WRANGLE REPORT

As introduced in the introduction, I will explore four different questions:

- Which dog stages receive the highest rating
- What is the relationship between breeds and rating
- What are the most popular dog names
- What dog stages have the highest favorite count

I will identify the tidiness and messiness of the dataset, but will not address all of them because not all of the tidiness, messiness problems relate to the research questions above.

ISSUES

Messiness Issues

df:

(1) Missing data in columns: in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, expanded_urls

(2) This dataset includes retweets, but I will not use the retweet dataset. Thus, I have to filter out those rows that are retweets

(3) Timestamp and retweeted_status_timestamp is an object. Need to be fixed into timestamp format

(4) The expanded_url includes all possible links within that URL. If one wants to use this link, they will have to filter out one link only

(5) Some dog names are suspiciously not names. I will have to delete them out of the dataset.

(6) The fact that the rating numerators are greater than the denominators does not need to be cleaned. This unique rating system is a big part of the popularity of WeRateDogs. I will process this to calculate the true rating

Img_df:

(1) dog breeds are funnily not consistent in all p1,p2,p3 in at least 20 first data. I created a subset of random 20 rows, the same thing happens. I will have to arbitrarily decide how to choose values from thse columns.

df_tweet_json: tweet_id format (currently object)

Tidiness Issues

- **df**: 'doggo','floofer','pupper','puppo' columns should be binary (Yes, No/ 0,1) rather than repeating the whole column name/ Nan
- **img_df and df_json**: Need to merge with the df dataset
- **img_df**: There are three prediction algorithms. I will have to create a new algorithm to choose the best prediction out of all possible values.

ACTION

1) All timestamps to be changed into datetime format
2) One column for dog stages: doggo, floofer, pupper, puppo
3) Dog breed: Depending on the algorithm function. If 1st algorithm performs too bad, move to the next one. Threshold: 30% for each. If none reaches 30%, it's empty
4) Change the prediction dog breed to all lower case (so that it can be standardized among different inputs)
5) Delete retweets
6) Dog ratings get standardized, calculate by using numerator divided by denominator
7) Merge the copied df_clean, img_df_clean, and json_clean data frame
8) Remove columns no longer needed: in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, and retweeted_status_timestamp
9) Tweet_id format for merging
10) Remove rows with non-dog names