# Wrangle Report

Wrangle and Analyze Data Project Submission
Data Analyst Nanodegree Program
Jessica Ertel

## Data Gathering

There were three elements of data gathering involved in this project. The WeRateDogs Twitter archive was provided by Udacity and downloaded manually. Using the tweet IDs from the WeRateDogs Twitter archive, I queried the Twitter API for each tweet's JSON data and wrote the data to a text file. This was the most time-consuming portion of the data gathering phase. In order to use the scraped tweets in the text file, I imported them into a Pandas DataFrame. Finally, I programmatically downloaded tweet image predictions according to a neural network from Udacity's servers using the Requests library.

## Data Assessment

After gathering the data, I assess them visually and programmatically for quality and tidiness issues. While there are many areas that can be cleaned, I limit this project to addressing 2 tidiness issues and 8 quality issues, including those that satisfy the Project Motivation.

## Data Cleaning

### Tidiness Issues

I merged the three gathered data sets into a single DataFrame because they each contain information from the same observational unit, in this case tweets. The second tidiness issue The original data contained four different columns documenting the dog stage as either doggo, pupper, puppo or floofer. Some dogs fell into multiple categories. I decided to consolidate this information into one column.

### Quality Issues

The first step was to ensure the new master DataFrame only included original tweets, no retweets or replies. In removing the non-empty retweet and reply rows, I was left with several columns that were only null values. These columns were dropped. Addressing incorrect data types was an important data quality issue. I changed 'tweet_id' to a string and updated the two timestamp columns to a pandas datetime object. After comparing the timestamp columns to ensure consistency, I dropped the duplicate.

Next I addressed several inaccurate dog names that had been identified from the 'text' column. I found the invalid dog names since they had only lowercase characters – it was clear from the list that these were not likely dog names, so I replaced them with 'None'. As described in the Project Motivation, the fact that the rating numerators are greater than the denominators does not need to be cleaned. However, there were several records with denominators over 10. After closer inspection I realized that some of the scores were incorrectly pulled from the 'text' column. I updated the scores where appropriate.

Next I moved to image prediction columns. It was important to change the strings in all the columns containing predictions to lowercase, so that golden_retriever wouldn’t be classified separately from Golden_Retriever. Finally, I changed some of the column names in the DataFrame to make them more intuitive.