# Wrangle Report

# Project Details

this project is as follows:

- Data wrangling, which consists of:
    - Gathering data (downloadable file in the Resources).
    - Assessing data
    - Cleaning data
- Storing, analyzing, and visualizing your wrangled data
- Reporting on 1) my data wrangling efforts and 2) my data analyses and visualizations

# Gathering Data

Gathering Data for this Project

Gather each of the three pieces of data as described below in a Jupyter Notebook titled wrangle_act.ipynb:

1. The WeRateDogs Twitter archive. Download this file (twitter_archive_enhanced.csv
2. The tweet image predictions, i.e., what breed of dog (or another 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
3. Twitter API for each tweet's JSON library and store in a file called tweet-json.txt file.

# Assessing and Cleaning Data

1. **Assessing Twitter archive data results**

- in_reply_to_status_id, in_reply_to_user_id, retweeted_status_user_id and retweeted_status_timestamp have high percentage of missing values so we will drop them
- tweet_id is stored as int , it should stored as object
- timestamp is stored as object , it should stored as datetime
- extract source as text and stored it as catogray
- convert Nona and a values in name to nan
- extract dogs stage from text, then drop doggo, floofer, pupper and puppo columns
- the fact that the rating numerators are greater than the denominators does not need to be cleaned. This unique rating system is a big part of the popularity of WeRateDogs.

**2. Assessing Image predictions data results**

- tweet_id is stored as int , it should stored as object
- img_num is stored as int , it should stored as category
- p1_cof always has the highest confidence percentage than p2_cof and p3_cof
- creat two columns 'confidence' and 'dog_breed' with confidence percentage refer to dog breed

## 3. Assessing Twitter API data data results¶

- tweet_id is stored as int , it should stored as object