Temporal Trends in Music Popularity

Matthew Menten, Kieren Ng, Braden Holmes, Tanner O’Rourke

First Steps

Getting a song set:

- Million Songs Dataset
  - https://labrosa.ee.columbia.edu/millionsong/
- How it was made
  - Integrated from Echo Nest and the CAL500 Dataset
  - “Getting the 200 top terms from The Echo Nest, then using each term as a descriptor to find 100 artists, then downloading as many of their songs as possible”
  - “A random walk along the similar artists links starting from the 100 most familiar artists”

Data Cleaning - Before

```
TRMMMQN128F4238509<SEP>SOGNNYL12A6D4F910B<SEP>Prince & The Revolution<SEP>Raspberry Beret (LP Version)
TRMMMKQ128F92EBCB5<SEP>SOOLRHW12A8C142643<SEP>Kreator<SEP>All of the same blood
TRMMMQK128F428AB1F<SEP>SOMKNZS12A8C13DEBC<SEP>Hall Of Fame<SEP>One Little Too Little
TRMMMPU128F42B134D<SEP>SOYKVON12A8C14097E<SEP>Frank Chacksfield<SEP>Cockleshell Heroes
TRMMMYP128F429A5E4<SEP>SOEXAWC12AB01817E6<SEP>Voyage<SEP>Trancesequence
TRMMMKN12903CB44A5<SEP>SOKQCDA12A6D4FA556<SEP>Wishbone Ash<SEP>Wonderful Stash
TRMMMUZ128F4238E64<SEP>SOFWVLR12A8C1328DC<SEP>Malavoi<SEP>La filo
TRMMMTP128F4276193<SEP>SOLRKBF12A8C13C88A<SEP>Cobra Verde<SEP>Throw It Away
TRMMMSA128F425FA38<SEP>SOBDLRM12A8C13A0AC<SEP>Lisa Brokop<SEP>Before He Kissed Me
TRMMMHS128F42A0971<SEP>SOMXBRA12AB017C656<SEP>The Emotions<SEP>Blessed
TRMMMJI12903CC3E5C<SEP>SODJAOH12A58A77B23<SEP>Safi Connection<SEP>Goa Amsterdam
TRMMMDJ128F92CC9F6<SEP>SOCPTIN12A8C14265F<SEP>Hawthorne Heights<SEP>Disaster [Demo Version]
TRMMMBU128F9305AC3<SEP>SODVOFJ12AB0181EE6<SEP>The Maytals<SEP>Night And Day
TRMMMFJ128F92E15AC<SEP>SODDEQU12AAF3B2FC8<SEP>Neffa<SEP>Passione
TRMMMNI12903CE0AF1<SEP>SONYUEW12AB018B373<SEP>Lil<0x19> O<SEP>My Everything [Screwed] (feat. Trae The Truth)
TRMMMFG128F425087B<SEP>SOWWQXA12A81C1F9CA<SEP>Billy Idol<SEP>Scream
TRMMMBM128F933EDF7<SEP>SONNHYN12AB018538B<SEP>The Vichy Government<SEP>The Man Delusion
TRMMMLV128F14821BE<SEP>SOCWWYB12A6D4F9AF4<SEP>Out Of The Grey<SEP>He Is Not Silent (Out Of The Grey Album Version)
TRMMMDN12903CD37AD<SEP>SOIRTNX12AB018B41D<SEP>Orquesta Sonara La Habana<SEP>Pepe El Mañoso
TRMMMLF128F4228771<SEP>SOLGCKQ12A8AE47CFA<SEP>Leo Marini<SEP>Acércate Más
TRMMMXD128F932AD86<SEP>SOZOAEF12AB017F0A9<SEP>Sheb Wooley<SEP>Mule Boogie
TRMMMQA128F14A454A<SEP>SOHMYGC12A6D4FAC4B<SEP>Diana Krall<SEP>Dancing In The Dark
TRMMMGL128F92FD6AB<SEP>SOHSSPG12A8C144BE0<SEP>Clifford T. Ward<SEP>Mad About You
TRMMMJM128F42648D0<SEP>SOCBSOR12A8C1314DD<SEP>Ray Conniff;Billy Butterfield<SEP>Heartaches
TRMMMHB128F92F0E3C<SEP>SOOKLXI12A8C142FB0<SEP>Weekend Nachos<SEP>Pain Over Acceptance
```

Data Cleaning - After

```
Prince & The Revolution - Raspberry Beret
Kreator - All of the same blood
Hall Of Fame - One Little Too Little
Frank Chacksfield - Cockleshell Heroes
Voyage - Trancesequence
Wishbone Ash - Wonderful Stash
Malavoi - La filo
Cobra Verde - Throw It Away
Lisa Brokop - Before He Kissed Me
The Emotions - Blessed
Safi Connection - Goa Amsterdam
Hawthorne Heights - Disaster
The Maytals - Night And Day
Neffa - Passione
Lil<0x19> O - My Everything
Billy Idol - Scream
The Vichy Government - The Man Delusion
Out Of The Grey - He Is Not Silent
Orquesta Sonara La Habana - Pepe El Mañoso
Leo Marini - Acércate Más
Sheb Wooley - Mule Boogie
Diana Krall - Dancing In The Dark
Clifford T. Ward - Mad About You
Ray Conniff - Heartaches
Weekend Nachos - Pain Over Acceptance
```

- Removed extraneous information from song and artist names
  - E.g. (LP Version), artist features, semicolons as separators, underscores used as commas
- Format to make spotify searches more successful

Getting Track Info from Spotify API

- Spotipy package used to make requests
  - Built in search function used to read in cleaned names + artists, pull ID
  - Spotify ID pulls attributes per song
- Pick out relevant information from JSON
  - Track ID, name, artists, release date, duration (ms), explicit
- Out of the million songs, ~850k successfully obtained from API
- Track ID allows us to get audio features from the API
- Rate limiting
  - Initial searches took approx. 48 hours

Data Cleaning - cont.

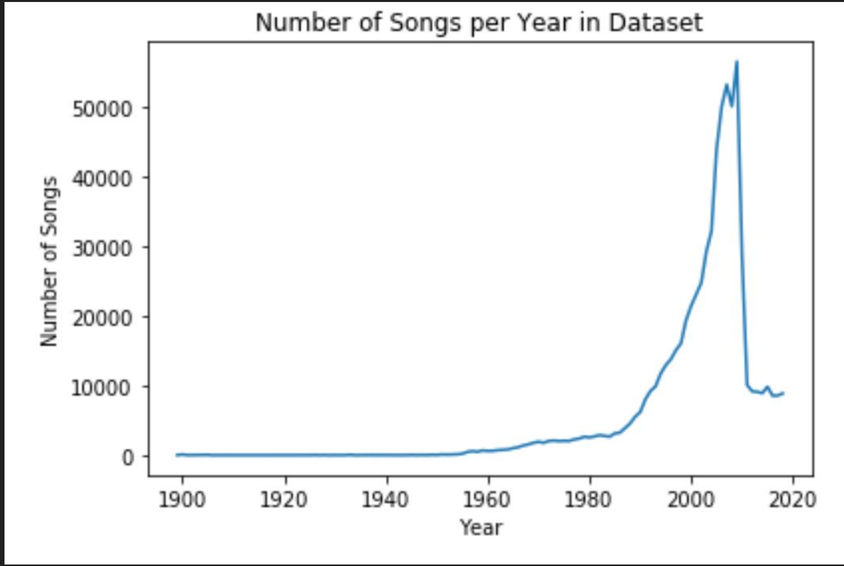
- Spotify data was still a little messy, either their estimates were incorrect or the person who submitted the song inputted wrong information
  - Date, Tempo, Time Signature, Duration
- Implemented SQL query to only pull necessary data, which increases performance and cleans the data
- Query also categorizes each attribute, in order to be used in analysis and to provide context to each datapoint

```
0|4|3|102.07|1|264200|0.704|0.65|0.777|-6.458|0.0245|0.212|0.00997|0.155
0|4|5|113.122|1|153040|0.511|0.912|0.527|-13.665|0.0513|0.679|0.619|0.101
0|4|9|130.01|1|198720|0.519|0.47|0.826|-6.012|0.0313|0.00336|0.139|0.0572
0|4|11|145.844|1|213693|0.648|0.904|0.273|-4.897|0.0472|0.0941|0.0924|0.0853
0|4|10|90.51|0|185000|0.764|0.384|0.652|-8.399|0.228|0.0234|0.0|0.309
```

```
2010|2010s|0|clean|time signature: 4|3|D sharp/E flat|102.07|moderate|1|major|264200|long|0.704|moderate/high danceability|0.65|moderate/high valence|0.777|high energy|-6.458|loud|0.0245|just music|
0.212|low acousticness|0.00997|less instrumental|0.155|studio recording
2013|2010s|0|clean|time signature: 4|5|F|113.122|moderate|1|major|153040|short|0.511|moderate/high danceability|0.912|high valence|0.527|moderate/high energy|-13.665|loud|0.0513|just music|0.679|mod
erate/high acousticness|0.619|moderate instrumentalness|0.101|studio recording
1998|1990s|0|clean|time signature: 4|9|A|130.01|moderate|1|major|198720|medium|0.519|moderate/high danceability|0.47|low/moderate valence|0.826|high energy|-6.012|loud|0.0313|just music|0.00336|low
acousticness|0.139|less instrumental|0.0572|studio recording
2009|2000s|0|clean|time signature: 4|11|B|145.844|moderate|1|major|213693|medium|0.648|moderate/high danceability|0.904|high valence|0.273|low/moderate energy|-4.897|loud|0.0472|just music|0.0941|lo
w acousticness|0.0924|less instrumental|0.0853|studio recording
2008|2000s|0|clean|time signature: 4|10|A sharp/B flat|90.51|slow|0|minor|185000|medium|0.764|high danceability|0.384|low/moderate valence|0.652|moderate/high energy|-8.399|loud|0.228|just music|0.0
234|low acousticness|0.0|less instrumental|0.309|studio recording
```

Data Statistics

- 850k → 690k with duplicates removed
- 675k with secondary data cleaning
- Songs ranging from 12-31-1899 to 11-09-2018
- 81k distinct artists
- Longest song: 97.17 minutes
- Shortest song: 1 second
- Speechiest song: Stjernene by Agnar Mykle



Artists with the most songs

Joan Baez- 149 songs

Bruce Springsteen- 138

Daniel Johnston- 128

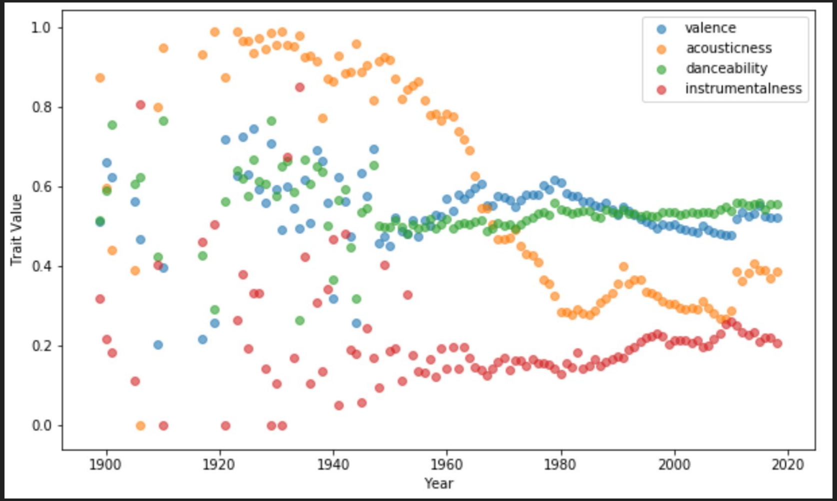
Gilberto Santa Rosa- 128

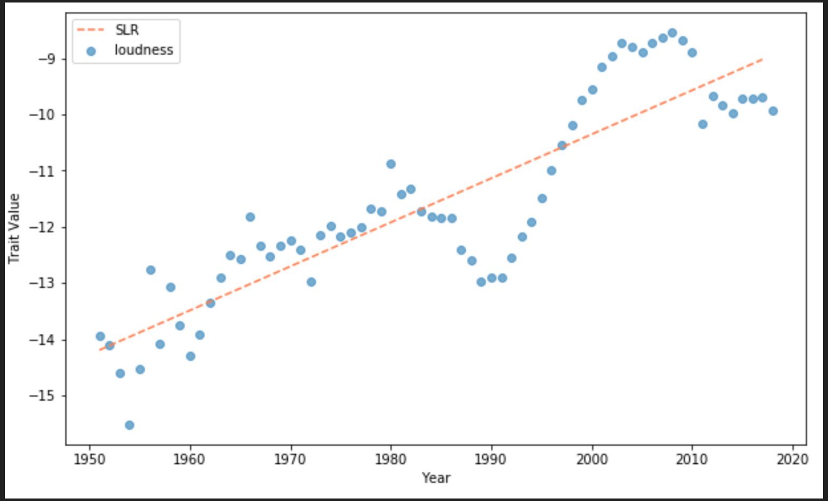
Bob Dylan- 128

Other notable artists: Johnny Cash, The Beatles, Stevie Wonder, Aretha Franklin The Cure, The Rolling Stones

Temporal Analysis - Regression

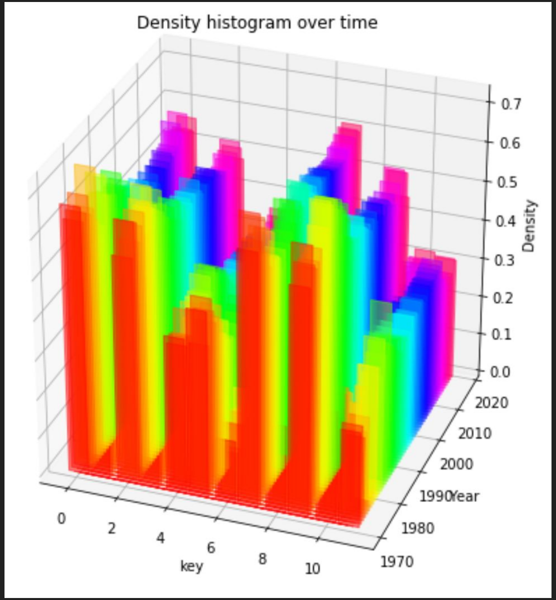
- Mean of trait for each year
- Simple Linear Regression

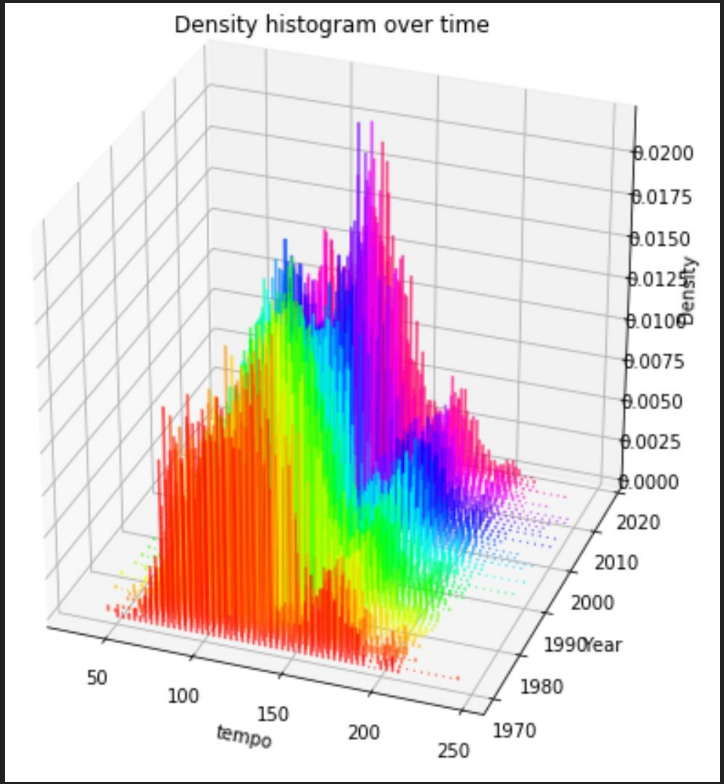




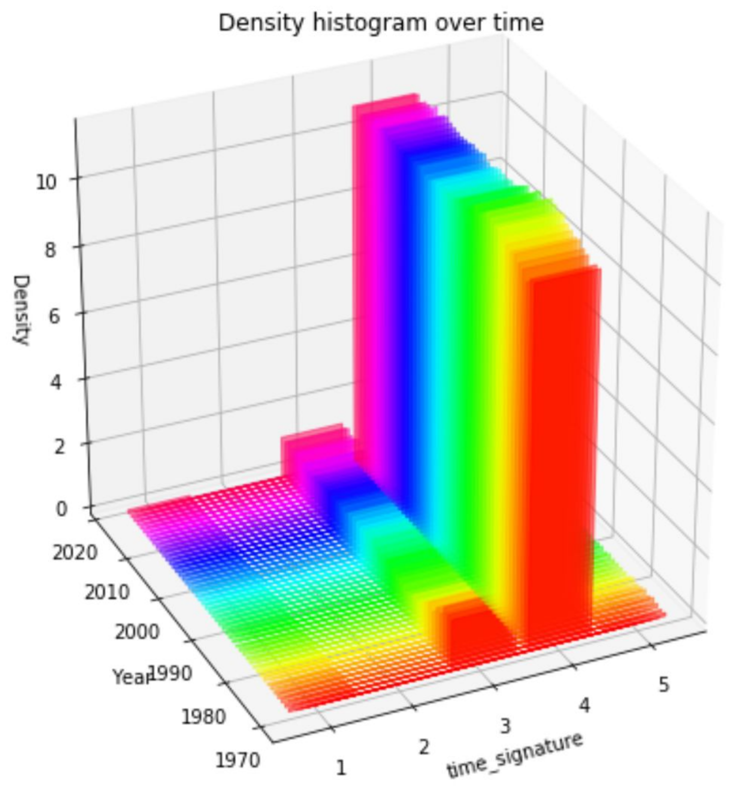
Temporal Analysis - 3D histograms

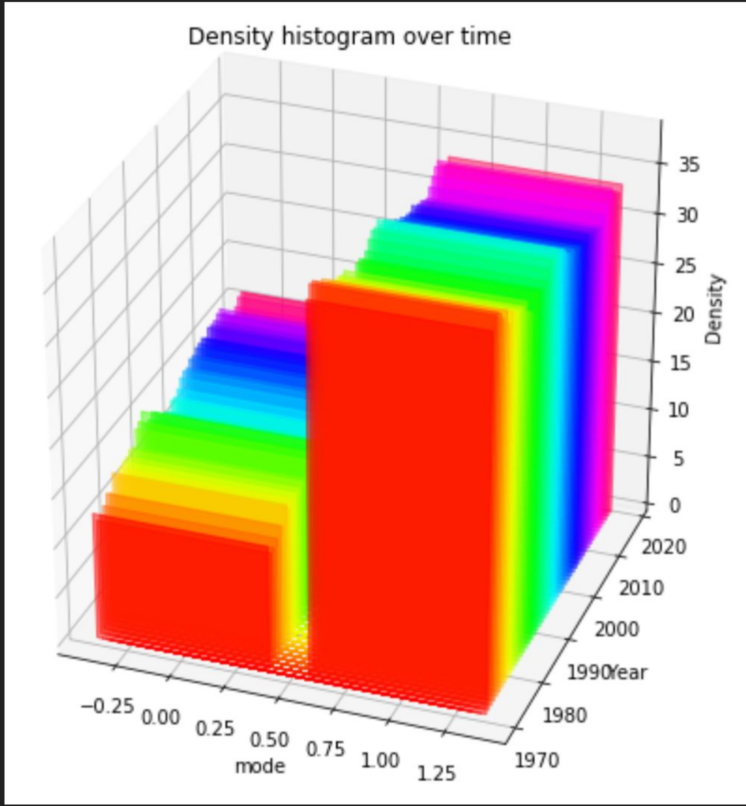
- Histogram of trait for each year
- Revealed trends in key and tempo





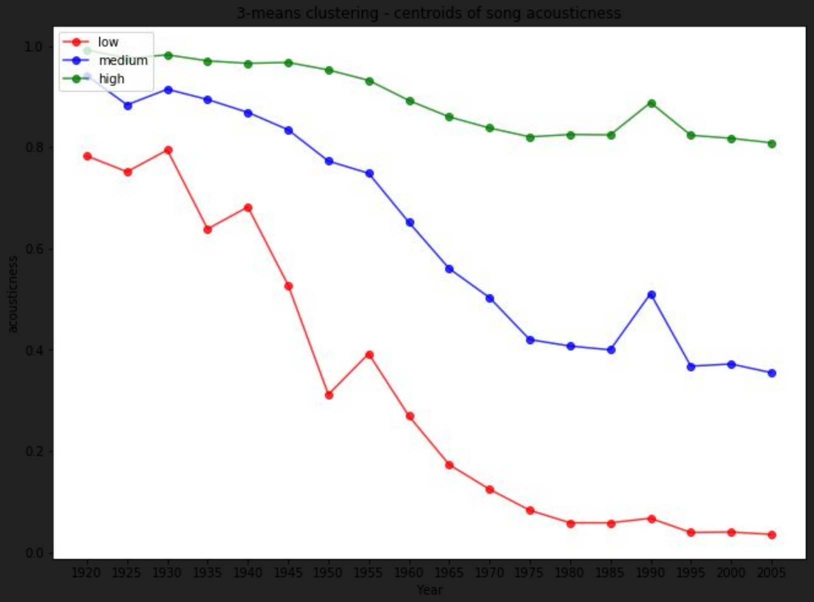
Temporal Analysis - 3D histograms

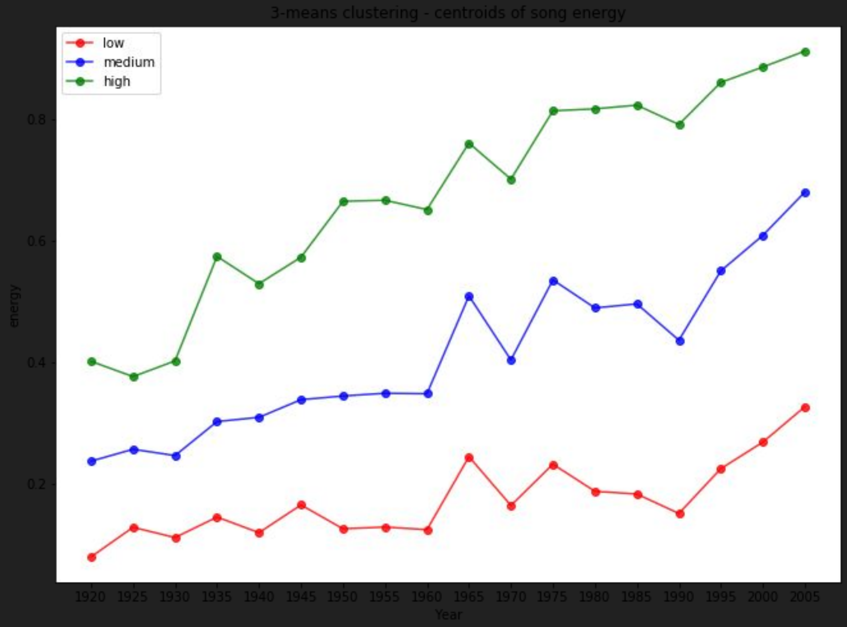




Temporal Analysis - 3-Means Clustering

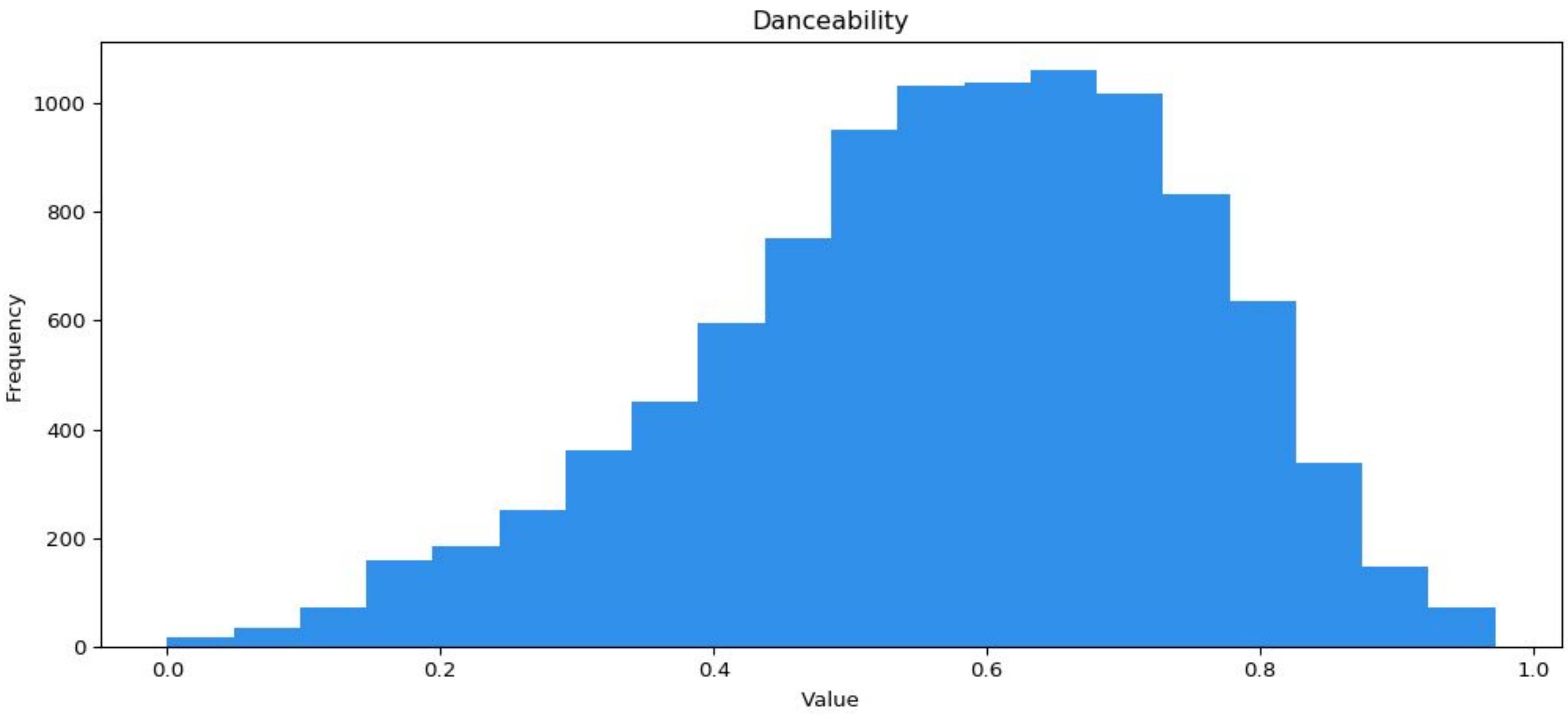
- 3-Means Clustering for 5 year periods
- Allows for more advanced analysis of trait's temporal attributes





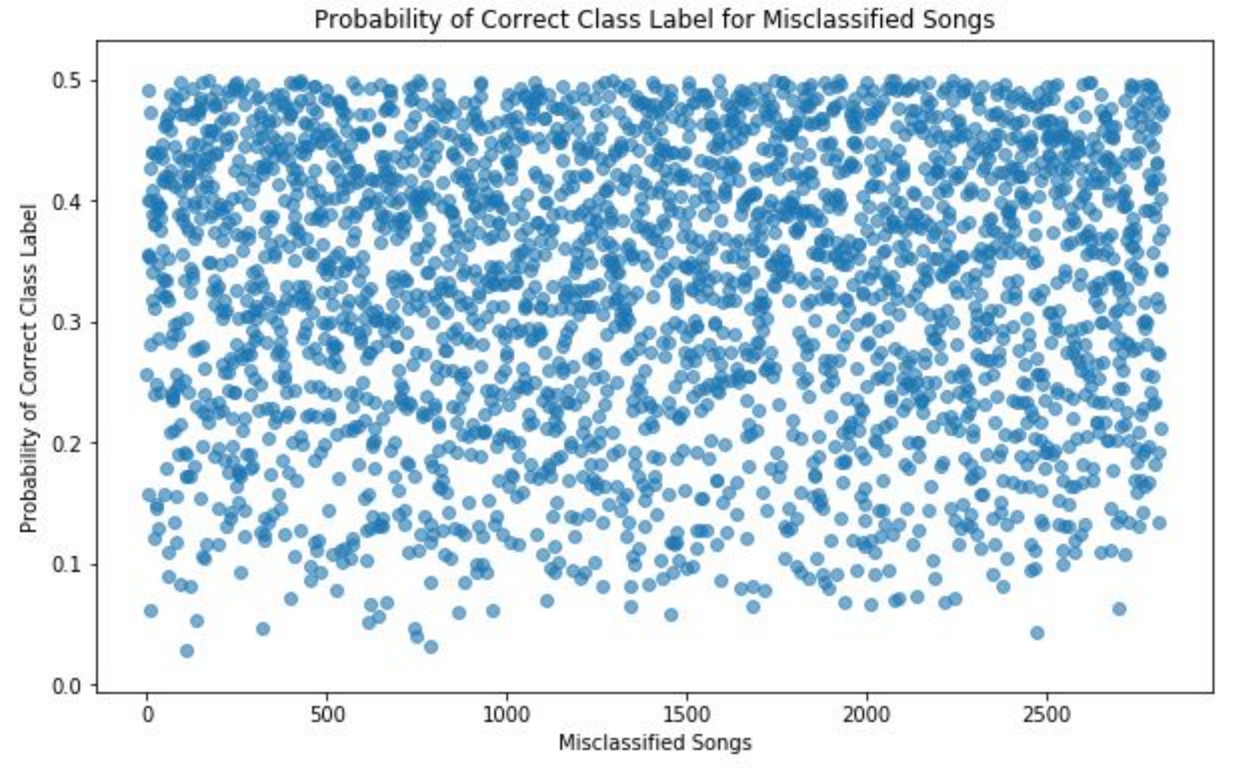
Danceability Classification/Prediction

- SKLearn
  - Logistic Regression
  - K Nearest Neighbors
- Full attribute list
  - Release year, explicit, time signature, key, tempo, mode, energy, loudness, speechiness, acousticness, instrumentalness, liveness, valence
- Goal: classify low or high danceability based on features
- Accuracy
  - Tested repeatedly with random samples of 10,000 rows from our dataset
  - Average of 73% accurate


Danceability
Frequency
1000
800
600
400
200
0
0.0
0.2
0.4
0.6
0.8
1.0
Value

Danceability Classification/Prediction cont.

- Feature Reduction
  - Used recursive feature elimination to trim down our model for logistic regression
- Reduced attribute list
  - 6 attributes are most significant for classifying danceability
  - Time signature, energy, speechiness, acousticness, liveness and valence
- Accuracy
  - 72% accurate in repeated sampling
  - No significant difference b/t full model and reduced model
- KNN classification
  - Curse of dimensionality when full feature list was used


Probability of Correct Class Label for Misclassified Songs
Probability of Correct Class Label
0.0
0.1
0.2
0.3
0.4
0.5
0
500
1000
1500
2000
2500
Misclassified Songs

Lessons Learned

- Data rates can drastically affect data collection time
- Database structure is important, can affect runtimes
- Data cleaning is important and time consuming, can cause problems with analysis
- Make sure you have the attributes to do the analysis you want
  - Popularity missing
- Make informed decisions about what analysis methods to use beforehand

Conclusion - Key Findings

- Some keys and tempos are more common than others
  - Near 100 bpm most common
  - Major keys are more popular than minor keys
  - 4 / 4 is the most popular time signature
- The rise of modern electronic music may be correlated to..
  - Increasing musical energy
  - Decreasing musical acousticness
  - Decreasing correlation of tempo vs. valence
- Some traits help predict danceability more than others
- Important features for prediction danceability