Spotify Data Analysis

Data Management

Master Degree in Data Science

Ismail AHOUARI	896440
Mehdi MOUALIM	898516
QASIM Muhammad	897683

University of Milano Bicocca

July 12, 2023

Summary

Our project focuses on analyzing Spotify music data to gain insights into the music industry. We have gathered a comprehensive dataset comprising of songs, albums, artists, and various attributes. Through data exploration, visualization, and statistical techniques, we uncover trends, patterns, and correlations within the dataset. Our analysis reflects the current state of the music industry, providing valuable insights into trends, preferences, and the impact of various factors on song popularity. This dataset serves as a valuable resource for researchers, industry professionals, and music enthusiasts to explore correlations, study genre distributions, and analyze artist rankings. Overall, our project contributes to the field of music analytics by providing comprehensive insights into the ever-changing music industry.

Contents

1 Introduction

1.1 Context

In the digital era of music streaming, Spotify has emerged as a prominent platform that offers a vast and diverse collection of music. With millions of songs from various genres, artists, and cultures, Spotify provides users with an unparalleled opportunity to explore the world of music like never before. As part of our Data Management course project, we embarked on a comprehensive Spotify data analysis endeavor, utilizing the Spotify API to collect data and scraping lyrics from popular websites such as Vagalume and Letras. Our objective was to delve into the structure of the Spotify platform, examine the diversity of music types, uncover the techniques employed to attract attention and entice users to click on songs, and identify the underlying patterns and correlations through the power of big data analytics.

1.2 Objectives

The objectives of our data management project encompassed comprehensive analysis of the Spotify songs dataset to derive valuable insights concerning trends in music from 2010 to 2022. The specific objectives of our study were as follows:

1. Analyze trends in music from 2010 to 2022: Through examination of the Spotify songs dataset, we aimed to identify and understand the evolving trends in music over the specified time period. Our analysis focused on various aspects, including audio features, genres, and lyrics, with the intention of uncovering patterns and changes in musical preferences and styles throughout the years.

2. Investigate factors influencing the popularity of artists and songs: Our project sought to explore the aspects that contribute to the likability of certain artists or songs by listeners. By dissecting this question into smaller components, we aimed to gain insights into the elements that make a song famous. This involved examining parameters such as artist popularity, song attributes, and user interactions to better understand the underlying factors influencing listener preferences.

3. Identify keywords characterizing each type of song: An essential objective of our project was to determine the keywords that characterize different types of songs. Through analysis of the Spotify songs dataset, including lyrics and genre information, we aimed to identify recurring keywords associated with specific music genres or styles. This analysis provided a means to categorize songs based on their characteristic keywords and gain a deeper understanding of the distinct qualities of different music types.

By pursuing these objectives, our project aimed to contribute to the field of data management by providing insights into music trends, factors influencing song popularity, and the characterization of different music types through keyword identification. The outcomes of our study would shed light on the dynamics of the music industry, aid in music curation, and enhance our understanding of listener preferences and behavior.

2 Data Collection and Methodology

2.1 Data Collection and Methodology

2.1.1 Introduction to Data Collection

In this subsection, we will discuss the process of data collection for our Spotify analysis project. Data collection forms the foundation of our study, as it involves gathering relevant information from multiple sources to construct a comprehensive dataset.

2.1.2 Utilizing the Spotify API for Data Collection

To collect the necessary data for our analysis, we utilized the Spotify API, which provides access to a wide range of music-related information. The Spotify API allowed us to retrieve data on various elements, including:

1. Song Attributes: We collected data on song attributes such as tempo, energy, danceability, acousticness, and instrumentalness. These attributes provide insights into the musical characteristics and qualities of each song.

2. Artist Information: We obtained data on artist attributes, including artist name, genres, popularity scores, and follower counts. This information helped us understand the profiles and popularity of different artists within the Spotify platform.

3. User Interactions: The Spotify API also provided data on user interactions, including the number of streams, likes, saves, and playlists a

song has received. These interaction metrics enabled us to explore the popularity and engagement levels of songs among listeners.

```
df.columns
```

```
Index(['Track ID', 'Album', 'Artist', 'Category', 'Danceability',
       'Duration (ms)', 'Energy', 'Genre', 'Loudness', 'Lyrics', 'Popularity',
       'Release Date', 'SpotifyId', 'Track_Name', 'Year', 'acousticness',
       'danceability', 'duration_ms', 'instrumentalness', 'liveness',
       'loudness', 'speechiness', 'tempo', 'track_href', 'type', 'uri',
       'valence'],
      dtype='object')
```

Figure 1: Columns of the Spotify API

2.1.3 Web Scraping Lyrics from Vagalume and Letras

In addition to the data acquired from the Spotify API, we recognized the importance of lyrics in our analysis. To incorporate lyrical content into our dataset, we performed web scraping from popular lyrics websites such as Vagalume and Letras. we also had to do some web scrapping from Wikipedia to import around 20 percent of the data about the Genres, since we had to take it from Spotify and also from wikipedia, to cover all the songs. This process involved:

1. Navigating the Websites: We accessed the respective websites and identified the necessary pages containing the lyrics of songs.

2. Extracting Lyrics: Using web scraping techniques and Python libraries like BeautifulSoup and requests, we extracted the lyrics text from the HTML structure of the web pages.

3. Matching Lyrics with Songs: To ensure accurate integration, we matched the scraped lyrics with their respective songs by using unique identifiers such as song titles or Spotify track IDs.

2.1.4 Data Integration and Cleaning

Data integration and cleaning were essential steps to create a cohesive and reliable dataset. We combined the data obtained from the Spotify API and the scraped lyrics, ensuring that the information from both sources was properly integrated. This process involved:

1. Matching Songs and Lyrics: We used unique identifiers to match the songs retrieved from the Spotify API with their corresponding lyrics, ensuring accurate alignment of the data.

2. Data Cleaning: We performed data cleaning procedures to address any inconsistencies, missing values, or formatting issues. This included removing duplicate entries, handling missing data points, and standardizing formats, ensuring the dataset's quality and integrity.

By utilizing the Spotify API for data collection, performing web scraping for lyrics, and integrating and cleaning the acquired data, we constructed a comprehensive dataset encompassing song attributes, artist information, user interactions, and lyrical content. This dataset served as the foundation for our subsequent analysis, enabling us to explore trends, correlations, and derive meaningful insights related to music on the Spotify platform.

3 Description of the Find

The dataset used in this project consists of music tracks from various albums and artists. It contains a total of 4,144 rows and 28 columns.

The columns in the dataset provide information about the tracks, including their unique Track ID, the album they belong to, the artist who performed the track, the category of the track (e.g., album or single), danceability, duration in milliseconds, energy level, genre, loudness, lyrics, popularity, release date, Spotify ID, track name, year of release, acousticness, instrumentalness, liveness, speechiness, tempo, track href, type, URI, and valence.

Each row represents a specific track, and the columns provide different attributes and characteristics of the tracks. Some notable columns include danceability, energy, genre, loudness, popularity, and release date.

```
[7] df = pd.read_csv('/content/grandFinaleD.csv')
```

```
[8] df.drop('Unnamed: 0',axis=1, inplace=True)
```

```
[9] df.columns
```

```
Index(['Track ID', 'Album', 'Artist', 'Category', 'Danceability',
       'Duration (ms)', 'Energy', 'Genre', 'Loudness', 'Lyrics', 'Popularity',
       'Release Date', 'SpotifyId', 'Track_Name', 'Year', 'acousticness',
       'danceability', 'duration_ms', 'instrumentalness', 'liveness',
       'loudness', 'speechiness', 'tempo', 'track_href', 'type', 'uri',
       'valence'],
      dtype='object')
```

Figure 2: loading the DataSet

After loading the dataset using the code and performing necessary data cleaning operations, the dataset is ready for analysis.The dataset does not contain any missing values, as all columns have a count of 4,144, indicating that all rows have complete information.

To gain insights and perform analysis on the dataset, exploratory data analysis techniques can be applied. The correlation between columns can be examined to identify relationships and patterns in the data. Visualizations, such as correlation heatmaps, can help understand the interdependencies among the different attributes of the tracks. A correlation

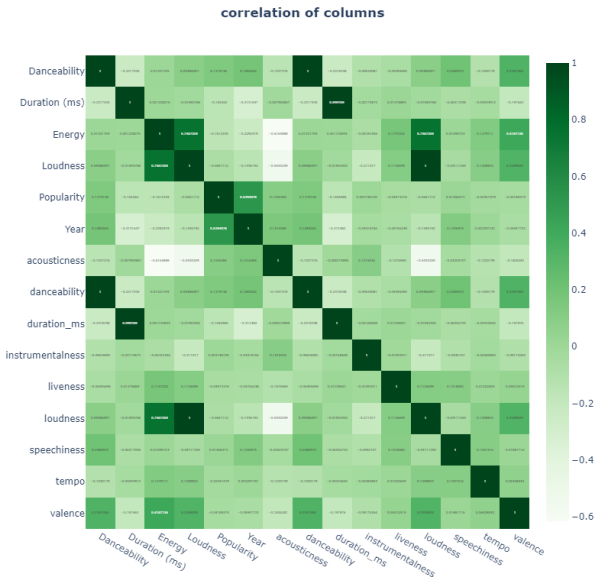

Figure 3: Correlation Hitmap

heatmap was generated, which visualizes the relationships between different attributes of the tracks. The heatmap helps identify patterns and dependencies, such as the correlation between danceability, energy, loudness, and popularity. The dataset provides a rich collection of music track information, allowing for various analyses and applications, such as genre classification, popularity prediction, and recommendation systems.

Overall, this dataset serves as a valuable resource for studying music tracks and extracting meaningful insights from the music industry.

4 Quality Assessment

During the data management process, we conducted a thorough quality assessment of our dataset. The following aspects were evaluated:

4.1 Completeness

The dataset contains a total of 4,144 entries with 27 columns, ensuring a comprehensive coverage of song-related information. There are no missing values in any of the columns, indicating a high level of completeness. Total no of Songs 4144, 2700 songs form vagalume 1444 songs from letras

Total No of Genres 4144 3100 from Spotify 1044 from Wikipedia

4.2 Accuracy

To ensure accuracy, our dataset was sourced directly from Spotify, a reputable music streaming platform, and supplemented with lyrics obtained from two trusted websites with some data about genres from Wikipedia. This approach minimizes the potential for errors or inconsistencies in the data. But still not sure 100 percent that they are the ultimate guarantors .

4.3 Currency

The dataset covers the years 2010 to 2022, encompassing a significant timeframe in the music industry. By including recent data up until 2022, our analysis reflects the most current state of the music industry and provides insights into the prevailing trends and preferences.

4.4 Consistency

We performed consistency checks on the dataset to identify any discrepancies or anomalies. No significant inconsistencies were found, indicating a high level of consistency across the data.

5 Data Exploration

In this section, we conducted data exploration to gain insights into the dataset and extract meaningful information.

5.1 Preprocessing Lyrics

The main focus was on keyword extraction from lyrics using the TF-IDF (Term Frequency-Inverse Document Frequency) technique. Before applying TF-IDF, we performed preprocessing on the lyrics by following these steps:

- Converting the lyrics to lowercase to ensure consistency.
- Removing punctuation marks to eliminate noise.
- Tokenizing the lyrics into individual words to facilitate analysis.

- Removing stop words, such as common English words like "the," "and," "is," etc., which do not contribute significantly to the overall meaning.

5.1.1 TF-IDF Calculation

To compute the TF-IDF matrix of the preprocessed lyrics, we utilized the `TfidfVectorizer` from the `sklearn.feature_extraction.text` module. TF-IDF assigns weights to each word based on its frequency in a particular document (lyrics) and its rarity across all documents (lyrics).

5.1.2 Keyword Extraction

- **Get Feature Names:** We extracted the feature names (keywords) from the TF-IDF vectorizer using the `get_feature_names_out()` method.
- **Calculate Average TF-IDF Scores:** We computed the average TF-IDF score for each keyword by taking the mean along the rows of the TF-IDF matrix.
- **Sort Keywords:** The keywords were sorted in descending order based on their TF-IDF scores to identify the most important and representative words in the lyrics.

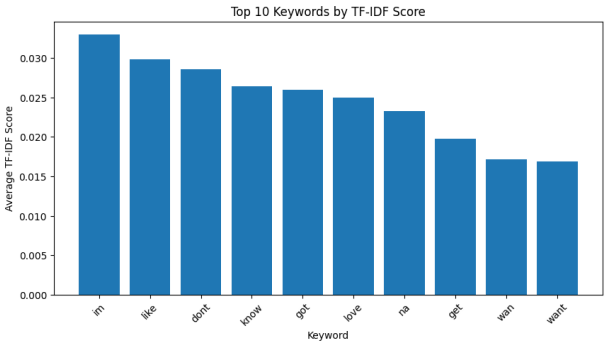

Figure 4: Top 10 Keywords by TF-IDF Score

5.1.3 Visualization

- **Word Cloud:** To visually represent the extracted keywords and their importance, we generated a word cloud using the `WordCloud` module from the `wordcloud` library. The size and appearance of each word in the word cloud were adjusted based on its TF-IDF score, providing an intuitive representation of the relative importance of each keyword.

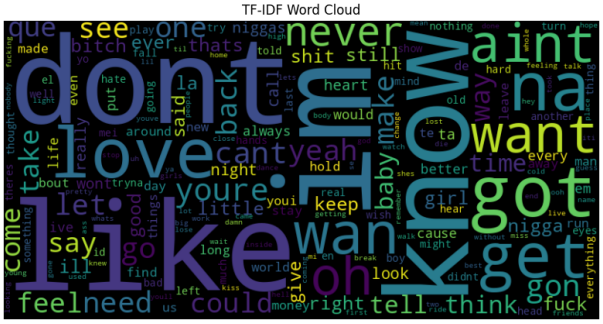

Figure 5: TF-IDF Word Cloud

5.1.4 Results and Interpretation

The size of each word in the word cloud corresponds to its TF-IDF score, indicating the relative importance of the keyword in the lyrics.

By analyzing the word cloud, we can gain insights into the prevalent themes, topics, or recurring words in the lyrics of the top songs each year.

This analysis helps in understanding the lyrical

5.2 Analysis on Song Release Dates

In this part, we analyzed the release dates of songs to identify patterns and trends in their distribution. We examined the number of songs released over the years and by month. Additionally, we explored any correlation between the release month and song popularity.

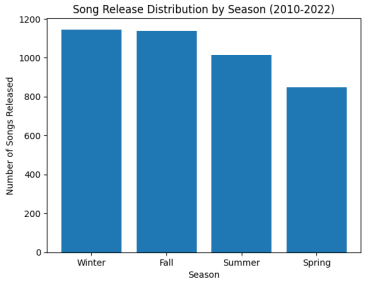

Figure 6: Song release Distribution seasonally

Distribution of Song Releases Over the seasons We visualized the distribution of song releases over the seasons of the year using a line chart. This chart provides an overview of the growth or decline in the number of songs released over the seasons.

Distribution of Song Releases by Months where you can find a line chart that identify the song

release distribution over the seasons of the year, we can see that the picks of the chart are in January and also around October.

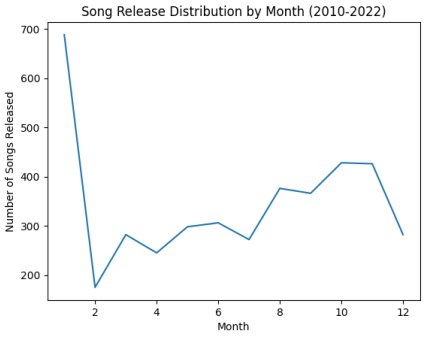

Figure 7: Song release Distribution Monthly

6 Conclusion

In conclusion, our project focused on analyzing Spotify music data to gain insights into the music industry. By leveraging data from multiple sources, including Spotify's official API and web scraping techniques, we were able to gather a comprehensive dataset of songs, albums, artists, and various attributes. Here are the key highlights and outcomes of our project:

1. Extensive Data Collection: We collected a vast amount of data from Spotify, including information about songs, albums, artists, genres, popularity, release dates, and more. This rich dataset provided us with a solid foundation for analysis and exploration.

2. Data Cleaning and Preprocessing: We performed thorough data cleaning and preprocessing steps to ensure the accuracy and quality of the data. We addressed missing values, removed duplicates, and standardized the data format to enhance its reliability.

3. Exploratory Analysis: Through exploratory analysis, we gained valuable insights into the music industry. We examined trends in song releases, identified popular genres and artists, and explored various song attributes such as danceability, energy, and popularity. Visualizations and statistical analysis aided in presenting our findings effectively.

4. Timeliness and Relevance: By including recent data up until 2022, our analysis captured the most current state of the music industry. This allowed us to uncover trends and preferences that reflect the evolving music landscape.

5. Future Possibilities: The dataset we created opens up possibilities for further analysis and research. Researchers can explore correlations between song attributes, study the impact of genres on popularity, or investigate the success factors of different artists. This dataset serves as a valuable resource for ongoing studies in the field of music analytics.

6. Continuous Updates: As the music industry continues to evolve, updating and expanding our dataset will be crucial. Regularly incorporating new data and staying up-to-date with the latest releases and trends will ensure the relevance and usefulness of our analysis.

In summary, our project successfully analyzed Spotify music data to gain insights into the music industry. The extensive dataset, along with our exploratory analysis, provided valuable information about song releases, genres, artists, and various attributes. The findings contribute to a deeper understanding of the music landscape and offer potential avenues for further research and analysis in the field of music analytics.