1/29/2024

# Statistical Analysis of Spotify's Top 2023 Songs

SHREYA KUSUMANCHI, DAVID ATTAR, PRITHVI ADIGA

#SPOTIFYWRAPPED

## Abstract

# Why are certain songs popular?

For our project, we chose to employ statistical and textual analysis using R to understand why certain songs are "popular". We are quantifying popularity using the number of streams as a metric. We are using two datasets, one containing the 1000 most streamed songs of 2023 and another containing the lyrics to 18,000 popular songs. Both datasets have access to Spotify metadata such as energy, speechiness, dancability, valence, liveness, and were sourced from Kaggle.com. We hope to understand what lyrics and musical attributes are synonymous with popular songs.

Spotify Metadata Used

# Description of metrics from Spotify Metadata

The variables which we employed for our testing, via linear regression, were:

1. **Valence** – A variable which Spotify creates to measure the "musical positiveness conveyed by a track"; Ranges from 0 to 100%
2. **Danceability** – A variable which measures how much someone can dance to a song; Ranges from 0 to 100%
3. **Energy** – A variable that measures how energetic a song is; Ranges from 0 to 100%
4. **Speechiness** – A variable that detects the presence of spoken words on a track; Ranges from 0 to 100% (for example, a podcast would have a speechiness score of almost 100)
5. **Instrumentalness** - A variable which acts as an indicator to determine whether a song has words or not; Ranges from 0 to 100%
6. **Acousticness** – A variable which acts as an indicator of whether a song is acoustic or not; Ranges from 0 to 100%
7. **Liveness** - A variable that detects the presence of an audience during a recording; Ranges from 0 to 100%
8. **BPM** – Beats per minute; Ranges from 65 bpm to 186 bpm

Spotify Metadata Used

# Hypothesis

**Null Hypothesis:** There is no statistically significant association or correlation between any of the Spotify metadata metrics shown in previous slide and an increase popularity or a larger number of streams.

**Alternative Hypothesis:** There is an association or correlation between any of the Spotify metadata metrics shown in previous slide and an increase popularity or a larger number of streams.

Furthermore, we hypothesize that songs that are categorized as pop songs will show a higher correlation in number of streams. Our research reveals that pop songs have higher than average bpm, energy and danceability. (2)

Linear Regression

## Cleaning Data

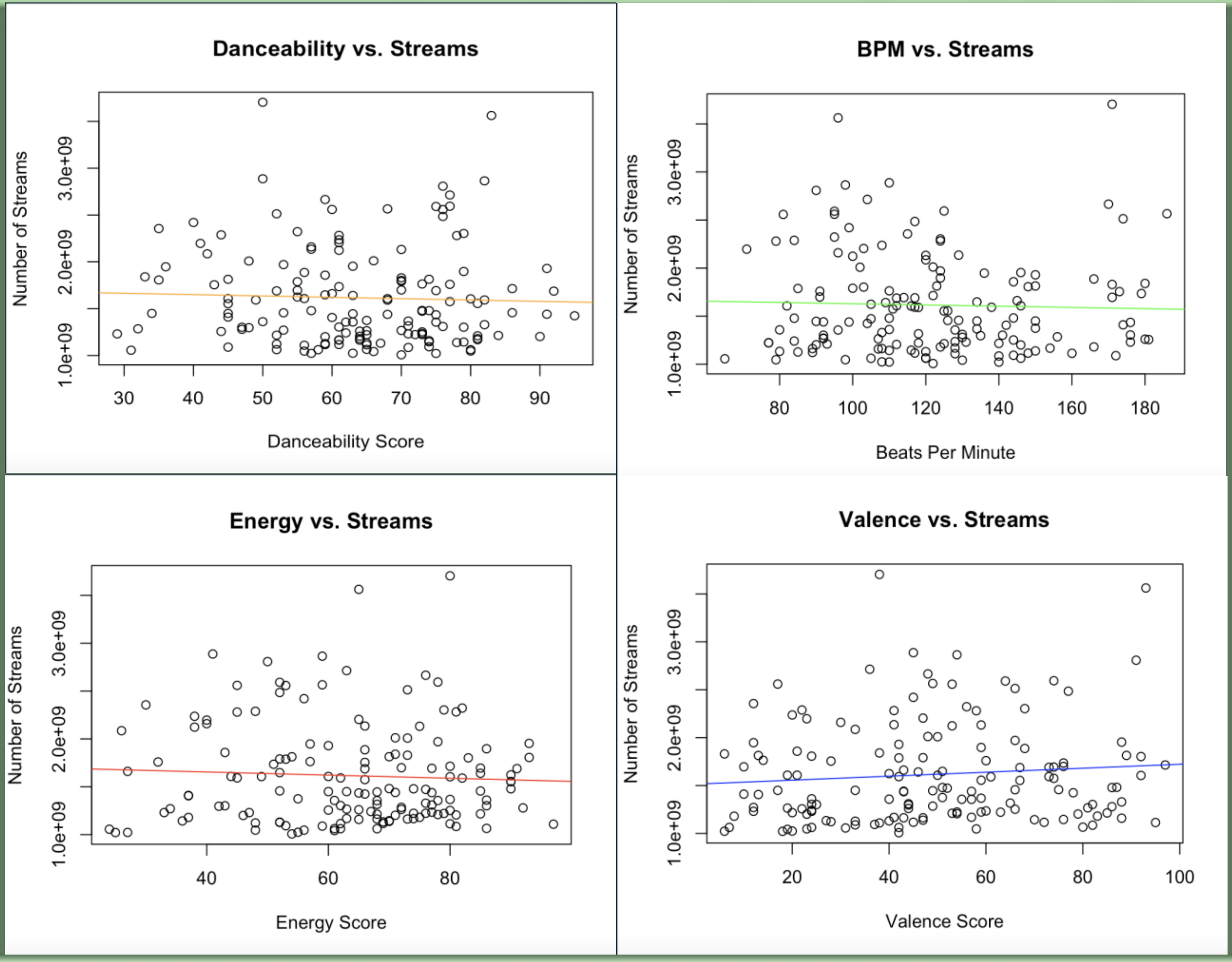
Before plotting data, we had to vigorously clean the dataset from Kaggle.com.

1. Remove unnecessary columns
   - The columns we removed were artist name, key, mode, charts other than Spotify and release year.
2. We also filtered with a forward pipe for songs that had more than 1 million streams
   - A truly popular song would have more than 100 million streams (1), so this filtration left many outliers in the dataset
3. We converted all the values into numeric format rather than character or string

Linear Regression

# Plotting

We plotted each of the eight musical attributes described against the number of streams. However, we decided to remove instrumentalness, acousticness, liveness and peechiness. For multivariable linear regression, one can get a better estimate by limiting the number of variables as discussed in class. We found that these variables were not as important in our analysis and hypothesis about pop songs. Additionally, the p-value of the linear model



Linear Regression

# Descriptive Statistics

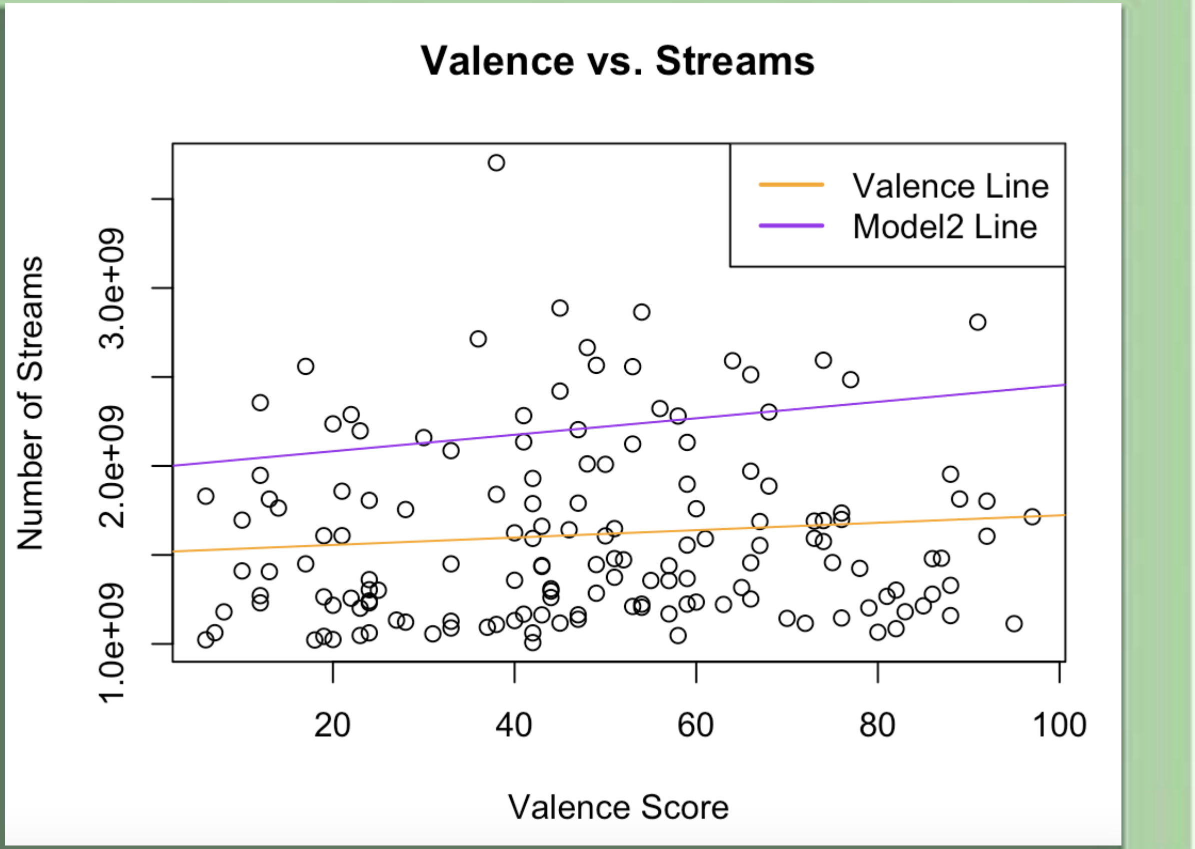
Since our dataset contains the top 1000 songs with the most streams on Spotify, the mean for each musical attribute would be the expected value for this sample dataset. From the previous plotting, the datasets do not seem very skewed. More sampling would be needed to find the true population mean for each of the musical attributes.

| | Average | Minimum | Maximum |
|---|---|---|---|
| Valence (%) | 49 | 6 | 97 |
| BPM (%) | 122 | 65 | 186 |
| Energy (%) | 64 | 24 | 97 |
| Danceability (%) | 64 | 29 | 95 |

Linear Regression

# P-value, Results and Interpretation

After some evaluation, we decided to remove some of the variables to have a more efficient linear model. Specifically, I removed the acousticness and liveliness variables since the p-values were extremely high. With this new model, the p-value for valence was 0.0487. This falls below the common significance value of 0.05. This might indicate that there is statistically significant evidence to reject the null hypothesis for the valence variable. However, it's important to note that statistical significance does not necessarily imply practical significance, and the choice of significance level (e.g., 0.05) is somewhat arbitrary. Therefore, more investigation with different datasets and more sampling is needed to conclude. This is weak evidence to show that there is an association between valence and number of streams.



Shows the model linear regression line against scatterplot of valence and streams

Sentiment Analysis

# Using Sentiment Analysis to test popularity of songs based on mood

We hypothesized that songs with very positive and negative lyricality that maintain a decently high valence (Spotify metadata describing musical positivity on a score of 0-1) tend to be more popular. To calculate this, we quantified the sentiment of song lyrics using the Syuzhet library, which contains a sentiment analysis algorithm. We also created an algorithm that quantified sentiment using the Afinn dictionary. We then made graphical representations of the top and bottom ten songs in positivity, excluding low-valence songs and sorting by popularity. We filtered for pop songs only since the data set used for linear regression was the 1000 most popular songs of 2023 and we wanted the lyrical data we used for Sentiment Analysis to resemble it closely.

Sentiment Analysis

# Sentiment Analysis Algorithms (Syuzhet + Our Homebrewed Algorithm)

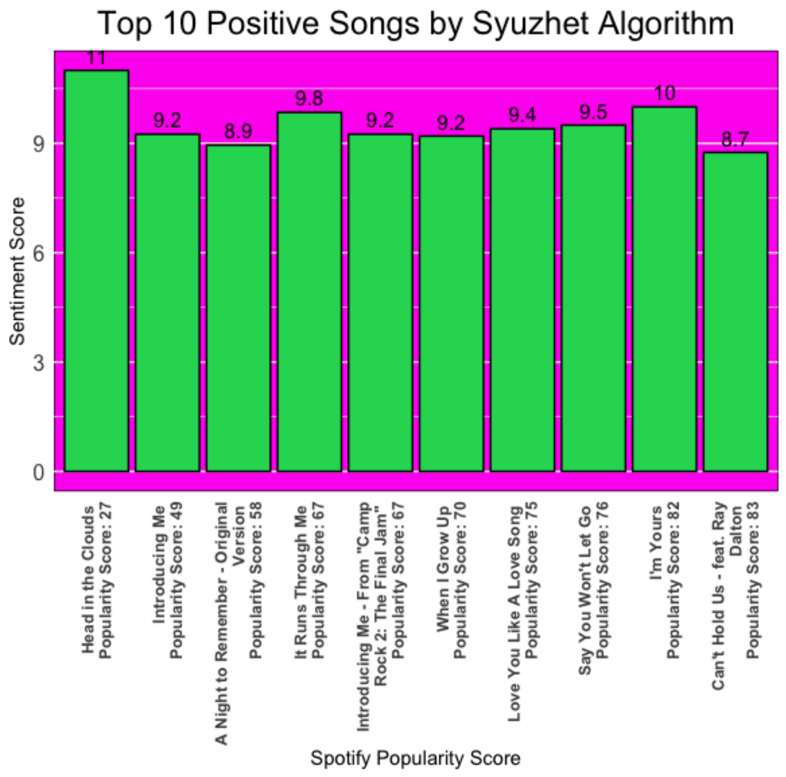
After looking for suitable sentiment analysis models, we came across the Syuzhet library on Cran.r. Syuzhet contains four different Sentiment Dictionaries including three that we learned about in class. It also contains tools to access a Sentiment Extraction tool developed by the NLP group at Stanford, which we ended up using as one of our main tools.

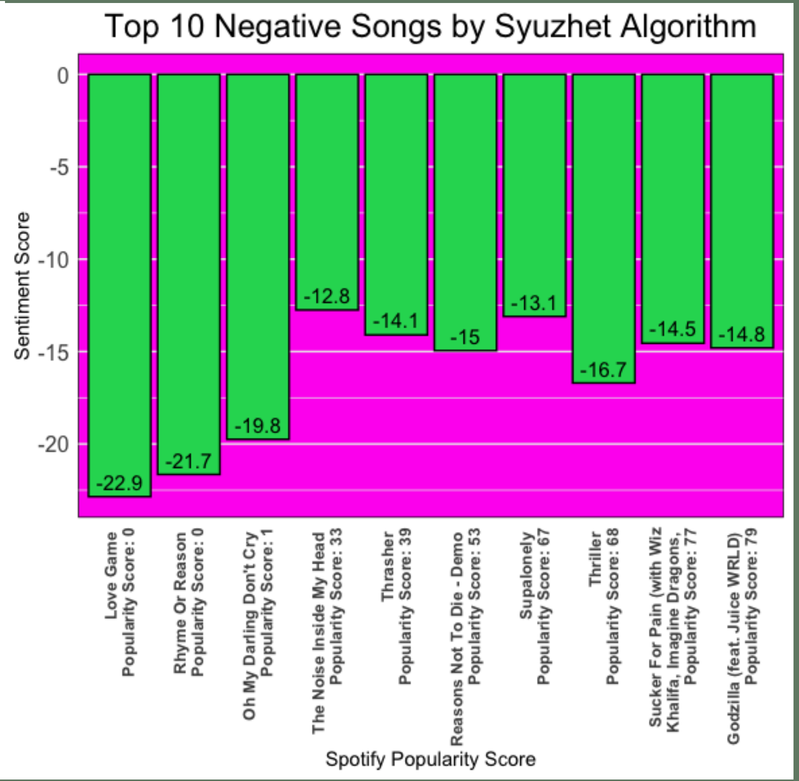
We also created our own Sentiment Analysis algorithm. We created a word counting function and then added a word counter column to the dataset. We then tokenized by word and calculated the Afinn score for each word. We decided Afinn would be best since it calculates quantitatively, the same as the tools created by the NLP group. We then repeating words that were adjacent to each other with a for loop (to remove choruses that skewed data) before grouping it back to songs after performing a sum function on the Afinn sentiments, creating scores unique to each song.

Finally, we created graphics of our findings from both models.

Sentiment Analysis

# Syuzhet findings


Top 10 Positive Songs by Syuzhet Algorithm


Top 10 Negative Songs by Syuzhet Algorithm

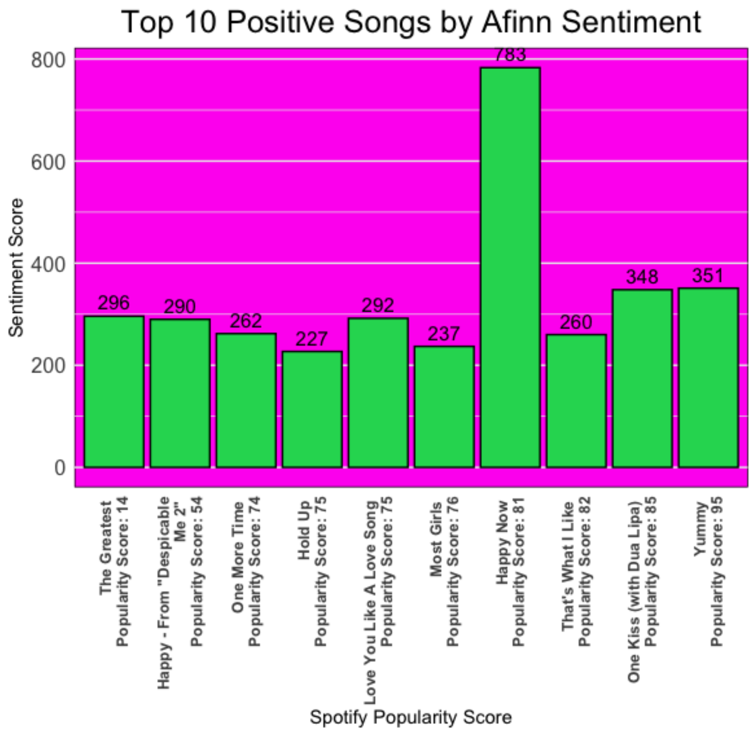
The Syuzhet Algorithm created sentiment scores with a range of 34, with the songs skewing towards negativity. The Positive songs were overwhelmingly popular with a median of 68.5 while the negative songs, while not sharing the same levels of popularity still skewed towards being popular with some notable outliers, causing the median to be 46.

The popularity scores provided by Spotify range between 0-100 and consider the number of streams as well as their recency

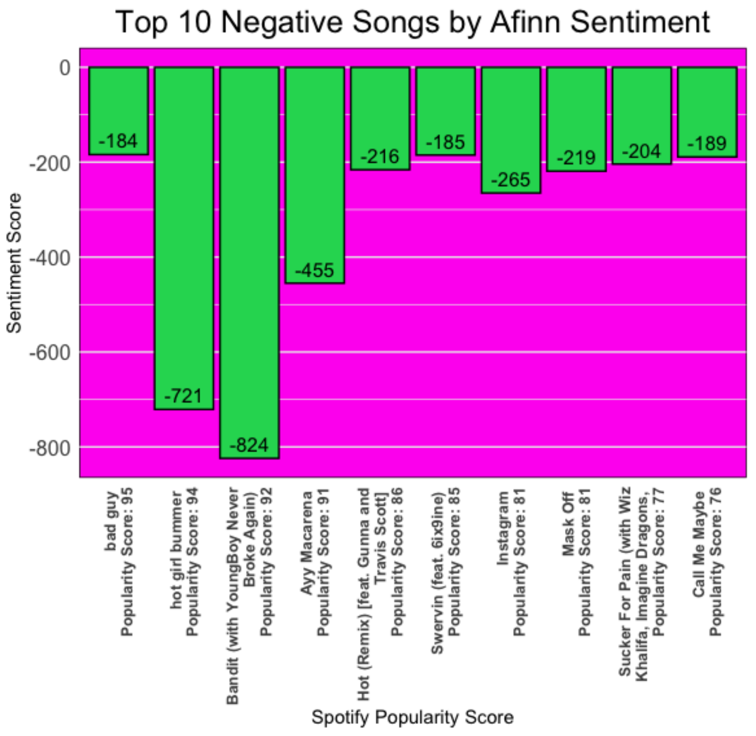
Sentiment Analysis

# Afinn findings

The Afinn algorithm was made by our group, where we used the Afinn Dictionary to sum the sentiments of the lyrics after cleaning them and removing the choruses. This gave us a numerical score for each song with a total sentiment range of 1607. The positive songs had a median of 75.5 in popularity while the negative songs had a 85.5, showing that they are very popular. The data and analysis reveals that the more positive or negative a song is, the higher the likelyhood that it is popular.



Top 10 Positive Songs by Afinn Sentiment



Top 10 Negative Songs by Afinn Sentiment

The popularity scores provided by Spotify range between 0-100 and consider the number of streams as well as their recency

In Summary

# Conclusion

While there were some limitations of the datasets that we employed, our evidence strongly indicates that songs that are "happier" are strongly correlated with the number of streams.

In the multi-variable regression: "Danceability" and "Liveness" both had P values within the threshold to accept them as positively impacting the number of streams.

In the single-variable regression: Valence was approaching the statistical threshold to accept it as positively impacting the number of streams. (0.0506, where 0.05 is the significance level).

In sentiment analysis, there is an overwhelmingly clear pattern of songs to the extreme ends of positivity or negativity lyricality end up being popular. Between the two analyses, positive songs are more reliable with them being more likely to end up popular.

Spotify Metadata Used

# Resources

1. Spotify Popularity Index Definition: https://chartmasters.org/spotify-most-popular-artists/
2. Syuzhet Library: https://cran.rproject.org/web/packages/syuzhet/vignettes/syuzhet-vignette.html
3. Spotify Popular 2023 Songs Dataset: https://www.spotify-song-stats.com/about
4. Spotify Lyrics Dataset: https://www.kaggle.com/datasets/imuhammad/audio-features-and-lyrics-of-spotify-songs
5. Spotify metadata definitions: https://www.spotify-song-stats.com/about