We Rate Dogs Tweets Wrangle Report

- ## Data gathering
    - **twitter_archive** dataset, manually download from Udacity.
    - Use the Requests library to download the tweet image prediction (**image_predictions.tsv**)
    - Use the Tweepy library to query additional data via the Twitter API (**tweet_json.txt**)
    - These three files are then loaded into three dataframes, **twitter_archive, df_image** and **df_data.**
- ## Data accessing
    - ### Quality issues

      twitter_archive

1. "timestamp" should be datetime type rather object(string) and "tweet_id" should be string rather int.
2. "source" column contain unnecessary html residues, only text part should be parsed. They can be simplified as 4 types:
    - Twitter for iPhone
    - Vine - Make a Scene
    - Twitter Web Client
    - TweetDeck
3. Drop all retweets (78) and reply tweets (181).
4. Columns of "in_reply_to_status_id", "in_reply_to_user_id", "retweeted_status_id", "retweeted_stat

us_user_id"',"retweeted_status_timestamp"` are not about original tweets, so they should be dropped.

5. 109 inaccurate names are recorded in "names" column need to be dropped. 59 "None" names should be changed to "NaN".
6. There are 639 double links presents in the "expanded_urls" column.
7. Ignoring retweets or replies tweets, There are 17 tweets with denominators that aren't 10. 4 tweets need to be manually corrected and other 13 tweets will be dropped given multiple dogs present in the original tweets.
    - **tweet id**: 740373189193256964 **original rating**: 9/11 **new rating**: 14/10
    - **tweet id**: 716439118184652801 **original rating**: 50/50 **new rating**: 11/10
    - **tweet id**: 682962037429899265 **original rating**: 7/11 **new rating**: 10/10
    - **tweet id**: 666287406224695296 **original rating**: 1/12 **new rating**: 9/10
8. Ignoring reply or retweets, there are 5 tweets whose rating_numerator is far bigger than 15 given its rating_denominator is 10. Among these 5 tweets, 3 tweets will be correcting and rest will be dropped.
    - **tweet id**: 786709082849828864 **original rating**: 9.75/10 **new rating**: 10/10
    - **tweet id**: 778027034220126208 **original rating**: 11.27/10 **new rating**: 11/10
    - **tweet id**: 680494726643068929 **original rating**: 11.26/10 **new rating**: 11/10

image_prediction

1. "tweet_id" should be string type
2. "p1", "p2" and "p3" columns, writings are not consistent (upper case or lower case).
3. There are 324 tweets whose dogs can’t be recognized by the algorithm.

tweet_json

1. "id" should be string type

Tidiness issues

1. doggo, floofer, pupper, puppo all describe one property "dog stage" which violates tidiness rule, they should exist in one column.
2. Columns with numerical data are located to the far right of the tweet_archive table, which makes it difficult to readily see the data that will be used for analyses.
3. The dog breed prediction with the highest confidence level can be combined with the tweet_archive table to make the information more comprehensive
4. json dataframe should be combined with first table