Data Wrangling Report

This document summarizes the data wrangling efforts performed to produce a master Twitter archive dataset for the WeRateDogs Twitter account. This project was performed using Python in Jupyter Notebooks.

For the first step of the data wrangling process, data was gathered. Enhanced Twitter archive data in .csv format was provided which contained basic tweet data such as the ID, URL, timestamp, text, reply status and retweet status. Additionally, from the text of the tweet, the dog names, stages and ratings were extracted in this data. To provide more data for our analyses, we used the Tweepy library to query Twitter's API and write each tweet's JSON data into a text file. From the JSON data in the text file, we then obtained the retweets and likes counts to add on to our data analyses. Finally, we were also provided with image prediction data in .tsv format which used a neural network to classify the breed of the dog from the image in the tweet. This data included the top 3 predictions and their corresponding confidences.

In the second step, the data was assessed for data quality and tidiness issues using visual and programmatic assessments. The issues that were found are as follows:

Quality

- The 'timestamp' column has an object datatype instead of a datetime datatype in the 'archive_df' table
- We only want original tweets that are not replies or retweets, or the rows where the 'in_reply_to_status' and 'retweeted_status' columns are null
- The 'expanded urls' column is missing rows
- There are invalid dog names such as 'a', 'an', 'the' and etc.
- The 'rating_numerator' column is of type int and is improperly extracted because of the decimal
- For the 'image_df' table, some of the predictions are not dog breeds and the column names are not very descriptive
- The image prediction words are separated by underscores and are inconsistently capitalized
- There are 17 errors for the 'api_df' table, resulting in 17 missing rows
- There are only 2075 rows in the 'image_df' table compared to 2356 rows in the 'archive_df' table

Tidiness

- Data about the dog's stage is in 4 separate columns in the 'archive_df' table and some dogs are in multiple dog stages
- Each table does not contain one observational unit — it makes sense to include all the information in one dataset since they all pertain to the dog's tweet data

The final step was to clean the data based on the assessments. Using a variety of Python and pandas functions, all the data was combined into a master Twitter archive dataset. This master dataset was saved as a .csv file, and the file and the Jupyter Notebook was submitted as part of Udacity's Data Analyst Nanodegree Program.