Capstone 1: Analysis of Top 200 Spotify stream data – Data Wrangling

Data for the Top 200 song analysis was aggregated and collected using the following sources:

|  |  |
| --- | --- |
| Data: | Source: |
| 1. Daily Top 200 songs (ranked 1 to 200 by highest stream numbers) | Web Scrape |
| 1. Song audio features on Top 200 songs | Spotify API |
| 1. Song audio analysis data | Spotify API |
| 1. Song meta data (genre, lyrics etc.) | MusixMatch API |
| *See appendix for specifics on data* |  |

1. Data for the top 200 songs was collected by web scraping [Spotify Charts](https://spotifycharts.com/). Using the Python library requests, a [short script](https://github.com/abelpd/Top_200_Spotify_Analysis/blob/master/Data/spotify_csv_scape.py) was utilized to download the daily csv files for the past two years and write to a local folder. Error handling was added to capture any dates where the csv did not download correctly. The date of these files were written to a specific folder for review.

Csv files for 4 dates were unable to be downloaded due to Spotify not having any data for those dates. Due to the top ranked songs not changing significantly day to day, the missing data was linearized using the previous day.

Once downloaded locally, another [short script](https://github.com/abelpd/Top_200_Spotify_Analysis/blob/master/Data/combine_top_200_csvs.py) was utilized to amalgamate the csv files by using a helper function to read the csv data to a list and using the Pandas “.Append” method to amalgamate to one dataframe.

Once summarized in one dataframe, a mask was used to remove all the headers from the csv files as well as any inconsistencies.

The final cleaned dataframe was written to a csv for analysis.