Wrangling Report: @WeRateDogs Twitter Data

The @WeRateDogs twitter account posts images of dogs along with a text blurb description and a rating for the dog. The data for this analysis comes from three sources:

First, a .csv file of tweet data including the tweet ID, the url, the text of the tweet, retweeting status id and user id, and other information about the content of the tweet. This was uploaded using the pd.to_csv function.

Second, a .tsv file from the internet that was saved to the Jupyter Notebook with the python requests library. This file contains information about each tweet and a prediction of the images included in the tweet.

Finally, I used the twitter's API tweepy to gather additional data about each tweet in the .csv archive of tweets.

Following the gathering of the data, I performed a visual and programmatic assessment of all three dataframes to identify and fix quality and tidiness issues to prepare the data for analysis.

Quality Issues:
1. There should be no retweets or replies in dataset: remove tweets/rows from dataset that are retweets or replies
2. Texts of tweets that start with "RT" are manual tweets and should be removed from dataset.
3. `archive["text"]` (same as `tweets_api["text"]`) have too long of strings to be readable: extract useable info from string
4. Missing `"expanded_urls"` values in 4 rows: meaning no image with tweet.
5. `df["timestamp"]` is string, not datetime object.
6. Column names in `images` unclear
7. `name` in `archive` should be titlecase; fix any obvious name errors (ie: "a")
8. Image predictions in columns `p1`, `p2` and `p3` have underscores.

Tidiness Issues:
9. Data in `tweets_api` can be joined with archive to supplement missing data and because it contains data for the same tweet ids.
10. Columns for `"in_reply_to_status_id"` and `"in_reply_to_screen_name"` contain no data once retweets and replies are filtered out of dataset and joined with archive.
11. `"rating_numerator"` and `"rating denominator"` columns are two columns for data point: rating. Convert to integer value.
12. Columns: `['Doggo', 'floofer', 'pupper', 'puppo']` can be condensed to one column: dog type.

Following the cleaning and wrangling of this data, I had two separate dataframes that were saved into .csv files for analysis: `twitter_dogs.csv` and 'dog_predictions.csv`.