Data Wrangling of 'WeRateDogs' Twitter data

In this document, I will describe my efforts in wrangling 'WeRateDogs' Twitter data. 'WeRateDogs' is twitter account for rating dogs. It has a unique rating system for rating dogs out of 10. After retrieving its basic tweets information, we will collect some more data from other sources and we will assess them then clean them. This is for the purpose of making insights from these data. Let's begin.

Step 1: gathering data

- I downloaded and loaded 'twitter_archive_enhanced.csv' into pandas data frame.
- The file (image_predictions.tsv) contains image predictions of dog breed for 'WeRateDogs' tweets and is hosted on Udacity's servers, so I downloaded it programmatically using the 'Requests' library and saved it into data frame.
- Each tweet's retweet count and favorite count was queried from the Twitter API for each tweet's JSON data using Python's Tweepy library. I loaded the JSON Data line by line into data frame.
-

Step 2: Assessing data

I have assessed the data visually using 'Numbers' App and also programmatically using pandas functions.

The following **Quality issues** was found :

in twitter_archive_enhanced.csv as (archive_df):

(1) 'source' columns : there is tags and unnessesary url, can be converted to : 'iPhone', 'Web client','Vine', ‘TwitterDeck'
(2) There is rows that represent retweets and replies. We only want original tweets so they need to be removed.
(3) 'retweeted_status_id', 'retweeted_status_user_id', 'in_reply_to_status_id', 'in_reply_to_user_id', 'retweeted_status_timestamp' columns : a lot of missing values. No need for those columns since we only want original tweets.
(4) 'rating_denominator': there are some values other than 10, some tweets contain more than one dog. it is better if all ratings are of 10.
(5) 'rating numerator': in row (47) is 5 while in tweet it is 13.5, this need to be corrected.
(6) 'name' column: some values are incorrect. It seem to be extraction issue. For example, row (58) dog name is 'a'.
(7) 'name' column: missing values are denoted 'None'. It should be NaN to identify it as missing.
(8) 'rating_numerator', 'rating_denominator' need to be converted to float64 to account for decimals.
(9) 'expanded_urls' column : 59 missing values.

(10) 'timestamp' column : need to be converted to 'datetime' type
(11) There is outliers in 'rating numerator' : for example 1775

I found one quality issue in image_predictions.tsv as (predictions_df):

(1) 'jpg_url' column : has 66 dublicated values

For **Tidiness issues** :

(1) in predictions_df we need to summrize the dog preed prediction into one column instead of p1, p2, p3
(2) in 'archive_df' dog stage information is found over 4 columns. It needs to me merged into one column named 'dog_stage'
(3) 3 dataset need to be merged into one dataframe by ‘tweet_id'

Step 3 : Cleaning data

For cleaning step I addressed each one of the issues mentioned in the assessment step and I fixed them one by one as I saw appropriate. Some erroneous values were dropped and unneeded columns were removed, some incorrect data types were converted. Also the separate dataframes were marked into one file and saved. For tidiness issues, I have summarized dog preed prediction into one column ‘dog_breed_prediction’. I have merged the cleaned Dataframes and I have merged dog stage information into one column named ‘dog_stage' by extracting the dog stage from the tweet again and make sure to account for tweets with multiple stages.

Finally, these efforts does not cover all cleaning process and data still has work to be done. For the purpose of visualizing the data, wrangling work I have done were adequate.