Spotify Music trends based on extracted features

—

Miles Devaney, David Richardson, Emma Van Der Werf

Introduction

Some initial questions we had:

- Are there noticeable trends in song features over time?
- Are there distinct clusters that appear to match up with existing genres?
- Can music features be used to predict how popular a song is?

Initial Data Source

Over 1.2 million songs, with 24 audio features scraped from the Spotify API

```
music.dtypes
```

id	object
name	object
album	object
album_id	object
artists	object
artist_ids	object
track_number	int64
disc_number	int64
explicit	bool
danceability	float64
energy	float64
key	int64
loudness	float64
mode	int64
speechiness	float64
acousticness	float64
instrumentalness	float64
liveness	float64
valence	float64
tempo	float64
duration_ms	int64
time_signature	float64
year	int64
release_date	object
dtype: object	

explicit	danceability	energy	key	loudness	mode	speechiness	acousticness	instrumentalness	liveness	valence	tempo	duration_ms	time_signature
False	0.470	0.978	7	-5.399	1	0.0727	0.02610	0.000011	0.3560	0.503	117.906	210133	4.0
True	0.599	0.957	11	-5.764	1	0.1880	0.01290	0.000071	0.1550	0.489	103.680	206200	4.0
False	0.315	0.970	7	-5.424	1	0.4830	0.02340	0.000002	0.1220	0.370	149.749	298893	4.0
True	0.440	0.967	11	-5.830	0	0.2370	0.16300	0.000004	0.1210	0.574	96.752	213640	4.0
False	0.426	0.929	2	-6.729	1	0.0701	0.00162	0.105000	0.0789	0.539	127.059	205600	4.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
False	0.264	0.966	5	-6.970	0	0.0672	0.00935	0.002240	0.3370	0.415	159.586	276213	4.0
False	0.796	0.701	11	-6.602	0	0.0883	0.10400	0.644000	0.0749	0.781	121.980	363179	4.0
False	0.785	0.796	9	-5.960	0	0.0564	0.03040	0.918000	0.0664	0.467	121.996	385335	4.0
False	0.665	0.856	6	-6.788	0	0.0409	0.00007	0.776000	0.1170	0.227	124.986	324455	4.0
False	0.736	0.708	2	-9.279	0	0.0539	0.01680	0.296000	0.2790	0.204	117.991	304982	4.0

Features

- Acousticness: A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic
- Danceability: Describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable
- Energy: Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity.
- Instrumentalness: Predicts whether a track contains no vocals. The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content.

Features(cont.)

- Liveness: Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live.
- Speechiness: Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words.
- Valence: Valence measures the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).

Potential Problems with the Data

- Spotify listeners may not be representative of the population of all music listeners
  - 58% of Spotify users come from Europe or North America, under-representing other regions, especially Africa and Asia, which together only represent a combined 20% of users
  - 55% of Spotify listeners are under the age of 35
- The data is a little out of date
- The songs chosen for this dataset are likely not representative of all songs on Spotify
  - The songs in the dataset were all songs uploaded to MusicBrainz catalog
  - Are likely to represent the more popular song and over represent English language music
- Spotify is vague about how they determine some song features, so it's unclear how accurate they are
  - Danceability, energy, acousticness, liveliness, valence
- Spotify has been accused of accepting payment in exchange for featuring songs on popular playlists and has problems with fraudulent streams, potentially making some songs seem more popular than they really are

Ethical Issues

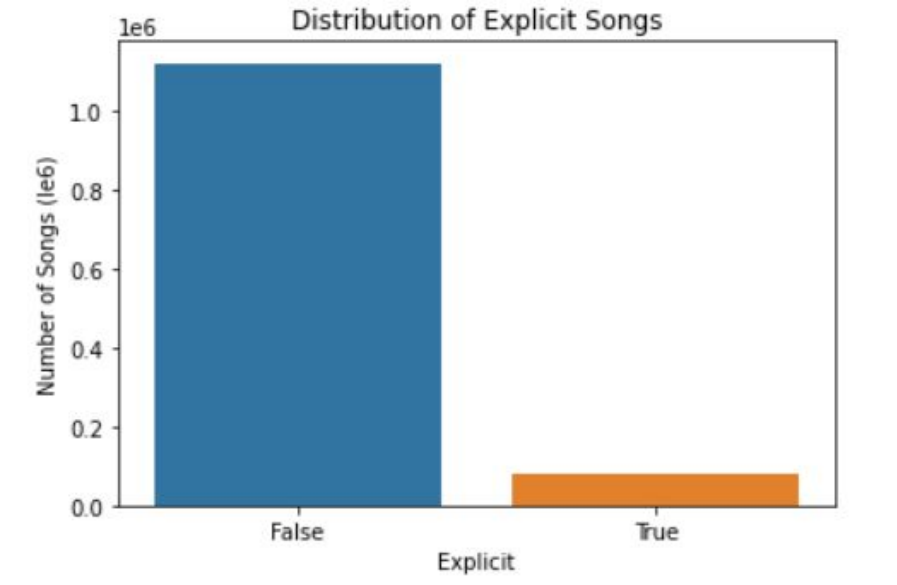
- It’s possible that, in order to become as popular as possible, musicians will change their music to reflect the trend of the most popular indicators, making music less authentic
- Many contend that in allowing access to its users’ data, Spotify is violating the privacy of its users
- There are concerns that access to Spotify users’ data could enable advertisers to track users’ emotions and target ads to them accordingly

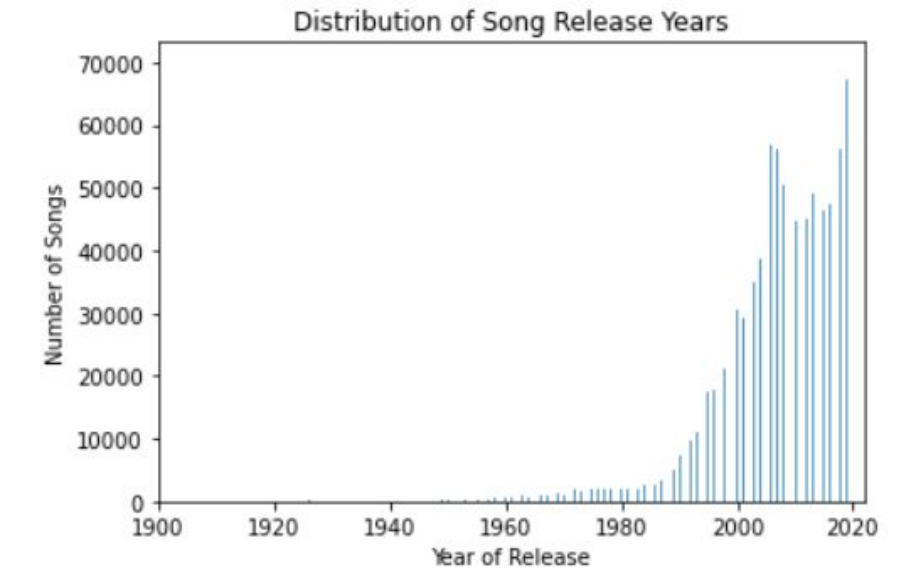
Data Cleaning

- 10 songs were listed incorrectly with a release year of 0, the actual release years were found online and inserted instead
- 2777 song were listed incorrectly with a tempo of 0, which were changed to Null values in order to not skew numeric analysis involving the variable

EDA

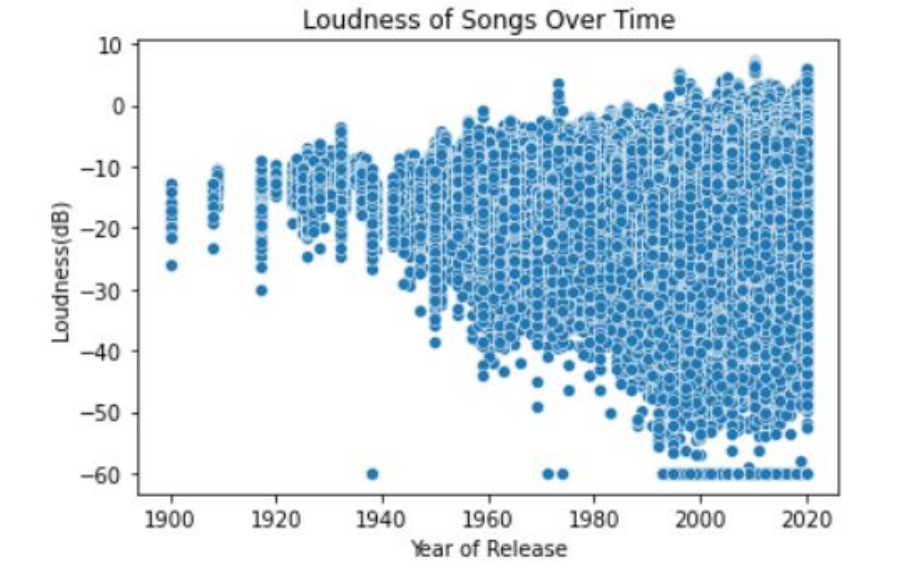
- Far more non-explicit than explicit songs
- For this set of data, song releases peaked in the mid-2000s and the late 2010s

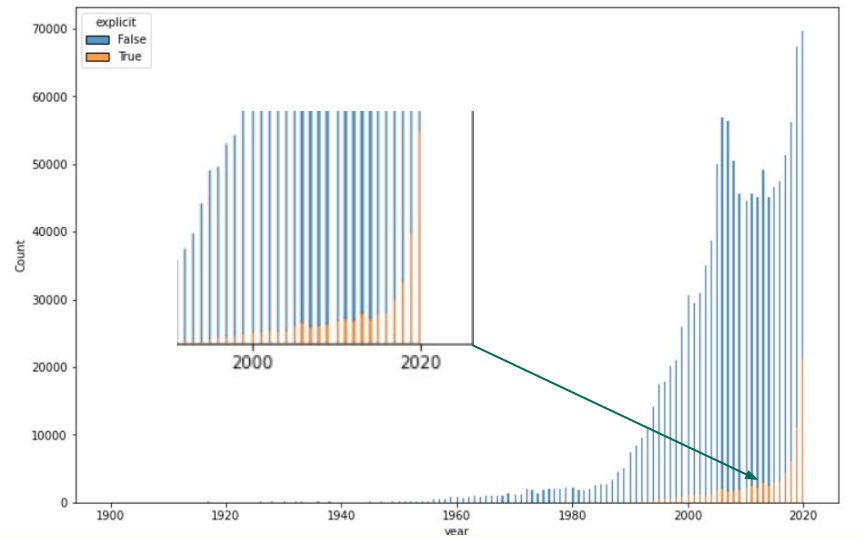




Continued EDA

- Rather than a strict increase in loudness over time, the overall range widens
- The number of explicit songs spikes dramatically at the end of the 2010s

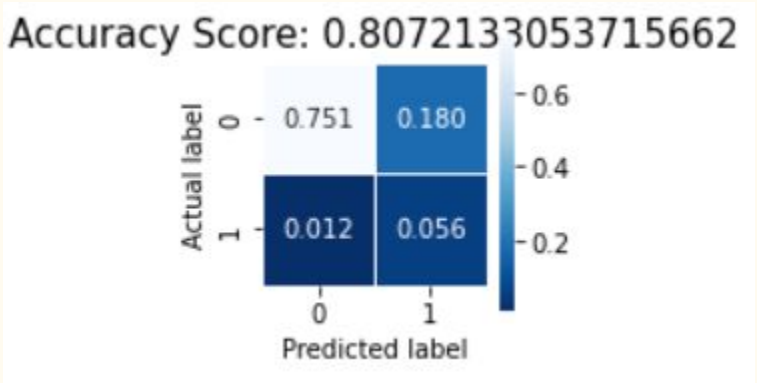




Classification: Explicit

- Used audio features to classify whether or not a song was explicit
- Many more non-explicit songs, so class weights had to be balanced

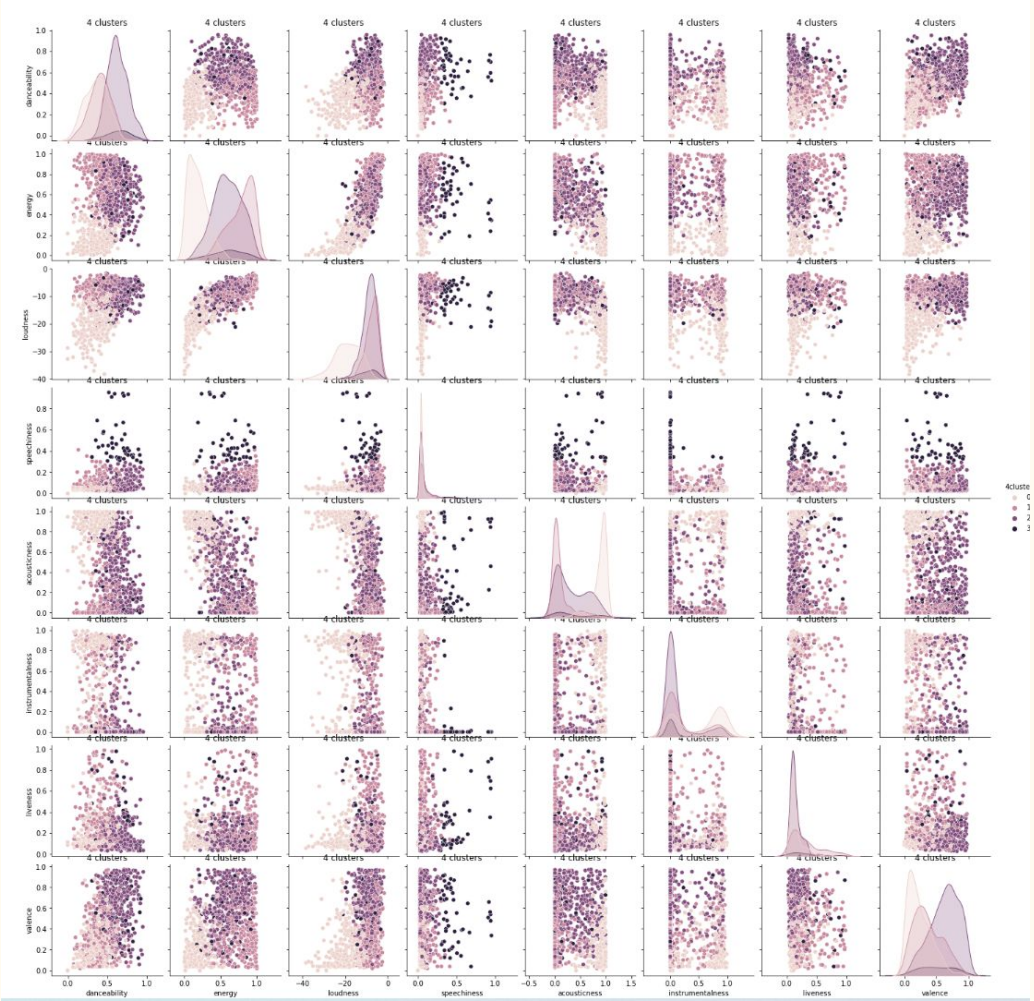
```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, stratify=y, random_state = 21)
lg = LogisticRegression(class_weight = 'balanced', solver = 'liblinear')
lg.fit(X_train, y_train)
score = lg.score(X_test, y_test)
```



Clustering

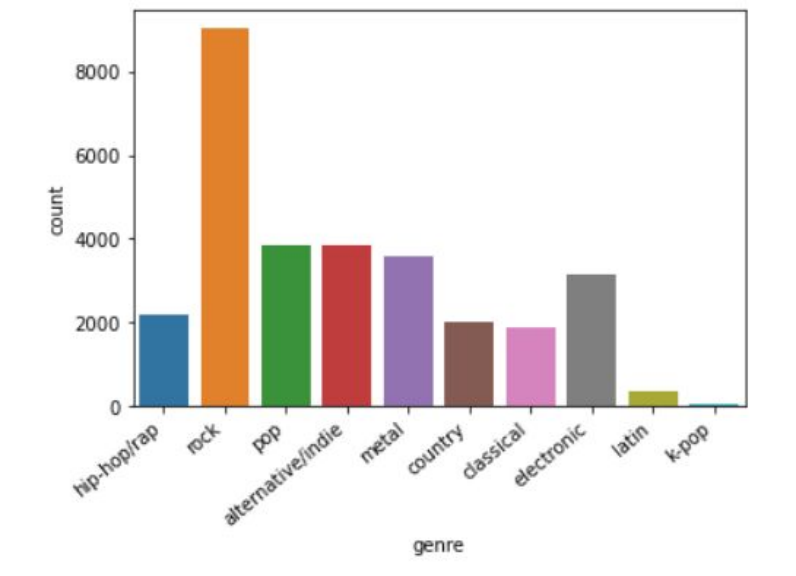
The idea behind clustering was to see if clustering based on audio features would produce clusters similar to music genres.

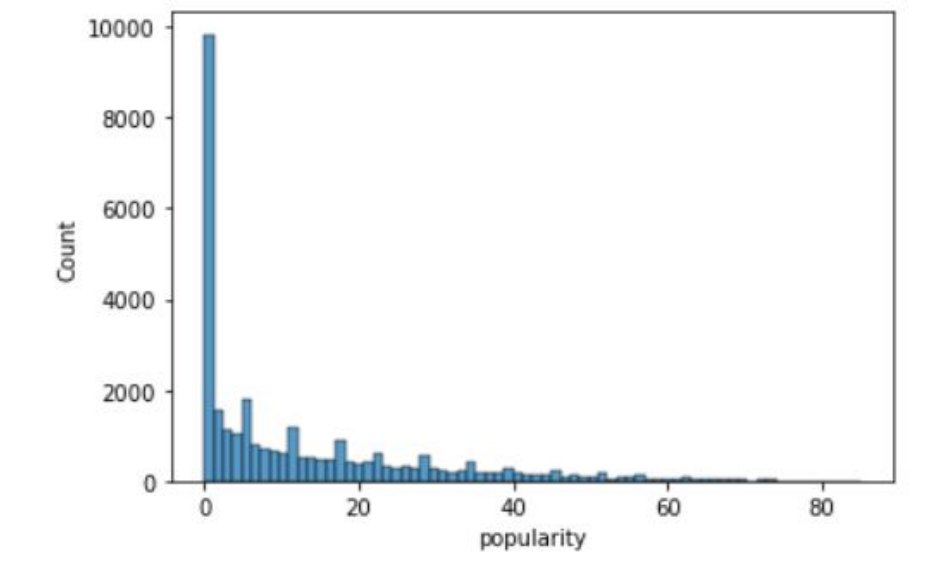
They don't.



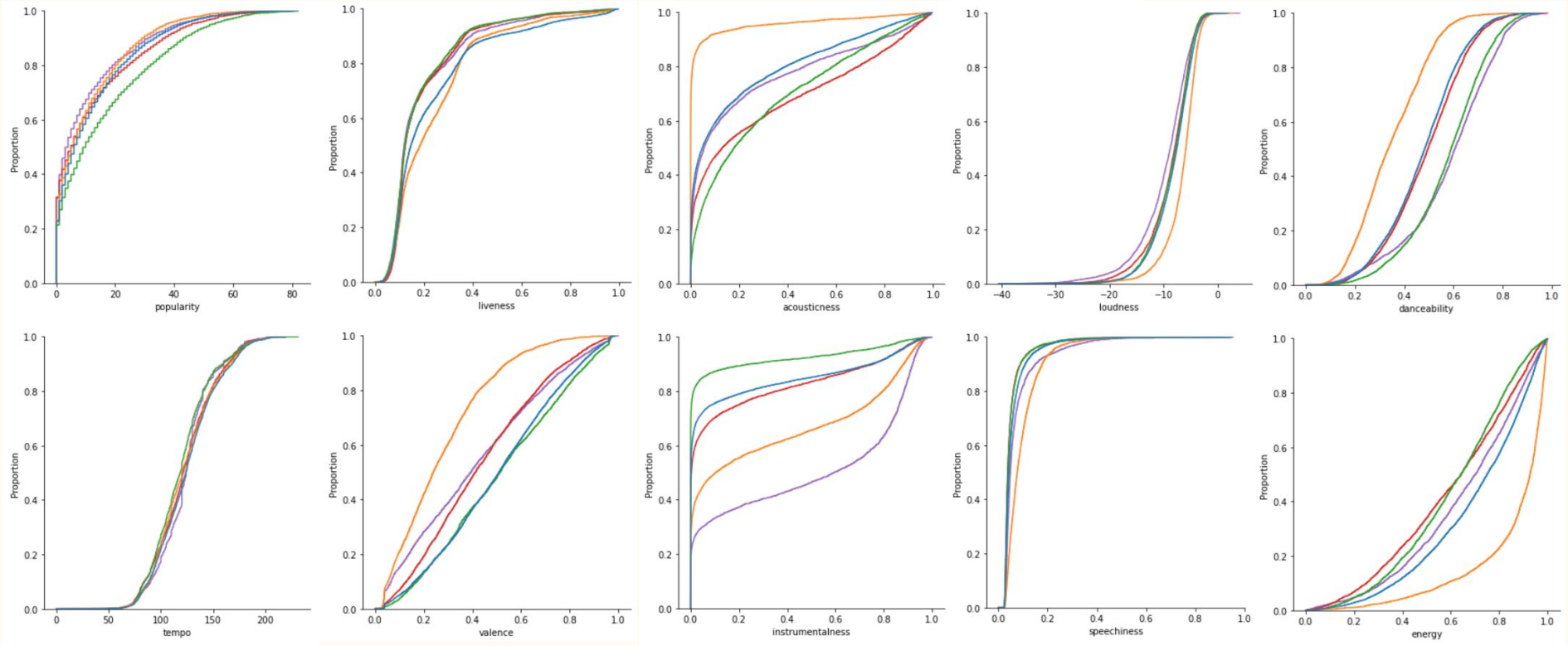
Retrieving Genre and Popularity

- Genres were generated based on tags from last.fm
- Popularity is a value assigned by Spotify





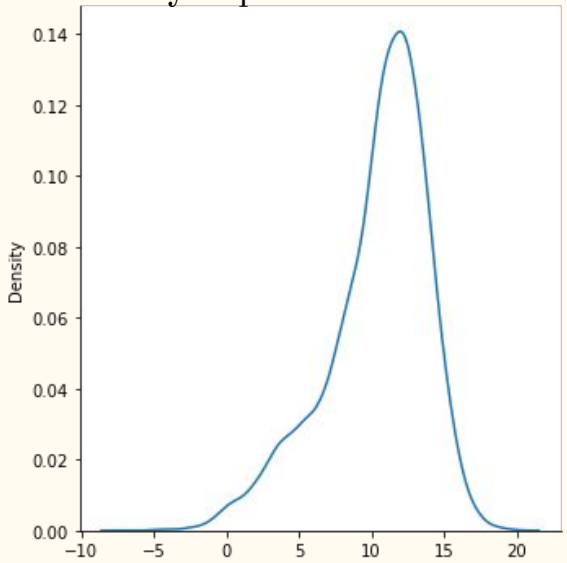
Distribution of Features by Genre



Using Regression to Predict Popularity

- Initially used the 8 music features and tempo
- Best models were Linear Regression, Decision Tree, and KNeighbors
- All had low scores (~.08)

Density of predictions for LR



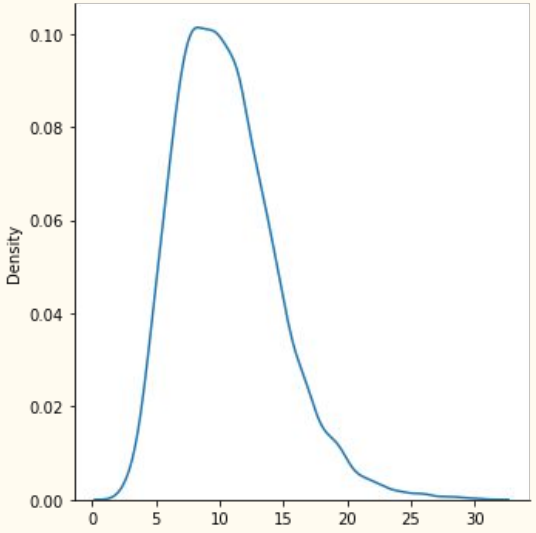
LR Positive Coefficients:

- Danceability
- Loudness

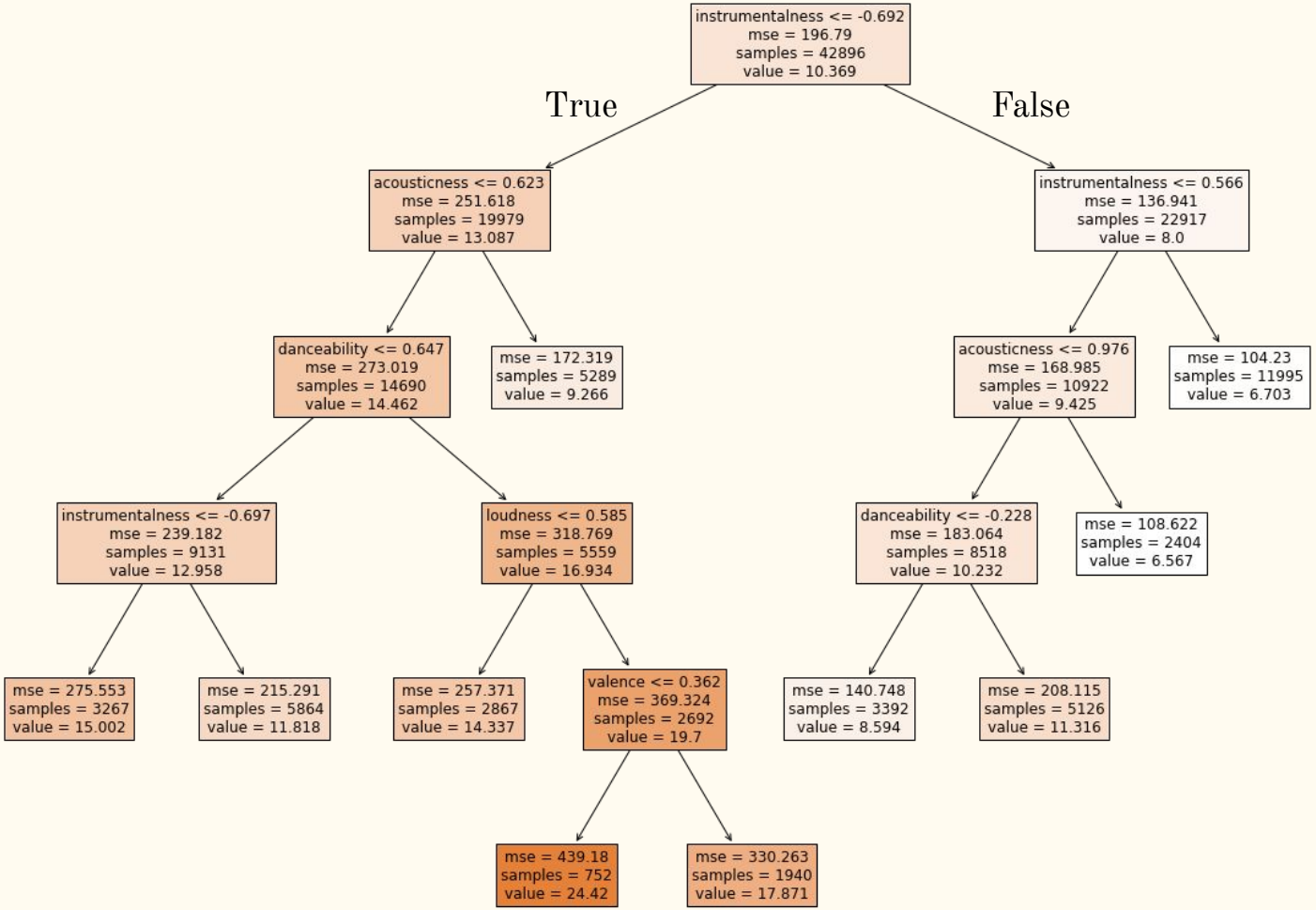
LR Negative Coefficients:

- Acousticness
- Instrumentalness
- Valence
- Energy

Density of predictions for KNeighbors



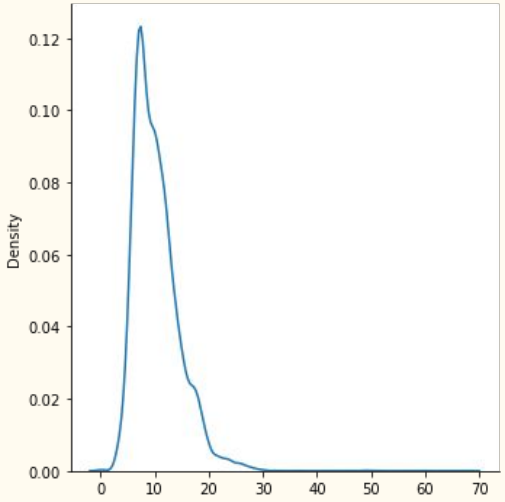
Decision Tree for Popularity



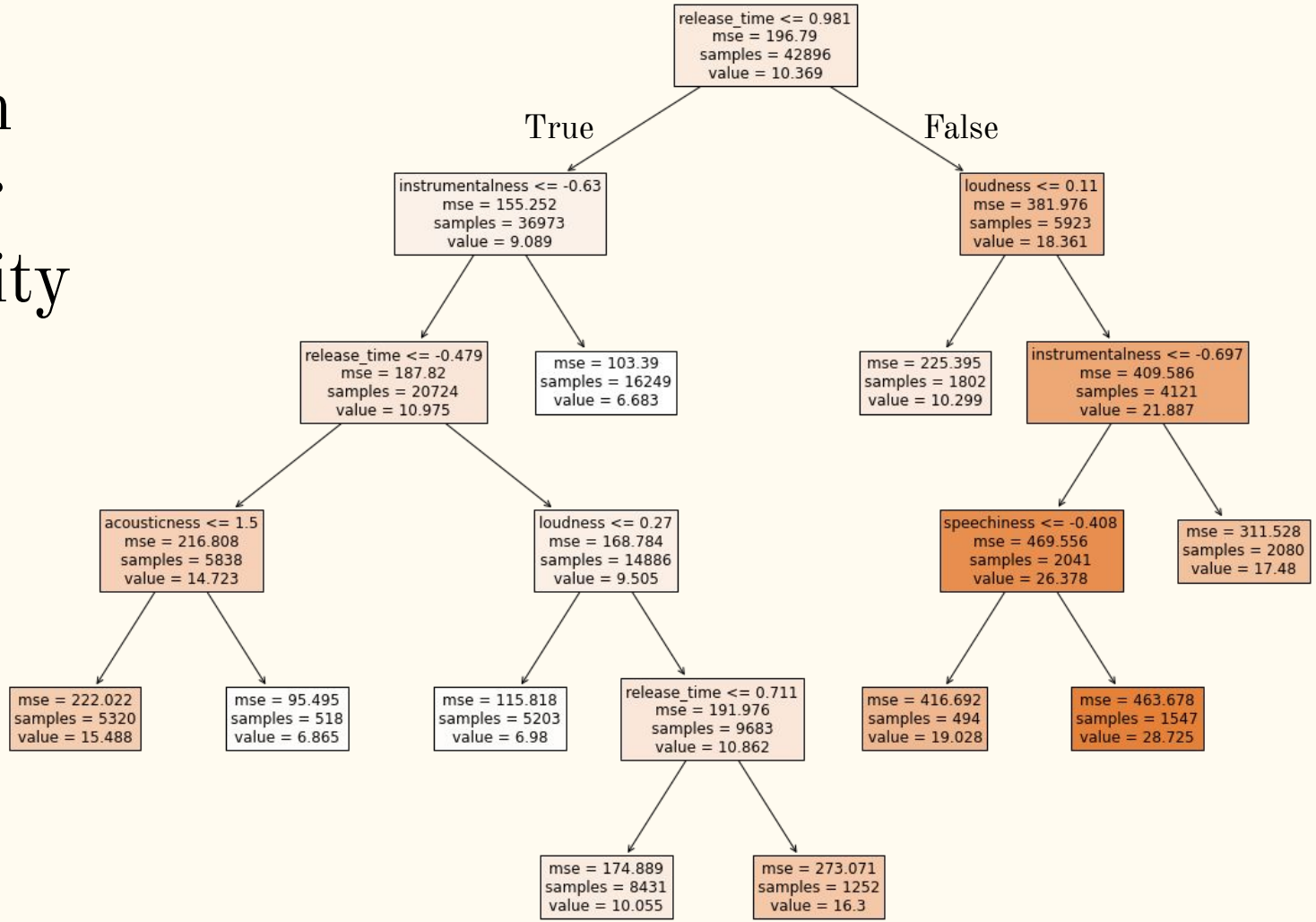
Using Regression to Predict Popularity (2)

- Included duration of song and release time of song as features
- Scores for the decision tree and KNeighbors improved (~.12 and .15)

New KNeighbors prediction density



Decision Tree for Popularity (2)



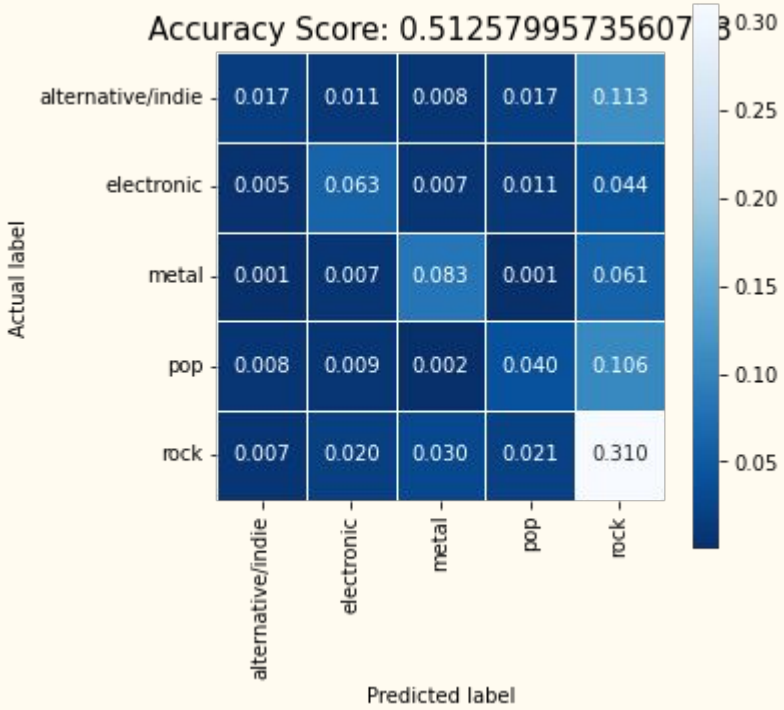
Analysis of Regression Results

- Of the features, high instrumentalness, acousticness, and energy seemed to correspond with lower popularities.
- High loudness and danceability seemed to correspond with higher popularities.
- Speechiness, liveness and tempo had little impact
- Songs that were either released recently (after August 2015) or a while ago (December 1998) were more likely to be popular.
- Popularity of artist is likely a much better predictor than the features we used.

Classification of Genres

- Accuracy around 50%, regardless of model

SVM:



Possible Next Steps

- Retrieve playlists from spotify API and try to predict the next song in a playlist
- Run a similar analysis on the lyrical content of songs to see what factors affect their popularity
- Determine what features are popular when grouped in a song together
- See how artist popularity would affect ability of regression models to predict song popularity

Purposes for this Analysis

- Could be used to predict the popularity of a song based on its features
- Could be used to recommend songs and artists to users of Spotify and other streaming services
- Could help artists to select songs off of an album to release as singles
- Track how the features of popular songs from each decade changed and predict how popular music will continue to change in the future
- Could be used to construct playlists of music with similar features to appeal to fans of that type of music

Things we learned from working on the data

- Our initial plans for clustering based on genres turned out to be unfeasible.
- Because music genres are not strictly defined, it was difficult to get authoritative labels of music genres.
- Consider the value of features and dataset source early.

Thank You

Any Questions?