Capstone Project: Data Wrangling

Introduction:

For my capstone project, I am doing an analysis on Spotify's "Top 100 Tracks of 2017" playlist. This playlist is a Spotify featured playlist which includes the top 100 streamed songs for the year 2017 in the United States. I will also use the worldwide "Top 100 Tracks of 2017" playlist. Might be interesting to compare how the list of songs differs on each of these lists.

Data Wrangling Steps:

For the Data Wrangling process, I used Spotify to find the 2 playlists I will be using for this analysis. I used the Spotify API and their documentation to understand all the data features available for download from each song, artist, & playlist. Once I determined the specific data and information I wanted to use for my analysis, I used a few YouTube videos, Spotipy, Sublime Text, & the Mac Terminal to prepare the necessary code to download the data I wanted to download for this analysis.

The way that the Spotify API is setup to download data, it was necessary to download this data in the following steps:

> 1) Using the Sublime Text editor & help from a few YouTube tutorials on how to use Spotipy, I wrote code & downloaded the artist(s) name, track name, track id #, & album name information as a JSON file. I then transferred this JSON data into a CSV file. Included with this data were links to the specific webpages for each song, artist(s), & album.
>
> 2) In order to download the audio features of each song, it was necessary to get the track id #'s for each song from the playlist data I downloaded in the 1st step above. Then I used similar code as in step 1 to download the audio features of each song as a JSON file. I then transferred this JSON data into a CSV file.
>
> 3) Since the data had to be downloaded as 2 different CSV files, my next step was to open these 2 files and combine them into 1 CSV file. Once I was able to combine these, it was time to inspect our data and decide which columns & rows may need to be rearranged and/or removed.

What kind of cleaning steps did you perform?

After close inspection of this data, I was able to determine that we had no missing data in any of the columns or rows. I then arranged the columns in the data in an order that made more sense and would make it easier to analyze. I also deleted any columns that had webpage links since this data in not necessary for this analysis.

A few of the songs from the playlists feature guest artists on certain songs. When there are featured artists, Spotify lists a separate row for each featured artist along with their artist profile links and other data specific to each artist. To clean up our data, I added featured artists after the song title in parenthesis and deleted these artist specific rows.

How did you deal with missing values, if any?

There were no missing values in this data. However, as I mentioned in the cleaning steps above, if a song features a guest artist there were separate rows for each featured artist along with their artist profile links and other data specific to each artist and these rows were removed.

Were there outliers, and how did you decide to handle them?

There were no outliers in this data.