Predictive Power of Song Attributes on Music Genre and Popularity

Spotipythons

December 20, 2022

Annabelle Huether
amh9750@nyu.edu

Mary Nwangwu
mcn8851@nyu.edu

Allison Redfern
amr10211@nyu.edu

Isha Slavin
ivs225@nyu.edu

1 Introduction

Every December, millions of Spotify users receive and share their “Spotify Wrapped”, a personally-curated analysis of your top artists, songs, and genres. With the millions of songs that Spotify has to offer, we may ask, what are the attributes of a song that determine its genre or popularity? This report explores the impact of 9 different song attributes by determining the differences between genres given these attributes, which attributes can make a song more popular, and whether these attributes allow us to classify a song’s genre to ultimately approach the broader question: what makes a genre a genre?

To avoid any search algorithm bias, we started by creating a new Spotify account and searched for each Spotify-defined genre of interest. The 12 genres studied were: Anime, Broadway, Classical, Country, Dance/Electronic, Disney, Happy Holidays, Hip Hop, Jazz, Latin, Pop, and Rock. We selected the top playlists within each genre’s main page and combined the songs from each playlist into 12 genre-titled playlists (see Appendix 1). Next, we used “Organize Your Music” (see Appendix 2) to extract a dataset containing rows of songs and columns of attributes behind the songs as well as the Spotify-defined genre labels. The 9 attributes that are explored throughout this report are: Release Date, Beats per Minute (BPM), Energy, Danceability (Dance), Loudness (Loud), Valence, Length, Acousticness (Acoustic), and Popularity (Pop.) (see Appendix 3). Our dataset at this time included 15,525 samples.

There are obviously many songs that may not fit into just one genre. Therefore, we removed any songs that were included in more than one genre to only analyze music that fits distinctly within its genre. For ease of analysis, any songs that had missing values were also removed from the data set and song lengths were converted to seconds. Our dataset was reduced to 14,892 samples following data cleansing. Random seed 177669368 was used for any randomization.

2 Inference

2.1 Question

To begin, a natural first question would be to determine how the genres are different from each other based on the given attributes. This topic was explored by answering the following sub-questions: Which attributes have the most differences and similarities between genres? Which genre has the most attributes that are different from other genres? Which genre has the least attributes that are different from other genres? These questions can be explored through hypothesis testing.

2.2 Approach

In order to choose a hypothesis test, the assumption of equal variance must be tested for each genre per attribute. A Levene test (which assumes independent samples) with $\alpha = 0.005$ was used to test if the variance was the same across all genres for each attribute. Since every Levene test produced significant results (meaning the variance cannot be considered equal), Welch’s t-tests were chosen for comparison between each genre. By using Welch’s t-tests, it is assumed each attribute’s value is able to be reduced to its sample mean and that the data is normally distributed, but equal variance cannot be assumed among the attribute scores across genres. Welch’s t-tests were iterated across all 66 pairs of the 12 genres for 8 target attributes: Beats per Minute (BPM), Energy, Danceability, Loudness, Valence, Length, Acousticness, and Popularity. Release Date was not considered. The null hypothesis for each test is that the attribute is the same between two genres. The p-value from each test under the α-level of 0.005 is considered a significant result and allows us to reject the assumption that the null hypothesis is true.

Also within each test, the power ($1 - \beta$) was calculated using TTestIndPower() to determine the probability of the test finding a significant result if it is present. The effect size, measured by Cohen’s d, was calculated to determine the magnitude of difference between the means of the attribute between each genre pair. Power and effect size were key indicators in filtering down the results to only focus on larger and highly replicable significant differences between genres for each attribute. A large effect size was considered to be an absolute value of Cohen’s d over 0.8 and high power as $1 - \beta$ greater than 0.8.

2.3 Analysis

The Levene test for each attribute across all genres led to significant values since the p-value for each was below 0.005 (see Jupyter notebook for full results). Therefore, the variance among genres per attribute are not the same. As mentioned in the Approach section, the non-equal variance assumption led to the use of Welch’s t-tests to assess the differences between each pair of genres for each attribute (see Jupyter notebook for full results). A summary of the significant results are shown in Table 2A.

Table 2A. Significant Welch’s t-test results for 66 genre pairings per attribute

Result Type	**BPM**	**Energy**	**Danceability**	**Loudness**	**Valence**	**Length**	**Acousticness**	**Popularity**
Significant	78.8% (52)	97.0% (64)	89.4% (59)	89.4% (59)	81.8% (54)	89.4% (59)	97.0% (64)	93.9% (62)
Significant & High Power	71.2% (47)	89.4% (59)	89.4% (59)	86.4% (57)	78.8% (52)	83.3% (55)	92.4% (61)	90.9% (60)
Significant, High Power, & High Effect Size	**4.5%** (3)	**56.1%** (37)	**43.9%** (29)	**47.0%** (31)	**25.8%** (17)	**3.0%** (2)	**54.5%** (36)	**53.0%** (35)

Once significant results are filtered to only those with high effect sizes and power, 5 of the 8 attributes still have over 40% of the genre pairs as significant, meaning that these genre pairs can be interpreted as different within these attributes. Energy had the most genre pairs (37 of 66) different from others. The pair with the highest effect size in Energy was Anime vs. Classical (p-value: 0.000, test statistic: 90.7, effect size: 1.95, power: 1.00). See Figure 2A for a histogram of the results. Length of song was the most similar among genres as it only has 2 pairs of genres that are significant with high effect and power: Anime vs. Pop (p-value: 3.52×10^{-63}, test statistic: 18.1, effect size: 0.893, power: 1.00) and Anime vs. Disney (p-value = 2.28×10^{-47}, test statistic: 15.3, effect size: 0.861, power: 1.00). The histogram of the Anime vs Pop results are shown in Figure 2B.

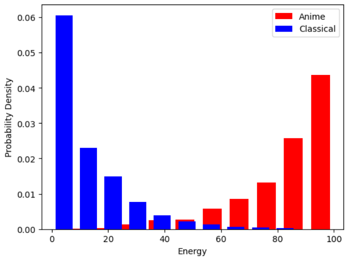


Figure 2A: Histogram of Energy for Anime vs Classical genres

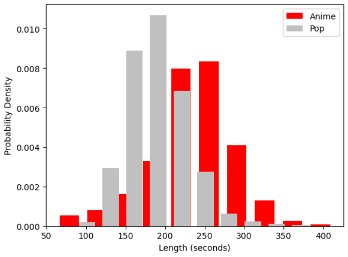


Figure 2B: Histogram of Length for Anime vs Pop genres

Using the Welch's t-tests results, genres pairs containing the most differences among attributes can be examined. Table 2B shows the count of attributes each genre pair differed (with high effect size and high power). The higher the value, the more attributes they differ in. Classical is the most different genre compared to others for the most of the 8 attributes since it on average differed in 5.73 of the 8 attributes. Country is the least different among genres with an average of 1.82 different attributes from other genres.

Overall from hypothesis testing, it can be concluded that there are significant differences between genres given attributes. Therefore, it can be explored whether these attributes can be predictors of other attributes and genre itself.

Table 2B. Counts of significant, high power, and high effect size attribute differences among genres

	Anime	Broadway	Classical	Country	Dance/Elec	Disney	Holidays	Hip Hop	Jazz	Latin	Pop	Rock
Anime		4	6	1	0	4	3	3	4	3	3	0
Broadway	4		6	4	5	3	0	5	0	6	5	4
Classical	6	6		6	6	5	5	6	5	6	6	6
Country	1	4	6		0	0	1	1	3	1	0	3
Dance/Elec	0	5	6	0		2	2	0	4	3	2	1
Disney	4	3	5	0	2		0	2	1	3	1	3
Holidays	3	0	5	1	2	0		3	0	3	1	3
Hip Hop	3	5	6	1	0	2	3		4	0	0	3
Jazz	4	0	5	3	4	1	0	4		4	2	4
Latin	3	6	6	1	3	3	3	0	4		1	5
Pop	3	5	6	0	2	1	1	0	2	1		4
Rock	0	4	6	3	1	3	3	3	4	5	4	
Average:	**2.82**	**3.82**	**5.73**	**1.82**	**2.27**	**2.18**	**1.91**	**2.45**	**2.82**	**3.18**	**2.27**	**3.27**
Total:	**31**	**42**	**63**	**20**	**25**	**24**	**21**	**27**	**31**	**35**	**25**	**36**

3 Prediction

3.1 Question

The next question of interest was: which attributes from the Spotify dataset are correlated with higher streams? Additionally, can this information be wielded to predict the popularity of a song? Regularized regression was utilized while controlling for confounds to tackle these questions.

3.2 Approach

During initial analysis, a slightly positive trend in popularity over time based on release year was evident (see Figure 3A). It was also evident from the figure that the minimum and maximum values of popularity change over time as well. The only features contained in the Spotify dataset that are relevant to time are 'Length' (length of song) and 'Release' (date in which song was released). In an effort to capture all information related to time, an additional seven attributes were generated: minimum pop. score, maximum pop. score, count of songs released, mean pop., median pop. which is robust to outliers, standard deviation of pop., and sum of popularity scores, all calculated by year. The 'Release' attribute was converted from timestamp to integer format of the year of release, and values for each calculated attribute were matched to each data point on this column.

When visualizing various feature values, differences in patterns of popularity by genre became evident (see Figure 3B). Different genres clearly displayed distinct trends in overall and median popularity shifts (for example: Anime versus Jazz), which is indicative of a potential strong correlation or dependency. To capture these possible effects, the categorical attribute Genre was converted into 12 indicator variables, each one representing a different genre. The indicator variables were binary with a value of '0' meaning the data point (unique song) was not of the genre and '1' meaning otherwise. These attributes were combined with the time-related calculated features as well as with the beats per minute, energy, danceability, loudness, valence, song length, and acousticness attributes to create a new feature set, and was cleaned in preparation for training and testing a regression model.

Using the aforementioned features, regression was utilized to predict the popularity of a song, which was normalized to range from 0 to 1. Certain measures were taken in order to avoid confounds. Firstly, in an effort to avoid overfitting a penalty factor was introduced to the weights found by regression through L-2 regularization, a method known as Ridge Regression. Ridge was chosen over Lasso Regression due to the small-sized nature of the dataset. This regularization term is multiplied by a lambda value which can range from 0 to infinity. In order to find the optimal value for the problem, hyperparameter tuning was implemented through the K-Fold cross-validation method with 10 splits. To remove the confound of certain attributes being measured at different scales, after dividing the dataset into an 80/20 train/test split a built-in scaler from the scikit-learn library was used to standardize the feature set, which transformed all data to resemble standard normally-distributed data. The regularized model was fit to the scaled training data and validated against unseen testing data to determine the efficacy of predictions.

3.3 Analysis

Upon examining a residual plot to validate that the model's estimates are unbiased (see Figure 3C), model performance was then assessed using validation metrics. The Ridge regularized regression model had a root mean squared error (RMSE) value of 0.12335 and a Coefficient of Determination of 0.50973. Therefore, the Ridge Regression model that was built had high accuracy in predicting popularity scores as the error in average difference between predicted and actual values was very low (~12%). Additionally, it was able to explain approximately 51% of the variance in popularity index, representing how well the model fit the data and how much variance in popularity it was able to capture.

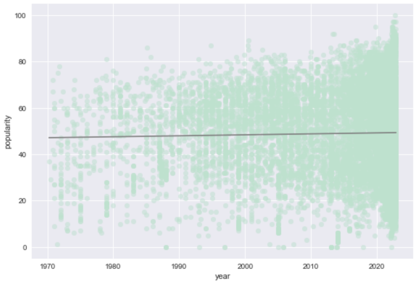


Figure 3A: Scatterplot of Pop. Over Time

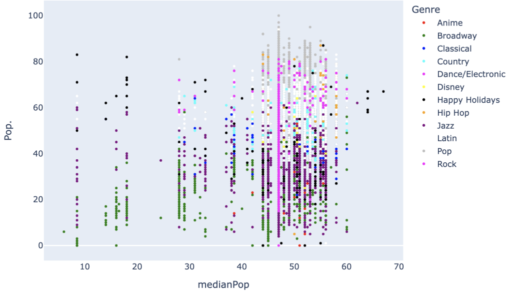


Figure 3B: Pop. by Genre

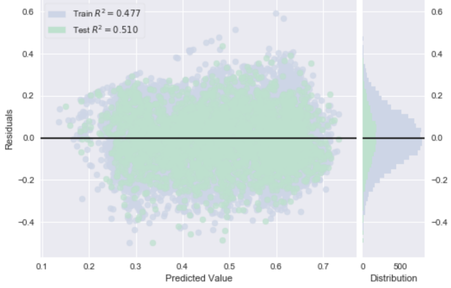


Figure 3C: RidgeCV Residual Plot

In order to run a regression analysis using p-values and F-statistics to determine coefficient significance, an additional Ordinary Least Squares regression model was built. It was trained and tested on the same split as the regularized regression model mentioned above (see Appendix 4 for full regression analysis and further explanation). Analysis showed that 18 out of the 27 total independent variables in the regression model were significant ($p<0.005$), including all genre attributes except for 'Genre_Rock'; information relating to the mean and median popularity by year were also significant. Furthermore, the F-statistic essentially representing the significance of regression coefficients in the multiple linear regression model was 425.8 and had a p-value of 0.00 showing extremely strong significance, leading to a rejection of the null hypothesis that all regression coefficients are 0.

On top of regression analysis, feature analysis was conducted for purposes of interpretation by ranking β values found by the Ridge Regression model. There were a total of 27 features in the feature set (12 genre-related, 7 time-related, 8 original or modified). Feature importance was determined by comparing magnitude of coefficients found by the model - when run on the regularized regression model, 'sumPop' (sum of popularity index by year) had the highest importance while 'Genre_Broadway', 'Genre_Jazz', 'countYear' (# of songs released by year) 'Genre_Pop', and 'Genre_Latin' followed closely behind.

Based on regression investigations, validation metrics, and feature importance, it can be asserted that genre is a very important attribute in predicting the popularity of a song. Furthermore, a model was created that is able to predict the popularity of a song based on information relating to gender, popularity shifts over time, and the other attributes defined in Organize Your Music that can explain over half of the variance in popularity values with low error. Due to these findings, a natural continuation of this investigation was to further explore genre and its possible classification.

4 Classification

4.1 Question

Given the results from inference and prediction, the data collected from Spotify's "Organize My Music" elicits a question of methodology: does Spotify use the obtained attributes to categorize its music library into broad genres? Furthermore, are these 8 attributes at all unique to these particular genres? This question can be explored using unsupervised machine learning methods of 1) Principal Component Analysis (PCA) and 2) K-Means Clustering in addition to the common supervised machine learning methods of 1) neural networks, 2) random forest classifiers, 3) boosting, and 4) support vector machines.

4.2 Approach

For the purposes of this analysis, the following 8 attributes were focused on: BPM, Energy, Dance, Loud, Valence, Length, Acoustic, and Popularity. Release Date was not considered. To help visualize and understand the dataset, PCA and K-Means Clustering were used to reduce dimensionality and to find natural groupings which have not been explicitly labeled in the data. PCA assumes a linear relationship between features, correlation between features, and no missing values. K-Means assumes that clusters are spherical, clusters are of similar size, and all attributes have the same variance. To then attempt to classify data into one of 12 genres, a neural network was built, in addition to a random forest classifier, a boosting classifier, and an SVM. Neural networks, random forests, and boosting do not have many formal assumptions since they are nonparametric models but they all generally assume independence of input features. SVM is generally parametric and assumes independence of input features and that data is identically distributed.

To approach this question from an unsupervised learning perspective, the labels for the Spotify generated genres were separated from the attributes of interest. Since PCA is a variance maximizing method, the attribute data was standardized in order to overcome skew due to the scale of variable value. The covariance matrix showing the pairwise feature correlation was obtained and the eigendecomposition of the matrix was performed to calculate the eigenvectors and their eigenvalues. The eigenvectors (principal components) are the directions of the axes where there is the most variance, while the eigenvalues give the amount of variance carried in each principal component. To find the optimal number of principal components to be considered, the Kaiser criterion was used to drop the components for which the eigenvalues of the standardized data were less than 1. Variance explained by the principal components selected was also computed. Once the principal components were found, the standardized attribute dataset was then visualized along with their original labels.

Silhouette analysis was used to measure the goodness of a clustering technique. The silhouette score is a measure of how similar a data point is within a cluster (cohesion) and compared to other clusters (separation). Various cluster numbers (from 2 to 9) were evaluated to determine the optimal number of clusters to be used for K-Means clustering analysis. Once the number of clusters *k* was determined, the standardized attribute dataset was visualized with these cluster groupings on the PCA determined axes using the K-Means clustering algorithm. Clusters formed were described based on a set of rules automatically generated by training a decision tree model (pruning level = 0.05) that used the original attributes and clustering result as the label.

To approach this question from a supervised learning perspective, a neural network was built and trained using the 8 attributes to classify data into one of 12 genres: Anime, Broadway, Classical, Country, Dance Electronic, Disney, Happy Holidays, Hip Hop, Jazz, Latin, Pop, and Rock. This network was a feedforward network using stochastic gradient descent to update weights and biases based on the error calculated through backpropagation. The activation function used for this neural network was the Sigmoid function with a one-hot encoder. The input layer contained 8 neurons for the 8 features, the hidden layer contained 10 neurons, and the output layer contained 12 neurons. Before training the network, the feature data was standardized in order to improve accuracy of the model. In addition, the data was split into a training set and a test set (and then a validation set) using an 80/20 split generated using the random state 177669368 to avoid overfitting. The neural network was trained and cross validated on 8 features and all 12 classes and tested on held out data; different hyperparameters for the network were also tested to see which improved accuracy the most. The final network utilized 30 epochs, mini-batch sizes of 10, and a learning rate of 3.0. A one-vs-rest scheme was used to build 12 ROC curves to additionally evaluate the neural network's performance on multiclass classification. In this process, the data for each of the 12 genres was separated out of the main dataset and labeled with a 1 and a random sample of data of equal size was taken from the rest of the dataset and given a label of 0. This data was then passed through the neural network after a reasonable test/train split and the performance of each genre was evaluated.

In order to corroborate the results from the neural network, random forest classifiers, boosting, and support vector machines were also utilized. For the random forest classifier, the data was split using an 80/20 test/train split to avoid overfitting. The data was left unstandardized for the random forest classifier due to standardization not affecting the results of the model. Hyperparameter tuning was performed with GridSearch; the optimal number of features to consider at each split was 0.3, the optimal minimum number of samples required at each leaf node was 300, and the optimal number of trees in the random forest was 55. For the boosting classifier and the SVM, the standardized data was used in this case since it maximized accuracy and was split using an 80/20 test/train split to avoid overfitting. For the boosting model a learning rate of 0.1, a max depth of 2, and 300 estimators were used. For the SVM, the optimal parameters were a c (or error) value of 10, a degree of 2, and an rbf kernel. For all 3 models, a K-fold 5 split cross validation method was used to improve the efficiency of the learning process. The performance of each of these models was consequently evaluated on held out test data.

4.3 Analysis

As PCA assumes correlation between features, associations between features were computed and visualized in a correlation matrix (Figure 4A). It is noted that attributes Loud and Energy have a strong positive correlation at 0.82 and Acoustic and Energy have a strong negative correlation at -0.81. Prior to dimensionality reduction, the attribute dataset's dimensions were (14892, 8). Once the attribute data was standardized and the corresponding eigenvectors were rearranged in order of decreasing eigenvalue, the Kaiser Criterion determined that only 2 principal components should be considered (Figure 4B). The dataset after dimensionality reduction became (14892, 2). The variance explained by the 2 principal components was 57.4%. The magnitude of the corresponding values in the eigenvectors informs us about the importance of each feature in the PCA output; the larger their absolute value, the more a specific feature contributes to that principal component. As observed in Table 4A, Energy, Loud, and Acoustic attributes are most important for PC1, while Length, BPM, and Popularity are most important for PC2.

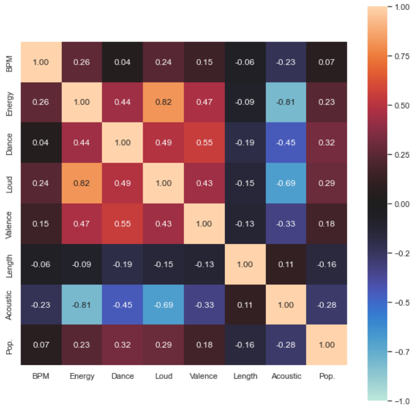


Figure 4A: Attribute Correlations

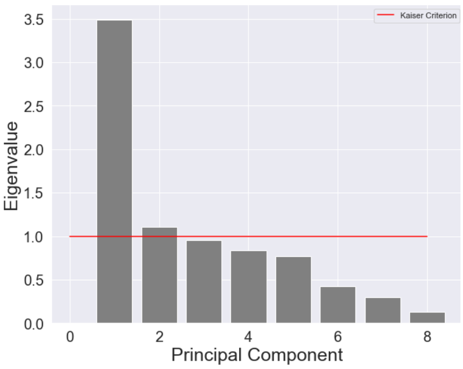


Figure 4B: Kaiser Criterion

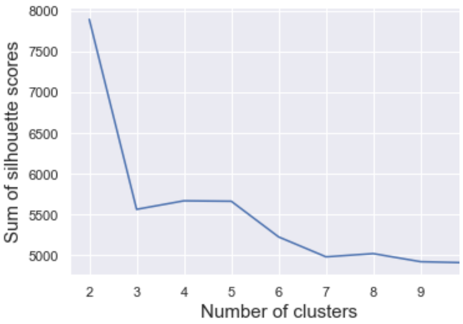


Figure 4C: Silhouette Analysis

Table 4A: PCA Components and Feature Importance

Attribute	BPM	Energy	Dance	Loud	Valence	Length	Acoustic	Popularity
PC1 Contribution	-0.171	-0.473	-0.375	-0.464	-0.346	0.131	0.445	-0.236
PC2 Contribution	0.465	0.260	-0.354	0.147	-0.176	0.572	-0.211	-0.409

PCA results were visualized using the standardized attribute data and the two principal components found using the Kaiser Criterion (Figure 4D). The target names (genre) for each data point were also included in the visualization to provide us a better understanding of where each song in each genre falls in accordance with the principal components.

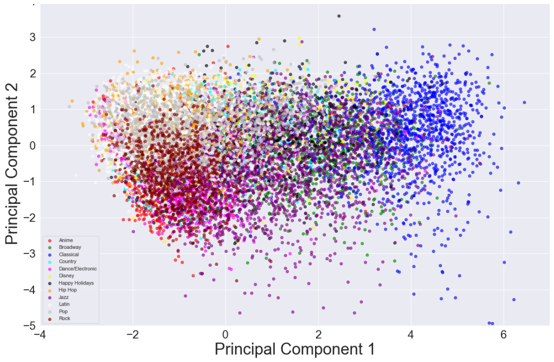


Figure 4D: Principal Component Analysis for Music Genres

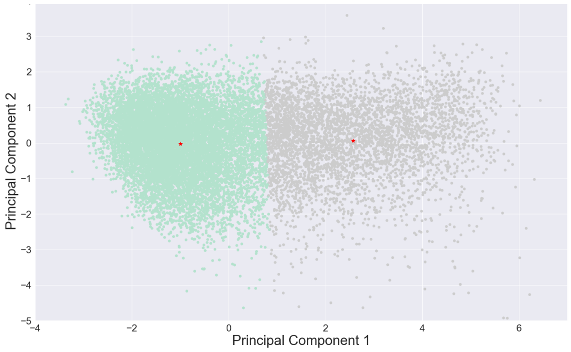


Figure 4E: K-Means Clustering with 2 Clusters

Silhouette analysis determined the optimal cluster number to be k=2 (Figure 4C). K-Means clustering of standardized attribute data with 2 clusters is shown in Figure 4E. Principal Component 1 has a higher spread of variation in comparison to Principal Component 2. Decision tree training and rules extraction method found that membership in each cluster was primarily defined by the Energy attribute. Samples with an energy level greater than 41.5 were placed in cluster 1 (mint) and those with an energy level less than or equal to 41.5 in cluster 2 (gray). Comparing the PCA and K-Means visualizations, the 2 defined clusters seem to mainly separate Classical music from most other genres available, where Classical music falls in cluster 2.

The neural network built for classification had an accuracy score of about 50% on held out data - indicating that only half of unseen data is classified correctly by the network. This result is better than chance - as the network guessing classes randomly would most likely result in an accuracy of about 12% - but by no means does this result characterize the model as one of good quality. For the one-vs-rest ROC analysis, the accuracies and AUC values are reported in Table 4B below. Of all 12 genres, the model built for Classical had the most accuracy when it came to classification, as seen in the ROC curve in Figure 4F. On the other hand, Disney had one of the worst accuracies (seen in the ROC curve in Figure 4G), with Anime just ahead in terms of accuracy. This supports the findings in previous sections about Classical being the most distinct genre in terms of attributes. In Figure 4H, the micro-averaged ROC curve is plotted. This plot implies that when the network makes classifications based on features a genre does not possess, rather than on the features it does, it is slightly more accurate. Table 4C reports the accuracies of the four classification methods on held out test data. Across all methods of supervised learning classification methods, the models only achieved about 47-56% accuracy when it came to classification of new song data into one of the 12 genres based on song features. Notably, this is better than random guessing, but it is far from good. The interpretation of this result is that machine learning models can only do an adequate job of determining which of 12 Spotify genres a song belongs to based solely on the features it possesses due to these features not playing a vital role in genre categorization.

Table 4B: One-vs-Rest ROC Analysis

Genre	**Anime**	**Broadway**	**Classical**	**Country**	**Dance/Elec**	**Disney**	**Holidays**	**Hip Hop**	**Jazz**	**Latin**	**Pop**	**Rock**
Accuracy	73.8%	84.2%	94.0%	75.7%	81.7%	65.3%	77.7%	76.9%	81.7%	80.3%	76.4%	81.7%
AUC	0.733	0.844	0.942	0.760	0.819	0.654	0.777	0.770	0.819	0.807	0.767	0.821

Table 4C: Classification Method Performance

Model	**Neural Network**	**Random Forest**	**Boosting**	**SVM**
Accuracy	50.1%	47.1%	54.7%	55.6%

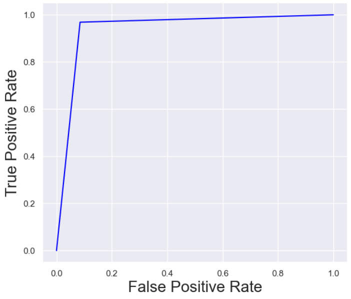


Figure 4F: ROC Curve for Classical

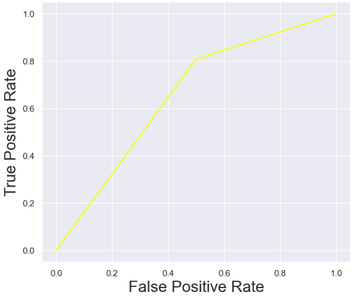


Figure 4G: ROC Curve for Disney

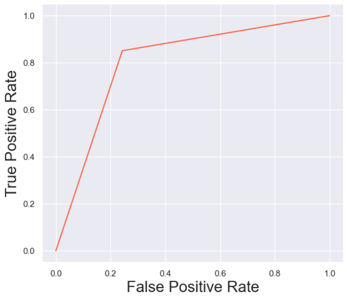


Figure 4H: Micro-Averaged ROC Curve

5 Conclusion and Limitations

When listening to a song, you may not consider the underlying attributes that make up that song, or even guess that they exist in the first place. Metrics like BPM or valence aren't immediately obvious or necessarily discernible to an average listener, but when extracted as data they can reveal much about the characteristics and individuality of a song that can be useful for categorization and comparison. In this study, we set out to determine if predefined Spotify genres were significantly different in terms of these attributes, and exactly how important these underlying song attributes were to song popularity and genre categorization.

Through Welch's t-tests on all pairs of genres per attribute we determined if there are any significant differences among genres for each attribute. We discovered that the Energy attribute has the highest number of differences between genres while Length has the least. When looking specifically at individual genres, Classical stood out as significantly different for most attributes when compared to most genres, while Country had the least number of different attributes compared to each genre. From this initial exploration, we determined that these attributes did have merit in distinguishing between genres.

When analyzing our dataset further through visualizations, we noticed varying levels of popularity through time and across genres. We sought to predict the popularity of a song based on statistics relating to the song's year of release as well as genre, and using these features - as well as BPM, energy, danceability, loudness, valence, song length, and acousticness - we were able to build a regression model that could predict popularity with relatively low error rates. Through regression analysis and feature importance calculated on an Ordinary Least Squares as well as regularized Ridge Regression model respectively, we calculated statistics that supported our hypothesis that almost every feature relating to genre had a significant effect on popularity prediction. Our regularized regression model's low error showcased its high accuracy in ability to predict the popularity of songs based on attributes taken or calculated from the original Spotify dataset. Overall, we were able to determine the importance of genre on popularity prediction and were able to build a high-performing model to predict this value. These findings and correlations were factors in demonstrating that the attributes we were exploring could be used to determine a song's genre.

After determining the potential predictive nature of the song attributes through inference and prediction, we explored unsupervised and supervised machine learning methods to see if we could classify a song's genre based on these attributes. PCA and K-Means clustering techniques were used to visualize and understand the natural groupings of the dataset. Through these methods, it was found that Classical had the clearest separation from other genres based on attribute contributions, consistent with the hypothesis tests. Since the dataset already came with predefined labels (genres), using these unsupervised learning methods would not generally be needed/recommended. However, since we do have these labels, interpretation of the K-Means clustering results is a lot easier. The 2 principal components defined by PCA only explain about 57% of variance in the dataset. Supervised learning on our preexisting genre labels corroborates these results. We interpret a 47-56% 12-genre classification accuracy to mean that song attributes are somewhat involved in Spotify's process of classifying songs into broad genres, but as the genres are categorized, songs within these labels do not possess unique enough features in order to be easily separated. In other words, classic machine learning models cannot easily be trained to pinpoint which of 12 Spotify genres a song belongs to based solely on the features it possesses. It is concluded, therefore, based on the K-Means clustering, PCA and supervised learning models that attributes alone do not define genre categorization; there could be underlying information that is not possessed in this study that Spotify uses instead of or alongside song features in order to categorize their music library.

This conclusion introduces the first limitation to our study, which is the fact that there are only 9 song attributes available for us to draw conclusions from about the dataset using inference, prediction, and classification. It is noted that perhaps there are other features such as additional song attributes, artist preference, underlying algorithms etc, that might factor into whether a song is placed into a certain genre. In addition, this study is generalizable only to the predefined genres in the Spotify music library search, i.e. it is only examining samples and labels predetermined by the algorithm that Spotify uses to categorize songs and playlists.

We also note the limitation of the imbalanced sample groups for each genre: since the Spotify library is only so large and their algorithm does not necessarily have to place equal amounts of playlists and songs into the top playlists for a particular genre, each genre had a different amount of playlists and songs which led to unequal sample sizes. In addition to these limitations, our study only involves 12 self-chosen genres, implying that there may be nuances in other genres that this study does not account for.

The limitations to our analysis imply how this study could be improved. In an ideal world, we would have exclusive access to information about Spotify's specific methods for categorization (perhaps by working for them) and would be able to extract the missing attributes and expand the number of samples in the study. This way, we could add missing attributes and data to our analysis and likely arrive at conclusions with higher significance, accuracy, and predictive power. Regardless, there is still merit in the conclusions drawn from this study. Though not a high enough amount for extremely accurate classification, we find that there is still a considerable amount of separation and difference between genres - pointing to the predictive power of the underlying attributes of songs. In this way we can attempt to use data to answer the age old question: what makes a genre a genre? The results of this study reveal that there is certainly something to be said about how song attributes define genres. Maybe country songs aren't defined by how many times the word "truck" or "beer" appears in a song, but rather by a meaningful combination of beats per minute, energy, danceability, loudness, valence, length, acousticness, and popularity.

Something remarkable that we were also able to extract from the dataset was the significant popularity difference between songs released by multi-genre and single-genre artists. The granularity of our data is at the song level, so our original dataset included the name of the Artist for each particular song. In an effort to explore the ubiquity of individual artists in the dataset, we investigated artists that released across numerous genres. Out of 7,651 total unique musical artists, 9 artists appeared in four genres (the max. # of genres for one artist to be a part of), 48 artists appeared in three genres, and 459 artists appeared in two genres; the rest were single-genre artists. This comprised a total of 516 multi-genre artists, which made up 2,497 released songs (and thus that many popularity scores). We randomly selected 2,497 popularity scores from songs released by single-genre artists and performed a Welch's t-test under the assumption that the sample variances differed to determine whether the population means of our two samples differed. The test showed a highly significant difference between mean popularity scores for songs released by single-genre vs. multi-genre artists ($t=22.7$, p-value=$5.03 \times 10^{-109} << 0.005$). This was an interesting find as we did not use artist information while building regression models to predict the popularity of a song, yet we would consider including quantitative data relating to artists such as genre count information in the future due to these results.

See Appendix 5 for Author Contributions.

Appendix

1. Details Behind Spotify Playlists Used Per Genre

Spotify Genre	Number of Songs	Spotify Playlists
Anime	629	***Popular Anime Playlists:*** Shonen, Anime Now, Anime Hits, Anime on Replay, This is MAPPA, Anime Rewind ‘80s, Anime Rewind ‘90s, Anime Rewind ‘00s, Anime Rewind ‘10s
Broadway	1101	***Broadway Playlists:*** Best of Broadway, Off Broadway, Tony Awards: Best Musical Winners, Broadway Belts, Showstoppers, Opening Numbers, Broadway Ballads & Duets, Broadway in Love!, I Want My Broadway, Broadway Family Sing Along, Broadway Overtures, Les Miserables, Wicked, The Phantom of the Opera, Mamma Mia, Cabaret, Cats, West Side Story, My Fair Lady, Chicago, A Chorus Line, Hello Dolly!(Original Cast), Into the Woods, The Color Purple, The Lion King, The Best of Rent, Miss Saigon, Jersey Boys, Hello Dolly!(New Cast), Annie, Evita, Tony Awards Best Original Score
Classical	1338	***Classical Featured Playlists:*** Classical Essentials, Classical New Releases, Calming Classical, Pop Goes Classical, Dark Academia Classical, Classical Piano, Classical Cooking, Reflection, Feel Good Classical, Shimmering Strings, Soft Piano, Royalcore, Baroque Classics, Light Academia, Soundtracked, Medieval Vibes, Classical Romance, Creative Writing, ClassicalX, Mellow Cello, Classical Garden, Mellow Classical, Dramatic Classical, Hopeless Romantic, Soundtracks for Studying
Country	1109	***Popular Country Playlists:*** Hot Country, New Boots, 90’s Country, Chillin’ on a Dirt Road, Country, Worldwide Hot 50, Country Gold, 2010s Country, 2000’s Country, Country’s Greatest Hits, Breakout Country, Big Country, Energy Booster: Country, New Music Friday Country, 80’s Country
Dance/Electronic	1615	***The Best of Dance & Electronic Today:*** Mint, Dance Rising, Housewerk, Bass Arcade, Tantra, Night Rider, Happy Beats, Brain Food, Dubstep Don, Dance Hits, Dance Party, Bass Lounge, Techno Bunker, Fresh Dance Pop, Metropolis, Planet Rave, Operator, EDM, Jersey Club Heat
Disney	471	***Disney Playlists:*** Disney Channel Hits, Disney Favorites, Disney Hits, Disney Love Songs, Disney Classics, Disney Princess, Disney Sing Alongs
Happy Holidays	1161	***It’s Never Too Early…:*** Christmas Classics, Christmas Jazz, Christmas Pop, Country Christmas, Electronic Christmas, Hip Hop Christmas, Holiday Magic, Indie Christmas, Latin Christmas, Merry & Bright, Navidad Cristiana, New Music Holiday, Rock Christmas, Soulful Christmas, Acoustic Christmas, Spotify Singles: Holiday Collection
Hip Hop	1447	***Popular Hip-Hop Playlists:*** RapCaviar, Feelin’ Myself, Most Necessary, I Love My 2010s Hip-Hop, Get Turnt, Signed XOXO, State of Mind, Out The Mud, No Cap, CST, Westside Story, B.A.E, Mind Right, Alternative Hip-Hop, Workout Twerkout, Rap Workout, Hip Hop Controller, Mellow Bars, Hip-Hop Drive, New Joints, Beast Mode Hip-Hop, City to City, Pressure, I Love My Underground Classics, I Love My Midwest Classics, I Love My East Coast Classics, I Love My Down South Classics, I Love My West Coast Classics, Power Gaming, Spilled Ink
Jazz	1966	***Featured Jazz Playlists & Shapes of Jazz:*** State of Jazz, Fresh Finds Jazz, Orbit, Jazz Classics Blue Note Edition, Pocket, All New Jazz, Vocal Jazz, EQUAL: Jazz, Easy Jazz, Jazz Funk, Jazztronica, Fusion Fest, Acid Jazz, Prog Jazz, Latin Jazz, String Theory, Bottoms Up, Avant-Jazz, Shimmer, Jazz Live & Loud, Jazz Piano Today, 21st Century Jazz
Latin	2103	***Popular Latin Playlists:*** Viva Latino, Baila Reggaeton, Old School Reggaeton, La Lista Pop, MANSION REGGAETON, Los Que Mandan, Exitos USA, Bachata Classics, Fuego, Salsa Classics, Rock en Espanol, Latin Hit Mix, Cumbia Sonidera, #SadCuhHours, 100% Cumbia, Bachata Lovers, Mal de Amores, Super Cumbias, Canciones del Recuerdo, Puro Perreo, Latin Party Anthems, Mexican Party Anthems, Dembow Pegao, Salsa Nation, Perrear Y Llorar, Trap Land, Workout Latino, Baladas Romanticas, Bad bunny perreo mix, Latin Pop Classics, Ensenciales, Fiesta
Pop	1163	***Popular Pop Playlists:*** Today’s Top Hits, Hot Hits USA, Teen Beats, Pop Rising, Young & Free, Just Good Music, Sad Hour, Party Hits, New Music Friday, Pop Drive, Fresh Finds Pop, Chill Pop, SALT, Soft Pop Hits, K-Pop ON!, Fresh Dance Pop, Pop Sauce, BBE, Mega Hit Mix, OBSESSED, La Lista Pop, Hot Rhythmic, Dance Pop Hits, Fresh and Chill
Rock	789	***The Pulse of Rock:*** Rock This, Alt NOW, All New Rock, Rock Hard, The New Alt, New Noise, misfits 2.0, The Locker, Fresh Finds Rock, Rock Frequency, Legends Only, Alternative Noise, Hard Rock, Rock Rising

2. “Organize Your Music” Description

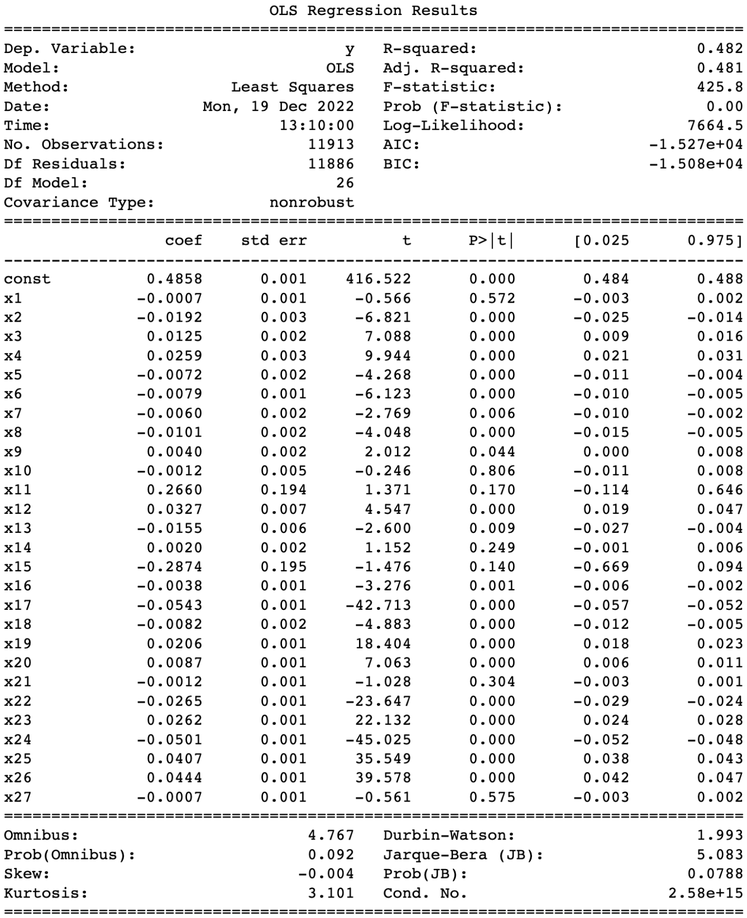
“Organize Your Music” takes Spotify playlists as input and outputs data and attributes behind each song in each playlist. The attributes that were available in “Organize Your Music” when the data used in this report was extracted (11/18/2022) are described in Appendix 3.

3. Definition of Musical Attributes from “Organize Your Music”

Beats Per Minute (BPM) - The tempo of the song
Energy - The energy of a song - the higher the value, the more energetic the song is
Danceability - The higher the value, the easier it is to dance to this song
Loudness - The higher the value, the louder the song
Valence - The higher the value, the more positive the mood for the song
Length - The duration of the song
Acoustic - The higher the value, the more acoustic the song is
Popularity - The higher the value, the more popular the song is

4. OLS Regression Analysis for Popularity Model (significance: alpha = 0.005)

Regression Analysis is available in MatLab, R, and Stata; however, it is not available in Python for regularized regression. In order to run this analysis, the statsmodels.api library needs to be utilized. Methods to fit the model include .fit() and .fit_regularized(), the latter of which is used for regularization. However when using .fit_regularized(), the only parameters that are calculated and visible to a user are the regression coefficients; at the moment, the current version of this package does not have the functionality of calculating coefficient significance. Thus unless calculated by hand, regression analysis in the form of p-values and F-statistics is only accessible when using .fit(), or Ordinary Least Squares regression. To analyze the Prediction component of this paper, we built and analyzed an OLS model to predict popularity with 27 attributes contained in the feature set. The following chart displays analysis results.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.482
Model:                            OLS   Adj. R-squared:                  0.481
Method:                 Least Squares   F-statistic:                     425.8
Date:                Mon, 19 Dec 2022   Prob (F-statistic):               0.00
Time:                        13:10:00   Log-Likelihood:                 7664.5
No. Observations:               11913   AIC:                        -1.527e+04
Df Residuals:                   11886   BIC:                        -1.508e+04
Df Model:                          26
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.4858      0.001    416.522      0.000       0.484       0.488
x1            -0.0007      0.001     -0.566      0.572      -0.003       0.002
x2            -0.0192      0.003     -6.821      0.000      -0.025      -0.014
x3             0.0125      0.002      7.088      0.000       0.009       0.016
x4             0.0259      0.003      9.944      0.000       0.021       0.031
x5            -0.0072      0.002     -4.268      0.000      -0.011      -0.004
x6            -0.0079      0.001     -6.123      0.000      -0.010      -0.005
x7            -0.0060      0.002     -2.769      0.006      -0.010      -0.002
x8            -0.0101      0.002     -4.048      0.000      -0.015      -0.005
x9             0.0040      0.002      2.012      0.044       0.000       0.008
x10           -0.0012      0.005     -0.246      0.806      -0.011       0.008
x11            0.2660      0.194      1.371      0.170      -0.114       0.646
x12            0.0327      0.007      4.547      0.000       0.019       0.047
x13           -0.0155      0.006     -2.600      0.009      -0.027      -0.004
x14            0.0020      0.002      1.152      0.249      -0.001       0.006
x15           -0.2874      0.195     -1.476      0.140      -0.669       0.094
x16           -0.0038      0.001     -3.276      0.001      -0.006      -0.002
x17           -0.0543      0.001    -42.713      0.000      -0.057      -0.052
x18           -0.0082      0.002     -4.883      0.000      -0.012      -0.005
x19            0.0206      0.001     18.404      0.000       0.018       0.023
x20            0.0087      0.001      7.063      0.000       0.006       0.011
x21           -0.0012      0.001     -1.028      0.304      -0.003       0.001
x22           -0.0265      0.001    -23.647      0.000      -0.029      -0.024
x23            0.0262      0.001     22.132      0.000       0.024       0.028
x24           -0.0501      0.001    -45.025      0.000      -0.052      -0.048
x25            0.0407      0.001     35.549      0.000       0.038       0.043
x26            0.0444      0.001     39.578      0.000       0.042       0.047
x27           -0.0007      0.001     -0.561      0.575      -0.003       0.002
==============================================================================
Omnibus:                        4.767   Durbin-Watson:                   1.993
Prob(Omnibus):                  0.092   Jarque-Bera (JB):                5.083
Skew:                          -0.004   Prob(JB):                       0.0788
Kurtosis:                       3.101   Cond. No.                     2.58e+15
==============================================================================
```

5. Author Contributions

Allison Redfern: Inference Analysis & Report
Isha Slavin: Prediction Analysis & Report
Mary Nwangwu: Classification - Unsupervised Learning Analysis & Report
Annabelle Huether: Classification - Supervised Learning Analysis & Report
All: Data Discovery, Data Cleansing, Ideation, Introduction, Conclusion, Extra Credit, Merging Files on GitHub, Editing and Reviewing Report