This report will note the data cleaning and transformation efforts effected on the WeRateDogs Twitter archive data, the tweet image prediction data, and the archive supplemental data downloaded via Twitter's API. Upon downloading the three files, I discovered a number of quality and tidiness errors.

Gathering Data

There were three procedures to gather the data.

The WeRateDogs Twitter archive data was on a local hard drive and only required setting the path to the local directory and using pandas's read_csv function.

The tweet image prediction results data was hosted on Udacity servers and required the requests library. I used requests.get to retrieve the data and pandas's read_csv, with tabs as the separator to gather the prediction results.

I acquired data from the twitter API using the Tweepy library. Retrieving the actual Twitter data was done using a function that read in a list of the relevant tweet ids from the archive data, broke up the lists into chunks, and used Tweepy's status_lookup function to query for tweet ids. A JSON file was created from the results, and pandas's read_json to turn it into a dataframe.

Assessing Data

Both visual assessments, using .head() and .sample(), to review the data's layout, and .describe(), .info(), and slicing filters to review data types and identify duplicate and/or missing data. After using these, I identified the followed Quality and Tidiness errors from the three data sets.

From the WeRateDogs Twitter archive data:

Quality

- timestamp data is an object, should be a timestamp
- all status_ids, and user_ids are floats, should be objects
- name column contains values that are not names
- value in floofer field should be a yes/no, and category type

Tidiness

- doggo, pupper, puppo - should be one column and category type

From tweet image prediction data:

Quality

- tweet_id is an int, should be an object
- rename columns p1, p1_conf, and p1_dog to more clear column names
- for p1 column values, replace underscore with space and make title case

From the Twitter API:

Quality

- id is an integer, should be an object
- Retweet and favorite counts are changed to decimal, revert to integer

Tidiness

- all data is in a tweepy.model.resultset and must be converted to a useful json before it can be loaded into pandas

Data Cleaning

Cleaning used a variety of methods, ranging from quick methods such as to_timestamp to change a data entry to a datetime object, as astype to change a column from one type to another, to moderately complex methods like applying a lambda function to a column, to more complex means, such as writing functions and applying them to a dataframe. For example, resolving the names required using regular expressions to capture text after a word and overwriting the error entries. Columns that contained data that was not useful or analysis was dropped to keep the dataframes easy to read. Some issues ended up generating errors that could only be resolved through reviewing similar errors on Stack Overflow. While some significant quality and tidiness errors were found and resolved, others remain.