For our research project, we asked the question: What are the factors that make a song popular? To answer this question, we created and trained linear regression models and also made use of sentiment analysis using the Afinn dictionary. Both approaches made use of datasets that contained several columns of Spotify metadata, each of which contained a specific characteristic of a song that the preexisting Spotify algorithms had quantified down to a percentage point. They are Acousticness (confidence that the track is acoustic), Danceability (suitability of the track for dancing based off several internal metrics), Energy (Intensity and Activity), Instrumentalness (whether the track contains no lyrics), Liveness (whether the track was recorded with a live audience), Loudness (loudness in dB), Popularity (a metric based off recency of streams), Speechiness (amount of the track that is spoken), Valence (musical positivity of a song). We found the average for each these values after cleaning our dataset and removing outliers. Since all the songs in the dataset are popular, the mean would show the most expected value for each attribute for popular songs.

We hypothesize songs with attributes to pop genre be more popular. We measure popularity using streams and popularity (a Spotify metadata category that combines stream count as well as the recency of the streams to assign songs an exponential value between 0-100). Pop genre songs tend to have positive lyrics, and have high energy, danceability, and bpm. We wanted to work on a topic we could conduct both quantitative and qualitative analysis on.

Linear Regression:

We used Linear Regression to measure the objective musical qualities of each song and track which qualities align most with the most popular songs. This was the quantitative analysis of song attributes using a dataset from Kaggle.com. Linear regression provided an easy way to determine different relative weights of variables in each song. We ran this code by determining which key variables recorded in each song would play a role in determining how popular a song was. We cleaned the dataset

to remove outliers (songs under 10,000,000 streams) and converted everything to numeric. Using the cleaned data set, we tested variables such as BPM and discarded variables such as “in_apple_music.” We then set two variables, X and Y, as data for our linear regression. Y represented our dependent variable, which was the number of streams, and x represented the variables which we believed impacted Y. They include: “danceability”, “valence”, “energy”, “acousticness”, “instrumentalness”, “liveness”, and “speechiness”. We also cleaned our data by converting the “streams” variable from a string of characters to an integer using gsub. I used the set.seed function to guarantee the reproducibility of random results. We trained indices, and set our proportions to 80/20, using 80% of our data for testing and 20% for training, as is the traditional standard in machine learning. We then created a basic linear regression model and used our training data as the variable.

Next, in order to effectively communicate our findings, we graphed our results using ggplot2. We set x to actual and y to predicted and graphed a line of best fit through the actual vs predicted values on the x and y axis. Our results indicated a key finding: if every value was maxed out the predicted value for the number of streams would be billions more than the actual number of streams. This is because the maximization of some of these variables would be incongruent with the maximization of other variables. For example, it would be impossible to maximize speechiness, a variable which determines how much of a song has spoken words, and acousticness, which determines how acoustic a song is. We did, however, determine that there were two variables in the multivariable regression which had P-values of less than 0.05: Danceability and Liveness. Their P-values both indicate that they have a positive impact on the number of streams that a song has. When evaluating variables individually, we also determined that valence was approaching the bound of the metric for impacting songs. The p-value was 0.0006 larger than it should have been to draw this conclusion.

We also did a traditional linear regression model with valence_. + energy_. + bpm + instrumentalness_. + danceability_ against number of Spotify streams. All of the characteristics except

valence had an insignificant p-value. Since our null hypothesis is that there is no association between musical characteristics and streams, a statistically significant p-value would give evidence to reject the null hypothesis. Valence gave a p-value of 0.0506 which is on the borderline for significance. This could be slight evidence against the null hypothesis, or for the belief that there is an association between valence and Spotify streams. To make a stronger conclusion, more investigation is needed with other datasets or songs.

Sentiment Analysis:

Next, we decided to conduct a sentiment analysis of song lyrics to understand whether the positivity or negativity of the lyrics in a song could affect its popularity. We needed to use a different dataset for sentiment analysis since the dataset we were using for linear regression did not have lyrics data. We ended up using a dataset that contained 18,000 songs and their lyrics. To ensure that it was as similar as possible to the previous dataset of popular songs, we filtered the genre to only include pop, and then removed duplicates, and filtered for English songs only. It also didn't have a streams field and instead had a popularity field. Popularity is a Spotify metadata field that quantifies a song's popularity on a scale of 0-100 by taking into account total streams as well as the recency of these streams. We also wanted to exclude songs with extremely low word counts since they'd be impossible to conduct sentiment analysis with, so we created a word count function that made use of regular expressions to create a field for word count and then remove any rows where the word count was less than 30. This concluded the data cleaning of the lyrics dataset.

We found a library on the internet called Syuzhet that contains Sentiment dictionaries including all three we learned about in class. It also contained a Sentiment processing algorithm that was created by the Natural Language Processing group at Stanford. We applied this algorithm to the lyrics dataset and added the sentiment score of each song as a column to the overall dataframe. We then subsetted the

top and bottom 10 and filtered out songs with a valence below 0.2, since it corresponds to the musical positivity. We then graphed them using the ggplot2 library, with the x being the songs themselves, the y being the sentiment score, and the order being their popularity. After using the predefined algorithm, we decided to use the Afinn dictionary we learned about in class with our reasoning being that it would be the best for quantitative analysis. We took the cleaned data and tokenized it, and then used a for loop to remove any word at index i that is equal to index i-1. This was done to remove the chorus from songs, which can often skew the data. We then summed the scores and assigned each song a song sentiment score, which we then graphed the top 10 most and least positive songs using ggplot2 with the x being the songs themselves, the y being the sentiment score, and the order being their popularity. We found that a significant portion of these songs were popular, with the median of 3 of the 4 graphs made being over 50 (Spotify popularity score is calculated exponentially, with Spotify heavily promoting songs that have a popularity score greater than 30). Both positive graphs had a median over 50, which asserts that there is at least some tangible correlation between the strength of a song's lyrical emotions and its popularity.

Ultimately, we assert that there is significant evidence to conclude that upbeat songs are more popular given that variables like: "liveness", "danceability", and "valence" all were within the statistical range for assuming that they had a positive impact on the number of streams. Additionally, given the sentiment scores of the top 10 most popular songs, it is a foregone conclusion that happiness has a tacit impact on a song's popularity. While negative songs should not be discarded, as our research in sentiment analysis indicated that songs on either end of the spectrum were more likely to be popular, the "happiness aspect" of a is a key variable when estimating popularity

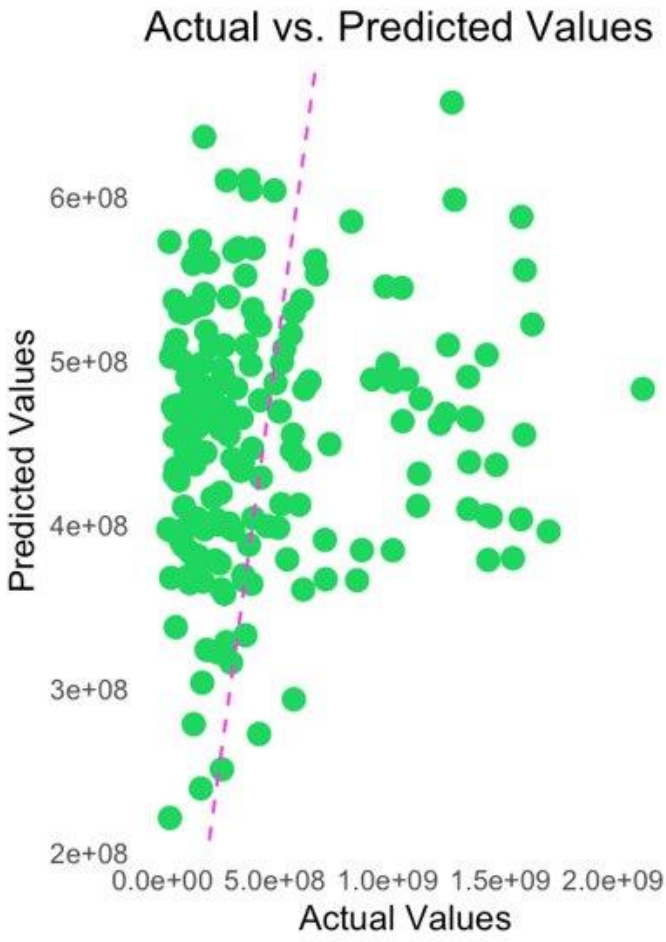
Division of Labor

David- I did a multi-variable linear regression running all the pertinent variables against a target variable which I made the total number of streams that each pop song. I did this by cleaning my data, which was no small undertaking, creating target and test variables, then creating testing data, and using the lm function to create a linear regression. I discovered that in the multiple variable regression danceability and liveness both had P-values which were small enough to justify them as having a positive impact on the number of streams that different songs had. I then used ggplot2 to graph my findings, making sure my work was color coded to match Spotify. Also, I created the slides pertinent to my part of the project and the conclusion to bring everything together.

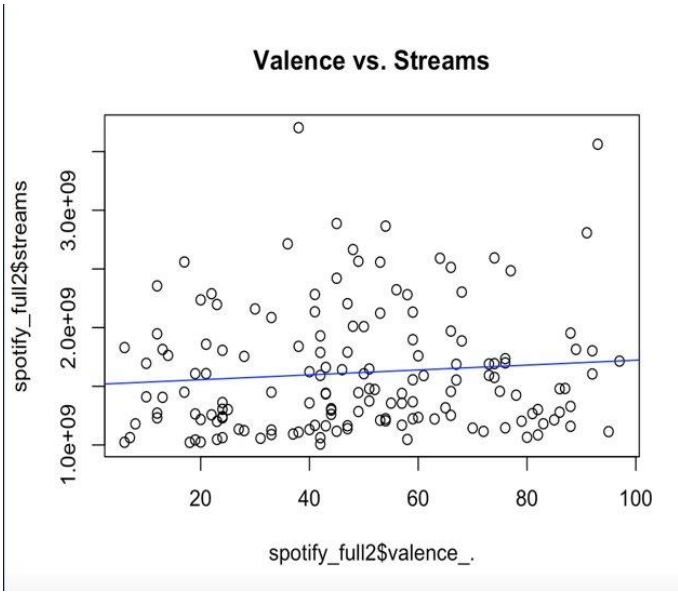
Shreya – I extensively cleaned that dataset. I removed variables that were unnecessary and there were many outliers or values with flaws. I removed these rows entirely and made was entirely and made sure there were no outliers to skew the dataset. I did the linear regression plotting and found the p-values. I also found the descriptive statistic lots of brainstorming for what analysis should be done for the linear regression and extent of what interpretation can be made from our results.

Prithvi- I found the lyrics dataset and the Syuzhet library and then conducted the sentiment analysis on the dataset that I cleaned with the syuzhet library's algorithm as well as creating our own Afinn analysis and then generating graphs for each using ggplot2. This all amounted to approximately 200 lines of code. I then worked on the Sentiment Analysis sections in the Power Point and Writeup. I also designed the graphical layout of the final presentation.

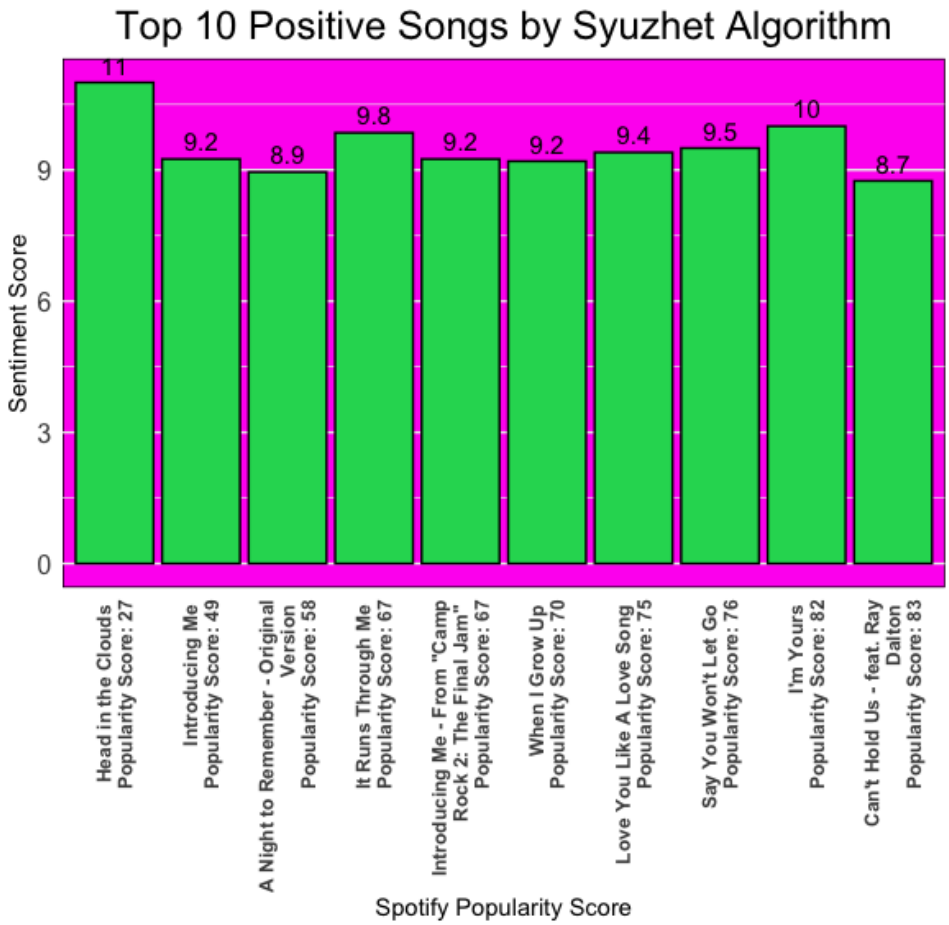
Pertinent Graphs:



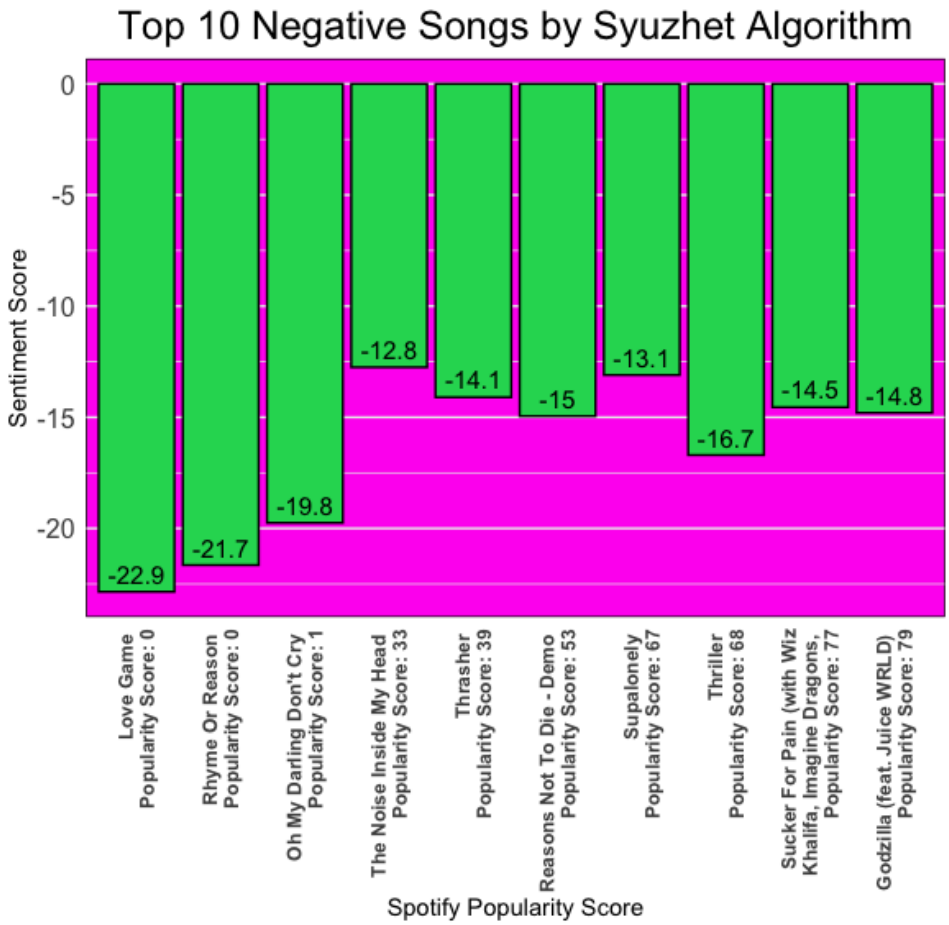
Predicted streams vs actual streams in multi-variable linear regression (David)



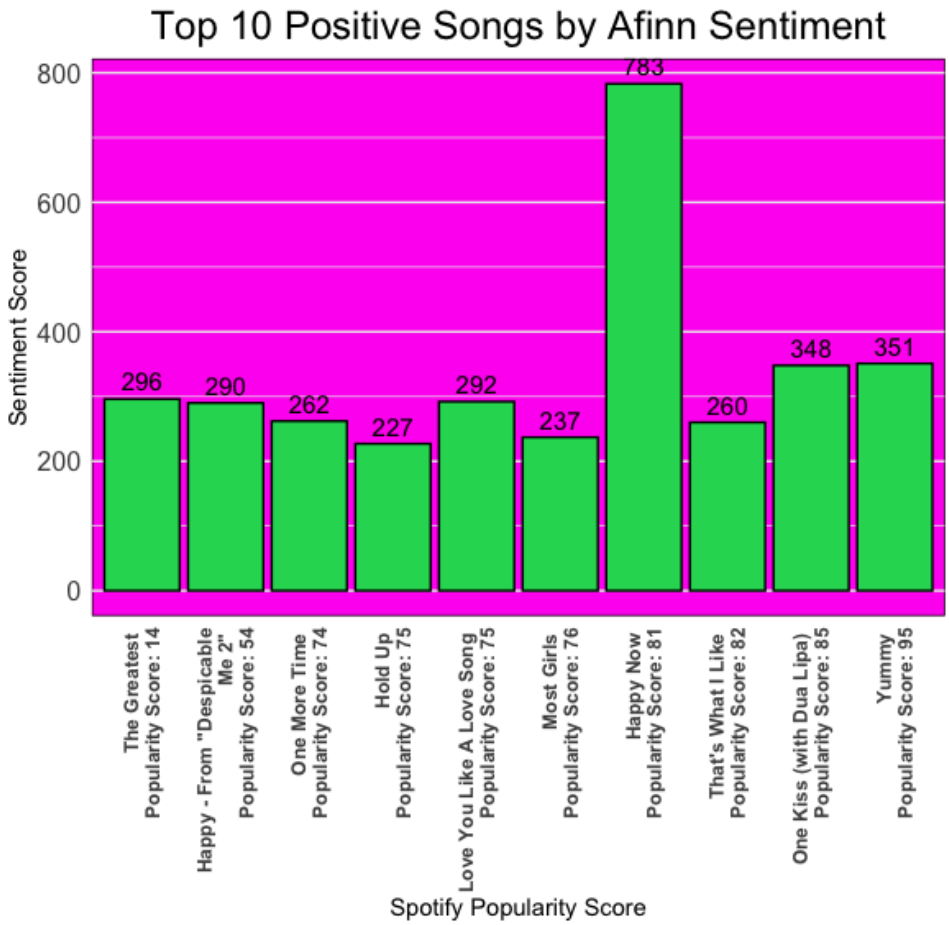
Shreya – valence vs streams



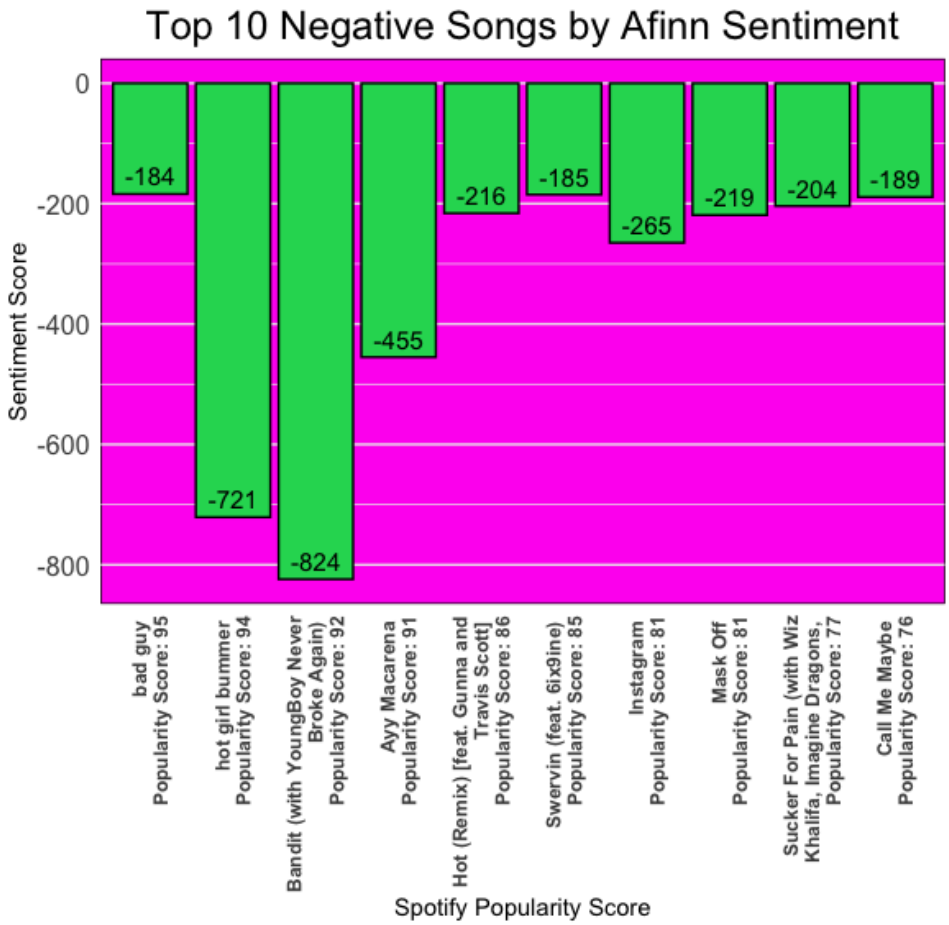
Top 10 songs using Syuzhet algorithm



Bottom 10 songs using Syuzhet algorithm



Top 10 songs using the method we created using Afinn



Bottom 10 songs using the method we created using Afinn

Sources:

Syuzhet Library: https://cran.r-project.org/web/packages/syuzhet/vignettes/syuzhet-vignette.html

Spotify Popular 2023 Songs Dataset: https://www.spotify-song-stats.com/about

Spotify Lyrics Dataset: https://www.kaggle.com/datasets/imuhammad/audio-features-and-lyrics-of-spotify-songs

Spotify metadata definitions: https://www.spotify-song-stats.com/about

Spotify Popularity Index Definition: https://chartmasters.org/spotify-most-popular-artists/