# Wrangle Report

## Introduction:

Gather data from a variety of sources and in a variety of formats, assess its quality and tidiness, then clean it. This is called data wrangling.

The dataset I worked on 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 wrangling Process go through some steps:

- Gathering data
- Assessing data
- Cleaning data

And ends with Storing, analyzing, and visualizing the wrangled data.

## Step 1: Gathering.

There are 3 resource of data

- WeRateDogs Twitter archive (CSV file)
- tweet image predictions (Using the Requests library)
- tweets retweet count and favorite using twitter APIs (Text file)

In my case unfortunately twitter reject my request, so I used twitter Json file provided to me by Udacity.

## Step 2: Assessing.

After gathering all necessary data, I start assessing data for Quality & Tidiness Issues.

### Quality Issues:

- timestamp & retweeted_status_timestamp type in twitter_archive_enhanced should be datetype.
- tweet_id type in twitter_archive_enhanced should be string
- p1,p2,p3 in image_predictions table should be renamed to clear meaning
- wrong names in twitter_archive_enhanced ( like : a , an , the)
- source column content in twitter_archive_enhanced contain HTML link tags surrounding the text.

- Some tweets are not original tweets "retweets"
- rename id in tweet_rt_fav table to tweet_id and covert it to string
- doggo, floofer, pupper, and puppo columns have values with None instead of NaN
- missin rows in tweet_rt_fav and image_predictions.(2 missing row in tweet_rt_fav and 281 missing row in image_predictions )
- missing expanded_urls in twitter_archive_enhanced
- extraction of ratings of some rows are not correct
- The rating_numerator column should of type float.

### Tidiness Issues:

- doggo, floofer, pupper, and puppo column in twitter_archive_enhanced better to be one column with this value (doggo, floofer, pupper, and puppo).
- combine 3 data resources to be one dataset.

## Step 3: Cleaning.

- This step is last process of wrangling data after assessing data. I follow this process of cleaning (Define, Code, Test).
- change timestamp & retweeted_status_timestamp type to datetime.
- Separate timestamp column into 2 columns date and time.
- change tweet_id type to string and rating numerator to float.
- Correct wrong extracting of rating.
- Correct wrong names by replacing wrong names by NaN.
- renaming p1, p2, p3 column names to clean names to become (1st_prediction, 2nd_prediction, 3rd_prediction).
- Correct source column content in twitter_archive_enhanced table to be without <a> tag.
- Dropping retweets tweets.
- Rename id column to tweet_id.
- Change doggo, floofer, pupper, and puppo have values with "None" to NaN.
- Combine all 3 dataframes together.
- Store datafram to Csv file.