Colin Michael

DSC 540 Data Preparation

Milestone 1:

Part 1

Music is universally loved by people across all backgrounds and cultures. Music can entertain, unite, and inspire people in incredible ways. I am extremely passionate about music and am practically always listening to something. There is an extremely vast amount of data available about music, which makes it a highly intriguing topic for me to study. I started out by finding three different data sources in CSV, website, and API format.

I used Kaggle.com to find a CSV data source provided by Spotify: <https://www.kaggle.com/datasets/rodolfofigueroa/spotify-12m-songs>. The CSV contains over 1 million rows of data with 24 columns. Some of the most important columns include track title, album title, artist name. It also includes interest detail columns like energy, danceability, and tempo. I believe this data source will serve as the foundation for my analysis.

Next, I turned to Wikipedia.com to find information about popularity of songs in pop-culture. I found this table that shows the number of *Billboard Hot 100* songs: <https://en.wikipedia.org/wiki/List_of_Billboard_Hot_100_number-one_singles_of_the_2010s>. I am very interested to see what songs were the most popular and what common trends exist among them. I have never used Python to pull data from a Wikipedia table, so I am very excited to learn this process. I am a little intimidated, but I think it will be a great skill to have.

I also wanted to call on an API to pull in additional supporting detail. This API from Songkick provides data about live concert events: <https://www.songkick.com/developer/response-objects>. I believe this will be very useful towards my analysis as it provides a layer of information apart from pure music streaming. I have always enjoyed going to concerts and am very excited to dive into this data.

Part 2

My goal in this project is to merge the CSV data, Wikipedia data, and API data into one cohesive dataset. I believe this will create an extremely compelling dataset on music that covers streaming data, music popularity data, and live event data. I am excited to create a wholistic data set that should allow for excellent analysis

I believe that I will need to create common ids between the datasets to join them together. First, I want to combine the CSV streaming data from Kaggle with the Wikipedia *Billboard* data. I should be able to do this by joining on song name. I believe my biggest challenge will be making sure that the song names are consistently formatted between the two sources. I am thinking I will be able to cast them all as lowercase and remove any quotations or special characters but am sure I will learn the proper techniques in this course. Using specific examples has always helped me find a solution on joining datasets.

Next, I will need to bring in the live music information from the Songkick API. This is a little different from the first two datasets because it focuses on performers, rather than individual songs. I believe I will need to join on artist name. Like before, I will need to be sure that the artist names are uniformly formatted to ensure that they join properly. I also envision a scenario where there is API data on performers that are not in my larger dataset, and performers from the Spotify CSV that are not in the concert API dataset. This is completely fine, and I believe I will do a full outer join, which means that data without exact matches will not be excluded.

I do not see any clear ethical data issues in my analysis. It does not contain any personal data information, such as the names of people who streamed content from Spotify or the names of people attending live events. I will need to be sure that I properly reference the datasets from Spotify, Wikipedia, and Songkick in my analysis. It would be a major ethical violation if I passed all this data off as my own findings. I will be sure to give proper acknowledgement to the original sources of my data.