# Wrangle Report

## Introduction

The purpose of this project is to put in practice what I learned in data wrangling data section from Udacity Data Analysis Nanodegree program. The dataset that is wrangled 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 comments.

## Project Tasks

1. Gathering data
2. Assessing data
3. Cleaning data

## Gathering Data

- Twitter archive file: the twitter_archive_enhanced.csv was provided by Udacity and downloaded manually.
- The tweet image predictions, i.e., what breed of is present in each tweet according to a neural network. This file (image_predictions.tsv) is hosted on Udacity's servers and was downloaded programmatically using the Requests library and URL information.
- Twitter API & JSON: by using the tweet IDs in the WeRateDogs Twitter archive, I queried the Twitter API for each tweet's JSON data using Python's Tweepy library and stored each tweet's entire set of JSON data in a file called tweet_json.txt file. I read this .txt file line by line into a pandas dataframe with tweet ID, favorite count, retweet count.

## Assessing Data

Once the three tables were obtained I assessed the data as following:

- Visually, I used two tools. One was by printing the three entire dataframes separate in Jupyter Notebook and two by checking the csv files in Excel.
- Programmatically, by using different methods (e.g. info, value_counts, sample, duplicated, query, etc).

# Assessment Issues

## Twitter Archive Dataset

### Quality

- We only wan tweets" that have images in Image Predictions dataset .
- We only want original ratings no retweets nor replies
- There are columns that won't be used for analysis.
- Consistency issue nan is written sometime as None or $nan.
- Weird values in name column like ('this', 'unacceptable','all',"such") .
- Timestamp column data_type is string (object) instead of DataTime.
- Some denominators values are > or < than 10 or = 0.
- Some numerator values seem outliers.
- Missing data in multiple columns.

### Tidiness

- The dog stage has 4 stages (values in headers).

### Image Predictions

- Undescriptive columns names. (p1,p2,p2,p1_conf,p2_conf,p3_conf,P1_dof,p2_dog,p3_dog).
- Duplicates.
- There are columns that won't be used for analysis.

### Tidiness

- There is 3 columns for classifications algos and another three for algos confidentiality (values in headers).

### Cleaning Data

This part of the data wrangling was divided in three parts: Define (Assessments issues), code and test the code. These three steps were on each of the issues described in the assess section.

First and very helpful step was to create a copy of the three original data frames. I wrote the codes to manipulate the copies. If there was an error, I could create a new copy from the original. Whenever I made a mistake, I could create another copy of the data frames and continue working on the cleaning part.

Then I solved every issue I found in the assess section. The most challenging part for me Was denominator and numerator values ,because I don't how to begin correctly (e.g. I thought I should drop rows with values >10 for denominator) and when I tried to merge the 3 data frames using merge function and got too many extra rows!