**Capstone Project 1 – Data Wrangling**

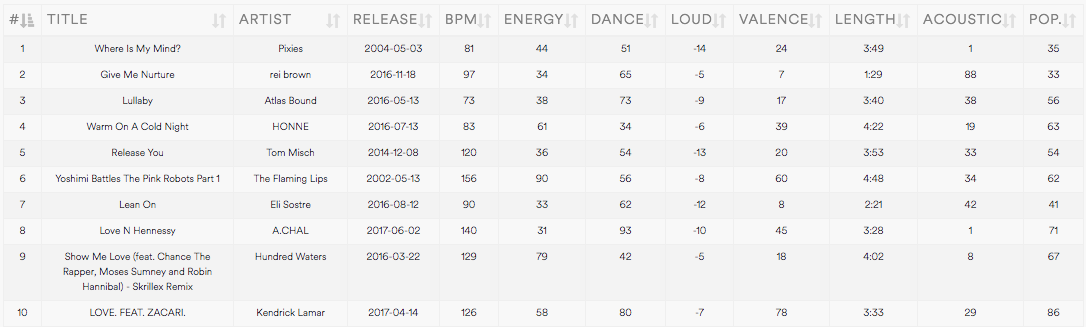
**Data Acquisition**

The dataset includes three Excel files:

1. Liked playlist (‘LIKES.xlsx’) – (630 rows/songs, 11 columns)
2. Disliked playlist (‘DISLIKES’.xlsx’) – (537 rows/songs, 11 columns)
3. Today’s Top Hits playlist (‘TOP\_HITS.xlsx’) – (125 rows/songs, 11 columns)

I compiled the ‘Liked’ and ‘Disliked’ playlists myself. I put together the ‘Today’s Top Hits’ playlist from three of Spotify’s playlists: ‘Global Top 50’, ‘United States Top 50’, and ‘United States Viral 50’. There was some overlap between Spotify’s three playlists so that’s why there are only 124 songs and not 150 songs in the ‘Today’s Top Hits’ playlist.

After creating these three playlists, I used a web app from Echo Nest called [Sort Your Music](http://sortyourmusic.playlistmachinery.com/index.html)1 that pulls the audio features of each song from Spotify. Sort Your Music pulls 8 audio features from Spotify and generates a neat table of songs and the values of their audio features (example below).



Once I have these tables generated, I copied the tables to a spreadsheet and saved them as Excel files. I then loaded the files into Jupyter notebook as a data frame using Pandas.

liked = pd.read\_excel('LIKES.xlsx')

disliked = pd.read\_excel('DISLIKES.xlsx')

tophits = pd.read\_excel('TOP\_HITS.xlsx')

The limitation of the dataset is that it only pulls 8 out of the 13 audio features of a song. Descriptions for each audio feature can be found on [Spotify’s API](https://developer.spotify.com/web-api/get-audio-features/)2.

**Data Wrangling**

Since the Sort Your Music web app generated clean tables with the appropriate column headings, data wrangling was minimal. After inspecting the dataframes, the only issues I found were some null values and the column name ‘POP.’ containing a period.

Removing null values:

liked = liked.dropna().reset\_index(drop=True)

disliked = disliked.dropna().reset\_index(drop=True)

tophits = tophits.dropna().reset\_index(drop=True)

* Liked playlist: removed 10 rows with null values (from 630 to 620 rows)
* Disliked playlist: removed 8 rows with null values (from 537 to 529 rows)
* Today’s Top Hit playlist: removed 1 row with null values (from 125 to 124 rows)
* I also reset the index in each dataframe

Removing the period in the ‘POP.’ column:

liked = liked.rename(columns=lambda x: x.replace('.', ''))

disliked = disliked.rename(columns=lambda x: x.replace('.', ''))

tophits = tophits.rename(columns=lambda x: x.replace('.', ''))

The full python code of my data gathering is available at Part1\_DataWrangling.ipynb