability in streaming data, especially for higher values. The model seems to consistently underestimate actual streams, which could be due to missing influential features, an overly simplistic model, or inherent unpredictability in streaming performance. While the scatter plot is effective for visualizing discrepancies between actual and predicted values, it highlights the need for improvements in the prediction model.

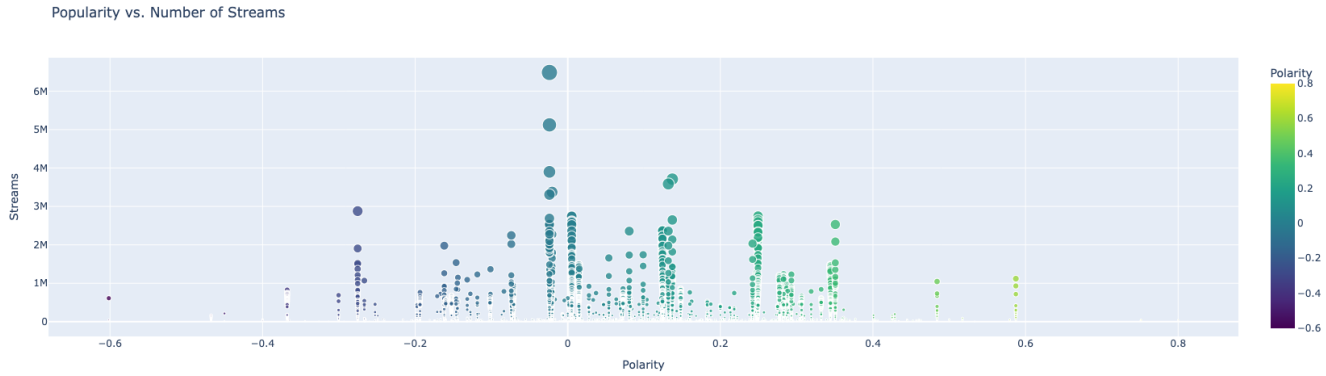

Fig 3.8 Relation between popularity, streams, and sentiment polarity

Figure 3.8 provides a scatter plot illustrating the relationship between popularity, streams, and sentiment polarity (indicated by color). The x-axis represents polarity, ranging from negative to positive sentiment, while the y-axis measures streams. The color gradient emphasizes polarity, with darker colors indicating negative sentiment and lighter shades representing positive sentiment. The plot reveals no strong correlation between sentiment polarity and streams; both positively and negatively polarized songs achieve a wide range of streams. However, the clustering of points around the center suggests that songs with neutral or slightly positive sentiment are more prevalent, and some of these achieve higher popularity. This could imply that moderately neutral sentiments resonate more with listeners, but sentiment alone is insufficient to predict streams reliably.

Together, these visualizations highlight critical insights. The scatter plot in *Figure 3.7* underscores the limitations of the current predictive model for streaming data, particularly its inability to handle variability in higher stream counts. On the other hand, *Figure 3.8* shows that sentiment polarity has minimal direct influence on streams, suggesting that other factors, such as genre, marketing, or artist reputation, might play a more significant role in determining popularity. These insights suggest a need for refining both prediction models and feature selection to better capture the complexities of streaming success.

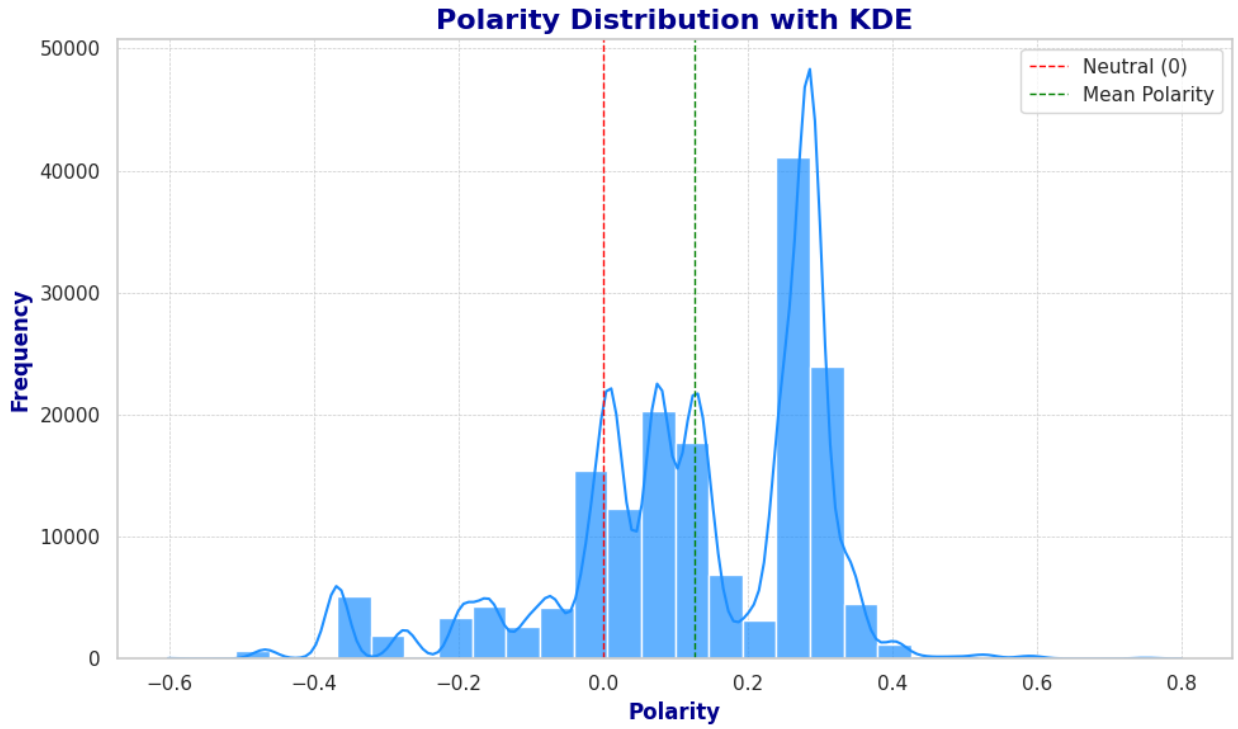

Fig 3.9 KDE of Polarity Distribution

The KDE curve helped us see how sentiment gradually shifts across the dataset. In general, the overall sentiment skew toward positive sentiment with polarity values generally in the range of -0.4 and 0.6. The most frequent polarity value is around 0.2. While negative values do exist, particularly those as low as -0.6, their occurrence is much less frequent

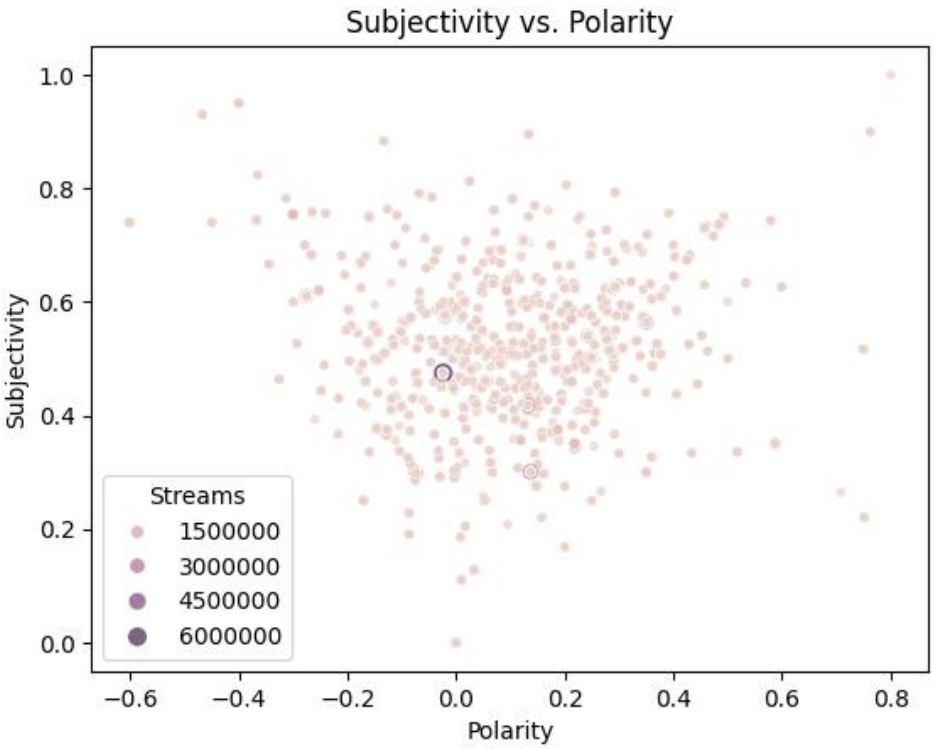

Fig 3.10 Relation between song lyrics subjectivity and sentiment polarity

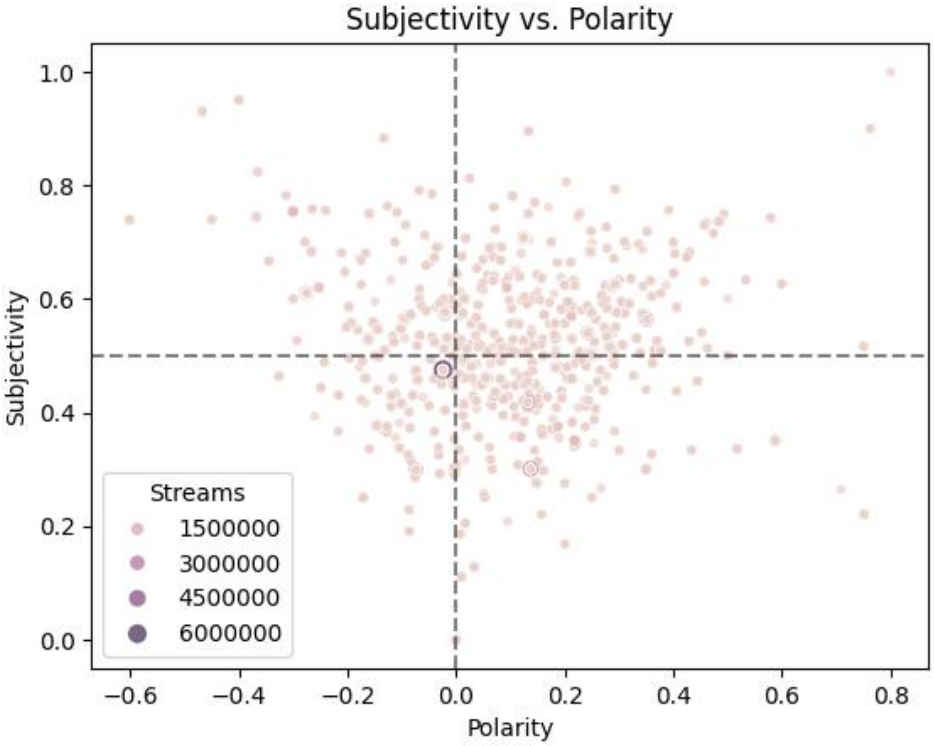

Fig 3.11 Fig 3.10 Relation between song lyrics subjectivity and sentiment polarity

The analysis of sentiment polarity and its connection to subjectivity in song lyrics has revealed a nuanced relationship. Both highly positive and highly negative polarities tend to indicate moderate subjectivity, suggesting that songs in these categories often express personal experiences, opinions, or emotional perspectives. This aligns with the understanding that music serves as a medium for artists to convey their emotions—whether it be joy, love, sadness, or anger—in a way that resonates with listeners on a personal level.

Outliers observed in the data likely represent songs with strong subjectivity, where highly personal or intense emotions dominate. These outliers could be impactful for niche audiences or evoke strong reactions but might not have the same broad appeal as moderately subjective songs. For instance, a highly emotional or controversial theme might resonate deeply with a small group of listeners while alienating others.

4.Discussion

The findings of this project provide a comprehensive understanding of the relationship between lyrical complexity, sentiment polarity, subjectivity, and song popularity. The analysis of streaming data has revealed several critical insights into modern music consumption trends. Firstly, the relationship between lyrical complexity, as measured by tools like the Flesch Reading Ease and Gunning Fog Index, and song popularity appears to be weak. Most streamed songs cluster within a moderate range of complexity, suggesting that audiences prefer lyrics that strike a balance between simplicity and sophistication. While highly complex lyrics can achieve significant popularity, they are exceptions rather than the norm. This indicates that lyrical complexity alone is not a primary determinant of a song's success.

Sentiment analysis further highlights the nuanced role of emotional content in song lyrics. Both highly positive and highly negative polarities were found to correspond with moderate subjectivity. This suggests that songs expressing strong emotions, whether positive or negative, often draw from personal experiences or perspectives, making them relatable to listeners. The clustering of songs with neutral or slightly positive sentiment in higher popularity brackets implies that moderately neutral sentiments resonate more broadly with audiences. Outliers, characterized by highly subjective or emotionally intense lyrics, may appeal to niche audiences, evoking strong reactions but lacking widespread appeal.

The visualizations offer additional insights. The scatter plot comparing actual and predicted streams revealed significant limitations in the predictive model. The clustering of predictions within a narrow range, despite wide variability in actual streams, underscores the need for a more robust model. Similarly, the scatter plot exploring the relationship between sentiment polarity, streams, and popularity demonstrated minimal direct influence of sentiment on streams, suggesting that external factors such as genre, marketing strategies, or artist reputation play a more substantial role. The correlation heatmap further reinforced the weak

relationship between textual complexity and streams, while the word cloud highlighted recurring themes of introspection, longing, and relational dynamics in popular songs.

These results suggest that while lyrical complexity is not the sole driver of success, it interacts with other factors such as genre, artist reputation, and emotional content to shape listener preferences. This interplay reinforces the notion that song success cannot be attributed to a single characteristic but rather emerges from a nuanced combination of multiple elements.

5.Conclusion

This study demonstrates that lyrical content, sentiment polarity, and subjectivity contribute to the overall appeal of songs but are not standalone determinants of popularity. While moderately complex lyrics and neutral or slightly positive sentiments tend to resonate with a broader audience, other factors such as artist reputation, genre, and marketing significantly influence streaming success. The weak correlation between lyrical complexity and streams suggests that textual characteristics alone cannot predict popularity, emphasizing the multifaceted nature of audience preferences.

Additionally, the analysis of subjectivity reveals that personal experiences and emotional expressions are central to lyrical relatability. Songs that strike a balance between generality and personal emotion tend to achieve greater success, reflecting the role of music as a shared emotional medium. However, the limitations of the current predictive model highlight the complexity of streaming performance, which involves a multitude of interacting factors beyond textual analysis.

6.Future work

Future research could address several areas to build upon the findings of this project. First, refining the predictive model by incorporating additional features such as genre, artist reputation, marketing efforts, release timing, and social media activity could improve its accuracy. Exploring machine learning approaches with richer datasets could provide deeper insights into the determinants of streaming success.

Second, expanding the analysis to include more diverse datasets, such as lyrics in non-English languages or emerging artists, could provide a more global perspective. Understanding regional variations in musical preferences and their impact on streaming trends would be particularly valuable for international marketing strategies.

Third, integrating audio analysis with lyrical content could offer a holistic view of a song's appeal. Features like melody, rhythm, and instrumentation could be analyzed alongside lyrical characteristics to better understand their combined influence on popularity.

Finally, investigating the role of playlist curation and algorithmic recommendations on streaming performance could provide actionable insights for artists and record labels. As streaming platforms increasingly influence listening behavior, understanding the interplay between content and algorithms will be essential for navigating the evolving music industry.

In conclusion, this study highlights the complexity of factors shaping modern music consumption and provides a foundation for future research into the interplay between lyrical content, sentiment, and audience engagement.

7.References

[1] Somme SVD, Sogancioglu G, Paperno D. Popularity of music tracks based on lyrics. Utrecht University; 2021.

[2] The Role of a Record Company. url: https://powering- the- musicecosystem.ifpi.org/.

[3] M. A. Casey et al. “Content-Based Music Information Retrieval: Current Directions and Future Challenges”. In: Proceedings of the IEEE 96.4 (2008), pp. 668–696. doi: 10.1109/JPROC.2008.916370.

[4] Thomas G. Bever The cognitive basis for linguistic structures In: Cognition and Language Development (New York: Wiley & Sons, 1970), pp. 279–362. doi: 10.1093/acprof:oso/9780199677139.003.0001

[5] Varun Sardana Veritas AI Tuning into Trends: Machine Learning Models for Song

Popularity Prediction on Spotify in: The National High School Journal of Science 2023

[6] Jaehyun Kim Music Popularity Prediction Through Data Analysis of Music’s Characteristics International Journal of Science, Technology and Society 2021; 9(5): 239-244 http://www.sciencepublishinggroup.com/j/ijsts doi: 10.11648/j.ijsts.20210905.16 ISSN: 2330-7412 (Print); ISSN: 2330-7420 (Online)

[7] https://www.kaggle.com/datasets/edumucelli/spotifys-worldwide-daily-song-ranking

[8] Dorien Herremans, David Martens, and Kenneth S¨orensen. “Dance Hit Song Prediction”. In: Journal of New Music Research 43.3 (2014), pp. 291– 302. doi: 10.1080/09298215.2014.881888. eprint: https://doi.org/10.1080/09298215.2014.881888. url: https://doi.org/10.1080/ 09298215.2014.881888.

[9] Sumiko Asai. “Factors Affecting Hits in Japanese Popular Music”. In: Journal of Media Economics 21.2 (June 2008), pp. 97–113. doi: 10.1080/ 08997760802069895.

[10] Myra Interiano et al. “Musical trends and predictability of success in contemporary songs in and out of the top charts”. In: Royal Society Open Science 5.5 (May 2018), p. 171274. doi: 10 . 1098 / rsos . 171274. url: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5990848/.

[11] Onder C¸ oban. “Turkish Music Genre Classification using Audio and Lyrics ¨ Features”. In: Suleyman Demirel University Journal of Natural and Applied Sciences (May 2017). doi: 10.19113/sdufbed.88303.

[12] Ricardo Malheiro et al. “Emotionally-Relevant Features for Classification and Regression of Music Lyrics”. In: IEEE Transactions on Affective Computing 9 (Jan. 2018), pp. 240–254. doi: 10.1109/TAFFC.2016.2598569.

[13] Markus Schedl. “Genre Differences of Song Lyrics and Artist Wikis: An Analysis of Popularity, Length, Repetitiveness, and Readability”. In: The World Wide Web Conference. WWW ’19. San Francisco, CA, USA: Association for Computing Machinery, 2019, pp. 3201–3207. isbn: 9781450366748. doi: 10 . 1145 / 3308558.3313604. url: https : / / doi . org / 10 . 1145 / 3308558.3313604.

[14] Kahyun Choi et al. “Music subject classification based on lyrics and user interpretations”. In: Proceedings of the Association for Information Science and Technology 53.1 (2016), pp. 1–10. doi: https://doi.org/10.1002/pra2.2016.14505301041. eprint: https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/pra2.2016.14505301041. url: https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/pra2.2016. 14505301041.

[15] Mauch, M., et al. (2015). The evolution of popular music: USA 1960–2010. Royal Society Open Science.

[16] Arora, S., Rani, R. Soundtrack Success: Unveiling Song Popularity Patterns Using Machine Learning Implementation. SN COMPUT. SCI. 5, 278 (2024). https://doi.org/10.1007/s42979-024-02619-5

[17] Yee YK, Raheem M. Predicting music popularity using spotify and youtube features. Indian J Sci Technol. 2022;15:1786–99. https://doi.org/10.17485/ijst/v15i36.2332.

[18] Brooks, A., et al. (2019). Sentiment Analysis of Lyrics in Popular Music. Journal of Music Psycholgy.

[19] Flesch, R. (1948). A new readability yardstick. Journal of Applied Psychology.

[20] https://www.kaggle.com/datasets/edumucelli/spotifys-worldwide-daily-song-ranking

[21] https://www.kaggle.com/datasets/joebeachcapital/57651-spotify-songs/data

[22] https://raw.githubusercontent.com/johan/world.geo.json/master/countries.geo.json