# Wrangle Report

## Introduction

In this project, the tweet archive of Twitter user @dog_rates, also known as WeRateDogs will be wrangled and analyzed. This account rates people's dogs with a humorous comment about the dog and typically a rating out of 10. WeRateDogs asks people to send photos of their dogs which are rated on a scale of one to ten, but are invariably given ratings that exceeds the maximum.

## Steps

- Data wrangling, which consists of:
  - Gathering data
  - Assessing data
  - Cleaning data
- Storing, analyzing, and visualizing our wrangled data
- Reporting on 1) our data wrangling efforts and 2) our data analyses and visualizations

## Gathering

- Enhanced WeRateDogs Twitter archive: twitter_archive_enhanced.csv
- 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 should be downloaded programmatically using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv
- Additional data via the Twitter API: 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. The Twitter APIs was reached twice.

## Assessing

After gathering data from a csv file, tsv file, and Twitter's API, the data was later assessed for quality (i.e. content) and tidiness (i.e. structural) issues. The following issues were found:

*Quality*:

- In the twitter_archive table, some dog names are missing (labeled as 'None')
- In the twitter_archive table, there are incorrect dog names
- In the twitter_archive table, there are retweets
- In the twitter_archive table, the timestamp column is a string instead of date/datetime/timestamp
- In the twitter_archive table, the sources are not readable
- In the twitter_archive table, the texts contain links
- In the twitter_archive table, in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, source, retweeted_status_user_id, and retweeted_status_timestamp are not needed and need to be removed
- In the tweets_json table, the 'id' column name does not match the name in the other 2 tables

- In the image_predictions, there are tweets with no images
- In the image_predictions table, some of the dog breeds are lowercase instead of uppercase

*Tidiness*:

- In the twitter_archive table, the dog "stage" variable is in four columns (doggo, floofer, pupper, puppo) instead of one
- 3 separate datasets which should be combined into one

## Cleaning

The previously mentioned quality and tidiness issues in the assessment step were fixed and resolved. First, copies of the 3 data frames were created before cleaning and then followed the process: define, code, test. The final combined data frame was stored as a csv file called 'twitter.csv'. At that point, the data was wrangled and ready for the data analysis and visualization.