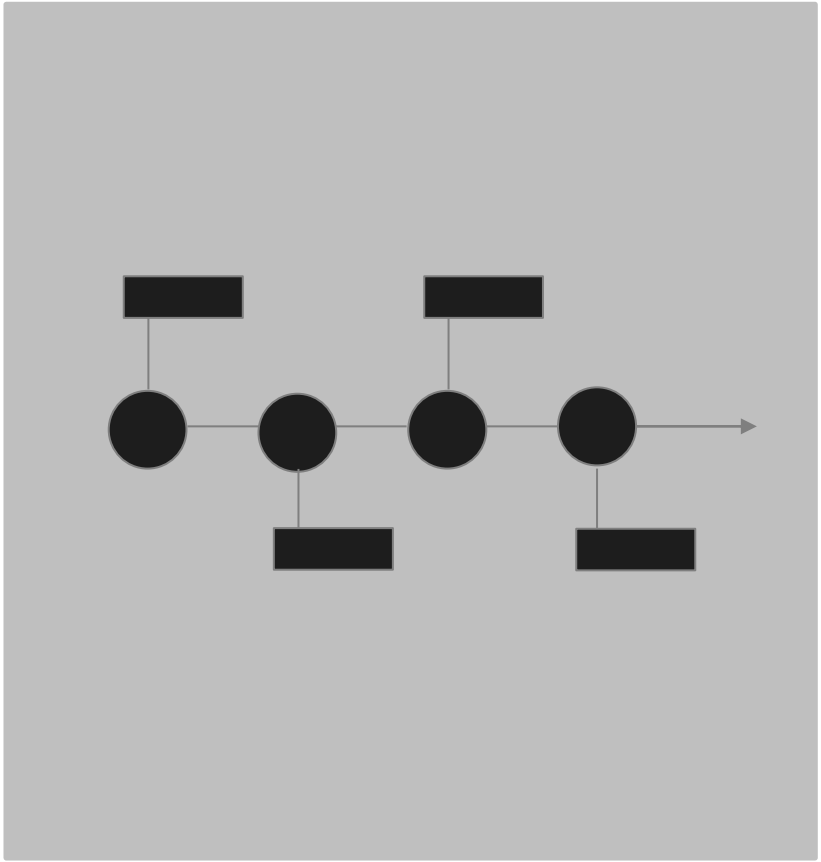
Music Recommendation Service

Group 2: Jason Ingram & Isaiah Martinez

Table of contents

01

Project Design



Project Description

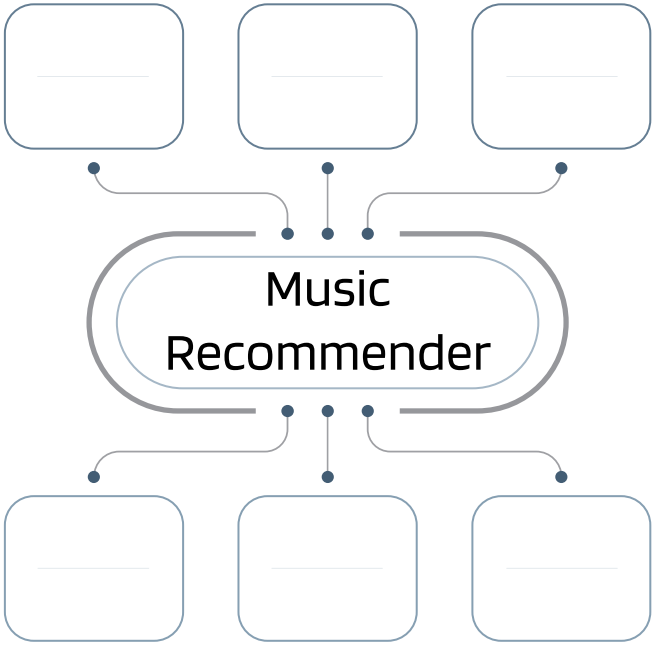
Music Recommendation Service:

- 1 Source for Music and User Data

Source for Data

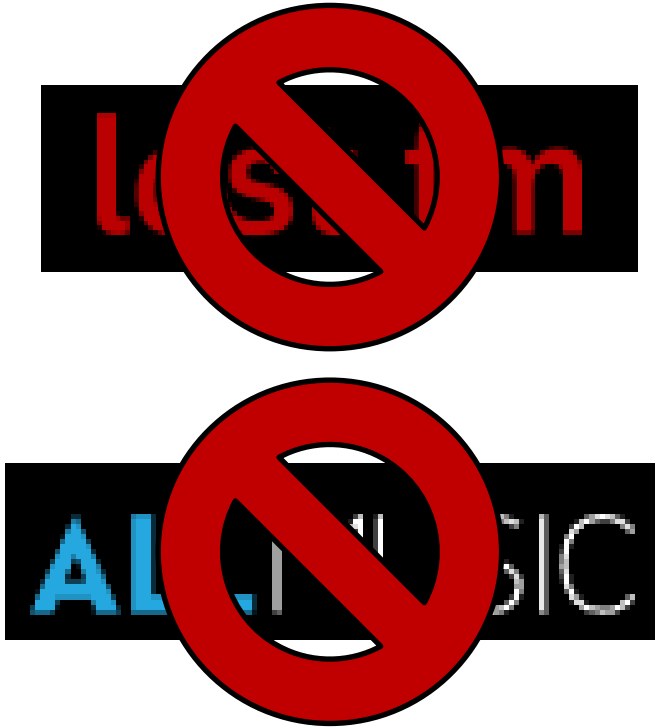
Project Design

Combined
Music and User Data



Why did we change the dataset used?

- Not easily usable:
  - Required manual cleaning or advanced automation techniques
  - Over 40,000 malformed entries out of a total 200,000 entries (~20%)
  - No relational data for the songs
  - Blacklisted from Last.fm and All Music



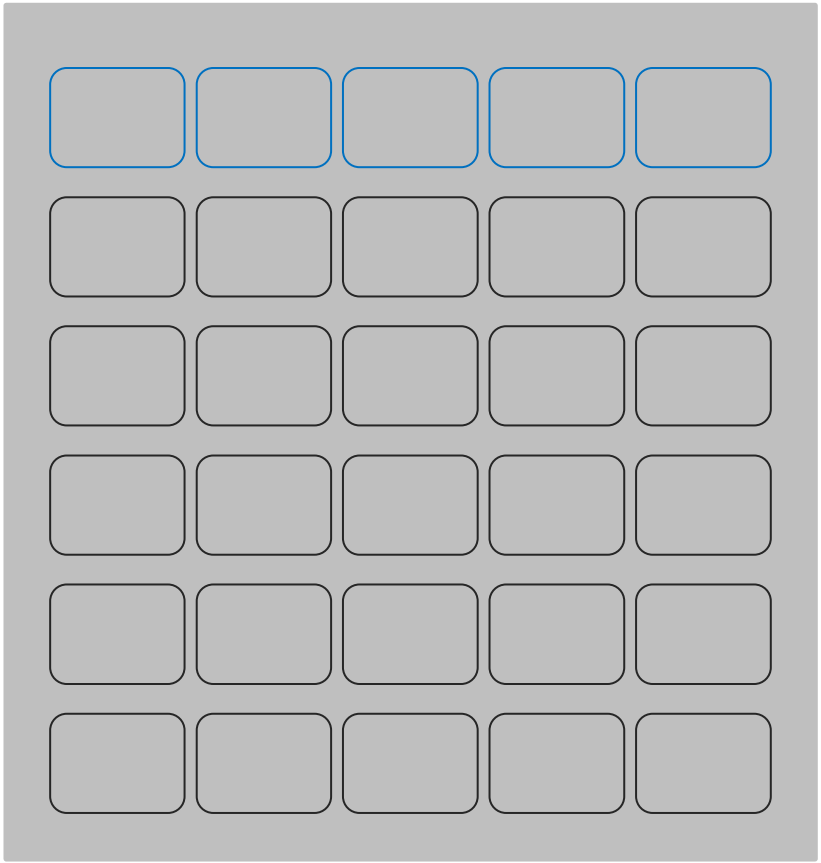
Previous data

	Song.Name	Song.Length	Album	Artist	Date.Released	Tags
1	Welcome	3:47	Age Of Aquarius	Villagers of Ioannina City	2019	stoner rock;psychedelic rock;progressive rock;folk;folk rock;r...
2	Age of Aquarius	8:09	Age Of Aquarius	Villagers of Ioannina City	2019	stoner rock;psychedelic rock;folk;folk rock;rock
3	Part V	10:01	Age Of Aquarius	Villagers of Ioannina City	2019	stoner rock;psychedelic rock;folk;folk rock;rock
4	Dance of Night	8:46	Age Of Aquarius	Villagers of Ioannina City	2019	stoner rock;psychedelic rock;progressive rock;folk;folk rock;r...
5	Arrival	2:26	Age Of Aquarius	Villagers of Ioannina City	2019	stoner rock;psychedelic rock;progressive rock;folk;folk rock;r...
6	Father Sun	7:19	Age Of Aquarius	Villagers of Ioannina City	2019	stoner rock;psychedelic rock;progressive rock;folk;folk rock;r...
7	Millennium Blues	8:18	Age Of Aquarius	Villagers of Ioannina City	2019	stoner rock;psychedelic rock;folk;folk rock;rock
8	Cosmic Soul	8:30	Age Of Aquarius	Villagers of Ioannina City	2019	stoner rock;psychedelic rock;progressive rock;folk;folk rock;r...
9	For the Innocent	5:51	Age Of Aquarius	Villagers of Ioannina City	2019	stoner rock;psychedelic rock;folk;folk rock;rock
10	Sparkle out of Black Hole	2:30	Age Of Aquarius	Villagers of Ioannina City	2019	stoner rock;psychedelic rock;progressive rock;folk;folk rock;r...
11	Kalesma	1:29	riza	Villagers of Ioannina City	2014	stoner rock;psychedelic rock;folk;folk rock;rock
12	Echoes	3:39	riza	Villagers of Ioannina City	2014	stoner rock;psychedelic rock;folk;folk rock;rock
13	Nova	7:37	riza	Villagers of Ioannina City	2014	stoner rock;psychedelic rock;folk;folk rock;rock
14	Jiannim	7:45	riza	Villagers of Ioannina City	2014	stoner rock;psychedelic rock;folk;folk rock;rock
15	Tabourla	Add lyrics on Musixmatch	riza	Villagers of Ioannina City	2014	stoner rock;psychedelic rock;folk;progressive rock;folk rock;r...
16	Krasi	6:47	riza	Villagers of Ioannina City	2014	stoner rock;psychedelic rock;folk;folk rock;rock
17	Tsakko	9:20	riza	Villagers of Ioannina City	2014	stoner rock;psychedelic rock;folk;folk rock;rock

Post-Cleaning for previous data

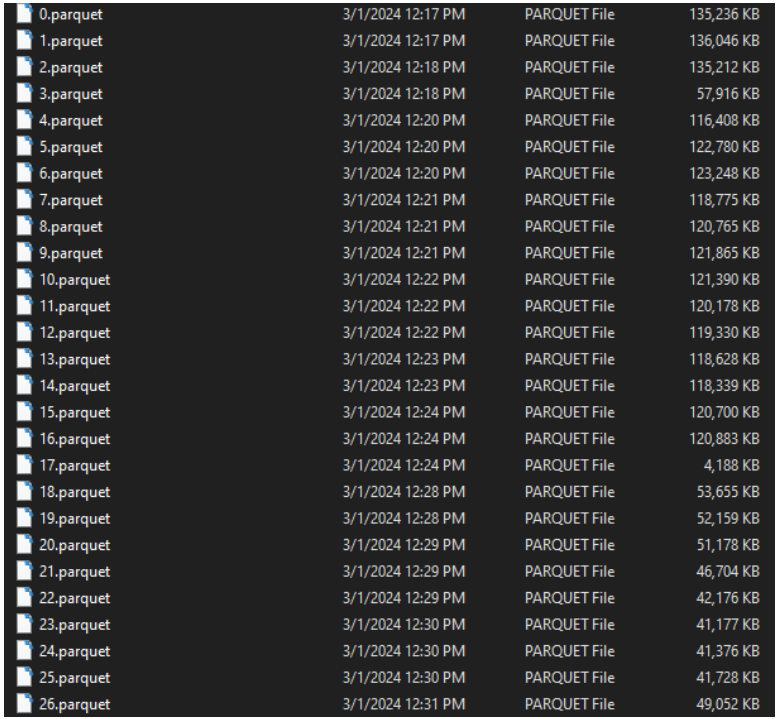
02

Data Visualization



Data Files

- Located in parquet Files from MusicBrainz
- Total of 789 parquet files with size of ~42GB
- First parquet has over 1 million entries



0.parquet	3/1/2024 12:17 PM	PARQUET File	135,236 KB
1.parquet	3/1/2024 12:17 PM	PARQUET File	136,046 KB
2.parquet	3/1/2024 12:18 PM	PARQUET File	135,212 KB
3.parquet	3/1/2024 12:18 PM	PARQUET File	57,916 KB
4.parquet	3/1/2024 12:20 PM	PARQUET File	116,408 KB
5.parquet	3/1/2024 12:20 PM	PARQUET File	122,780 KB
6.parquet	3/1/2024 12:20 PM	PARQUET File	123,248 KB
7.parquet	3/1/2024 12:21 PM	PARQUET File	118,775 KB
8.parquet	3/1/2024 12:21 PM	PARQUET File	120,765 KB
9.parquet	3/1/2024 12:21 PM	PARQUET File	121,865 KB
10.parquet	3/1/2024 12:22 PM	PARQUET File	121,390 KB
11.parquet	3/1/2024 12:22 PM	PARQUET File	120,178 KB
12.parquet	3/1/2024 12:22 PM	PARQUET File	119,330 KB
13.parquet	3/1/2024 12:23 PM	PARQUET File	118,628 KB
14.parquet	3/1/2024 12:23 PM	PARQUET File	118,339 KB
15.parquet	3/1/2024 12:24 PM	PARQUET File	120,700 KB
16.parquet	3/1/2024 12:24 PM	PARQUET File	120,883 KB
17.parquet	3/1/2024 12:24 PM	PARQUET File	4,188 KB
18.parquet	3/1/2024 12:28 PM	PARQUET File	53,655 KB
19.parquet	3/1/2024 12:28 PM	PARQUET File	52,159 KB
20.parquet	3/1/2024 12:29 PM	PARQUET File	51,178 KB
21.parquet	3/1/2024 12:29 PM	PARQUET File	46,704 KB
22.parquet	3/1/2024 12:29 PM	PARQUET File	42,176 KB
23.parquet	3/1/2024 12:30 PM	PARQUET File	41,177 KB
24.parquet	3/1/2024 12:30 PM	PARQUET File	41,376 KB
25.parquet	3/1/2024 12:30 PM	PARQUET File	41,728 KB
26.parquet	3/1/2024 12:31 PM	PARQUET File	49,052 KB

Some of the parquet files downloaded

Reading the Data

- Requires package to extract information (Arrow)

- 10 attributes:
  - release_name (Album Name)
  - recording_name (Song Name)
  - recording_mbid (used for server-side data processing
  - release_mbid (unique album id)
  - recording_msid (song id)
  - artist_credit_id (artist id)
  - artist_credit_mbids (server-side artists credited)
  - listened_at (YYYY-MM-DD HH:MM:SS)
  - user_id
  - artist_name

Package to install in R

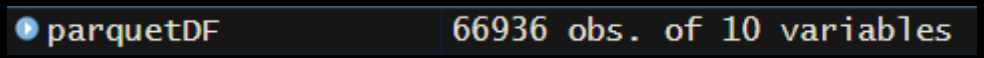
Data Visualization

	listened_at	user_id	recording_msid	artist_name	artist_credit_id	release_name	release_mbid	recording_name	recording_mbid
1	2006-11-29 13:19:10	16493	3fd94a93-a1e1-4847-8f73-b5f6595c4e36	Greg MacPherson Band	248984	Good Times Coming Back Again	8a673254-05bb-426e-9d9c-bd1fab1160b6	Numbers	22e96a0e-40a3-4
2	2006-11-29 13:52:16	8793	2002138e-1244-416d-afd3-e22962db13b8	Wolfgang Amadeus Mozart	11285	The World of Sacred Music	bf33ee03-ee6f-4439-8eee-e997179b7af3	Ave Verum Corpus	ac888a7e-a043-4
3	2006-11-29 13:59:42	6263	1a522e27-1e6a-4723-942f-803ccdcf5128	Japan	30345	Tin Drum	df4ddfb5-2db2-48c3-9679-b0c060c7fe7f	Ghosts	85f8df46-83bb-4
4	2006-11-29 13:55:42	5838	504f7a2c-8a04-416a-a484-d73759a82e49	Enigma	116	The Cross of Changes	b80b7a66-e51f-4687-82bf-3d77fe27ef18	Age of Loneliness (Carly's Song)	c146a166-725c-4
5	2006-11-29 14:04:29	1061	16dda417-2715-493a-a96c-757cc01e3c39	Paul Simon	1773	Graceland	74dd464c-d693-401f-9731-2a15223b4ad0	All Around the World or the Myth of Fingerprints (early versi...	0521d9fe-9268-4
6	2006-11-29 14:39:42	5838	504f7a2c-8a04-416a-a484-d73759a82e49	Enigma	116	The Cross of Changes	b80b7a66-e51f-4687-82bf-3d77fe27ef18	Age of Loneliness (Carly's Song)	c146a166-725c-4
7	2006-11-29 14:44:38	8115	3c02fde3-5ed9-4270-9f8b-c5c9a1ab77d4	Pixies	249	Surfer Rosa	9aebe9d4-18ac-4a43-9b00-19e2003aa076	I'm Amazed	62b1eff7-b615-4
8	2006-11-29 15:07:16	1061	16dda417-2715-493a-a96c-757cc01e3c39	Paul Simon	1773	Graceland	74dd464c-d693-401f-9731-2a15223b4ad0	All Around the World or the Myth of Fingerprints (early versi...	0521d9fe-9268-4
9	2006-11-29 15:14:45	6419	0a5e52bf-f3ae-4934-95fc-bd91648cc714	Partizani	*NA*	*NA*	*NA*	Pociva jezero v tihoti	*NA*
10	2006-11-29 15:26:37	4685	0bd598f8-52e0-43ea-8201-9fa454caa372	Red Hot Chili Peppers	12389	Stadium Arcadium	d9eff86a-d4c3-4b1f-bb2a-f2ed473318c6	Dani California	8ebb47f5-d7d5-4

Initial View of Data

> 60,000

Total listens within parquet file 17

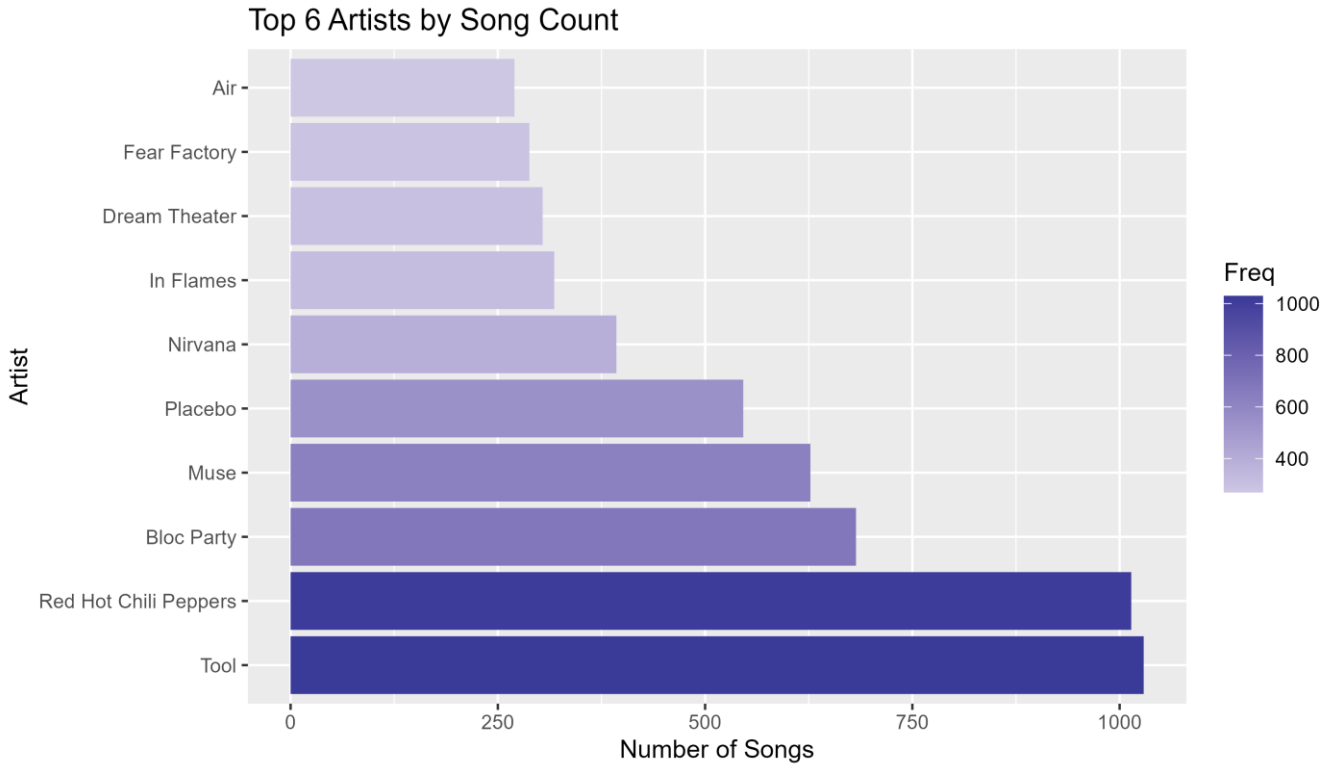


Unique Artists

```
 [975] "The Quantic Soul Orchestra"                    "Accelera Deck"
 [977] "Hecate"                                        "Side B"
 [979] "Émilie Simon"                                  "Cocteau Twins"
 [981] "Above & Beyond"                                "Stiff Little Fingers"
 [983] "The Residents"                                 "Patricia Elliott, Victoria Mallory, Harold Hastings"
 [985] "Baptiste Trotignon"                            "梶芽衣子"
 [987] "Tim Hart & Maddy Prior"                        "Cypress Hill"
 [989] "Tiësto"                                        "Local H"
 [991] "Ed's Redeeming Qualities"                      "The Bothy Band"
 [993] "The Beach Boys"                                "The Notorious B.I.G. feat. Diddy, Eminem, Obie Trice"
 [995] "Leningrad Cowboys"                             "Hocico"
 [997] "Lamb"                                          "Messer Chups"
 [999] "Scott Henderson, Steve Smith, Victor Wooten"   "Universal Principles"
 [ reached getOption("max.print") -- omitted 9503 entries ]
> length(unique(parquetDF$artist_name))
[1] 10503
```

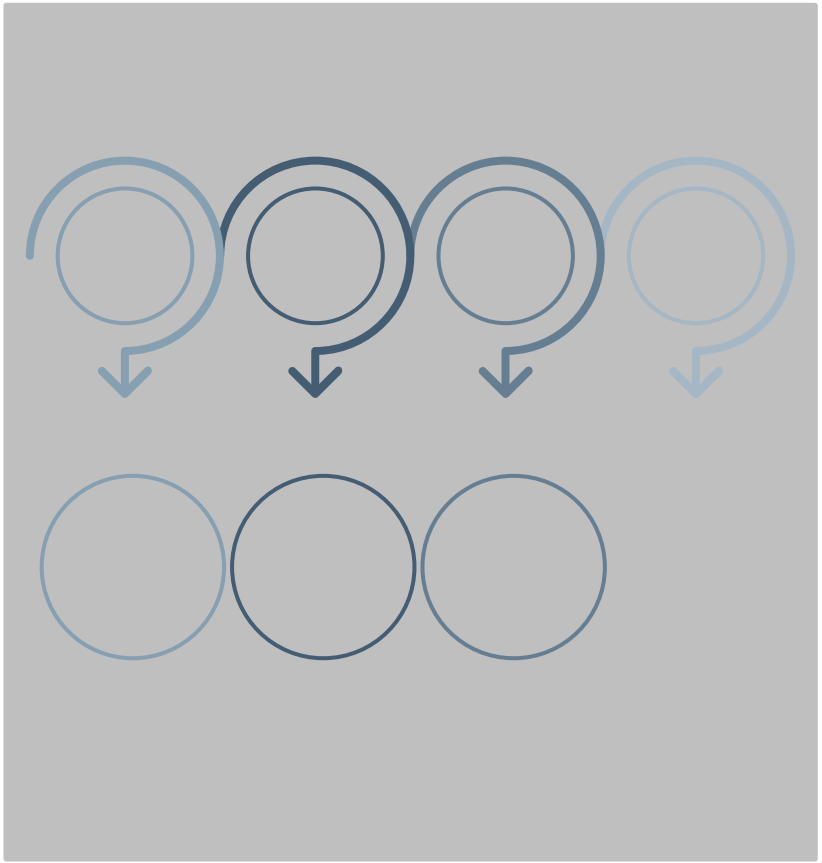
Over 10,000 artists obtained

Top Artists by Listens



03

Data Cleaning



Irrelevant Attributes

```
12 #recording_mbid, release_mbid, recording_msid are not needed:
13 #   recording_mbid - id used for the song in the db for data retrieval/storage
14 #                    not relevant for the scope of our project
15 #
16 #   release_mbid - unique id for each album that is released (eg. diff versions)
17 #                  unnecessary since each album will have a diff name
18 #
19 #   recording_msid - id for this particular song. Redundant since we will use
20 #                    song, album, and artist for identifying the song
21 #
22 #   artist_credit_id - id for the artist. Redundant since we will use the
23 #                      artist name
24 #
25 #   artist_credit_mbids - artists credited with release. Irrelevant since they
26 #                         don't match the artist_credit_id's => some other set of
27 #                         id's used internally for tracking
28
29 parquetDF <- parquetDF[,!(names(parquetDF) %in% c("recording_mbid",
30                "release_mbid", "recording_msid", "artist_credit_id",
31                "artist_credit_mbids"))]
```

Removal of attributes

Missing/NA Values

```r
#replace NA's
if (sum(is.na(parquetDF$user_id)) > 0) {
  #there exist NA's for user_id
  #drop these rows
  parquetDF <- parquetDF[!is.na(parquetDF$user_id),]
}
if (sum(is.na(parquetDF$artist_name)) > 0) {
  #there exist NA's for artist_name
  #rename to Unknown Artist
  parquetDF <- parquetDF %>% replace_na(list(artist_name = "Unknown Artist"))
}
if (sum(is.na(parquetDF$release_name)) > 0) {
  #there exist NA's for release_name
  #rename to Unknown Album
  parquetDF <- parquetDF %>% replace_na(list(release_name = "Unknown Album"))
}
if (sum(is.na(parquetDF$recording_name)) > 0) {
  #there exist NA's for recording_name
  #rename to Unknown Song
  parquetDF <- parquetDF %>% replace_na(list(recording_name = "Unknown Song"))
}
```

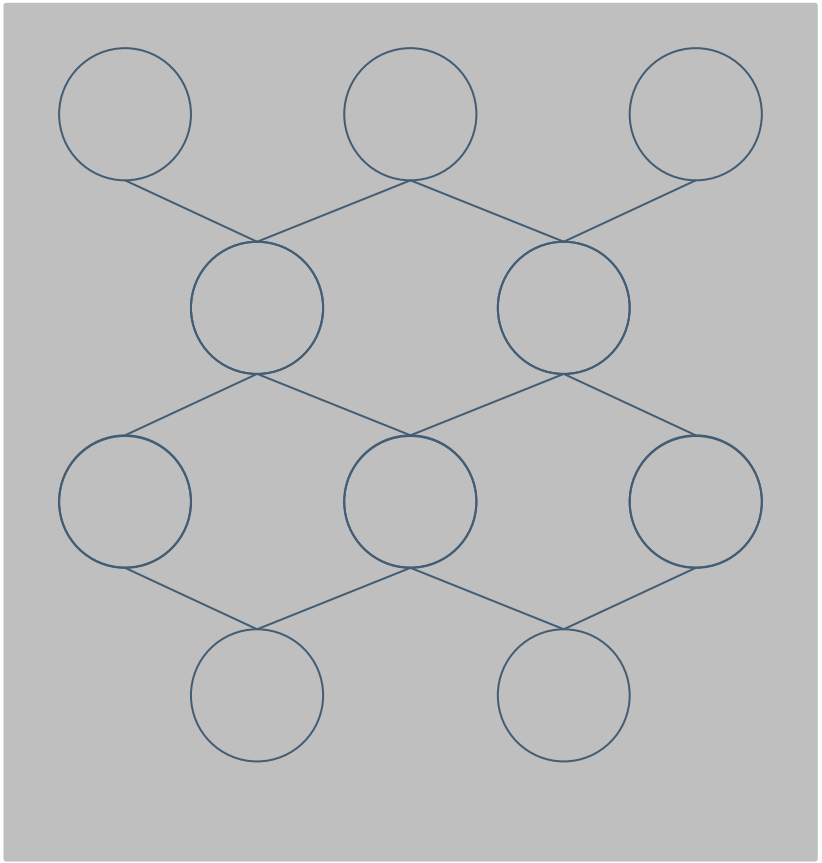
Replace Missing Artists with "Unknown Artist"

Replace Missing Album with "Unknown Album"

Replace Missing Song Name with "Unknown Song"

04

Data Selection



Feature Engineering

	listened_at
1	2006-11-29 13:19:10
2	2006-11-29 13:52:16
3	2006-11-29 13:59:42
4	2006-11-29 13:55:42
5	2006-11-29 14:04:29
6	2006-11-29 14:39:42
7	2006-11-29 14:44:38
8	2006-11-29 15:07:16
9	2006-11-29 15:14:45
10	2006-11-29 15:26:37

Before

Split `listened_at` into 2 different columns: `date` and `time`

date	time
2006-11-29	13:19:10
2006-11-29	13:52:16
2006-11-29	13:59:42
2006-11-29	13:55:42
2006-11-29	14:04:29
2006-11-29	14:39:42
2006-11-29	14:44:38
2006-11-29	15:07:16
2006-11-29	15:14:45
2006-11-29	15:26:37

After

Final Dataset

	user_id	artist_name	release_name	recording_name	date	time
1	16493	Greg MacPherson Band	Good Times Coming Back Again	Numbers	2006-11-29	13:19:10
2	8793	Wolfgang Amadeus Mozart	The World of Sacred Music	Ave Verum Corpus	2006-11-29	13:52:16
3	6263	Japan	Tin Drum	Ghosts	2006-11-29	13:59:42
4	5838	Enigma	The Cross of Changes	Age of Loneliness (Carly's Song)	2006-11-29	13:55:42
5	1061	Paul Simon	Graceland	All Around the World or the Myth of Fingerprints (early versi...	2006-11-29	14:04:29
6	5838	Enigma	The Cross of Changes	Age of Loneliness (Carly's Song)	2006-11-29	14:39:42
7	8115	Pixies	Surfer Rosa	I'm Amazed	2006-11-29	14:44:38
8	1061	Paul Simon	Graceland	All Around the World or the Myth of Fingerprints (early versi...	2006-11-29	15:07:16
9	6419	Partizani	Unknown Album	Pociva jezero v tihoti	2006-11-29	15:14:45
10	4685	Red Hot Chili Peppers	Stadium Arcadium	Dani California	2006-11-29	15:26:37

Questions?

Thanks for Listening