Wrangling Report

For the data wrangling portion of this project, I gathered, assessed, and cleaned data.

For the gathering portion of the project, I acquired three different tables.

- The first table was provided by Udacity as a .csv file, which I uploaded to the Jupyter notebook local directory and read into the notebook.
- The second table, I downloaded from a website in the Udacity server using the Requests library.
- For the third table, I used the first table to acquire the tweet_ids and I used the Tweepy API to extract the data for these tweets using a for loop and saved it as .txt file in the directory. I then extracted the .json data line by line with another loop and turned it into a Data Frame for later use.

For the assessing portion, I used sample(), info(), describe(), value_counts() and visual inspection to try to determine 8 quality issues and 2 tidiness issues for the cleaning portion as required. I ended up with the following list:

Quality Issues:

1. DF1 timestamp data type is a string instead of datetime.
2. DF1 index 2034, 2066, and 2116 names are incorrect. Should be 'Jacob', 'Rufus', and 'Spork' respectively.
3. DF1 names 'a','an',and 'the' are either incorrect or not descriptive enough. Since they indicate a lack of name identification, these are not consistent with the existing 'None' value.
4. DF2 dog breed names are not case consistent.
5. DF2 dog names have different characters for spaces. Mainly, - or _.
6. Retweet and reply rows exist in DF1 and need to be removed.
7. For DF1, where denominator is not 10, most are not actual ratings but dates, names (e.g. 7/11), etc.
8. DF1 source column has tags and other unecessary coding.

Tidiness Issues:

1. DF3 can be merged with DF1
2. Once DF3 and DF1 are merged, reply and retweet indicator columns can be removed as only original tweets remain. Additionally, text_y column can be removed, too.
3. DF1 doggo, floofer, pupper, puppo column is unnecessary and can be collapsed into one column.

For the cleaning portion, I laid out my notebook cells as ‘Define’ and ‘Code’ blocks for each issue. I tested while coding, but did not include it as part of the notebook as it would make the notebook bulky.

During the cleaning portion, I merged two of my tables into one and ended up with two tables. One for the twitter data and another one for the dog predictions. I saved both files into the directory as 'twitter_archive_master.csv' and as 'tweet_prediction_master.csv' respectively.