WeRateDogs Wrangle Report

The data wrangling efforts started by gathering the three required datasets for the WeRateDogs Twitter project. I started by downloading and reading the provided Twitter_archive.csv dataset. Then I programmatically downloaded the image prediction dataset by making use of the requests library. The third dataset was gathered using Twitter API, accessed through the tweepy library with my access keys to get the tweet IDs, retweet count, and favorite count. This was then stored in a json.txt file and converted to a pandas data frame. After storing the data, I renamed the 'id' column from the Twitter API queried data as `tweet_id` because it is the unique identifier across the three datasets. This will prove useful when merging the data.

After visually accessing the data using a spreadsheet and programmatically accessing the data, there were eight data quality issues and two tidiness issues identified in the dataset that needed to be cleaned so that I could have accurate and clean data for analysis. Firstly, I dropped rows that were replies or retweets as I wanted original ratings that had images. I also dropped columns that would not be useful for the analysis of the data. Secondly, there were incorrect dog names both for names provided and names not provided. I cleaned the incorrect dog names using regular expression (I used an online tool that could capture text strings to help ensure my regular expression was correct and capturing the right information) to extract the correct names given and I replaced the incorrect dog names that were not given to 'None'. Incorrect dog stages was also cleaned. The dog stage `doggo` column had values erroneously captured as 'None' instead of doggo. Some of the rating numerators and denominators were incorrect so I first converted the datatype of the rating numerator to float so that the correct rating numerators could be captured, then used a regular expression to extract the rating numerators that were in decimals. Ratings with incorrect denominators, that is ratings that were not equal to 10 were cleaned by either dividing such ratings by the common factor as such ratings were given to images that had more than one dog or they were replaced with the correct rating.

The timestamp column was converted to a datetime datatype and the tweet_id was converted to a string. The doggo, floofer, pupper, and puppo columns were concatenated to form the `dog_stage` column. Finally, I merged all three cleaned datasets into a master dataset using the `tweet_id` column as the unique identifier. The cleaned dataset was then saved and used for analysis.