# Examining the Keys to Successful Music

Jasmine Kellogg | Jake Kaihewalu

INT 15 - Spring 2019 Final Project

Professors Kate Kharitonova and Alexander Franks

June 14th, 2019

## Abstract

Music is such a huge factor of popular culture. Anywhere you go you hear music coming from cars, in stores, on the radio, the list goes on. Part of what makes music such a driving factor is it's unique ability to evoke our emotions. When you find that one song or album that just makes you feel something, happy or sad or any other strong emotion, no amount of plays can feel like enough. In this project, we explored a variety of musical features in an effort to uncover patterns that make songs successful in pop culture and in critical reviews. Using PCA, we found that the top components were associated features such as danceability, valence, and energy and play a role in the score received by critics. We also found that the language used by critics to review albums can be used to loosely predict popularity but generally reviews from critics do not align very well with being able to predict what ends up on the top album charts. As the old saying goes, beauty is in the eye of the beholder and in this case, good music is in the ears of the listener.

## 1. Introduction

Because there are so many factors, internal and external, that contribute to a song or album's success, our goal was to analyze a case from each. Internally, we looked at acoustic attributes of songs to identify features with the most effect on critic ratings. Externally, we analyzed album review sentiment to see how influential (if at all) the language critics used in describing an album was associated with the overall score. Next, we sought to answer the question of how accurately we could fit a model to predict an album's success on the charts public based on the language used in the reviews from critics.

The 'acoustics' dataset contains various features used in musical composition such as acousticness, tempo, key, valence, instrumentalness, etc, as well as standard measurements such as duration of a song in milliseconds, and time signature of a song. A majority of these features are scaled from 0 to 1, while 'mode' is binary with major as 1 and minor as 0, tempo values fall on a scale of about 50 to 200, loudness values range from -60 to 0, and finally key which has Integers mapped to pitches using pitch class notation. These features assess the composition of songs and certain features such as valence and danceability can be used to explain why a certain song is so catchy and why people respond well to it.

In order to measure album reception we used data from two different sources. The dataset titled 'reviews' gave us ratings and reviews written by music critics from popular online magazine *Pitchfork*. We also used the dataset 'albums' to get information on the popularity of the song. This dataset gives weekly chart standings for the Billboard Top 200, a record chart that ranks the 200 most popular music albums across all genres. Rankings are derived from sales and streaming activity.

After deciding what directions we wanted to explore we initially set out to find out if any related work had been done. Because Spotify has made it very easy for virtually anyone to use their API to collect acoustics data, there are many different reports that people have constructed that explore what can be said about their music from the attributes. One of the more popular ones that comes up in a quick search on the web titled "What does your Spotify music sound like? Data Science with Spotify" by Alvin Chung explores a similar idea to ours by analyzing potential patterns of songs that chart on the top 100 list (Chung, 2018). His findings suggest that the factors 'danceability' and 'speechiness' are significant in a songs popularity (Chung, 2018). This project indicated that our questions were pointed in the right direction and motivated us to dive deeper into exploring album success and comparing different measures of this 'success'.

## 2. Data and Methods

### 2.1 Initial Data Analysis

The dataset for acoustics came from Spotify's collaboration with Echonest in which Echonest broke down the metrics for various songs. The datasets for reviews and albums came respectively from csv files containing written reviews and scores for albums, and the billboard top 200 by week. Assessing the principles of measurement for each of the datasets, relevance of data and cost of measurements are not of concern. Each dataset contains useful information that can be used to answer our questions. Cost is also not relevant as there aren't any ethical or social issues at hand and the data is open to the public and thus, doesn't cost anything. Precision and distortion however are relevant principles of measurement. Because music *is* such a subjective topic of study, it is impossible to get 100% precision on predicting music scores by acoustics alone. Distortion is also a concern as the written reviews and scores in the reviews dataset are by humans who are inherently biased in their opinions. The various reviews are also written by different critics and because each critic has their own take on the quality of an album, the reviews and scores an album receives could be distorted based on who reviewed it.

In our initial exploration of the datasets, we checked for missing values in each. In the reviews dataset, every variable except for genre had few or no missing values while genre had 2305. The acoustics dataset showed 5 observations with missing feature values, and 51 observations with the name of the artist missing. Finally, the albums dataset contained the most missing values, with 65 artist values missing, 81400 missing length values, and a staggering 105080 missing track_length values. We decided to drop some variables in the that were not relevant to answering our questions. For our analysis of acoustics and review polarity and their effect on scores, we merged the two datasets by album and artist and took the means for each of the features to get a cleaner, organized dataset with acoustic measurements for each album and their corresponding review polarity.

### 2.2 Methods

The first text mining method in our analysis was used in order to get a metric for the overall polarity of each review from *Pitchfork.* To do this we chose to use VADER (for Valence Aware Dictionary for sEntiment Reasoning) lexicon because it is a model that uses an "empirically validated gold standard master lexicon" that is accurate and simple to apply (Hutto & Gilbert, n.d.). This allowed us to quickly calculate a numerical value for the review polarity and incorporate it in our analysis along with the acoustic features and see how these features were related to critic rating.

After creating a metric for polarity of an album's review and adding it to our list of features, we did exploratory analysis on the relationships (if they existed) between the various features. Because there were so many unique values for the score variable, we grouped rating scores into 5 categories using a custom classifier. After we grouped our scores, we created a general pairplot that compared each feature to each other feature and examined the resulting scatterplots for linearity. In addition to the pairplot, we created a correlation matrix that calculated the coefficient for each pair of features. We used these plots to decide what type of analysis we wanted to do going forward, and, seeing a lack of linear relationships, we opted out of using a linear regression model and ultimately decided on PCA for dimension reduction, and fitting a model using the Random Forest Classifier in the sklearn package. We felt that because there were a large number of features to be analyzed as well as lack of linearity, PCA would be a good method to use to identify key features that affect rating scores. Before we completed PCA, we did a transformation on the features to normalize them. Specific variables such as loudness, polarity, and tempo had very large scales. If we did not do this preprocessing, our results from the dimension reduction would have been inaccurate. After standardizing the data, we were able to complete our PCA and examine our Random Forest fit.

In order to try and predict whether or not an album would appear on the Billboard Top 200 from the language used in the reviews we used a second text mining approach with the Naive Bayes Classifier from the Natural Language ToolKit (NLTK) package. This classification model uses a training set and calculates a prior probability for each label (keyword) of the training set from it's frequency (Bird, Klein, & Loper, 2009, pp. 245). Then, each keyword's contribution is measured by how many times it occurs in each feature or defined category (Bird, Klein, & Loper, 2009, pp. 245). A likelihood estimate for each keyword is then constructed by multiplying the prior probability with the contribution and these likelihood estimates of each keyword form the model that is then able to classify a given set of keywords (Bird, Klein, & Loper, 2009, pp. 245). We chose this model because it is a simple yet accurate supervised machine learning model that is easy and efficient when working with large sets of data like we have with the lengthy *Pitchfork* reviews.

## 3. Results

### 3.1 Pitchfork Review Sentiment Analysis

The initial evaluation of our results from the VADER lexicon sentiment analysis are based on simple summary statistics and related plots. The polarity values for our dataset ranged from -131.1 to 139.9 and had an approximately normal distribution centered on the mean of 25.59 seen in Figure 1. A majority of the review sentiment was positive which aligns with the overall distribution of the critic scores that had over half of the reviews in the 6.1-6.8 range. Figure 2 shows the relationship between polarity and critic score which rules out the possibility linear relationship between the two values. Instead, the scatterplot loosely suggests that the higher scored albums had more extreme values of polarity. This indicated to us that the results of model were most likely not representative of the opinion of the critic.

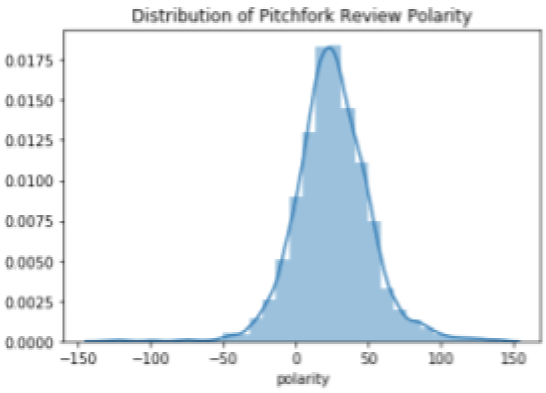


*Figure 1: Univariate distribution of polarity values.*

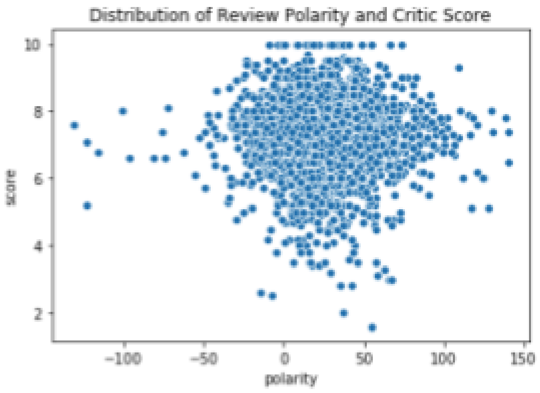


*Figure 2: Scatterplot showing relationship between polarity and critic score.*

Analysis of the most negative review sentiment derived from the VADER model provides more insight on what could be a better interpretation of the sentiment values. The album "Until

Death Call My Name” is a Rap album by the artist YoungBoy Never Broke Again that received a good critic rating of 6.9 while also having the most negative sentiment analysis. Looking into the actual review we see that much of this can be attributed to the critics focus on discussing violence that appears in the artist's songs as well as crimes and other negatively associated topics. This brings up potential issues in using this method of tokenized sentiment analysis in an attempt to relate the review sentiment to score and highlights the difficulty in having a simple model like this discern between the opinion and the subject matter that the critic is discussing.

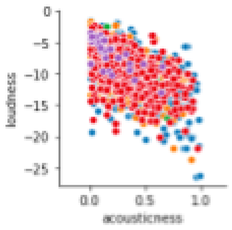


*Figure 3:*
*Pairplot showing relationship between loudness and acousticness.*

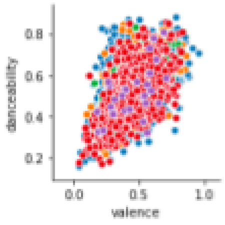
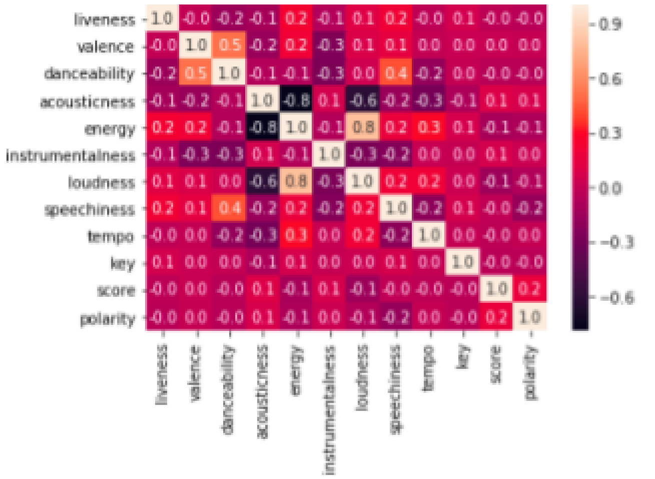
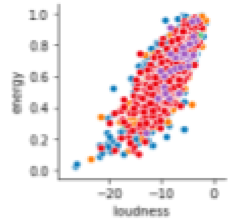


*Figure 4:*
*Pairplot showing relationship between valence and danceability.*





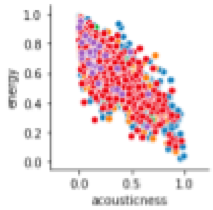


*Figure 7:*
*Correlation matrix displaying the relationships between each pair of features.*

*Figure 6:*
*Pairplot showing the relationship between energy and loudness.*

*Figure 5:*
*Pairplot showing the relationship between acousticness and energy.*

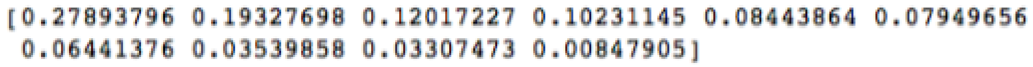
### 3.2 Pairplot Examination

After adding polarity as a feature to the dataset and reviewing the general pairplot which contained all relationships between features, we noticed four relationships that displayed linearity, while the rest showed little to no correlation.

Looking at Figures 3 and 5, we see negative relationships between acousticness and loudness and acousticness and energy, both of which are understandable as they describe opposite aspects of music. In Figures 4 and 6 we also see positive relationships between loudness and energy and valence and danceability. In order to confirm the relationships above, we used a correlation matrix as shown in Figure 7. As suspected each of the relationships above exists; their correlation coefficients are shown in black and orange boxes below.

### 3.3 Score Classification using Principal Component Analysis (PCA)

After examining our pairplots and assessing linearity, we decided that there weren’t enough linear relationships between the variables to warrant a linear regression model. So we moved forward with PCA. After transforming all variables to normal distributions and splitting the data into test and training sets (80% training, 20% test), we performed PCA on 10 features in the dataset - liveness, valence, danceability, acousticness, energy, instrumentalness, loudness, speechiness, tempo, and key, with the target feature as score label. Using the explained_variance_ratio_ function of our PCA model, the fraction of variance explained by each primary component is shown in Figure 8 below.



```
[0.27893796 0.19327698 0.12017227 0.10231145 0.08443864 0.07949656
 0.06441376 0.03539858 0.03307473 0.00847905]
```

*Figure 8:*
*Fraction of variance explained by each component in our PCA model.*

To get a better visualization of the fractional variance, we plotted a scree plot as well as a cumulative sum of variance plot for the 10 primary components.

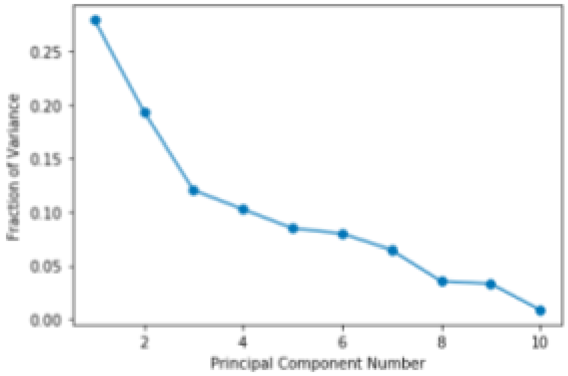


*Figure 9: Fraction of variance explained by each principal component.*

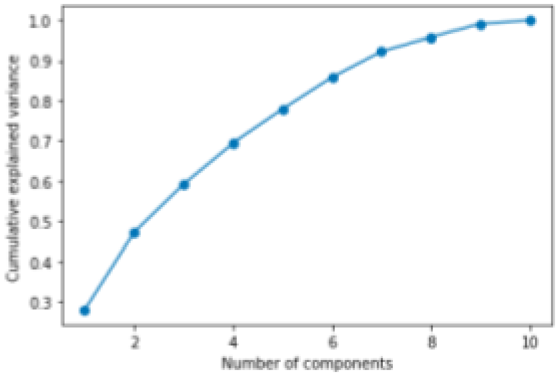
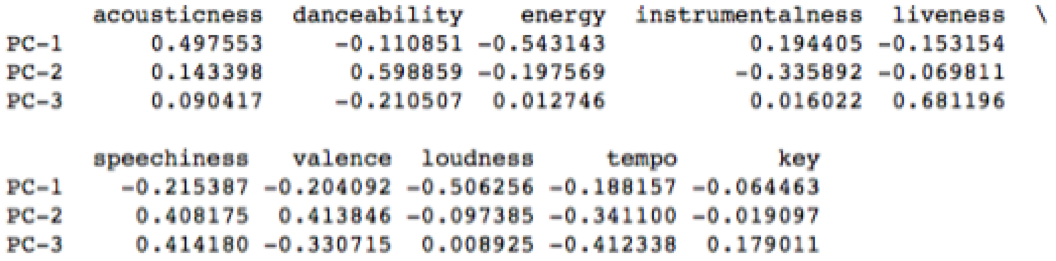


*Figure 10: Cumulative explained variance for each successive primary component.*

As shown by Figure 9, the scree plot elbows at PC3, indicating that after the 3rd component, the remaining components do not account for much of the variance. Figure 10 displays the cumulative explained variance. With only 3 PC's, we can explain over half of the variance in the data. The plots shown above correspond to a PCA analysis of acoustic features alone. We originally included polarity scores as a feature, but after running PCA *with* polarity, our accuracy decreased by about 4 percent. So, we omitted polarity from our final model. We ultimately decided on a model with 3 PC's. Not only does the scree plot elbow after the 3rd component, but in multiple trials, we altered the number of components in the model and found that our accuracy stopped increasing after the addition of the 3rd component. After settling on PCA with 3 components, we took a look at each feature's correlation with the top 3 components.



```
      acousticness  danceability    energy  instrumentalness  liveness  \
PC-1      0.497553     -0.110851 -0.543143          0.194405 -0.153154
PC-2      0.143398      0.598859 -0.197569         -0.335892 -0.069811
PC-3      0.090417     -0.210507  0.012746          0.016022  0.681196

      speechiness   valence  loudness     tempo       key
PC-1    -0.215387 -0.204092 -0.506256 -0.188157 -0.064463
PC-2     0.408175  0.413846 -0.097385 -0.341100 -0.019097
PC-3     0.414180 -0.330715  0.008925 -0.412338  0.179011
```

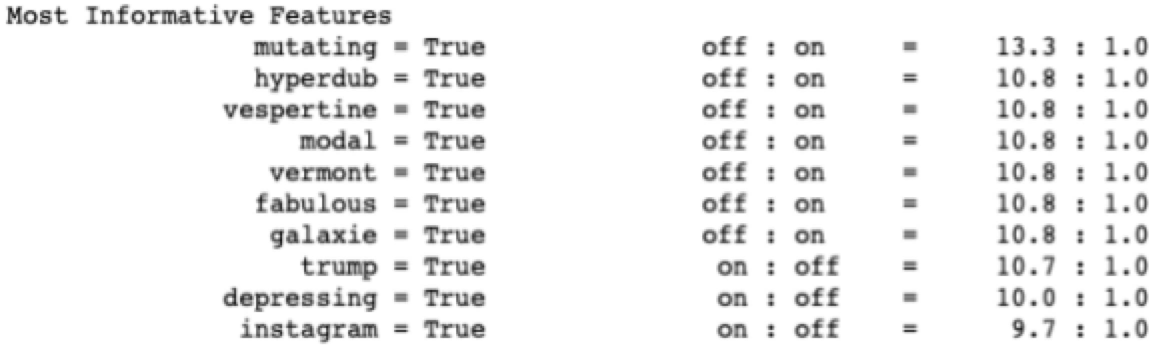
*Figure 11:*
*Feature correlation with each PC*

Figure 11 shows that for PC1, acousticness, energy, and loudness are the most correlated features. For PC2, danceability and valence are the most, and finally, for PC3, liveness, tempo, and speechiness are the most. We can conclude that features such as the ones listed for the first few PC's have the strongest effect on rating scores, while features such as instrumentalness, key, and tempo have little to no effect. This analysis was interesting as we expected features such as valence to have the highest correlation and acousticness to have rather low correlation. Nonetheless, this dimension reduction is useful as it allows us to gain insight about our data using very few components.

As we mentioned previously, we settled on 3 components for our final model after we tested our model with various numbers of components. The model we used was the Random Forest Classifier which used a combination of decision trees on samples from the training dataset to build a predictive model. Running the classifier on our PCA model multiple times, with a different number of primary components in each PCA model, we got the highest accuracy with 3 components at about 56 percent. Increasing the number of components after 3 did not affect our accuracy score. Our score was definitely not as high as we wanted it to be, but it allowed us to draw conclusions about rating scores. We concluded that while acoustic features do play a role in determining an album's rating score, they don't play as big of a role as we thought prior to doing this analysis. Our accuracy score of 56 percent indicates that other factors not accounted for in this project play a big role in an album's reception.

**3.4 Predicting Album Chart Success from Pitchfork Reviews**

In search of a more accurate method of analyzing the review sentiment to predict success of an album we decided to train our own Naive Bayesian Classification model from the NLTK package. Popularity in this case is indicated by the occurrence of the album at any spot on the Billboard Top 200 chart since 1963 (none of the albums from the reviews dataset were released before then). This gave us a binary classifier for our model to predict and after running the comparison 65.5% of the albums in our dataset have appeared on Billboard Top 200.



```
Most Informative Features
                 mutating = True              off : on     =     13.3 : 1.0
                 hyperdub = True              off : on     =     10.8 : 1.0
               vespertine = True              off : on     =     10.8 : 1.0
                    modal = True              off : on     =     10.8 : 1.0
                  vermont = True              off : on     =     10.8 : 1.0
                 fabulous = True              off : on     =     10.8 : 1.0
                  galaxie = True              off : on     =     10.8 : 1.0
                    trump = True               on : off    =     10.7 : 1.0
               depressing = True               on : off    =     10.0 : 1.0
                instagram = True               on : off    =      9.7 : 1.0
```

*Figure 12: Sample list of the top 10 most informative key words in the model and their corresponding frequencies in their indicated categories of "on" The Billboard 200 vs. "off".*

After cleaning normalizing the raw review strings and forming the test and training data from the bags of words for each class, the model was trained and tested. The model had a consistent accuracy of around 70% which indicated to us that it had done a moderately good job at classifying based on the language used in the review. Figure 12 above displays the top 10 influential words for that round randomly chosen training data. The first 7 words with "off : on" are significant key words that showed up many times more in reviews of albums that did not make it on the Billboard Top 200 list. These initially seem like arbitrary words but upon closer analysis we found some ways to interpret why these words in particular were found to be informative. For example, “hyperdub” may refer to the London-based electronic record label and its occurance of 10.8 times more in non-charting album reviews could be attributed to its low popularity in the context of popular music. There are three informative words “trump”, “depressing”, and “instagram” that show up 9.7-10.7 times more in *Pitchfork* reviews that have charted versus album reviews that have not. Generally, each of these words can be associated with popular subjects in modern pop culture and thus makes sense that these could show up more in discussions of albums that include these topics that have also charted. It is important to note, however, that none of these keywords can be easily associated to opinion besides “fabulous”, which counterintuitively appeared more on albums not on Billboard Top 200.

## 4. Discussion

After completing our analysis, we unfortunately were not able to fully achieve one of our goals, which was to predict the score an album received based on acoustic features and polarity of its review. We found that while acoustic features *do* play a part in the score an album receives, they are not the only factors that play a role. One of the challenges we faced was the uneven score data. A large portion of the albums included in the reviews dataset scored in the middle-upper range. Had the data been more evenly distributed, with more scores in the lower range, we may have gotten different results. The acoustic features were also weakly correlated, which made fitting a successful model difficult. Future work could include acoustics as well as data that shows how often an artist or album name appears in social media as publicity is a big factor in the reception and buzz of music. The combination of these two could provide more accurate insights and models for predicting rating scores.

Another one of our primary conclusions besides the acoustics modelling had to do with the ability of text mining of the reviews to give insight on the success of an album. From our results we conclude that VADER lexicon is not sufficient in capturing the opinion of album reviews and that the Bayesian Classifier that was created has the potential to be a moderately reliable way to get information on the popularity of an album. However, this model has a lot of room for improvement before we can confidently say that it captures the opinions of the critics. The lack of influential words we would normally associate with analysis of the album from the critics can be attributed to a variety of factors. It could indicate that there is in fact a dissociation between critic opinion and the acceptance of the general public. Future work could make use of techniques that improve accuracy such as n-grams that relate chunks of words of text in order to make more concrete conclusions.

## References

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media, Inc.

Chung, A. (2018, September 27). What does your Spotify music sound like? Data Science with Spotify (Part 1). Retrieved June 14, 2019, from Towards Data Science website:https://towardsdatascience.com/data-science-and-machine-learning-with-spotify-841225bfb5d0

Hutto, C. J., & Gilbert, E. (n.d.). *VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text*. 10.

Malik, Usman. "Implementing PCA in Python with Scikit-Learn." *Stack Abuse*, Stack Abuse, 10 May 2018, stackabuse.com/implementing-pca-in-python-with-scikit-learn/.