Wrangling Report

Wrangling data is a streamlined iterative process that all Data Analysts must excel in order to be prosperous in the IT industry. I took this opportunity to increase my skills and further hone my data wrangling knowledge by working with WeRateDogs twitter data.

I was tasked with getting data from various different sources, such as downloading the twitter_archive given by Udacity directly, downloading the image predictions programmatically using a get request, and using a twitter API to gather the retweet and favorite counts of WeRateDogs tweets.

Sifting through our three datasets I began assessing visually first through pandas handy .info() calls. I immediately noticed missing rows for several columns such at 'in_reply_...' and 'retweeted_status_...'. I then began to look through the datatypes of several variables and noticed some were out of place such as 'timestamp' which was a pandas object (safely assume it was string), and 'created_at' which was also an object. I then decided to look for inconsistences in the columns, and I've only bumped into one significant one which was the names for the predictions, some were capitalized and some weren't. Some of the columns also contained information that I did not want in my master dataset, such as the 'source' column which had html tags, would be hard to sort through the column if I wanted to find a source of a tweet_id. Also the 'text' column contained hyperlinks which again causes problems when we want to sort through and see what the actual tweet was for each tweet_id. For the final quality issue I noted something that bothered me which was the rating numerators, now I know WeRateDogs has their own way of doing ratings I felt as if there was no scale, or nothing that determined what actually was a very high rating compared to a lower one, some of the ratings were absurd and went above 400/10.

As for tidiness issues the two obvious ones that came to mind when I looked through each dataset were:

- Dog stages were multiple variables as columns
- We only need one master data set because the observational unit for each table was the same, tweets.

After my assessing efforts it was time to clean the data set and make it tidy enough for the user to understand. I tackled the tidiness issues first since based off previous lessons it makes all the other operations easier. I ended up creating a function that I could reuse to simply replace any string or NaN with a string of my choosing, and that helped a lot in getting my columns ready to be merged into one. I then applied a lambda function to join any empty strings to my 'maturity' column. I ran into another issue which I didn't notice before hand was that the dataset had several maturity tupules such as 'doggopupper' a simple reassignment to doggo, pupper helped organize my 'maturity' column. After resolving the first tidiness issue it was time to merge all the dataframes together using .merge from pandas. Since all the tables were linked by tweet_id as their primary keys we could use 'left joins' essentially from SQL to merge each dataframe on tweet_id. And so I did that and had a master dataset ready to be further cleaned. After those efforts I performed various cleaning techniques to solve my issues such as dropping duplicates, changing data types, using functions along with regex to remove html tags or hyperlinks, and setting a maximum for the ratings to be 15.

I thorough enjoyed working on this wrangling project as not only did I gain more skills in data wrangling but I honed some aswell.