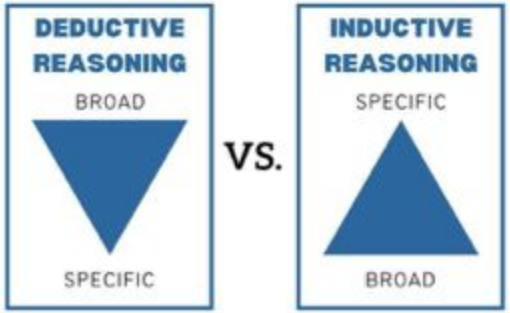
Cracking the Spotify Popularity index

Martina Paez Berru, Yuwei Ding, Lydia Ying, Andrea Cuadros Vera, Yongwen Yuan

Problem Statement

Why we picked this project?

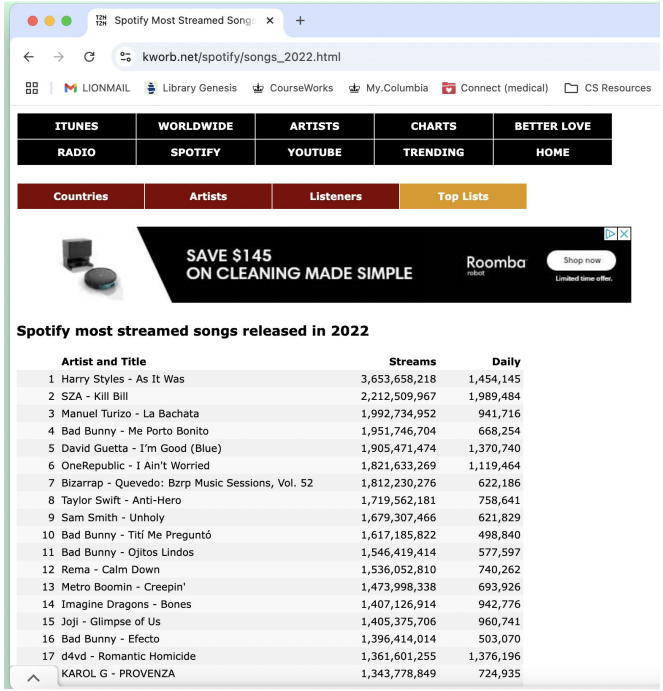


General Overview

- Deductive reasoning: started with an idea to predict an algorithm
- Inductive reasoning: After understanding popularity, our research dived into similarities across years 2018-2023 (lyrics)
- Analyzed how trends change over time in popularity dynamics
- Led to more questions
  - What characteristics drive a song's popularity score? Genre? Audio features? The artist?
  - Can the public sentiment on a song's performance (love it or hate it) predict a song's popularity?
  - What similarities can we spot in popular songs across years? What topics are recurrent?

Part 1 Regression Analysis and Neural Network

Data Collection & Preprocessing

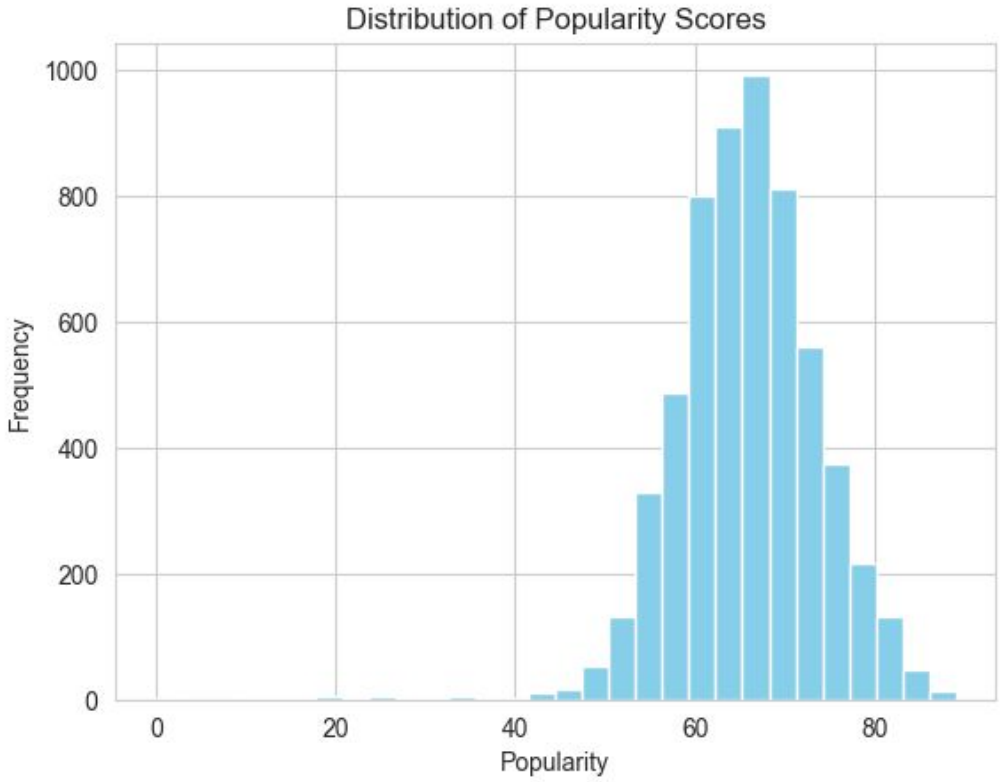


	Artist and Title	Streams	Daily
1	Harry Styles - As It Was	3,653,658,218	1,454,145
2	SZA - Kill Bill	2,212,509,967	1,989,484
3	Manuel Turizo - La Bachata	1,992,734,952	941,716
4	Bad Bunny - Me Porto Bonito	1,951,746,704	668,254
5	David Guetta - I'm Good (Blue)	1,905,471,474	1,370,740
6	OneRepublic - I Ain't Worried	1,821,633,269	1,119,464
7	Bizarrap - Quevedo: Bzrp Music Sessions, Vol. 52	1,812,230,276	622,186
8	Taylor Swift - Anti-Hero	1,719,562,181	758,641
9	Sam Smith - Unholy	1,679,307,466	621,829
10	Bad Bunny - Tití Me Preguntó	1,617,185,822	498,840
11	Bad Bunny - Ojitos Lindos	1,546,419,414	577,597
12	Rema - Calm Down	1,536,052,810	740,262
13	Metro Boomin - Creepin'	1,473,998,338	693,926
14	Imagine Dragons - Bones	1,407,126,914	942,776
15	Joji - Glimpse of Us	1,405,375,706	960,741
16	Bad Bunny - Efecto	1,396,414,014	503,070
17	d4vd - Romantic Homicide	1,361,601,255	1,376,196
	KAROL G - PROVENZA	1,343,778,849	724,935

BeautifulSoup

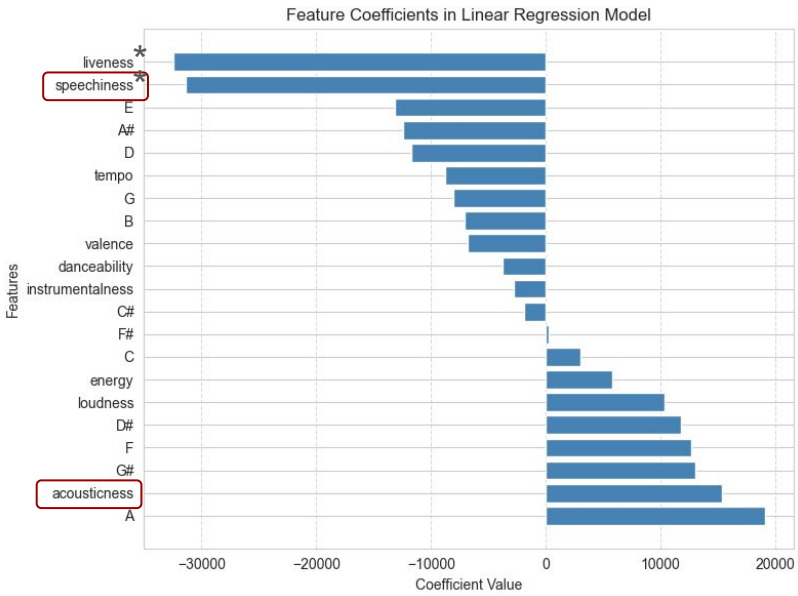
Spotify for Developers

Regressions on Kworb Data



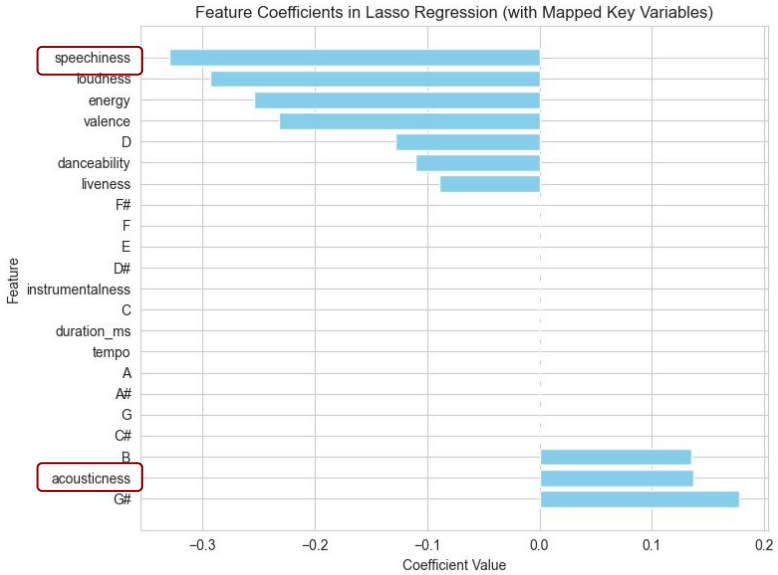
Regressions on Kworb Data

Linear Regression



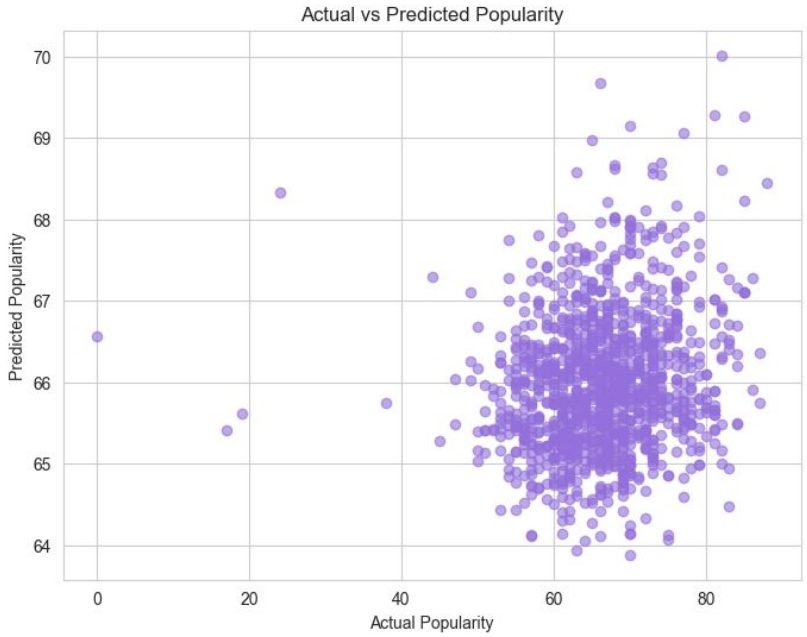
MSE: 46342350376.14
R^2 Score: 0.0448
Prob(F-statistics): 0.0897

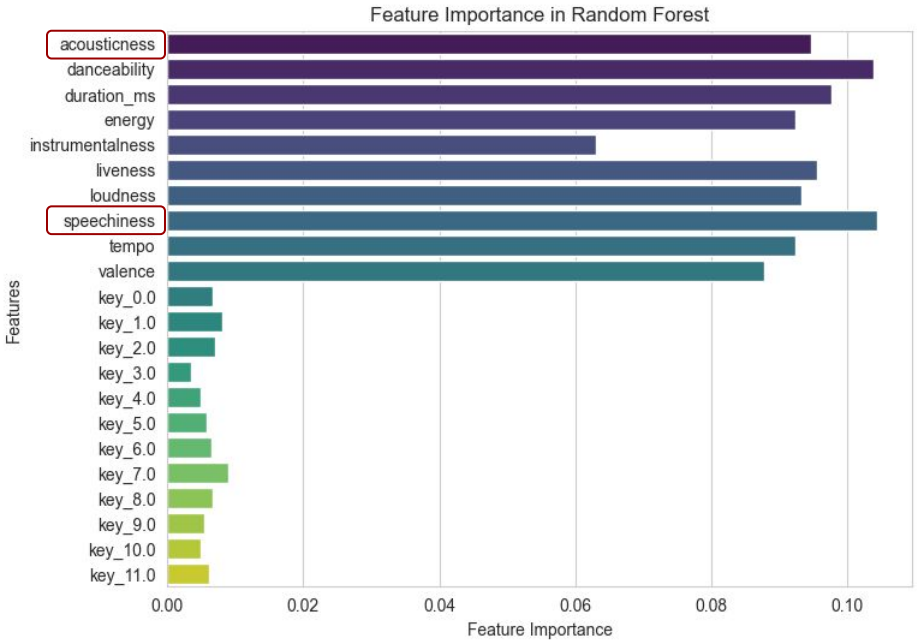
Lasso Regression



Best alpha: 0.1
Mean Squared Error: 62.28
R^2 Score: 0.02

Regressions on Kworb Data – Random Forest





Mean Squared Error: 55.98
R^2 Score: 0.12

Results with Spotify audio features were not promising…

How about directly analyzing the audio instead?

Extracting Features from Audio File

Get music preview

Music previews were scraped from Spotify API, 30s for each song. 2022-2023 songs selected.

Extract musical features from Yamnet

Including 521 features, each representing an audio elements. E.g. roll, R&B, rattle, drum roll…

Random forest: decide x-variables

Based on the result of random forest, choose most important features as independent variables

Neural network: predict popularity

Exclude outliers

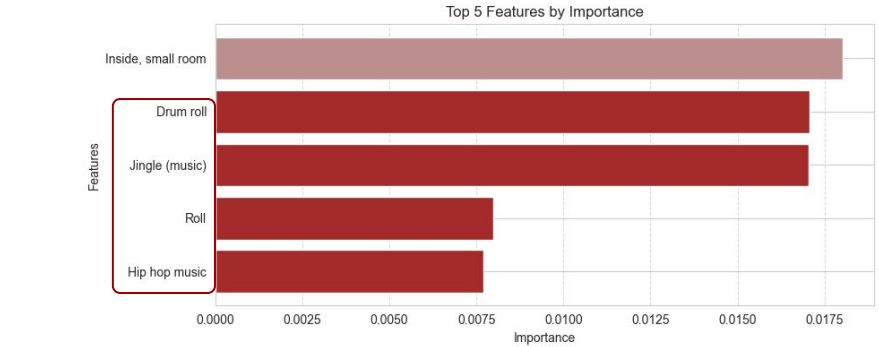
Input:
Features
Time variable

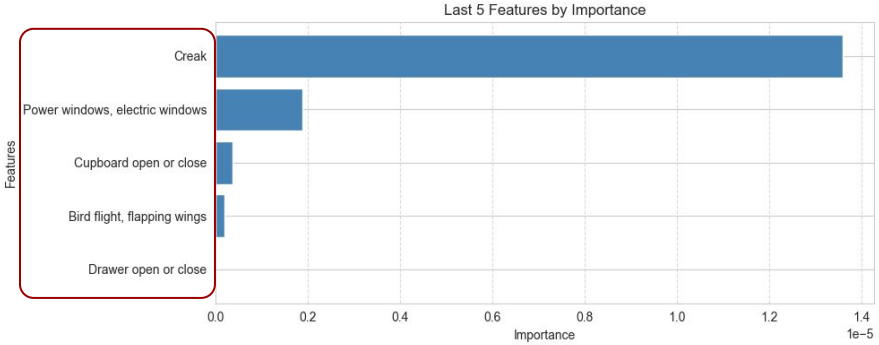
Neural Network

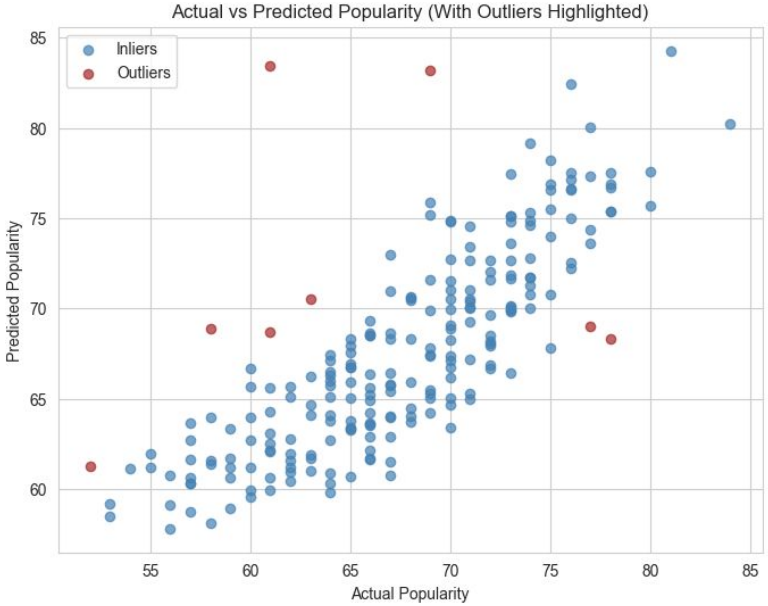
Output:
Popularity Score

*** Add time variable to to capture the temporal dynamics of the song's popularity.*

Extracting Features from Audio File



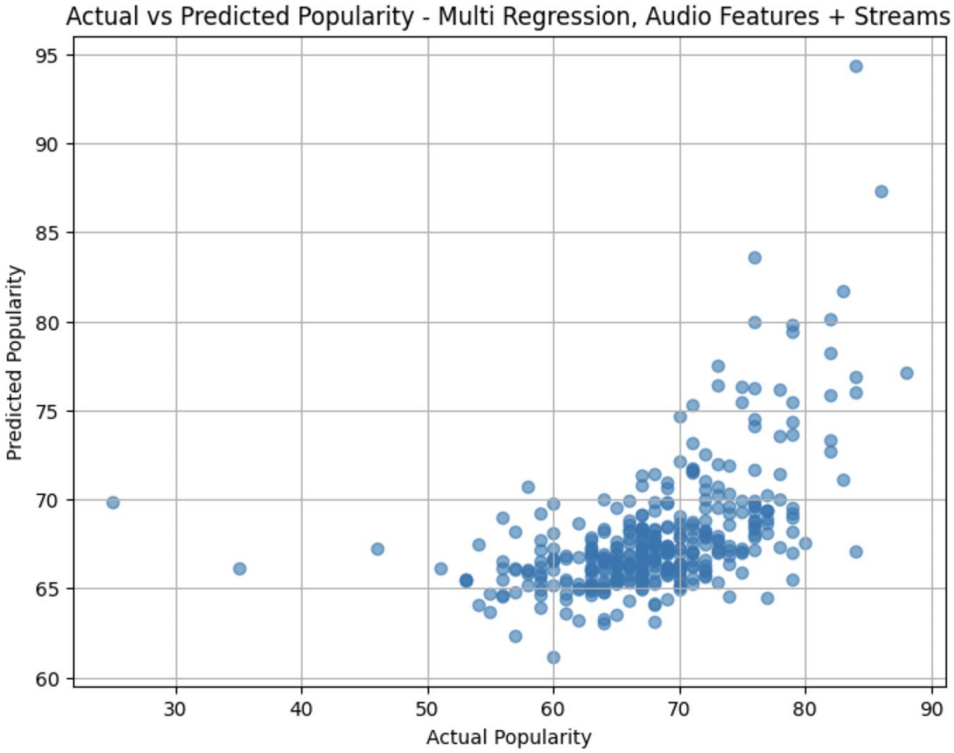




Before excluding outliers R^2 Score: ~0.5
After excluding outliers R^2 Score: ~0.7

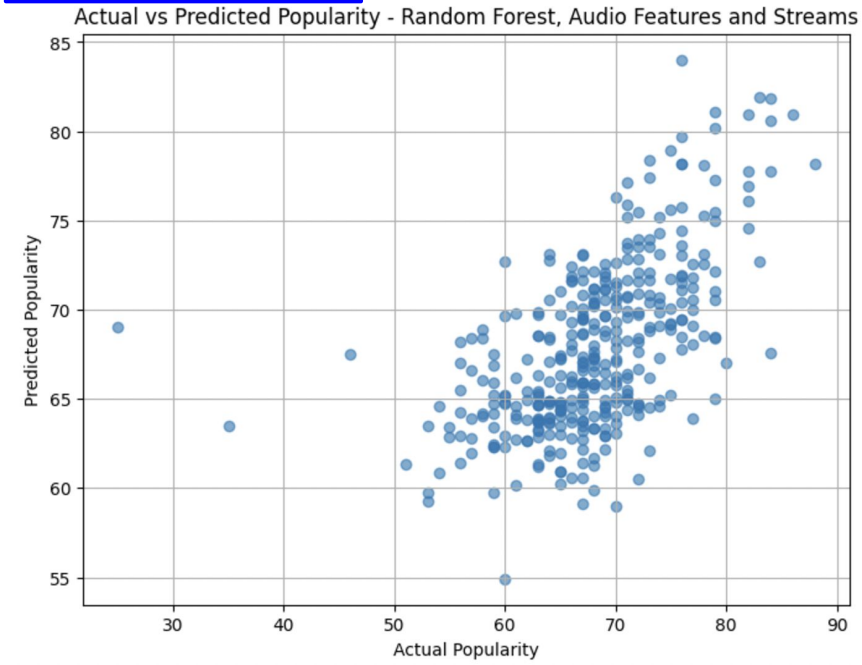
Regressions on Kworb Data with Streams

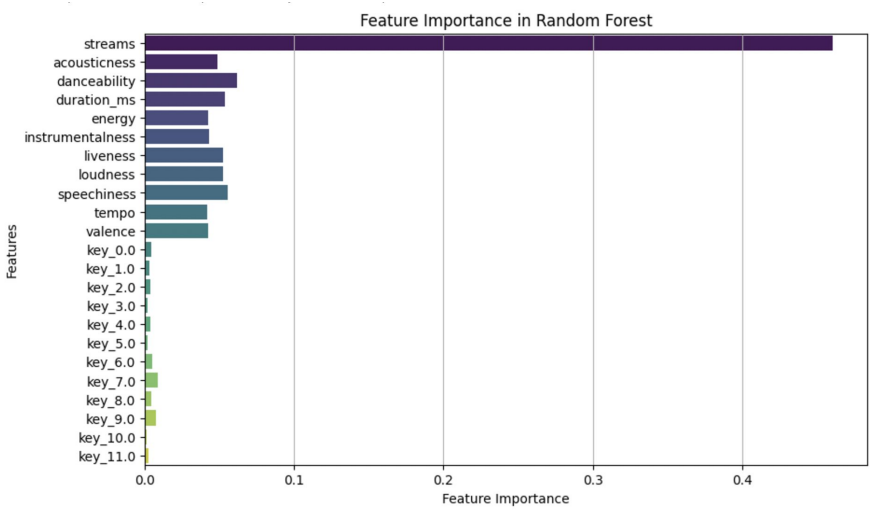
```
(MSE): 36.51732080212758
(R^2): 0.2908992062186594
```



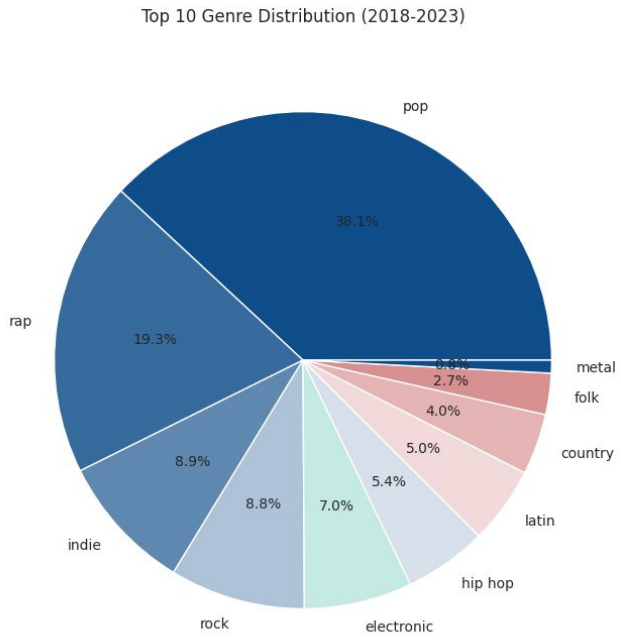
Regressions on Kworb Data - Random Forest with Streams

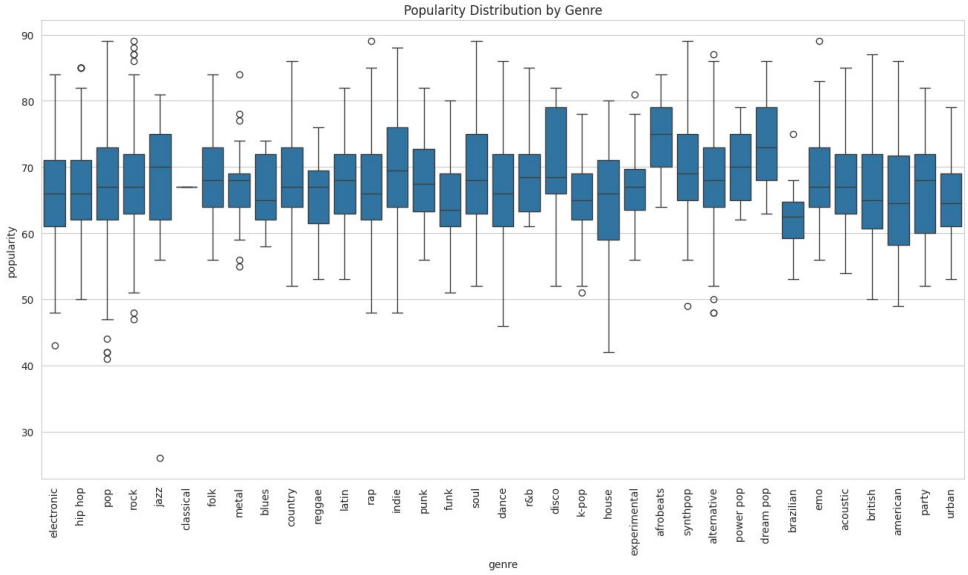
```
Mean Squared Error (MSE): 34.24
R² Score: 0.34
```





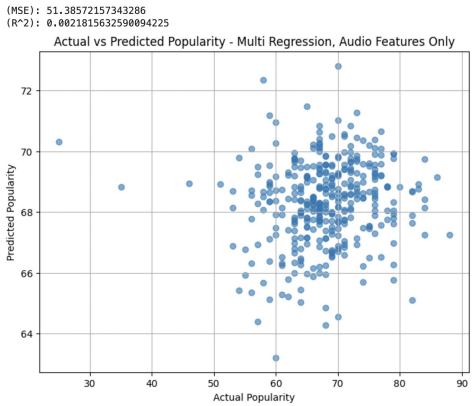
Impact of Genre: Trends by Year

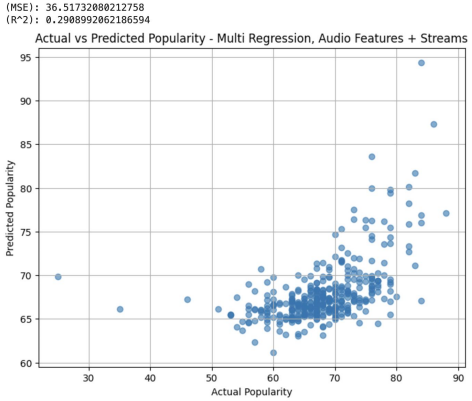




Impact of Genre: Regressions on Kworb Data

Without Genre:

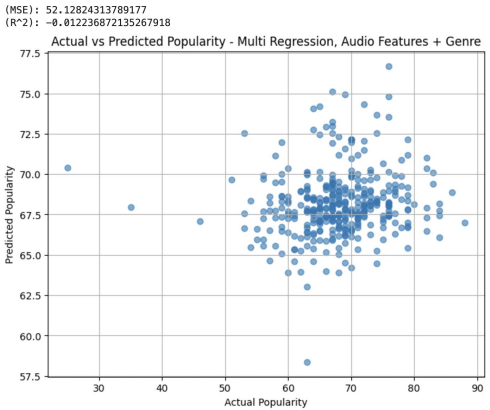


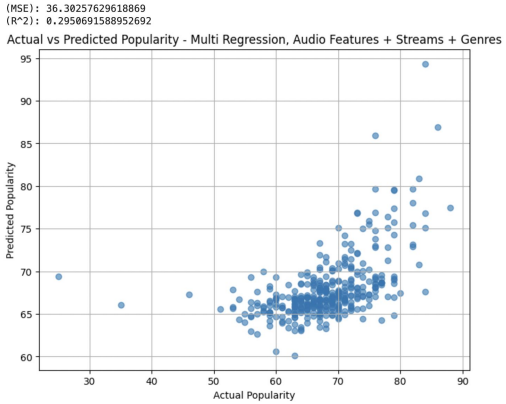


MSE: +0.7425
R^2: -0.0144

MSE: -0.2147
R^2: +0.0042

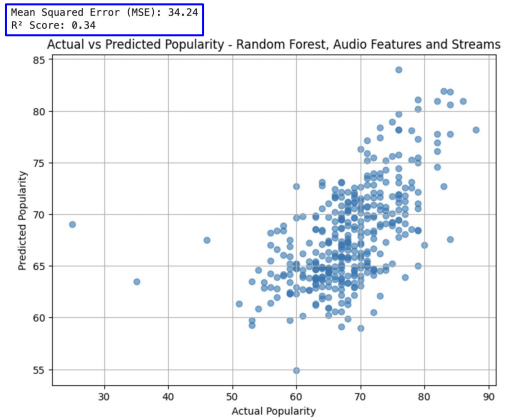
With Genre:

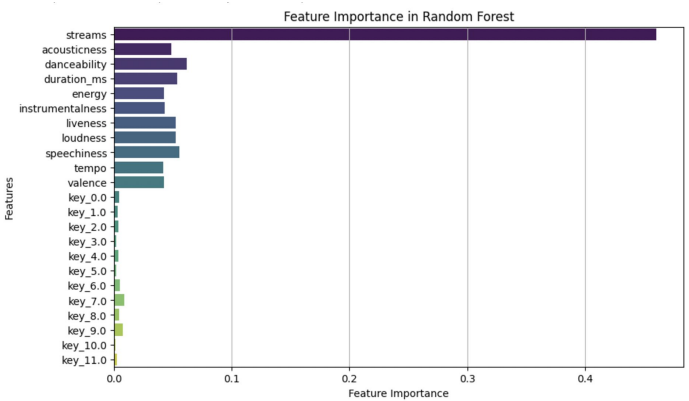




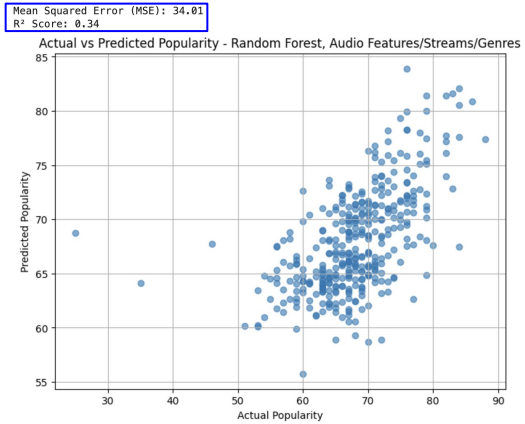
Impact of Genre: Regressions on Kworb Data (cont.)

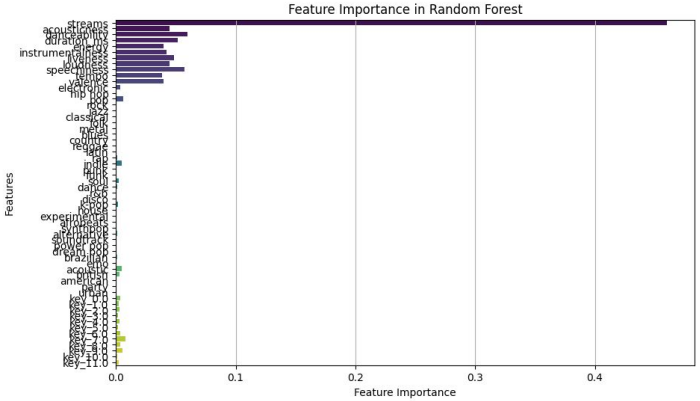
Without Genre:





With Genre:





Part 2: LDA & LSI

Data Collection

- Used data set for popular song collection from Part 1
- Lyrics retrieved using Lyrics.ovh and Genius APIs for top 25 songs each year (2018–2023)
- Music Board- YouTube APIs to retrieve top 40-50 comments of music videos from the artists' official Youtube channel

LYRICS.OVH
ONLY THE LYRICS

YouTube
API

Question to tackle

We want to study texts (song lyrics) to discuss:

- Similarity of songs across year
- Sentiment trend of songs across year

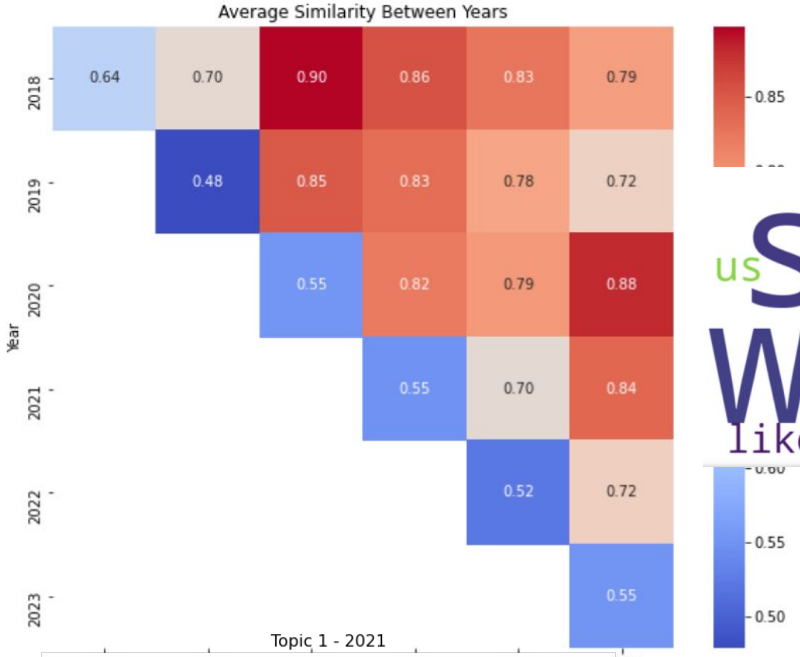
Apply the same framework to two case studies, alongside listener comment

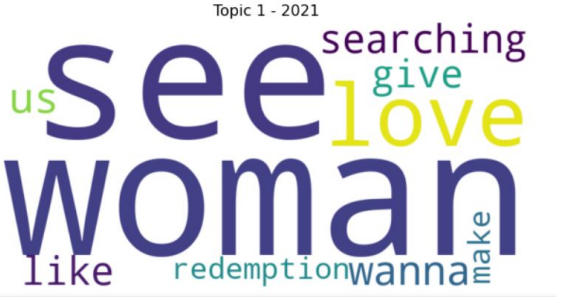
- Similarity of Taylor Swift, Billie Eilish song across year
- User sentiment of the two artists
    - Comparing with song qualities such as lyrics sentiment, and popularity score

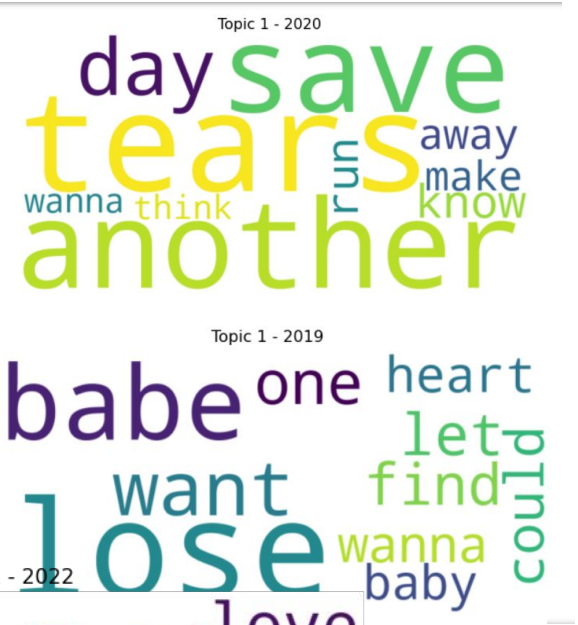
General Overview of Song Sentiment 2018-2023

Year-to-Year Similarity:

- Compared LSI topics of one year with another using cosine similarity → how similar the lyrical themes are between the two years.

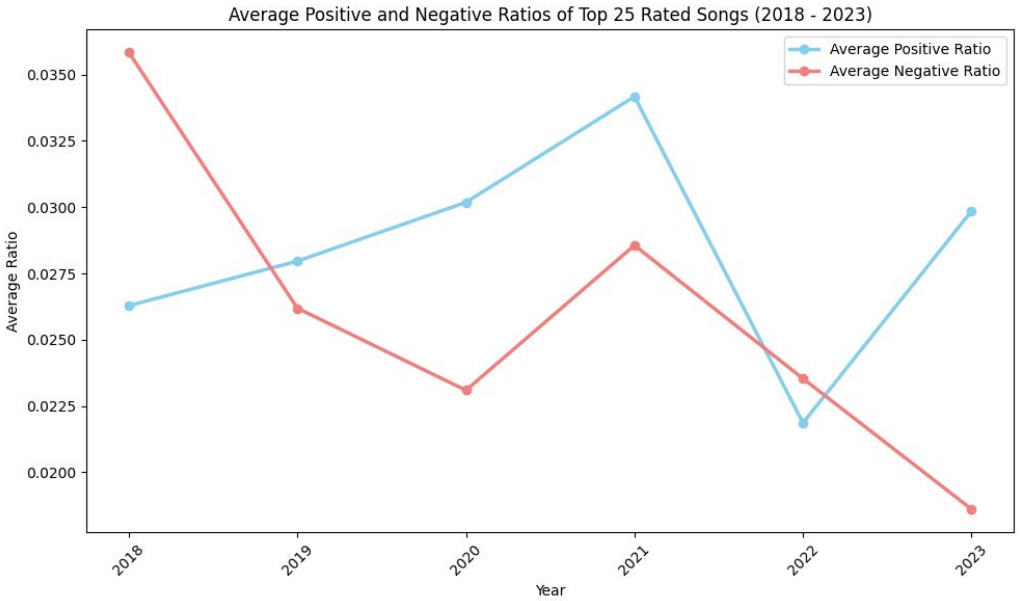


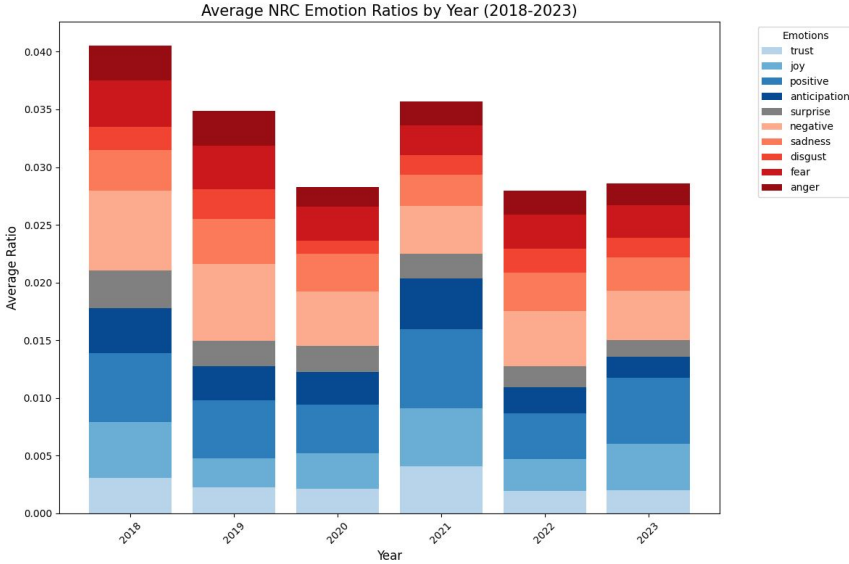




General Overview of Song Sentiment 2018-2023

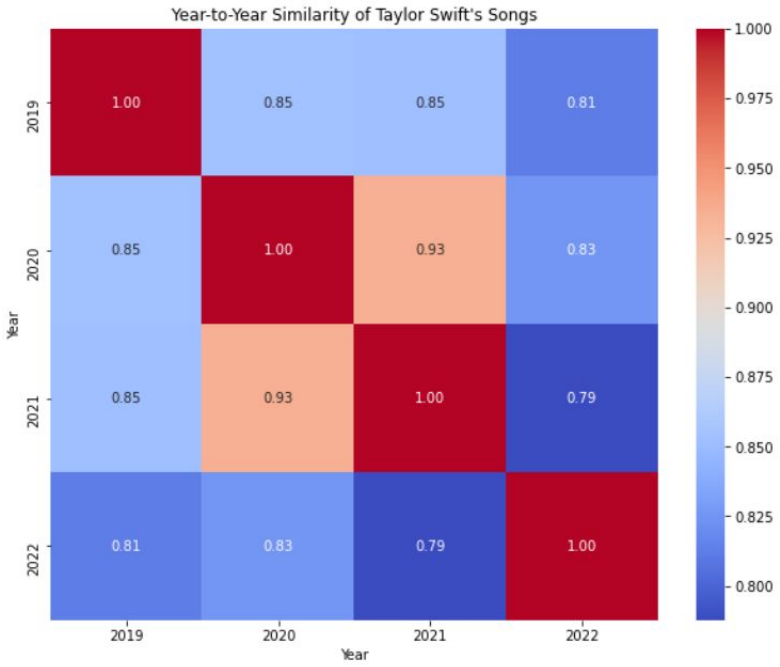
- **Naive pos/neg ratios by year**
- **NRC emotion ratios by year**

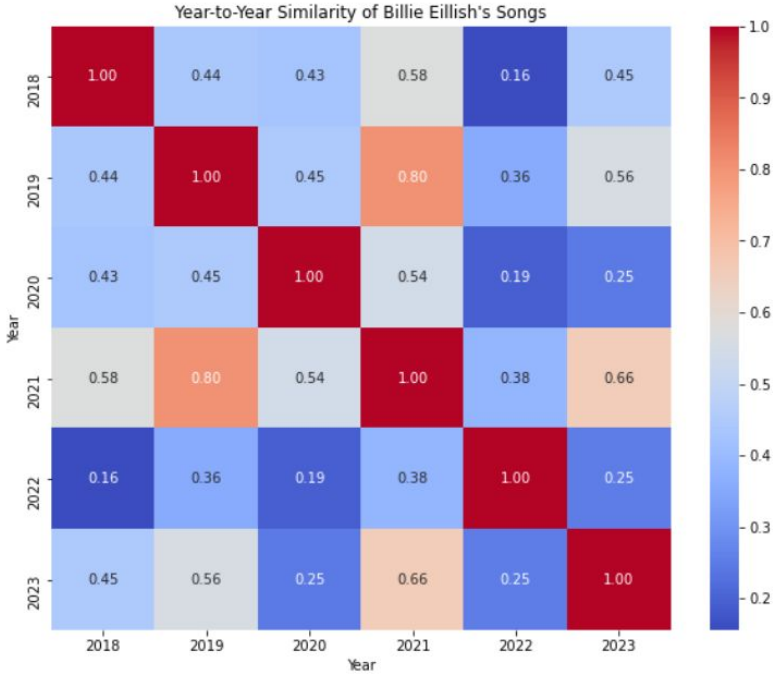




Case Studies Taylor Swift & Billie Eilish

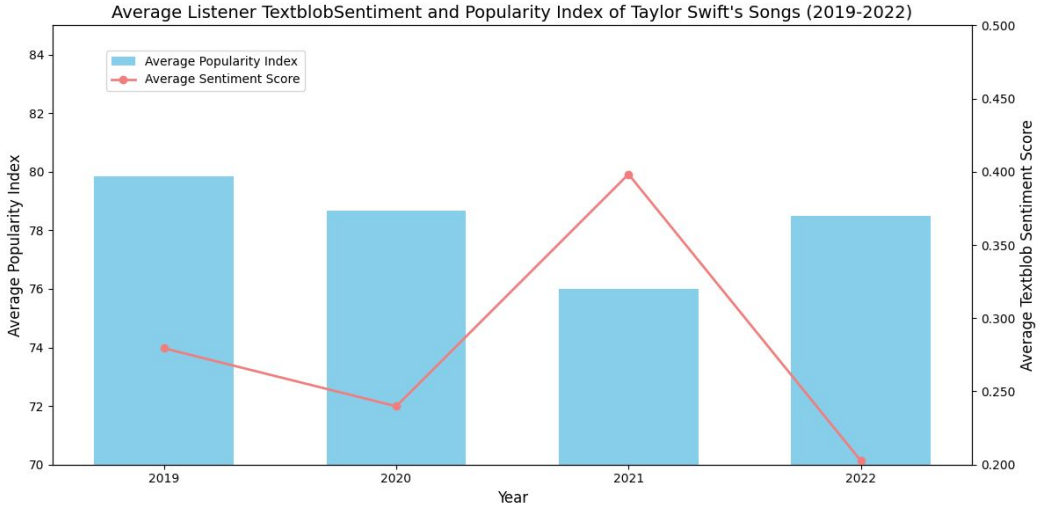
Similarity Score for the 2 artists

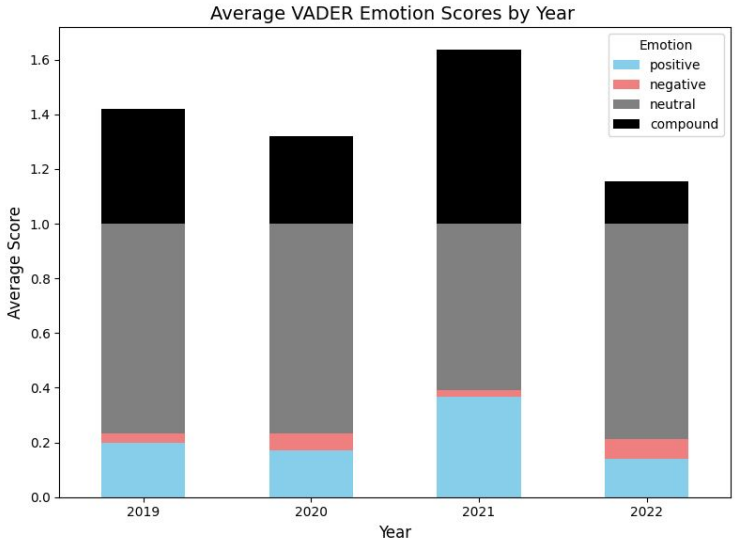




Sentiment Analysis of Youtube Listener Comment

- Textblob
- VADER





Key Insights & Conclusions

- Part 1: Genre and Spotify features do not play a large part in predicting a song's popularity. However, Stream significantly does.
- Part 2: Across the years, popular songs don't show a big shift in lyrical trends

This research will be published on our Github for any music enthusiast to continue collecting data to perfect the now not so secretive Spotify's popularity index score and understand how despite the evolving trends in music production and cultural shifts over the years, the core ways in which humans communicate and express emotions through lyrics remain remarkably consistent.

Sentiment analysis reveals that, at their heart, the same themes of love, struggle, joy, and longing continue to resonate even now, demonstrating that the fundamental emotions conveyed in music transcend temporal boundaries.