CAPSTONE PROJECT

MAY 2024

KATIE CARLSON

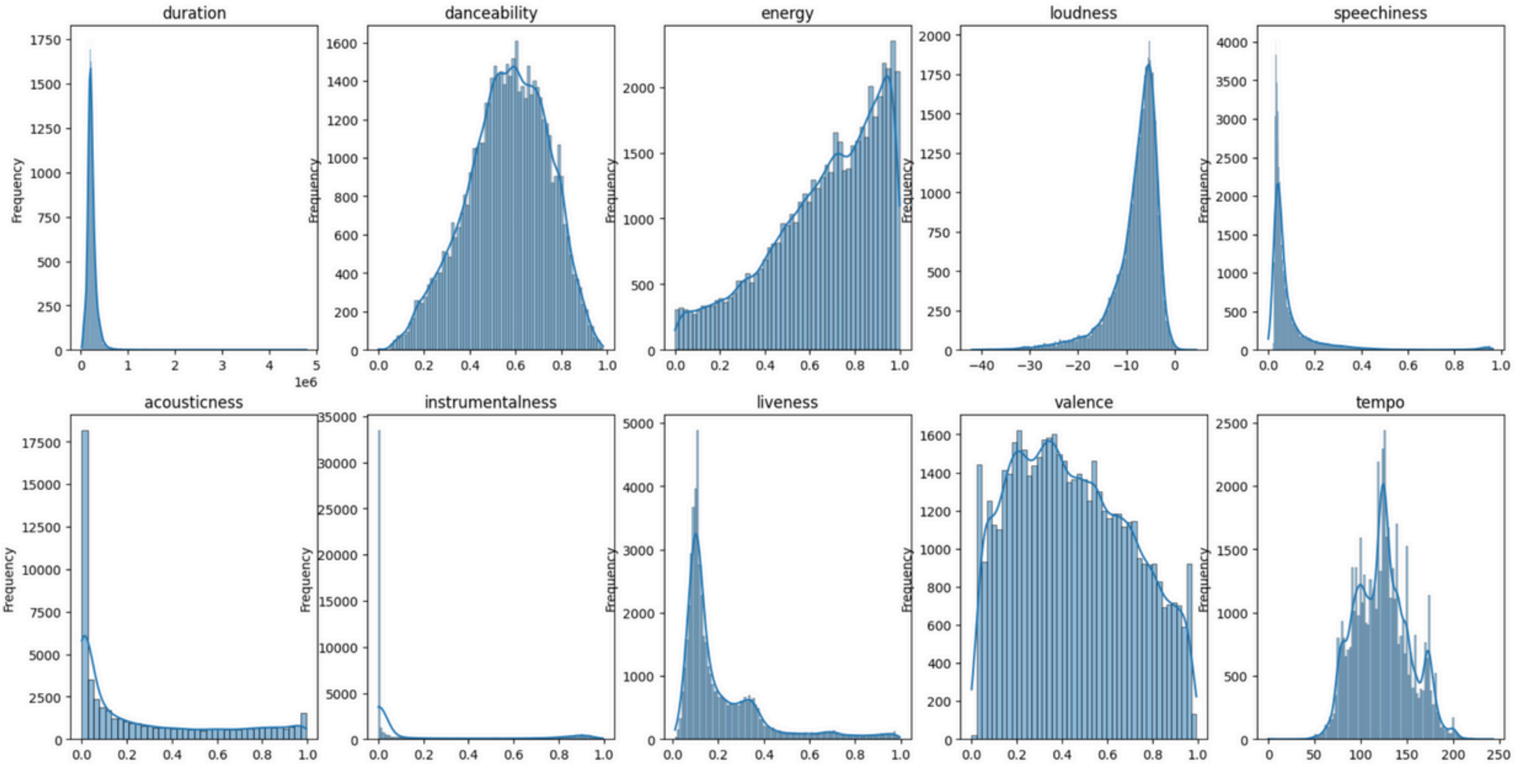
Introduction

For this project, I utilized the following Python libraries: Pandas, Numpy, Seaborn, and Sci-Kit Learn. I loaded the data by reading the CSV file using Pandas. I removed any potential missing values using the "dropna" method and then proceeded. I also flattened the array inputs before making histograms. Before performing a principal component analysis (PCA), I scaled the data with a standard scalar function. I seeded the random number generator (RNG) by setting the random-state argument of the sklearn model selection method to my student number, N16594496.

The following sections will address each step in my analysis.

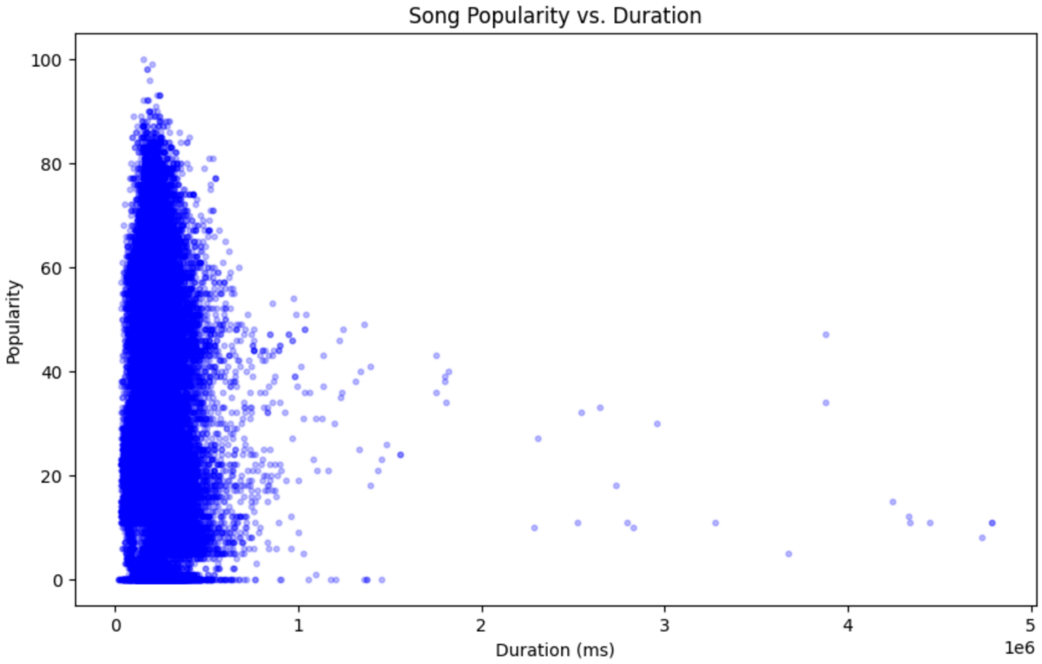
Feature Distributions

For each song feature, I plotted a histogram of the distribution across the data set. Next, I visually inspected for potentially normal distributions. Out of the 10 distributions, the "danceability" and "tempo" features appear to be reasonably normally distributed. The other distributions appear to be skewed in various ways. The following distributions are displayed below:



Length and Popularity

To determine the relationship between song length and popularity, I created a scatterplot of the popularity versus duration of each song. Then, I visually inspected the plot for a clear relationship between the two variables. The scatterplot is displayed below:

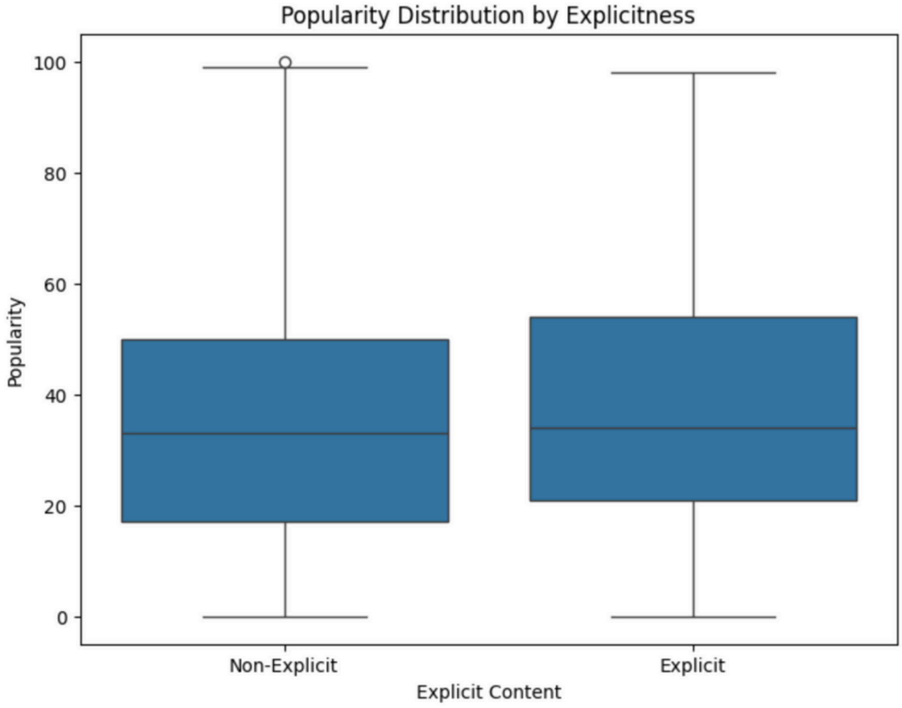


Based on visual inspection, there appears to be no relationship between song length and song popularity. At any given duration, the song popularity takes on a wide range of values.

*Note: Spotify quantifies song "popularity" as an integer on a scale of 0 to 100, with higher numbers corresponding to more plays on Spotify.

Explicitness and Popularity

To determine whether explicitly rated songs are more popular than non explicit, I performed the Mann-Whitney U test, a non-parametric test that does not assume a normal distribution in the inputted data set. I previously determined the explicit feature distribution to be non-normal. Using the conventional alpha level of 0.05, there appears to be a statistically significant difference between the popularity of a song based on its explicit rating. The test statistic used was equal to 139361273.5 and the resultant p-value was equal to 3.07e-19 (negligibly small). Below, I have provided box plots of the distributions of explicit and non-explicit songs for visual comparison.

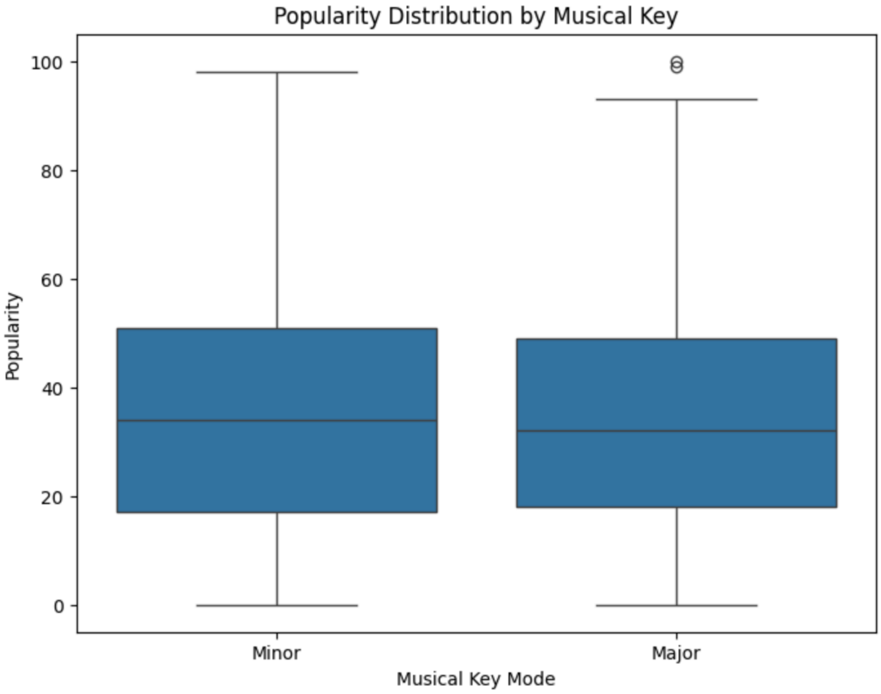


*Note: The explicit rating of a song is quantified here as a Boolean categorical variable. It is true if a song contains explicit language.

*Note: A visual inspection of the distribution of popularity indicates that the median for explicit songs is *slightly* higher.

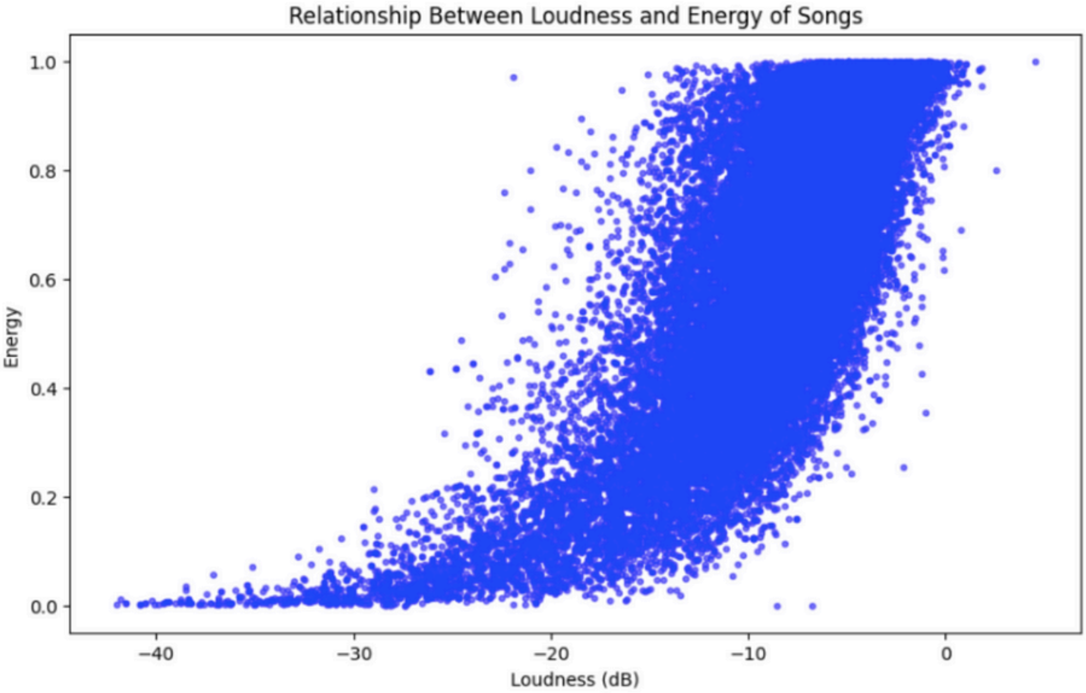
Key and Popularity

Similar to explicitness, the "key" feature of a song appeared to be non-normally distributed based on the previously displayed histograms. Therefore, I performed the Mann-Whitney U test again in order to determine whether major key songs are more popular than minor key songs. Setting the conventional alpha level to 0.05, the test results indicate that there is a statistically significant difference in song popularity based on the key. The test statistic used was equal to 309702373.0 and the resultant p-value was equal to 2.018e-6 (negligibly small). Below, I have provided box plots of the distributions of major and minor key songs for visual comparison.



Energy and Loudness

For this section, I was tasked with either substantiating or refuting the claim that the “energy” feature is believed to largely reflect the “loudness” feature of a song. To answer this question, I created a scatterplot of loudness and the energy of each song in the data set. This scatterplot is displayed below:



Upon visual inspection, the scatterplot suggests a positive correlation between the “loudness” and “energy” of a song. As the song gets louder, the energy tends to increase. The Pearson Correlation coefficient between “loudness” and “energy” is equal to 0.775, indicating a strong, positive correlation. This substantiates the claim that the energy tends to reflect the “loudness” of a song. Since there does appear to exist a relationship between the two, the energy does reflect the loudness of a song. Therefore, I would substantiate the initial claim.

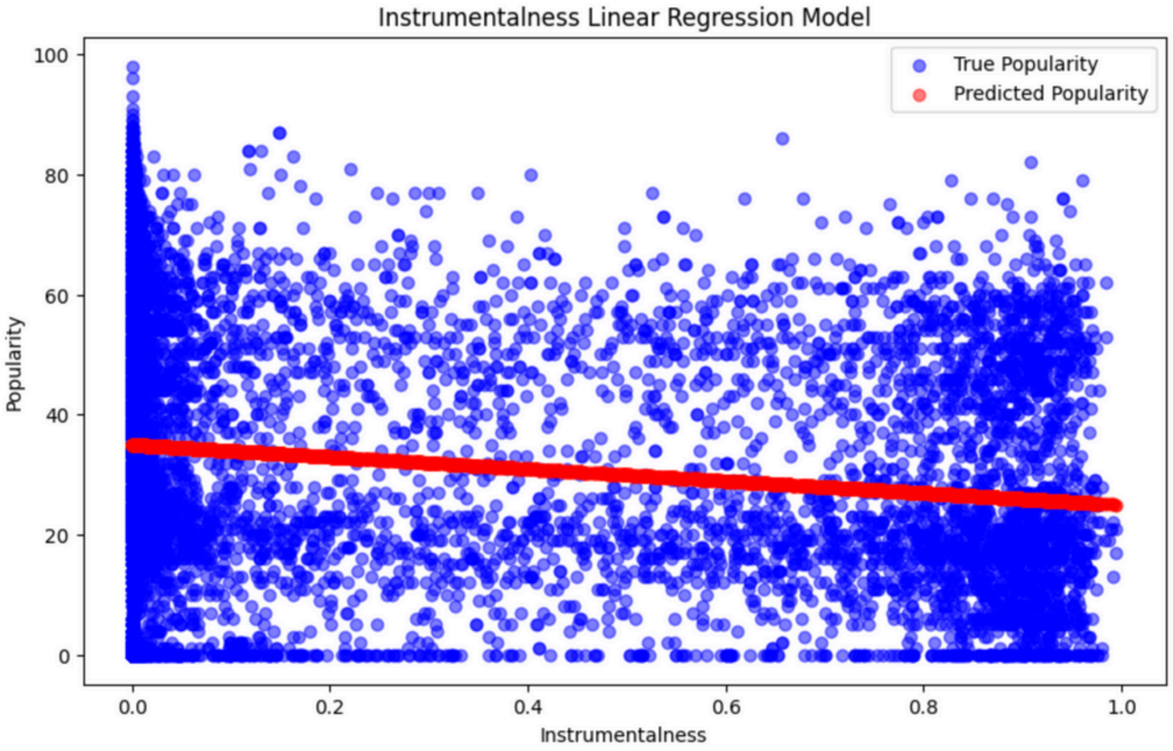
Best Predictor of Song Popularity

To determine which of the 10 song features best predicts popularity, I performed a simple linear regression for each song feature using the popularity as a response variable. The R-squared values for each regression model are displayed below:

Song Feature:	R-Squared Value:
Duration	0.0023361
Danceability	0.001554
Energy	0.003495
Loudness	0.001690
Speechiness	0.002321
Acousticness	0.0006777
Instrumentalness	0.01718
Liveness	0.001887
Valence	0.001514
Tempo	-7.586e-05

Overall, all of the models display a very weak predictive power. Based on the R-squared values, “instrumentalness” appears to be the best predictor of popularity out of the 10 song features, able to explain about 1.718% of the variance in popularity.

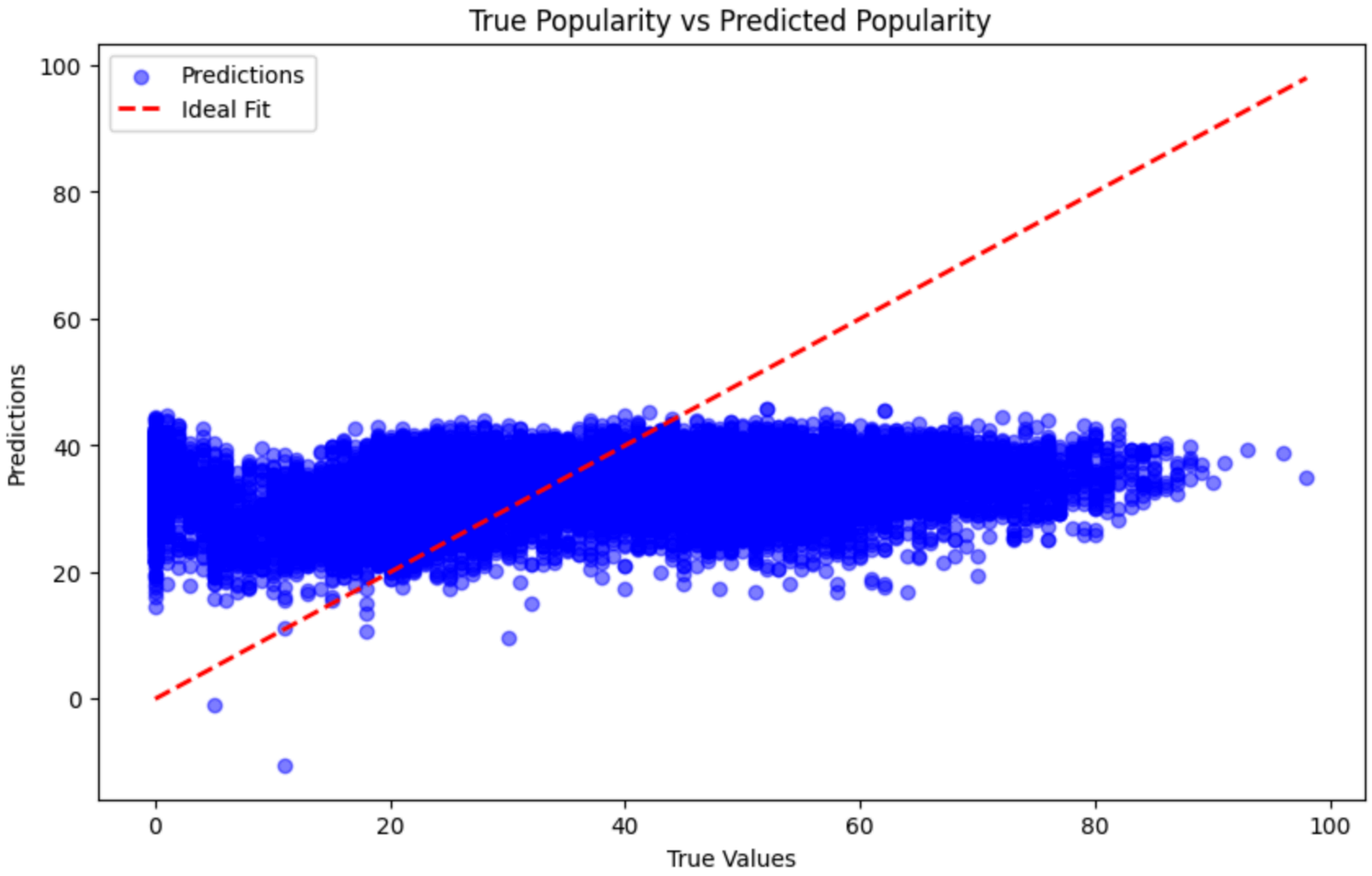
Next, I isolated the “instrumentalness” model and created a scatterplot of the true and predicted popularity values. This scatterplot is displayed below:



This scatterplot confirms my previous findings about predictive power. Upon visual inspection, it is clear that there exists an almost negligible relationship between the true and predicted popularity values. The true values are equally scattered above and below the predicted value line.

Predicting Popularity

In this section, I sought out to build a model utilizing all of the song features to predict popularity. Here, I used a multiple linear regression. The R-squared value for this model, 0.0418 indicates that about 4.2% of the variance in song popularity can be explained by all of the features. This is still a weak predictive power; however, it is significantly better than the best predictive model out of the individual features. Graphed below is the true value for song popularity compared against the predicted value for a song's popularity score. The red fit line displays the shape of the curve where the predicted values match the true values.



Visually, it is evident that this model has weak predicting power. Still, the comparison between the multiple versus simple linear regressions suggests that other features play a role in determining a song's popularity.

*Note: In comparison to the model for "instrumentalness", this model has improved by 225%.

Principal Component Analysis (PCA)

I was able to extract 7 meaningful principal components when considering the 10 song features. I set my threshold so the components are able to explain at least 95% of the variance in the data. The proportion of variance in the data explained through each principal component is described by the table below:

Component:	Explained Variance Ratio:
1	0.3670
2	0.2427
3	0.1422
4	0.0934
5	0.0504
6	0.0419
7	0.0283

Component 1 explains roughly 36.7% of the variance in the data, the most out of the components. Based on the explained ratios for each song feature, Component 1 likely differentiates loud and energetic songs from instrumental and acoustic songs.

*Note: Cumulatively, the first 3 principal components are able to explain roughly 75% of the variance in the data.

Key Prediction

In this section, I was tasked with predicting whether a song is in a major or minor key—quantified by Spotify as the "mode"—based on the valence. To accomplish this, I built a logistic regression model.

My model received an accuracy score of 0.6231, meaning it is able to accurately predict the key about 62.31% of the time. The model's ROC-AUC score is equal to 0.505, indicating that valence is not a strong predictor of key.

Genre Prediction

Finally, I investigated which of the following is a better predictor of whether a song is classical music: duration or the principal components I previously extracted. To prepare the data, I converted the genre label to a binary numerical of classical (1) or not classical/everything else (0). Then, I built two logistic regression models, one for the duration and one for the principal components to compare against each other's predictive capabilities. I judged these models based on their accuracy and ROC-AUC scores.

The duration model was highly accurate, able to successfully predict the classical or not classical music genre around 98.15% of the time. The principal components model was also highly accurate, able to predict the classical or not classical genre around 98.21% of the time.

By contrast, the duration model received an ROC-AUC score of 0.5690, indicating a significantly weaker ability to distinguish between the classes. This vastly differs from the principal components model, which received an ROC-AUC score of 0.9677. Based on this evidence, I would conclude that the principal components are a better predictor of whether or not a song is classical music.