# Wrangle Report

## 1. *Gather*

The act of gathering the data was achieved using three different techniques.

a. The WeRateDogs Twitter Archive

The twitter archive was provided to us from WeRateDogs through Udacity, and it was imported into the jupyter notebook using pandas and was stored on dataframe.

b. The Tweet Image Predictions

The predications for the dog breeds was provided through Udacity, but this time I have to download the data programmatically using the requests library.

URL for the data:

*https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv*

c. Each Tweet's Retweet Count and Favorite Count

The information for the tweets wasn't provided from Udacity. I used the Tweepy library for querying the twitter API for more information regarding the tweets.

First I had to set up a developer, which I already had. Then use the credentials that twitter provided me with to query the Twitter API using the Tweepy library.

## 2. *Asses*

The act of assessing the data was done both visually and programmatically, visually I used Excel to view the data we have easily, and programmatically I used several built-in functions from pandas to view and get more information about our datasets.

Issues that were detected in the assessing phase:

## *Quality*

### twitter_archive

- Some tweets have rating denominator != 10

- Some tweets have rating numerator < 10
- Some tweets have rating numerator that are outliers.
- Missing some expanded URLs.
- Some dogs have incorrect names.
- Reply and Retweet columns are not needed.

### image_prdes

- Missing rows (2075) instead of (2356).
- Inconsistency with naming of dog breed.
- Remove entries that have p1_dag as false.
- Remove columns(p2, p3, p2_dog, p3_dog, p2_conf, p3_conf) as we will only be using the p1 related columns.

### df_tweets

- Missing rows (2327) instead of (2356).

## *Tideness*

- All three datasets should be joined.

### twitter_archive

- Merge dog stages(doggo, floofer, pupper, puppo) into one column.

## *3. Clean*

The act of cleaning our data was done programmatically using many pandas built-in functions, and some that I created to help with cleaning our data set.

In this phase all issues that were discussed in the assessing phase have been solved.