Wrangle report

1. Gather

WeRateDogs Twitter archive, tweet image predictions:
With pd.read_csv() function and the url addresses, I directly download the datasets and transform those datasets into pandas.DataFrame.

Data from Twitter API:
Using tweepy, I could get data into Python dictionary format about each tweets. I stored those dictionaries into a Python list and after finishing scraping all data from Twitter API, I transformed the list into pandas.DataFrame. Below part is the code that I used.

2. Assess

The assessment was done visually and programmatically. Quality and tidiness issues of each dataframe were described as below. The common tidiness issue was that the three dataframes should be merged into one dataframe because each row in the dataframes is about the same single tweet.

WeRateDogs Twitter archive

Quality Issue

- There are retweets. (in the project motivation: You only want original ratings (no retweets) that have images. Though there are 5000+ tweets in the dataset, not all are dog ratings and some are retweets.) retweets start with 'RT' in the text column or have values in columns regarding retweets.
- Some ratings are not properly extracted especially decimals.
- The type of in_reply_to_status_id and in_reply_to_user_id columns should be int, not float.
- The type of retweeted_status_timestamp and timestamp columns are string, not np.datetime64.
- There are missing values in in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, and expanded_urls.
- Missing values in the name, doggo,pupper, puppo, and floofer are represented as 'None'
- There are strange dog names such as 'a', 'an', and 'the'.
- There are some strange values in rating_numerator and rating_denominator columns such as '1776'.
- URLs in ‘source’ column contains 'a' tag.

Tidiness Issue

- doggo, floofer, pupper, and puppo columns should be merged into one column dog_type.

tweet image predictions

Quality Issue

- There are somethings, not dogs (p1_dog==False).
- The number of rows is 2075. That means some tweets in df_dog_rating are not corresponded to the prediction dataset.

Data from Tweeter API

Quality Issue

- The number of rows is 2331. That means some tweets in df_dog_rating are not corresponded to the retweet_count and favorite count columns in this dataset.

Tidiness Issue

- doggo, floofer, pupper, and puppo columns should be merged into one column dog_type.

3. Clean

The issues that were identified in the assessment were transformed into a to-do list below.

WeRateDogs Twitter archive

- Remove (drop) retweet rows
- Drop unnecessary columns (here I decided to drop in_reply_to_status_id, in_reply_to_user_id, source, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, and expanded_urls).
- Change types of timestamp columns to np.datetime64.
- Change missing values in the name, doggo,pupper, puppo, and floofer into np.nan.
- Change strange dog names (e.g. 'a') into np.nan.
- Correct ratings that were extracted from decimal texts
- drop the rows where the denominator is not equal to 10.
- drop the row where the numerator is more than 15.
- merge name, doggo,pupper, puppo, and floofer into one column

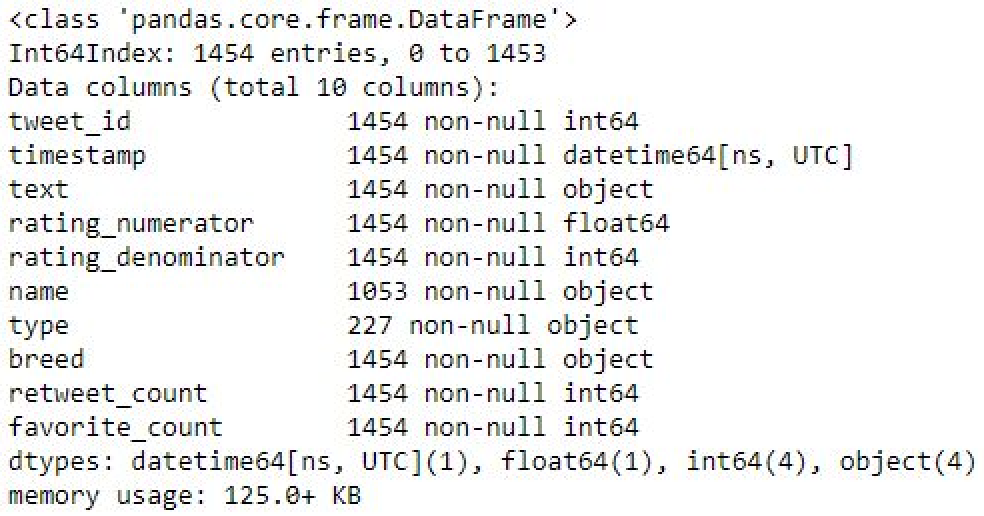
tweet image predictions

- drop the rows where the pictures were not predicted as a dog.

Common

- merge the three dataframe into one dataframe using an inner join.

Here, I decided not to remove NaN values in `name` and `type` columns because to get maximum records as much as possible regarding each column. Keeping those rows will be helpful when we investigate the dataframe without `name` and `type` columns. The number of values in each column like below.



```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1454 entries, 0 to 1453
Data columns (total 10 columns):
tweet_id              1454 non-null int64
timestamp             1454 non-null datetime64[ns, UTC]
text                  1454 non-null object
rating_numerator      1454 non-null float64
rating_denominator    1454 non-null int64
name                  1053 non-null object
type                  227 non-null object
breed                 1454 non-null object
retweet_count         1454 non-null int64
favorite_count        1454 non-null int64
dtypes: datetime64[ns, UTC](1), float64(1), int64(4), object(4)
memory usage: 125.0+ KB
```