Data wrangling project (WeRateDogs)

The goal of this project is to wrangle, analyze and visualize the tweet archive of Twitter user @dog_rates, also known as WeRateDogs, which is a Twitter account that rates people's dogs with a humorous comment about the dog.

Gathering the data:

First, I gathered each of the three pieces of data as described below in a Jupyter Notebook titled (wrangle_act.ipynb):

1-The WeRateDogs Twitter archive given by Udacity (twitter_archive_enhanced.csv) and imported it into a dataframe using panda's 'read_csv()'.

2-The tweet image predictions (image_predictions.tsv) hosted on Udacity's servers which display what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network. I used the Requests library to download the tsv file hosted on Udacity server and write it to 'image_predictions.tsv'.

3-Each tweet's retweet count and favorite count, Using the tweet IDs in the Twitter archive, by querying the Twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON data in a file called (tweet_json.txt) file. Each tweet's JSON data is written to its own line. Then the .txt file is read line by line into a pandas dataFrame with tweet ID, retweet count, and favorite count.

Assessing the data:

I used visual and programmatic assessment on each dataframe, and I came up with some quality and tidiness issues:

Quality issues:

Twitter archive data:

- *`tweet_id`* is int, should be boject (string).
- *`timestamp`* is object, should be datetime.
- *`in_reply_to_status_id`, `in_reply_to_user_id`, `retweeted_status_id`, `retweeted_status_user_id`* and *`retweeted_status_timestamp`* are not needed as they are not original tweets.
- `expanded_urls` contain only (2297) records, which means some records are missing.
- Invalid ratings in *`rating_numerator`* (max = 1776), *`rating_denominator`* (max = 170) and (min = 0) in both.
- *`name`* column contain invalid / missing names (none, a, an, the, lowercase names = 115).
- Null values are (none) in *`name`* column and dog breeds columns.
- Dog stages are objects, should be category dtype.
- *`source`* column is difficult to read.

Image prediction data:

- *`tweet_id`* is int, should be object (string).
- There are missing photos for some tweets (2075 instead of 2356).
- *`p1`, `p2`* and *`p3`* columns are object, should be category dtype.
- some predictions are lowercase, and some are uppercase.

Twitter API data:

- *`tweet_id`* is int, should be object (string).
- There's missing data (2331 out of 2356).

Tidiness issues:

- All dataframes should be combined in one dataframe.
- The *`doggo`, `floofer`, `pupper` and `puppo`* columns should be merged into one column.
- Image prediction dataframe columns headers are values, not variable names.
- *(p1, p2, p3), (p1_conf, p2_conf, p3_conf)* and (p1_dog, p2_dog, p3_dog) are in 3 columns, should be one.

Cleaning the data:

I created a copy of each dataframe that ends with (_cleaned) to perform the cleaning process on, then I used programmatic and manual cleaning on the identified issues:

Tidiness issue: **The *`doggo`, `floofer`, `pupper`* and *`puppo`* columns should be merged into one column.**
replacing *'none'* records with '' and creating a new column *`dog_stage`* containing the combination of the 4 columns, then dropping the original 4 columns and fixing some combined names.

Tidiness issue: **All dataframes should be combined in one dataframe.**
merging the three dataframes into one dataframe using pandas merge function.

Quality issue: ***in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id* and *retweeted_status_timestamp* columns are not needed as they are not original tweets.**
dropping *`in_reply_to_status_id`, `in_reply_to_user_id`, `retweeted_status_id`, `retweeted_status_user_id`* and *`retweeted_status_timestamp`* columns.

Quality issue: **expanded_urls contain only (2297) records, which means some records are missing.**
exclude tweets with empty `expanded_urls` record.

Quality issue: ***name* column contains invalid / missing names (none, a, an, the, lowercase names = 115).**
replacing invalid names with the correct names or np.nan if there is not.

Quality issue: **source column is difficult to read.**
simplifying the source column.

Quality issue: **Invalid rating values in *rating_numerator* (max = 1776), *rating_denominator* (max = 170) and (min = 0) in both.**
correcting `rating_denominator` and `rating_numerator` values based on `text` info.
changing the dtype to (int) which changed after merging.

Tidiness issue: **Image prediction dataframe columns headers are values, not variable names. / *(p1, p2, p3), (p1_conf, p2_conf, p3_conf)* and *(p1_dog, p2_dog, p3_dog)* are in 3 columns, should be one.**
changing the columns header and combining the columns into one column using 'wide_to_long' function.

Quality issue: **some predictions are lowercase, and some are uppercase.**
capitalize all records in `prediction` column.

The rest of cleaning process involved changing the incorrect dtypes.

Storing the data:

After finishing the cleaning process, I stored the *archive_cleaned* dataFrame as *twitter_archive_master.csv* file.