WeRateDogs Twitter Archive - Wrangle Report

In this report I will briefly describe my wrangling efforts to gather, assess, and clean the data required for analysis of the WeRateDogs Twitter Archive in order to o create interesting and trustworthy analyses and visualizations.

Data Gathering

I gathered data from 3 sources:

1. WeRateDogs Twitter Enhanced archive, manually downloaded from the Udacity servers.
2. The image predictions file, programmatically downloaded from the Udacity servers.
3. The entire set of each tweets' JSON data, downloaded by querying the Twitter API using the Tweepy library from where I extracted programmatically The favourite_count and retweet_count.

I loaded the 3 raw data files into separate tables: archive, predictions and extra, respectively.

Data Assessing & Cleaning

After gathering all three pieces of data, I assessed them visually and programmatically for quality and tidiness issues. I detect and document **nine (9) quality issues** and **three (3) tidiness issues**.
Before I perform the cleaning, I made a copy of the original data then I used the define-code-test framework and clearly documented it.

Quality issues and solutions:

1. Unnecessary 181 retweets
2. Unnecessary 78 reply tweets
   → All rows containing non-null values in the retweeted_status_id retweeted_status_user_id and retweeted_status_timestamp , and also in the in_reply_to_status_id and in_reply_to_user_id columns were dropped
3. The timestamp column is string rather than date
   → The timestamp column was converted to datetime data type
4. Invalid 109 lower case strings in the name column: "a", "an", "in"
   → All the lowercase words were replaced with 'none'
5. Empty retweets columns
6. Empty reply columns
   → These columns were then also dropped

7. Unreasonable high values in the rating_numerator
   → Tweets with large numerators (>20) were dropped
8. Sources can be simplified and divided into 4 subsets: Twitter for iPhone, Vine - Make a Scene, Twitter Web Client, and TweetDeck
   → The html strings in the source column were replaced with the display portion of itself.
9. 281 missing image predicted compared to the number of tweets in the archive
   → The rows with missing images were dropped

Tidiness issues and solutions:

1. The 4 columns doggo, floofer, pupper, puppo can be resumed in one column called dog_type
   → The 4 dog stage columns were melted into the stage column: dog_stage
2. All rating_denominators are the same (10) this column is no longer needed.
   → I dropped the rating_denominators column (it contained only ‘10’s)
3. The dog breed prediction with the highest confidence level can be combined with the archive table as the twitter table contains information that is all about the dog in the tweet
   → The best prediction for breed and associated confidence level were extracted and merged into the archive table.

I finally merged individual pieces of data according to the rules of tidy data. The result was a high-quality and tidy master pandas DataFrame: “twitter_archive_master.csv”.