Christian Shadis
CS-453
Dr. Braynova
Due 12/15/2021

What Makes a Song Popular?

Abstract

A song can be described as a set of numerical attributes representing different features of the sound. The goal of this project is to use a dataset of 50,000 songs and machine learning techniques to determine whether a song's popularity or genre can be predicted from those numerical attributes. The project found that predicting a song's popularity using the given numerical attributes is not possible with the data, nor is classification from a set of ten genres. However, it was found that machine learning techniques can be used to differentiate between Rock and Jazz songs.

1. The Dataset

The dataset chosen for this project consisted of 50,000 instances, each representing one song on the popular music streaming service Spotify. There were 5,000 instances (songs) for each of the genres included in the dataset. The dataset in full contained eighteen attributes, but six were discarded for analytical purposes (See: Data Preprocessing). The remaining twelve attributes are summarized in the table below:

Attribute	Type (# categ.)	Values	Example
Popularity	Numeric	[0, 99]	65
Acousticness	Numeric	[0, 1)	.99
Danceability	Numeric	(0, 1)	.001
Energy	Numeric	(0, 1)	.5
Instrumentalness	Numeric	[0, 1)	0
Liveness	Numeric	(0, 1]	1
Loudness	Numeric	(-48, 4)	-3
Speechiness	Numeric	(0, 1)	.2
Valence	Numeric	[0, 1)	.99
Key	Nominal (12)	A, A#, B, etc.	F#
Mode	Nominal (2)	Major, Minor	Minor
Music Genre	Nominal (10)	Pop, Blues, etc.	Jazz

Popularity, as its name suggests, refers to the relative popularity of a song on a scale of 0-99. Acousticness represents the confidence that a given track could be described as “acoustic”. Danceability, energy, instrumentalness, liveness, speechiness, and valence (or happiness) are all similar attributes that range from 0.0 to 1.0 and use elements such as tempo, timbre, and beat strength to determine the confidence that the song possesses the specified quality. Music Genre is a nominal attribute containing ten values: “Rock”, “Jazz”, “Blues”, “Electronic”, “Anime”, “Hip-Hop”, “Rap”, “Alternative”, “Country”, and “Classical”.

2. Data Preprocessing

The dataset was complete in the sense that there were few missing values in the significant attributes mentioned in Section 1. There were six additional attributes that were problematic for various reasons.

A unique identifier “instance_id” was removed for analysis because looking up certain songs by that unique identifier was never necessary, and it provides no insight into the data. Two String attributes “artist_name” and “track_name” were removed from consideration in the analysis. String attributes are difficult to work with, and there are unlikely to be any interesting patterns involved with the names of artists or songs.

Among those three, the other variables omitted from analysis were “duration_ms”, “tempo”, and “obtained_date”. Duration was unlikely to offer any valuable insight into a track’s popularity, so the existence of some missing values was enough justification to omit the attribute entirely. Similarly, tempo had many missing values as well, while also being accounted for in other more complex measures such as danceability and energy. Thus, the attribute was removed. The date each song was obtained from Spotify seemed to serve no purpose for this analysis and was thus removed.

The final modification made to the dataset was the discretize the “Popularity” variable. The original range of values was 0-99, so the values were divided into five bins, each 20 units in length, labelled “Unpopular” (0 – 19), “Slightly Popular” (20 – 39), “Moderately Popular” (40 – 59), “Popular” (60 – 79), and “Very Popular” (80 – 99).

3. Summary Statistics

Most variables have five-number summaries like those we might expect from variables like these.

Popularity has a center close to 0.5 and a very wide spread, as one might expect.

Acousticness has center far below 0.5, which indicates that there are more non-acoustic songs than acoustic songs. The third quartile of Acousticness is slightly above 0.5, which means that nearly 75% of all songs in the dataset were non-acoustic songs.

Danceability is centered above 0.5, and has a relatively small standard deviation, indicating that most songs are danceable. The first quartile is 0.44, meaning the vast concentration of the songs in the data are relatively danceable.

There is no interesting information in the statistics for the energy attribute. It is centered above 0.5, indicating that the majority of songs are energetic, but there is a high standard deviation and wide spread.

Instrumentalness is heavily skewed right, with at least 25% of the data concentrated at the minimum value, and 75% of the data concentrated below 0.16. Yet the maximum value is 0.99 and a high standard deviation indicates a wide spread.

Liveness, like instrumentalness, is heavily skewed right, but less extremely so. The vast majority of songs are non-live, and the spread of the data is more narrow than most other attributes.

Loudness, as one would expect, has its middle 50% of the data centered in the range from -10 db to -5 db, with some spread on either side, and some loud and quiet outliers.

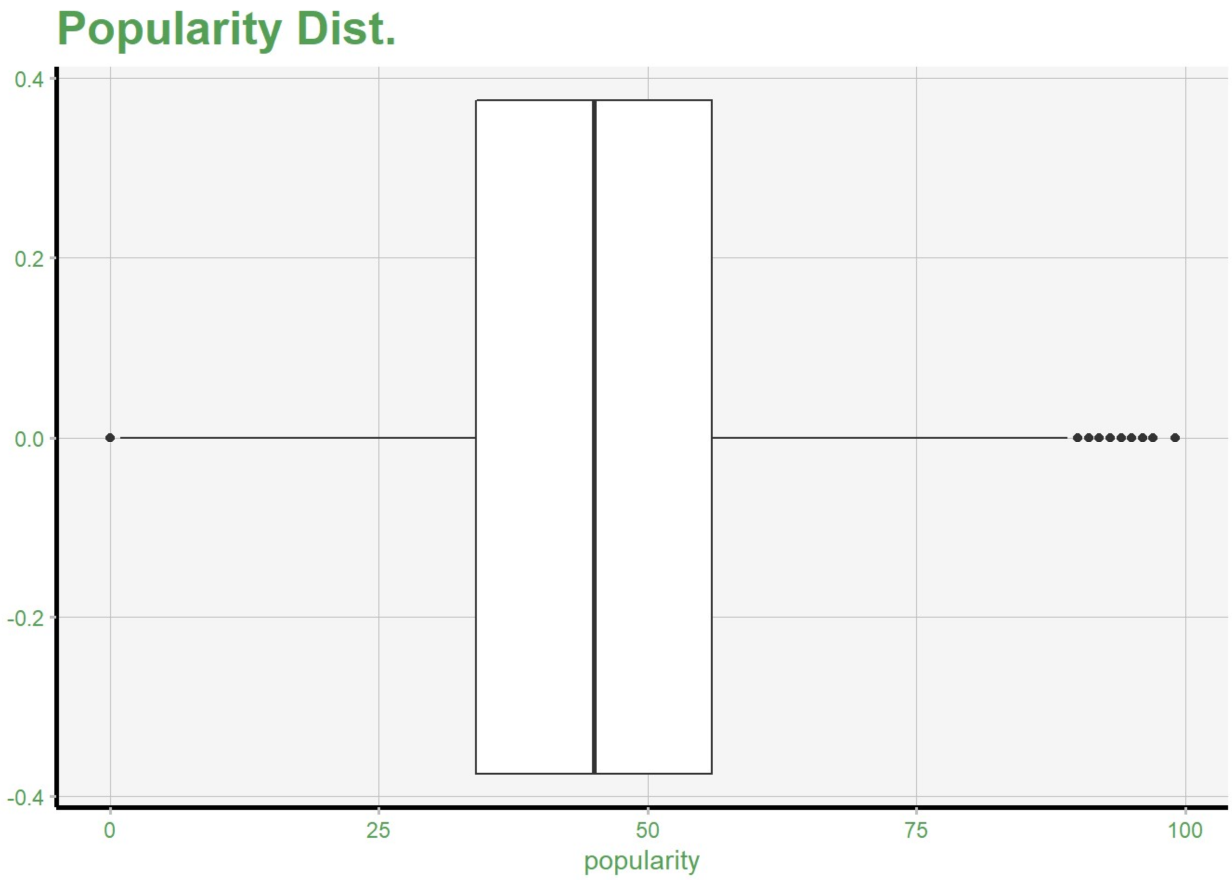
Speechiness is also among those attributes with a heavy right-skew. As expected, 75% of the songs have a speechiness value of .10 or under. This follows intuition as music typically contains minimal speech.

Valence appears nearly normal, with a center around 0.5 and a standard deviation of 0.25. There is nothing noteworthy in its summary statistics.

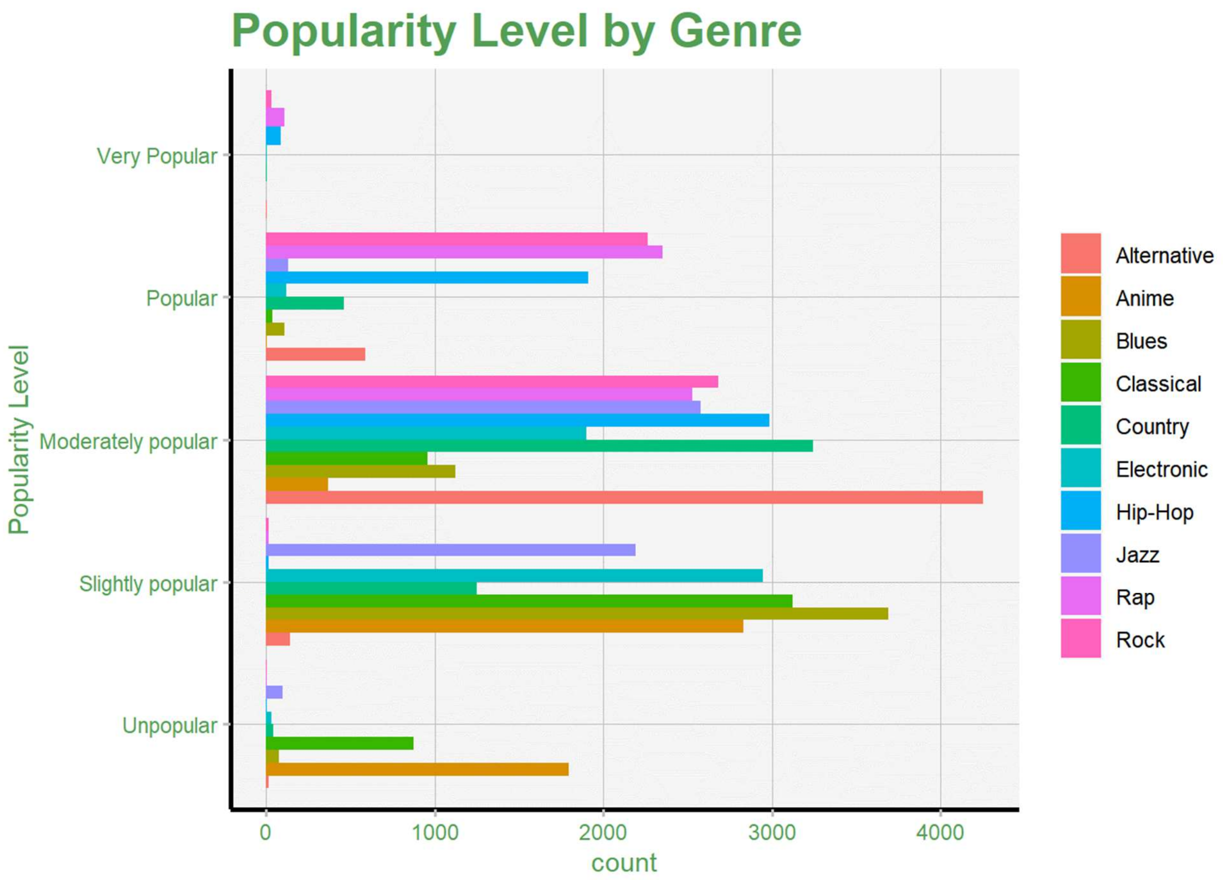
Variable	Min.	Q1	Median	Q3	Max.	Mean	Standard Deviation
Popularity	0	34	45	56	99	44.22	15.54
Acousticness	0	0.02	0.14	0.55	0.99	0.31	0.34
Danceability	0.06	0.44	0.57	0.69	0.99	0.56	0.18
Energy	7.92e-4	0.43	0.64	0.82	0.99	0.60	0.26
Instrumentalness	0	0	1.58e-4	0.16	0.99	0.18	0.33
Liveness	0.01	0.10	0.13	0.24	1	0.19	0.16
Loudness	-47.05	-10.86	-7.28	-5.17	3.74	-9.13	6.16
Speechiness	0.02	0.04	0.05	0.10	0.94	0.09	0.10
Valence	0	0.26	0.45	0.648	0.99	0.46	0.25

4. Distribution Visualizations

Popularity



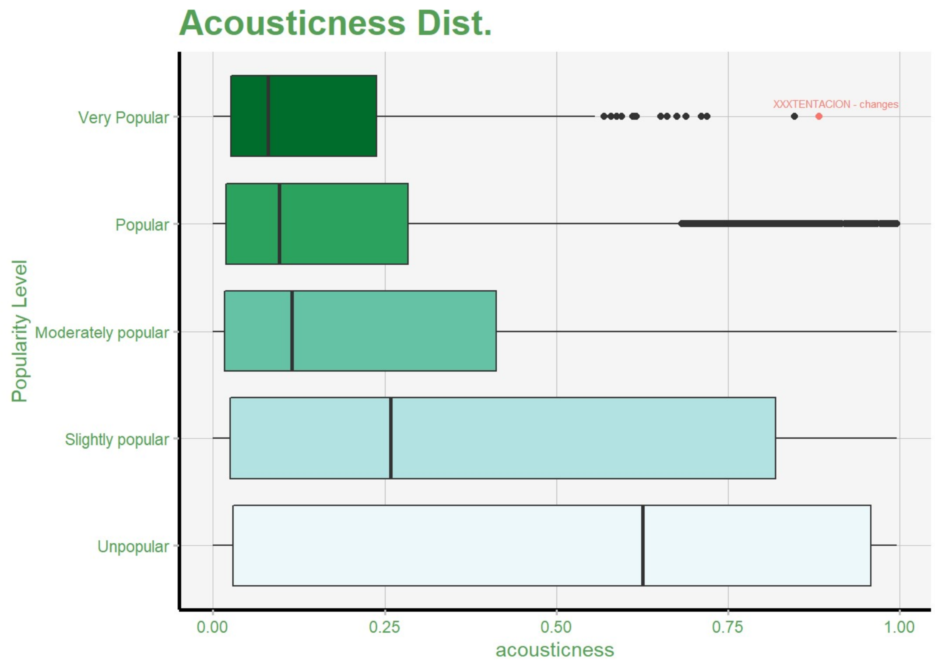
There are 694 outliers with 0 popularity, represented by a single point on the graphic above. The outliers with popularity over 95 are "Taki Taki" by DJ Snake, "Sunflower" by Post Malone, "Wow." By Post Malone, and "MIDDLE CHILD" by J. Cole.

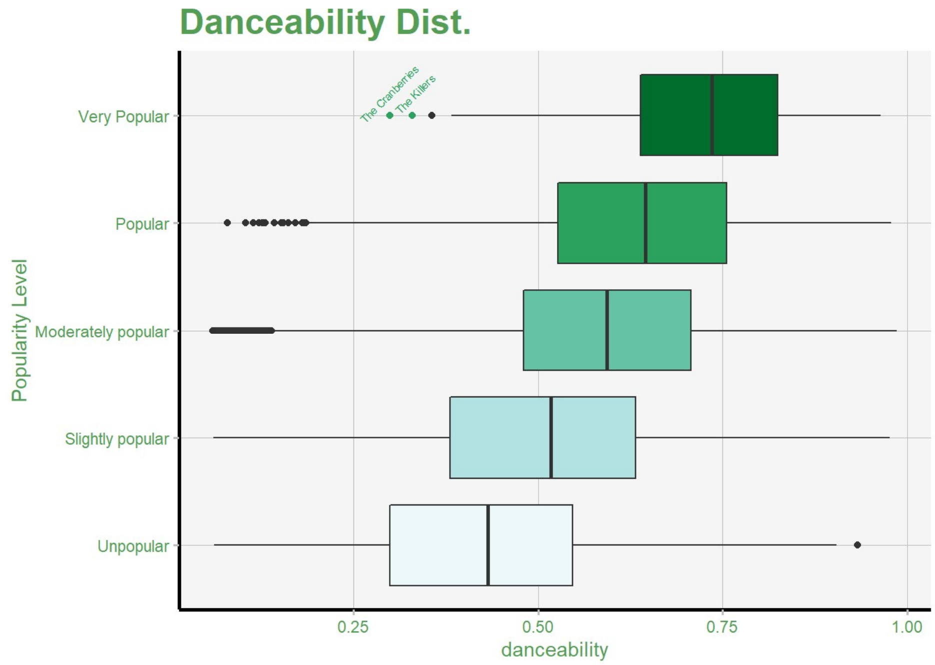


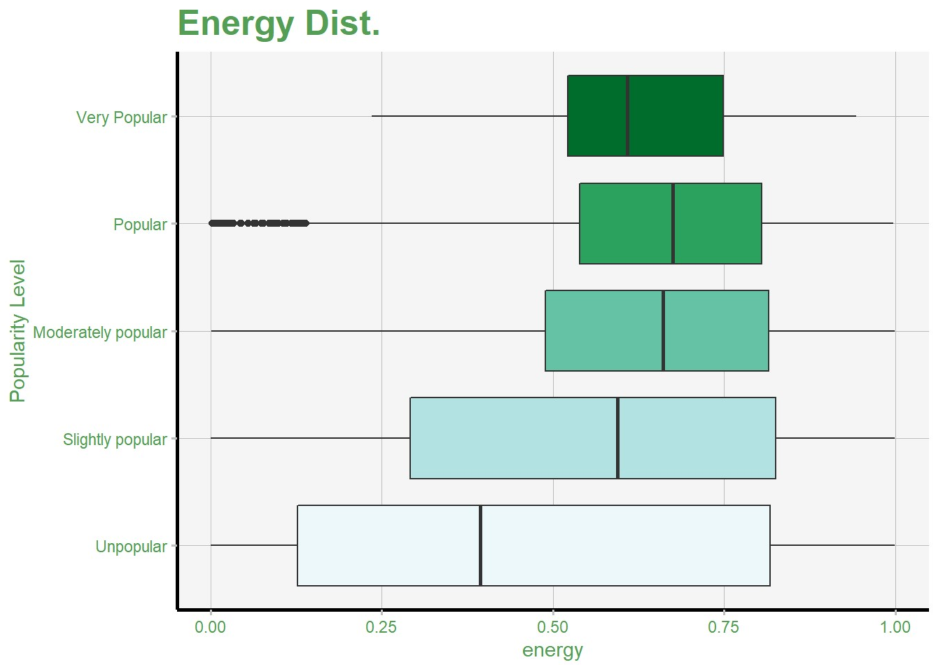
As one might expect, the largest proportion of songs with a relatively high popularity are Rock, Hip-Hop, and Rap, while Anime, Classical, and Blues make up a large proportion of songs with a relatively low popularity.

Boxplots

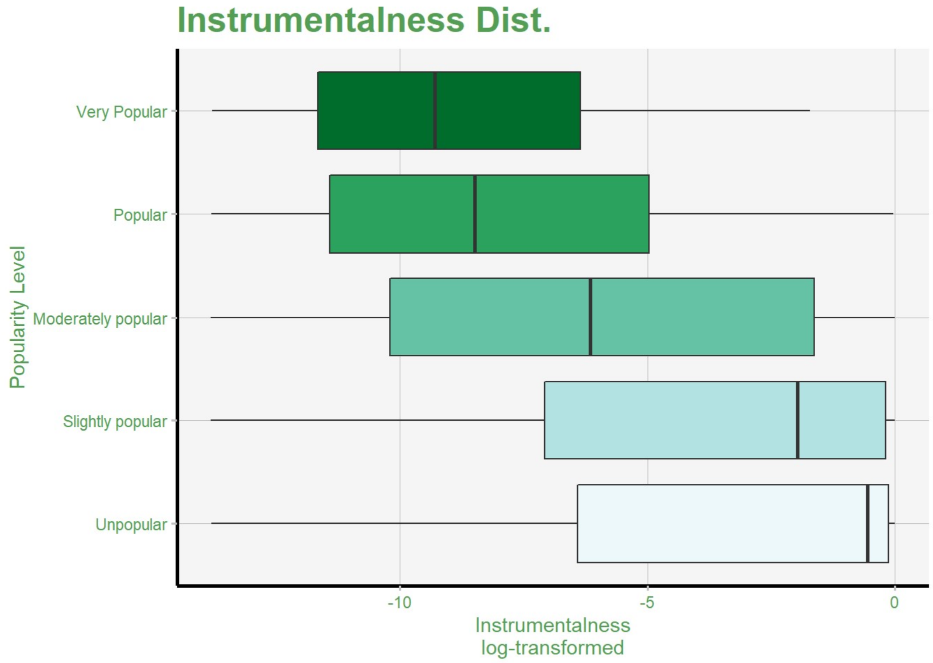
The boxplots of the remaining attributes resemble their summary statistics from Section 3, but the data is separated by popularity.



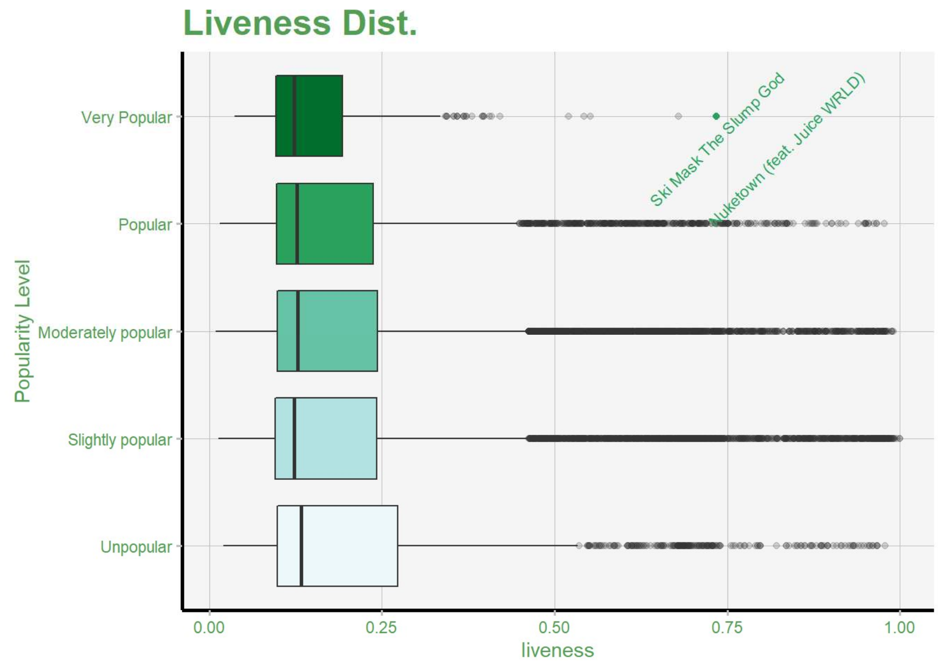




Very popular songs tend to have relatively high energy and are never extremely non-energetic. All other popularity groups have a much wider spread of values.

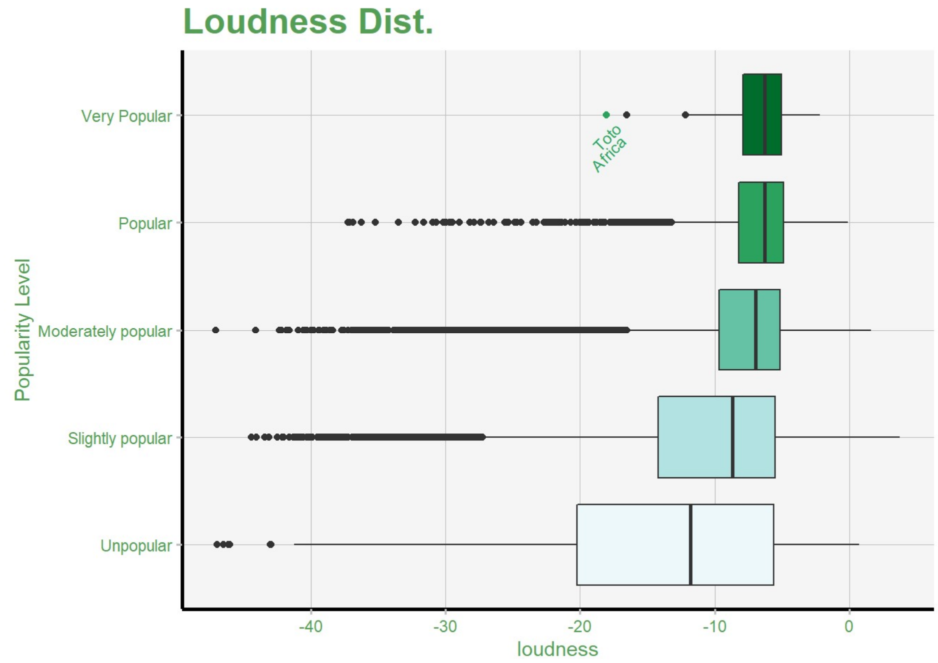


The instrumentalness of more popular groups of songs have measures of center lower than their less popular counterparts. More successful songs tend to have less instrumentalness.



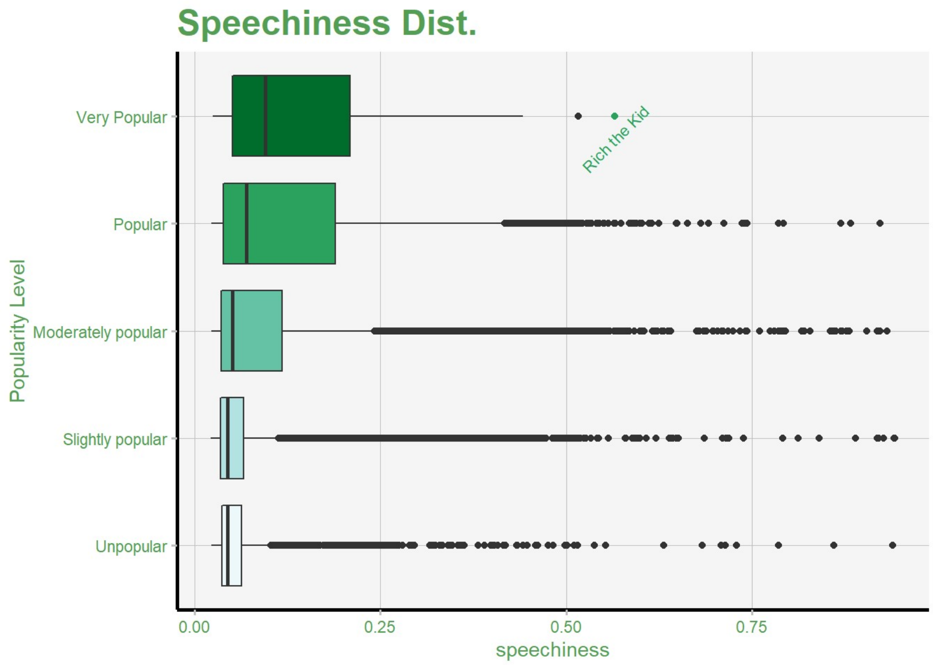
Liveness Dist.

Most songs on the platform are non-live, and most very popular songs are non-live.

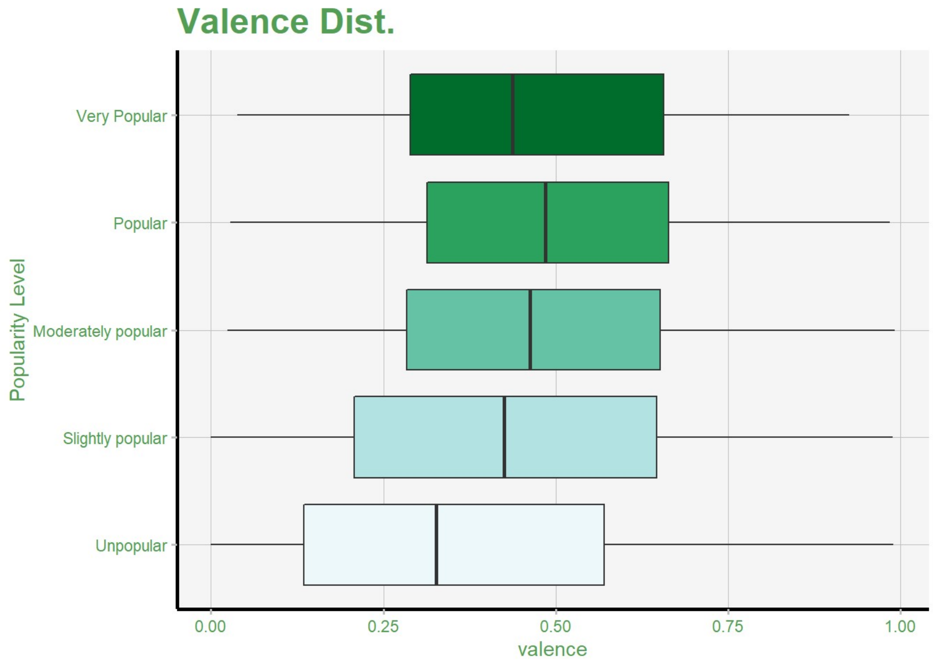


Loudness Dist.

Songs that are abnormally quiet are unlikely to be very popular.



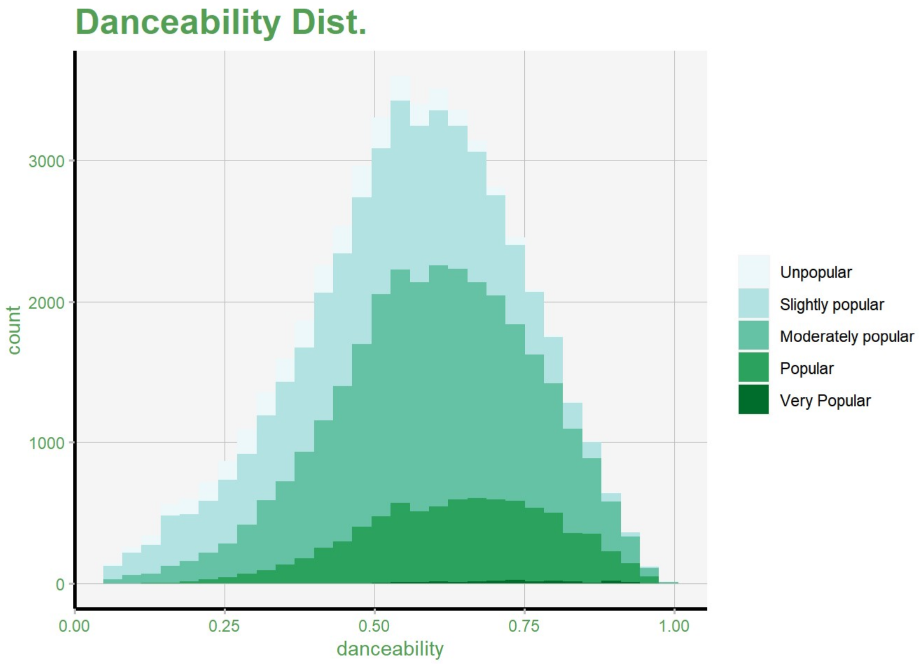
Very few speechy songs are Very Popular, and only slightly more are Popular.

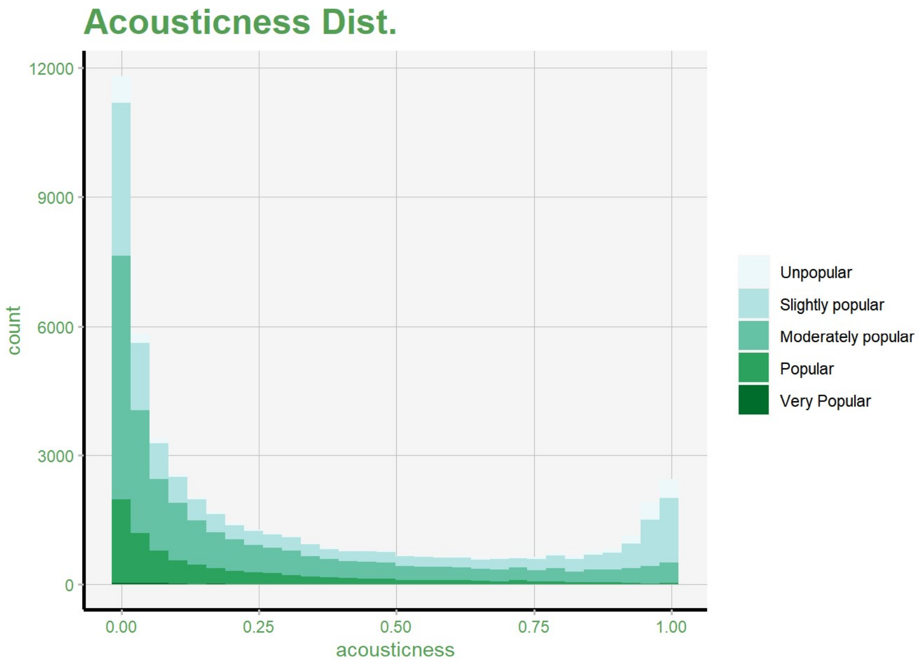


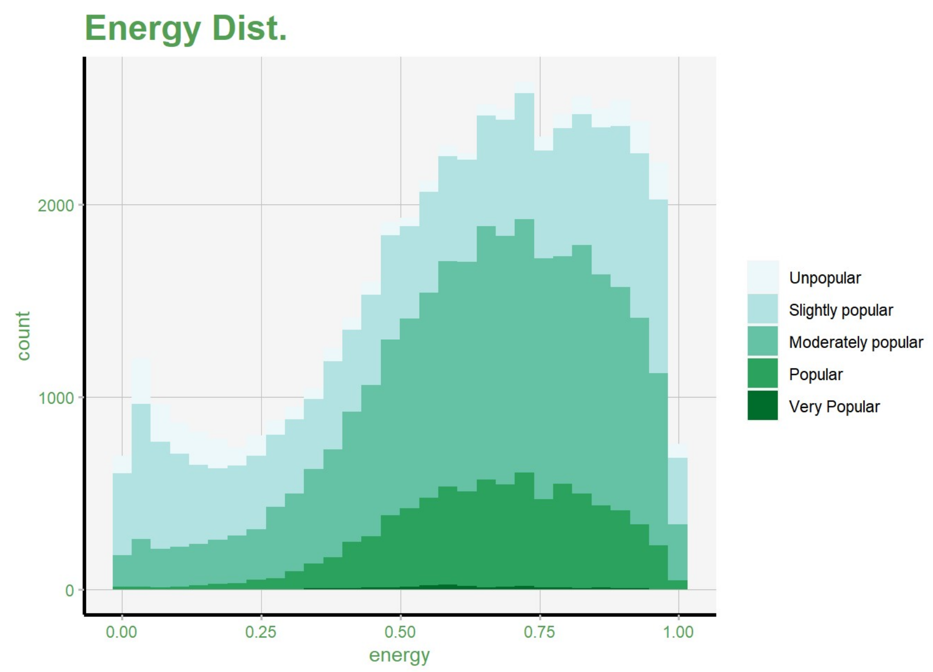
The valence of songs in all popularity categories is evenly distributed.

Histograms

Histograms were produced for some of the more interesting distributions described above.



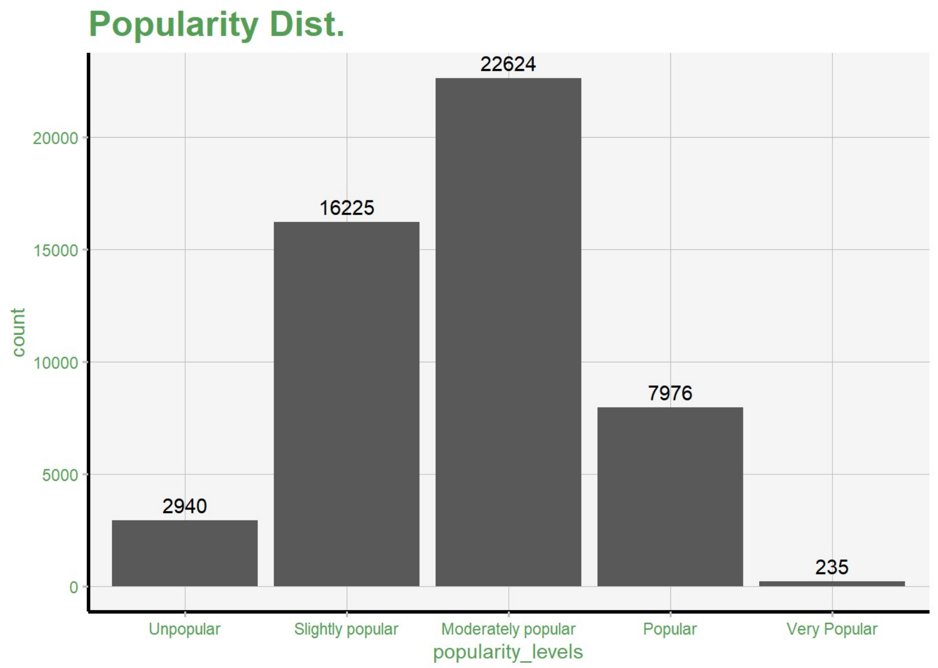


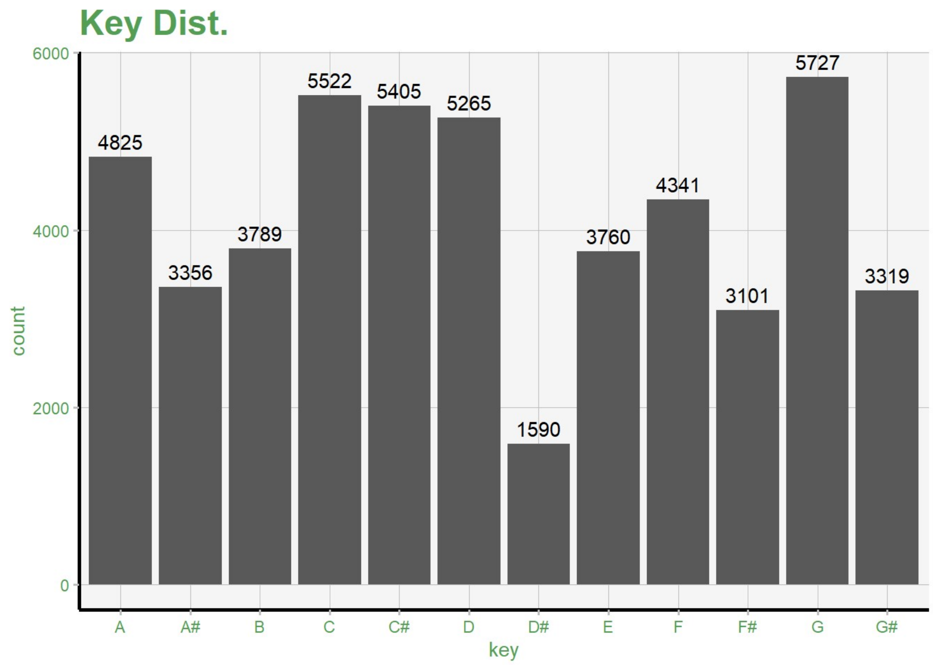


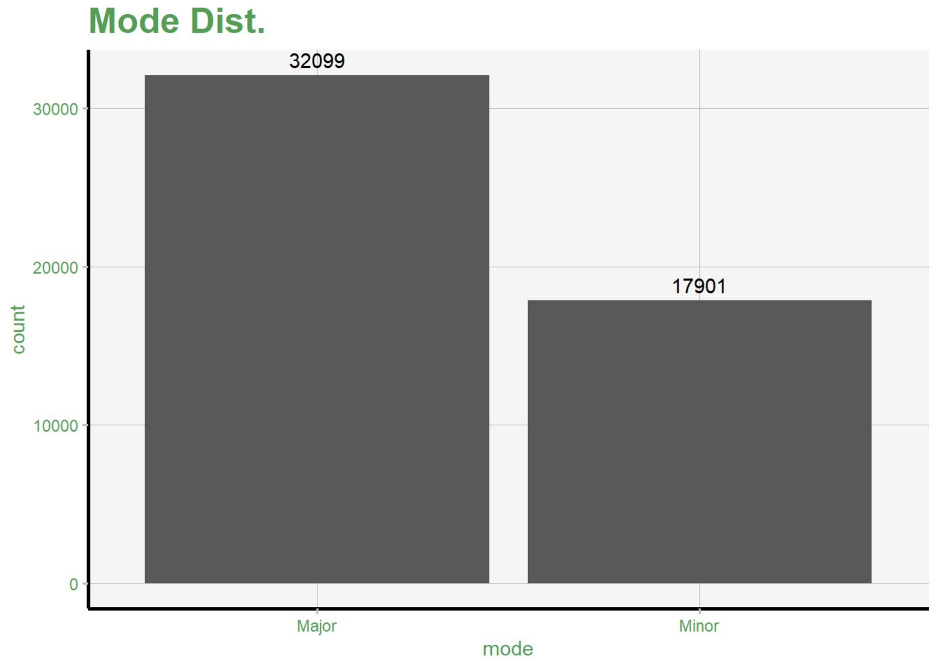
These histograms seem to illustrate that the distributions of variables seem to be similar amongst the popularity groups.

Bar Charts

A Bar Chart was constructed to analyze the distribution of each categorical variable.



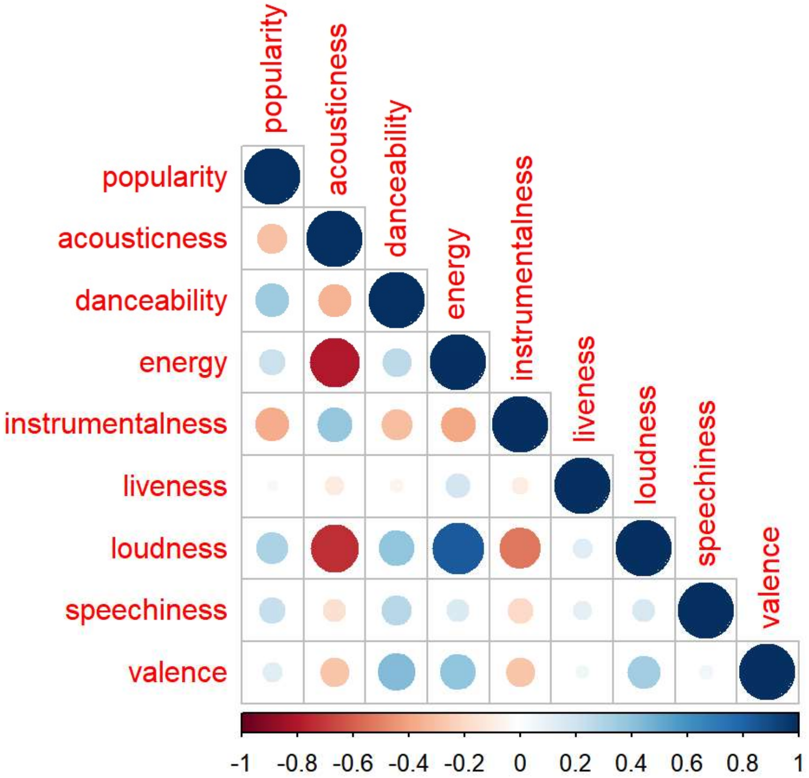




These bar charts offer little noteworthy information other than the largest popularity groups are "Moderately Popular" and "Slightly Popular", and that the Major scale is used more commonly than the Minor scale in the data.

Correlation

The following is a correlation map representing the Pearson Correlation measure **r** between each of the numerical attributes.



The relationships illustrated by the correlation map are unsurprising. There is a strong negative correlation between acousticness and energy, which is intuitive in that acoustic songs are more downtempo. There is also a strong negative correlation between acousticness and loudness for a similar reason. The strongest positive correlation is between loudness and energy, which is again intuitive.

5. Data Mining

Data Mining Questions

This project focused primarily on answering two questions: Can we predict a song's popularity based on its numeric attributes, and can we predict a song's genre based on those same attributes? Numerical prediction is used to predict the song's *numeric* version of "popularity", and then classification is used to predict the song's popularity *category* (i.e. "Very Popular"). To predict music genre, classification is used.

Knowledge Representation Models

For numerical prediction, multiple linear regression is used. For classification, Decision Trees, Bayes Classifiers, and Rule-based Classification are used.

Algorithms Used

The multiple linear regression algorithms used were Weka's Simple Linear Regression and traditional Linear Regression, in addition to the R Programming Language's linear model function.

The Decision Tree algorithms used were the J48 tree, REPTree, Random Tree, and Random Forest.

The Bayes classifiers used were the NaiveBayes classifier and a BayesNet.

The Rule algorithms used were OneR and PART.

Train/Test Sizes

When using a train/test split during analysis, three different split values were chosen: 66/34, 80/20, and 90/10. These are common values chosen for train/test splits. Each model was run with each split.

Efficiency Results

Numerical Prediction:

Using WEKA, the Relative Absolute Error (RAE) for Linear Regression was 84.6%. The RAE for Simple Linear Regression was 90.9%. Thus the WEKA linear regression models do a poor job of predicting accuracy. In R, a package called 'Leaps' was used to determine the best subsets of attributes to use in the multiple linear regression model. The three best models were chosen; one contained all attributes, one contained all attributes except energy, and the third contained all attributes except valence. The maximum R^2 value among those models was 0.233. Thus only 23% of the variation in Popularity was described by the model containing all attributes.

Classification:

Popularity Classification Results

Test Option Algorithm	J48	RepTree	Random Tree	Random Forest	Naïve Bayes	Bayes Net	OneR	PART
Cross – Validation: 5	49.83%	51.18%	48.45%	57.75%	35.37%	45.66%	45.84%	51.68%
Cross – Validation: 10	50.15%	51.40%	49.04%	58.36%	35.42%	45.59%	46.13%	51.83%
Cross – Validation: 20	50.55%	51.63%	49.37%	**58.63%**	35.38%	45.66%	46.36%	51.92%
Percent Split: 66	49.11%	50.73%	47.13%	56.68%	34.94%	46.14%	45.49%	51.33%
Percent Split: 80	50.16%	51.97%	48.11%	58.11%	35.53%	46.22%	45.65%	52.07%
Percent Split: 90	50.76%	52.16%	48.52%	58.06%	35.44%	46.48%	46.28%	52.22%

Genre Classification Results

*Note: RandomForest excluded due to computing limits

Test Option	**Algorithm**	**J48**	**RepTree**	**Random Tree**	**Naïve Bayes**	**BayesNet**	**OneR**	**PART**
Cross – Validation: 5		33.55%	39.14%	29.79%	33.57%	40.11%	21.08%	33.33%
Cross – Validation: 10		33.60%	39.36%	29.21%	33.54%	40.17%	21.00%	33.54%
Cross – Validation: 20		33.52%	39.72%	29.10%	33.53%	**40.29%**	21.06%	33.64%
Percent Split: 66		33.75%	39.41%	29.53%	33.60%	40.24%	20.86%	33.71%
Percent Split: 80		33.57%	38.29%	29.38%	33.22%	39.42%	21.12%	33.05%
Percent Split: 90		33.82%	39.14%	29.10%	32.98%	39.12%	21.02%	33.32%

Analysis

The most successful linear model was a multiple linear regression model predicting Popularity using Acousticness, Danceability, Energy, Instrumentalness, Liveness, Loudness, Speechiness, and Valence. That linear model only accounted for 23.3% of the variance in the Popularity attribute. Thus linear regression is ineffective for predicting a song's popularity.

The most successful popularity classifier was a RandomForest with 20-fold cross-validation, though its accuracy was just over 58%. Thus, classification of popularity level based on numerical attributes is also ineffective.

The most successful genre classifier was a BayesNet with 20-fold cross-validation, which only achieved a success rate of 40.3%. Thus it is ineffective to classify amongst the ten genres provided in the dataset. Intuitively, it seemed as though genres should be able to be distinguished based on numerical attributes. As a 'bonus' data mining question, I decided to explore whether we can distinguish between just two genres.

I filtered the original dataset to only those songs classified as 'Rock' or 'Jazz'. Rock and Jazz seem like genres that would have different levels of instrumentalness, loudness, danceability, energy, and maybe even valence. Passing this filtered dataset through the PART rule-based algorithm (with an 80/20 split) yielded accuracy of 81.6%.

The composition of these genre-based findings indicate that telling two genres apart via machine learning is quite simpler and much more feasible than distinguishing between ten.

6. Conclusion and Reflection

Through this analysis, it has become apparent that there is no 'formula' for a successful song. Had the results been different, the information gained would enable any musician to improve the potential popularity of their song by conforming to the numerical attributes contributing most to popularity. The results show, however, that there is no meaningful relationship between numeric attributes of a song and its popularity.

Distinguishing between genres is less of a useful skill, but it could have applications in automated playlist-generation engines or some sort of media player which changes in design based on the characteristics of the song playing.

Had I had longer to work on this project, I would have analyzed how 'distinguishable' each pair of genres provided were, and which attributes were most significant in those distinctions. I would also like to reapply my analysis to a larger million-song dataset provided by Spotify to determine if the results are consistent.

7. References

Literature

Witten, I. H., Frank, E., Hall, M. A., &amp; Pal, C. J. (2017). Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann.

https://developer.spotify.com/documentation/web-api/reference/#/operations/get-several-audio-features

Dataset

https://www.kaggle.com/vicsuperman/prediction-of-music-genre?select=music_genre.csv

Tools

RStudio IDE
WEKA
Tidyverse package (https://www.tidyverse.org/)
GridExtra package (https://cran.r-project.org/web/packages/gridExtra/index.html)
Corrplot package
(https://www.rdocumentation.org/packages/corrplot/versions/0.84/topics/corrplot)
Leaps package
(https://www.rdocumentation.org/packages/leaps/versions/3.1/topics/leaps)
Ggrepel package (https://cran.r-project.org/web/packages/ggrepel/ggrepel.pdf)

Languages

R Programming Language