Data Wrangling Report

I had to obtain the necessary data from three sources in various ways before I could start the data wrangling procedure. For the first and most straightforward technique, I read a CSV file into a dataframe called archive after downloading it from the Udacity website. I used the requests library to programmatically download a TSV file from a URL provided by Udacity, which was the second data source. I then read this file into a predictions dataframe. Since I was unable to access the Twitter API, I used a tweet json file that was provided by udacity for the final method. Using the json library, I was able to read it line by line and generate the tweet_data

I used both programmatic and visual assessments during the evaluation process. Numerous problems in the archive dataframe were caused by the erroneous extraction of dog stages, ratings, and names from the text column. Additionally, there were several columns for various factors in both the archive and predictions dataframes, which made the data unorganized. Due to the deletion of some tweets, there were some missing data points in the tweet data dataframe. I did remark that this information couldn't be found elsewhere, though. Overall, I also required to simplify the data by having one dataframe for each observational unit—tweet_data, dog data, and predictions and to limit each dataframe to just original content with photographs.

I started the cleaning process by making duplicates of the dataframe, then I iterated through defining the cleaning operation, coding it, and verifying the outcome. I chose to start by limiting the dataframes to only original content with photos because I thought that could solve some of the other problems. The data was then organized, starting with archive clean and predictions clean, which were simpler to organize because the data were generally reliable and correct. However, I had to re-extract the names, ratings, and dog stages before I could tidy up dogs_clean by splitting it from archive_clean. Since it was not possible to check the accuracy of each tweet individually, this was the most challenging step in the cleaning process. I acknowledge that there may still be some problems with dogs clean. I was able to quickly fix any lingering quality issues after tidying up the three new dataframes.

I saved each dataframe to a separate CSV file to put an end to the data wrangling procedure.