Data Wrangling Report

Jin Lim, Udacity Data Analyst Nanodegree

I. Gather

- I gathered from three different data sources.
- The first was a CSV file called 'twitter-archive-enhanced.csv' and was provided by Udacity. This was sent from WeRateDogs own twitter archives, and was wrangled to extract ratings, dog names, and dog "stages" to make an "enhanced" version.
- The second was a TSV file called 'image-predictions.tsv' which I downloaded programmatically from Udacity's servers. This contained breed predictions from supplied images of dogs.
- The third was a text file called 'tweet_json.txt', and was scraped from Twitter's API and written into the text file. This contained tweet ids, favorites and retweets counts.

II. Assess

- Before starting anything, I made sure to look over the data where I could in Excel. By using Excel's various functions to sort and filter, I could visually assess the data firsthand.
- I used pandas functions such as describe(), info(), value_counts(), etc. to further visually assess. Programmatically, I also searched for null values, duplicates, and odd or invalid values (such as where ratings = 0).
- Here were the issues I found:

Data Quality Issues

1. Remove retweets, as these are not original ratings
2. Remove missing images and breed predictions

3. Tweet_archive has the erroneous datatypes for the following columns: tweet_id (int -> str), timestamp (-> datetime)
4. Several names are not actually names, such as "A, An"
5. Several tweets have been deleted (missing API data indicated by N/A during Tweepy query)
6. Several columns have values of "None", could be changed to NaN
7. Several other variables, such as numerators and denominators, also have invalid datatypes
8. Rating numerator and denominator columns sometimes have incorrect values

Data Tidiness Issues

1. Doggo, floofer, pupper, puppo should be combined into one column
2. All three dataframes (info, tweet_archive, images) should be combined

III. Clean

- I attempted to fix each issue outlined above to the best of my ability. I made a copy of each dataframe before commencing cleaning.
- The broad strokes of my process included merging redundant columns, merging all three dataframes together, fixing incorrect datatypes, using regex expressions to find appropriate tweets to modify values, and more.