Spotify Data Analysis

Gadge Uday

Final Project for Statistical Programming in R

Contents

Aim

Dataset

Procedure

Results

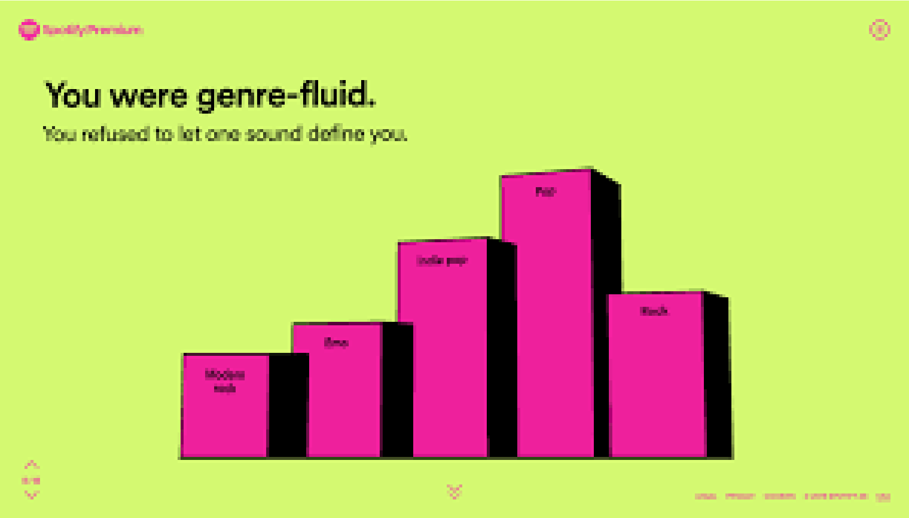
- Genre Analysis
- Recommendations

Scope

Aim

The project has two parts:

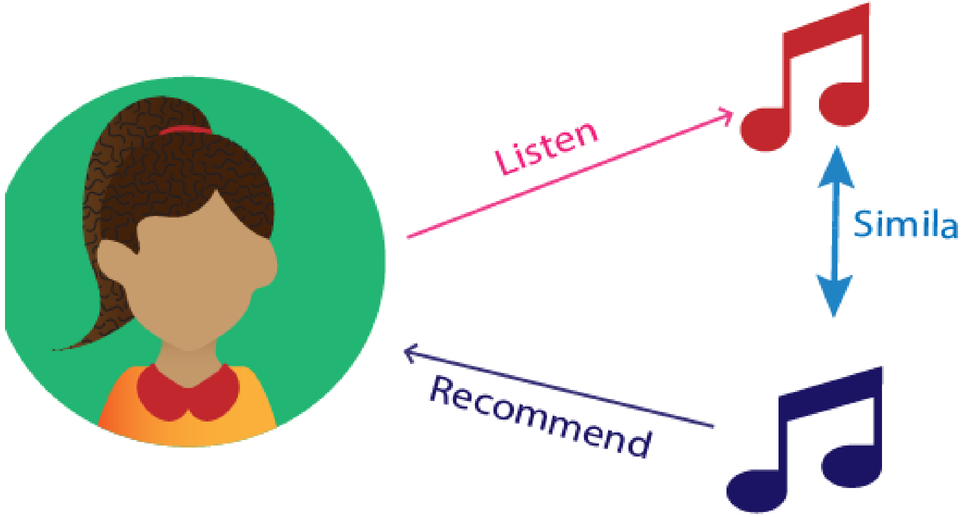
- Genre Analysis
- Building a Recommendation model



Spotify Premium

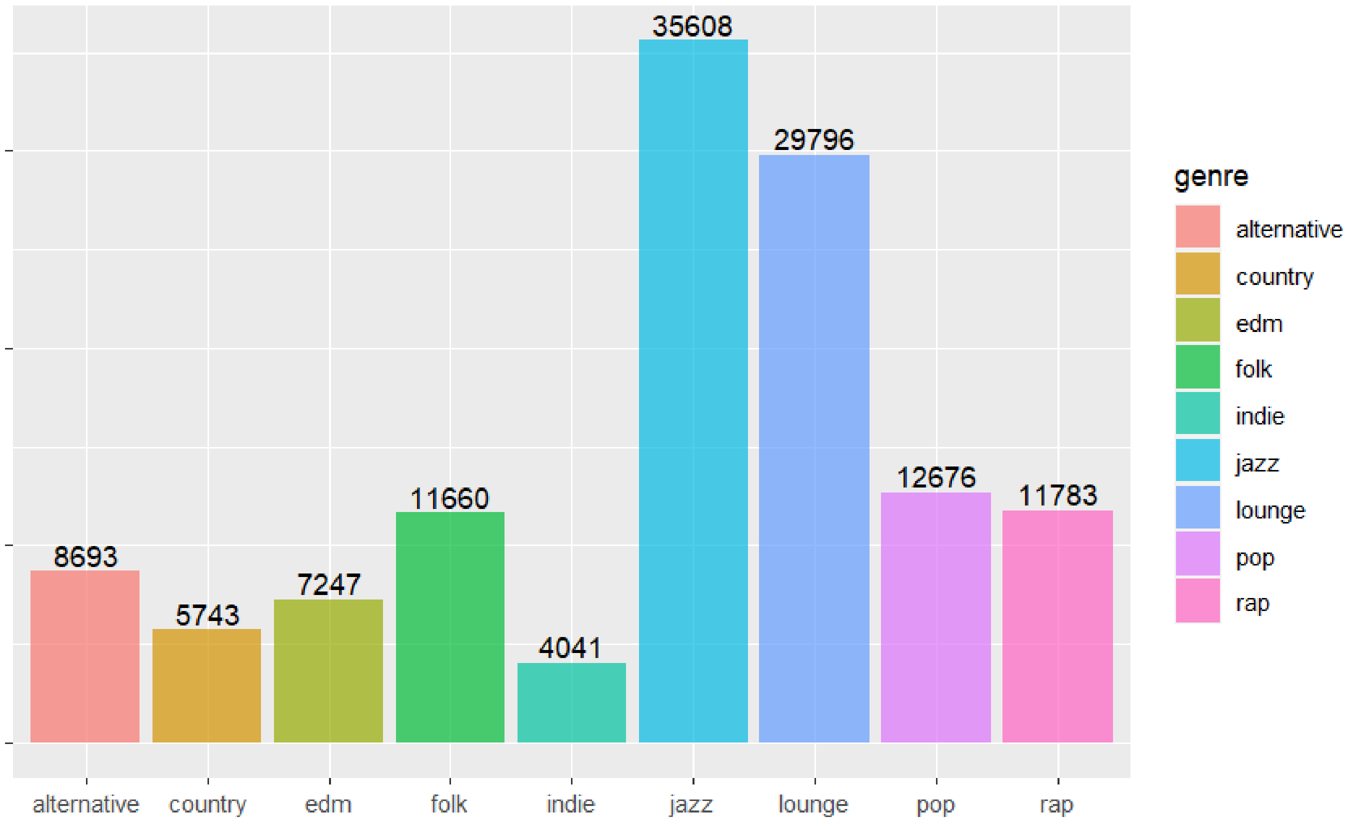
You were genre-fluid.

You refused to let one sound define you.



Dataset

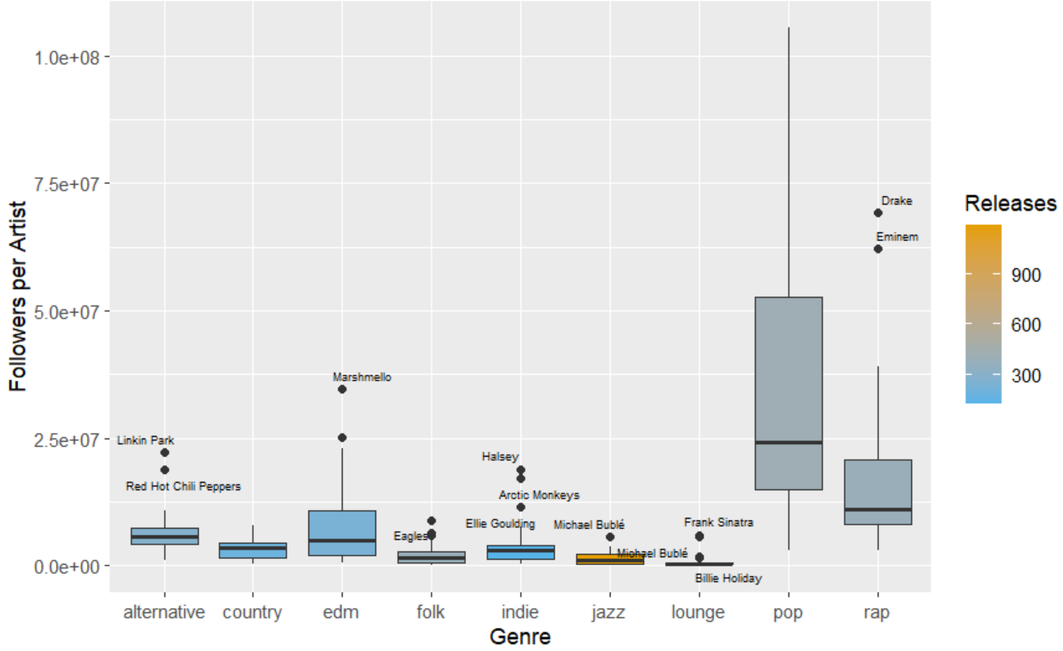
Genre Count of the dataset

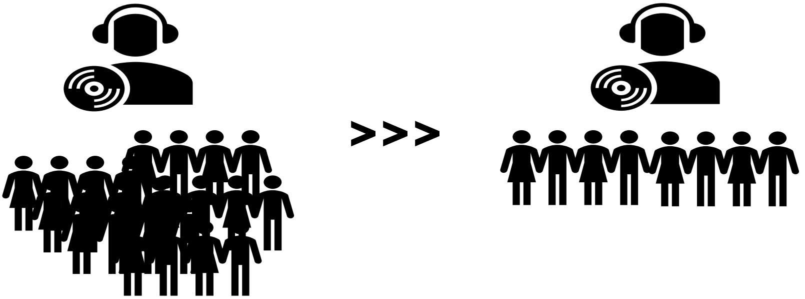


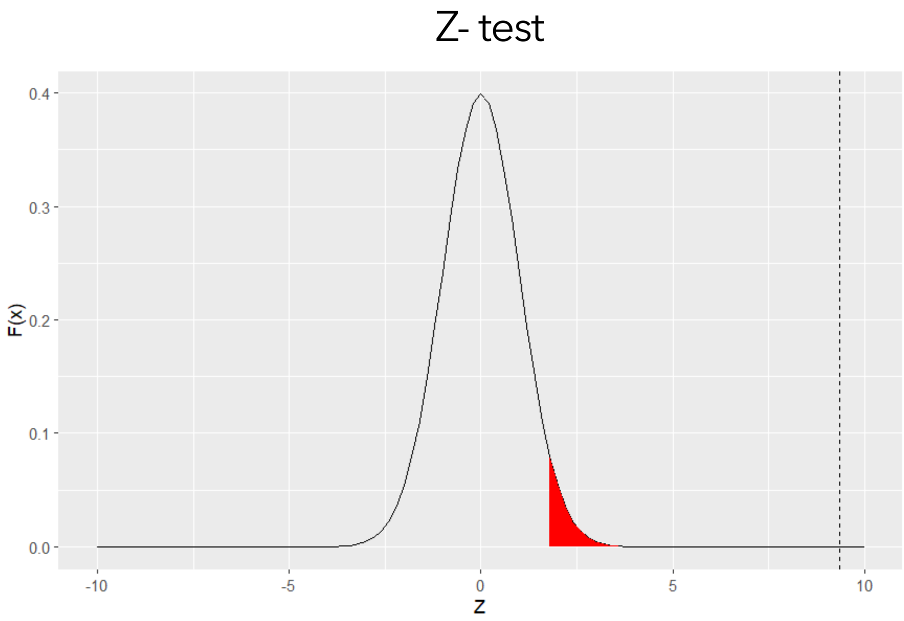
Genre Analysis

Hypothesis testing

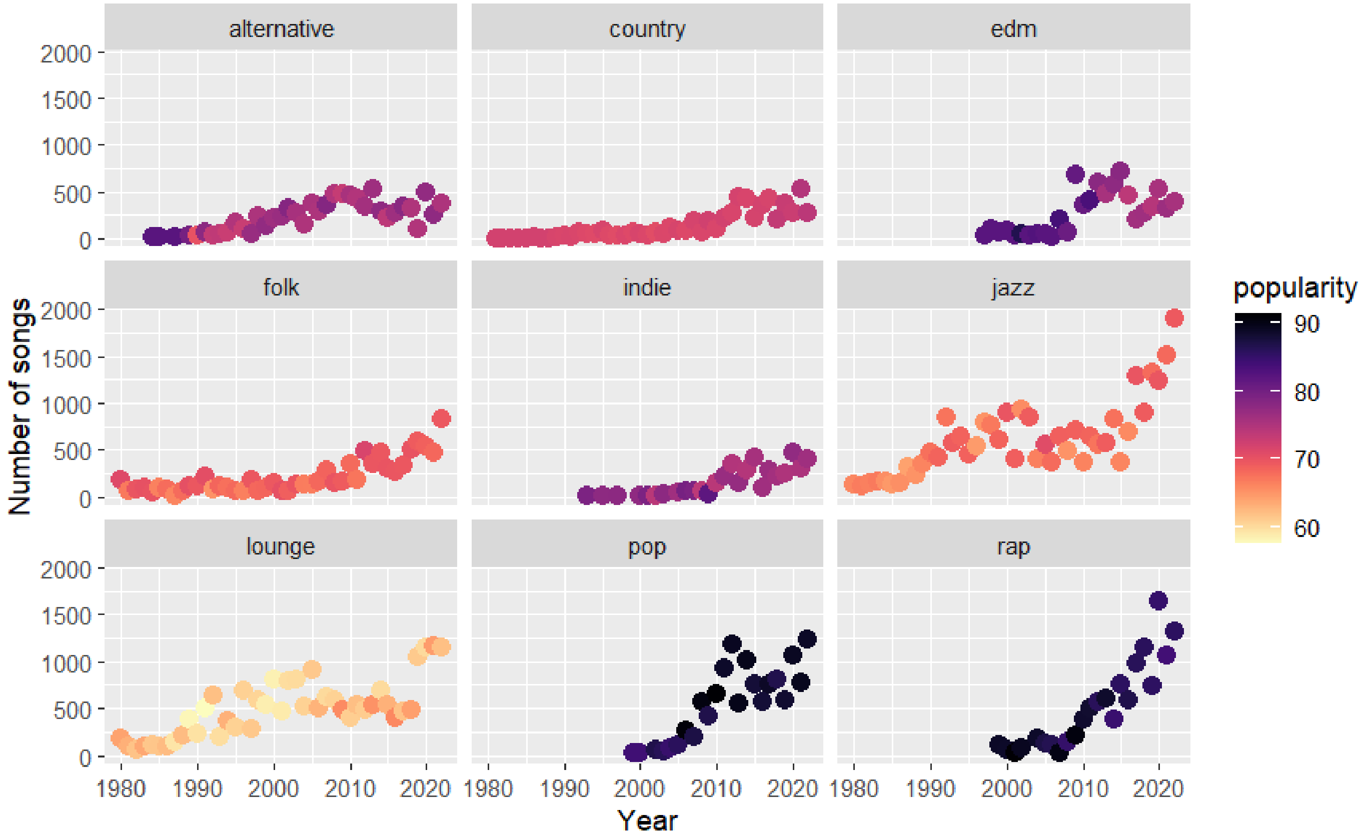
Do pop artists have more followers than other genre artists?







Popularity over time

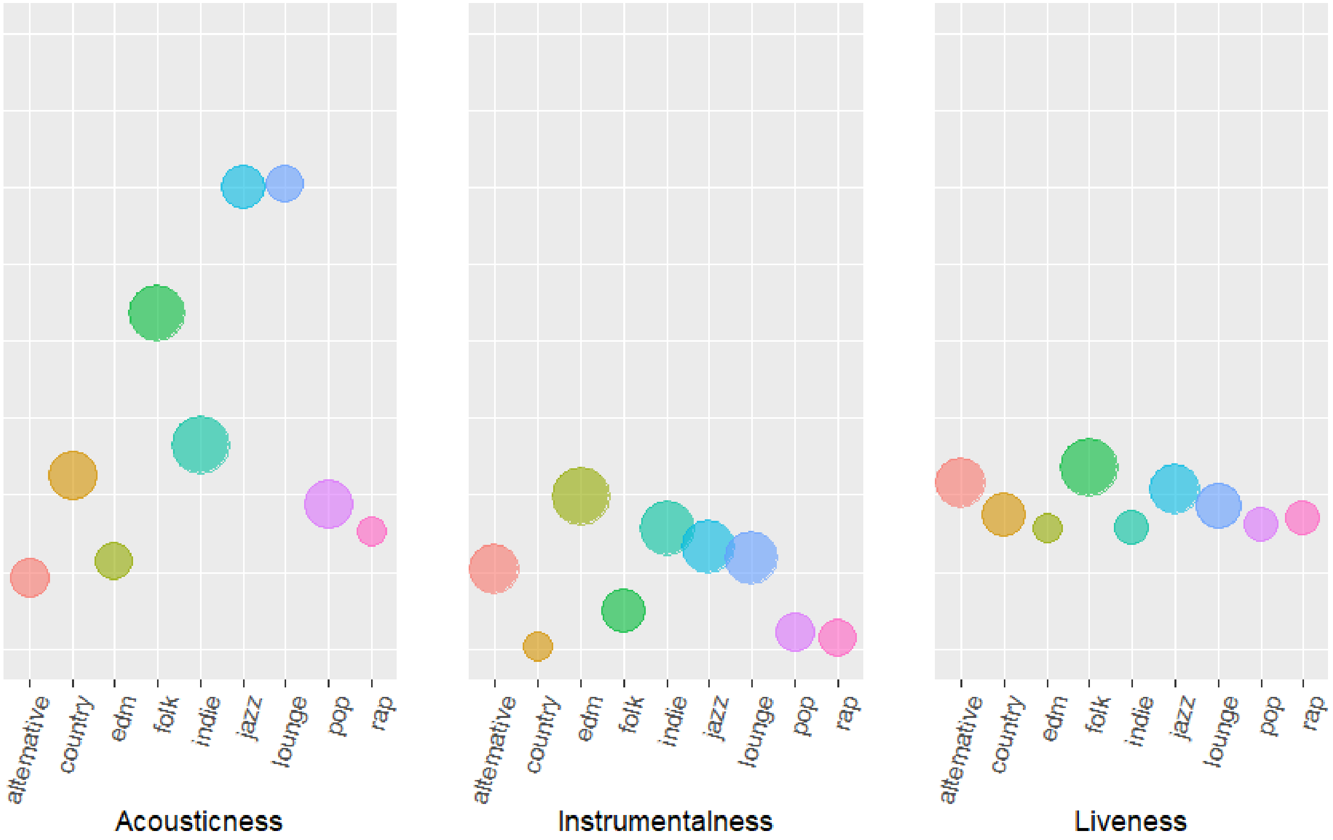


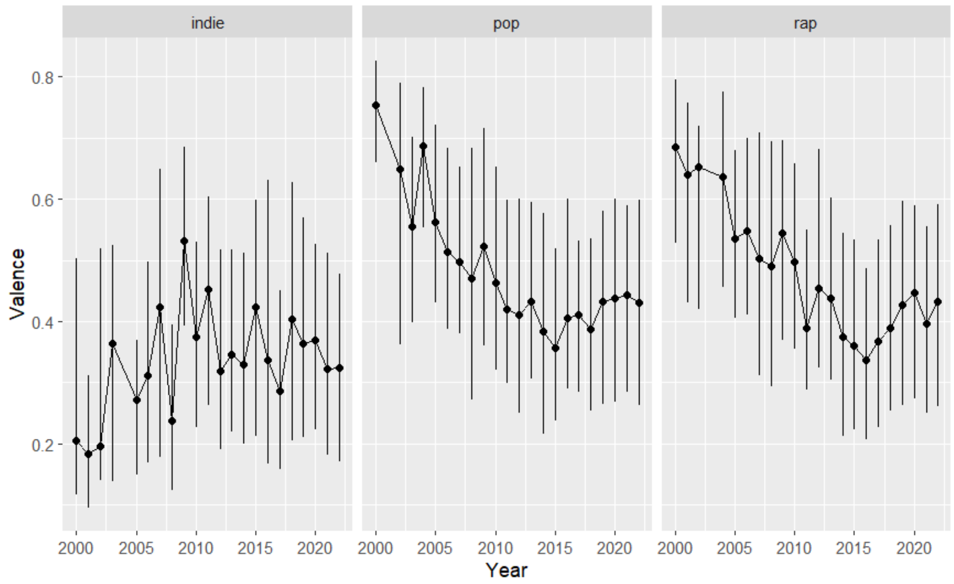
- Pop and rap music are significantly more popular than other genres.
- Lounge music had a lull in the 2000s

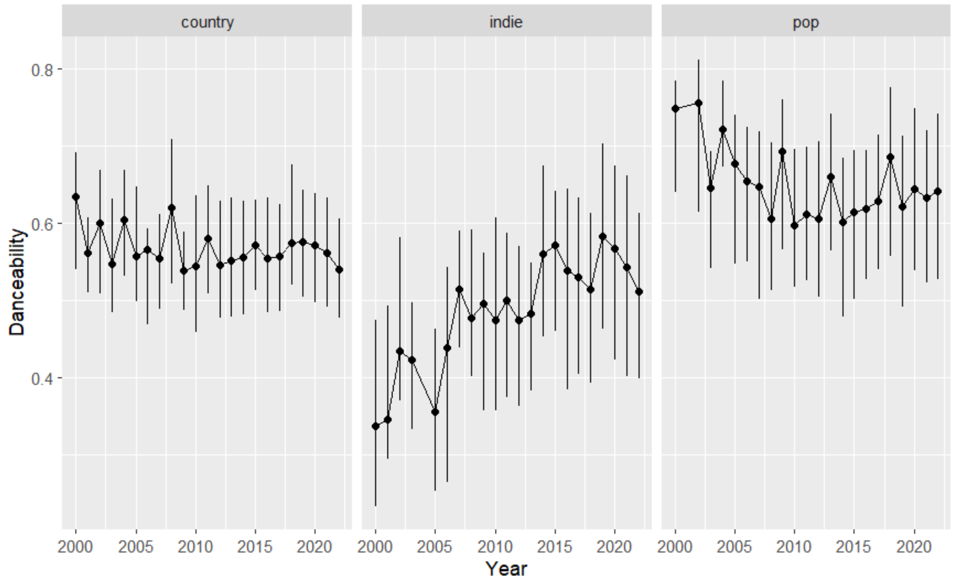
Key Characteristics of genres

- Jazz and lounge genres have a high degree of similarity

- Country music has a low instrumentality and also is less variant in this aspect.



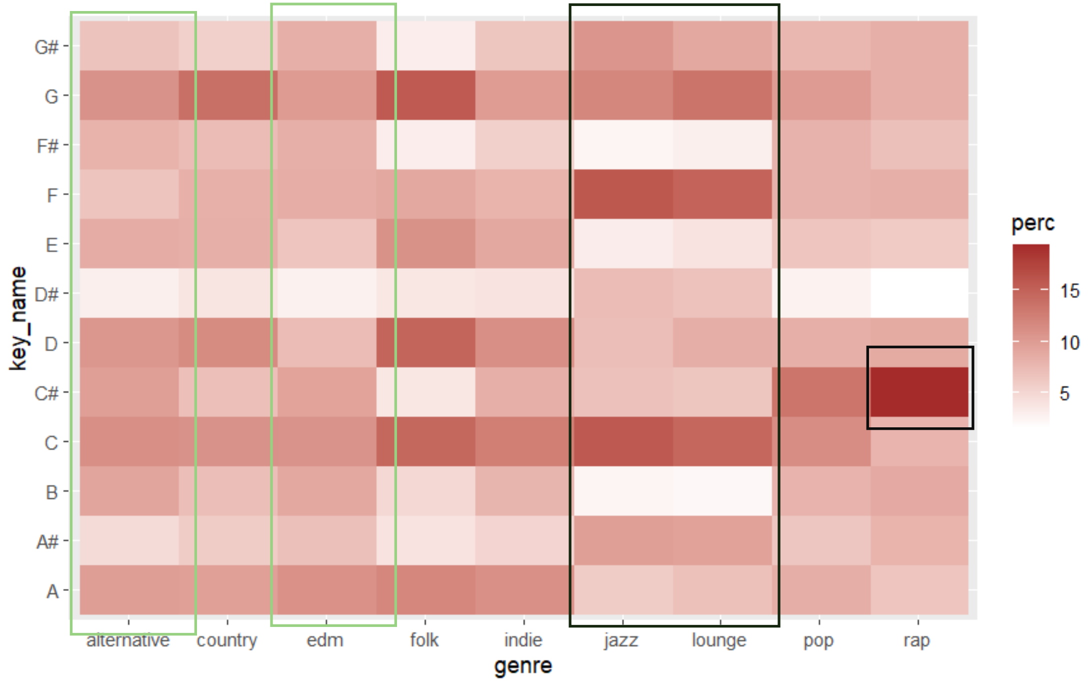


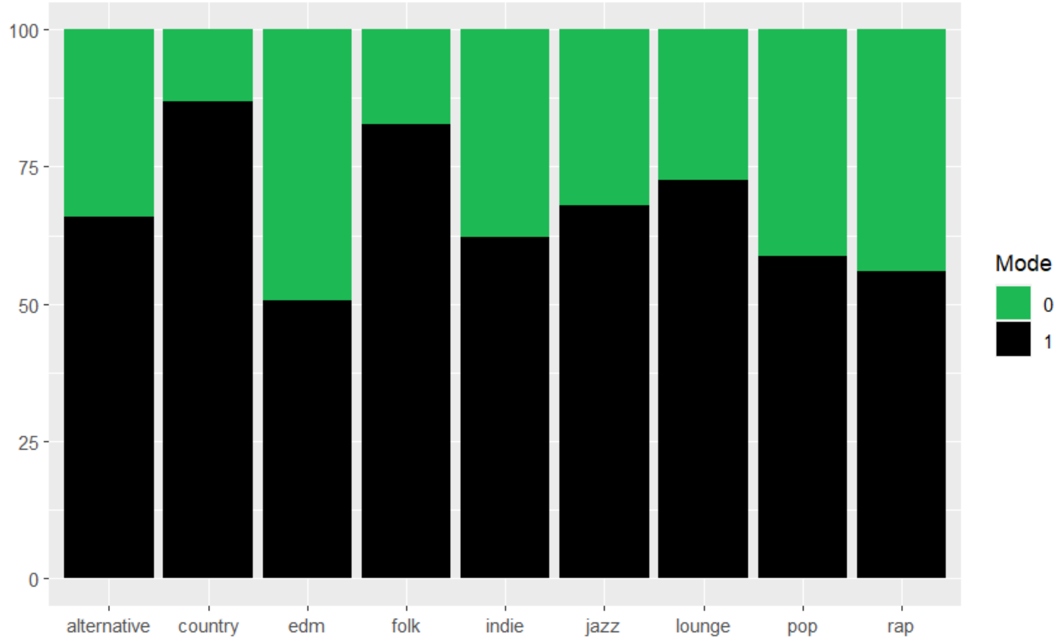


Valence and danceability

- Valence is a measure of positivity in a song (through words)
- Decline in both the valence and danceability in pop music
  - More heartbreak music?
  - Irony in music?

Key in which it's played

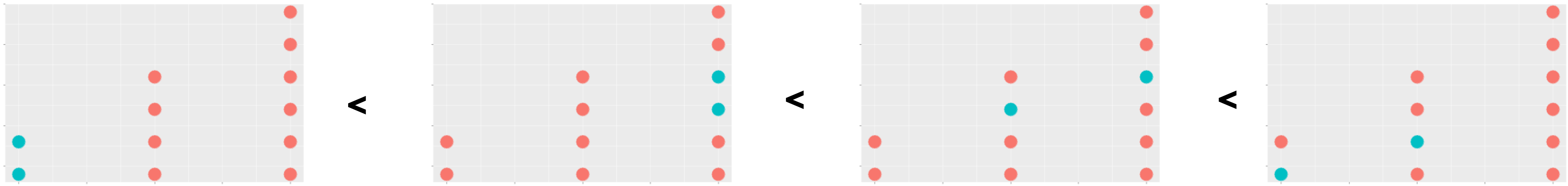




- The keys also suggest a close similarity between jazz and lounge.
- It also suggests a close similarity between alternative and edm.

Distance metric

$$Dis(X,Y) = \sqrt{\underbrace{\sum_{num}(x_i - y_i)^2}_{\text{Euclidean}} + \sum_{cat}\delta_{i,j}P_iP_j + (1-\delta_{i,j})(1-P_iP_j)}$$

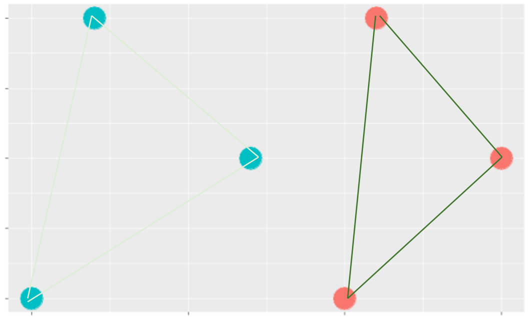


This distance is still biased towards categorical features. In cases with a high number of categories with low probabilities, it does tend to be 0 and 1.

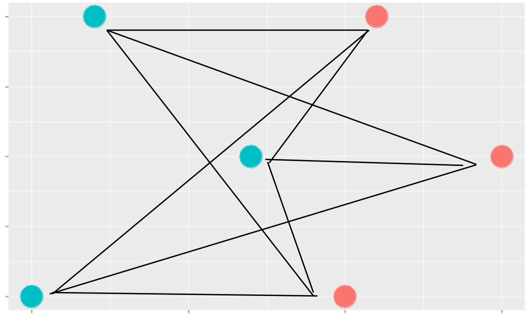
However, it does the job of scaling different combinations correctly.

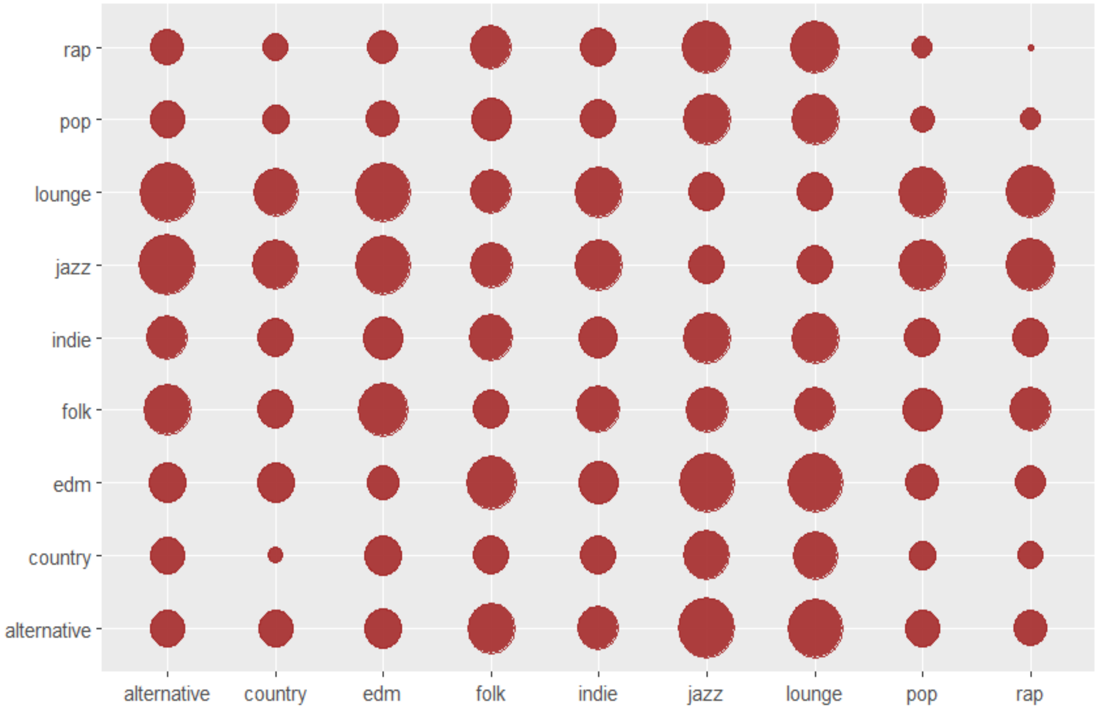
Genre Similarity

Intra genre distance



Inter genre distance



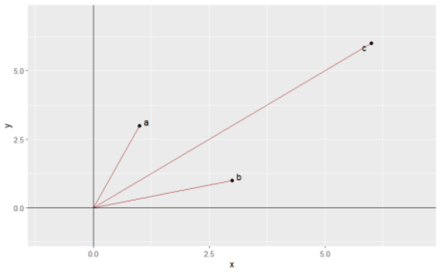


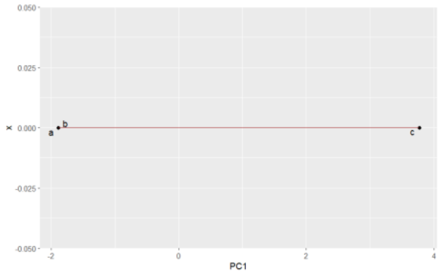
- The Intra genre distance for rap is extremely low followed by country music indicating that the songs are not that different from each other in terms of the structure (fair enough)

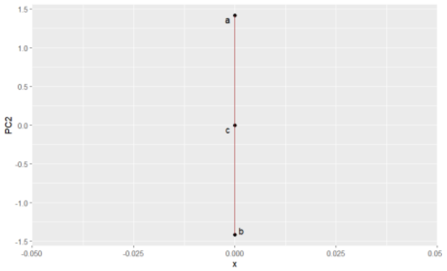
- Jazz and Lounge have similar patterns here as well.

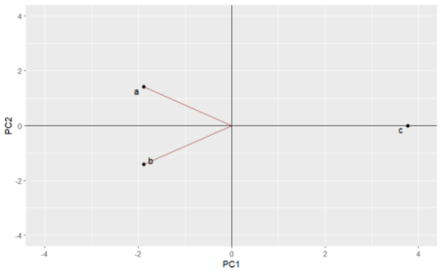
- It does give a decent idea if how close these genres are to each other.

Principal Component Analysis





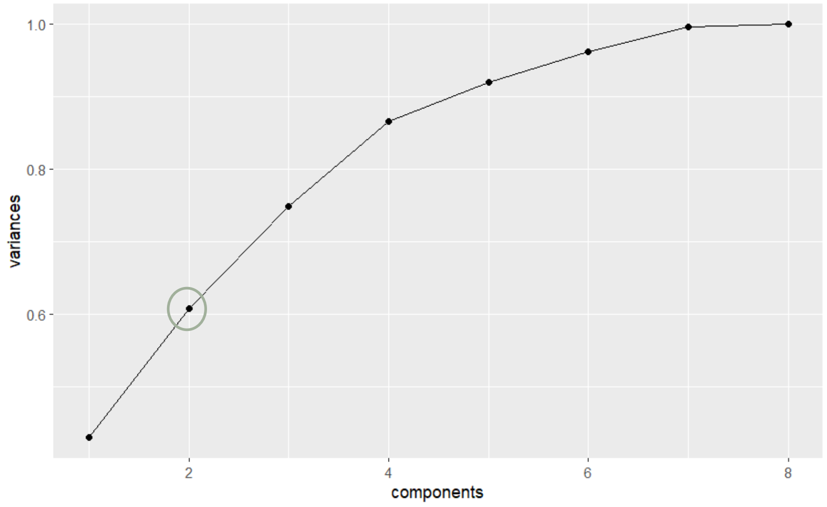


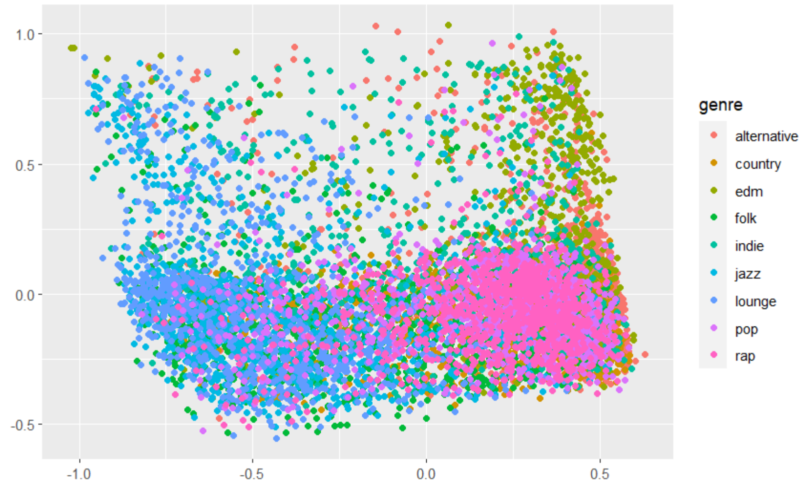


- Dimensionality reduction
- Explain more variance with limited features
- Applicable only to numerical features.

Principal Component Analysis

PC vs Explained Variance



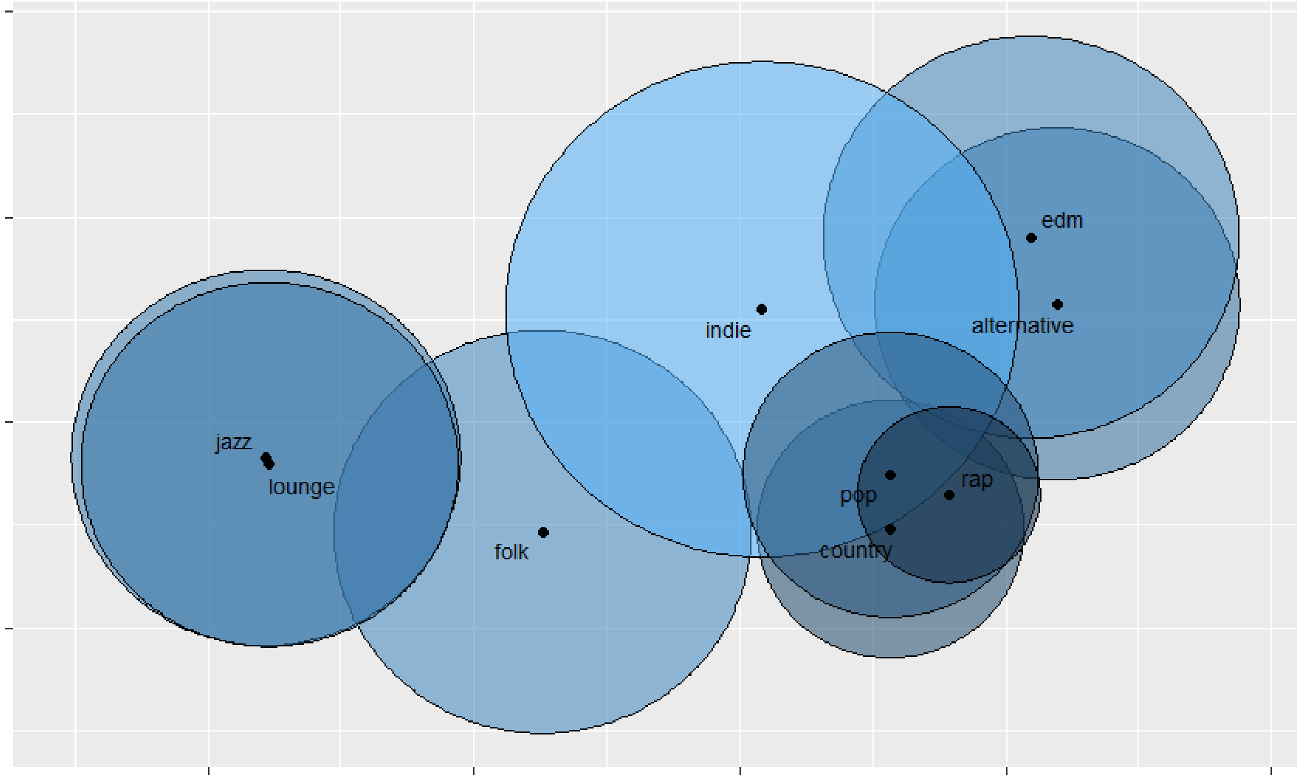


MESSSS!!!!

The Big Picture

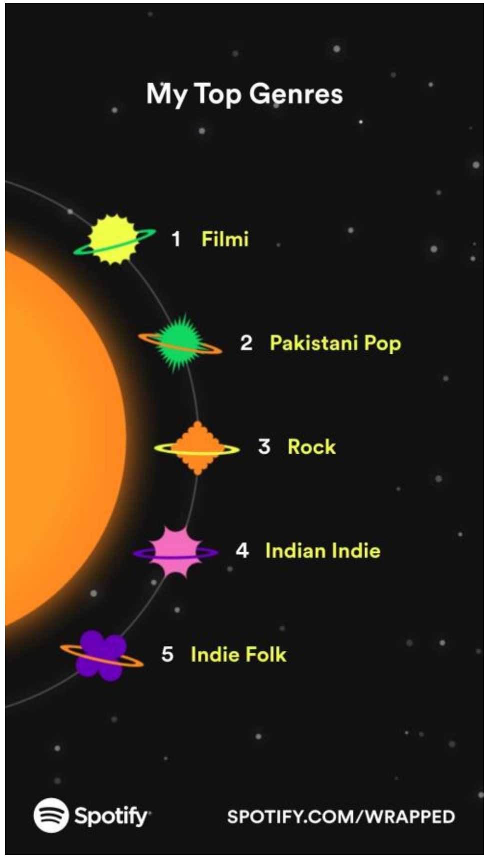
The centroids of each genre in a PCA plot with the radius as the average distance to this centroid do give a broad overview of how the genres are relative to each other. This also confirms a lot of inferences made earlier.

- Jazz ~ Lounge
- EDM ~ alternative
- Indie overlaps with many genres.
- Rap has less variation and overlaps with pop quite a bit.



Recommendations

The library



My Top Genres

1 Filmi

2 Pakistani Pop

3 Rock

4 Indian Indie

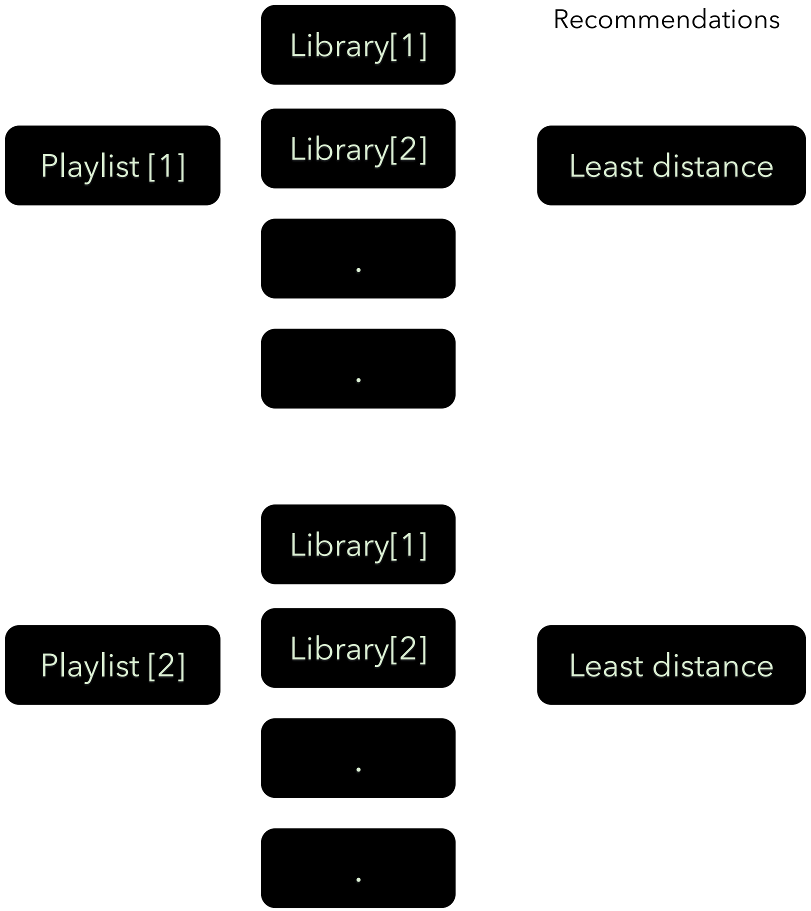
5 Indie Folk

Spotify SPOTIFY.COM/WRAPPED

Artists from the playlist (100 songs)

Artists similar (40k songs)

Individual similarity


Recommendations

Playlist [1]

Library[1]

Library[2]

.

.

Least distance

Playlist [2]

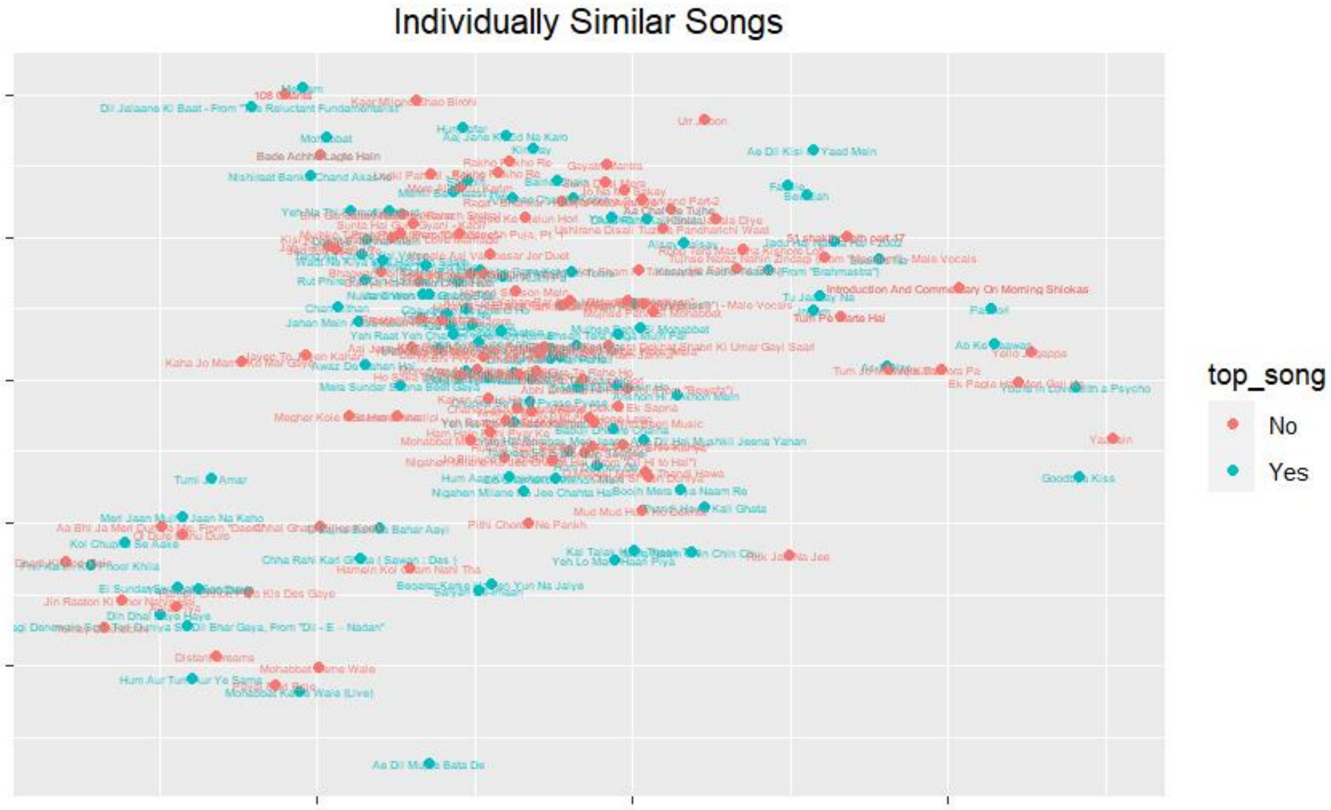
Library[1]

Library[2]

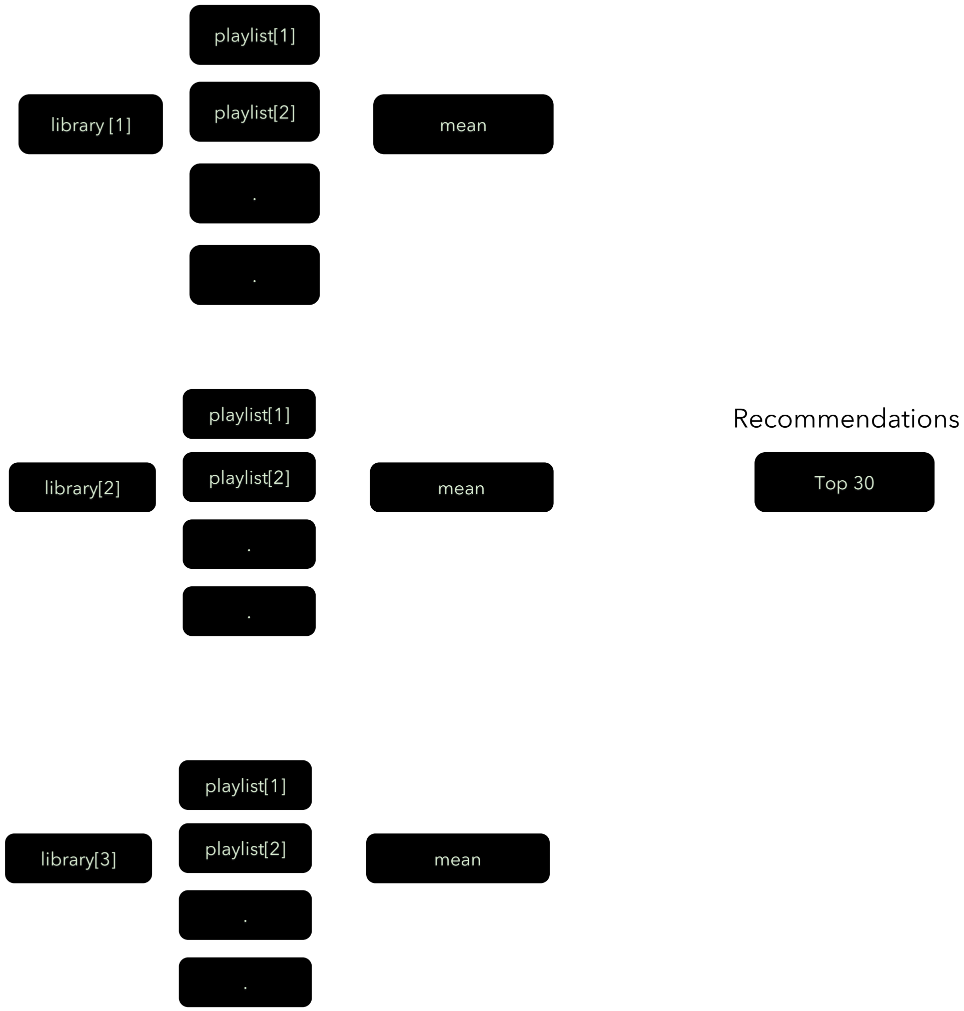
.

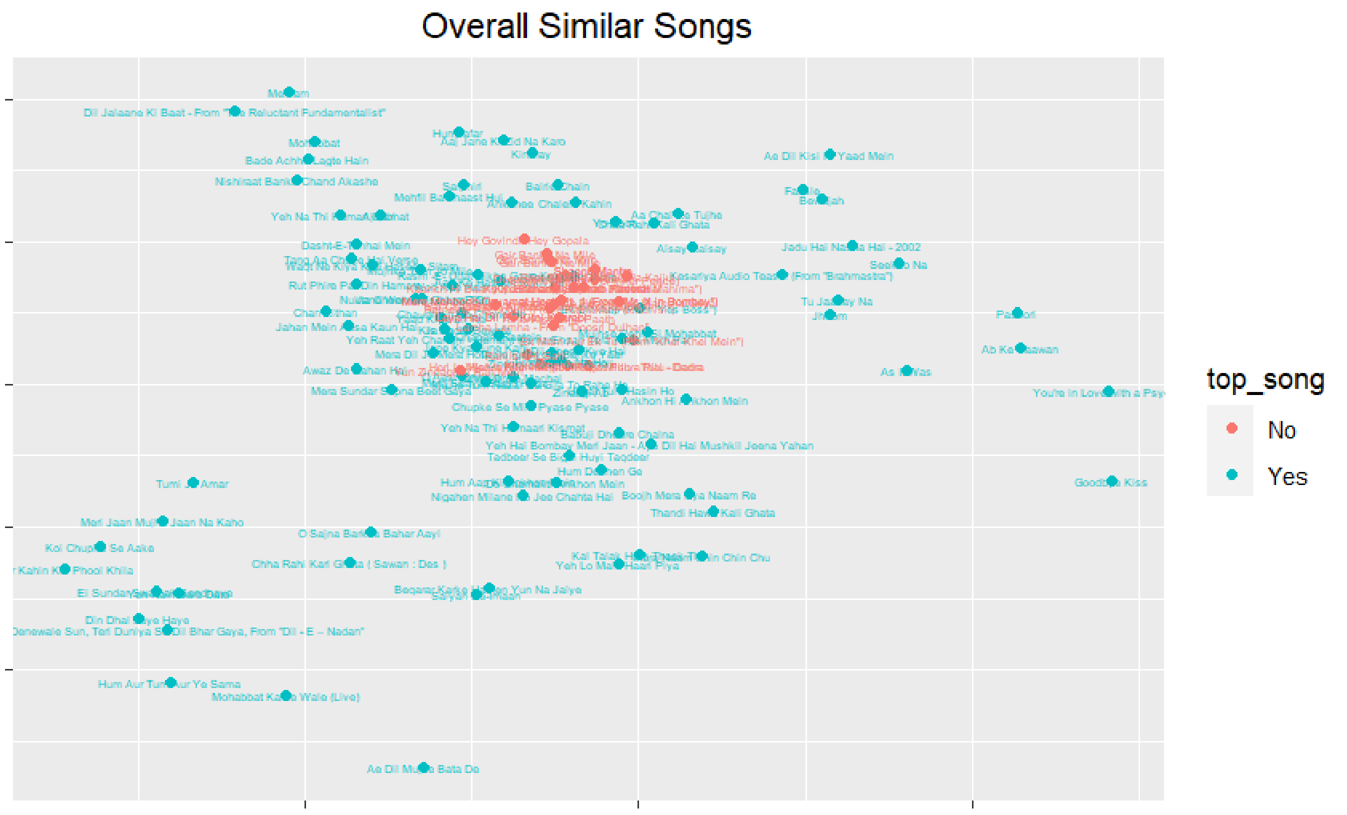
.

Least distance



Overall similarity





Scope

+ Exploring other distance metrics like Gower or podani.

+ Getting dimensionality reduction tools that incorporate categorical features.

+ Get more playlists to work with which enables testing the accuracy of the recommendation system.

Thank You

udga1318@colorado.edu