Wrangle and Analyze Data Report
John Zaharick

I found the difficulty in this project split between learning new skills and the length of time involved in applying known skills. Gathering the data from Twitter's API was the most challenging and I engaged in a lot of trial and error figuring out the code for downloading data associated with a tweet id and figuring out where to set parameters such as wait_on_rate_limit. Making sure each tweet was written to a new line in tweet_json.txt also took some research. I didn't know if there was a built in method for this or if I had to construct my own command and insert \n in the file after each tweet (I ended up using the latter). When I initially tried to set the parameters for waiting on the Twitter rate limit, I placed them in the wrong location and kept throwing my except function. I did this experimenting with a small subset of the tweets (just 5 tweet ids) so I could repeatedly call on the API without long waits between learning if my code was working as intended. Reading the downloaded tweets back into Jupyter Notebook and storing them in a pandas data frame was also tricky. I again expected a pre-existing function could do this, but ended up writing my own loop to perform the task. My instinct is to find code that can accomplish tasks in one line after an experience on a project outside of this class where I built a series of Python functions to clean a data set, and then later learned the pandas package and discovered that I could perform the same tasks with less code in seconds instead of the minutes my home brewed loops took.

Assessing and cleaning were a different kind of challenge. I knew what code to employ from the prior lessons in this class, but the sheer amount of work involved was daunting. I've spent many hours on this project. Assessing was also hard at first as I didn't notice very many issues with the data and felt lost. As I spent more time looking at summaries of the data, reading individual rows, and thinking about the prior lessons, I noticed more and more problems however. At this point, the work became exciting because I knew what the problems were and how to solve them.

Cleaning the data was the most time consuming, mostly in terms of formatting the Jupyter Notebook with markdown code and keeping everything organized for others to understand my work. There's a temptation to just write code and solve problems one after the other, but communicating the code and work done to others is of vital importance. I also found myself assessing the data again in the middle of cleaning, specifically in the case of identifying incorrect scores based on unusual denominators. I had dropped many rows by this point, so some incorrect scores identified earlier no longer existed in the data frame. I also had to re-index the data frame to be able to properly target individual rows.

I spent a lot of time on this project and had numerous frustrations, but it feels very rewarding to see the final csv file of the cleaned data open in a spreadsheet and to see the analysis I conducted thanks to the cleaned data.