Description of Data Wrangling for WeRateDogs Project

The first step of data wrangling for this project was to gather the data. There were 3 pieces of data I gathered. The Twitter archive, twitter_archive_enhanced.csv, was given to me to simply read in. The tweet image predictions, image_predictions.tsv, were downloaded from Udacity's servers via Requests. Lastly, I queried the Twitter API for additional information about each tweet in the archive and programmatically saved it to tweet_json.txt to read in separately in Pandas.

The second step of the wrangling process was to asses the data for quality and tidiness issues. I used Pandas commands including head()/sample() as well as opening the files in Excel to visually look for problems in the data. Next, I used .info(), .describe(), and .value_counts() to look for inconsistencies, missing data, and incorrect data. There was a lot of incorrect data. I also saw several columns in each of the tables I could remove.

The final step of the process was to clean the data based on the assessments I made. Almost all the cleaning was done on the twitter archive table. The first thing I did was to remove all the retweets, but finding which tweets had a non-null retweet status ID. I then dropped the 3 columns related to retweets from that table. I also chose to remove all the replies to tweets, in favor of analyzing only original WeRateDogs tweets. I then dropped columns detailing replies as well. The next issue I fixed was incorrect ratings. I chose to re-extract all numerator and denominator ratings from the text. After fixing some manually, I forced all denominators to be 10 by reducing fraction ratings that were out of multiples of 10 to rate multiple dogs. The way I extracted the ratings also accounted for decimal numerators, which it originally had not. Next, I removed an original tweet that was not a dog rating, but a post about plagiarism (it was given a rating of 0/10 and would have skewed my data). Another variable I modified was timestamp, which I changed to datetime format. Next, I re-extracted dog names from the text column, so that there weren't any "the", "a", "an" or other one-off incorrect names. I had to fix a few (with apostrophes and accent marks) manually. In the image predictions table, I changed dog breeds to have consistent capitalization. Before merging all the data, I removed a few columns from this table and the tweet archives table, as well as all but ID, favorite count, and retweet count from the Twitter API table. I made 2 final tidiness cleaning steps. The first was to combine the dog stages into one column in the Twitter archive table. The final was to merge all three tables into one.