Cassandra Clemente

Professor Pascal Wallisch

Principles of Data Science

20 December 2023

Capstone Project: Analysis of Spotify Song Data

Preprocessing

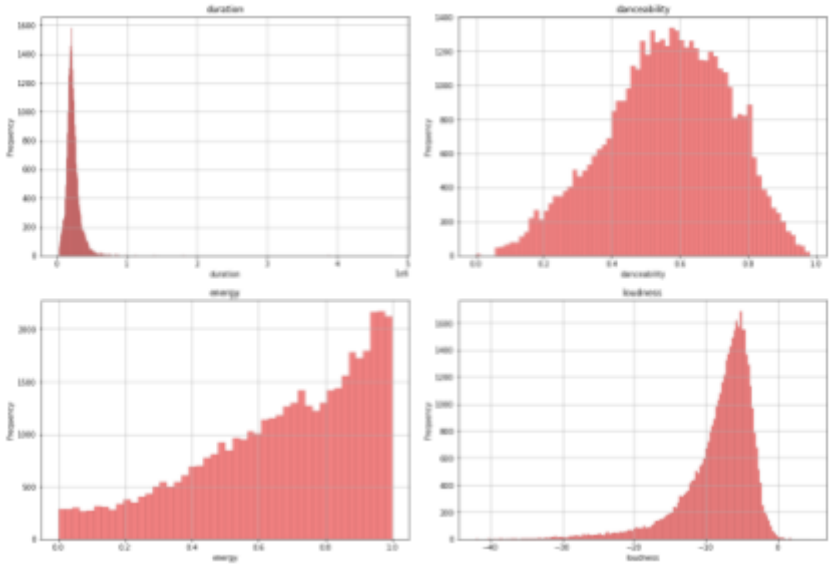
I used pandas to import the data set as a csv. To handle duplicates, I sorted the dataset by popularity and then dropped duplicates with all the same other features. I thought about keeping duplicates with all other columns the same except for ‘track_genre’ because songs can be placed into multiple genres, but there were about 5000 duplicates I would be keeping that would affect all of the other columns, so I decided to drop those as well. Since I sorted the data set by popularity, the highest popularity was kept as songs might be uploaded to spotify as a single, ep, music video, song in an album. These popularities for duplicates are understated, but combining their popularity (1-100 so I am assuming it is a type of percentile calculation) would not be possible without knowing the sample size of popularity.

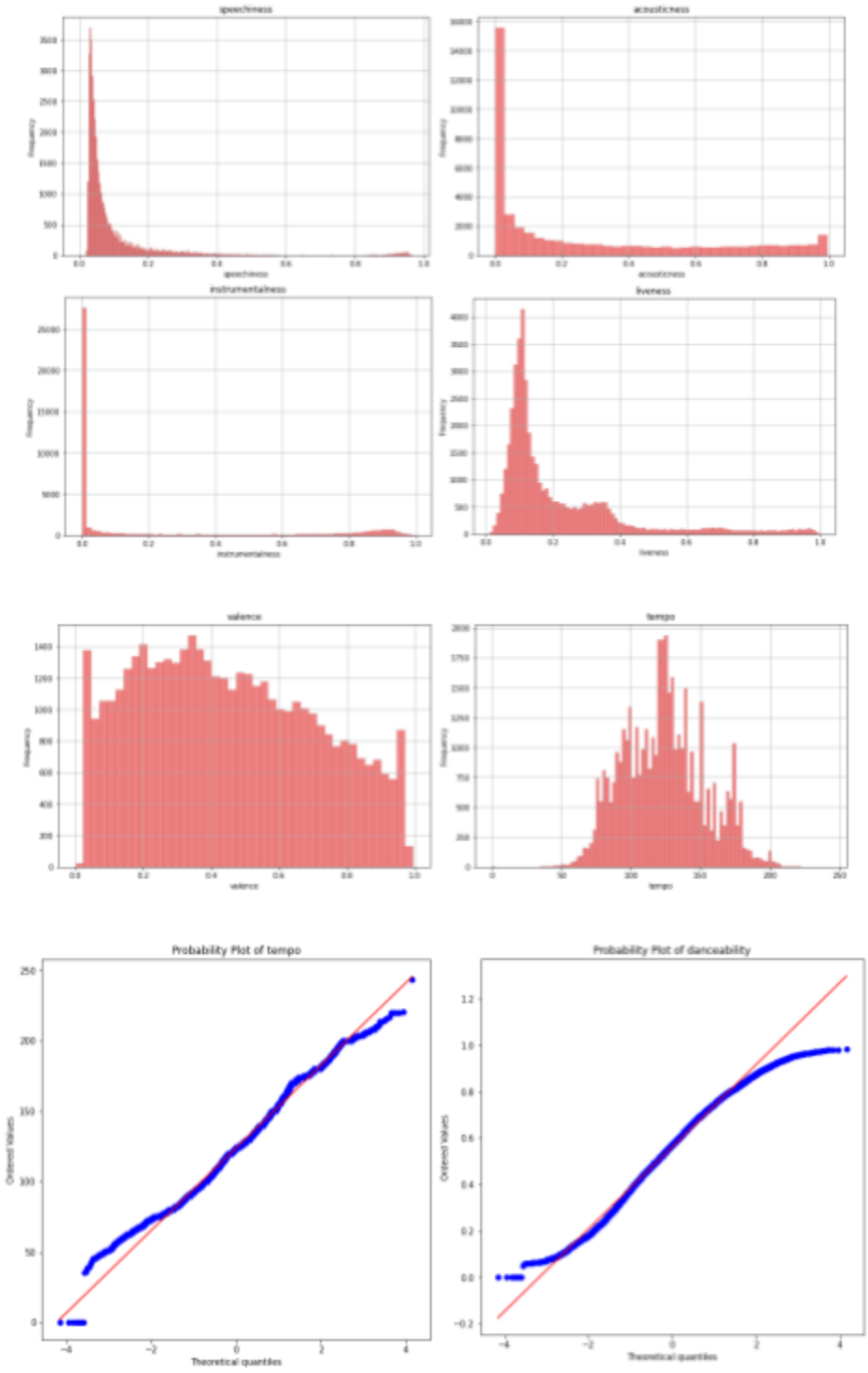
I seeded the RNG with my N-number and set a random_seed to my N-Number, so that random calls, and random states during train-test split would be unique to me.

Question 1

Consider the 10 song features duration, danceability, energy, loudness, speechiness, acousticness, instrumentalness, liveness, valence and tempo. Are any of these features reasonably distributed normally? If so, which one?

Tempo and danceability are reasonably normally distributed. At initial inspection, using a histogram, duration, energy, loudness, speechiness, acousticness, instrumentalness, liveness, and valence were eliminated due visual characteristics of their distributions that do not resemble normal distributions. This left tempo and danceability to further investigate. A probability plot showed tempo as normally distributed and danceability as normally distributed with slightly thicker tails.

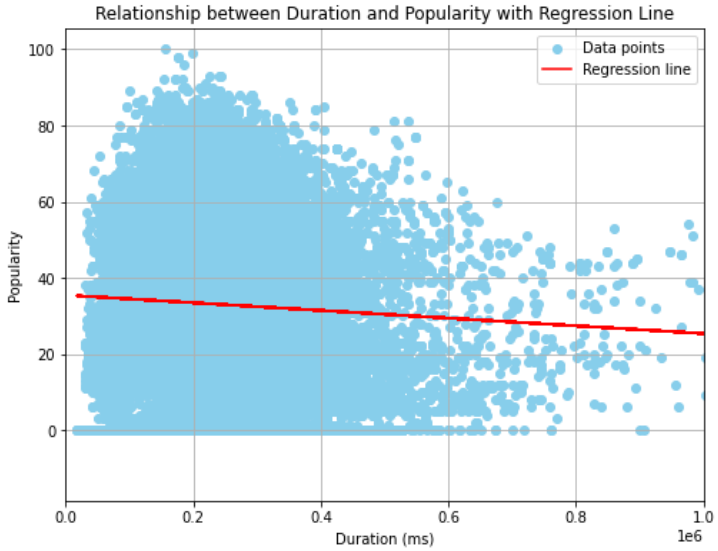




Question 2

Is there a relationship between song length and popularity of a song? If so, is the relationship positive or negative?

There is a very small negative correlation between duration of a song and popularity of a song. The pearson's correlation coefficient is -0.0547 showing that the longer the duration, the less popular the song is. The majority of songs are less than 1,000,000 ms (16.666 min) long, so the scatterplot below shows only songs less than ~17 minutes long. Pandas .corr was used to calculate the correlation.

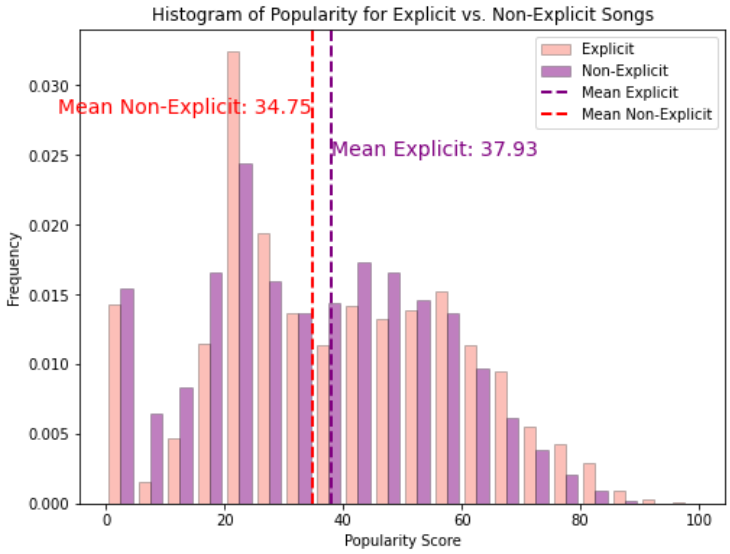


Question 3

Are explicitly rated songs more popular than songs that are not explicit?

To determine this, I used a t-test because popularity is reasonably normally distributed aside from the many zeroes. Since the t-test also assumes the same standard deviations, I ensured that the standard deviations are similar. The standard deviation of popularity for explicit songs in this sample is 20.78 and for non-explicit songs is 19.69. The null hypothesis is that there is no impact of explicitness on popularity. The test statistic was T = 10.45, which is considerably significant with a p-value of $1.58e^{-25}$. This means we can reject the null hypothesis and conclude that

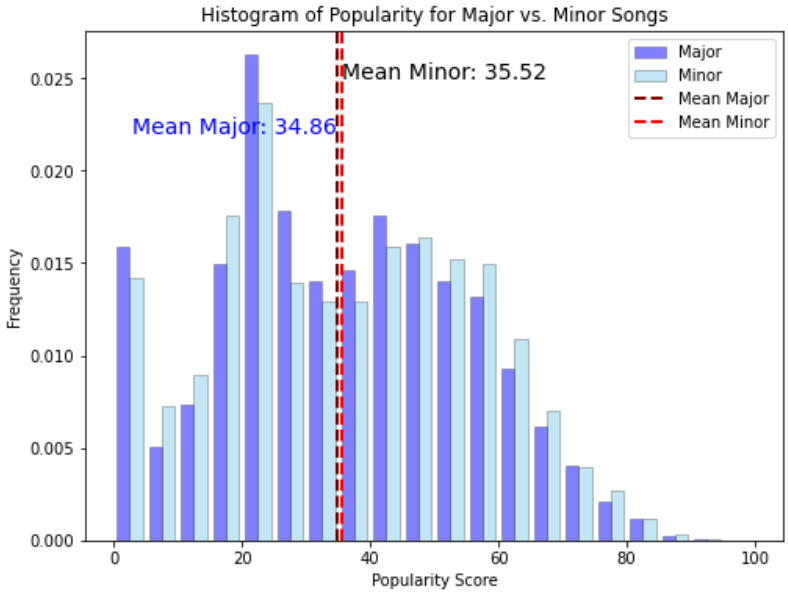
explicit songs have a tendency to have higher popularity. The below histogram shows both distributions and their means. This result aligns with society's preferences in today's times.



Question 4

Are songs in major key more popular than songs in minor key?

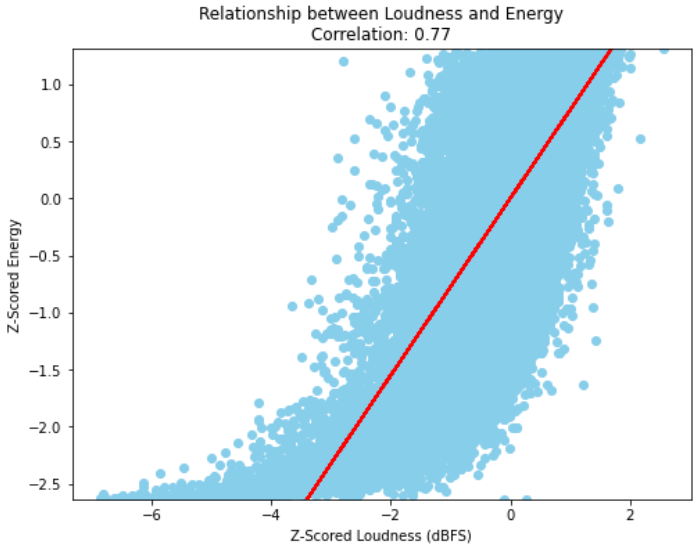
Similarly to above, I conducted a t-test to determine if songs in a major key are more popular than songs in a minor key. The standard deviations of popularity for major key songs and minor key songs were 19.59 and 20.24, respectively. The null hypothesis is that being in a major key does not influence the popularity of a song. The alternative hypothesis is that a song in a major key will influence its popularity. Interpreting the information using a two-tailed t-test, the critical T value is 1.64. The T test statistic was -3.38 which is less than -1.64 and corresponds to a 0.0007 p-value. This is statistically significant, and in the negative direction, meaning that we reject the null hypothesis, and conclude that a song in major key negatively influences the popularity of a song.



Question 5

Energy is believed to largely reflect the "loudness" of a song. Can you substantiate (or refute) that this is the case?

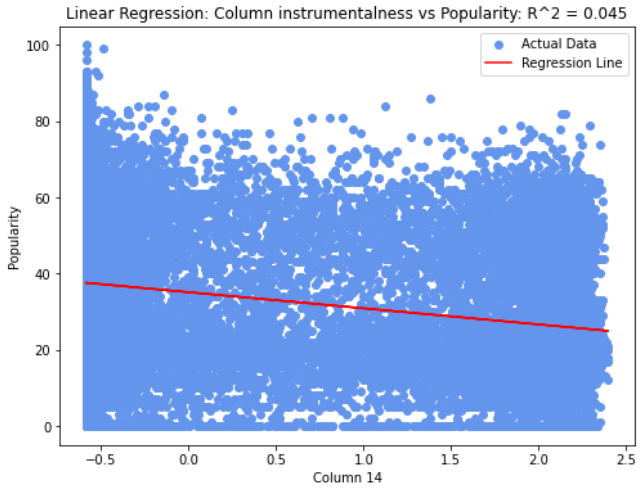
To investigate the relationship between loudness and energy in songs, I conducted a visual analysis using Matplotlib to create a scatter plot. I plotted loudness on the x-axis and energy on the y-axis for a set of songs. Upon visual inspection, the scatter plot revealed a clear trend suggesting a positive association between loudness and energy. To quantify this relationship, I calculated the correlation coefficient using Pandas' `.corr` method. The calculated correlation coefficient was 0.77, indicating a strong positive correlation between loudness and energy. This value is close to 1 which suggests that as loudness increases, energy tends to increase as well. This result makes sense because many people listen to loud songs when they need more energy (ie. falling asleep or before a high energy social gathering to 'hype themselves up').



Question 6

Which of the 10 song features in question 1 predicts popularity best? How good is this model?

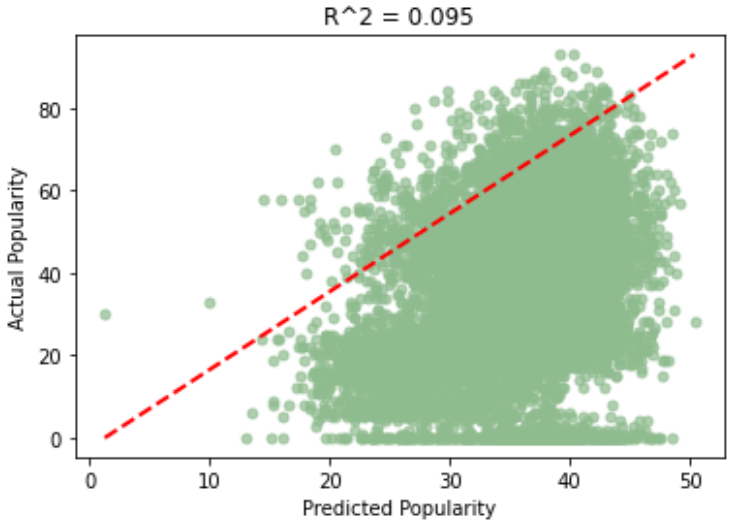
I put all 10 song features through a linear regression model and no one feature alone did well predicting popularity. The best predictor of the 10 was instrumentalness with the highest R^2 of 0.045 and the lowest RMSE of 19.38. All others scored under .009 for R^2 and above 19.70 for RMSE. This suggests that no one feature can account for variability of popularity.



Question 7

Building a model that uses *all* of the song features in question 1, how well can you predict popularity? How much (if at all) is this model improved compared to the model in question. How do you account for this?

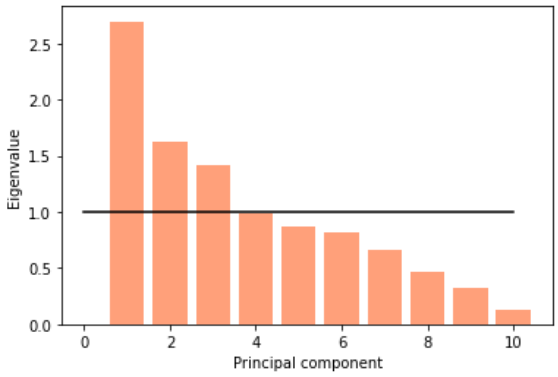
Performing a multiple regression with all 10 numerical song features in this data set, the score doubled from an R^2 of 0.045 to 0.095. This is not much better at all, as the model can only account for 9.5% of variance. This makes sense as there are most likely thousands of factors that go into any individual song's popularity, and much of it is luck of being noticed.



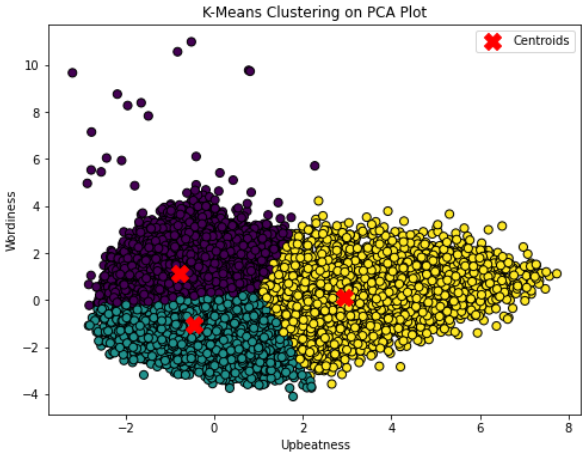
Question 8

When considering the 10 song features above, how many meaningful principal components can you extract? What proportion of the variance do these principal components account for? Using the principal components, how many clusters can you identify?

When using the Kaiser criterion (principal components above 1, shown as a black line on the graph below), 3 meaningful principal components can be extracted, but they only account for 57.495% of the variance. Because of this, I would choose enough principal components to account for 90% of the variance, so for this data there are 7 meaningful principal components that account for 90.87%. The first principal component is based on valence and danceability, I interpret this as upbeatness. The second is based on speechiness and liveness. The third is based on duration, so length of song. The fourth is based on tempo, so the speed of the song.



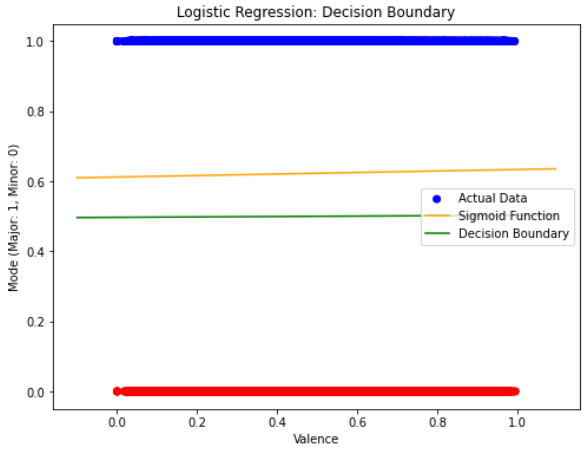
I created a scatterplot using the first 2 principal components and noticed there is a longer skinnier part of the data, and a wider portion. I chose 3 clusters for a K-means analysis. I don't believe there is clear enough separation to make meaning or significance of this analysis.

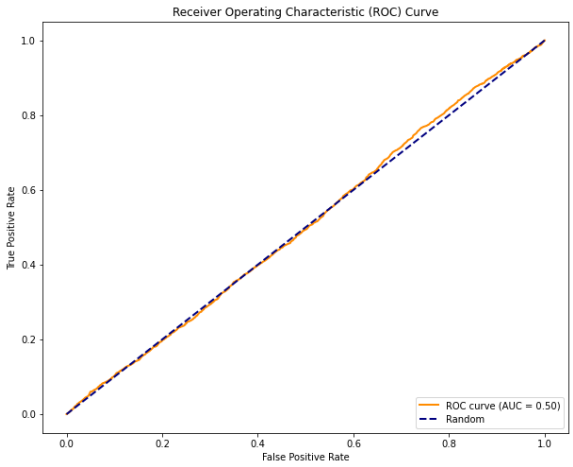


Question 9

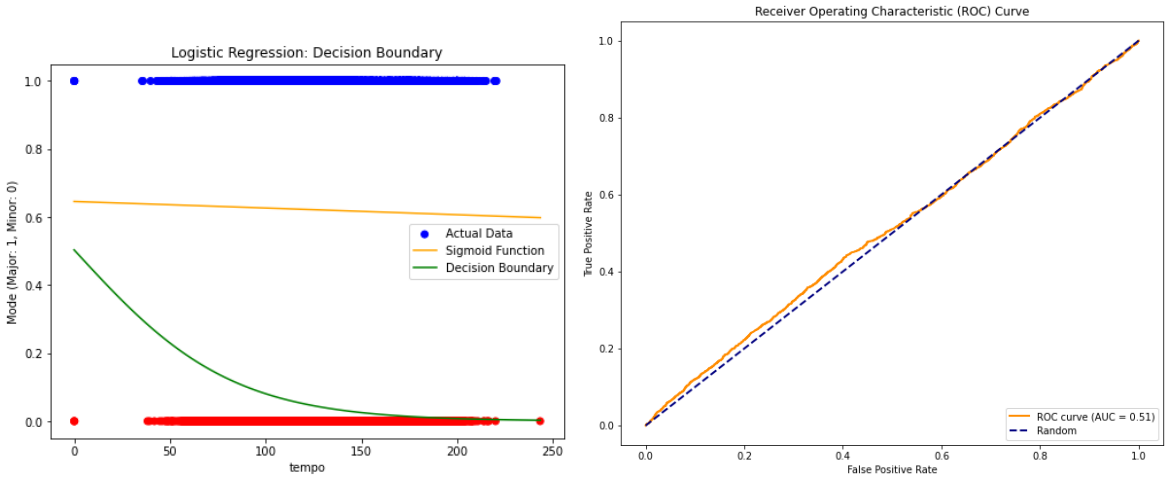
Can you predict whether a song is in major or minor key from valence? If so, how good is this prediction? If not, is there a better predictor? [Suggestion: It might be nice to show the logistic regression once you are done building the model]

I used scikitlearn's logistic regression package to show a logistic regression of valence to predict whether the song is in a major or a minor key. Valence does not seem to have any relationship with determining the mode of a song as it is distributed relatively easily across when mode is major and when mode is minor. The corresponding ROC curve has an AUC of 0.50 which means the model does no better than random chance.





Intuition told me that tempo of all the others might do better, because slow songs tend to be sad, which tend to be blues or ballads, but it did not do better by much. The tempo is concentrated in the middle which makes predicting for the model even more difficult. The AUC is 0.51, which is only slightly better than the random classifier.



10) Can you predict the genre, either from the 10 song features from question 1 directly or the principal components you extracted in question 8?

I used a classification tree with the 10 numeric song features to predict genre. First, I had to assign numerical values to each genre. The model has an accuracy of 0.23, which is low, but for such a complicated outcome, I am not so disappointed. The highest precision, recall, and f1-score, was for predicting the drum-and-bass genre with scores of 0.51, 0.52, and 0.51 respectively. This is most likely because drum-and-bass has more identifiable characteristics due to its instrumental and often heavier musical backtracking.

I ran the model again with a data set containing duplicates with different genres because the same numerical song features could apply to different genres in hopes that this model would do

better. But the accuracy was 0.20 for this model. However, it still predicted the drum-and-bass genre the best with scores of 0.50, 0.45, and 0.48.

Classification Report:

	precision	recall	f1-score	support
acoustic	0.13	0.14	0.14	197
afrobeat	0.17	0.16	0.17	204
alt-rock	0.06	0.07	0.06	123
alternative	0.02	0.01	0.01	68
ambient	0.32	0.35	0.34	161
anime	0.15	0.14	0.15	197
black-metal	0.35	0.33	0.34	190
bluegrass	0.23	0.23	0.23	193
blues	0.07	0.06	0.06	139
brazil	0.09	0.09	0.09	157
breakbeat	0.22	0.22	0.22	183
british	0.09	0.10	0.09	167
cantopop	0.23	0.24	0.23	180
chicago-house	0.34	0.35	0.35	184
children	0.31	0.36	0.33	163
chill	0.25	0.21	0.23	187
classical	0.46	0.44	0.45	150
club	0.12	0.12	0.12	191
comedy	0.77	0.78	0.78	199
country	0.11	0.10	0.11	125
dance	0.10	0.12	0.11	60
dancehall	0.18	0.19	0.19	162
death-metal	0.22	0.22	0.22	178
deep-house	0.12	0.13	0.12	159
detroit-techno	0.37	0.34	0.35	177
disco	0.15	0.16	0.15	156
disney	0.34	0.34	0.34	202
drum-and-bass	0.51	0.52	0.51	193
dub	0.07	0.07	0.07	95
dubstep	0.15	0.13	0.14	125
edm	0.10	0.10	0.10	104
electro	0.15	0.17	0.16	103
electronic	0.06	0.06	0.06	180
emo	0.13	0.13	0.13	182
folk	0.12	0.12	0.12	156
forro	0.48	0.46	0.47	213
french	0.12	0.12	0.12	188
funk	0.13	0.15	0.14	132
garage	0.10	0.10	0.10	189
german	0.13	0.15	0.14	137
gospel	0.21	0.23	0.22	160
goth	0.08	0.08	0.08	197
grindcore	0.61	0.62	0.62	187
groove	0.15	0.13	0.14	157
grunge	0.15	0.14	0.14	173
guitar	0.25	0.21	0.23	205
happy	0.35	0.37	0.36	169
hard-rock	0.14	0.14	0.14	170
hardcore	0.17	0.17	0.17	169
hardstyle	0.40	0.36	0.38	199
heavy-metal	0.19	0.19	0.19	204
hip-hop	0.13	0.13	0.13	149
accuracy			0.23	8588
macro avg	0.21	0.21	0.21	8588
weighted avg	0.23	0.23	0.23	8588