Data wrangling efforts.

The provided datasets were `twitter_archive_enhanced.csv`, `image_predictions.tsv` and `the tweet_jspn.txt`. These 3 datasets are riddled with Quality and Tidiness issues. The quality issues are as follows: The rating_denominator values were differing instead of all reading 10, I had to programmatically set all the values to 10. The `names` were mostly inaccurately extracted with invalid names like 'a'. I defined a function to extract the names from text as well as removing the invalid names. The `timestamp` column was of string dtype instead of datetime which I made sure to correct. Some of the tweets are retweets and therefore of no use. I removed them programmatically too. Some of the columns had a lot of null values, and had little meaning to the analysis so I dropped them. The columns `id` and `id_str` were duplicated columns and were no always equal, so I dropped id_str since it was the erroneous variable. The columns in the different dataset had overlaps whereby some shared the same data but with different column names. I renamed the column names to make later merging easier. The `source` variable contained irrelevant html tags, I hence removed the texts to extract the relevant information. In the last dataset, plenty of the columns were very much useless to my analysis so I dropped them too.

For the Tidiness issue, I discovered in the first dataset, there were four variables referring to the same variable `dog_stage`. I resorted to melting these columns to form a single variable 'dog_stage' with the values 'doggo', 'floofer', 'pupper' and 'puppo'. I then dropped the resulting duplicate variable. The other issue was the separate tables for the related data. Having renamed the similar columns across the tables to match each other, I used left merge to join the first two data sets on the variables: tweet_id', 'timestamp', 'text', 'source'. I then did the same with the last table using `tweet_id` to form the master dataset twitter_archive_master.csv