Data Wrangling for the Dog Rating Dataset

The data wrangling process consists of assessing, cleaning and setting up the data we will be using for analysis. I break it down into three steps when I perform my workflow:

- Assess: Identify what data is presented and get a general feel of the data.
- Identify: Once assessment is done, look deeper to identify specific issues that we might need to fix
- Clean: Fix the data we identified thus far to prepare for analysis.

Assessing

One of the first things I did was figure out what exactly is going on within the dataset. A lot of the time, the data presented is messy or unstructured - in fact, I assume this to be the default state for all incoming data. In our case, we notice the number of columns, what each column does, what kind of data is present in each column and other basic structural assessments.

Identifying

Once I've gotten a hang of the data, it's time to specify what issues we see and what we might want to do to fix it. Here:

- Multiple values in same column: https://twitter.com/dog_rates/status/882045870035918850/photo/1,https://twitter.com/dog_rates/status/882045870035918850/photo/1,https://twitter.com/dog_rates/status/882045870035918850/photo/1,https://twitter.com/dog_rates/status/882045870035918850/photo/1
- The four columns for dog classificaton (doggo, floofer etc.) can be condensed to one
- The data contained in all the dataframes are all part of one set of observations, and can be condensed into one dataframe.
- Lots of missing data in: (1)doggo, floofer, fluffer, puppo columns. We can probably merge data from the predictions into this one.
- Expanded URL's has replicated data: https://twitter.com/dog_rates/status/888554962724278272/photo/1,https://twitter.com/dog_rates/status/888554962724278272/photo/1,https://twitter.com/dog_rates/status/888554962724278272/photo/1,https://twitter.com/dog_rates/status/888554962724278272/photo/1
- Retweets can be deleted
- Since we don't care about retweets, we don't need these columns: retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp,in_reply_to_status_id, in_reply_to_user_id
- Since we don't have predictions for posts after August 1st, 2017, remove those.
- Since the rating denominator is the same for all ratings, we can remove the rating_denominator column.
- Decimals don't seem to be handled properly. Row 342 for example, has a 9.75 rating in the text, but the integer is 75.

Now that we've identified specific issues to fix, focusing on and getting them cleaned is less of a hassle. In terms of the specific issues above, here's how I fixed them:

- Delete rows that have in_reply_to_status_id values, meaning its not a retweet
- Delete retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp,in_reply_to_status_id, in_reply_to_user_id & rating_denominator
- Clean up the sources column to only include the text.
- Remove duplicated links; split on ',' and only keep the first for the expanded_url's column
- Parse and remove the urls from the text, and put it into it's own column.
- Use the prediction dataset to put in data in the columns if the prediction was true
  We'll be transposing the #1 prediction onto the dataset, since that's the highest confidence prediction. There can be more cleaning done for this though, namely, the p1 prediction is not always true, but one of the others might be. I didn't do that here.
- Remove the four dog classification columns and replace it with one
- Try to find classifications for the dogs from the tweet text and set it for the classification column. If not present, then no problem.
- Instead of using "this is.." as the precursor, use "named.." (and other such words that indicate that the next word is a Name) and get the word after that to get names for the dogs. We will atleast get a few more names.
- Find which rows have floats instead of ints for the rating, and fix them manually (there are only a few)