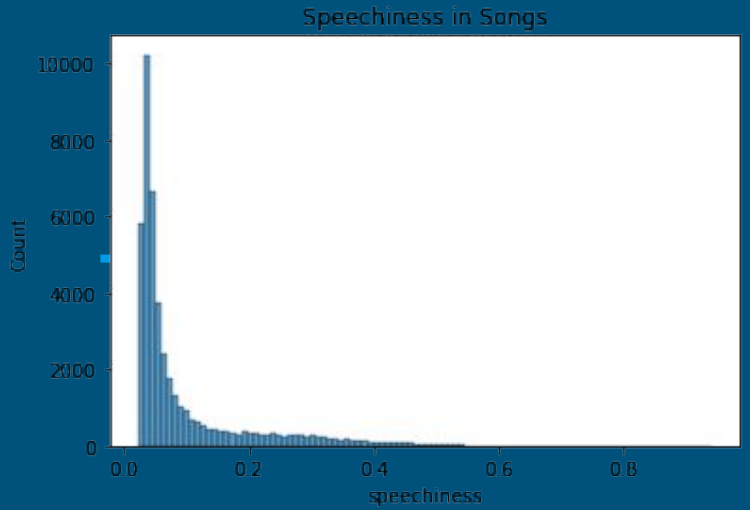
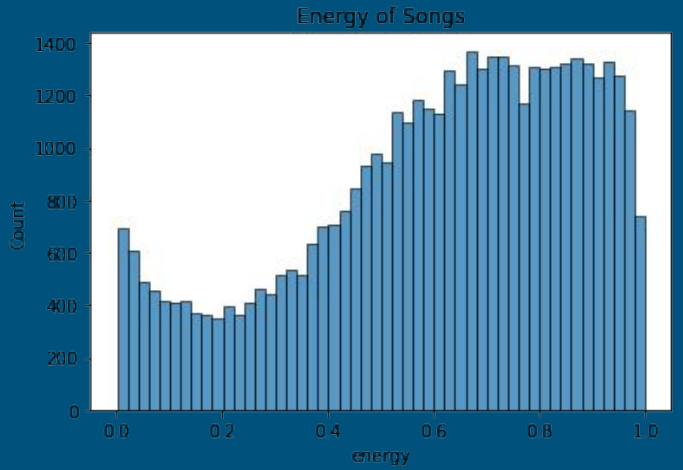
Predicting Musical Genres

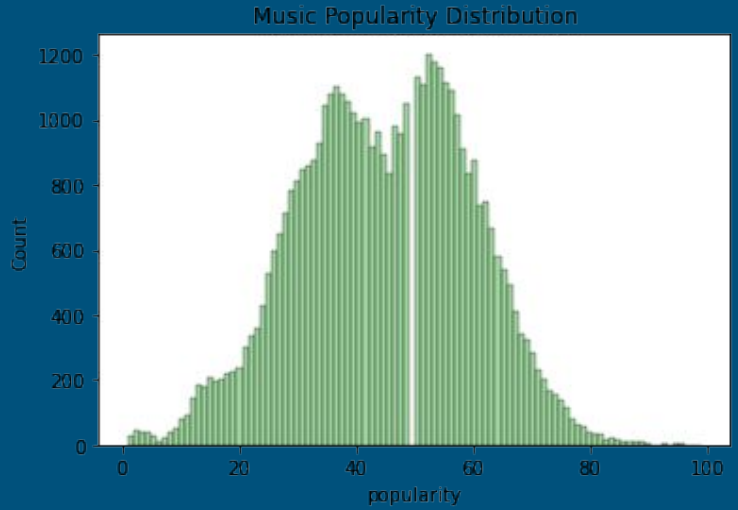
Codecademy Data Science Final Project

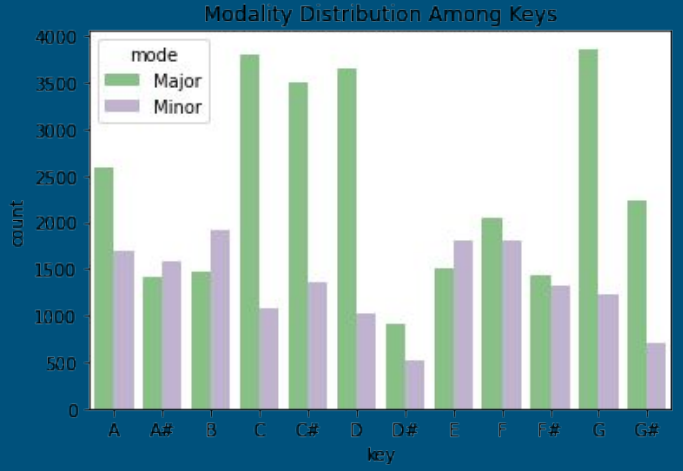
The hypothesis was that by analyzing enough disparate factors about a song, a machine learning algorithm could be trained to correctly predict the genre of that song.

The dataset provided multiple factors with which to categorize the data; among them were duration, tempo, overall popularity, and other less specific categories such as acousticness, danceability, and speechiness.

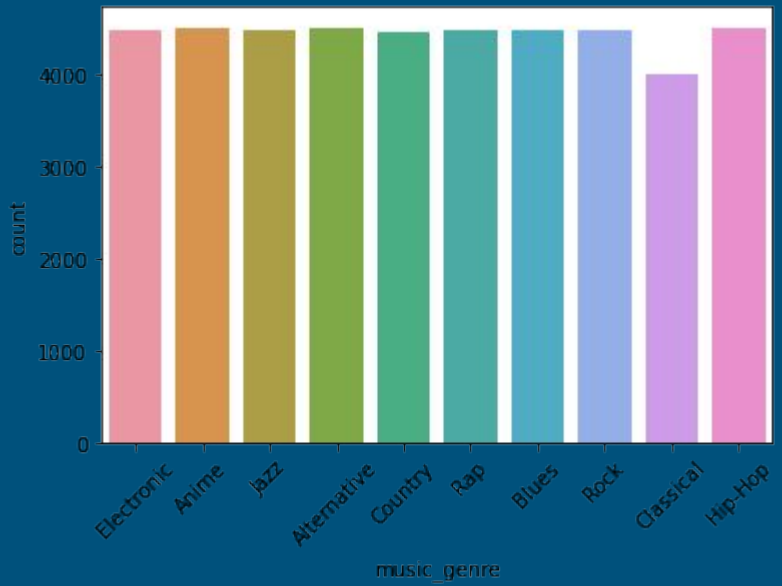

Speechiness in Songs
Count
speechiness
0
2000
4000
6000
8000
10000
0.0
0.2
0.4
0.6
0.8

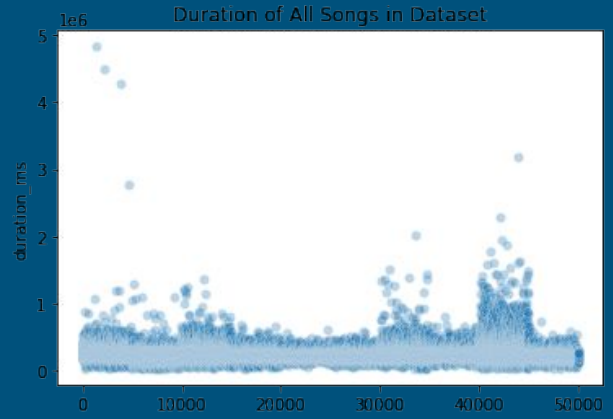

Energy of Songs
Count
energy
0
200
400
600
800
1000
1200
1400
0.0
0.2
0.4
0.6
0.8
1.0

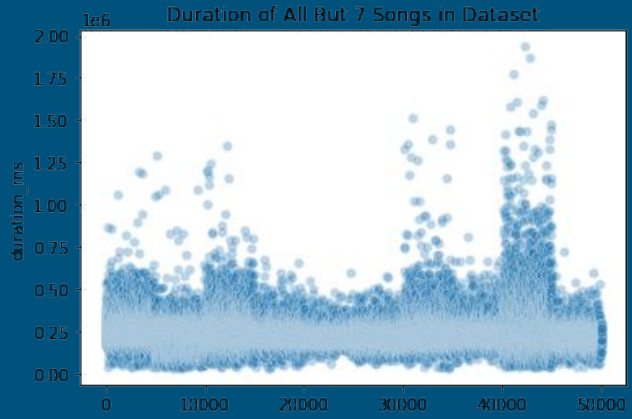

Music Popularity Distribution
Count
popularity
0
200
400
600
800
1000
1200
0
20
40
60
80
100


Modality Distribution Among Keys
mode
Major
Minor
count
key
0
500
1000
1500
2000
2500
3000
3500
4000
A
A#
B
C
C#
D
D#
E
F
F#
G
G#

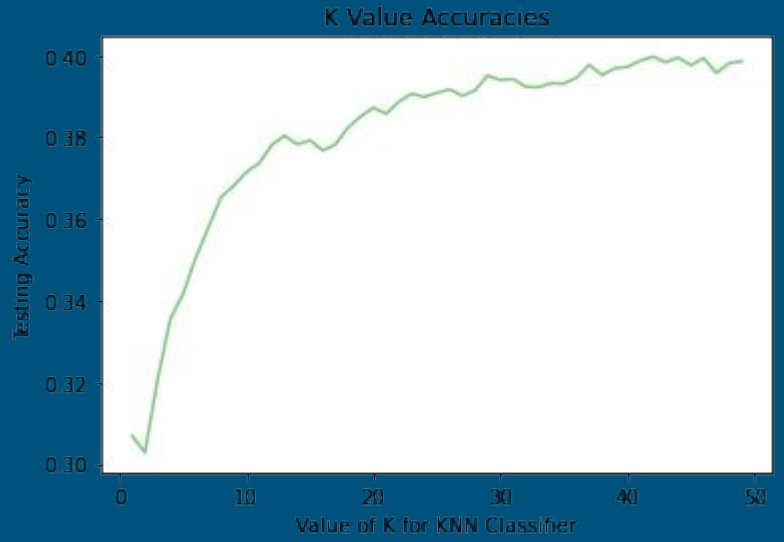
The songs were approximately evenly spread across the ten genres chosen by the dataset, but the durations had some notable outliers which had to be culled.







After cleaning the data, two classifiers were tried, a random forest and a K Nearest Neighbor. Both reached approximately a 39% accuracy rating; better than a blind guess, but not good enough to be confident in.



With a more comprehensive dataset with more features, it’s possible the prediction could grow more accurate. However, it is also possible that an art form as diverse as music covers too much ground to accurately train a machine learning algorithm to analyze this broadly.