Wrangle Report

In the wrangling process, I followed three major steps: data gathering, data assessment, and data cleaning.

In the data gathering step, I obtained three datasets. The twitter_archive_enhanced.csv has data from the We Rate Dogs twitter archive. It has fields such as twitter id, timestamp, source, text etc. The second dataset is image_predictions.tsv which I obtained using a python get request. This data derives from a machine learning model which predicts the twitter picture. The data shows the top three predictions, the confidence for those predictions, and whether or not that prediction is a dog. The last dataset was obtained from querying twitter's api. The api call returned a series of jsons and I pulled the retweet and like counts.

In the data assessment step, I checked the data for quality and tidiness. As I was checking for quality, I checked the data for missing values, inaccurate values, and duplicates. When I was assessing the data for tidiness, I checked to see if variables were columns, observations were rows, and observational units were tables.

In the data cleaning step, I corrected the following data errors that I found:

1. Some observations are retweets. I dropped observations with retweet_status id.
2. Some observations are replies. I dropped observation with a reply_status id.
3. The timestamp in the twitter csv is a string. This is harder to manipulate so I changed it to a date time object.
4. Dog type (doggo, puppo, etc) should be one column. According to rules of tidy data, the dog type is one variable so it should be a single column instead of dummy variables. I collapsed this data into one column.
5. Some dogs don't have a specified type. Some dogs had designated types such as "floofer" or "puppo", but many were listed as none. Additionally, some dogs were listed as multiple types. So those are now categorized as "multiple."
6. Inaccurate twitter denominator column. This should be 10, but some were not 10. These observations were also dropped.
7. In the twitter csv, some of the dog names have weird values such as "a" or "his" etc. I filtered these out by only keeping proper nouns, aka values that started with a capital letter.
8. Source URL is clunky so I split the value to extract only the source without the rest of the url. It will now say "twitter for iphone" instead of the whole url, for example.
9. Not all images are dogs. Some of the predictions listed that the observations were not dogs. These were dropped from the dataset.
10. The id headers should have the same name. I changed the id column label to say "twitter_id" for each data frame.