DATA WRANGLING REPORT

INTRODUCTION

The dataset that is wrangled is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. This Twitter account rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10 and numerators always greater than 10. The twitter account has over 4 million followers and has received international media coverage. This report documents my wrangling efforts as part of the ALX/Udacity data analysis program.

GATHERING THE DATA

Three different datasets were used to in this project:

The WeRateDogs Twitter Archive: This file was provided by Udacity and was downloaded manually. Once downloaded, it was uploaded it and read into a pandas DataFrame.
The tweet image predictions: This file was downloaded programmatically using the requests library from the Udacity servers.
Additional data from the Twitter API: Using the tweet IDs in the WeRateDogs Twitter archive, the Twitter API was queried for each tweet's JSON data using Python's Tweepy library. Each tweet's entire set of JSON data was stored in a file called tweet_json.txt. The code template used for this was provided by Udactiy.

ASSESSING THE DATA

The data, was assessed visually and programmatically for quality and tidiness issues:

Visual assessment: Each piece of gathered data is displayed in the Jupyter Notebook for visual assessment purposes.
Programmatic assessment: Pandas' functions and methods are used to assess the data. Some of the identified issues include:

Quality

The WeRateDogs Twitter Archive:

- The following columns contain missing values: in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp and expanded_urls.

- tweet_id dataype should be an object and not int64
- timestamp should be a datetime64 dtype and not object

- Not all in rating_denominator column equals ten

- Not all in in rating_numerator column greater than ten

- Missing values in the name column indicated as 'none'. Also names that look incomplete such as 'a' are present.

Additional data from the Twitter API:

- Tweet IDs in the tweet_count table are less than those in the twitter_archive table.

The tweet image predictions:

- The datatype of the tweet_id column should be object instead of int64.
- Tweet IDs in the image_pred table are less than those in the twitter_archive table.

Tidiness

The WeRateDogs Twitter Archive:

- The dog stage is one variable and hence should form single column. But this variable is spread across 4 columns doggo, floofer, pupper, and puppo.
- Information about one type of observational unit (tweets) is spread across three different files/dataframes.

CLEANING THE DATA

The issues documented while assessing the data were cleaned. In cleaning the data, a copy of the original data was made. During cleaning the define-codetest framework was used. Cleaning included merging individual pieces of data resulting in a high-quality and tidy master pandas DataFrame.