The wrangling process:

Gathering Data:

First, I had three sources to gather the data from, the twitter_archive_enhanced.csv which was delivered by Udacity resources and can be read directly using pd.read_csv, the second is the image_predictions.tsv file, which had to be downloaded within the program first, then opened using the same method as before.

Additionally, more data is gathered using twitter API, at this stage I faced a problem verifying my developer account in twitter which is required to get access to the API. To solve this problem, I used the tweet_json.txt file provided from Udacity.

Assessing Data:

After gathering data, the next stage was to assess it looking for quality and tidiness issues.

Quality issues:

1-so many missing values in these 5 columns: in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id,retweeted_status_timestamp.

2-timestamp is string not datetime object.

3-"None" instead of Nan in dog breeds columns and name column.

4-rating numerator values aren't correctly extracted as they're integers in the rating_numerator column.

5-"rating_numerator" has a minimum value of 0.

6-"rating_numerator" has a maximum value of 1776 and 'ratings_denominator" has a maximum value of 170.

7-"+0000" in timestamp column.

8-the denomenator is inconsistent (isn't always 10).

9- 55 dogs names as "a".

10-the source column mostly has the same value for all entries.

11-(project requirement) some tweets don't have images and this is observed by finding that there are missing values in the expanded_urls column. moreover, some tweets don't have images in the image_predictions data frame.

Tidiness issues:

1-"doggo", "floffer", "puppo" "pupper" are in 4 columns instead of 1.

3-the image_predictions dataframe should be reshaped.

2-the twitter_archive and twitter_api data frames should be merged.

Cleaning Data:

This was the most time demanding part of the project.

- Made copies of the two data frames that have quality or tidiness issues, then I started the cleaning process by dropping some irrelevant columns with so many missing values,as well as the “source” column which has almost the same value in all rows using drop function .
- Replaced “+0000” in timestamp column with empty space using replace function.
- Converted timestamp column to datetime object using pd.to_datetime.

- Iterated over the name column and the breed columns and changed any "None" to np,nan using replace function.
- I merged the clean_image_predictions jpg_url column to the clean_tw_arch df based on tweet_id column to determine which tweets have no photos and drop them.
- Replaced all values in name column that is ”a” to be NaN.
- Tried extracting the rating_numerator column from the text column and converted it to float instead of int in the process. This didn't solve the problem entirely, so I had to go through some of them manually to check for validation and dropped all that aren't valid. Then I noticed that when the denominator is greater than 10, this means it's a rating for a group of dogs, so I took the mean of each dog rating and placed it as the numerator. Then I dropped the denominator column.
- Melted the 4 dogs breeds column into 1 column using melt function.
- Reshaped the clean_img_pred df using pandas wide_to_long function.
- Merged the clean_tw_arch df with api_df in a new master_df.

Storing Data:

This was the easiest part of the project, I saved the new master_df as well as the reshaped clean_img_pred data frames into two csv files using pandas to_csv function.