what’s the vibe?

an analysis of spotify song data

Tianna + Mohsin + Deen

introduction + questions of interest

- ★ increasingly so, machine learning techniques are used to make claims about music taste and listening habits
- ★ can we use machine learning to classify the *emotion* behind a song?
- ★ are only “happy” songs popular?
- ★ have popular songs reflected the times?

the spotify api explained…

★ **the Spotify API provided us access to live song data:**

- public user playlist songs (that we used for training our model)
- top 100 Billboard songs (this is the data in the “wild” that we classify)
- features of individual songs that ended up as our model variables

the data:

★ 3,946 total songs ★

905 “happy” and 1,114 “sad” songs

“billboard top 100” songs since 2000

in order to classify a song…

★ ***spotifyr::get_audio_features()***

- **tempo:** overall estimated tempo of a track in beats per minute (BPM)
- **mode:** Major (generally happy) or Minor (generally sad)
- **valence:** tracks with high valence sound more positive (e.g. happy), while tracks with low valence sound more negative (e.g. sad). 0.0 to 1.0 describing the musical positiveness.
- **danceability:** how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. 0.0 is least danceable and 1.0 is most danceable

we scraped 2,019 songs

from 28 sad and happy playlists

and 24 users

user:mohsin

user:12128612688

user: tiannacouch

- **random forest + model building**
  - ★ a Random Forest Model was built on **user playlist data** to classify the songs into happy or sad, using the following parameters: **tempo, valence, mode, danceability**
  - ★ **our final model included a RF with 201 trees with an mtry = 1 (found through cross validation method).**

error rates + model stuff…

on 201 trees with an mtry = 1

OOB Error = 0.179

test Error = 0.26

we hope we can confidently predict wild data given our ~ 80% accuracy

★ happy & sad song profiles ★

“I’m Still Standing”

by Elton John

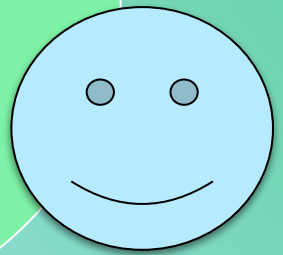
Tempo: 176.808 bpm

Valence: .772

Mode: Major

Danceability: .504

Classification: Happy



“When I Was Your Man”

by Bruno Mars

Tempo: 72.795 bpm

Valence: .387

Mode: Major

Danceability: .612

Classification: Sad

(from the test data)

classification of billboard data…

- ★ **our billboard data includes:**
  - top 100 songs for each year between 2000-2020
  - using the song information **we obtained track features for each of these songs**
  - this data was put in to our random forest model
- ★ **here's what we found …**

our model isn’t perfect…

“Bad Day”

Daniel Powter

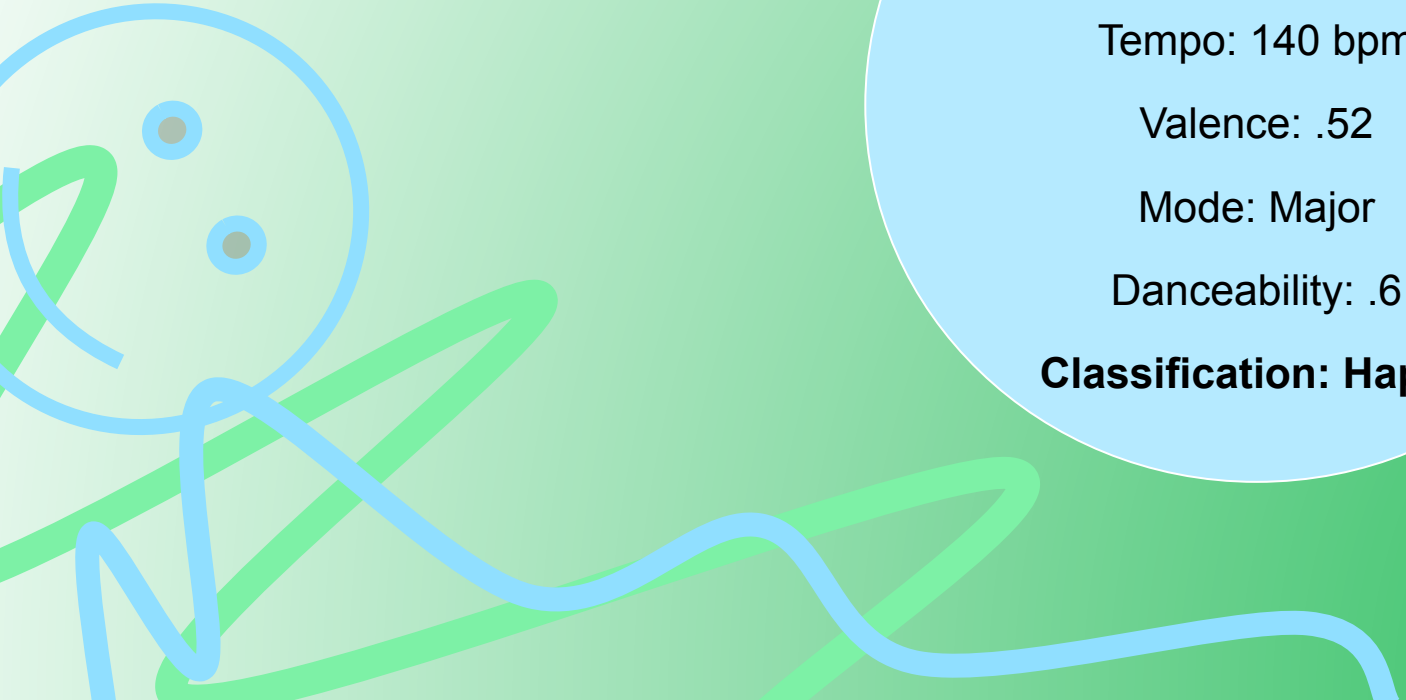
Tempo: 140 bpm

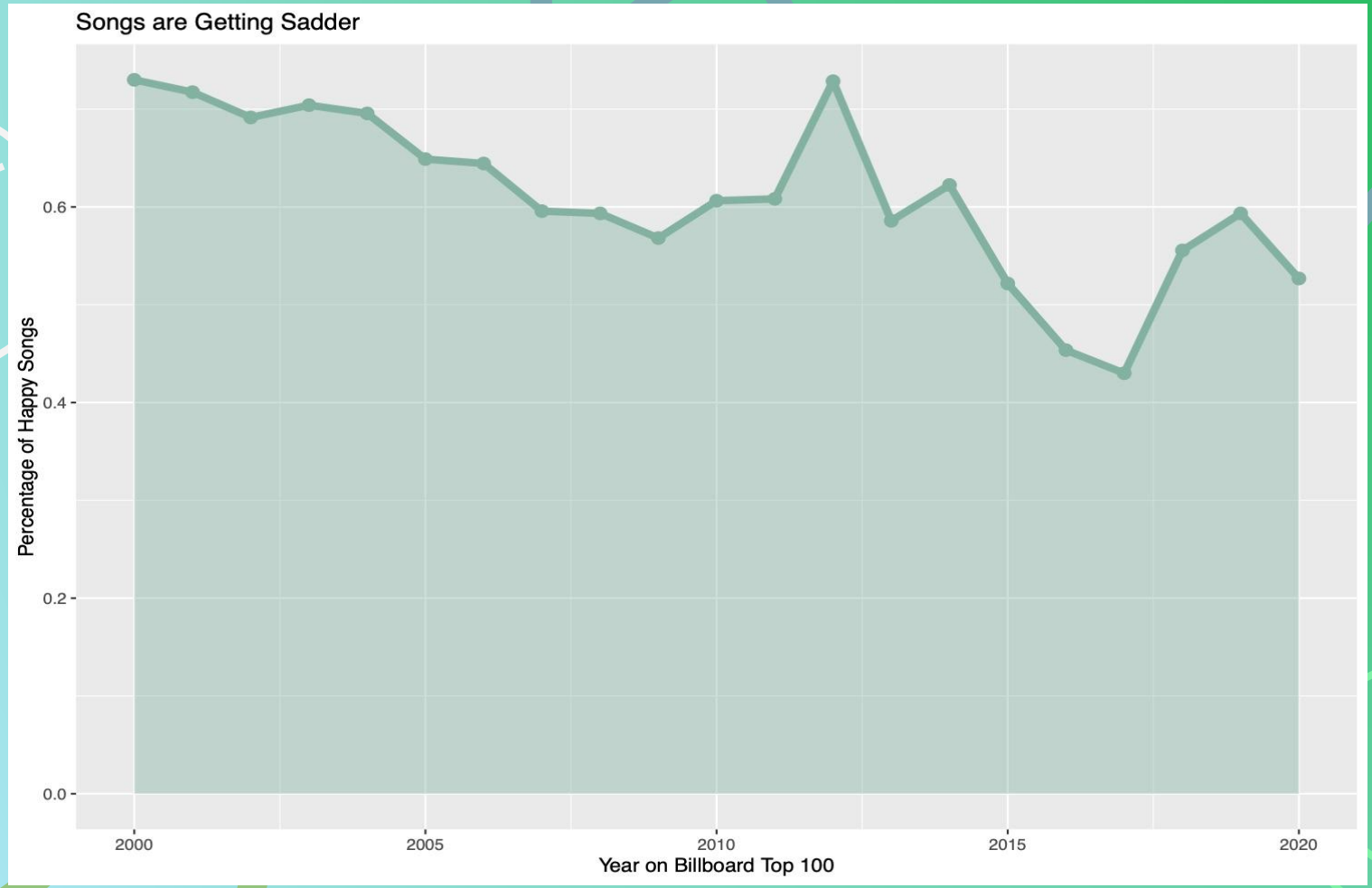
Valence: .52

Mode: Major

Danceability: .6

Classification: Happy




Songs are Getting Sadder
Percentage of Happy Songs
0.0
0.2
0.4
0.6
2000
2005
2010
2015
2020
Year on Billboard Top 100

looking closely…

2020

Top Song: **Blinding Lights - The Weeknd**
Classification: **Sad**
Overall % of Happy Songs: **50.5%**

Notable Events: COVID Pandemic

2016

Top Song: **Love Yourself - Justin Bieber**
Classification: **Happy**
Overall % of Happy Songs: **45.5%**

Notable Events: Election of Trump, Death of Harambe

2008

Top Song: **Low - Flo Rida**
Classification: **Sad**
Overall % of Happy Songs: **62.7%**

Notable Events: Great Recession

conclusions, limitations + what we wish we did

- ★ clear trend of songs getting sadder over past 20 years!
- ★ more areas of classification! not all songs are just happy or sad
- ★ can we assume our data is representative? can we assume tianna's breakup playlist is what defines sadness?
- ★ lyrics are incredibly important to a song's "mood". however scraping lyrics for these songs is morally muddy. a sentiment analysis of lyrics could inform this model.

ethics + further discussion

- ★ the obsession with classification – maybe humans can't use a computer to pick a sad song out of the bunch. *vibes cannot be quantified.*
- ★ isn't it weird spotify can access all of this data? you can pull profile photos using the spotify API. should we all have access to this?
- ★ although we found alternative routes for gathering lyrics data, we ultimately decided not to so that we could respect the artists' copyrights

citations

Charlie Thompson, Daniel Antal, Josiah Parry, Donal Phipps and Tom Wolff (2021). spotifyr: R Wrapper for the 'Spotify' Web API. R package version 2.2.3. https://CRAN.R-project.org/package=spotifyr

Garrett Grolemund, Hadley Wickham (2011). Dates and Times Made Easy with lubridate. Journal of Statistical Software, 40(3), 1-25. URL https://www.jstatsoft.org/v40/i03/

Kuhn et al., (2020). Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles. https://www.tidymodels.org

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

Yihui Xie (2021). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.33.

thanks!