**CS703 3.0 Data Preparation**

*Madeline Lin*

**Task 3.1: Selecting Data**

**Deliverable: Data Rationale Report**

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

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

In the Data Understanding phase, the three datasets were collected and verified. Here I will specify which portions of the data collected will be used for data mining.

My data mining goal is exploring the behind algorithm of Spotify’s recommendation system using two methods:

1. Content-based filtering method – building a statistical model with the best performance
2. Collaborative filtering method – finding the variable with the least distance of the target variable

* Individual Dataset
* Spotify sent a package of my data as shown below. However, only two datasets are related to my data mining goal.
* Rational For Inclusion and Exclusion

|  |  |  |
| --- | --- | --- |
| (Sub) Dataset | Inclusion/Exclusion | Rationale |
| Follow | Exclusion | This includes information of:  1) The number of followers the account has;  2) The number of other accounts this account is following;  3) The number of other accounts this account is blocking.  The information is not related to my data mining goal. |
| Identifiers | Exclusion | This includes the identifier type and identifier value information. It shows how I register for Spotify (e.g. email or phone number).  The information is not related to my data mining goal. |
| Identity | Exclusion | This includes my account name, the profile image and whether I am a tastemaker, and whether I am verified.  The information is not related to my data mining goal. |
| Inferences | Exclusion | This includes my interests and preferences based on my usage of the Spotify service. Spotify uses this data to identify a list of market segments with which I am currently associated. Depending on my settings, Spotify may serve interest-based advertising to me within the Spotify service.  The information is not related to my data mining goal. |
| Marquee | Exclusion | In the Spotify app I might periodically receive messages about new releases by artists and creators. The datasets regarding ‘marquee’ relate to these messages. They are based on my activity in Spotify app to be as relevant to me as possible. The information is not related to my data mining goal. |
| Payments | Exclusion | This includes the information of my subscription payment, which is not related to my data mining goal. |
| Playlist | Exclusion | This includes the information about my playlist. It is how I group the similar songs into one playlist, but not related to recommendation system.  The information is not related to my data mining goal. |
| Search Queries | Exclusion | This includes a list of search made, such as the data and time the search was made; type of device/platform used (i.e. iOS, desktop).  The information can be indirectly related to my music taste and preference. For example, if I like a song, I may search it and save it in my library. However, there may also be a possibility that I search a song because I see its name from an ad or some friend recommends me. I will not necessarily like the song. Thus, I won’t include this dataset. |
| Streaming History 0  Streaming History 1  Streaming History 2 | Inclusion  (These three datasets are to be merged) | This includes the items (e.g. songs, videos, and podcasts) listened to or watched in the past year, including the data and time when the stream ended in UTC format, name of the creator, name of the item, msPlayed.  This information correlates to my music taste and preference, and it is important for me to achieve my data mining goal. |
| User Address | Exclusion | This includes the information of the user address.  The information is not related to my data mining goal. |
| User Data | Exclusion | This includes the information of username, email, country, whether it is created from Facebook, Facebook user ID, birthdate, gender, postal code, mobile number, mobile operator, mobile brand, creation time.  The information is not related to my data mining goal. |
| Your Library | Inclusion | This includes a summary of the content saved in Your Library (e.g. songs, episodes, shows, etc.), including the entity names, creators, URIs. The songs in my library can represent my music taste and preference. They are related to my data mining goal. |

\*Further information of each dataset mentioned above can be found [here](https://support.spotify.com/us/article/understanding-my-data/)[[1]](#footnote-1).

* Global Weekly Top 100 Songs
* This dataset is retrieved from Spotify Charts[[2]](#footnote-2).
* This is a single dataset which I will include. The rationale for inclusion is that this dataset serves as a test dataset after I determine the optimal prediction model based on my individual dataset. This dataset is crucial for my data mining goal.
* Generalized Dataset
* This dataset is retrieved from Kaggle ([link](https://www.kaggle.com/code/vatsalmavani/music-recommendation-system-using-spotify-dataset/data)).
* This is a single dataset which I will include. The rationale for inclusion is that this dataset is used for the collaborative filtering method. I will apply neighborhood collaborative filtering to use the similarity metrics method by calculating the distance using all audio features available in this dataset and find the neighbor songs which have closet distance. This dataset is crucial for my data mining goal.

As discussed in Task 2.1, I have checked these three datasets and confirmed the data quality can support the data mining goals of the project and there are no major technical issues being identified.

**Task 3.2: Cleaning Data**

**Deliverable: Data Cleansing Report**

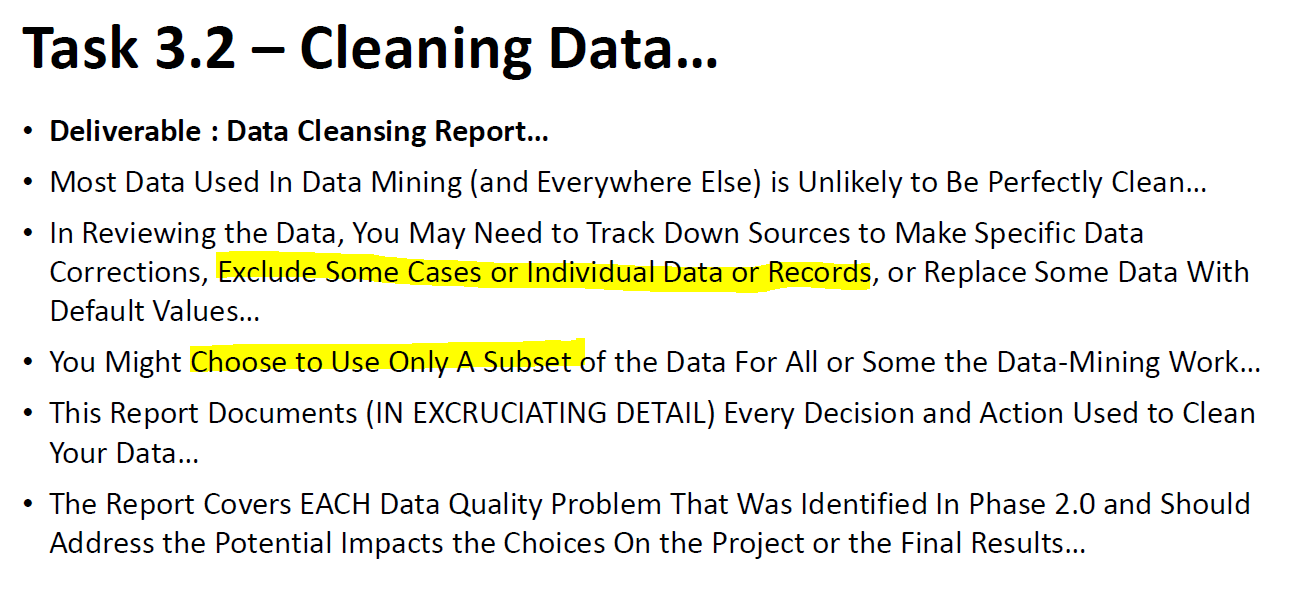
From my perspective, these three datasets are decently clean as I performed preliminary screens in Data Understanding phase. I did not track down sources to make specific data corrections, replace some data with default values. However, I excluded some cases or individual data or records, chose to use only a subset of the data for all or some the data-mining work as well as other steps.

Prof feedbacks: really should review all the data…you never know if later data is clean…

Answer: I agree. However, I guess I meant I did the preliminary screens during my Data Understanding phase so the datasets had been decent clean already before I started in Data Preparation phase.

Why “excluded some cases or individual data or records”?

Answer: this is actually retrieved from the CRISP-DM guide. What I was referring were I removed some unnecessary rows. For instance, for the individual dataset – streaming history, I removed rows which are 2022-01-16 and 2022-01-17, so the time range is the exact whole year. Hope it clarifies!



As discussed in Task 2.1, there should not be much data quality problems for these three datasets. My choices of how I clean the data are consistent with my data mining goals and final outcome of this project.

Below documents in excruciating detail about every detail and action used to clean my data.

* Individual Dataset – Streaming History
* Merged Streaming History 0, Streaming History 1 and Streaming History 3 into Streaming History
* Removed rows which are 2022-01-16 and 2022-01-17, so the time range is the exact whole year (i.e. 2022-01-18 ~ 2023-01-17)
* Removed the rows of 0 msPlayed
* Renamed the columns for consistency throughout three datasets
  + Column1.endTime 🡪 endTime
  + Column1.artistName 🡪 artist
  + Column1.trackName 🡪 track
  + Column1.msPlayed 🡪 msPlayed
* Individual Dataset – Your Library
* Chose a subset of the data - Only selected tracks and removed other rows (i.e. albums, shows, episodes, bannedTracks, artists, bannedArtists)
* Removed Name column
* Originally, the dataset includes a column (i.e. Name) which indicates tracker, albums, shows, episodes, etc. As I only kept tracks column, it would not be necessary to keep the Name column as all are tracks.
* Renamed the columns for consistency throughout three datasets
* Value.artist 🡪 artist
* Value.album 🡪 album
* Value.track 🡪 track
* Value.uri 🡪 URI
* Global Weekly Top 100 Songs
* Removed rows from 102 to 201
  + Originally, the dataset includes 200 pieces Global Top songs. As I only wanted first 100 songs as a test dataset, I removed rows from 102 to 201 and only kept the first top 100 songs.
* Renamed the columns for consistency throughout three datasets
* artist\_names 🡪 artist
* track\_name 🡪 track
* streams 🡪 msPlayed
* duration 🡪 duration\_ms
* Generalized Dataset
* Removed explicit column
  + For the above first two datasets, I need to pull the audio features of each song, which I will discuss further in the Constructing Data section. Spotify Web API[[3]](#footnote-3) allows us to be able to retrieve the audio features of a track based on the track URI. I found out that there is no “explicit” audio feature available in the page of Get Track’s Audio Features[[4]](#footnote-4). However, in generalized dataset, there is “explicit” audio feature for each song. In order to keep consistency throughout three datasets, I decided not to include the “explicit” attribute.
* Renamed the columns for consistency throughout three datasets
* Name 🡪 track
* Id 🡪 URI

**Task 3.3: Constructing Data**

**Deliverable 1: Data Attribute Report**

* Individual Dataset – Streaming History
* Created Sum\_msPlayed attribute
  + The Streaming History dataset serves the purpose of building the Your Library dataset. Specifically, I need to create a new field (i.e. Sum\_msPlayed column) of each track by using a pivot table. After that, I will use V lookup formula to transfer the Sum\_msPlayed data of each track to the Your Library dataset.
* Individual Dataset – Your Library
* There will be a new field (i.e. Sum\_msPlayed column) in the Your Library dataset.
* The Sum\_msPlayed column will be simplified as the msPlayed column moving forward.
* After transferring the msPlayed data, there are 46 pieces of track which does not have the data (i.e. #N/A) based on the Vlookup formula. I conducted a manual data validation screening and hard code the data. The reasons of #N/A include: 1) There is a bit nuance of the name of the track between Streaming History and Your Library; 2) I did not listen to some track for the period from January 18, 2022 to January 17, 2023.
* Pulled the audio features data and created these attributes
  + Spotify Web API is an interface that programs can use to retrieve and manage Spotify data over the Internet. In my case, I need to retrieve the audio features of each track. From the page of Get Track’s Audio Features, I am able to retrieve these by using the track’s URI. However, it would not be ideal if I manually retrieved the audio feature data for all the 946 tracks from my library. Based on public search, I used Python to perform audio feature data scraping from Spotify Web API to pull the data for all 946 tracks.
* So far, there will be a total of 17 attributes with 13 new attributes added in Your Library dataset.

Additionally, for the simplification, the Individual Dataset – Your Library will just be referred as Individual Dataset as we do not need Streaming History moving forward.

* Global Weekly Top 100 Songs
* Similar as Individual Dataset, I pulled the data of audio features of the top 100 tracks.
  + To create pred\_msPlayed attribute
    - Moreover, after determining the optimal prediction model, I will calculate the predicted msPlayed for the top 100 tracks and put them in descending order based on the predicted msPlayed. Then, I will be able to attain the 20 tracks with the top 20 predicted msPlayed. 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. So, I will add a new field – pred\_msPlayed.
* So far, there will be a total of 22 attributes with 13 new attributes added in this dataset.
* Generalized Dataset
* This dataset has included all the attributes that serve for my data mining goal. Therefore, there will be no further construction needed.

**Deliverable 2: Data Generation Report**

Based on my current standpoint, ~~I do not think there will be any new rows constructed.~~ there are no new rows which should be constructed. All the datasets include sufficient number of records.

Prof feedback: You don’t ‘think’ or there are no new ones?...

Answer: There are no new ones…

**Task 3.4: Integrating Data**

**Deliverable: Merged Data Set**

* Individual Dataset
* Previous integration performed.
  + The original streaming history data that Spotify sent are in three separate dataset (i.e. Streaming History 0, Streaming History 1, Streaming History 2). I guess the reason why Spotify did that is due to the size of the dataset as the whole streaming history includes a user’s whole year streaming history. These three streaming history datasets are merged as I’ve mentioned in the Cleaning Data section. Also, I have performed some data constructing steps and transferred the data to Your Library, so there is no further step needed for integration.
* Global Weekly Top 100 Songs
  + This is a single dataset. No integration needed.
* Generalized Dataset
  + This is a single dataset. No integration needed.

**Task 3.5: Formatting Data**

**Deliverable: Final Formatted Dataset**

So far, I have made all three datasets consistent in terms of the attribute naming, data format (i.e. all csv) and so forth. Based on my understanding, there is no further reformatting action needed at this point. These three datasets are ready for modeling.

Below is an overview of what these three datasets will look like.

* Individual Dataset
* 946 x 17 matrix (946 pieces of tracks, 17 attributes for each track)

|  |
| --- |
| Attributes |
| artist |
| album |
| track |
| URI |
| danceability |
| energy |
| key |
| loudness |
| mode |
| speechiness |
| acousticness |
| instrumentalness |
| liveness |
| valence |
| tempo |
| duration\_ms |
| msPlayed |

* Global Weekly Top 100 Songs
  + 100 x 22 matrix (100 pieces of tracks, 17 attributes for each track)

|  |
| --- |
| Attributes |
| rank |
| URI |
| artist |
| track |
| source |
| peak\_rank |
| previous\_rank |
| weeks\_on\_chart |
| msPlayed |
| danceability |
| energy |
| key |
| loudness |
| mode |
| speechiness |
| acousticness |
| instrumentalness |
| liveness |
| valence |
| tempo |
| duration\_ms |
| pred\_msPlayed |

* Generalized Dataset
  + 170,653 x 18 matrix (170,653 pieces of tracks, 18 attributes for each track)

|  |
| --- |
| Attributes |
| year |
| artists |
| track |
| URI |
| release\_date |
| danceability |
| energy |
| key |
| loudness |
| mode |
| speechiness |
| acousticness |
| instrumentalness |
| liveness |
| valence |
| tempo |
| duration\_ms |
| popularity |

1. Additionally, the information of Identifiers, Identity, Inferences, User Address and User Data are based on my review of the data. The information of Marquee is based on my inquiry of Spotify Support team. The rest are from Understanding my data. [↑](#footnote-ref-1)
2. https://spotifycharts.com [↑](#footnote-ref-2)
3. https://developer.spotify.com/documentation/web-api/ [↑](#footnote-ref-3)
4. https://developer.spotify.com/console/get-audio-features-track/ [↑](#footnote-ref-4)