Reporting: wragle_report

Gathering Data

There are three sources of dataset for the analysis of WeRateDogs datasets, which are listed and explaine below:

1. **twitter-archive-enhanced.csv:** This archive is in csv file format and contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017. This was downloaded from thier twitter archive for the purpose of this project and read into the dataframe.

1. **image-predictions.tsv:** This contains table full of image predictions alongside each tweet ID, image URL, and the image number that corresponded to the most confident prediction (numbered 1 to 4 since tweets can have up to four images). This is hosted on Udacity's servers, it is in tsv file format and was downloaded programmatically using the
Requests library.

2. **tweet-json.txt:** This is the resulting data from twitter_api.py. This was read read line by line into DataFrame with (at minimum) tweet ID, retweet count, and favorite count.

```
        On assessing the 3 dataframes, the following are the issues encounter
and how they were dealt with.
```

Issue 1: Rows that have non-empty retweeted_status_id, retweeted_status_user_id, and retweeted_status_timestamp are
duplicated rows. **Solution 1:** Only rows that have empty 'retweeted_status_id' column were used for the dataframe.

Issue 2: From Df1_cleaned (twitter-archive-enhanced) In_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp and expanded_urls columns are not necessary.

Solution 2: Drop unnecessary columns from Df1_cleaned (twitter-archive-enhanced) using panda drop method.

Issue 3: Timestamp column is in 'object' dtype.

Solution 3: 'timestamp' in df1_cleaned to was converted to 'datetime' dtype using 'pd.to_datetime'.

Issue 4: Some values in rating_numerator column are less than "10".

Solution 4: 'rating_numerator' values that are lees than '10' was made atleast "10" using slice method.

Issue 5: Nulls are represented as (none) in name, doggo, floofer, pupper, and puppo column.

Solution 5: 'None' were replaced with 'NaN' in the name, doggo, floofer, pupper, and puppo columns using replace
function.

Issue 6: jpg_url shows duplicated counts of 66 when 'nunique' function were run on it.

Solution 6: All the 66 duplicated rows in 'Df2_clean (Image-predictions) were dropped using pandas drop function.

Issue 7: 'date_created' as it is identical to 'timestamp' on Df1.

Solution 7:'date_created' in Df3 (tweet-json) were dropped using pandas drop function.

Issue 8: doggo, floofer, pupper, and puppo column represents a single variable.

Solution 8: doggo, floofer, pupper, and puppo were merged into one column (dog_stage) in Df1 (twitter-archive-
enhanced) using bfill function.

Issue 9: The three datasets are supposed to be presented in one dataframe.

Solution 9: The three datasets where put in one dataframe using merge funtion.

Issue 10: 'tweet_id' column of the df_combined is a qualitative variable, but the dtype is in 'int'.

Solution 10: 'tweet_id' column dtype was converted to 'str'using astype.str function.

Issue 11: Dtype for img_num, favorite_count and retweet_count has changed after merging.

Solution 11: Dtype for img_num, favorite_count and retweet_count were converted to 'int' using astype function.

_

Issue 11: Dtype for img_num, favorite_count and retweet_count has changed after merging.

Solution 11: img_num, favorite_count and retweet_count were converted into 'int' dtype using astype function.

In [ ]: