Analysing Factors Influencing the Popularity of Top 50 Spotify Songs in 2023

Jan Eljasiak, Kamila Fido, Weronika Pyrka

Research Question:

We wanted to find out what makes a song popular on Spotify. Specifically, we asked: "What kind of words, feelings and topics are common in the top 50 Spotify songs in 2023 and how are they impacting their popularity?"

Project components and their roles

Datasets and acquisition methods:

For our project, we focused on getting three main types of data: song titles, lyrics and detailed song metrics.

- **Getting song titles from Spotify:**
  - We used Spotify's API to access a playlist created by Spotify itself. This playlist had 50 most popular songs of the year.
- **Fetching lyrics with Zylalabs API:**
  - Than to get the lyrics for these songs, we used the Zylalabs API.
  - This step was crucial because the lyrics are key to our analysis of sentiments and themes in popular songs.
- **Enriching our dataset with Kaggle:**
  - Besides the basic song data and lyrics, we also used a dataset from Kaggle to get more in-depth information about trends in the most popular songs in 2023. For some songs that were in the Spotify list (28 common positions), it was a great way for data enrichment. We also decided to analyze songs that didn’t occur in Spotify API to have a different perspective on the problem of finding variables that are making songs more popular also from the perspective of the different dataset.
  - Track name, artists name and the number of artists (artist_count).
  - Release date details (released_year, released_month, released_day).
  - The song's presence in various charts and playlists across platforms like Spotify, Apple Music and Deezer (in_spotify_playlists, in_apple_playlists, etc.).
  - Streaming counts (streams).
  - Musical features such as BPM (beats per minute), key, mode and percentages representing danceability, valence, energy, acousticness, instrumentalness, liveness and speechiness.

Role of data acquisition in our project:

The data acquisition phase was foundational for our project. Each type of data contributed uniquely:

- **Song titles from Spotify:** This was our starting point, helping us identify which songs were most relevant for our 2023 analysis.
- **Lyrics from Zylalabs:** Analyzing the lyrics was crucial for understanding the themes and emotions in popular songs. It allowed us to perform sentiment analysis and thematic exploration.
- **Detailed metrics from Kaggle:** This enriched our analysis by providing deeper insights into each song's popularity and musical characteristics. The additional data points like BPM, key and danceability helped us examine if and how these features correlate with a song's popularity.

By combining these diverse data sources, we were able to create a comprehensive dataset that not only told us what was popular, but also gave us the tools to explore why these songs might be popular.

Data integration:
To accomplish data integration, we had to address descriptive conflicts. Therefore, we changed the name of the relevant column ad merged columns from all three datasets. The outcome was then stored in MongoDB.

Data quality:

- Consistency - we verified the uniformity of data formats and structures across the diverse Spotify, Zylalabs and Kaggle datasets.
- Accuracy - we ensured that data from all of the datasets precisely represented the top 50 Spotify songs in 2023. We also resolved any descriptive conflicts for precision.
- Completeness - we examined the original datasets to identify and address any missing values or essential information gaps in song titles, lyrics or metrics.
- Timeliness – we conducted a thorough assessment of the timeliness of data acquisition, guaranteeing that the information collected accurately reflected the relevant timeframe for 2023.

Data cleaning:
We manually cleaned lyrics for accuracy by leaving just one version of lyrics.

Exploratory analysis:

Utilizes pandas and matplotlib for initial data inspection.

Sentiment and subjectivity analysis:
Conducted using TextBlob.

Topic modeling:
LDA through Gensim to identify prevalent themes.

Popularity correlation analysis:
Statistical methods are used to explore the relationship between NLP features and song popularity.

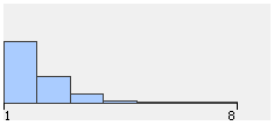
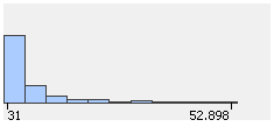
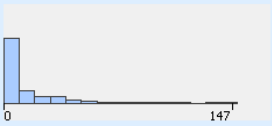
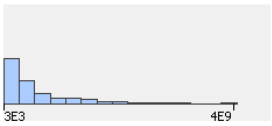
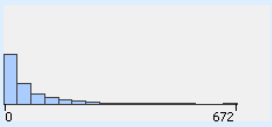
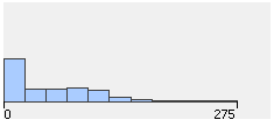
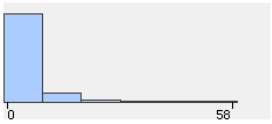
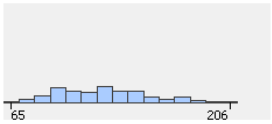
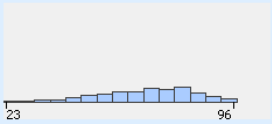
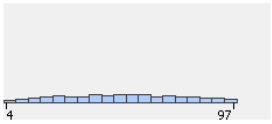
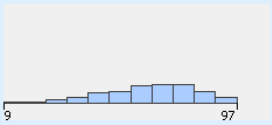
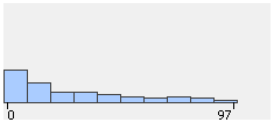
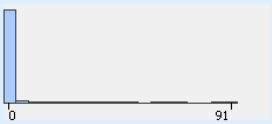
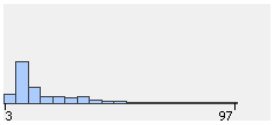
Software architectures:

- Python-based framework, utilizing libraries like nltk, Gensim, TextBlob, pandas and matplotlib
- MongoDB for data storage, chosen for its scalability and ease of integration with Python

Results:

- **Feeling the Lyrics:** Most songs had a happy mood. They also had a mix of personal feelings.
- **Themes:** Common themes were love, life and challenges. LDA showed us specific topics like dreams and overcoming hard times.
- **Link to Popularity:** We found that happier songs and songs about love and dreams were usually more popular.
- The project revealed that popular songs in 2023 often share positive sentiment and themes of love and aspiration.
- The analysis suggests that certain emotional and thematic elements in lyrics may influence a song's popularity on Spotify.

We utilized KNIME Analytics Platform to create charts for our project. KNIME is data management platform that allowed us to generate a variety of charts. KNIME's visualization tools significantly contributed to the comprehensive understanding of the factors influencing the popularity of the top 50 Spotify songs in 2023. This following image shows statistical moments such as minimum, maximum, standard deviation, variance, mean, overall sum, number of missing values and row count across all numeric columns from our database.

Column	Min	Mean	Max	Std. Dev.	Skewness	Kurtosis	Histogram
artist_count	1	1,5561	8	0,893	2,544	10,3667	
in_spotify_playlists	31	5.200,1249	52.898	7.897,609	2,9291	9,8761	
in_spotify_charts	0.0	12,0094	147	19,576	2,5805	8,5076	
streams	2.762	5,14E8	3,70E9	5,67E8	2,0006	4,3709	
in_apple_playlists	0.0	67,8122	672	86,4415	2,474	7,922	
in_apple_charts	0.0	51,9087	275	50,6302	1,0352	0,8905	
in_deezer_charts	0.0	2,6663	58	6,0356	3,7661	19,0211	
bpm	65	122,5404	206	28,0578	0,4132	-0,399	
danceability_%	23	66,9696	96	14,6306	-0,4359	-0,3336	
valence_%	4	51,4313	97	23,4806	0,0082	-0,9393	
energy_%	9	64,2791	97	16,5505	-0,4464	-0,26	
acousticness_%	0.0	27,0577	97	25,9961	0,9525	-0,1921	
instrumentalness_%	0.0	1,5813	91	8,4098	7,1242	56,6356	
liveness_%	3	18,213	97	13,7112	2,1043	5,7144	
speechiness_%	2	10,1312	64	9,9129	1,9347	3,3744	

Conclusions:

- The analysis of the top 50 Spotify songs in 2023 consistently revealed a prevalent positive sentiment, with a majority of songs conveying a happy mood. This aligns with the finding that songs centered around themes of love and dreams tended to be more popular. The project suggests a strong correlation between positive emotional content and a song's popularity on Spotify.
- The project's results imply that the emotional and thematic elements within lyrics hold predictive power for a song's popularity. Specifically, the association between happier tones and themes of love or aspiration and increased popularity suggests that listeners are drawn to music that elicits positive emotional responses.
- The common themes we found, like love, life and challenges, show things that everyone can relate to. We used a technique called Latent Dirichlet Allocation (LDA) to identify specific topics, such as dreams and overcoming tough times. This suggests that songs touching on basic human emotions and experiences usually become more popular.

Possible improvements and future developments:

- Incorporating more complex machine learning models could enhance the project's predictive capabilities. Advanced predictive modeling could further explore and quantify the impact of specific lyrical and musical features on future song popularity trends.
- Incorporating advanced natural language processing (NLP) techniques could offer deeper insights into lyrical content. Techniques such as sentiment analysis, named entity recognition and more sophisticated topic modeling approaches may provide a nuanced understanding of the lyrical intricacies influencing song popularity.
- To improve the generalizability of findings, future developments should focus on expanding the dataset. A multi-year, genre-inclusive analysis would enable a more comprehensive understanding of evolving patterns in popular music, accounting for potential shifts in listener preferences over time.