# Wrangle Report

First of all, this project is based on a dataset from Twitter, a tweet archive on WeRateDogs. Which is a Twitter account that rates people’s dogs with comments and photos. Often, some very amusing comments and eye-catching dog pictures.

The Wrangle\_Act project goals are as following:

* Wrangling the twitter data through the following processes:
  + Gathering data
  + Assessing data
  + Cleaning data
* Storing, analyzing, and visualizing your wrangled data
* Report of the wrangling process
* Report on the data analysis and visualizations

## Gathering Data

Data wrangling can be a process. This project proved to be challenging in a few areas. However, I felt it enhanced the learning process. Most notable was accessing the json files. I was unable to get a developer’s license from Twitter. I used the json files provided by Udacity.

Wrangle\_act project consisted of gathering data from the following sources:

* The WeRateDogs Twitter archive. The twitter\_archive\_enhanced.csv file was provided to Udacity students. WeRateDogs is a Twitter archive and sent it to Udacity for use in this project. This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017. It is a real-world example of data sets today. The project reinforces the importance of developing data wrangling skills.
* The tweet image predictions, i.e., what breed of dog or sometimes random pictures of other animals and or things. Images are present in each tweet according to a neural network. An additional file provided to Udacity students.
* Json files based on each tweet’s retweet count and favorite count. Other variable data was provided but at a minimum, tweet ID, retweet count and favorite count had to be utilized. I had some trouble accessing the json data. I had to do some research in Udacity’s community Knowledge Center. I did manage to get connected.

## Assessing Data

At this point, the data is loaded into Pandas data frames and ready to make an assessment on quality and tidiness.

Assessment Findings

archive:

* Tidiness:
  + Multiple columns for dog stages which include 4 classifications: doggo, floofer, pupper and puppo
  + Dog stages should be one column with 4 variable types; named dog\_type
  + Redundant headers with missing data in the following columns: in\_reply\_to\_status\_id, in\_reply\_to\_user\_id, retweeted\_status\_id, retweeted\_status\_user\_id, retweeted\_status\_timestamp, expanded\_urls
* Quality:
  + Extract dog stages from text to get accurate and updated classifications
  + dog names: some dogs have 'None' as a name, or 'a', or 'an.'
  + this dataset includes retweets, which means there is duplicated data (as a result, these columns will be empty: retweeted\_status\_id, retweeted\_status\_user\_id and retweeted\_status\_timestamp)
  + timestamp is an object
  + retweeted\_status\_timestamp is also an object (the other retweeted statuses are floats)
  + tweet\_id is an int (applies to all tables)

images:

* Tidiness:
  + p1, p2 and p3 columns are confusing headers, rename for quality
* Quality:
  + p1, p2 and p3 columns aren't consistent when it comes to capitalization: sometimes the dog breed listed is all lowercase, sometimes it is written in Sentence Case.

twitter\_counts\_df:

* Quality:
  + missing some data

## Cleaning Data

After the assessment, I cleaned the data through the following means:

Define, Code and Test

1. Merge the clean versions of archive, and api\_data
2. Correct the dog types
3. Create one column for the various dog types: doggo, floofer, pupper, puppo Remove columns no longer needed: in\_reply\_to\_status\_id, in\_reply\_to\_user\_id, retweeted\_status\_id, retweeted\_status\_user\_id, and retweeted\_status\_timestamp
4. Delete retweets
5. Remove columns no longer needed
6. Change tweet\_id from an integer to a string
7. Change the timestamp to correct datetime format
8. Correct naming issues
9. Create 2 separate csv files, one merged archive and api\_data, named twitter\_archive\_master.csv the second predictions table renamed df\_images\_clean.csv