WRANGLING REPORT

Submitted by Mohamed Abido as a part of the Data Analyst Nanodegree at Udacity.

Data wrangling consists of 3 steps:

- Gathering data
- Assessing data
- Cleaning data

Gathering:

Will be gathering each of the three pieces of datasets as described below :

1- The WeRateDogs Twitter archive. Download this file "twitter_archive_enhanced.csv" manually by clicking the following link: (https://d17h27t6h515a5.cloudfront.net/topher/2017/August/59a4e958_twitter-archive-enhanced/twitter-archive-enhanced.csv)

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. This file (image_predictions.tsv) is hosted on Udacity's servers and will be downloaded programmatically using the Requests library and the following URL: (https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv)

3- Each tweet's retweet count and favorite ("like") count at minimum, and any additional data you find interesting. Using the tweet IDs in the WeRateDogs Twitter archive, 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 be written to its own line. Then read this .txt file line by line into a pandas DataFrame with (at minimum) tweet ID, retweet count, and favorite count.

Assessing:

Quality

- Issue: Replies and Retweets should be removed.
- Issue: Some rows don't have images(expanded_urls).
- Issue: Some rows have invalid names in the name column.
- Issue: Some values in (p1, p2, p3) columns have '_' representing a space.
- Issue: In json data first column id should be tweet_id that can help when merging datasets later.
- Issue: tweet_id should be an object not an int.
- Issue: timestamp should be a date object not a str.
- Issue: Some columns are not useful.
- Issue: There's a problem with some of the extracted ratings from text. (tweet_id = 883482846933004288, 832215909146226688 and others).

Tidiness

- Issue: Dog stage values (doggo, floofer, pupper, puppo) should be under one column.
- Issue: Columns can have better arrangement
- Issue: There's no column that provides the gender of dogs. Create a gender column and detect the dog's gender from text.
- Issue: image_pred has many columns. only store the true value and its level of confidence.
- Issue: Some column names could have more descriptive names.

Cleaning:

Cleaning our data is the third step in data wrangling. It is where we fixed the quality and tidiness issues that we identified in the assess step. The final step of the cleaning process was the creation of a master Dataframe where we merged the 3 datasets and saved it in a .csv file.