Wrangle report of tweet archive 'WeRateDogs'

Gathering data

Gather each of the three pieces of data in a Jupyter Notebook titled `wrangle_act.ipynb`:

1. The WeRateDogs Twitter archive. We had to download the file `twitter_archive_enhanced.csv` manually from Udacity link.
2. The tweet image predictions, i.e., what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network. The file with this data named `image_predictions.tsv` is hosted on Udacity's servers and we had to download it programmatically using the **Requests** library with the **following URL**: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv
3. Using the tweet IDs in the WeRateDogs Twitter archive, we had to query 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 should have been written on its own line. Then we had to read this .txt file line by line into a pandas DataFrame with tweet ID, retweet count, and favorite count.
   Before running the API request code, we needed to set up our own Twitter app. To do this, we created a **Twitter developer account**.

Assessing data

After gathering each of the above pieces of data, we assessed it visually and programmatically for quality and tidiness issues. We also detected and documented quality issues and tidiness issues.

We identified the following issues:

Quality issues

`df_1` table (twitter_archive_enhanced.csv)

- `tweet_id` is an integer and not a string.
- `timestamp` has incorrect data type. It should be datetime data type.
- the column `name` of the dog contains incorrect data (like "a", "an", "by", etc.) and also some names are defined like "None".
- some tweets don't have images, so we don't have predicted data for them.
- the columns `in_reply_to_status_id, in_reply_to_user_id` , `retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp` and `expanded_urls` have missing values.

`df_2` table (image_predictions.tsv)

- `tweet_id` is a integer and not a string.
- some of the image predictions are not a breed of dog.
- some dog breeds are in lowercase (`p1`, `p2`, and `p3` columns).

`df_3` table (tweet_json.txt)

- `id` is a integer and not a string.
- rename column name `id` to `tweet_id`.

Tidiness issues

`df_1` table (twitter_archive_enhanced.csv)

- `doggo`, `floofer`, `pupper` and `puppo` should be in one column `dog_stage`

`df_2` table (image_predictions.tsv)

- merge df_1 and df_ 2 dataframes using the `tweet_id` column into a single dataframe

`df_3` table (tweet_json.txt)

- merge df_3 with above dataframes using the `tweet_id` column in a single dataframe

Cleaning data

To clean up the data, we used various methods and functions of Python's libraries such as Pandas, Numpy and Re. All of the above issues have been eliminated.

The cleaned data was saved to a file `twitter_archive_master.csv` for further analysis.