WeRateDogs WRANGLING DOCUMENTATION

Gathering:

We had 3 main sources of data:

- twitter_archive_enhanced.csv, which we obtained from the Udacity website. This file contained information on the tweets.
- image_predictions.tsv, which we downloaded programmatically from the Udacity server. This file contained neural net predictions of the images/videos in the tweets.
- tweet_json.txt, which we obtained by working with the Twitter API "tweepy". We obtained the favorite_count and retweet_count for every tweet from here.

The first 2 files were obtained easily and had no issues. While obtaining data from the API I noticed that 13 tweets could not be downloaded no matter the number of attempts. The tweet_id for these tweets are in the "failed" list in our notebook.

I also extracted "expanded_url" from the API despite there already being a copy of it in the twitter_archive_enhanced.csv. I did this because I noticed some tweets in the 'twitter_archive_enchanced.csv' file had missing expanded_urls.

Assessing:

We will mainly be assessing the image_predictions.tsv and twitter_archive_enhanced.csv files.

After gathering the data, we assess it visually and programmatically as required. We obtained the following quality and tidiness issues:

Quality:

Archive Table:

- Change tweet_id datatype to string
- Delete 'retweeted_status_id' and 'retweeted_status_user_id' and all retweet rows.
- Delete rows that have null in 'expanded_urls' . These rows have no images.
- Convert timestamp to datetime format.
- Data in 'in_reply_to_status_id' and 'in_reply_to_user_id' saved as float. Should have been saved as string.
- Delete row 1016, tweet_id = 746906459439529985. This row has no dog in it.
- Delete row 932, tweet_id = 754011816964026368 from the archive table.
- Clean up source.
- Increase column width to be able to read complete text or url.
- Delete rows with rating_numerator = 0 or rating_denominator = 0.

Image Table:

- Find rows that are not dogs and delete them
- Change tweet_id datatype to string

Tidiness:

- Add retweets_count and favorite_count to the archive table by merging archive and tweet_df

- Combine 'floofer', 'pupper', 'doggo', 'puppo' into 1 column, Age.
- Duplicate 'expanded_urls' column in tweet_df table made to check if we could fill any nulls in archive table. Delete this column later.

I am keeping the image predictions in a separate dataset since I feel it is a separate observation as per the rules of tidy data. The "twitter_archive_master" dataset has all the information related to the tweets from twitter, while the "image_predictions" dataset contains all neural net results that was obtained on the images.

Cleaning:

Now I cleaned the quality and tidiness based on the assessments I made using manual and programmatic steps.

I first made a copy of each dataset to test the changes. Each issue was resolved by breaking it down into 3 basic steps:

- Define
- Code
- Test

I have kept detailed comments in the notebook for each issue.

After going through all the issues, I saved them into csv files, “image_master.csv” and “twitter_archive_master.csv”.