Team 23: Team Project Write-Up – Spotify Hit Songs

Rashmi Gehi, Sarah Guo, Tushar Pandey, Ahmet Tas, Alexandra Wang

Business Understanding

Many artists, producers, and record labels in the music industry strive to create a hit song- but what is the perfect balance of danceability, energy, tempo, duration, and other factors that can ensure popularity? According to Business Insider[1], artists typically earn between $0.003 - $0.005 per stream, meaning it takes around 250 streams to earn just $1. This money comes from Spotify's subscription fees and revenue collected from running advertisements. Producers have found that the clearest path to generating a higher income from streaming is to ensure a song will become a hit, leading to a higher number of streams. We will aim to predict the popularity of unreleased songs based on various tangible parameters available (fields mentioned in the dataset). We will also aim to cluster popular songs based on different parameters and identify the parameters that play an important role in the popularity of a song. A data mining solution can create a model that can accurately formulate a mixture of different song factors to guarantee a hit song and ensure a strong income stream for an artist who uses this model through Spotify stream compensation.

Data Understanding

The data comes from Kaggle and is a merged set with a plethora of songs from decades from the 1960s until 2010s. There are 41106 songs represented in the set. The 20 variables in the data are track, artist, uri (Spotify's unique identifier), danceability, energy, key, loudness, mode, speechiness, acousticness, instrumentalness, liveness, valence, tempo, duration (in milliseconds), time signature, chorus hit, sections, target, and year. All the variables are numeric other than

[1] Ennica Jacob, "How Much Does Spotify Pay per Stream? What You'll Earn per Song, and How to Get Paid More for Your Music," Business Insider (Business Insider, February 24, 2021), https://www.businessinsider.com/guides/streaming/how-much-does-spotify-pay-per-stream.

track, artist, and uri. The descriptions of all the song features in detail are presented in Appendix A1.

Data Preparation and Exploratory Data Analysis

The data was first prepared by checking for null values and cleaning those out. We then did some exploratory data analysis to understand the data better and see which variables are the most impactful on popularity. Methods used include boxplot visualizations, word clouds, correlation plots, bar graphs, and scatter plots. First, we created a word cloud (Figure 1) using the song titles to visualize which words are most common in song titles. From this word cloud, it's easy to see that "love" is the most used word in song titles, followed by "don't", "rock", "live", "blue", and "remast".

Secondly, we created a correlation plot (Figure 1) with every variable other than track, artist, and uri, since those are character variables. Based on the correlation plot, we decided to take a closer look at the variables that seem to have the most positive or negative relationships with each other.

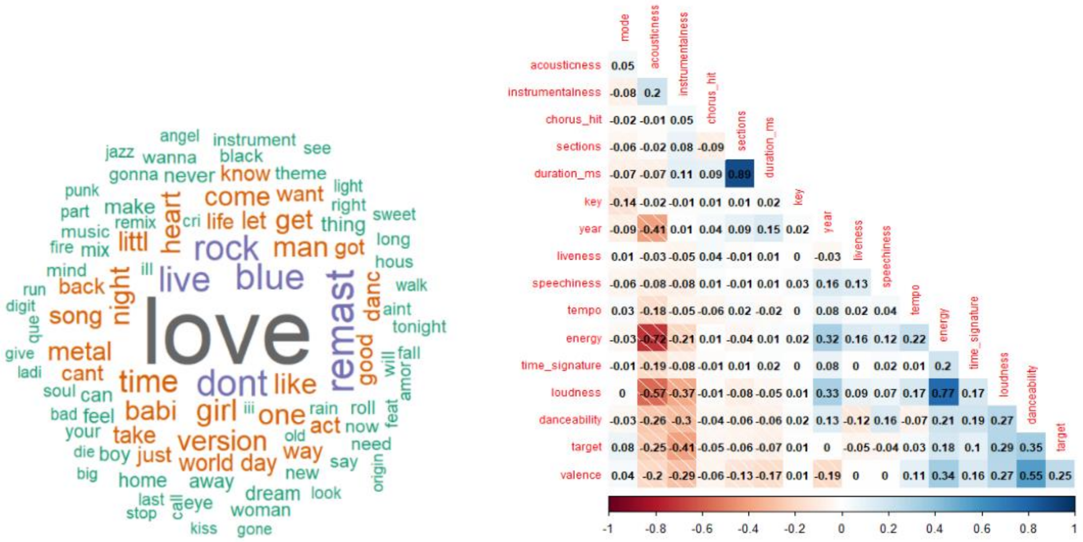


Figure 1: Word cloud using song name & correlation matrix between song

Sections and duration have a highly positive correlation of 0.89, meaning for songs with more sections, the longer it usually is. The dataset defines a section as one distinct part of the song, such as different verses, choruses, and bridges. This can be seen through a boxplot (Appendix A.2) which depicts the range, means, and quartiles for each number of sections, and song duration in milliseconds is on the y-axis.

Mode and song loudness had a correlation of 0, so we wanted to look at that next. Upon investigating a boxplot of the relationship between the two (Appendix A.3), we saw that distribution of loudness is the exact same regardless of mode- a binary variable that is 0 for minor and 1 for major. Minor and major describe the type of key a song is in, and different keys evoke different emotions- for example, sad songs are usually in minor keys. However, loudness, which is measured in decibels here, seems unaffected by different modes.

Next, we wanted to take a closer look at song energy and acousticness, which has a correlation of -0.72. Looking at a scatterplot (Appendix A.4), we can see the highly negative correlation between these two variables. This makes sense, as typically songs that are high energy will not feature acoustic aspects, such as a slow guitar, and will rather have electric synths and beats.

Lastly, we checked whether different decades had a different distribution of hit songs within our dataset, to which we found that for each decade, there were 50% hit songs and 50% flop songs.

Modeling

Unsupervised Machine Learning – PCA and Clustering

To better understand the data, we decided to perform a Principal Component Analysis on all song features (without the popularity variable) to classify the songs into groups. We found that we could explain 49.30% of all songs using only 4 components which we have defined below:

- High energy, loud songs

- Long-duration songs with multiple sections
- Less danceable, low-valence (sad) songs
- Songs in the minor mode with low beats-per-minute

Once we understood the song features that differentiated songs, we decided to perform k-means clustering to group the songs into segments. We found that 3 clusters could segregate the songs into distinct groups with similar characteristics in terms of their features. The songs song groups are in-line with the principal components described above. Below are three boxplots (Figure 2) of song valence and song energy and how they vary starkly across the groups created from k-means clustering.

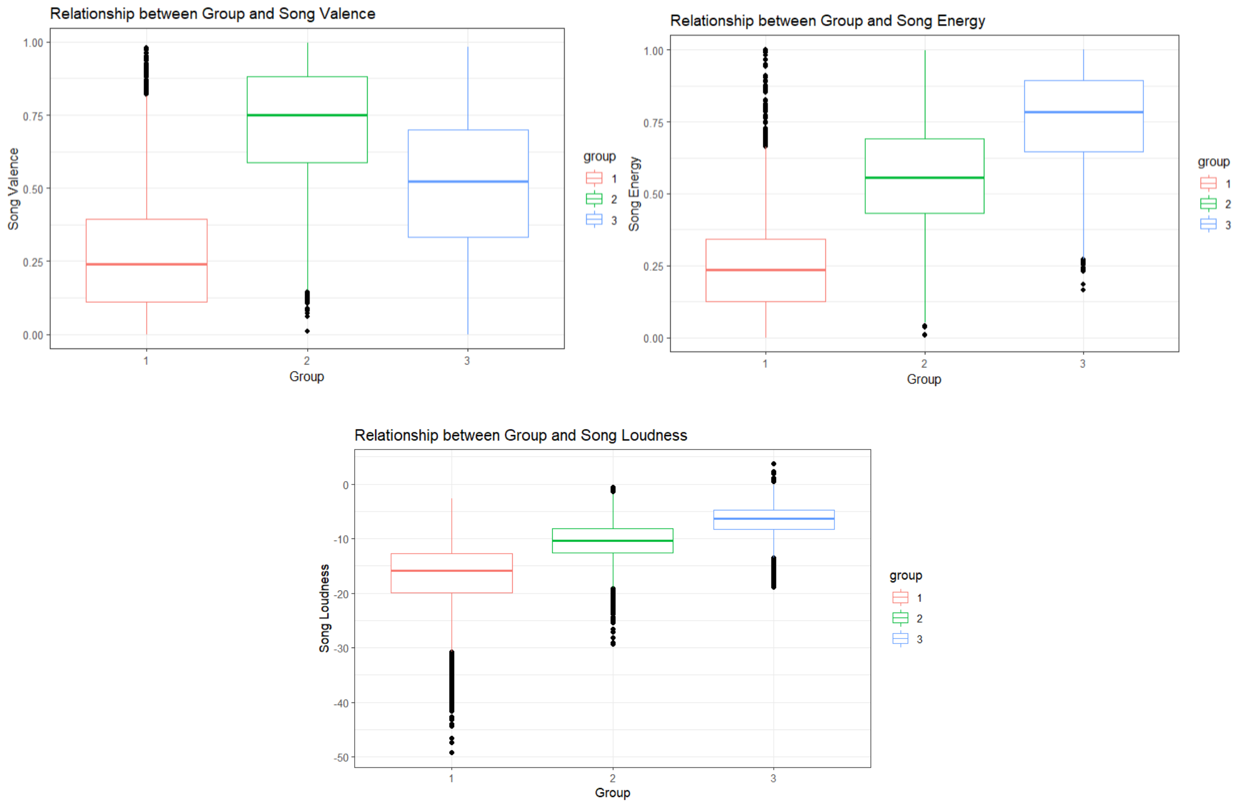


Figure 2: Boxplots showing different song features by clusters of k-means

Below is the distribution of songs from two different decades plotted as clusters with respect to the two most important principal components. We observe that the songs vary not only in terms of components but also by decade.

Predictive Analytics: Supervised Machine Learning

Our data has 20 features, as mentioned above. The variable which determines whether the song was hit on the platform or not is called ‘target’. Target is labeled as 1 if it is a hit song or 0 if the song is not a hit. Features such as track, artist, and Uri are all character variables. Hence, we are dropping these features to make the dataset compatible with predictive analytics. Below are the various models we ran to determine whether the song would be a hit song or not-hit song

- Logistic regression model – This gives us the log(odds) of whether a song would be a hit or not
- Least absolute shrinkage and selection operator (LASSO) model – This model was run as we have 16 different independent variables, which would need to be regularized in order to avoid bias and variance fit (avoid overfitting and underfitting of the variables). This model was run to account for the multicollinearity between the different independent variables
- Post-Lasso model – This model is like Lasso but only considers a minimum number of interactions between song features.
- Classification tree model
- Average model – This model was created using the average predictions of all the above models.

We run the following models to get the prediction and use the null model for the baseline inference of accuracy and performance. We checked various out-of-sample metrics to compare the above models and picked the model which would be used for predicting the probability of is_hit_song. In the next section, we also analyze what the best value of the threshold could be to convert the probability into a binary variable.

Evaluation

After performing 10-fold cross-validation, we evaluated the following average metrics to compare each of our models - Logistic Regression, Lasso Regression, Post Lasso Regression, Classification Tree, Average model, and Null Model.

- Out-of-sample Accuracy – From figure 3 below, we see that Lasso Regression gives us the highest out-of-sample accuracy (67.76%)

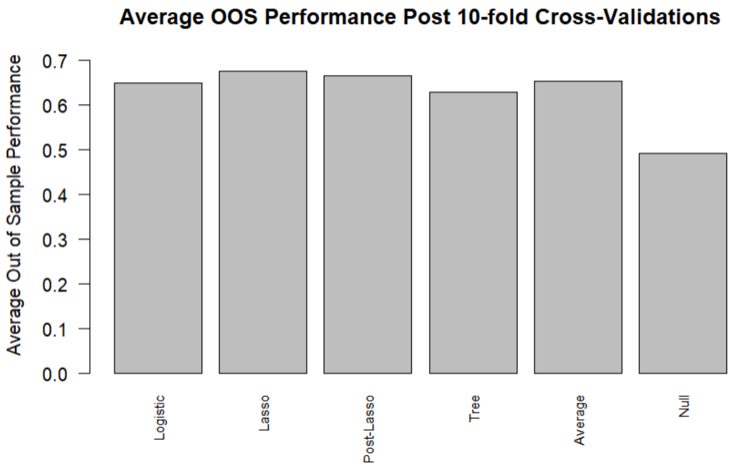


Figure 3: Comparison of Accuracy

- Out-of-sample R-squared – Similar to the average accuracy of 10-fold cross-validation, the average out-of-sample R-squared of Lasso Regression is the highest (30.29%)

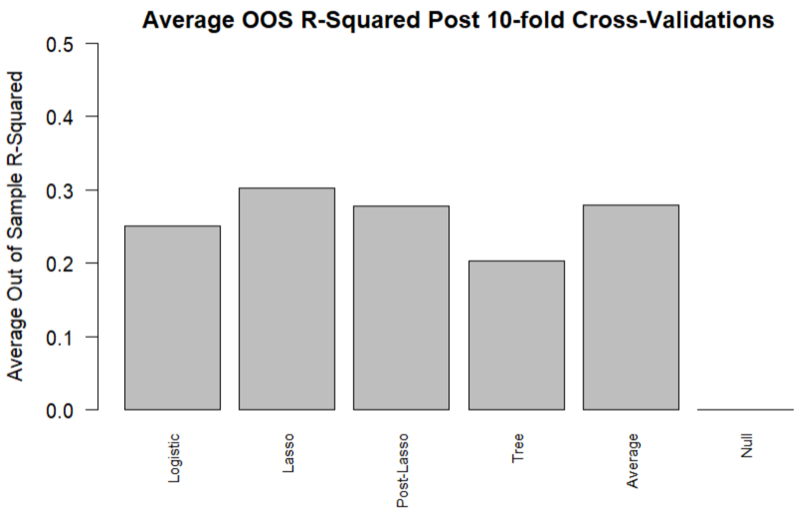


Figure 4: Comparison of R-Squared

- ROC curves – In addition to the accuracy and R-squared, we also plotted the ROC (Figure 5) of all the models (this was done after training the data on the whole dataset), and we see similar patterns of the Lasso regression out-performing the other models and has the largest area under the curve.

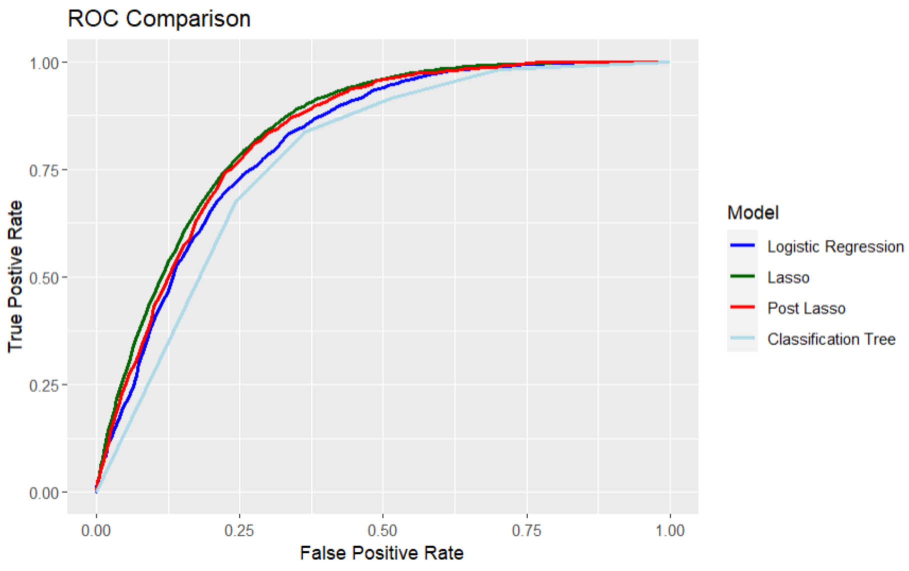


Figure 5: Comparison of ROC

Deciding threshold: For our business recommendations, we cannot afford to lose on the songs which are potentially hit songs. Put differently - it is essential that we maximize the correct predictions for hit songs. Hence, we decided to focus on optimizing the true positive rate. If we have any false predictions, we can still put some marketing budgets to make the record work. From the confusion matrix and ROC, we get the following threshold value table1:

Threshold	FPR	TPR	ACC	TP	FP	FN	TN
0.25	0.53	0.97	0.72	19,898	10,806	655	9,747
0.45	**0.35**	**0.89**	**0.77**	**18,335**	**7,289**	**2,218**	**13,264**
0.50	0.31	0.85	0.77	17,500	6,342	3,053	14,211
0.75	0.08	0.41	0.66	8,348	1,710	12,205	18,843

Table 1: Confusion Matrix at various threshold values

We chose the threshold value of **0.45**, which meets the business expectation. We chose to go forward with the Logistic Regression model with Lasso Reduction of features. Using the LASSO reduction, we got the dominant features (including interactions) which explain 86% variation of the dataset. Hence, we formed an equation of the expected value of the is_hit_song with betas of every dominant feature to decide the song's popularity. With this equation, we can predict the expectation of log odds and hence the probability of the unreleased song. As we finalize our logistic regression model with lasso-reduced features, we have below (Table 2) the top 10 interactions and the marginal effects on the log(odds) of a hit song. A few additional notes on the song feature when other features are kept constant.

- Increase in speachiness (number of spoken words) leads to a drastic **decrease** in log(odds) of popularity, but this is only when the instrumentalness and energy increase too. These can be considered descriptors of extremely wordy songs that are high-energy.
- Increase in energy leads to a drastic **increase** in log(odds) of popularity, but this is only when the acousticness and danceability increase too. These are descriptors of electronic dance music.

- Increase in valence (how happy a song is) leads to a drastic **increase** in log(odds) of popularity, but this is only when the instrumentalness and energy increase too. These are descriptors of happy vibe songs.

Interaction Name*	Coefficient
(Intercept)	45.74
speechiness:instrumentalness	-8.23
energy:speechiness	-7.99
energy:acousticness	5.75
danceability:energy	5.35
instrumentalness:valence	3.36
acousticness:valence	-2.96
danceability:valence	-2.94
energy:valence	2.41
mode:speechiness	-2.15
danceability:acousticness	-2.07

Table 2: Song Feature Interactions that are important in Lasso (in descending order of absolute value of coefficients)

Deployment

The deployment of our model, as mentioned in the business problem, is for songwriters, producers, and artists to understand how to guarantee a hit song and leverage this information to increase their revenue and profits. Based on the clusters we found through PCA, record labels can leverage the clusters to appeal to different demographics who prefer certain features in songs, such as those who prefer sad, low-energy songs and those who prefer songs with high valence and loud beats. The primary issue record labels should be aware of regarding the deployment of this model is the response that Spotify will have. If streams are increasing overall, then Spotify will begin to lose money, which would hurt its business. They may retaliate by decreasing the dollar per stream rate, undermining the effectiveness of our model.

One risk associated with the deployment of this model is that it would likely increase competition between entities in the music industry. Artists obviously want to make a livable wage, and if the clearest path to doing so is ensuring popularity, then there is no reason not to use our model and create a hit song and increase their number of streams. We could mitigate this problem by minimizing access to this model, perhaps by selling it at a high price or only working with artists who are not already popular.

Another risk is that many songs have the potential to sound the same. For example, if multiple record labels decide to implement this model in their songwriting process, then their songs would feature the exact same attributes, making new songs from the labels seem very cookie-cutter and predictable. However, our model leaves room for variance in terms of instruments, key, mode, tempo, and time signature, so the song's structure does not necessarily have to sound entirely like fit the model.

Lastly, one risk with our model is that we omitted any factors related to the artist's name. This is because we wanted our model to represent an equal opportunity for artists, regardless of popularity, to understand the type of music itself that will become popular.

BIBLIOGRAPHY

Jacob, Ennica. "How Much Does Spotify Pay per Stream? What You'll Earn per Song, and How to Get Paid More for Your Music." Business Insider. Business Insider, February 24, 2021. https://www.businessinsider.com/guides/streaming/how-much-does-spotify-pay-per-stream.

INDIVIDUAL CONTRIBUTIONS BY TEAM 23 MEMBERS

- Rashmi Gehi: I worked on coding all the different models in R that we tested out in this project. I also evaluated the different metrics to identify which model performed the best. Additionally, I performed PCA and k-means clustering, and finally, I worked on interpreting the different principal components and clusters that were the outcomes of the models.
- Sarah Guo: I worked on the business and data understanding, deployment, and some of the exploratory data analysis visualizations.
- Tushar Pandey: I performed the Modelling and Evaluation part. Majorly focusing on interpretation and evaluation of the best performing model through different performance metrices. Finalizing the final model and interpreting marginal impact on the expected value of the probability of a hit song.
- Ahmet Tas: I took part in the visualizations and presentation part of the project.
- Alexandra Wang: I supported the visualizations and scratched and finished the slide deck for the presentation.

APPENDIX A

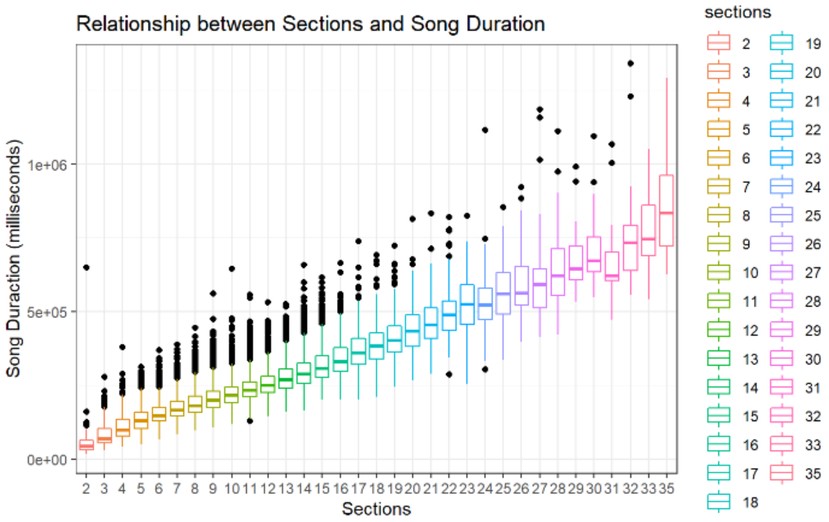
1. Song Features Description

- danceability: Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable.
- energy: Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. For example, death metal has high energy, while a Bach prelude scores low on the scale. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy.
- key: The estimated overall key of the track. Integers map to pitches using standard Pitch Class notation. E.g. 0 = C, 1 = C?/D?, 2 = D, and so on. If no key was detected, the value is -1.
- loudness: The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks. Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). Values typical range between -60 and 0 db.
- mode: Mode indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived. Major is represented by 1 and minor is 0.
- speechiness: Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks.
- acousticness: A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic.
- instrumentalness: Predicts whether a track contains no vocals. “Ooh” and “aah” sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly “vocal”. The closer the instrumentalness value is to 1.0, the greater likelihood the track contains

no vocal content. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0.

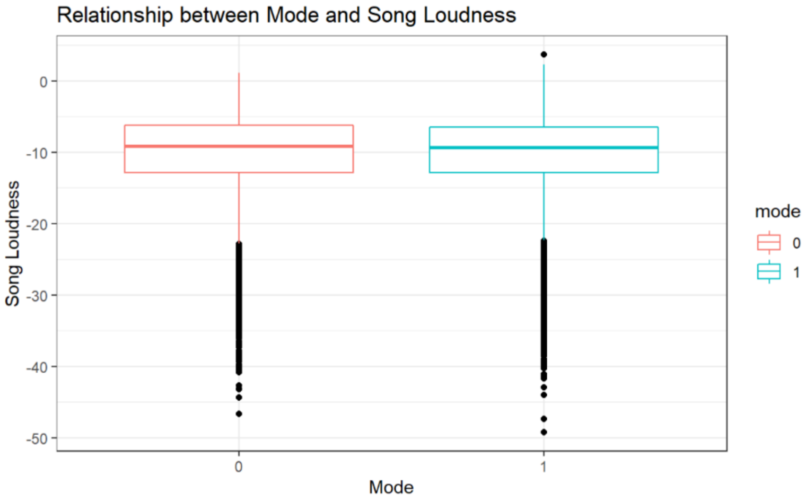
- liveness: Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live.
- valence: A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).
- tempo: The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration.
- duration_ms: The duration of the track in milliseconds.
- time_signature: An estimated overall time signature of a track. The time signature (meter) is a notational convention to specify how many beats are in each bar (or measure).
- chorus_hit: This the author's best estimate of when the chorus would start for the track. It’s the timestamp of the start of the third section of the track (in milliseconds). This feature was extracted from the data recieved by the API call for Audio Analysis of that track.
- sections: The number of sections the particular track has. This feature was extracted from the data received by the API call for Audio Analysis of that particular track.
- target: The target variable for the track. It can be either '0' or '1'. '1' implies that this song has featured in the weekly list (Issued by Billboards) of Hot-100 tracks in that decade at least once and is, therefore a 'hit'. '0' Implies that the track is a 'flop'.

2. Relationship between Song Section and Song duration

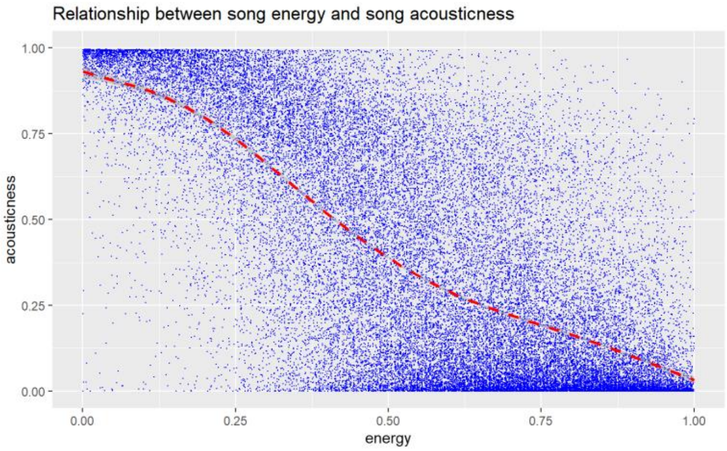


s

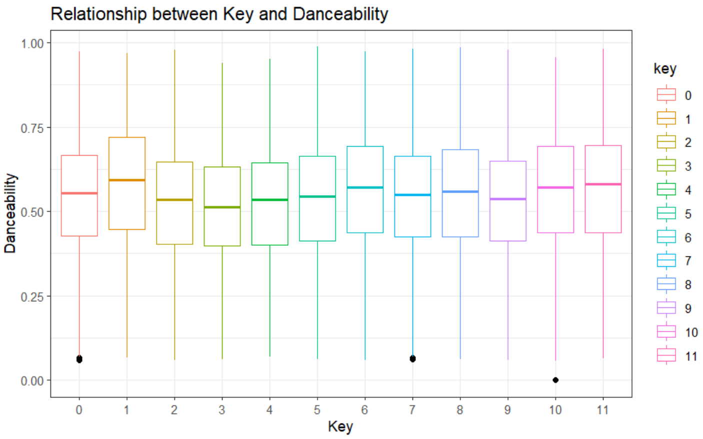
3. Relationship between model and loudness



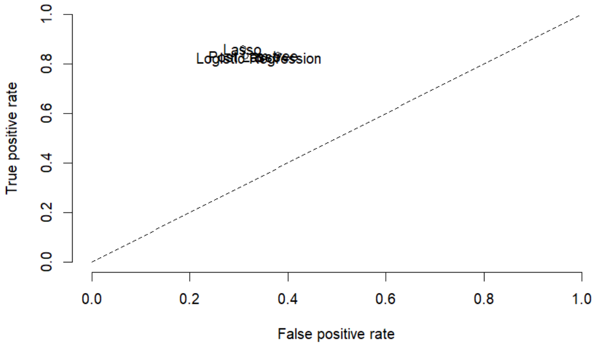
4. Relationship between song energy and song acousticness



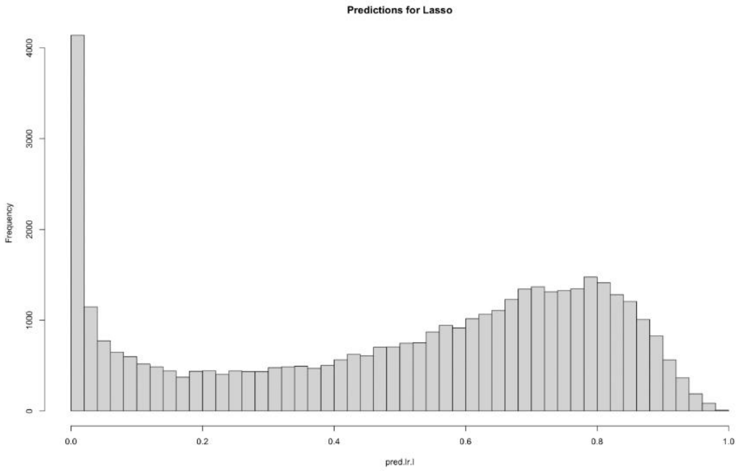
5. Relationship between song key and danceability



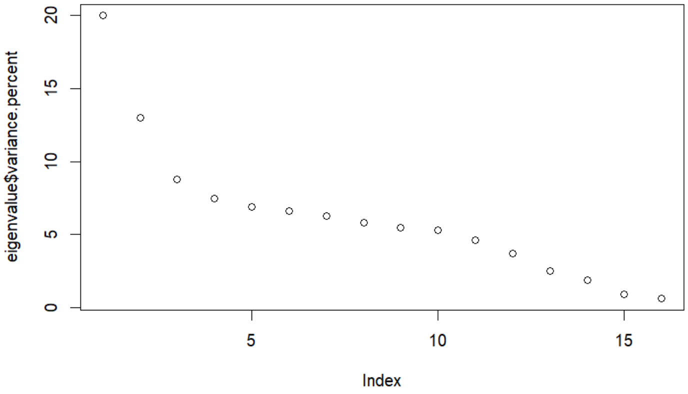
6. Average Accuracy of various models



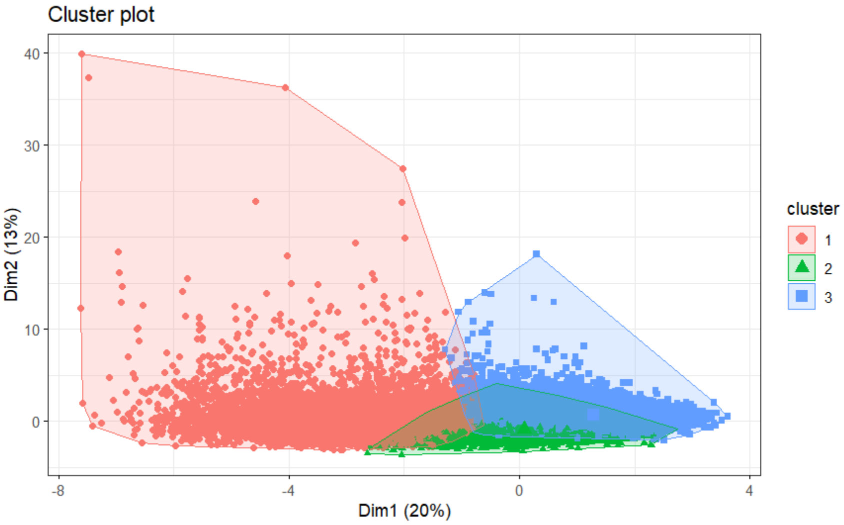
7. Histogram of predictions of Lasso



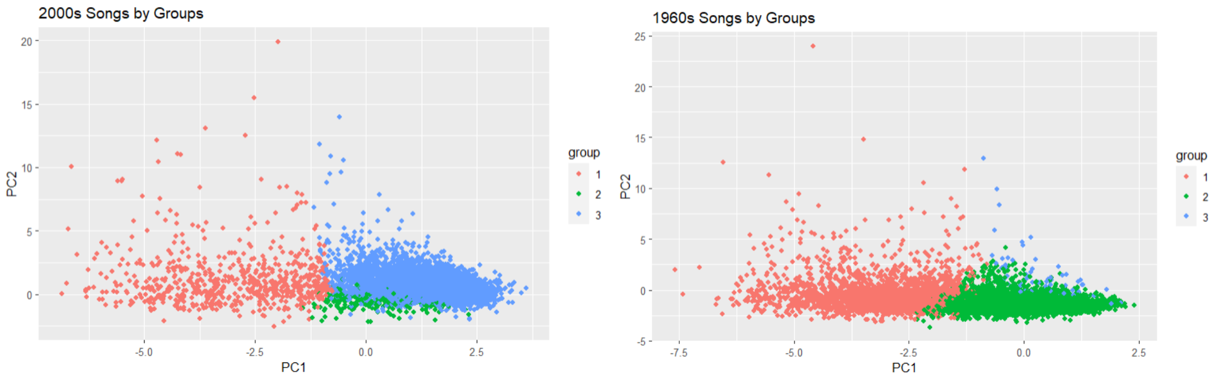
9. Variation explained by individual Principal Components



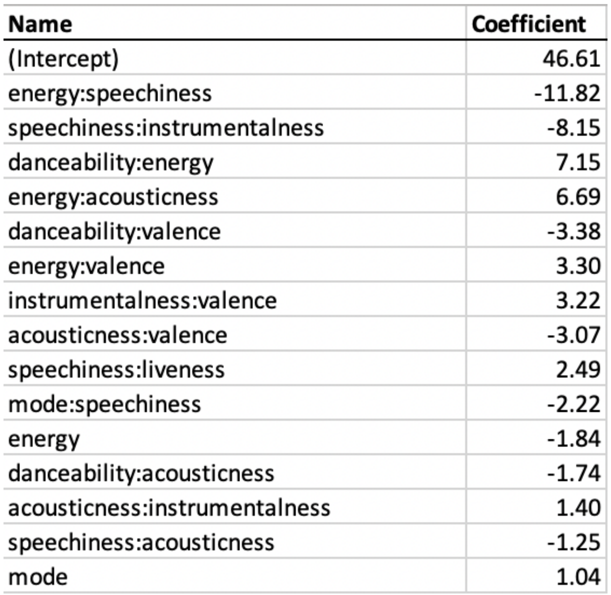
8. Cluster Analysis



9. Clusters plotted by decade (2000s and 1960s songs comparison)



10. Top song feature interactions identified by lasso regression (sorted in descending order of absolute value of coefficients). Lasso Features and their coefficients, coefficient cut-off value 1.



Name	Coefficient
(Intercept)	46.61
energy:speechiness	-11.82
speechiness:instrumentalness	-8.15
danceability:energy	7.15
energy:acousticness	6.69
danceability:valence	-3.38
energy:valence	3.30
instrumentalness:valence	3.22
acousticness:valence	-3.07
speechiness:liveness	2.49
mode:speechiness	-2.22
energy	-1.84
danceability:acousticness	-1.74
acousticness:instrumentalness	1.40
speechiness:acousticness	-1.25
mode	1.04