Spotify

Recommendation System

Data Mining Principles Group Project

Group 2

Devdutt Sharma, Scott Matsubara, Sanket Mayekar, Taylor Furry, Yi Chin Tzou

2023.03.08

Our Team

Devdutt Sharma

Scott Matsubara

Sanket Mayekar

Taylor Furry

Yi Chin Tzou

Agenda

#	Title	
1	Problem statement	00:00
2	Our Data	01:00
3	Data Exploration	02:00
4	Methodology	05:00
5	Conclusions	13:00

Problem Statement

The goal of our team is to analyze, build and improve the Spotify's recommendation system in order to increase customer satisfaction, retention, and time spent on the platform broadly.

Using different algorithms such as popularity based, Content-Based Filtering and K-Nearest Neighbors to build a recommendation engine that can suggest songs to the user based on their listening history and preferences.

Our Data

Variable	Data Type
track_id	String
artists	String
album_name	String
track_name	String
explicit	Categorical
key	Categorical
mode	Categorical
time_signature	Categorical
track_genre	Categorical

Variable	Data Type
popularity	Numeric
duration_ms	Numeric
danceability	Numeric
energy	Numeric
loudness	Numeric
speechiness	Numeric
acousticness	Numeric
instrumentalness	Numeric
liveness	Numeric
valence	Numeric
tempo	Numeric

Data Pre-Processing

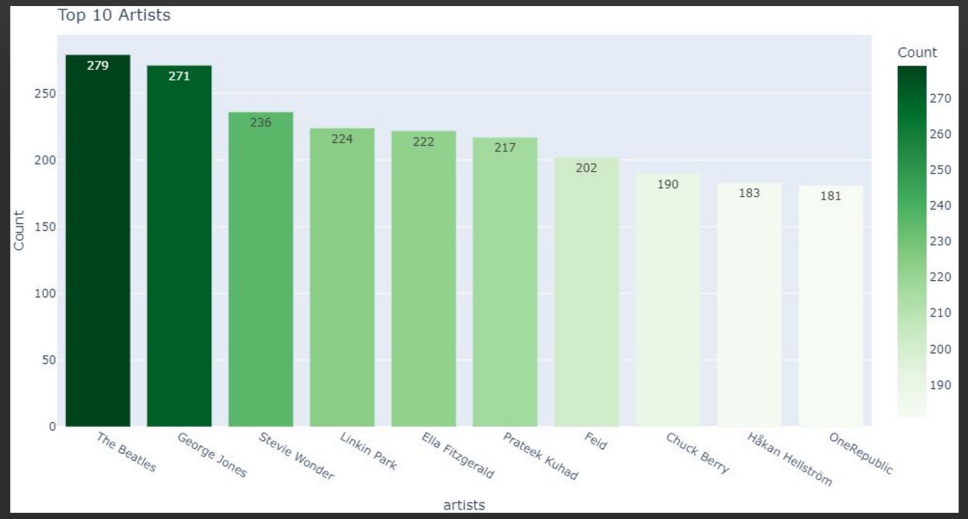
- Removals
  - Unnecessary columns such as mode, key, and time signature
  - Children's songs Genre
  - Duplicate track names

- Changes
  - Scaled numeric features
  - Created combined genre column
  - Added dummy variables for genre

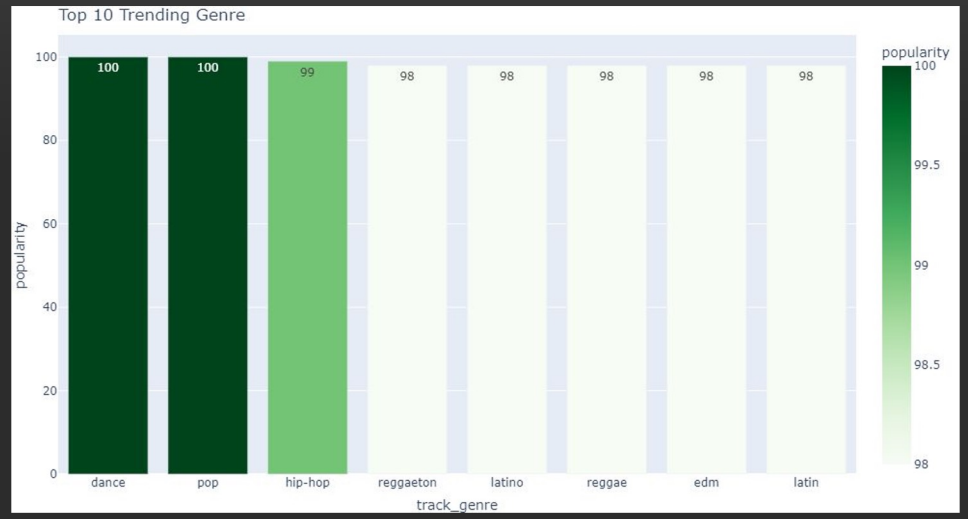
Data Limitations

- Limited data set (About 90,000 songs)
  - Some genres were not included entirely in our dataset (ex. Bollywood)
- No user or feedback ratings
- Had to use 2 different data sets
  - Not all English makes NLP problematic

Data Exploration

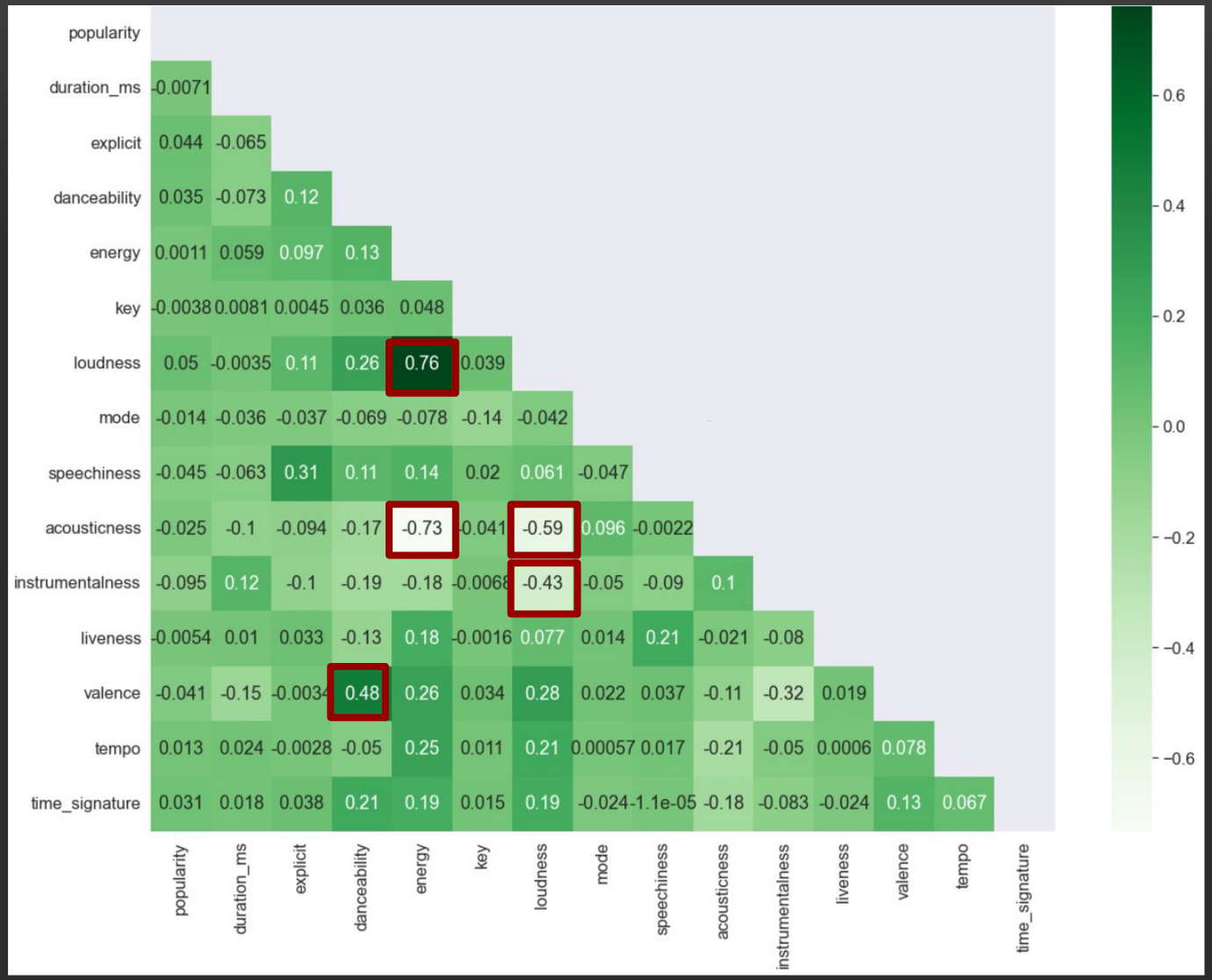


Top 10 Artist with the Most Songs



Top 10 Genres by Popularity

Data Exploration: Correlations



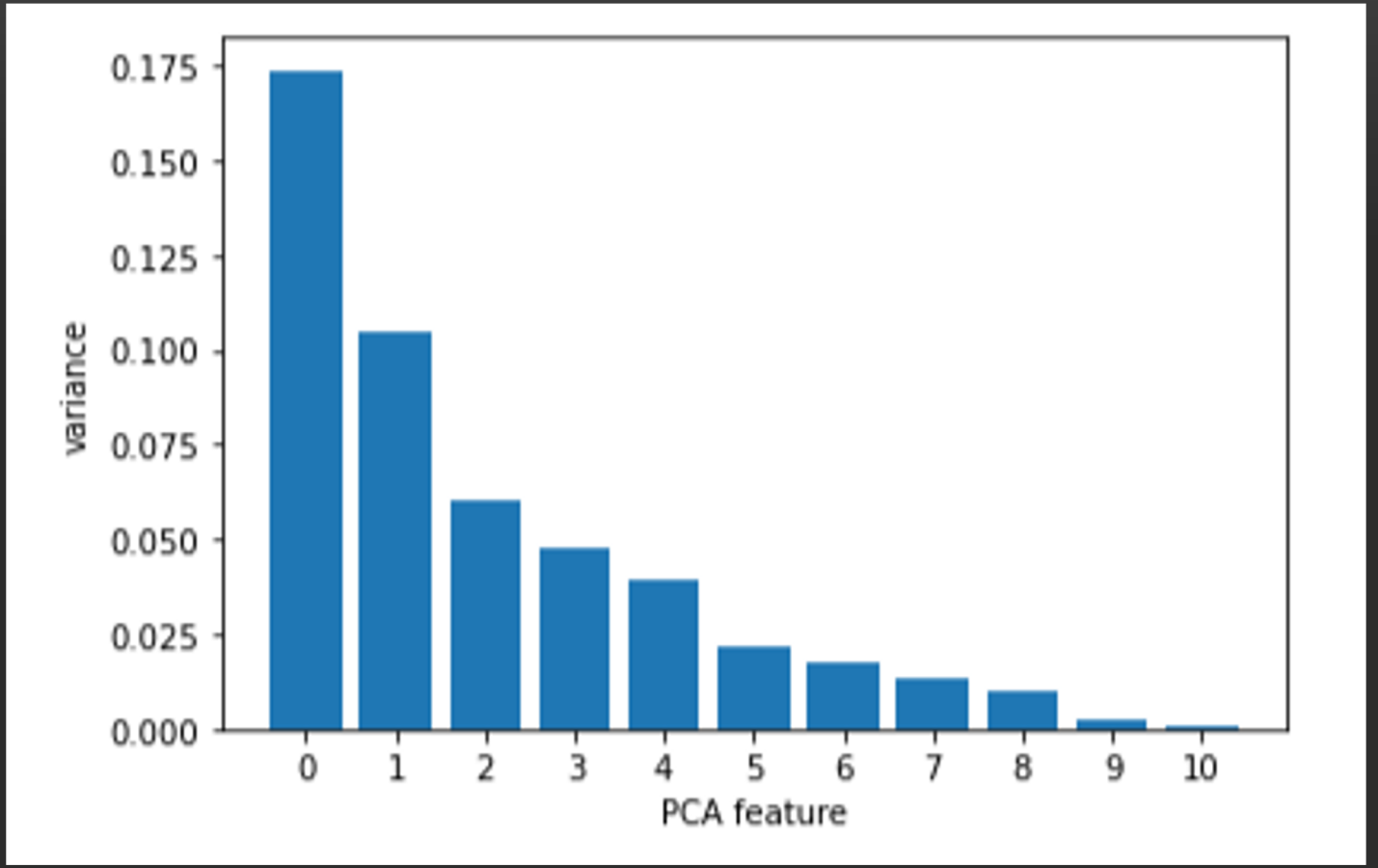
Negative correlation:

- ACCOUSTICNESS and ENERGY
- ACCOUSTICNESS and LOUDNESS
- INSTRUMENTALNESS and LOUDNESS

Positive correlation:

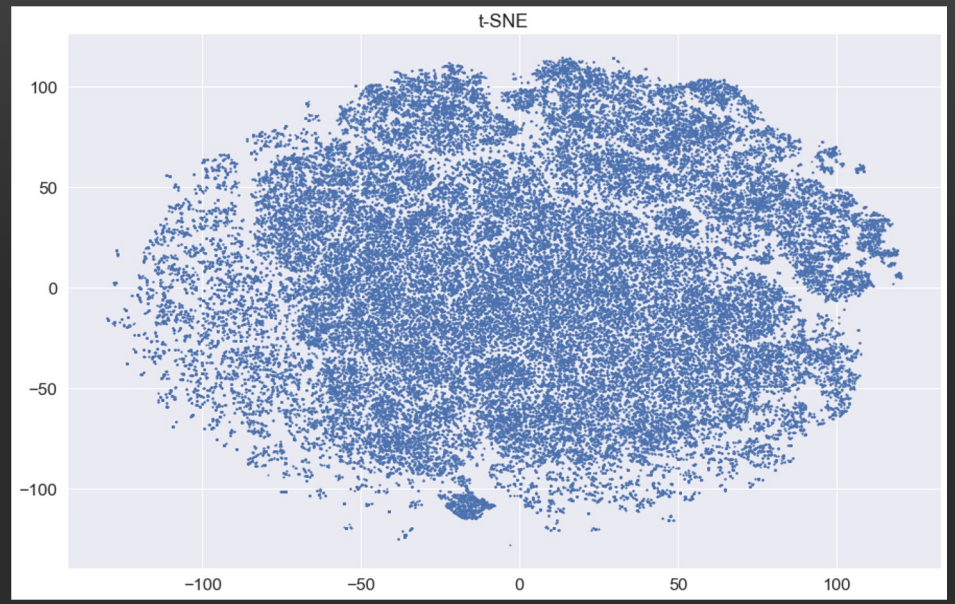
- LOUDNESS and ENERGY
- DANCEABILITY and VALENCE

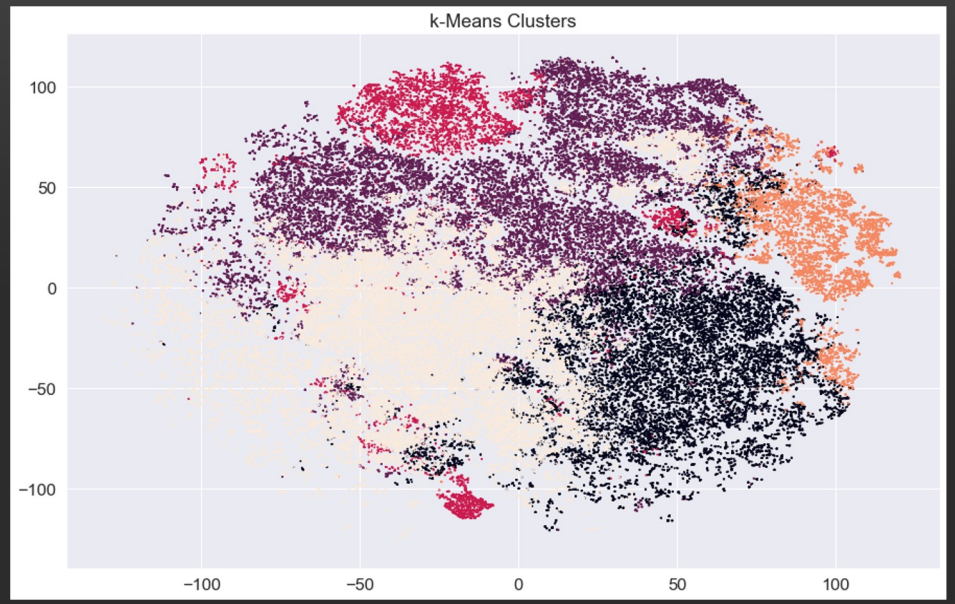
Data Exploration: Dimension Reduction



- Bar graph showing explained variance by features for PCA. We can see a lot of unexplained variance

Data Visualization: Clustering - (Exploration)





- Dimensionality Reduction using t-SNE
- Clustering using k-means

The purpose is to identify patterns and similarities between the tracks based on their features, and to group them into clusters based on those similarities.

1. Original features are transformed into new, lower-dimensional features using t-SNE that best preserve the pairwise similarities between data points.

1. Each point in the plot represents a track and its position reflects its similarity to other songs in the dataset. Each track was assigned a cluster label, which represented its similarity to other songs in that cluster. The color of each point indicates the cluster that it belongs to, as assigned by the k-means algorithm

However, it can be difficult to interpret the meaning of these new features because they are often complex combinations of the original features.

NOTE: We did not make use of this analysis in our model building step

Model 1-Popularity Based

- Fixes the cold start problem
- Can be used for new users that have not input a song or playlist of which they want recommendations
- Gives the most popular songs based on popularity score
  - Can be most popular overall or by genre
- Measured by given popularity column

S.No.	Artist	Track Name	Genre
1	The Killers	Runaway Horses - Abridged	alternative
2	Joy Division	Atmosphere	alternative
3	No Doubt	Just A Girl	alternative
4	Weezer	Island In The Sun	alternative
5	Nirvana	Come As You Are	alternative
6	The Smashing Pumpkins	Bullet With Butterfly Wings	alternative
7	Red Hot Chili Peppers	Nerve Flip	alternative
8	Andres Calamaro	Enganchate Conmigo	alternative
9	Los Abuelos De La Nada	Lunes Por La Madrugada	alternative
10	Hillsong UNITED	Oceans (Where Feet May Fail)	alternative

Model 1-Popularity Based

S.No.	Artist	Track name	Genre	Popularity
1	Sam Smith, Kim Petras	Unholy	Dance, pop	100
2	Bizarrap; Quevedo	Quevado: Bzrp Music Sessions, Vol. 52	Hip-hop	99
3	David Guetta; Bebe Rexha	I'm Good (Blue)	Dance, EDM, Pop	98
4	Manuel Turizo	La Bachata	Latin, Latino, Reggae, Reggaeton	98
5	Bad Bunny	Titi Me Pregunto	Latin, Latino, Reggae, Reggaeton	97
6	Bad Bunny, Chencho Corleone	Me Porto Bonito	Latin, Latino, Reggae, Reggaeton	97
7	OneRepublic	I Ain't Worried	Piano, Pop, Rock	96
8	Bad Bunny	Efecto	Latin, Latino, Reggae, Reggaeton	96
9	Bad Bunny, Bomba Estero	Ojitos Lindos	Latin, Latino, Reggae, Reggaeton	95
10	Bad Bunny	Moscow Mule	Latin, Latino, Reggae, Reggeaton	94

Model 2: Lyrics Based-Song Radio

Use Genius API to get the lyrics from our dataset

1. 58,913 songs' lyrics
2. Filtering english songs: Left with 40,745 songs
3. Filtering duplicate songs: Left with 35,015 songs

Data cleaning

1. Removing punctuations, and stopwords

Building Natural Language Processing(NLP) pipeline

1. Lemmatize
2. Tokenize
3. CountVectorizer
4. Use cosine similarity to calculate the similarity of songs

Input 1 song id, will output 10 songs based on the highest similarity score

Model 2: Lyrics Based-Song Radio

Insert:

Song ID	Track name	Artist
'01MVOl9KtVTNfFiBU9I7dc'	Days I Will Remember	Tyrone Wells

Result:

S.No.	Track name	Artist	Similarity score
1	Domino Dancing	Pet Shop Boys	0.6465
2	Another Day	White Reaper	0.6322
3	Some Days	Brent Morgan	0.6108
4	Once A Day	George Jones	0.5870
5	The Emperor	YUNGBLUD	0.5755
6	Day N Night	Afrojack;Black V Neck;Muni Long	0.5712
7	Lonely Day	System Of A Down	0.5637
8	Not A Day Goes By	Leatherface	0.5596
9	Good Day for Living	Joe Nichols	0.5427
10	Wait	Martin Jensen;Loote	0.5339

Model 3: Feature Based Recommendation

Prepare Dataset

- Optionally include genre feature and/or only include English song titles
- Scale numeric variables

Get Inputted Playlist Data

- Using Spotify API, get DataFrame of songs in playlist
- Scale inputted data
- Remove inputted songs from dataset

Find Similarity Using Distance

- Use Gower distance to include genre for comparison
- Use Euclidean distance for just similar sounding songs
- Find most similar songs to each song in the playlist
- Treat all inputted songs as cluster and find closest songs to centroid

Output Curated Spotify Playlist

- Create playlist in Spotify and output playlist URL

Live Demo

User Created and Inputted Playlist

Playlist

2000s bops

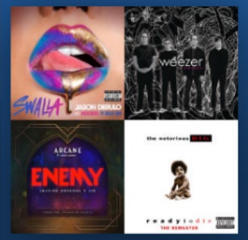
Scott Matsubara • 1 like • 157 songs, about 9 hr 30 min

#	Title	Album	Date added	
1	Love The Way You Lie E Eminem, Rihanna	Recovery	Feb 7, 2019	4:23
2	Stereo Hearts (feat. Adam Levine) Gym Class Heroes, Adam Levine	The Papercut Chronicles II	Feb 7, 2019	3:30
3	Pumped Up Kicks Foster The People	Torches	Feb 7, 2019	3:59
4	Bad Day Daniel Powter	Daniel Powter	Feb 7, 2019	3:53
5	The Man Who Can't Be Moved The Script	The Script	Jul 6, 2022	4:01
6	Big Girls Don't Cry (Personal) Fergie	The Dutchess	Feb 7, 2019	4:28
7	I Don't Want to Be Gavin DeGraw	Chariot	Feb 7, 2019	3:37
8	Bubbly Colbie Caillat	Coco	Feb 7, 2019	3:16
9	Moves Like Jagger - Studio Recording From "The Voice" Performance Maroon 5, Christina Aguilera	Hands All Over	Feb 7, 2019	3:21
10	Somebody That I Used To Know Gotye, Kimbra	Making Mirrors	Feb 7, 2019	4:04

Input and Output of Program


```
(venv) (base) scottsmacbook@Scotts-Air Spotify % python Spotify_Reccomendation_System.py
This Program will Generate a Playlist of Reccomendations Based on a Playlist that you Enter

Would you like most songs to be in English? (Y/N) y

Would you like the playlist to be based on Genre? (Y/N) n

To Get the Playlist URI:
1. Go into the desired playlist and make sure the playlist is public
2. Click on the 3 horizontal dots for more options
3. Scroll down to the share option, hover over Share, press option, and click the option: Copy Spotify URI

Note: this program will only generate recommendations based on the first 10 songs of the playlist.
Because of this, you may get songs that are already in your playlist that are after the first 10.

Please input a Spotify Playlist URI: spotify:playlist:3Ins6u27JrxxYbQaQ0SmDY
Got track 1 out of 88
Got track 2 out of 88
Got track 3 out of 88
Got track 4 out of 88
Got track 5 out of 88
Got track 6 out of 88
Got track 7 out of 88
Got track 8 out of 88
Got track 9 out of 88
Got track 10 out of 88
Gettings Recs for Song 1
Gettings Recs for Song 2
Gettings Recs for Song 3
Gettings Recs for Song 4
Gettings Recs for Song 5
Gettings Recs for Song 6
Gettings Recs for Song 7
Gettings Recs for Song 8
Gettings Recs for Song 9
Gettings Recs for Song 10

Your Playlist is Ready!
Here is the playlist URL: https://open.spotify.com/playlist/1VmngZf1r7WMcU1HMuy9Rr
```

Recommended Playlist is Generated in Spotify



Playlist

2000s bops Playlist Recs

Scott Matsubara • 33 songs, about 1 hr 45 min

#	Title	Album	Date added		
1	Enemy (with JID) - from the series Arcane League of Legends Imagine Dragons, JID, Arcane, League of Legends	Enemy (with JID) [from the series Arcane League of Legends]	59 seconds ago		2:53
2	Who Shot Ya? - 2005 Remaster E The Notorious B.I.G.	Ready to Die (The Remaster)	59 seconds ago		5:19
3	Swalla (feat. Nicki Minaj & Ty Dolla $ign) E Jason Derulo, Nicki Minaj, Ty Dolla $ign	Swalla (feat. Nicki Minaj & Ty Dolla $ign)	59 seconds ago		3:36
4	Beverly Hills Weezer	Make Believe	59 seconds ago		3:16
5	Love$ick (feat. A$AP Rocky) E Mura Masa, A$AP Rocky	Mura Masa	59 seconds ago		3:12
6	Snap Out Of It Arctic Monkeys	AM	59 seconds ago	♥	3:13
7	Sweet Dreams (Are Made of This) - Remastered Eurythmics, Annie Lennox, Dave Stewart	Sweet Dreams (Are Made Of This)	59 seconds ago		3:36
8	Lady - Hear Me Tonight Modjo	Modjo (Remastered)	59 seconds ago		5:07
9	Words - Original Version 1983 F.R. David	Words	59 seconds ago		3:29
10	13 Missed Calls E Zach Hood	13 Missed Calls	59 seconds ago		3:29
11	Waving Through A Window Ben Platt, Original Broadway Cast of Dear Evan Hansen	Dear Evan Hansen (Original Broadway Cast Recording)	59 seconds ago		3:56
12	There Is No Wonderwall Sylow, Pierre Leck, Kingsley Q	There Is No Wonderwall	59 seconds ago		3:38
13	Jab Koi Baat - Recreated Atif Aslam, Shirley Setia, Dj Chetas	Jab Koi Baat - Recreated	59 seconds ago		3:12
14	Po Ve Po - The Pain of Love Anirudh Ravichander, Mohit Chauhan	3 (Original Motion Picture Soundtrack)	59 seconds ago		4:14
15	I Wish (Christmas Version) Joel Corry, Mabel	Cuddle Up Christmas	59 seconds ago		3:20

Another Inputted Playlist

PUBLIC PLAYLIST

kpop jim

Sean Matsubara and audrey • 1 like • 88 songs, 4 hr 52 min

Custom order

#	Title	Album	Added by	Date added	
▶	Back Door Stray Kids	IN LIFE	Sean Matsubara	Aug 25, 2022	3:09
2	FIRST EVERGLOW	Last Melody	Sean Matsubara	Aug 25, 2022	3:32
3	God's Menu Stray Kids	IN LIFE	Sean Matsubara	Aug 25, 2022	2:48
4	Thunderous Stray Kids	NOEASY	Sean Matsubara	Aug 25, 2022	3:03
5	Kick It NCT 127	NCT #127 Neo Zone - The 2nd Album	Sean Matsubara	Aug 25, 2022	3:53
6	LOCO ITZY	CRAZY IN LOVE	Sean Matsubara	Aug 25, 2022	3:11
7	The Feels TWICE	The Feels	Sean Matsubara	Aug 25, 2022	3:18
8	WA DA DA Kep1er	FIRST IMPACT	Sean Matsubara	Aug 25, 2022	3:04
9	MANIAC Stray Kids	ODDINARY	Sean Matsubara	Aug 25, 2022	3:03
10	LOVE DIVE IVE	LOVE DIVE	Sean Matsubara	Aug 25, 2022	2:57

Playlist Generated Using Gower Distance

PUBLIC PLAYLIST

kpop jim Playlist Recs

Playlist Consisting of Songs Recommended by Data Mining Team 4's Program

Scott Matsubara • 25 songs, 1 hr 30 min

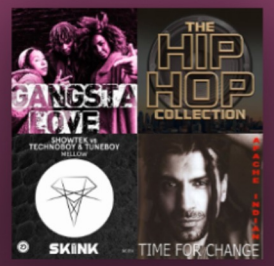
Scott Matsubara

Enhance

Custom order

#	Title	Album	Date added	
1	Chill Bro Dhanush	Pattas	1 hour ago	4:00
2	Stylish Thamizhachi Yuvan Shankar Raja, M.M.Manasi, Rubba Bend.Psycho Unit	Arrambam (Original Motion Picture Soundtrack)	1 hour ago	4:24
3	Wave ATEEZ	TREASURE EP.3: One To All	1 hour ago	3:24
4	MIC Drop BTS	Love Yourself 承 'Her'	1 hour ago	3:58
5	9 and Three Quarters (Run Away) TOMORROW X TOGETHER	The Dream Chapter: MAGIC	1 hour ago	3:32
6	FAKE LOVE - Rocking Vibe Mix BTS	Love Yourself 結 'Answer'	1 hour ago	3:58
7	The Eve EXO	THE WAR - The 4th Album	1 hour ago	2:56
8	Thiyagi Boys - From "Coffee With Kadhal" Yuvan Shankar Raja, Hiphop Tamizha	Thiyagi Boys (From "Coffee With Kadhal")	1 hour ago	3:26
9	Aathichoodi Vijay Antony, Dinesh	Tn07 Al 4777 (Original Motion Picture Soundtrack)	1 hour ago	4:11
10	CHEER UP TWICE	Page Two	1 hour ago	3:29
11	Thursday's Child Has Far To Go TOMORROW X TOGETHER	minisode 2: Thursday's Child	1 hour ago	3:31

Playlist Generated Using Euclidean Distance



Playlist

kpop jim Playlist Recs

Playlist Consisting of Songs Recommended by Data Mining Team 4's Program

Scott Matsubara • 35 songs, about 1 hr 45 min

#	Title	Album	Date added	
1	She's a Pro E Black Rob, Denaun	Gangsta Love	5 seconds ago	3:46
2	Cold Rock a Party E MC Lyte	The Hip Hop Collection	5 seconds ago	4:07
3	That Girl Apache Indian	Time For Change	5 seconds ago	3:24
4	Mellow Showtek, Technoboy, Tuneboy, TNT	Mellow	5 seconds ago	2:49
5	Chasing Colors (feat. Noah Cyrus) Marshmello, Ookay, Noah Cyrus	Chasing Colors (feat. Noah Cyrus)	5 seconds ago	3:15
6	Together As One The Pitcher	Together As One (feat. Sam LeMay)	6 seconds ago	3:06
7	When You're Gone Shawn Mendes	Autumn Vibes 2022	6 seconds ago	2:52
8	In Over My Head E grandson	20's Rock	6 seconds ago	3:18
9	My Songs Know What You Did In The Dark (Light Em Up) Fall Out Boy	Halloween 2022	6 seconds ago	3:06
10	No Stranger to Shame E Uncle Kracker	Human - Best Adult Pop Tunes	6 seconds ago	3:40
11	Bichotes Con Clase E Brray	Homecoming Latin Party	6 seconds ago	3:13

Conclusions

Recommendations:

- **Popularity based**: for new users that don't have listening history
- **Lyrics based**: when people are listening to one song, we can recommend others with a similar theme
- **Features based**: will recommend a new playlist, based on your current one

Our Recommendation Systems Value Add

1. **Increase User Retention**: Spotify can keep users engaged and coming back to the platform in order to build loyal user base

1. **Improved User Satisfaction**: This can lead to more positive brand image

1. **Increased Revenue**: If more users are on platform that can lead to increased revenue through advertising and subscription fees.

1. **Competitive Advantage:** By offering personalized recommendations that other platforms cannot match, Spotify can maintain a competitive advantage and attract new users.

1. **Switching Cost**: Because of our recommendation systems, more users may become attached to the platform, making it more difficult for the users to switch to another music platform. This can increase switching costs for users, as they may have to start from scratch with a new platform and build up their preferences.

Future Improvements

- Add more songs and genres
- User ratings dataset
- Days based recommendations
- Combine different models to a full scale end product
- Create a UI

Thank You!
Any Questions?

Closing Time

Semisonic

0:23 -4:34

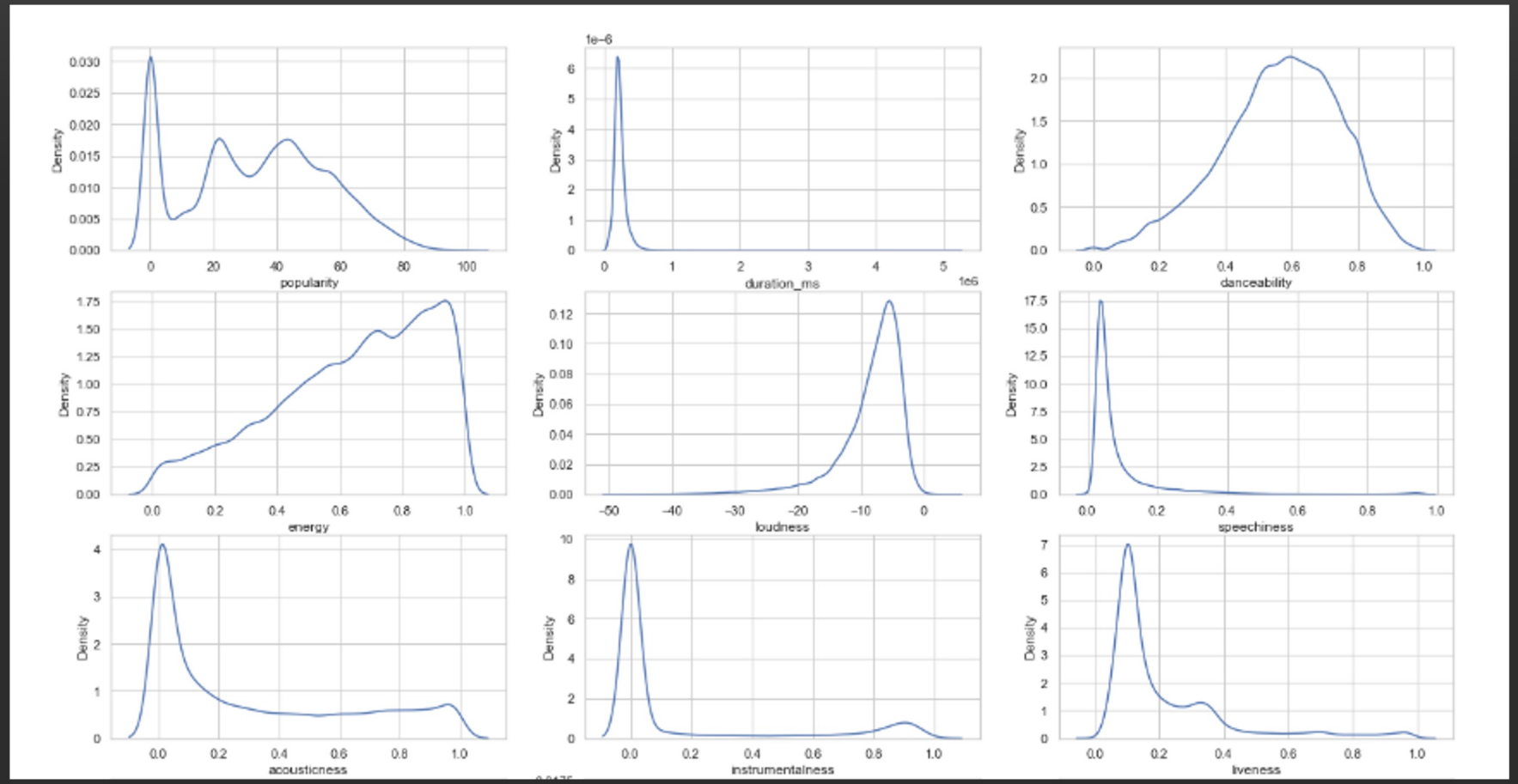
Appendix

Closing Time

Semisonic

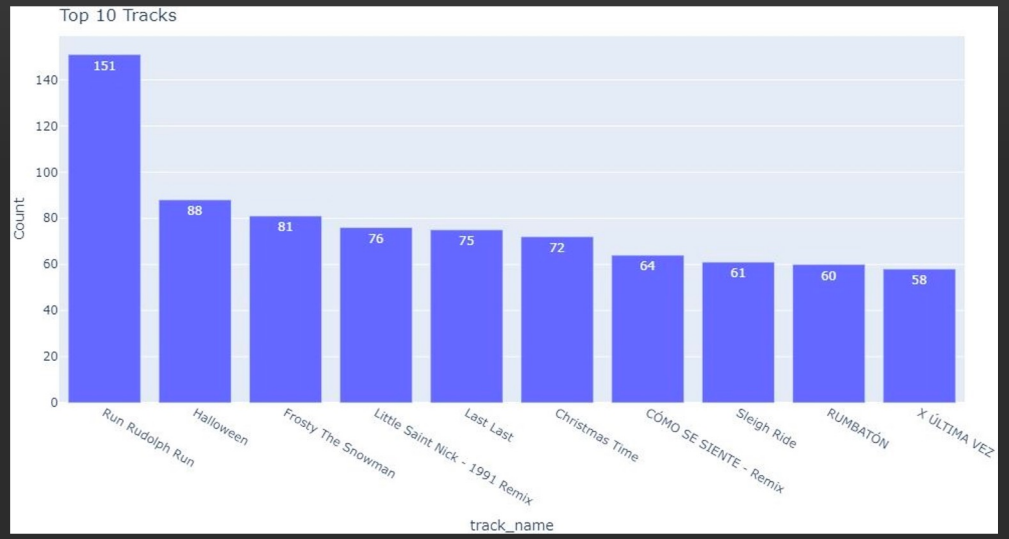
0:23 -4:34

Data Exploration: KDE

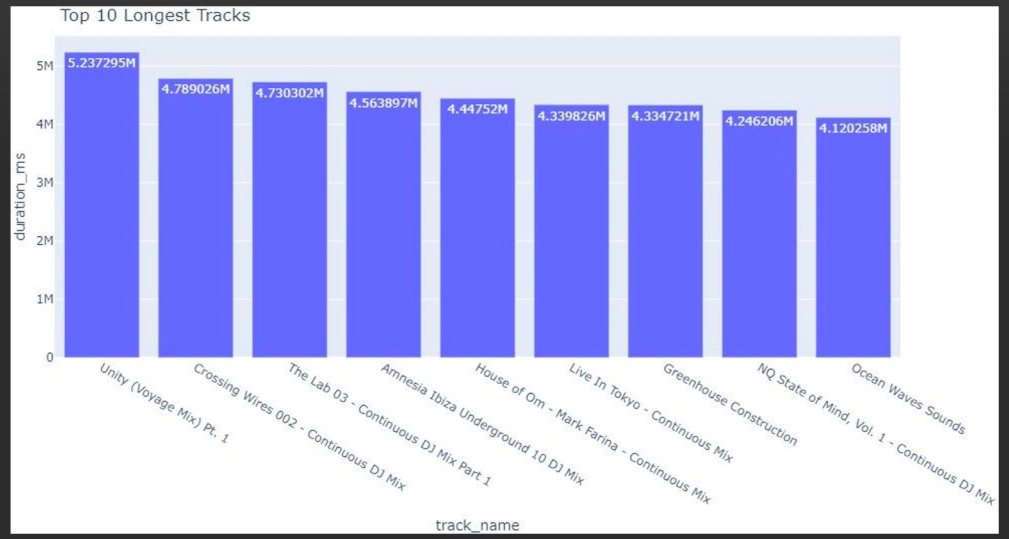


KDE Plots for All Numeric Features

Data Exploration

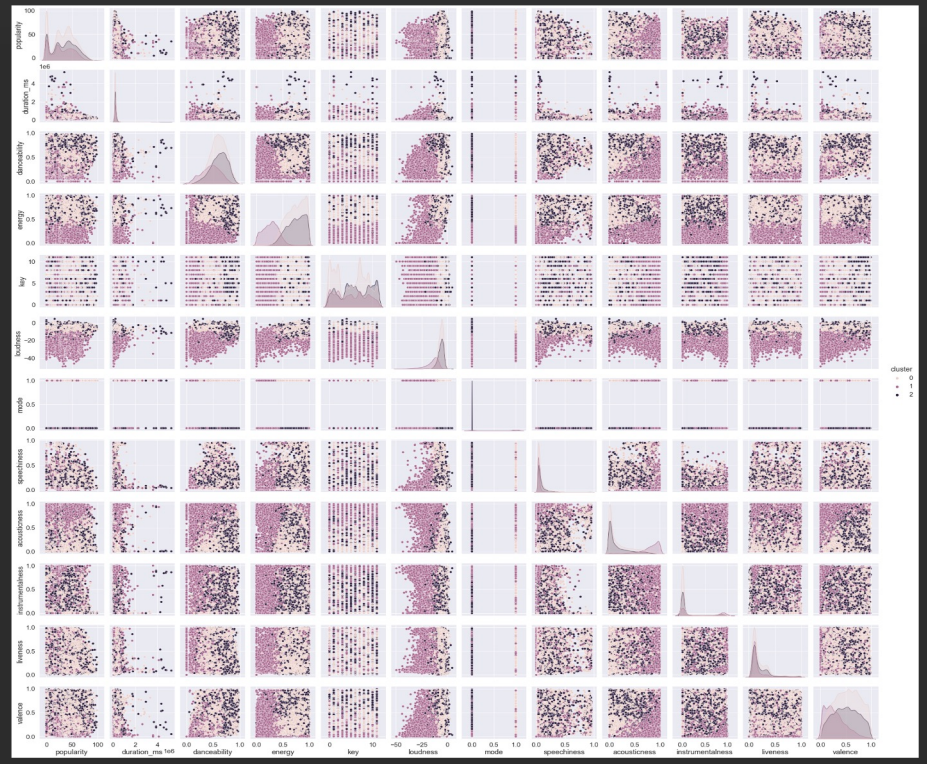


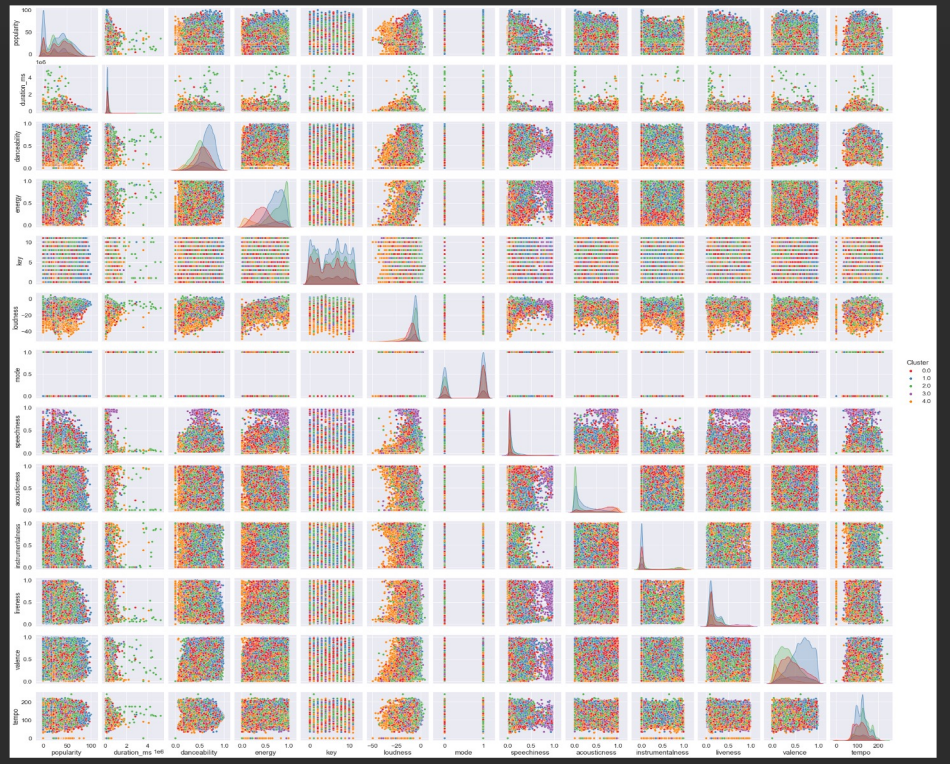
Top 10 most common tracks



Top 10 tracks with the longest duration

Data Visualization: Clustering





- Clustering using k-means (3 and 5 clusters)

Our Data

Dataset.csv (Observations: 89741 and Variables: 22**)**

The basic information of the song:

- String:
  - track_id, artists, album_name, track_name

The traits of the song:

- Numerical data which is calculated by Spotify algorithm:
  - popularity(0-100), danceability(0-1), energy(0-1), speechiness(0-1), acousticness(0-1), instrumentalness(0-1), liveness(0-1), valence(0-1)
- Numerical data:
  - duration_ms, loudness, tempo, time_signature
- Categorical:
  - explicit, key, mode, track_genre

The lyrics of the song:

- String