Wrangling efforts on data from We Rate Dogs

The following is a short summary of the wrangling I did on the WeRateDogs Twitter data. For further details, please check out the Jupyter notebook file wrangle_act.ipynb.

Gathering data:

Three pieces of data were downloaded into Jupyter Notebook wrangle_act.ipynb:

- **twitter_archive_enhanced.csv** (from Udacity) – saved and copied as DataFrame twit_archive_clean
- **image_predictions.tsv (**downloaded programmatically using the Requests library from Udacity's servers) – saved and copied as DataFrame image_pred_clean
- **tweet_json.txt** (I never received any response from Twitter to my requests for a Twitter API key, and so I had to download **tweet_json.txt** from the Udacity website.) -- saved and copied as DataFrame tweets_clean.

Assessing data:

I identified the following data quality issues:

1) In tweets_clean:
    a. "id" is integer format and should be a string
    b. "created at" is a string and not a datetime object
    c. there are columns that have either a large number of null values or are all null values.
2) In image_pred_clean:
    a. "tweet_id" value is integer, and should be string
3) In twit_archive_clean:
    a. "tweet_id" value is integer, and should be string
    b. "timestamp" is a string and should be datetime object
    c. Lowercase words such as "a", "an," and "the" appear as "name" values; in addition, "None" is a "name" value, even when there is a name in the tweet
    d. there are rows where the values for rating_numerator and rating_denominator are missing or do not match the values in the tweet.

I identified the following tidiness issues:

1) In tweets_clean:
    a. Rows relating to retweets and replies should be removed per project guidelines
    b. "id" and "id_str" are duplicative values in different formats
    c. the column "full_text" contains two items: text of the tweet and a link to the tweet that is not part of the original tweet. (I checked).
    d. there are columns that aren't useful to our analysis and clutter the dataframe, such as "truncated", "display_text_range", "source", "is_quote_status", "possibly_sensitive", "possibly_sensitive_appealable", "lang."
    e. the columns "entities", "extended_entities" and "user" contain multiple values that are in other columns, such as "id"," id_str", "created_at", and "lang".
    f. the column "expanded urls" had multiple duplicative urls instead of one per entry such as "source"

2) In image_pred_clean:
    a. column headers do not have descriptive names
3) In twit_archive_clean:

a. Rows relating to retweets and replies should be removed per project guidelines
b. Doggo, fluffer, pupper, puppo should not be separate column headers, but values in one column.
c. the column “text” contains two values: the text of the tweet and a link to the tweet that is not part of the original tweet. (I checked).
d. the “expanded_url” column has some values with duplicative (or more) urls instead of one

4) The DataFrames twit_archive_clean and tweets_clean, after cleansing, should be joined on “tweet id.”
    a. twit_archive_clean and tweets_clean have duplicative columns (or have duplicative information): “text” and “full_text”, “timestamp” and “created_at”
    b. “timestamp” in twit_archive_clean and “created_at” in tweets_clean are in different formats for date/time.

Cleaning data:

Data quality issues:

1) In tweets_clean:
    a. I converted “id” from integer to string format
    b. “created_at” was eventually dropped as part of merger with twit_archive_clean
    c. I filtered out/deleted the columns where almost of all or all of the values were null.
2) In image_pred_clean:
    a. I converted “tweet_id” from integer to string
3) In twit_archive_clean:
    a. I converted “tweet_id” from integer to string
    b. I converted “timestamp” in twit_archive_clean to a datetime object.
    c. I replaced the lowercase words such as “a”, “the” that appear as “name” value with the real name, if it was available in the text of the tweet. I fixed the rows where the value for “name” is missing or “None,” but is in the tweet.
    d. I fixed the rows where the values for numerator and denominator are missing or do not match the values in the tweet.

Tidiness issues:

1) In tweets_clean:
    a. I filtered the rows relating to retweets and replies because the instructions for this project stated that were not interested in analyzing retweets/replies, and then deleted the relevant columns for retweets and replies.
    b. I deleted the following columns because they contained multiple values in each entry, and these values were elsewhere in the dataframe: “entities", "extended_entities" and "user"
    c. I filtered and/or deleted “display_text_range”, “truncated”, “possibly_sensitive” and “possibly_sensitive_appealable”. “display_text_range” and “truncated” seemed like extraneous data that was available if needed from other columns and “possibly_sensitive” and “possibly_sensitive_appealable” seemed related to retweets, which we are not interested in.
    d. “id” was deleted b/c it was duplicative with “id_str,” which was renamed “tweet_id”
    e. the urls were removed from “full_text” and placed in a new column.
2) In image_pred_clean:
    a. column headers were renamed with more descriptive names
3) In twit_archive_clean:

a. I filtered the rows relating to retweets and replies because the instructions for this project stated that were not interested in analyzing retweets/replies, and then deleted the relevant columns for retweets and replies.
b. “doggo”, “floofer”, “pupper” and “puppo” became values in a new column, “stages”, and the original “doggo”, “floofer”, “pupper” and “puppo” columns were deleted.

4) twit_archive_clean and tweets_clean, after cleansing, were merged based on “tweet id.” Image_preds_clean was kept separate because its data did not directly pertain to specific tweets, but predictions from photos from tweets.