An Explanation of my Data Wrangling Process

In this project, I was tasked with data wrangling Twitter posts from the WeRateDogs account. Data wrangling consists of three stages: gathering, assessing, and cleaning. In the first stage, I gathered the WeRateDogs Twitter Archive, which consisted of about 2,400 Twitter posts, from 2015-2017. This archive was stored locally as a CSV file, and parsed using Pandas read_csv(). Stored remotely, were about 2,000 image predictions from posted images within the archive. Using Python’s Requests Library, I was able to fetch from a remote server the TSV file, which was also parsed using read_csv(). Lastly, I gathered Twitter meta-data for the selected posts. Unfortunately, I was unable to receive access to the Twitter API. So, I fetched a json file from a remote server, which was parsed with Pandas read_json().

In assessing the data, the two immediate issues in each dataset were a significant number of null values and columns unrelated to our future analysis. In the Twitter archive, I found that the Retweet and Reply status/user IDs contributed no valuable info. In the names column, all lower-case words were misidentified as names. And lastly, the rating system was inconsistent, if not erroneous. In the Image Predictions, columns like jpeg url or image number were unnecessary. Also, many predictions were not of dog breeds. Lastly, in the meta-data JSON I observed duplicated columns from the first Twitter Archive dataset and more unrelated columns to the future analysis.

To clean the data, I identified four Data Tidiness issues and eight Data Quality issues. The Data Tidiness issues addressed the removal of duplicate and unrelated columns, merging several columns into one, and merging all three datasets into one. The Data Quality issues included fixing misidentified names, filling null values, simplifying source locations, normalizing ratings, and data format conversions, among others. All of the cleaning steps are further described in the wrangle_act.ipynb notebook.