WeRateDogs Data Wrangling

To gather the data for the project, I have used three different sources:

1. The WeRateDogs Twitter archive was downloaded directly from the Udacity website and was sotred as 'archive'.
2. The tweet image predictions data was downloaded from the following link using Requests library and was stored as 'predictions': https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv
3. The final dataset includes tweet id, retweet and favorite count and the time of the tweet and was extracted directly form twitter website using Tweepy library and was stored in a text file called 'tweet_json'.

After assessing the data the following quality and tidiness issues were identified and addressed. See wrangle_act.ipynb for more details.

Quality

Archive Table

- tweet_id has integer datatype instead of string
- timestamp and retweeted_status_timestamp have the wrong datatype (object instead of date)
- there are retweets in the dataset (some of the expaned_urls are duplicates)
- many errors in the dog names (e.g. a, an), some dogs do not have a name and the name column is recorded as 'None' instead of np.nan making it hard to use some functions when exploring the data
- there are different rating_denominators that make the comparison difficult (there is a denominator of 0 in one case)

Predictions Table

- tweet_id has integer datatype instead of string
- most of the column names are not descriptive (e.g. p1 instead of prediction1)
- it has less datapoints than the archive table (2075 compared to 2356), indicating that a lot of the data in the archive table does not have an image and might be a retweet
- 324 images are not classified as dogs
- there are some duplicates of the same dog (same image url) tweeted with different ids

tweet_json Table

- id has integer datatype instead of string

Tidyness

Archive Table

- columns 'doggo', 'floofer', 'pupper', and 'puppo' should be combined in one column because they show the single variable dog_stage

Predictions Table & Tweet_json Table

- there is no need for these separate tables, can be combined with the archive table

After combining addressing the above issues and combining the dataset the following new issues were identified and addressed:

Quality

- columns in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, and retweted_status_timestamp, retweeted do not provide a lot of information
- column full_text (from the tweet_json table) has the same information as column text (from the archive table_
- the tweet source is in a url format and is hard to categorize and compare

Tidyness

- there are three different predictions for every dog which can make analysis confusing, it is better to store the best prediction (the prediction with the highest confidence level) in a separate