# Wrangling Report

## Intoduction

The purpose of this project is to practice what I learned in Data Wrangling section from Udacity Data Analysis Nanodegree program . 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.

## Project Goals

- Gathering data
- Assessing data
- Cleaning data

## Project Details

### Gathering Data :

The Data of this project was obtained from three different resources as following:

- **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, followers count, friends count, source, retweeted status and url.

### Assessing Data

Once the data was obtained, I started assessing data as following:

**Quality Issues:**

**'twitter-archive-enhanced-2.csv'**:

- **Completeness** : 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.
  Name column has (Nane , a, the ).
  Some tweets aren't original tweets.

- **Accuracy**: Erroneous datatypes (doggo, floofer, pupper and puppo columns).
  Incorrect data in rating_numerator and rating_denominator.

- **Consistency**: date and time, ids, representations of null values as string "none" ,
  These columns will be empty {retweeted_status_id, retweeted_status_user_id and retweeted_status_timestamp }.

**'image_predictions.tsv':**

- **Validity:** p1, p2 and p3 columns have invalid data , labeled photos as (starfish, paper towel , orange).

**'tweet_json':**

- **Completeness**: missing tweets.

**Tidiness Issues:**

**'twitter-archive-enhanced-2.csv':**

- The last 4 columns all relate to same category

**'image_predictions.tsv':**

- This data set is part of the same observational unit.
- 'twitter-archive-enhanced-2.csv' - one table with all basic information about the dog ratings.

**'tweet_json':**

- Also part of the same observational unit - one table with all basic information about the dog ratings.

## Cleaning Data

First step was to create a copy of the three original dataframes. I wrote the codes to manipulate the copies. If there was an error, I could create a new copy from the original
Cleaning was divided in three parts: Define, code and test the code.
on each of the issues described in the assess section as following :

- keep original ratings (no retweets) that have images.
- Delete columns that won't be used for analysis
- Melt the doggo, floofer, pupper and puppo columns to *dogs* and *dogs_stage* column
- Change the timestamp to correct datetime format.
- Separate timestamp into day - month - year (3 columns)
- change numerator and denominators type int to float to allow decimals
- Correct denominators and numerators.
- Editing name column and updating (none - weird names [a,aa,the])
- Drop duplicated images urls.
- Create 1 column for image prediction and 1 column for confidence level