

# Spotify Data Analysis

Uday Gadge

University of Colorado, Boulder

**Abstract** - The project has two parts. The first part is aimed at analyzing the trends across 9 popular genres of music and make inferences on the evolution over time. The second part involves getting music recommendations based on a playlist. The project uses the distance metrics and principal component analysis along with various plots from ggplot.

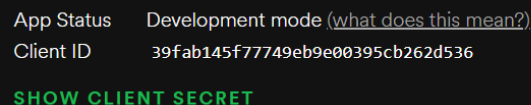
## 1. Introduction:

Spotify is the world's leading music streaming service platform serving 800 million users across the world. The music library has about 80 million tracks across 1000 genres and subgenres. Spotify also provides access to certain features of a song that can be mined through an API (Application Programming Interface). However, it doesn't give access to any user related info. It does inhibit exploring a lot of options for a recommendation system. The data is mostly clean and ready to use. This analysis is focused on nine genres - pop, rap, EDM, jazz, folk, country, lounge, alternative and indie. The data is collected for 30 most followed artists in each genre. This might create a certain bias especially in genres such as indie, country where the top artist's music might have an overlap with the pop genre. The whole analysis is also a good way to see if our perceptions of these genres hold truth. The recommendation system is built on a personal playlist (luckily the spotify wrapped came at the right time). This playlist is filled with music rooted deep in the Indian subcontinent and is wildly different from the

dataset used in the first part of the project. Hence, the second part used a different dataset based on the artists in the playlist instead.

## 2. Data:

Spotify provides data through a developer account. Once registered, it provides an access id and passwords.



App Status Development mode ([what does this mean?](#))  
Client ID 39fab145f77749eb9e00395cb262d536  
**SHOW CLIENT SECRET**

R has a package '*Spotifyr*' that generates an access token. This token can be used to pass queries to the server and extract data.

## 2a. Data for genre analysis

*get\_genre\_artists(genre, limit)*



```
{r}  
get_genre_artists(genre = "pop", limit = 30)
```

A tibble: 30 x 12

id	images	name	popularity
<chr>	<list>	<chr>	<dbl>
06HL4z0CvFAxyc27GXpf02	<data.frame [3 x 3]>	Taylor Swift	100
3TVXtAsR1Inumwj47259r4	<data.frame [3 x 3]>	Drake	97
1XyoaUu8XC12mMpaF05Pj	<data.frame [3 x 3]>	The Weeknd	95
66CXWjxzNUsdJx12dwnR	<data.frame [3 x 3]>	Ariana Grande	90
5pKCCKE2ajHZ9KAiaK11H	<data.frame [3 x 3]>	Rihanna	89
00FqB4jTyendYWa8pK0wa	<data.frame [3 x 3]>	Lana Del Rey	89
7bXgB6jMjp9ATfY66eO08Z	<data.frame [3 x 3]>	Chris Brown	88
6KlmcVD70vtQjWnq6nGn3	<data.frame [3 x 3]>	Harry Styles	90
1uNF0ZAH8GtllmzznPc13s	<data.frame [3 x 3]>	Justin Bieber	91
5cj0lljcoR7YOSnhnXOPo5	<data.frame [3 x 3]>	Doja Cat	87

1-10 of 30 rows | 3-6 of 12 columns

The dataset contains 30 artists per genre for the 9 genres.

*get\_artist\_audio\_features(artist\_Id)* to get the discography of an artist.

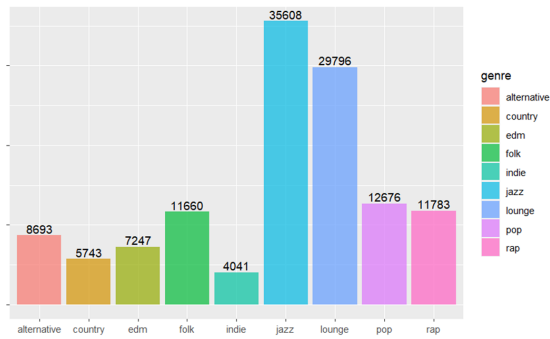


Fig 2a. The distribution of genres in the dataset

Since we took 30 artists from each genre, it can be said that the lounge and jazz artists release more music than other genres (fig 2a.). Indie music is the lowest. This is probably due to the fact that indie artists usually release singles instead of albums.

## 2b. Data for recommendations

Spotify releases the top songs heard in the year as a playlist. The tracks in this playlist can be used as the base for getting recommendations. Recurring artists in this playlist can be taken as the artists for building the library. Spotify also provides other similar artists. The combined discography is treated as the library.

`get_playlist(playlistID)`

`get_track_audio_features(trackID)`

`get_related_artists(ID)`

`get_artist_audio_features(ID)`

## 2c. Features:

The data collected has features associated with a song that are used widely in the analysis:

- Popularity - A popularity index from (0-100)
- Danceability - Danceability of a song (0 - 1)

- Energy - Energy of a song (0 - 1)
- Loudness - loudness of a song in db (-60 - 0)
- Acousticness - Acoustic measure (0 - 1)
- Instrumentalness - lack of vocals (0 - 1)
- Liveness - Recorded live (0 - 1)
- Valence - Positivity of the song (0 - 1)
- Tempo - Temp in beats per minute (BPM)
- Key\_mode - major or minor
- Key\_name - key in which it's played (A, A#,..)
- Key - Key\_name+Key\_mode

## 3a. Hypothesis testing

In order to test if pop music artists have a significantly higher number of followers on Spotify compared to other genres, Hypothesis testing was done.(fig.3a)

If the entire dataset of 270 artists is to be assumed as the population with a mean  $\mu$  and a standard deviation of  $\sigma$ , a sample mean of 30 entries is estimated to have a distribution of:

$$\bar{X} = N(\mu, \sigma/\sqrt{n}), n = 30$$

We can define null hypothesis and alternate hypothesis as:

$H_0$  = The number of followers for a pop artist is not significantly higher.

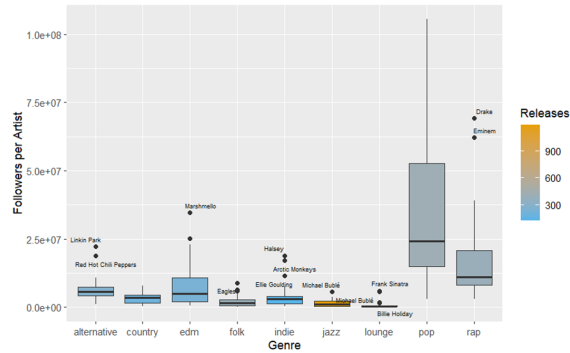


fig 3a. Followers per artist for different genres

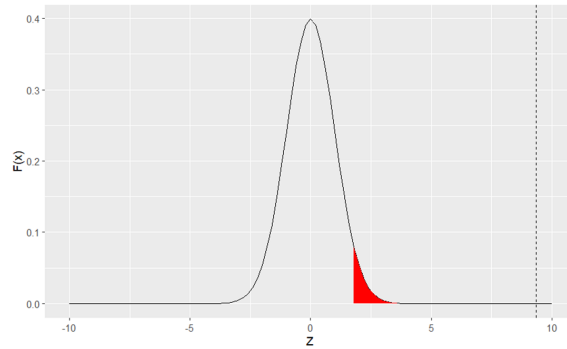


fig3b. Z-test

$H_1$  = The number of followers for a pop artist is significantly higher.

The critical Z value for  $\alpha = 0.05$  is around 1.64 and the obtained Z value for the sample mean is way higher (around 9). The null hypothesis is rejected and it can be concluded that pop music artists have significantly higher followers than other genre artists.(fig 3b)

### 3b. Inferences

- Pop and rap music are way more popular than other genres of music, while lounge music had a lull in popularity in the 2000s.
- Jazz and lounge music are similar to each other, while indie music has more variance as it tends to overlap with various genres.

- Valence, which is a measure of positivity in a song based on its lyrics, is on a steady decline, especially in pop and rap music. It can be because the algorithm used to predict valence is misclassifying songs due to the increased use of irony in recent years. In addition to this, pop music has shifted towards heartbreak themes compared to its earlier version had themes of love and empowerment.
- While pop music's danceability has gone down owing to the aforementioned reasons, indie has become more danceable.
- Key distribution heat maps help support the earlier claim that jazz and lounge music share similar compositions while rap music is dominated by a single key (implying little variation).

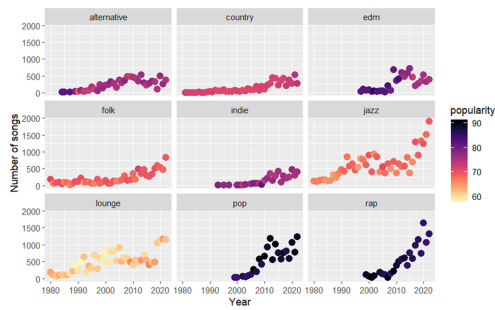


fig 3c. Popularity of songs based on release year



fig 3d. Acousticness, Instrumentalness and Liveness

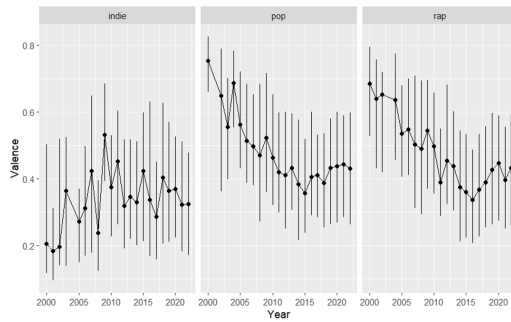


fig 3e. Valence over time

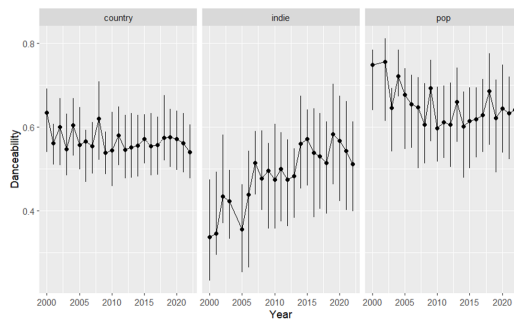


fig 3f. Danceability over time

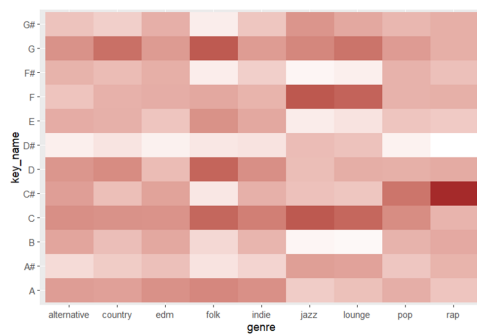


fig 3g. Presence of certain keys

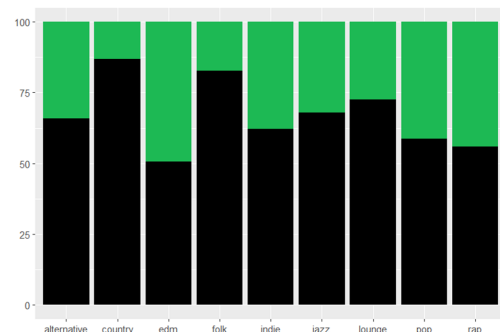


fig 3h. Major vs minor key

### 3c. Distance Metric

The analyses made were further expanded using distance metrics to get a general idea of how these genres are different from one another. Most of the features in the data, except the key, are numeric. Euclidean distance is a good way to

calculate distances between songs with numerical features. The analysis used min max scaling because most of the features are not normally distributed. Min max scaling fits all the values within a (0,1) interval by calculating,

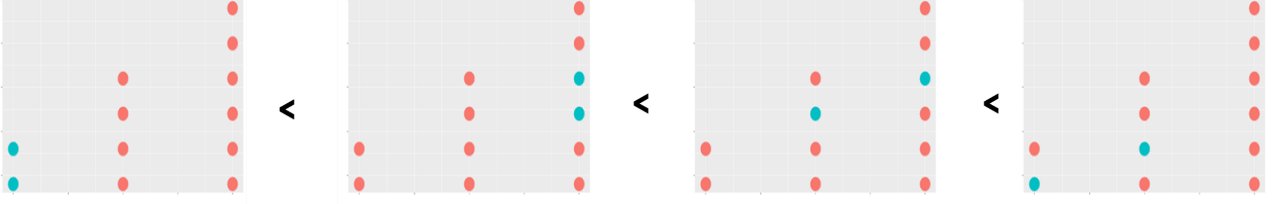


fig 3i

$$x_{scaled} = (x - \min(X)) / (\max(X) - \min(X))$$

A normal euclidean distance is defined for numerical features as

$$Dis(X, Y)_{euclidean} = \sqrt{\sum_i (x_i - y_i)^2}$$

The analysis modified this distance to accommodate categorical features using probabilities.

$$Dis(X, Y) = \sqrt{\sum_{num} (x_i - y_i)^2 + \sum_{cat} \delta_{ij} P_i P_j + (1 -$$

This enabled the distance to be smaller in cases where the probability of a feature being the same was low. It also maximized the distance where the probability of a feature being different was low for the given categories(fig.3i). Although this distance gives more weight to categorical features than numerical, the key in which a song is played is crucial to how it sounds.

Using this distance metric, the mean distance between two points in every genre was calculated. This was done in terms of both intra genre distances(fig.3j) and inter genre distances(fig.3k).

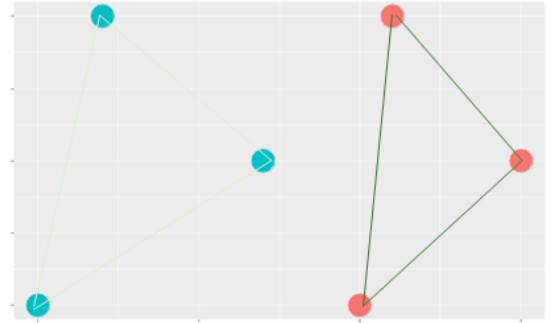


fig 3j. Intra genre distance

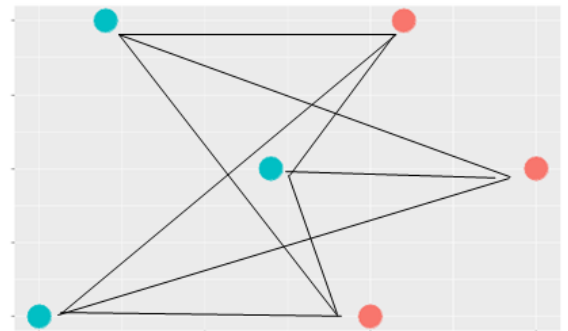
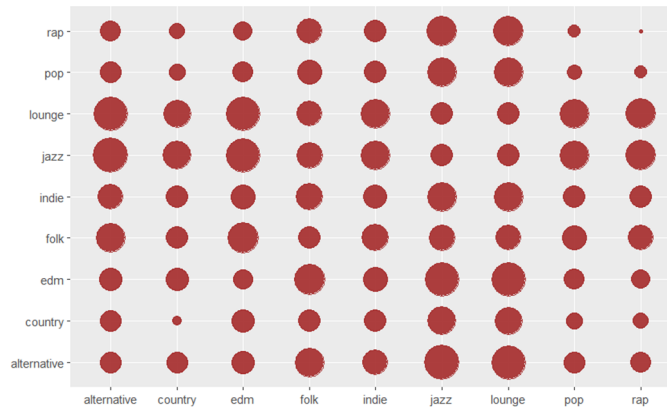


fig 3k. Inter genre distance



*fig 3l. Distance matrix of the genres*

This supported the following inferences made from earlier analyses.

- Rap and country music have minimum variation within themselves.
- Jazz and lounge are similar and have similar distances with other genres.
- Indie music overlaps with other genres.

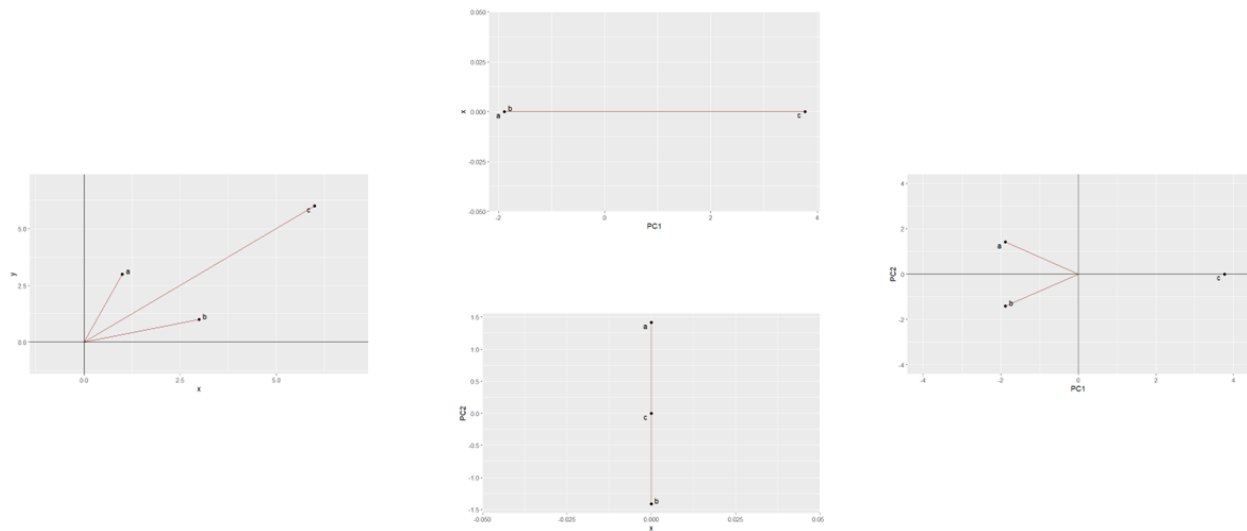
### 3d. Principal Component Analysis

Principal component analysis is a dimensionality reduction tool. It constructs new features as a linear combination of current features in a way that maximum information is stored in fewer dimensions. The first feature is constructed such that it accounts for the largest possible variance. The second feature is constructed in a way that it is uncorrelated with the first feature and explains the next largest possible variance and so on. There can be as many principal components as the features and the explained variance is cumulative(*fig.3m*).

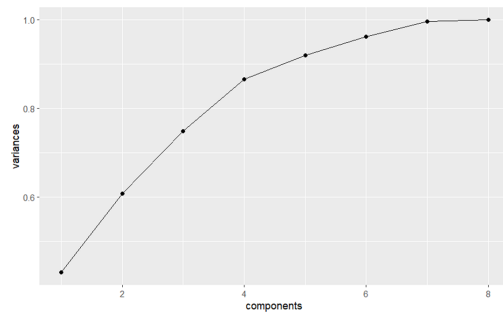
Principal Components are often a trade-off between variance, accuracy and simplicity. Although these components themselves can't be interpreted as they are a linear combination of scaled features, they are a great way to visualize the interaction between the data points in a more holistic view.

PCA can only be applied to numerical data. The analysis used numerical features to construct principal components. Although the first two components only explain about 60% of the total variance, they adequately visualize the relationship between the data points.

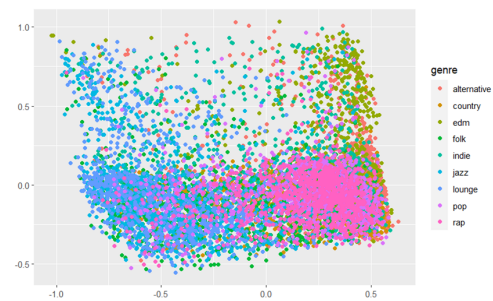
These principal components were visualized for ease of interpretation by calculating the centroids of each genre and the mean intra genre distance.



*fig 3m*



*fig 3n. Explained Variance*



*fig 3o. Principal component*

This confirmed the following inferences:

- Jazz and lounge are pretty much the same.
- Indie has a major overlap with a lot of genres barring jazz and lounge.
- Rap music has a very low spread and it is almost a part of pop music.
- EDM and Alternative music overlap to a major extent.

#### 4. Recommendation system

The second part of the project built recommendations based on the user's top 100 songs in 2022. The same distance metric mentioned earlier was used to compute distances between the top songs and the library mined.

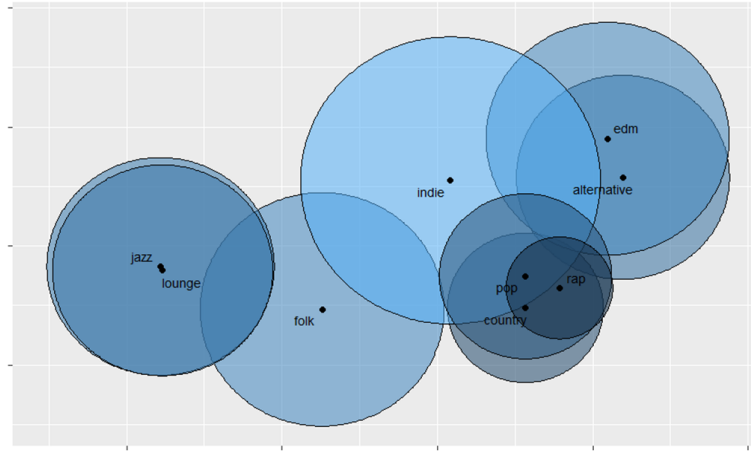


fig 3q. Visual representation of Principal Component

#### 4a. One-to-One recommendations

This recommended a track in the library that was closest to every song in the playlist (fig 4a).

#### 4b. Overall recommendations

This recommended the top 30 tracks with the minimum mean distance with the playlist(fig 4b)

#### 5. Scope:

- Although the distance metric used accounts for categorical features, it is biased towards them. Other mixed distances like gomer, podani or huang can be analyzed.
- The principal components only work with numerical features. There are other dimensionality reduction methods that can incorporate categorical features like factor analysis of mixed data.
- More playlists can be taken to build the base and recommendations can be

provided through a more detailed analysis of distance metrics.

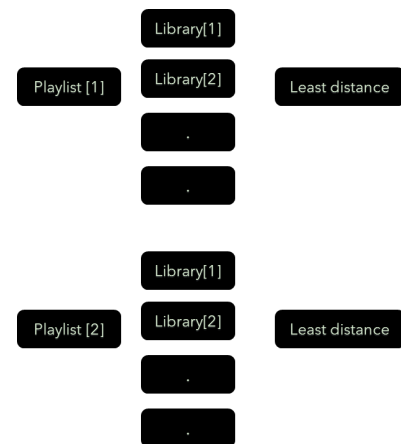


Fig 4.a

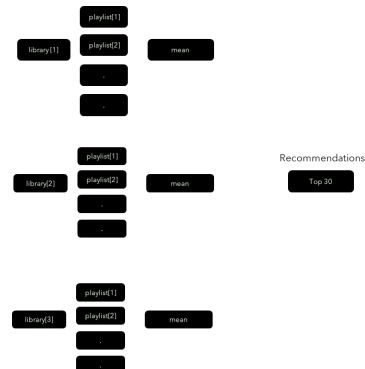


fig 4.b



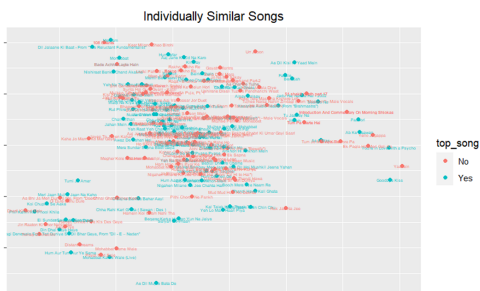


fig 4c. Individual similar songs

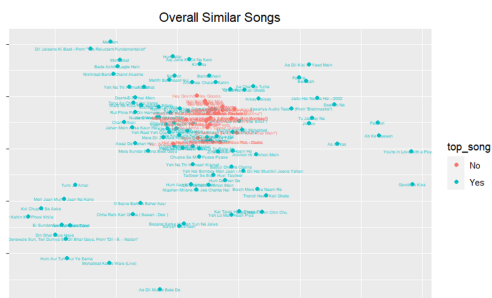


fig 4d. Overall similar songs