Wrangling Summary

Gathering

Twitter-archive-enhanced. I loaded into a dataframe using the pd.read_csv function.

Image-predictions.tsv. I had to get this file using the requests.get function and then I loaded it into a dataframe using the same function as the first file.

Tweet_json.txt. I got this using tweepy. I looped through all the tweet_ids in the first dataset to get all the available information from the API. Once the loop finished I used the pd.read_json function to load the responses into a dataset. I used this new dataset to create another dataset with only 3 columns: id, retweet_count and favorite_count.

Assessing

Quality issues

- There are 59 records without an associated image.
- There are 181 retweets that from this dataset perspective would be considered duplicate records.
- More than 800 records have an inaccurate name. 745 set up as None and 55 set up as 'a'.
- There is a record that has both doggo and floofer. It should only be floofer.
- Records with denominators different than 10.
- Records with numerator far from the median.
- Numerator should be 11 instead of 27. Ideally it should be 11.27 but for this exercise I will leave it as 11 because most of the numbers are integers.¶

Tidiness

- The columns doggo, floofer, pupper and puppo should be one column because it is one variable.
- The retweet_count and favorite_count should be on the master dataset.

Cleaning

Used the process explained: Define, Code, Test.

Fixing quality issues.

1. Remove the 59 records without an associated image. I removed the records using the isnull() function and saving the results in the copy of the dataset.

2. Remove the 181 retweets that from this dataset perspective would be considered duplicate records. I removed the records using the isna() function and saving the results in the copy of the dataset.

3. Replace all the names which start with a lowercase letter to 'None'. Used regex and the replace() function to replace all the words starting with lowercase to None.

4. Change to only doggo the record that has both fluffer and doggo. The word floofer was used to mention an owl. I used .loc to change the one field that required to be changed.

5. Remove records which description does not include a rating. Once the results were analyzed, that would be records 516 and 1662. I used the .drop function to drop the 2 rows that did not have ratings.

For tasks 6 through 9 I used the .loc function to modify the values. I created a mask to make the call of the .loc function more easy to read.

6. Correct tweet_id record 740373189193256964. The numerator and denominator should be 14/10.

7. Correct tweet_id 722974582966214656. The numerator and denominator should be 13/10.

8. Correct tweet_id 666287406224695296. The numerator and denominator should be 9/10.

9. Numerator should be 11 instead of 27. Ideally it should be 11.27 but for this exercise I will leave it as 11 because most of the numbers are integers.

Fixing Tidy Issues.

Tidy issue 1. The columns doggo, floofer, pupper and puppo should be one column because it is one variable. I created a function to populate a single column with the values of doggo, floofer, pupper and puppo. Then I removed the extra columns from the dataset.

Tidy issue 2. Retweet and favorite statistics should be merged into the twitter_archive dataset. I used the function merge on tweet_id to accomplish this task. Then converted the counts to integer datatype.

Tidy issue 3. Include image predictions into the twitter archive dataset. I used the function merge on tweet_id to accomplish this task.