WRANGLING REPORT

By Long Nguyen

I. Overview:

Initial data:

- **twitter_archive_enhanced.csv** - WeRateDogs Twitter Archive data
- **image_predictions.tsv** - Data that predicts dog breeds based on image through each tweet (through tweet ids) according to neural network
- **tweet-json.csv** - additional data via Twitter API collect favorite and retweet counts of each tweet

Wrangle:

- Gather data
- Access data
- Cleaning data

II. Wrangling

1. Gather data:

- **twitter_archive_enhanced.csv** has already been provided by Udacity
- Using the url provided and request library in Python, I write in **image_predictions.tsv** in binary mode
- For **tweet_json.csv,** I convert **tweet_json.txt** to csv file. The content in **tweet_json.txt** was written using Twitter's API and Tweepy library

2. Access data:

Quality:

1. **Twitter Archived Enhanced Data:**
    - Poor data format (**timestamp** shouldn't be object)
    - **source:** html format

- **name :** There are a lot of observations have 'None' as name (should be null) then there is name like 'a' (invalid name)
- retweeted_status_user_id: There are 181 retweets (not NAN values)
- **rating_numerator, rating_denominator:** There are a few unusual records for rating_numerator 1776,960,666 while their rating_denominators are relatively low. For rating_num = 960, its denominator = 0, which is invalid (row 313).
- **rating_denominator:** It seems like typical rating denominator should be 10 (2333 records). There are 23 records that have rating_deno not equal to 10

2. **Image Prediction Data:**
    - There are only 2075 observations while the archived Twitter data has 2356 records. This indicates there are 281 tweets will no image but not text

3. **Additional Data via the Twitter API:**
    - Only 2059 observations (missing IDs)

Tidiness:

1. **Twitter Archive Enhanced Data:**
    - There is no reason to have 4 separate columns of dog's types (doggo, floofer, pupper, puppo). We can create only 1 column called dog_type
2. **Image Prediction Data:**
    - missing entries
3. **Additional Data via the Twitter API:**
    - missing entries

4. **All:**
    - Merging 3 datasets into **twitter_archive_enhanced.csv**

3. Cleaning data:

Copy: archive_clean, image_df_clean, tweet_clean (3 copies of original data)

Define: (code, test are included in the ipynb. file)

- Remove 181 retweets and drop columns not in use
- 'None' to NA values in **name** column
- Drop 4 dog types columns
- Merge 4 redundant columns of dog types in 1 column - **dog_type**
- **timestamp** in datetime format
- Remove columns not needed

Merging: (twitter_archive_master)

- Right join **archived_clean** to **image_df_clean** because I only take 2075 ids that can be image-predicted
- Left join the above merge with **tweet_clean**
- Join on **tweet_id**

III. To CSV. file: Using pandas.to_csv() - **twitter_archive_master** dataframe to csv