Reporting: wrangle_report

- Create a **300-600 word written report** called "wrangle_report.pdf" or "wrangle_report.html" that briefly describes your wrangling efforts. This is to be framed as an internal document.

DATA WRANGLING PROJECT

- We Rate Dogs Data (@dograte)

Data Gathering :

I started my project by downloading the ‘twitter-archive-enhanced.csv’ file manually and read it into twitter_archive dataframe.

I created a folder named 'image_predictions' before I downloaded ‘image-predictions.tsv’ programmatically from Udacity's server using the requests library. Next I wrote it into image_predictions.tsv

Additional data was gotten from twitter API using Tweepy library after my application for developer access was granted. i made the request and got the tweet_json files needed to read the content into a dataframe. This was possible via reading the tweet_id from the ‘twitter-archive-enhanced.csv’ file, then I looped through each ID and query Twitter’s API with the ID to get each tweet’s JSON data. This took about half an hour to complete and returned some failed tweets which I handled with a try and except block.

Subsequently, I recorded the data in a text file named ‘tweet_json.txt’, with each tweet’s data written in a new line. After the query was completed and all the data was written in the text file, I read the text file line by line, obtained each tweet’s information (tweet ID, retweet count, favorite count, and followers count) using the json library, and appended the information into an empty list.

Finally, I convert the list of dictionaries to a pandas DataFrame and saved it into ‘tweets_api’

Assessing and Cleaning Data:

Some quality and tidiness issues were identified for the three tables. I created copies of all the dataframes and assessed the original while cleaning was done on the copies. Also created clones of the copies to test code methods along the line and compare results. This helped me avoid messing up the dataframes I was cleaning at that time.

Details of the issues identified are itemised below:

QUALITY

1 "twitter_archive" contains 181 retweets and 78 replies which are not needed and "tweet_id" is of type int64.

2 doggo, floofer, puppo and pupper columns have None for missing values

3 some dog names are clearly incorrect such as 'a', 'actually', 'an', 'very', 'all' etc. and None are used to represent missing values

4 'source' column contains html formatting

5 some columns not in the right datatype(s) - for example: IDs should not be int or float64 datatypes (for eg: tweet_id, in_reply_to_user_id, in_reply_to_user_id, retweeted_status_id and retweeted_status_user_id) and the 'timestamp' and 'retweeted_status_timestamp' not in datetime variable

6 some columns have plenty of missing data, for eg the in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp columns

7 some column names are not making sense - doggo, floofer, pupper and puppo for instance

8 some columns are not needed for our analysis.

Tidiness issues

1. doggo, floofer, pupper and puppo columns refer to the same thing and should be put in one column (traditionally called dog stage)

2. dataframe tracker shows there are three (3) dataframes (excluding copies), however, just one (1) comprehensive dataframe should be enough for our analysis.

Storing Data

Now that the data set is mildly clean and ready for analysis. I saved the master table to twitter_archive_master.csv, read it into a twitter_archive_master dataframe and began my analysis.

Data Visualization

I attempted a few plots and documented some of the insights I found from the data.

Plots:

- A Scatterplot of Twitter Likes and their Retweets
- Plot of The Correlation Between Likes & Retweets Of WeRateDogs
- Bar Plot showing the source distribution of tweets
- Plots On The Selection of Dog Names:
  - Plotting A Histogram To See The Most Common Dog Names
  - Plotting A Bar Chart To See The Most Common Dog Names
  - Plotting A Line Plot To See The Most Common Dog Names
- Bar Plot To See The More Common Dog Stages
- Plot of Changes Over Time
  - A Line Plot of The Retweet Count Over Time

        - A Line Plot of The Favorite Count Over Time

Insights:

1. pupper is the commonest dog stage

2. iPhone is the most used device by twitter users who engaged weratedogs account

3. Charlie and Lucy are quite popular dog names

Challenges:

Generally I didn't understand many things about this project at first and got confused a lot. I wasn't satisifed with my earlier submission though it had just a few corrections from the reviewer so I decided to redo the entire thing.

I still had to lean on the explanation from the session lead at the connect sessions, plenty of online materials and also tips from my colleagues on the nanodegree project to survive this phase and this made me unable to really communicate my findings with visualizations as I would have loved to or uncover insights from the data available.

Among the issues were:

- Breaking the contents of the text column to extract just a word or a group of words
- Reading and viewing images
- Melt() function
- Python regex
- Plotting just the top 5 values from a column in a dataframe
- Using cool data visualization techniques
- Using that Define-Code-Test method of analysis