# Wrangle and Analyze Data

# Project Overview

## Introduction

The dataset that I will be wrangling (and analyzing and visualizing) is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. Why? Because "they're good dogs Brent." WeRateDogs has over 6 million followers and has received international media coverage.

WeRateDogs downloaded their Twitter archive and sent it to Udacity via email exclusively to 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.

## What Software Do I Need?

- You need to be able to work in a Jupyter Notebook on your computer. P
- The following packages (libraries) need to be installed. You can install these packages via conda or pip. Please revisit our Anaconda tutorial earlier in the Nanodegree program for package installation instructions.
    - pandas
    - NumPy
    - requests
    - tweepy
    - json
- You need to be able to create written documents that contain images and you need to be able to export these documents as PDF files.

# Project Specifications

## Code Functionality and Readability

- All project code is contained in a Jupyter Notebook named wrangle_act.ipynb and runs without errors.
- The Jupyter Notebook has an intuitive, easy-to-follow logical structure. The code uses comments effectively and is interspersed with Jupyter Notebook Markdown cells. The steps of the data wrangling process (i.e. gather, assess, and clean) are clearly identified with comments or Markdown cells, as well.

## Gathering Data

Data is successfully gathered:

- From at least the three (3) different sources on the Project Details page.
- In at least the three (3) different file formats on the Project Details page.

Each piece of data is imported into a separate pandas DataFrame at first.

## Assessing Data

1. Two types of assessment are used:
    - Visual assessment: each piece of gathered data is displayed in the Jupyter Notebook for visual assessment purposes. Once displayed, data can additionally be assessed in an external application (e.g. Excel, text editor).
    - Programmatic assessment: pandas' functions and/or methods are used to assess the data.
2. At least eight (8) data quality issues and two (2) tidiness issues are detected, and include the issues to clean to satisfy the Project Motivation. Each issue is documented in one to a few sentences each.

## Cleaning Data

- The define, code, and test steps of the cleaning process are clearly documented.
- Copies of the original pieces of data are made prior to cleaning.
- All issues identified in the assess phase are successfully cleaned using Python and pandas.
- A tidy master dataset with all pieces of gathered data is created.

## Storing and Acting on Wrangled Data

- Save master dataset to a CSV file.

- The master dataset is analyzed using pandas in the Jupyter Notebook and at least three (3) separate insights are produced.
- At least one (1) labeled visualization is produced in the Jupyter Notebook using Python’s plotting libraries.

## Report

Two reports:

- Wwrangling efforts are briefly described in wrangle_report.pdf.
- The three (3) or more insights the student found are communicated in act_report.pdf including visualization.