**CS703 2.0 Data Understanding**

*Madeline Lin*

**Task 2.1: Gathering Data**

**Deliverable: Data Collection Report**

There are three datasets that I will be using for this project.

* Individual Dataset
* Global Weekly Top 100 Songs
* Generalized Dataset

I verified that I have acquired all the above three datasets, that I tested the data access process, and that the data exists. I loaded the data into Python and verified that the tools are compatible with the data.

*Outline Data Requirement*

* Individual Dataset
  + From Spotify, simply go to your Account - Privacy Settings - Download Your Data. You will be able to request the dataset there. I requested the dataset on January 15th and received my Account Data on January 18th.
  + The data mining goal is to predict the streaming time based on the audio features.
  + For Streaming History data, the time range is from January 16, 2022 to January 17, 2023. For Your Library data, the time range is up-to-date.
  + The original dataset from Spotify is in JSON format. I transformed into CSV.
* Global Weekly Top 100 Songs
  + This dataset is retrieved from Spotify Charts[[1]](#footnote-1).
  + The data mining goal is to find the tracks with the top 20 predicted streaming time. Those 20 tracks are regarded as the potential songs I will like. I will listen to these and determine whether they really match my taste.
  + Spotify updates the weekly top songs each Thursday. Thus, I downloaded the dataset of Week of January 12 (i.e. from January 6 to January 12) ([link](https://charts.spotify.com/charts/view/regional-global-weekly/2023-01-12)).
  + The format is in CSV.
* Generalized Dataset
  + This dataset is retrieved from Kaggle ([link](https://www.kaggle.com/code/vatsalmavani/music-recommendation-system-using-spotify-dataset/data)).
  + The data mining goal is to use the collaborative filtering method to find the song which has the closest audio features with the target song we’ve liked in our library.
  + The format is in CSV.

*Verify Data Availability*

I confirmed that the required data exist, and that they are available to be used.

*Define Selection Criteria*

* Individual Dataset: The original dataset received from Spotify is a personal dataset including information about user account, identity, payment, search queries, playlist, streaming history and library. For my project, I selected the relevant data, which are Streaming History and Your Library.
* Global Weekly Top 100 Songs: This dataset pulled from Spotify Charts is a dataset including information of rank, uri, artist, track, source, etc. of the global top 100 songs.
* Generalized Dataset: This dataset pulled from Kaggle is a dataset including the in formation of audio features, artist, name, uri, popularity of all tracks on Spotify from 1920 to 2020.

**Task 2.2: Describing Data**

**Deliverable: Data Description Report**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Individual Dataset | | | Global Weekly Top 100 Songs | Generalized Dataset |
| **Streaming History** | | **Your Library** |
| Source | Account | | | Spotify Charts | Kaggle |
| Formats | JSON🡪CSV | | | CSV | CSV |
| Number of Cases | 31,989 | 946 | | 100 | 170,653 |
| The Number of fields | 4 | 4\* | | 7\*\* | 18 |

\*The number of fields are to be updated after pulling the audio features of the track

\*\* The number of fields are to be updated after pulling the audio features and the predicted\_msPlayed of the track

|  |  |
| --- | --- |
|  | Description of the Fields |
| *Audio Features* | |
| danceability | Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable. |
| energy | Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. For example, death metal has high energy, while a Bach prelude scores low on the scale. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy. |
| acousticness | A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic. |
| valence | A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry). |
| instrumentalness | Instrumentalness predicts whether a track contains no vocals. “Ooh” and “aah” sounds are treated as instrumental in this context. Rap or spoken word tracks. |
| key | The estimated overall key of the track. Integers map to pitches using standard Pitch Class notation. E.g. 0 = C, 1 = C/D, 2 = D, and so on. If no key was detected, the value is -1. |
| liveness | This value describes the probability that the song was recorded with a live audience. A value above 0.8 provides strong likelihood that the track is live. |
| loudness | The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks. Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). Values typical range between -60 and 0 db. |
| mode | Mode indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived. Major is represented by 1 and minor is 0. |
| speechiness | Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks. |
| tempo | The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration. |
| duration\_ms | Duration\_ms is the duration of a track in milliseconds. |
| popularity | Spotify calculated the popularity of a track based on: 1) total streams of a track, 2) how recently a track has been played, 3) the frequency that a track has been played. |
| *General Features* | |
| artist | Artist represents the artist name of the track. |
| track | Track represents the track name of the song. |
| album | Album represents the name of the album that a track belongs to. |
| msPlayed | MsPlayed represents the streaming time of a track in milliseconds. |
| uri | The Spotify URI for the track. Uniform resource indicator is a link that you can find in the Share menu of any track, album, or Artist Profile on Spotify. |
| *Other Features* | |
| endTime | EndTime is the time when I end listening to the track. For example, 2020-01-16 23:31 means that I ended listening to a certain song at 23:31 on January 16, 2020. |
| rank | Rank represents the ranking of the tracks by a certain week. |
| source | Source is the music label of the track. |
| peak\_rank | Peak\_rank represents the highest rank of a track. |
| previous\_rank | Previous\_rank represents the previous rank of a track. |
| weeks\_on\_chart | Weeks\_on\_chart means how many weeks a track stays on chart. |
| year | Year indicates which year the track is released on Spotify. |
| released\_date | Released\_date indicates which date or year the track is released on Spotify. |

From my perspective, the data is suitable for my data-mining goals. The data includes the expected fields and the cases are sufficient based on the number of cases I’ve shown in the first table.

From Streaming History dataset, I can sum the streaming time of each track I’ve listened during this year (i.e. January 18, 2022 to January 17, 2023). Then I will pull the data of the streaming time to Your Library dataset. After that, I will run different statistical models based on the audio features of each track and choose the model with the best performance. Then, I will use the optimal model to test in Global Weekly Top 100 Songs dataset. Using the audio features of each track in Global Weekly Top 100 Songs dataset, I can calculate the predicted streaming time of each track and select the top 20 tracks, which are my potential liked songs. By listening, I will be able to tell how many tracks are my actual liked songs.

In Generalized Dataset, they include the audio features of each song, so I will be able to use the neighborhood collaborative filtering to use the similarity metrics method which calculates the distance using audio features in the dataset and find the neighbor songs which have relatively less distance.

**Task 2.3: Exploring Data**

**Deliverable: Data Exploration Report**

* Individual Dataset – Streaming History
* endTime
* Range: from 2022-01-18 to 2023-01-17 (a whole year)
* msPlayed
* Range: from 12 ms to 1,729,654 ms (i.e. 0.0002 min to ~28.9 min)
* Artist and track are categorical variables, which won’t be further analyzed in terms of the ranges/distributions/summaries
* Individual Dataset – Your Library
* Uri is used to pull the audio features of each track
* Artist, album and track are categorical variables, which won’t be further analyzed in terms of the ranges/distributions/summaries
* Global Weekly Top 100 Songs
* msPlayed
* Range: from 8,821,050 ms to 47,288,509 ms (i.e. ~147 mins to ~788 mins)
* Uri is used to pull the audio features of each track
* Rank related columns are not necessarily to be analyzed in terms of ranges/distributions/summaries
* Artist, track and source are categorical variables, which won’t be further analyzed in terms of the ranges/distributions/summaries
* Generalized Dataset
* Year, uri and release\_date are not necessarily to be analyzed in terms of ranges/distributions/summaries
* Artist and track are categorical variables, which won’t be further analyzed in terms of the ranges/distributions/summaries
* Summarized Table

|  |  |  |
| --- | --- | --- |
| Audio Features | Range | Distribution |
| Danceability | In theory, from 0 to 1  In our dataset, from 0 to 0.988 | N/A |
| Energy | from 0 to 1 | N/A |
| Explicit | 0 or 1 | Discrete |
| Acousticness | In theory, from 0 to 1  In our dataset, 0 to 0.996 | N/A |
| Valence | from 0 to 1 | N/A |
| Instrumentalness | from 0 to 1 | N/A |
| Key | from 0 to 11 | Discrete |
| Liveness | from 0 to 1 | N/A |
| Loudness | In theory, from -60 to 0  In our dataset, -47.046 to -0.671 | N/A |
| Mode | from 0 to 1 | Discrete |
| Speechiness | In theory, from 0 to 1  In our dataset, from 0 to 0.97 | N/A |
| Tempo | from 0 to 215.918 | N/A |
| Duration\_ms | From 10,371 ms to 2,525,293 ms  (i.e. ~0.173 min to ~42.088 min) | N/A |
| Popularity | From 0 to 100 (integer) | Discrete |

**Task 2.4: Verifying Data Quality**

**Deliverable: Data Quality Report**

I have checked these datasets and confirmed that the data quality can support the goals of the project.

In these datasets, as I have roughly examined all the rows and calculated the ranges of the numerical variables. I do not think there are any major quality issues. There may be some minor quality issues which I may encounter during my data modeling phases but I think I will be able to figure them out. Under the worst circumstance where I need to change the project goals or plans, I will consult with the professor in advance if there are any serious quality issues with no adequate solutions.

1. https://spotifycharts.com [↑](#footnote-ref-1)