FindMeFM(working title)

Final Project: Group 5

- Dorna Abdi
- Manuel Elizaldi
- Christian Koehlinger
- Cailyn Miller
- Mike Prue
- Joshua Rhodes

Project Outline

Topic: Can machine learning predict what songs a user will enjoy based on the audio features of songs they like?

Reason for Topic: Interest in the potential uses of the Spotify API and obtainable data.

Data Source: Kaggle dataset - *Spotify Dataset 1922-2021, ~600k Tracks*

“Audio features of ~600k songs released in between 1922 and 2021”

Dataset created using Spotify API

Data Structure: Tracks

tracks.csv (audio features of tracks, 600k rows)

Primary:
- id (Id of track generated by Spotify)

Numerical:
- acousticness (Ranges from 0 to 1)
- danceability (Ranges from 0 to 1)
- energy (Ranges from 0 to 1)
- duration_ms (Integer typically ranging from 200k to 300k)
- instrumentalness (Ranges from 0 to 1)
- valence (Ranges from 0 to 1)
- popularity (Ranges from 0 to 100)
- tempo (Float typically ranging from 50 to 150)
- liveness (Ranges from 0 to 1)
- loudness (Float typically ranging from -60 to 0)
- speechiness (Ranges from 0 to 1)

Dummy:
- mode (0 = Minor, 1 = Major)
- explicit (0 = No explicit content, 1 = Explicit content)

Categorical:
- key (All keys on octave encoded as values ranging from 0 to 11, starting on C as 0, C# as 1 and so on…)
- timesignature (The predicted timesignature, most typically 4)
- artists (List of artists mentioned)
- artists (Ids of mentioned artists)
- release_date (Date of release mostly in yyyy-mm-dd format, however precision of date may vary)
- name (Name of the song)

Data Structure: Artists

artists.csv (popularity metrics of artists, 1.1M rows)

- id (Id of artist)

- name (Name of artist)

- followers (Total number of followers of artist)

- popularity (Popularity of given artist based on all his/her tracks)

- genres (Genres associated with this artist)

Data Structure: Dictionary of Artist to Artist Relationships

dict_artists.json (artists related artists, represented by ids)

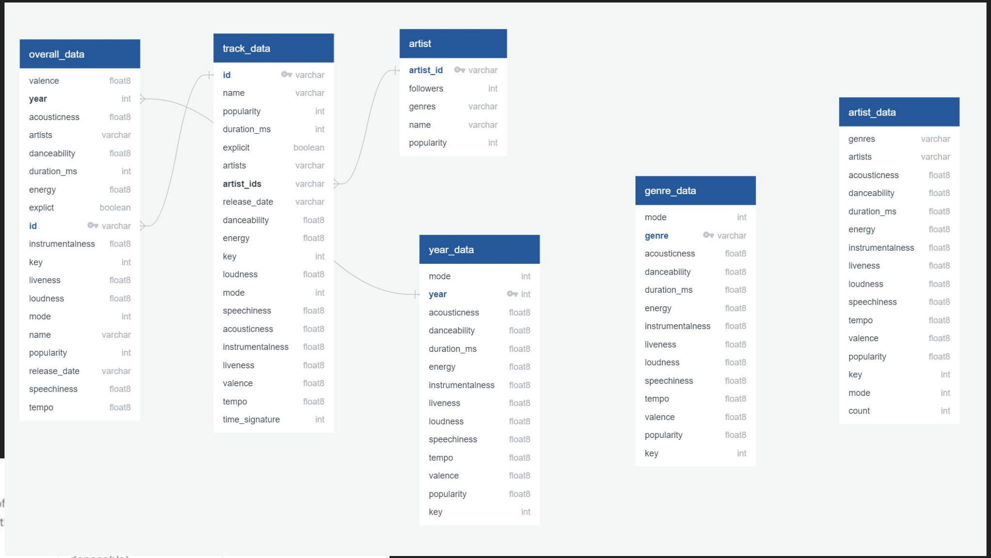
```
{
"any": [
"first",
"second",
"third",
…,
"nth"
],
"blank": [],
"first: [
"any",
"third",
"second
],
…
}
```

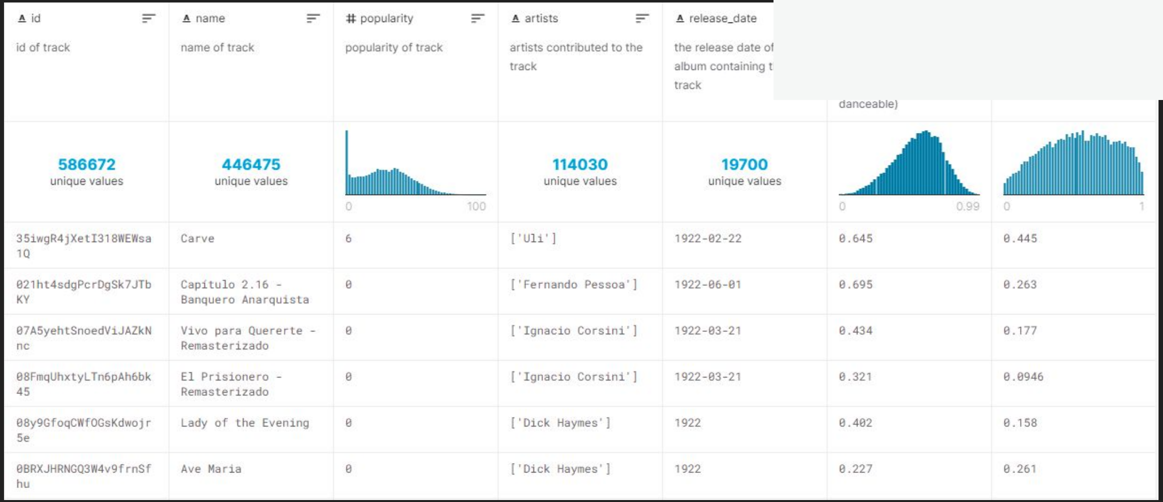
(The lists are sorted in descending order, so "first" is the most similar to "any" and "second" the second most and so on… Number of artists may vary from 0 to 20)

Questions This Data May Answer:

- Can we predict an artist's most popular song based on audio elements?
- Can we predict the most popular song off an album based on elements?
- Can we use audio elements to predict a song a listener would like?
- Can we use audio elements of an artist to predict another artist a listener would like?

Data Exploration Phase





id	name	popularity	artists	release_date		
id of track	name of track	popularity of track	artists contributed to the track	the release date of album containing t track	danceable)	
586672 unique values	446475 unique values	0 – 100	114030 unique values	19700 unique values	0 – 0.99	0 – 1
35iwgR4jXetI318WEWsa1Q	Carve	6	['Uli']	1922-02-22	0.645	0.445
021ht4sdgPcrDgSk7JTbKY	Capitulo 2.16 - Banquero Anarquista	0	['Fernando Pessoa']	1922-06-01	0.695	0.263
07A5yehtSnoedViJAZkNnc	Vivo para Quererte - Remasterizado	0	['Ignacio Corsini']	1922-03-21	0.434	0.177
08FmqUhxtyLTn6pAh6bk45	El Prisionero - Remasterizado	0	['Ignacio Corsini']	1922-03-21	0.321	0.0946
08y9GfoqCWfOGsKdwojr5e	Lady of the Evening	0	['Dick Haymes']	1922	0.402	0.158
0BRXJHRNGQ3W4v9frnSfhu	Ave Maria	0	['Dick Haymes']	1922	0.227	0.261

Analysis

Description of the analysis phase of the project

acousticness	danceability	energy	instrumentalness	liveness	loudness	popularity	speechiness	tempo
1.597267	-1.402608	-0.434803	0.527065	0.681875	-2.560544	0.401476	-0.384211	0.481201
-0.026886	-0.362256	0.666427	-0.519699	0.799534	-1.552361	1.193404	-0.071579	0.280969
3.212922	-1.633991	0.005412	1.118260	0.249359	-1.904048	-0.143154	1.033555	0.752712
-1.105168	-1.309363	-0.297432	0.131040	0.100094	-0.986246	-0.980184	-1.432502	0.749119
-1.037801	-0.886208	1.048483	1.097711	0.860369	-0.453503	0.875575	-1.344493	1.213830