Wrangling Report

Gathering:

The focus here was on reading the data into the jupyter notebook from 3 sources.

Firstly, manually downloading the WeRateDogs archive as given in the link provided by Udactiy

Then programmatically downloading the image-predictions data by requesting the url provided by Udacity and writing to a tab separated file image-predictions.tsv

Lastly, using the twitter api and the tweepy library to retrieve additional tweet information that was missing from the archives. Specifically, we needed to open a blank text file tweet_json.txt in write mode, then loop through each tweet id from the archive data, get the full tweet JSON using tweepy, and then dump the entire JSON information into tweet_json.txt. We then read back in the JSON information extracting only the tweet_id, favorite count and retweet count and storing this in a dataframe

Assessing:

After visually and programmatically assessing the data from the above 3 sources, the following issues were spotted.

Quality
Archive table
- timestamp is not datetime
- archive contains tweets that were deleted and therefore would not have full info on those tweets
- min denominator rating of 0, id 835246439529840640, this was actually a corrective tweet
- Drop columns we have no intentions of using
- Drop RT's
- Replace Nan with None in in_reply_to_user_id
- 891087950875897856 Missing dog name, (Marlo)
- 885518971528720385 Missing dog name, (Howard)
- Dog stage columns should be 1 for whether a dog of that stage is present in the tweet and 0 otherwise

image_predictions
- upper and lower case references to object names (p1,p2,p3)
- drop columns we don't intend to use

Tidiness
image_predictions
- Data should be in one table
- Duplicated columns such as text and source across the archiveand image_predictions table

Indeed, there are significantly more issues surrounding the quality and tidiness of thus dataset. For example, there are certainly instances of incorrect dog names and inappropriate listing of dog stages. However, these issues would take more time and thought to properly address.

Cleaning:

All of the issues listed above were fixed largely in part by the joining of the 3 tables. The RTs were dropped as specifically requested by the project motivation. Other methods include RegEx patterns and pandas string methods were used heavily.