Wrangling Project Report

Introduction:

This project is finished in fulfillment of DAND's data wrangling section by practicing the Data Gathering, Data Assessing, and Data Cleaning techniques learned in the section. The project focuses on wrangling data acquired from the twitter archive of WeRateDogs account, which is an account that rates people's dogs out of fun and enjoyment. Therefore the ratings are mostly out 10.

Gathering Data:

In this project I dealt with three datasets:

- **Twitter Archive:** named twitter_acrhive_enhanced.csv was provided and downloaded manually from Udacity.
- **Tweet's Image predictions:** This file is provided by Udacity programmatically using URL and Requests library.
- **Twitter API & JSON:** The Code to generate the json file was provided by Udacity to be put as a comment in the Jupyter file since I didn't generate a Twitter Developer Account. The json file provided by the name twitter-json.txt is used instead and read line by line into a dataframe with tweet ID, retweet count , and favorite count.

Assessing Data:

The Assessing stage consist of looking through the three datasets to find any quality or tidiness issues by using programmatical methods like (info, describe, sample, info, duplicated, etc). The issues I chose to address are mostly in the twitter archive and image predictions dataset. First **the Quality issues:**

- Some columns are float instead of integers (in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id)

- Some columns are object instead of datetime (timestamp and retweeted_status_timestamp)

- rating_numerator and rating_denominator have invalid values (denominator should always be 10 and numerator should always be greater than 10 yet not that great)

- The name column has invalid names ('a', 'an')

- Breed names are not capitalized

- there are 66 pictures(jpg_url) that are duplicated
- Breed names are seperated by '_'
- Some image predictions are not dog breed.

The Tidiness issues:

- Dog stages are spread on four columns (doggo, floofer, pupper, puppo)
- Seperate tables for image_prediction and tdf are not needed.

Cleaning Data:

In this stage I addressed each of the 10 issues mentioned above as I judged convenient using the three steps method I learned from the cleaning section which are:

- Define: I wrote the issue to be cleaned.
- Code: I wrote the code to reach the target end.
- Test: I checked that the above code did its job.