Summary of data wrangling performed for WeRateDogs® project

In this short report, we briefly describe the data wrangling process completed for the "WeRateDogs" project. Three datasets were gathered and cleaned:

1. `twitter_archive_enhanced.csv`. The udacity provided "Enhanced Twitter Archive" comma separated values (.csv) file contains basic tweet data for 2356 tweets, including the tweet text and ID. The archive is "enhanced" because each tweet also has the associated dog name, rating, and stage, all of which were extracted from the tweet text.

2. `tweet_json.txt`. Using the python module tweepy, additional data was acquired through the Twitter API: we downloaded all the metadata associated with each tweet ID in the `twitter_archived_enhanced` file. This data was saved as `tweet_json.txt`. This file was then read in, line by line, using the python json package. This was stored as a pandas dataframe under the name `dfTwitter`.

3. `image_predictions.tsv`. The udacity provided tweet image predictions tab separated (.tsv) file contains image predictions outputted by a neural network that classifies dog breeds.

Most of the wrangling efforts were devoted to cleaning the enhanced twitter archive.

The following issues of data quality and data tidiness were fixed from the `twitter_enhanced_archive` csv after it was converted to a pandas dataframe:

1. Retweets and associated columns were removed.
2. Replies to other users were removed.
3. Data from the '`source`' column was extracted into a human-readable form and the values associated with the two separate variables, '`source_name`' and '`source_location`' were stored in separate columns.
4. Data from the timestamp column was extracted into two separate columns, '`time`' and '`date`'. The datatype for these columns was set to the python datetime format.
5. The four columns '`doggo`', '`floofer`', '`pupper`' & '`puppo`' were removed; in their place, a single column was created, 'stage'. This column stored the stage of each dog, where the previous column names became the possible values that the stage could take.
6. Rows where the rating denominator was not 10 were removed once we recognized that incorrect values appeared because non-ratings such as dates were accidentally identified as ratings. This allowed us to then remove the `` `rating_denominator` `` column as all values were now 10. The `` `rating_numerator` `` column could then be renamed `` `rating_out_of_10' ``.
7. The row with the dog rating of 1776 was removed. On inspecting the associated text, we discovered that this rating was given merely as an amusing allusion to the American Day of Independence. On iterating the Assess-Clean-Analyse protocol, we found it useful also to a fix a few other outliers.

The following data quality issue was fixed from the `image_predictions` file after it was converted to a pandas dataframe:

1. Any image predictions that were not identified as dogs was removed.

The following data quality issue was fixed from the `dfTwitter` file:

1. The datatypes of the following columns were changed to int64: `` `tweet_id` ``, `` `retweet_count` `` and `` `favorite_count` ``.

The last change allowed the `dfTwitter` table and `twitter_enhanced_archive` table to be merged using `tweet_id` as the key, and the resulting table was stored as `twitter_archive_master.csv`. The cleaned image predictions were saved as `cleaned_image_predictions.csv`. Note: It would be better to save these as pickle files (instead of csv) in the future so that the datatypes are preserved.