Project: Wrangling data from Twitter archive known as 'WeRateDogs'

Dataset Introduction:

The dataset which will be wrangled is the tweet archive of Twitter user **@dog_rates**, also known as **WeRateDogs**.

WeRateDogs is a Twitter account that rates people's dogs. The denominator of the ratings is usually 10, but the numerators is most likely greater than 10.

What is required in the wrangle stage of this project:

1- Gathering Data for this Project

Data are gathered from three sources for this project:

a- WeRateDogs Twitter archive: the file twitter_archive_enhanced.csv is provided by Udacity and downloaded manually.
b- The tweet image predictions: this predicted what breed of dog is present in each tweet according to a neural network. This file (image_predictions.tsv) is hosted on Udacity's servers and is downloaded programmatically using the Requests library.
c- Twitter API and JSON file: using Tweepy library, tweet's retweet count and favorite count should be extracted for each tweet in the archive. this is then written to a JSON file (tweet_json.txt). The file is then read line by line into a pandas DataFrame.

2- Assessing Data for this Project

The result from gathering phase is three data frames. each data frame is assessed to quality and tidiness issues first visually (by printing them in Jupyter Notebook and try to find any issues) and then programmatically using different codes (info, describe, value_counts)

Quality issues:

twitter_archive:
1-Only keep original ratings (no retweets) that have images.
2-name column: rows 775 and 820 has partially extracted names.
3-name column has incorrectly extracted names like 'a' , 'an', 'the'
4-timestamp has object data type, should be changed to time stamp

5- Tweets have denominator other than 10 have incorrect extracted numerator and denominator
6- standardize the numerator and denominator ratings by dividing them and provide one value.
7- tweet_id has int type, should be of type object as no calculation is needed

image_predictions
8-66 duplicated jpg_url
9-inconsistent capitalization in p1, p2 and p3
10-tweet_id has float type, should be of type object as no calculation is needed

Tidiness issues:

1-twitter_archive: remove in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id,retweeted_status_user_id,retweeted_status_timestamp and source
2- image_predictions: dog breed prediction (p1, p2 & p3) and breed prediction confidence (p1_conf, p2_conf, p3_conf) could be combined into pred and pred_conf.
3- combine the three tables

3- Cleaning Data for this Project

In the cleaning stage all the addressed issues in the assessing phase are solved and tested. This resulted in three datasets which are ready for merge and analysis.

Some of the challenges in cleaning was the incorrect extracted names. There are many entries with incorrect names. Writing an algorithm to extract names was difficult as there is no common structure to indicate the name. The step for cleaning was to replace them with None. an alternative way to clean in may be using artificial intelligence algorithm which is beyond the scope of this course.

another issue which found need cleaning is some numerator has decimal which also resulted in wrong numerator value. these values where excluded from the analysis.