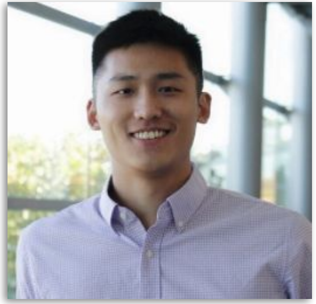
Visualizing Spotify Music Trends

Team Ex-penn-DATAbles

Aaron "*Shotgun*" Lee

Jeff "*Galgo*" Toussaint

Jeremy "*Christmas*" Strauch



Tianchi "*Yin Yang*" Zhang

Finding Data

- **Data Source:** Kaggle (Spotify Dataset 1920-2020, 160k+ Tracks)
- **Background:**
  - The story of sound recording, and reproduction, began in 1877, when the man of a thousand patents, Thomas Edison, invented the phonograph.
  - The commercialized music industry has only been around for about 100 years as the recording and radio technology emerged in the early 90s.
- **Core message:**
  - **Objective:** Uncover music trends. Understand how artists, genres, the features of audio and the popularity of songs developed through time.

- The Acoustic era (1877–1925)
- The Electrical era (1925–1945)
- The Magnetic era (1945–1975)
- The Digital era (1975–Present)

Source: Charm
Source: Wikipedia

Spotify

Developing Questions

1. From which era are the songs most popular now? (Tianchi)
2. Who are the most "*timeless*" artists? (Tianchi)
3. How do popular genres compare with each other? (Jeremy)
4. How have the audio features changed over time? (Jeff)
   a. How has people's music taste developed? (Jeff)
5. Is there any correlation between audio features? (Aaron)
6. Is there a way to predict popularity given audio features? (Aaron)
7. How can Spotify leverage the data? (Jeff)

Spotify

~175,000
Total Count of Songs

1921-2021
Data Timeframe

30,000+
Total Count of Artists

9
Audio Features

3,000+
Total Count of Genres

4
Beginner Analysts

Spotify®

Data Dimensions

19

Total Variables

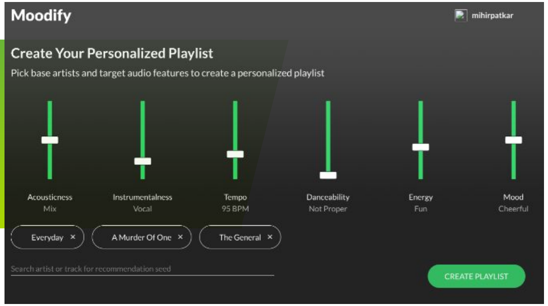
Audio Features

- Acousticness (Ranges from 0 to 1)
- Danceability (Ranges from 0 to 1)
- Energy (Ranges from 0 to 1)
- Speechiness (Ranges from 0 to 1)
- Instrumentalness (Ranges from 0 to 1)
- Valence (Ranges from 0 to 1)
- Liveness (Ranges from 0 to 1)
- Tempo (Float typically ranging from 50 to 150)
- Loudness (Float typically ranging from -60 to 0)

Objective Facts

- Id (Id of track generated by Spotify)
- Popularity (Ranges from 0 to 100)
- Duration_ms (Integer typically ranging from 200k to 300k)
- Year (Ranges from 1921 to 2020)
- Artists (List of artists mentioned)
- Name (Name of the song)
- Mode (0 = Minor, 1 = Major)
- Explicit (0 = No explicit content, 1 = Explicit Contents)
- Key (All keys on octave encoded as values ranging from 0 to 11, starting on C as 0, C# as 1 and so on…)
- Release_date (Date of release)

Audio Features



Acousticness - A confidence measurement from 0.0 to 1.0 of whether the track is acoustic.

Danceability - Describes how suitable a track is for dancing based on a combination of musical elements.

Energy - Measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity.

Instrumentalness - Detects whether a track contains no vocals. For instance, “Ooh” and “aah” sounds. The closer it is to 1.0, the greater the chances it contains no vocal content.

Speechiness - Detects spoken words in a track. Values above .0.66 describe tracks that are probably made entirely of spoken words.

Tempo - An estimated beats per minute. For example, the speed or pace of a track.

Valence - A measurement from 0.0 to 1.0 - Track will high valence are flagged as positive emotions whereas, low valence sound are more negative (i.e. sad, depressed).

Liveness - the presence of an audience in the track. For instance, above 0.8 produces a stronger likelihood that the track is live.

Data Cleaning/Descriptive Analysis

Data Cleaning:

- No missing data
- Data types are correct (int, float64, object)
  - No mixed data types
- Artists are lists of values
  - Lists vary depending on number of collaborators
- Delete unwanted fields

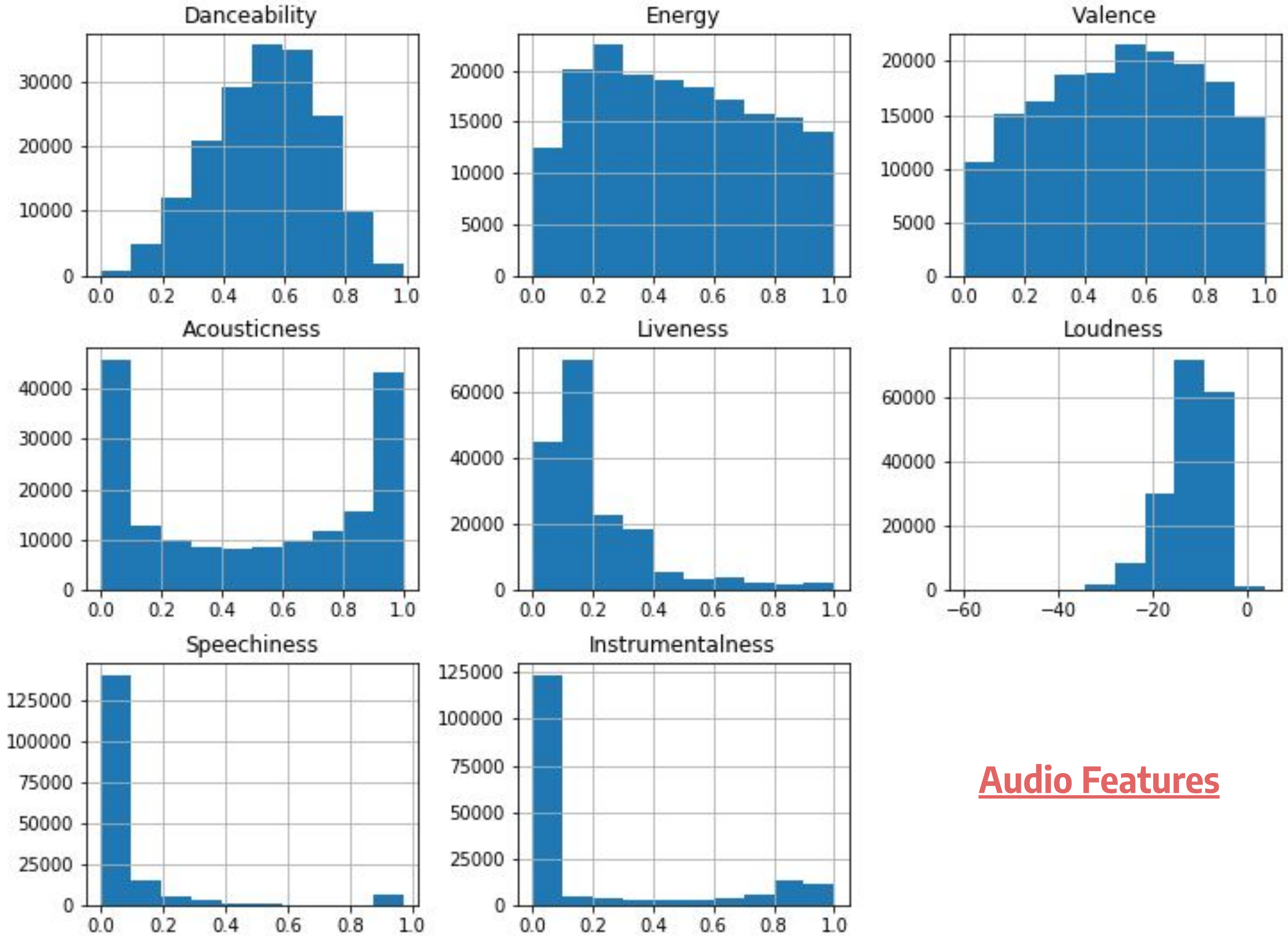
	acousticness	artists	danceability
0	0.991000	['Mamie Smith']	0.598
48	0.689000	['ST', '98', 'Niklas O']	0.690

```
df.isna().any()

acousticness        False
artists             False
danceability        False
duration_ms         False
energy              False
explicit            False
id                  False
instrumentalness    False
key                 False
liveness            False
loudness            False
mode                False
name                False
popularity          False
release_date        False
speechiness         False
tempo               False
valence             False
year                False
dtype: bool
```

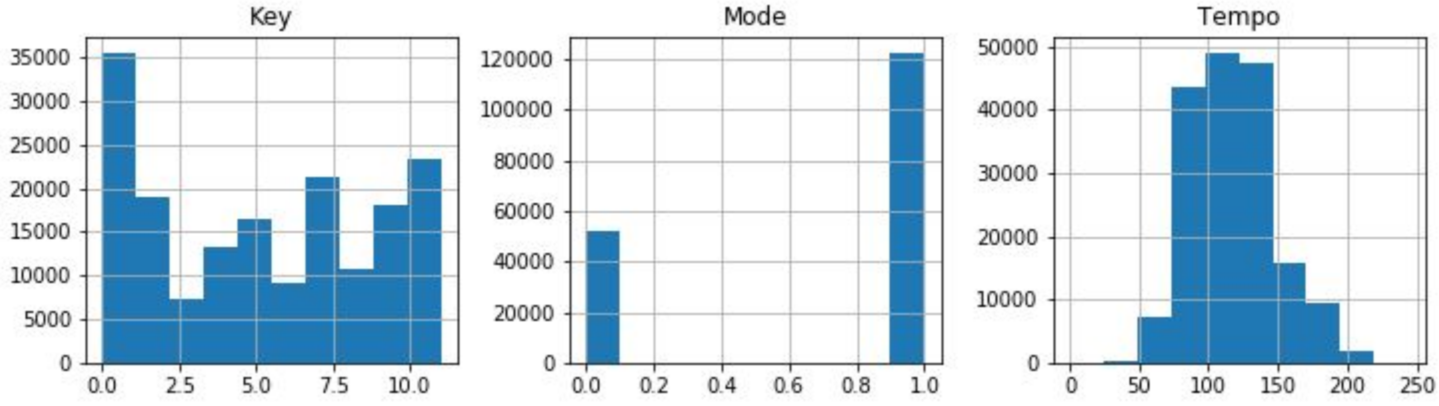
Descriptive Analysis:

- Histograms: audio features, music characteristics, other

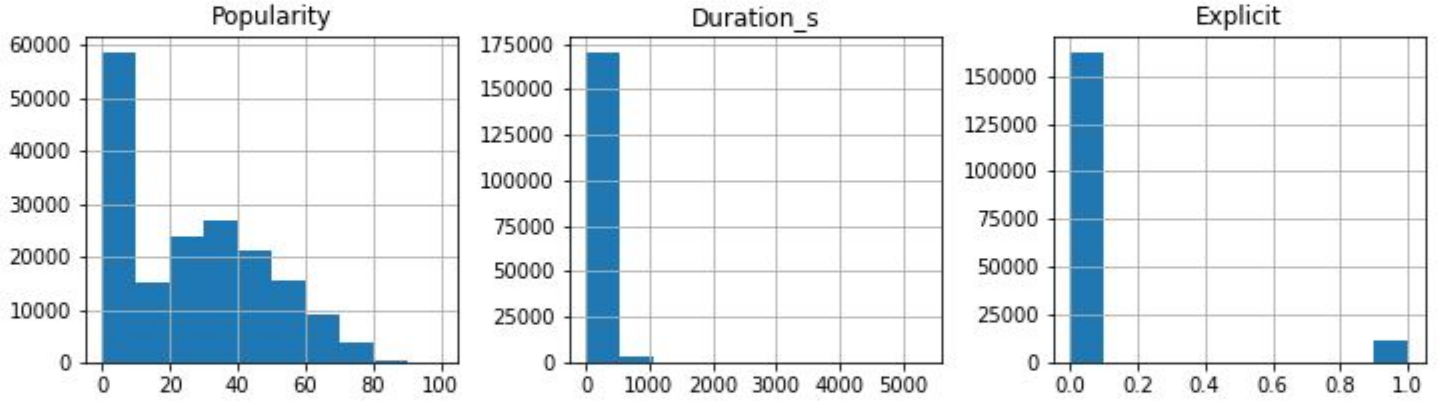


Audio Features

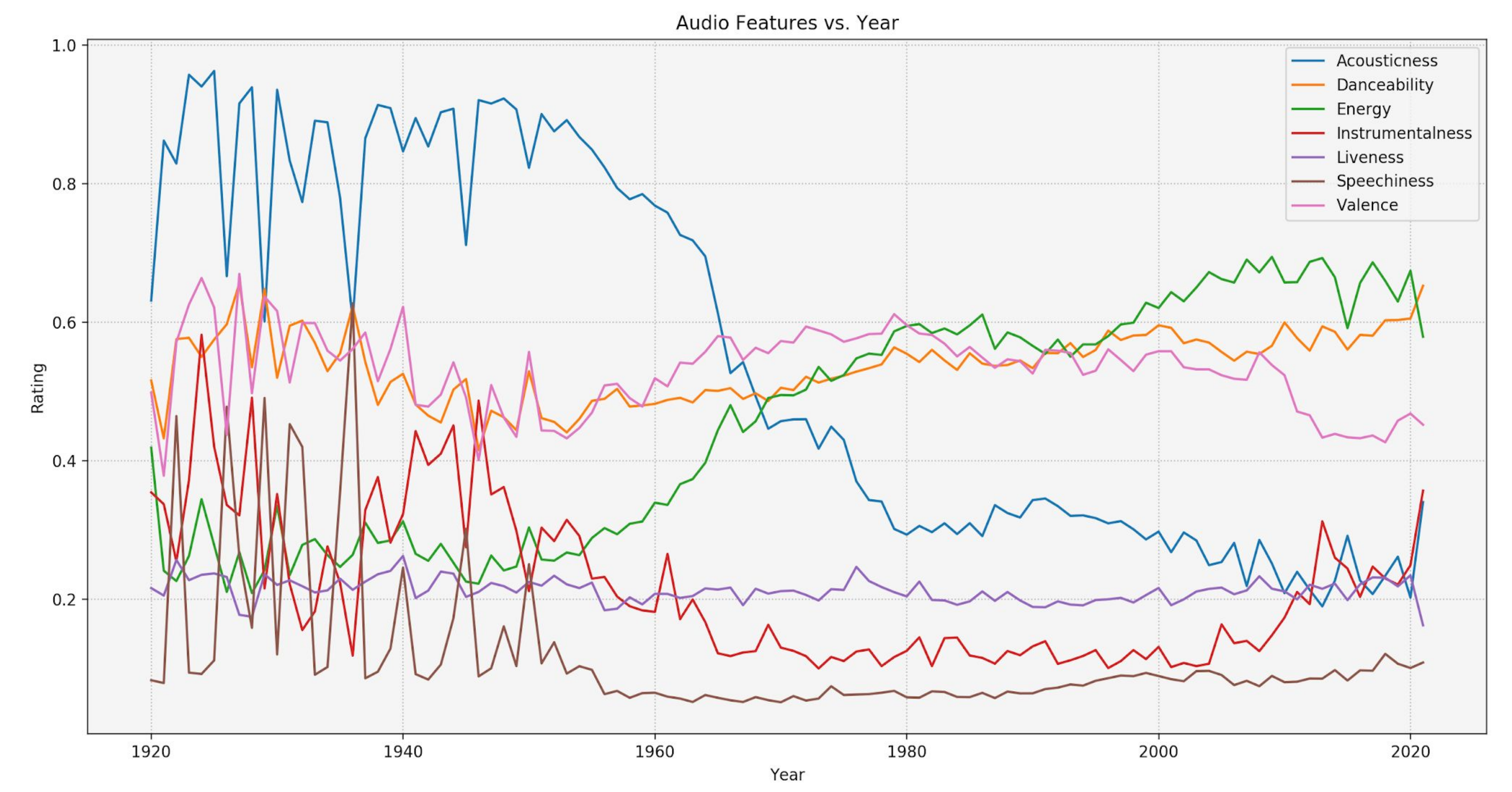
Music Characteristics

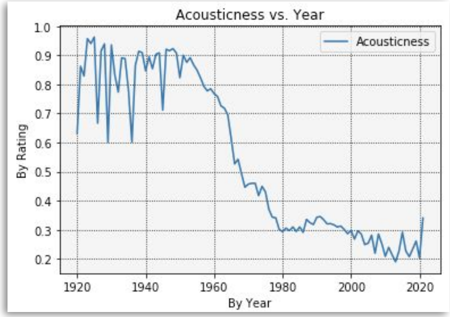


Other

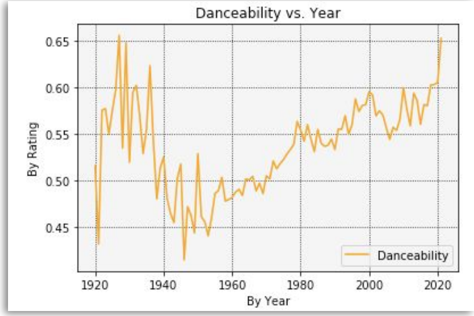


Analysis: Trends Over the Years

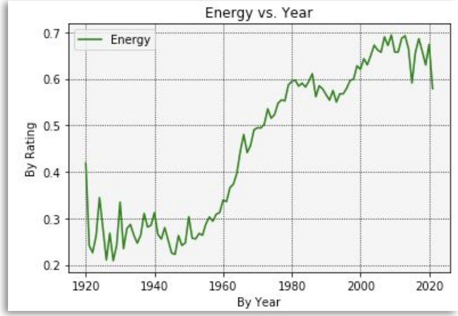




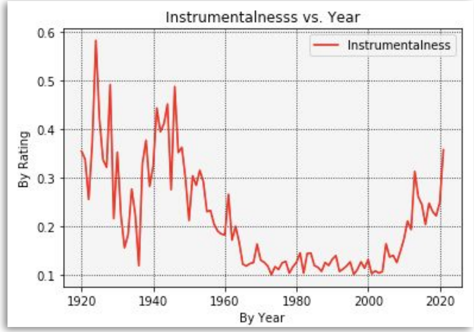
Acousticness dropped after 1960 after stabilizing around 30%



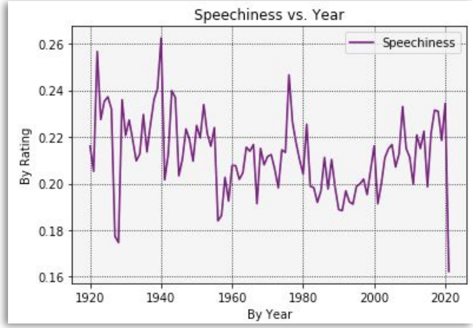
Danceability took a slight dip around 1940-1960, but later displays a positive uprise.



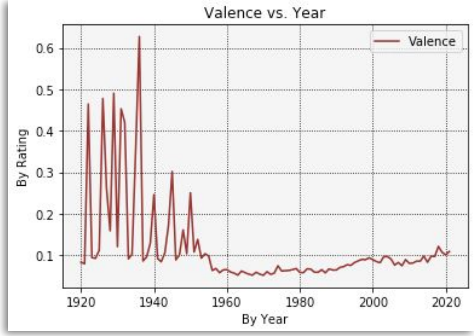
After 1940, Energy in within songs has been on the rise.



Instrumentalness proves to be sporadic between 1920-1960, but later stabilized with a increase spike.



Speechiness remains infrequent throughout the years with an average of 20%



Valence starts shifty but stays poised after 1960 with a baseline of 0.1

Time Series Analysis Dashboard

Analysis: Popular Genres

Major Question

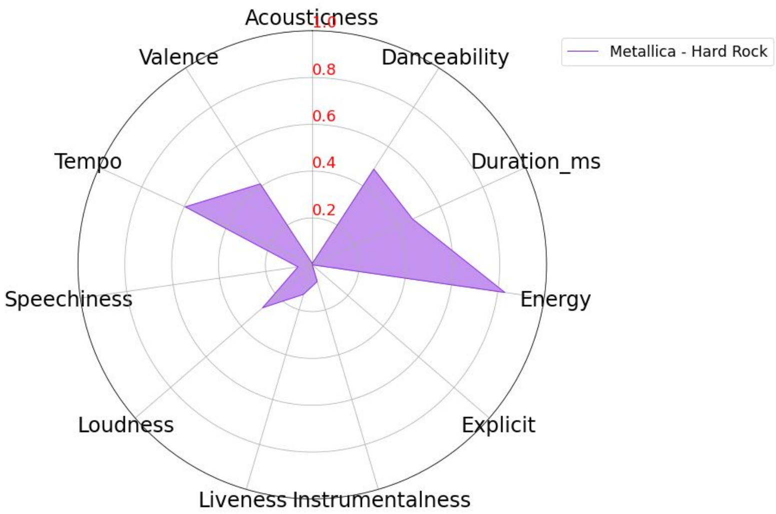
- How do song characteristics differ for major genres?

Radar Plot Comparison - Methodology

- 3,000+ genres defined in spotify; some artists span several genres
- Genre lists in alphabetical order, not by major themes
- Chose five of today's artists from Billboard 200
- Artists represent major genres themes
- Compared averaged values for top 20 songs
- Loudness, tempo, duration were normalized prior to plotting

Genres Explored

- Hard rock, EDM, rap, dance and country

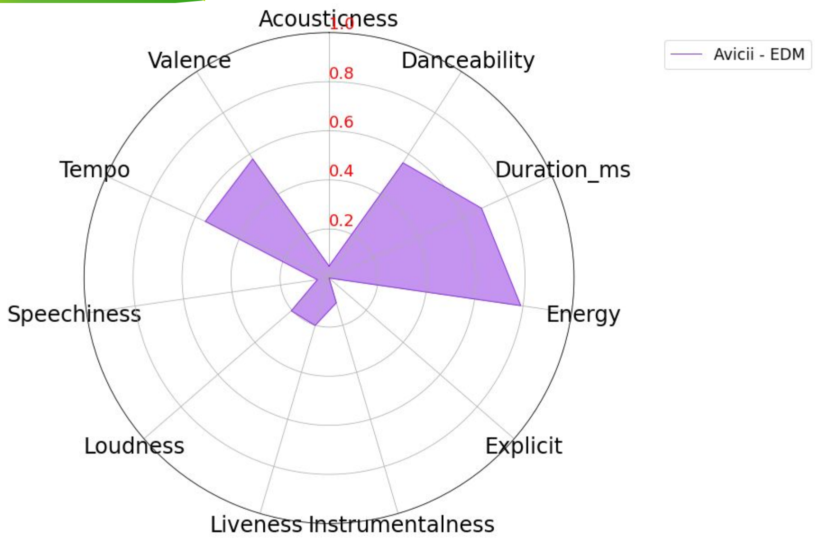


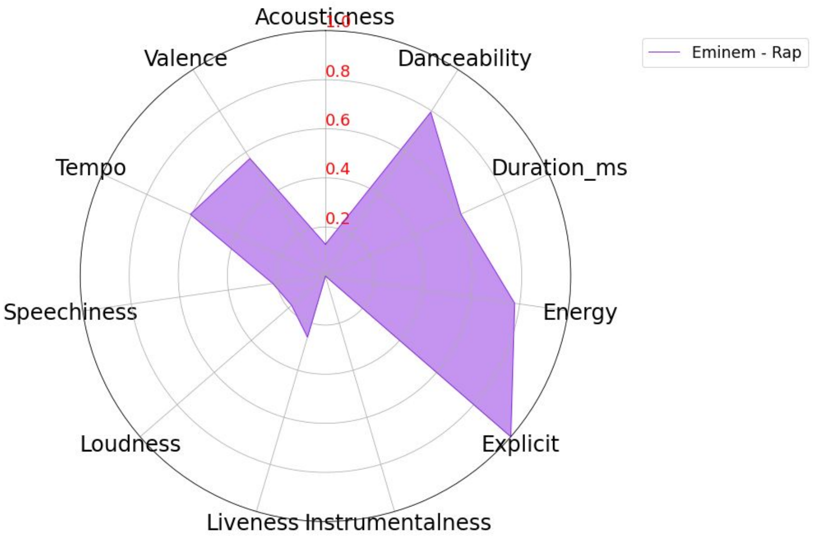
Metallica - Hard Rock

Surprises: danceability, explicit, loudness

Expected: valence, speechiness, accousticness, liveness, instrumentalness

Analysis: Popular Genres





Avicii - EDM

Surprises: loudness

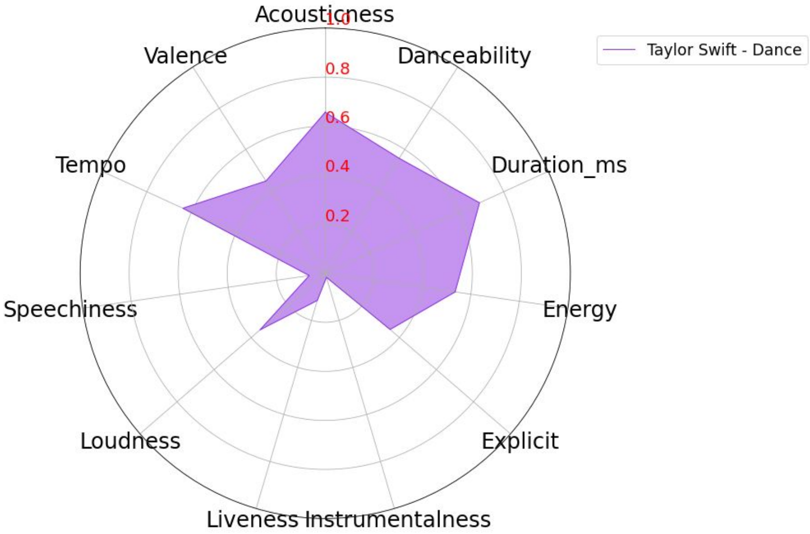
Expected: energy, speechiness, danceability, instrumentalness

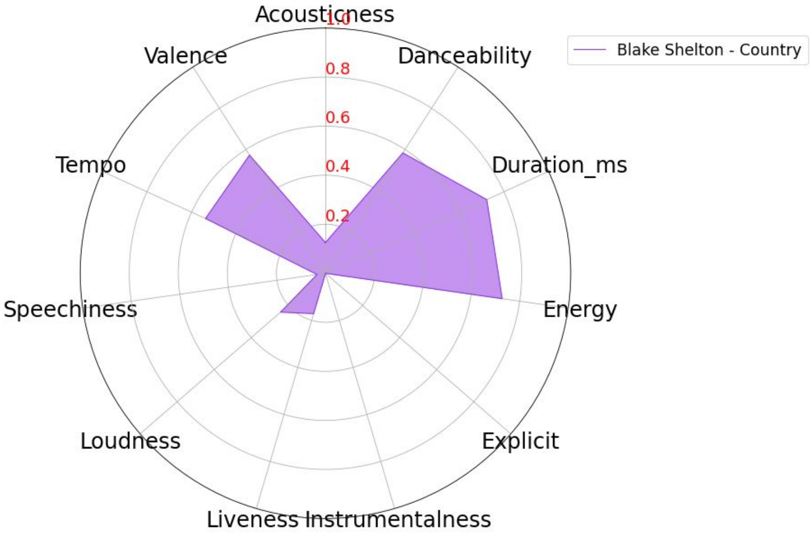
Eminem - Rap

Surprises: valence, speechiness

Expected: explicit, danceability, tempo, energy

Analysis: Popular Genres





Taylor Swift - Dance/Pop

Surprises: explicit, loudness, accousticness, speechiness

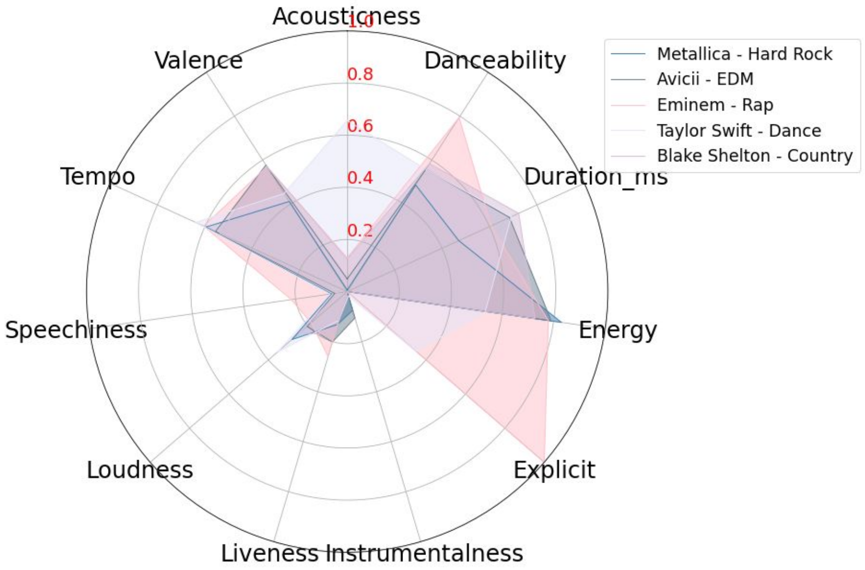
Expected: lenergy, danceability, tempo

Blake Shelton - Country

Surprises: loudness, duration, accousticness, valence, danceability

Expected: liveness, speechiness, explicit

Analysis: Popular Genres - Summary

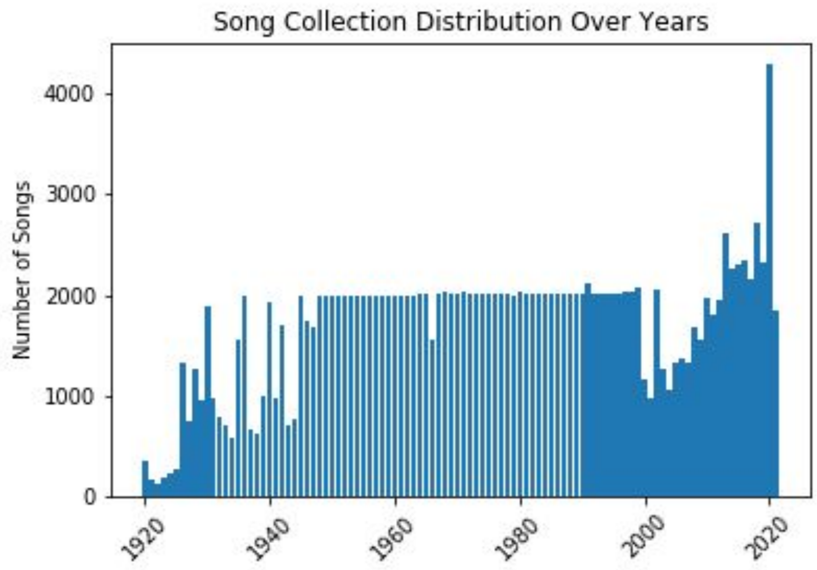


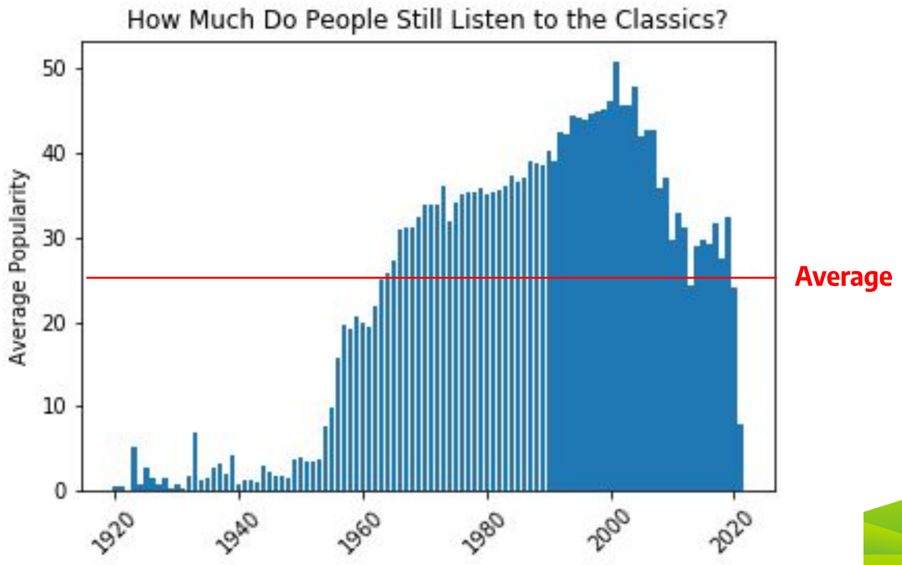
Summary of Popular Genre Characteristics

- All genres chosen have low speechiness and low instrumentalness
- Live recorded music doesn't appear to be popular option
- Major differentiation amongst explicit and accousticness categories for at least two artists
- Some delineation between valence and energy, but alignment amongst the genres
- Song characteristics: evolved or manufactured?

Analysis: How much do people still listen to the classics?

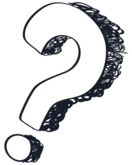
*As of March 2018, Spotify's user base was dominated by Millennials, with 29 percent of its users aged 25 to 34 and 26 percent aged between 18 and 24 years old.



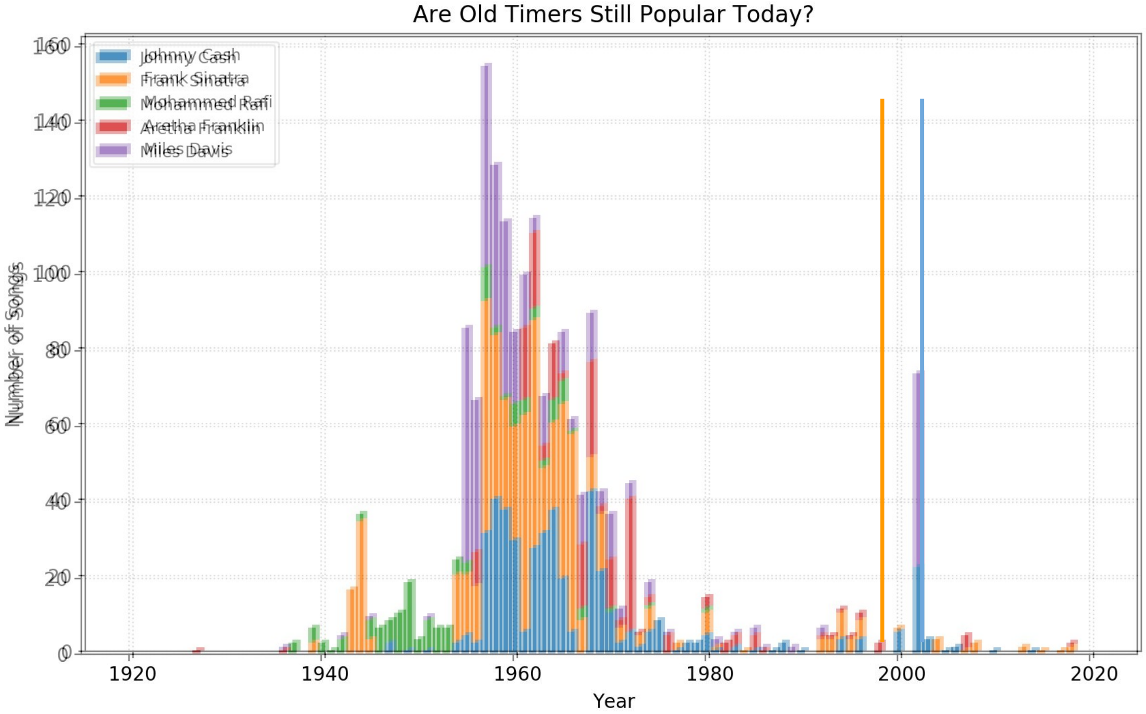


Source: [Statista]

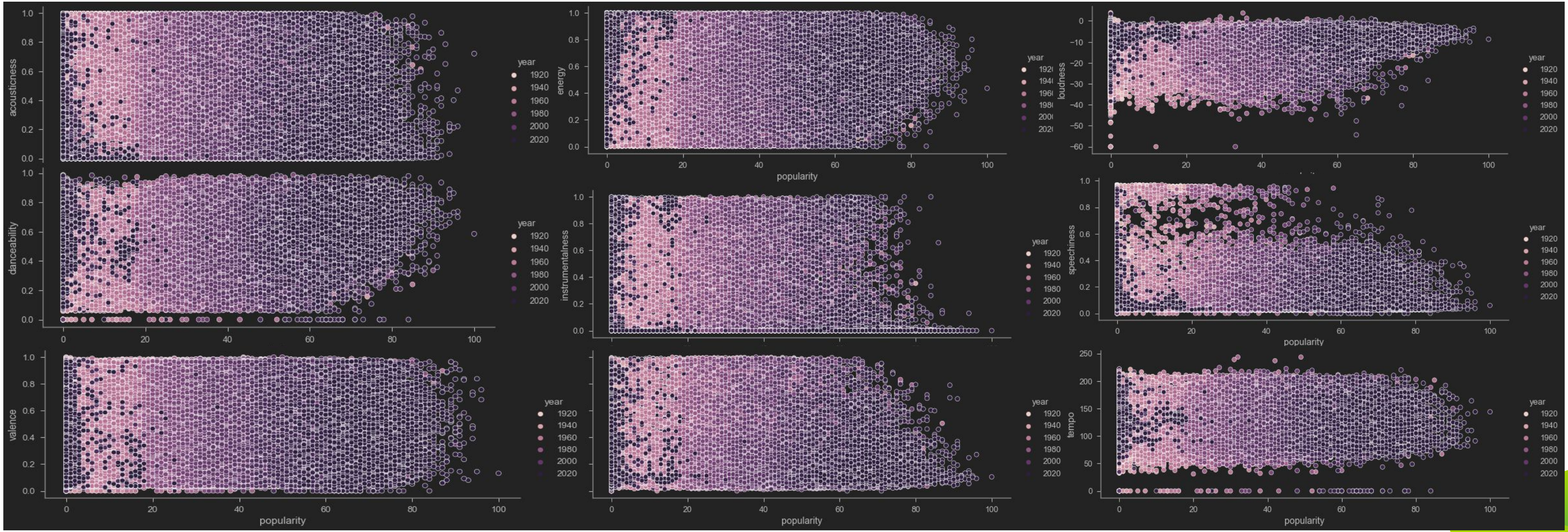
Analysis: Who is the music GOAT? (Who are the most timeless artists?)



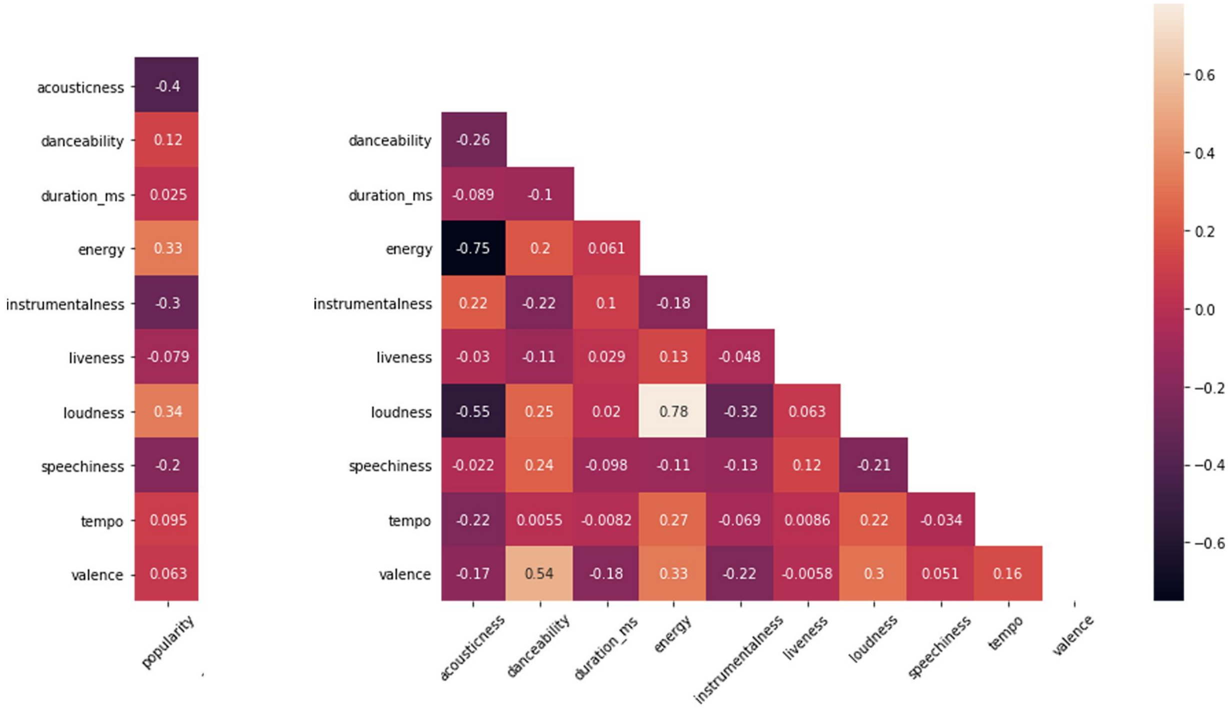
	artists	time span of songs	song count
0	Johnny Cash	43	456
1	Frank Sinatra	40	621
2	Mohammed Rafi	31	143
3	Aretha Franklin	30	201
4	Miles Davis	30	427



Scatter Plots - Popularity vs. Audio Features



Are there interdependent relationships between different audio features ?



- Acousticness vs Popularity
- Loudness and Popularity/Energy
- Acousticness vs. Energy/Loudness
- Loudness vs. Energy
- Danceability vs. Valence

Key Findings/Conclusions

1. Music features appear to have evolved over time and began stabilizing around 1960
2. Major genres have similar music characteristics like tempo, but some divergence noted for explicit and accousticness variables
3. No single music features can predict a song's popularity; some covariance found amongst the features
4. Multiple linear or nonlinear regression an interesting next step towards predicting music popularity but not explored.
5. Songs from the late 1990s to the early 2000s are the hottest on Spotify.
6. Music's legends continue to be popular today

Spotify

Thanks!

Any questions?

Appendix - Data Limitation

1. Lists of collaborators such as (['JAY-Z', 'The Notorious B.I.G.']) makes analysis tricky
2. Genres were loaded as lists of values and are sorted alphabetically, making it hard to determine popular genre trends
3. Popularity is a subjective measure. It's based on current episodes of listening by contemporary listeners
4. Data collection methodology not exactly clear, especially when accounting for noisy data in early 20th century

Yamac Eren Ay Dataset Creator • 8 months ago • Options • Report • Reply 1

Mean values: acousticness, danceability, energy, valence, instrumentalness, speechiness, tempo, loudness, duration_ms, liveness, popularity
Mode values: key, mode
Count values: count (total number of tracks)

max. 2000 songs are selected from each year but i am not really sure how 2000 of all songs are selected (chances are either "randomly" or "based on popularity", which in my opinion is "based on popularity").