Project: Wrangle and analyze data from WeRateDogs Twitter account – Wrangle report

The data wrangling process of this project consists of:

1. Gathering data from three different sources:

    a) The WeRateDogs Twitter archive (twitter_archive_enhanced.csv). This file was downloaded manually
    b) The tweet image predictions, i.e., what breed of dog (or another object, animal, etc.) is present in each tweet according to a neural network. This file (image_predictions.tsv) has been downloaded programmatically using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv
    c) Using the tweet IDs in the WeRateDogs Twitter archive, query the Twitter API for each tweets JSON data using Pythons Tweepy library and store each tweet's entire set of JSON data in a file called tweet_json.txt file.

2. Assessing data

    The three data frames have been assessed visually and programmatically through the following Pandas functionalities:

    a) .head
    b) .info
    c) .describe
    d) .isnull
    e) .value_counts
    f) .duplicated

    At the end of the assessing process, the following data issues have been found:

    a) Data quality issues

    twitter-archive-enhanced.csv:
    - rating_denominator: Data type is int64
    - rating_numerator: Data type is int64
    - rating_denominator: Some entries have a denominator != 10.
    - in_reply_to_status_id: Remove column as consequence of the tidiness assessment (see notes below)
    - in_reply_to_user_id: Same as above
    - retweeted_status_id: Same as above
    - retweeted_status_user_id: Same as above
    - retweeted_status_timestamp: Same as above
    - name: change column title to "dog_name"
    - name: 745 missing names and several "a", "an", "old" values
    - timestamp: data type is object

- tweet_id: data type is int64 and should be changed to object
- 181 entries are retweets and should be eliminated since each observation must form a row
- 78 entries are tweets replies and should be eliminated since only original tweets are analyzed

image-prediction.tsv:
- p2, p2_conf, p2_dog, p3, p3_conf, p3_dog: These columns can be removed since not useful for the sake of the analysis
- tweet_id: data type is int64 and should be changed to object

tweet_df.csv:
- tweet_id: 274 "na" values

b) Data tidiness related issues

twitter-archive-enhanced.csv:

- **doggo, floofer, pupper, puppo**: merge these columns into one

- Merging all the datasets into one master table

3. Cleaning data

twitter-archive-enhanced.csv, image-prediction.tsv

Entries with rating denominator different from 10 might be referred to tweet with multiple dogs. Those have been removed for consistency.
Retweets and tweet replies have been eliminated as well as columns not useful for the sake of the analysis. Proper data types have been assigned to columns.

tweet_df.csv

Data type of column tweet_id has been changed to int64 to enable merging with twitter-archive-enhanced.csv. Null values have been removed.

Finally, The cleaned data frames have been merged into twitter_archive_master.csv.