Wrangle Report

The dataset being wrangled in this project is the tweet archive of Twitter user **@dog_rates**, also known as **WeRateDogs**.

Objectives:

- Data wrangling, which consists of:

  - Gathering data

  - Assessing data

  - Cleaning data

- Storing, analyzing, and visualizing your wrangled data

- Reporting on data wrangling efforts and data analyses and visualizations

Gathering data

- The WeRateDogs Twitter archive 'twitter_archive_enhanced.csv'. I was given this file.
- 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. This file ('image_predictions.tsv') is hosted on Udacity's servers and should be downloaded programmatically
- 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' to gather each tweet's retweet count and favorite ("like") count at minimum, and any additional data you find interesting.

Assessing data

Quality issues

1- timestamp is type 'object' it should be datetime format

2- missing data in the following columns: in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, expanded_urls

3- HTML tags still in the source column <a>

4- tweet id is int instead of str (object)

5- this dataset has retweets, that means there is duplicates

6- dog names have None , a , an as names

7- retweeted_status_timestamp is type 'object'

8 - Dog rating are not correct and need to be standardized

Tidiness

-The kind of the dog has to columns it should be one

-im dataframe and df_json can be joind with df dataframe

Cleaning data

1. Converting tweet_id to str in all dataframes
2. merging all dataframe in one
3. make one column for all dogs kinds
4. Delete retweets(duplicates)
5. remove unnecessary columns
6. converting Timestamps to datetime format
7. correct dog's names issues
8. standardize dog rating
9. inaccurate data