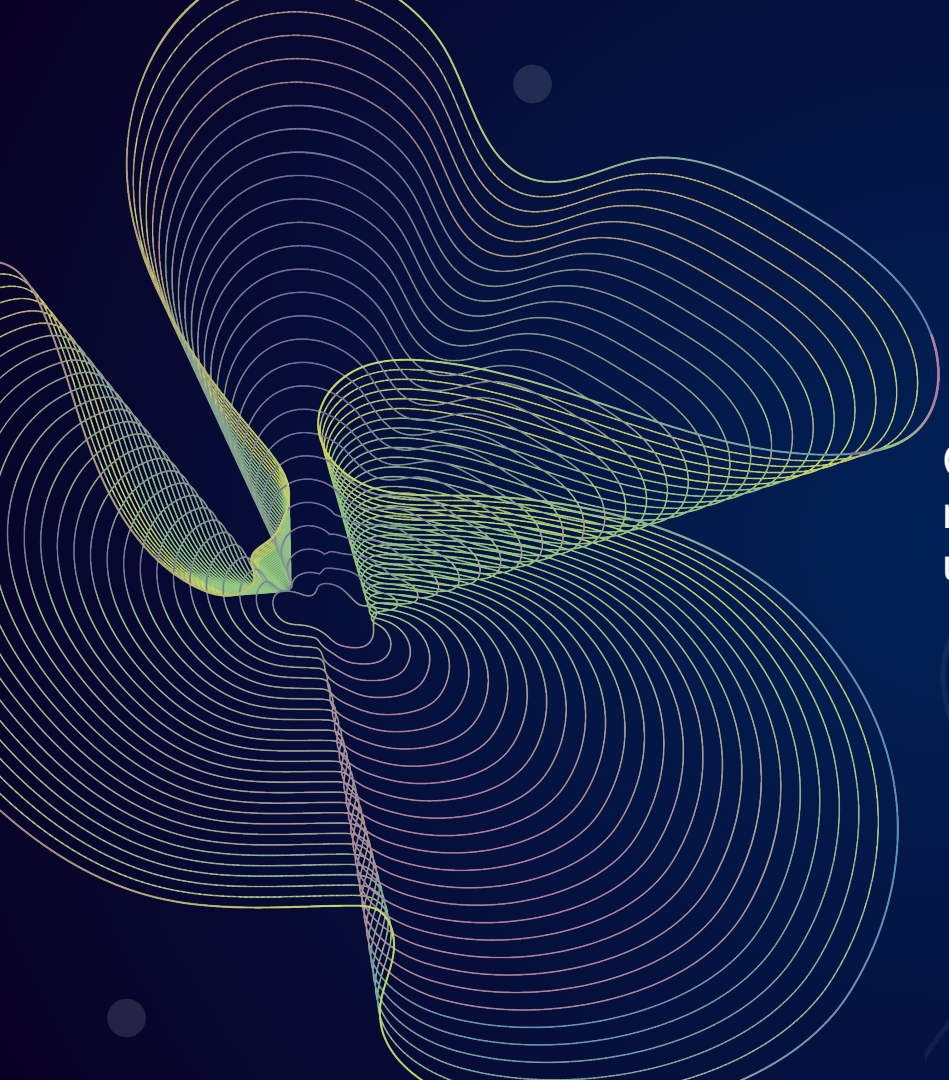
SPOTIFY DATASET MACHINE LEARNING

Isaac Vernon, Zain Khan, Anchit Sood, Hari Srikanth

W207



PROBLEM STATEMENT

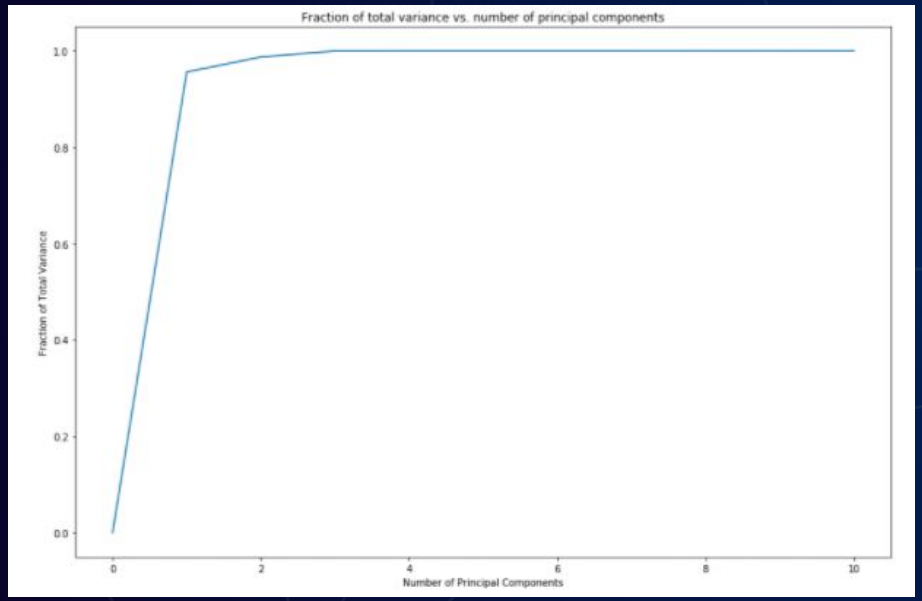
Can we extract genre from the various musical features at the song level, using an unsupervised learning technique?

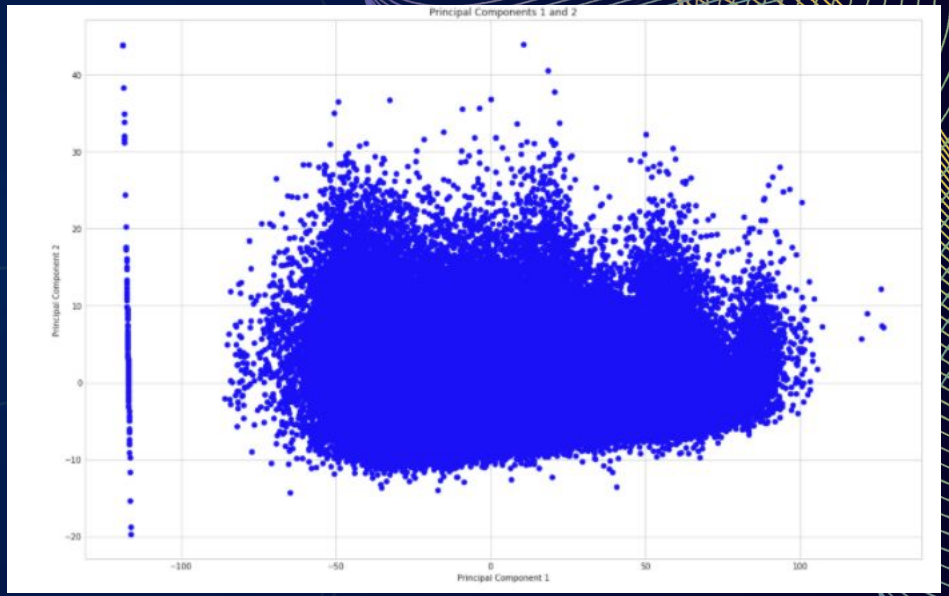
- We will attempt to answer this question via clustering

REINTRODUCTION TO THE DATA

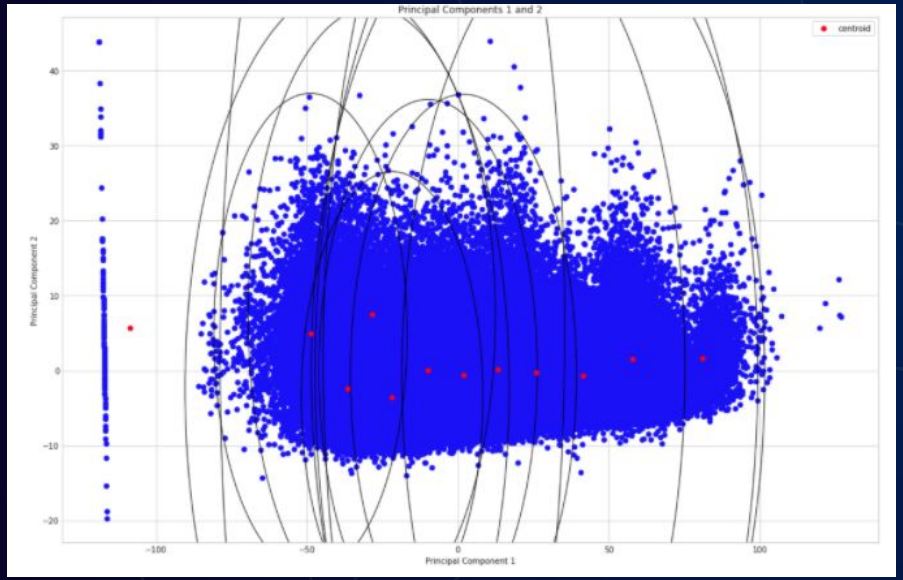
- The dataset came with 5 different tables
  - the main two of interest for clustering are the song level dataset, and the genre dataset
- Both datasets focus on the following features: **acousticness, danceability, duration_ms, energy, , instrumentalness, key, liveness, loudness, mode, popularity, speechiness, tempo, valence**
  - Spotify defines terms like “acousticness” or “danceability” using internal techniques, more info can be found at the following link
- The song level dataset does not include genre, but the genre dataset is an aggregation of the song features, presumably across all songs
- Our problem statement is framed around the lack of genre at the song level
  - We will consult the aggregated genre data in our clustering to provide some sense of how our clusters compare to genre data

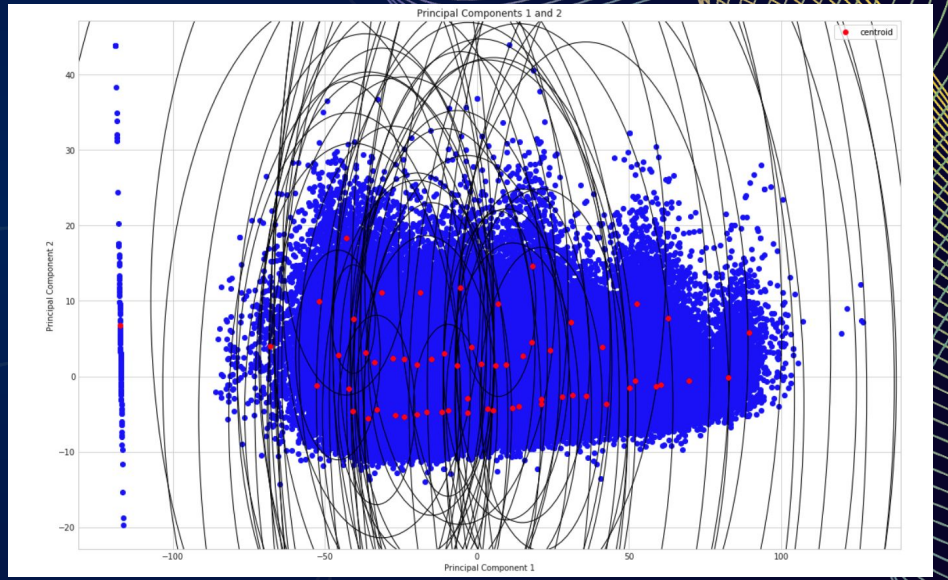
DIMENSIONALITY REDUCTION WITH PCA



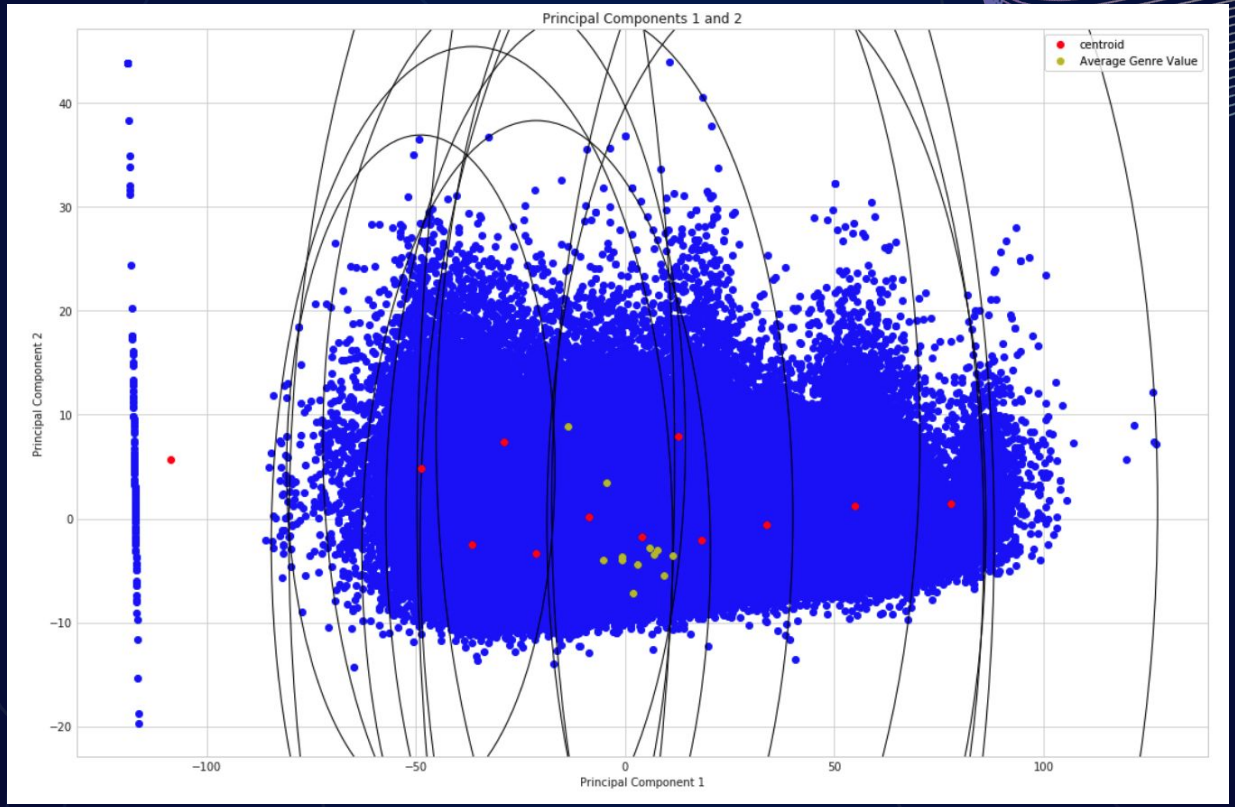


APPLYING K-MEANS CLUSTERING TO THE DATA

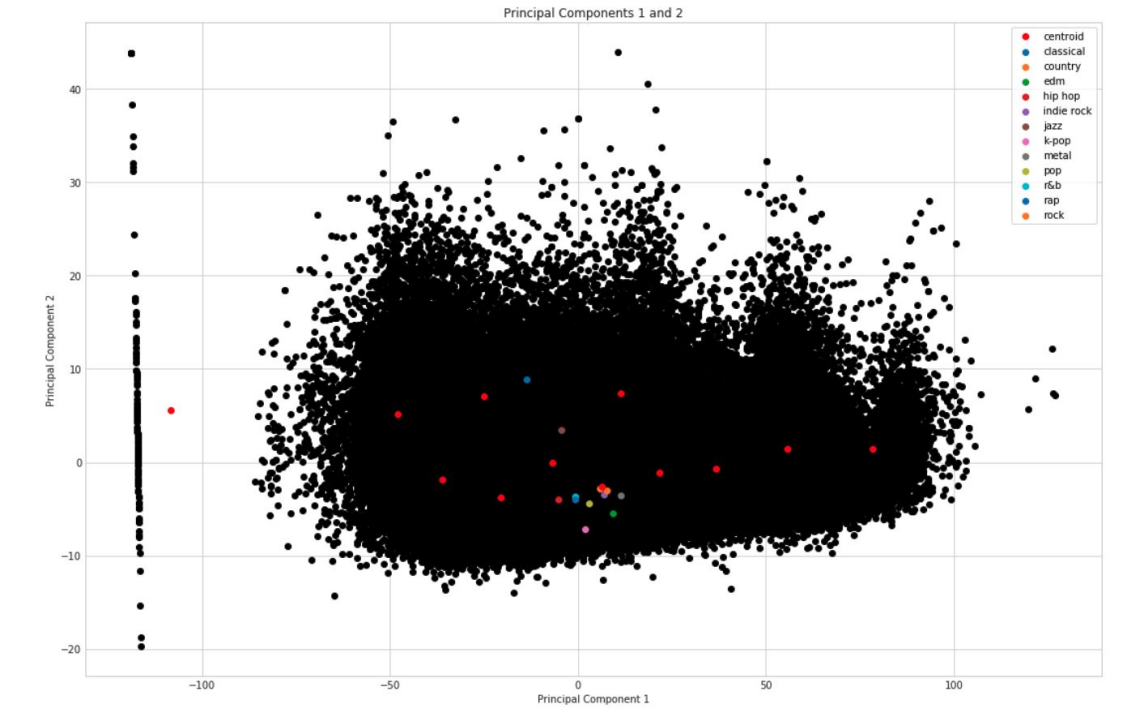




USING GENRE DATA ALONG WITH CLUSTERING



COMPARING CLUSTERS TO GENRE DATA



Cluster 1 is closest to the genre: metal
Cluster 2 is closest to the genre: classical
Cluster 3 is closest to the genre: metal
Cluster 4 is closest to the genre: classical
Cluster 5 is closest to the genre: classical
Cluster 6 is closest to the genre: jazz
Cluster 7 is closest to the genre: metal
Cluster 8 is closest to the genre: classical
Cluster 9 is closest to the genre: metal
Cluster 10 is closest to the genre: metal
Cluster 11 is closest to the genre: country
Cluster 12 is closest to the genre: classical

“SUPERVISED” CLUSTERING

PURE GENRES

Songs Guaranteed to be from the Genre

POSSIBLE GENRES

Songs have a Good Chance to be from the Genre (Artist has Produced songs in that genre before)

SUBSET AND FULL

Clustering of all 12 Genres as well as just a few

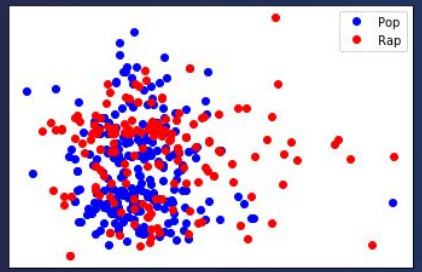
LOOKING AT TWO DATASETS

- Possible is an aggregation of all of the songs from artists that have contributed to the genre before
- Pure is an aggregation of all of the songs from artists who have only contributed to that genre before

Possible		Pure	
Rock	55,909	Pop	248
Pop	46,524	Rap	167
Jazz	18,235	Hip-Hop	80
Country	17,653	Country	70
Rap	15,387	K-Pop	25
Classical	13,441	R & B	20
Hip-Hop	12,053	Jazz	13
Metal	9,275	EDM	11
R & B	5,079	Classical	11
Indie Rock	2,244	Indie Rock	5
EDM	1,327	Metal	3
K-Pop	512	Rock	0

POP VS RAP PURE GENRES

2 WAY PCA



TEST ON PURE

5 Component PCA

1 GMM Component

Full Covariance

77.1% Accurate on Test Set

TEST ON POSSIBLE

5 Component PCA

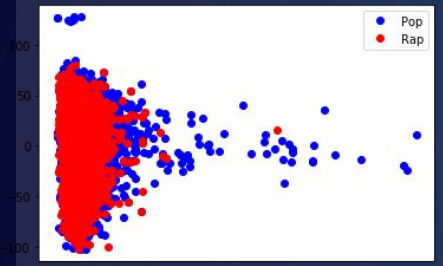
1 GMM Component

Full Covariance

39.8% Accurate on Test Set

POP VS RAP POSSIBLE GENRES

2 WAY PCA



TEST ON PURE

7 Component PCA

6 GMM Components

Tied Covariance

37.1% Accurate on Test Set

TEST ON POSSIBLE

7 Component PCA

6 GMM Components

Tied Covariance

77.5% Accurate on Test Set

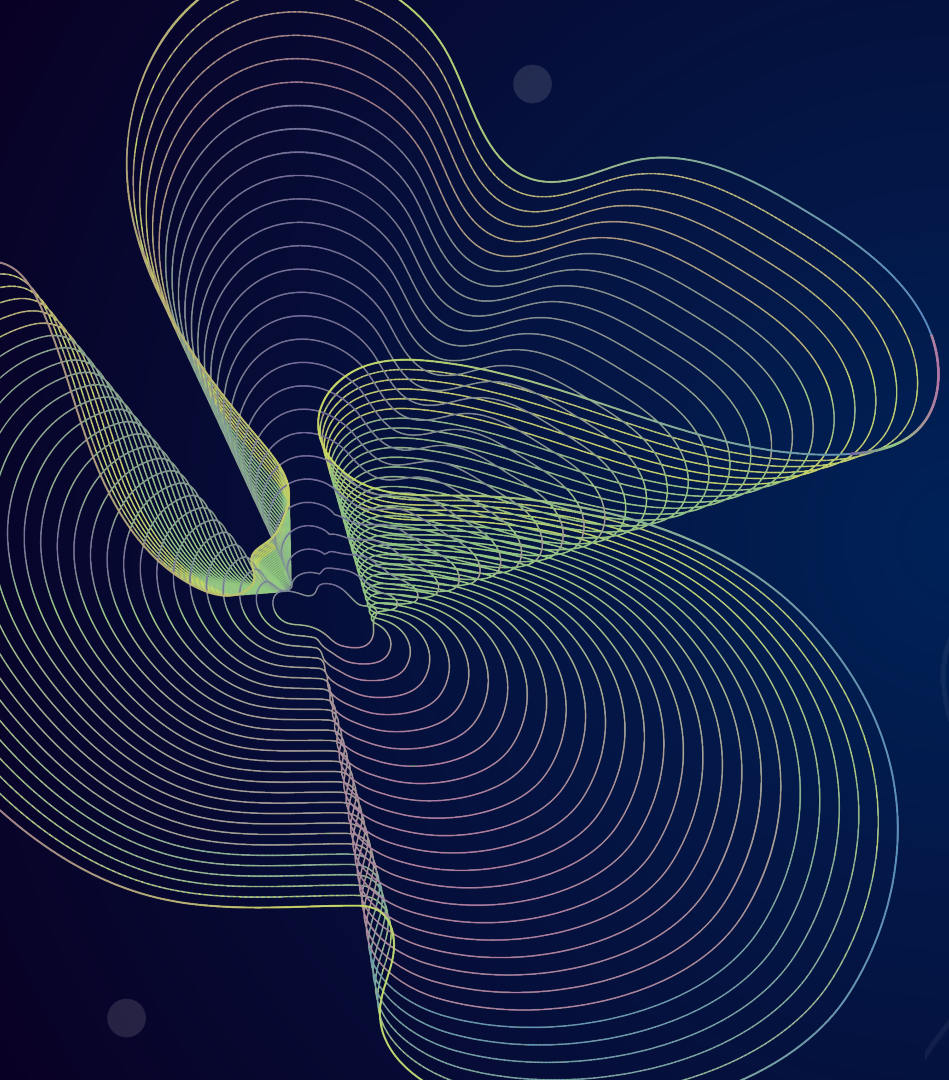
FULL GENRE CLUSTERING

12

Number of Genres used in Clustering

30.1%

The Accuracy on our test set for the fitted clusters



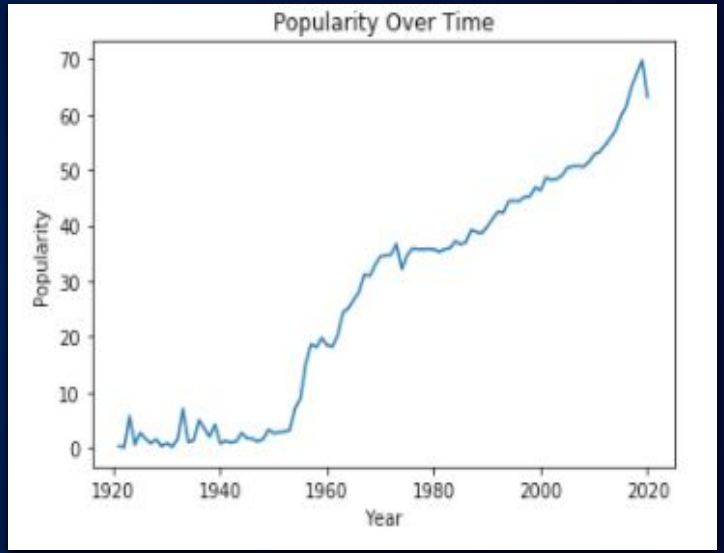
SUPERVISED LEARNING

- Can we predict popularity with the various musical features through different supervised learning techniques?

EXPLORATORY DATA ANALYSIS (PREDICTION)

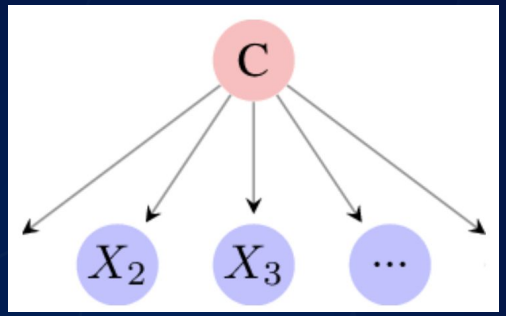
- After applying clustering, we wanted to see how we can utilize the various musical features in order to predict popularity.
  - Regression
  - Decision Trees
  - Naive Bayes

- Popularity appears to be better defined as “current popularity”



NAIVE BAYES

- Assumption of Independence among predictors



- **Our Intention**: Use Naive Bayes as a baseline for predicting popularity

	acousticness	danceability	energy	year	loudness	instrumentalness	liveness	speechiness	tempo	valence	key
acousticness	1.000000	-0.265950	-0.750283	-0.624550	-0.567072	0.335821	-0.023871	-0.056077	-0.204982	-0.185540	-0.021686
danceability	-0.265950	1.000000	0.220569	0.203430	0.294170	-0.281429	-0.105532	0.225305	-0.004872	0.560242	0.022599
energy	-0.750283	0.220569	1.000000	0.532419	0.782982	-0.287692	0.126293	-0.045226	0.249936	0.350086	0.029984
year	-0.624550	0.203430	0.532419	1.000000	0.490118	-0.291571	-0.055839	-0.120937	0.137892	-0.029304	0.012503
loudness	-0.567072	0.294170	0.782982	0.490118	1.000000	-0.417033	0.052985	-0.105796	0.211114	0.308418	0.021920
instrumentalness	0.335821	-0.281429	-0.287692	-0.291571	-0.417033	1.000000	-0.047397	-0.115735	-0.107570	-0.193929	-0.014268
liveness	-0.023871	-0.105532	0.126293	-0.055839	0.052985	-0.047397	1.000000	0.147667	0.008124	-0.000426	-0.000106
speechiness	-0.056077	0.225305	-0.045226	-0.120937	-0.105796	-0.115735	0.147667	1.000000	-0.010070	0.056383	0.015225
tempo	-0.204982	-0.004872	0.249936	0.137892	0.211114	-0.107570	0.008124	-0.010070	1.000000	0.171182	0.003148
valence	-0.185540	0.560242	0.350086	-0.029304	0.308418	-0.193929	-0.000426	0.056383	0.171182	1.000000	0.029064
key	-0.021686	0.022599	0.029984	0.012503	0.021920	-0.014268	-0.000106	0.015225	0.003148	0.029064	1.000000
mode	0.046475	-0.045306	-0.038355	-0.033084	-0.013147	-0.035051	0.005393	-0.057493	0.014539	0.014727	-0.112766
explicit	-0.253690	0.241891	0.142677	0.245227	0.152695	-0.138292	0.039272	0.413074	0.011484	-0.022327	0.008578
duration_ms	-0.079311	-0.134500	0.036396	0.076293	-0.014687	0.084814	0.034270	-0.058449	-0.028816	-0.198760	-0.003116
popularity	-0.593345	0.221077	0.497488	0.880724	0.466546	-0.299829	-0.075293	-0.135707	0.135047	0.009327	0.010675

CORRELATION OF POPULARITY VS. MUSICAL FEATURES

- Significant Correlation Coefficients
  - Year
  - Acousticness
  - Energy
  - Loudness
  - Instrumentalness
  - Danceability
  - Explicit

```
acousticness      -0.593345
danceability       0.221077
energy             0.497488
year               0.880724
loudness           0.466546
instrumentalness  -0.299829
liveness          -0.075293
speechiness       -0.135707
tempo              0.135047
valence            0.009327
key                0.010675
mode              -0.032854
explicit           0.214044
duration_ms        0.063292
popularity         1.000000
Name: popularity, dtype: float64
```

BEST NAIVE BAYES MODELS

Gaussian

- Var_smoothing = 0.01
- Accuracy = 19.0%
- RMSE = 11.96

Multinomial

- Alpha: No change between alpha variation
- Accuracy = 16.3%
- RMSE = 38.24

LINEAR REGRESSION

CLASSIC

.782

RIDGE

.778

Alpha:
1

LASSO

.782

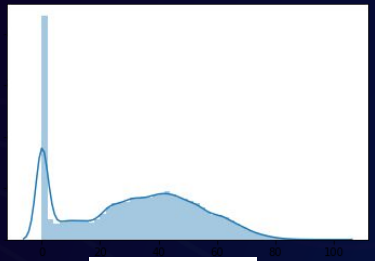
Alpha:
1e-6

ELASTIC NET

.782

Alpha:
1e-5

LINEAR REGRESSION GRAPHS

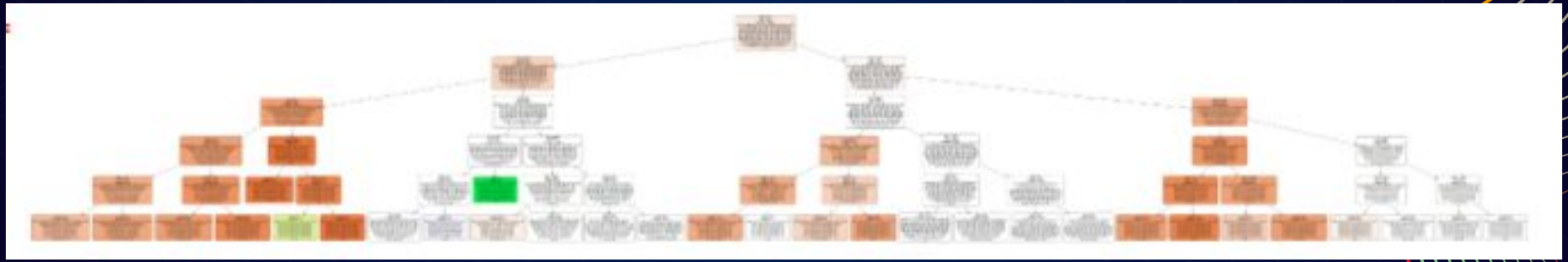


Popularity

LOGISTIC REGRESSION

BINARIZE	INTERPRET	ACCURACY
Popularity > 50	Predict if Popularity is in the upper half of the distribution	81.4% (Just Predicting 0 gives 78.9%)
Popularity > 31.56	Predict if Popularity will be above or below the mean	77.8% (Just Predicting 1 gives 53.2%)
Popularity > 33	Predict if Popularity will be above or below the median	76.8% (Just Predicting 0/1 gives 50%)

RANDOM FOREST



CONCLUSION

- Clustering appeared to be difficult to do well with the features we had. There didn't seem to be much distinguishing the different genres (could be due to the presence of middle ground genres -- acid rock, acid pop, acid rap, etc)

- Popularity prediction wasn't ideal when we were trying to predict the exact value (100 labels). When we binarized it and predicted a value off of that we had strong performance, as well as with Linear Regression