Spotify Dataset 1922-2021, ~600k Tracks

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

[Yamac Eren Ay](https://www.kaggle.com/yamaerenay)

Version 15

Content

* tracks.csv (audio features of tracks, 600k rows)
* artists.csv (popularity metrics of artists, 1.1M rows)
* dict\_artists.json (artists related artists, represented by ids)

More on Spotify audio features, [click here](https://developer.spotify.com/documentation/web-api/reference/tracks/get-audio-features/)  
More on other Spotify track features, [click here](https://developer.spotify.com/documentation/web-api/reference/tracks/get-track/)

Structure

1. tracks.csv

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…)
* - time*signature* (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)

1. artists.csv

* - 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)

1. dict\_artists.json

{  
"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)

Acknowledgements

I would thank to [Spotify Developers](https://developer.spotify.com/) website for creating a very good documentation for Spotify Web API, [Spotipy](https://github.com/plamere/spotipy) (Python module for Spotify web servers) for making the requesting process more flexible, and I want to especially thank to Kaggle Community for allowing me the opportunity to share this project!

If you want to analyze different Spotify markets, please check my other datasets:

[Spotify Tracks Dataset](https://www.kaggle.com/yamaerenay/spotify-tracks-dataset-19222021), size: 3GB  
[Spotify Artists Dataset](https://www.kaggle.com/yamaerenay/spotify-artists-dataset-19222021), size: 1GB