Reporting: wragle_report

- Create a **300-600 word written report** called "wrangle_report.pdf" or "wrangle_report.html" that briefly describes your wrangling efforts. This is to be framed as an internal document.

Gathering

The image_predictions.tsv file was downloaded from the server, and with the help of Twitter's API, I retrieved tweets from the WeRateDogs account using the tweet ids in the twitter-archive-enhanced.csv file. The retrieved tweets was saved in a file and then loaded to a dataframe called `tweet_df`. The image_prediction.tsv and twitter-archive-enhanced.csv files were also loaded in two separate dataframes called `img_preds` and `we_rate_dogs_archive` respectively.

Assessing

I majorly assessed the `we_rate_dogs_archive` dataframe. These were the following data quality issues I have found so far:

1. tweet records missing retweet and favorite counts (completeness issue)
2. The `expand_urls` column has more than one url (some of which are all the same) lumped up as a string (validity issue)
3. `source` column contains html tags (validity issue)
4. missing data represented as `None` in `name, doggo, floofer, pupper` and `puppo` columns (validity issue)
5. `in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id,` and `retweeted_status_user_id` expressed as float (validity)
6. some tweets also have their retweet records in this table with duplicate dog ratings (consistency issue) e.g the tweet_id: 873337748698140672 is a retweet of 873213775632977920 with duplicate dog ratings.
7. `timestamp` and `retweeted_status_timestamp` is an object instead of datetime (validity issue)
8. The `name` column has invalid names such as a, an, the, quite e.t.c (accuracy issues)
9. wrong ratings: `rating_denominator` not a multiple of 10 for some records. (validity issue)

and these were the tidiness issues I noticed:

1. `doggo`, `floofer`, `pupper`, `puppo` are values of a variable

2. two rating columns (`rating_numerator` , `rating_denominator`)

Cleaning

To fix the first quality issue, I merged `tweet_df` with `we_rate_dogs_archive` on the `tweet_id` with an inner join. For the second issue, I ensured each record has only one expanded url. The source url was extracted from the html tag to fix the third issue; "None" was replaced with "NaN" in the `name` , `doggo` , `floofer` , `pupper` and `puppo` columns. The `in_reply_to_status_id` , `in_reply_to_user_id` , `retweeted_status_id` , and `retweeted_status_user_id` columns were converted to strings. I also did the same to the `tweet_id` column for the sake of consistency.

Records whose `retweet_status_id` are the same as the `tweet_id` of other records were dropped. The `timestamp` and `retweeted_status_timestamp` columns were converted to datetime. Non-capitalized names were dropped from the `name` column. Similarly, records with invalid rating denominators were also dropped.

To fix the tidiness issues, I created a `dog_stages` column and dropped the `doggo` , `floofer` , `pupper` , and `puppo` columns. Then I divided `rating_denominator` by `rating_numerator` to form a single `ratings` column after which I dropped the `rating_numerator` and `rating_denominator` columns.