# WRANGLING REPORT

We Rate Dogs Data

## Gathering Data Phase:

The data was gathered from three sources:

1. The WeRateDogs Twitter Archive: is a file that we download manually and upload to the workspace. Once it is downloaded, we upload it and read it to dataframe.

2. The Tweet Image Predictions (image_prediction.csv): This is a file produced by running every image in the WeRateDogs Twitter Archive through a neural network. We downloaded this file programmatically using Requests library from a provided URL.

3. Additional data from the Twitter API: I was unable to use the tweeter api because I did not get the approval so I used the code provided by udacity to query twitter Api and download the last file. tweet_json.txt: This is the resulting data from twitter_api.py. Then read the tweet_json.txt file line by line into a panda DataFrame with (at minimum) tweet ID, retweet count, and favorite count.

## Assessing and Cleaning Data Phase:

I used both visual assessment and programmatic assessment to assess the data:

❖ Visual assessment: each piece of gathered data is displayed in the Jupyter Notebook for visual assessment purposes.

❖ Programmatic assessment: pandas' functions and/or methods are used to assess the data.

### Quality issues & solutions:

| Issues | Solution |
|---|---|
| Incorrect datatype in timestamp column | Convert the datatype using to_datetime function in pandas |
| A lot of error in dog names | Convert the error names to none |
| Some denominators have values that are not equal to 10 | Convert all the values to 10 |
| Some expanded URLs have two URLs | Extract the first URL from all rows with multiple URLs |
| Text column contains description and URL | Extract the URLs from the text column and removed them |
| Source column is in HTML format | Ectract the text from the HTML format |
| the text in P columns sometimes start with uppercase and other times start with lowercase. | Convert the text in the P columns to uppercase characters. |

### Tidiness issues & solutions:

| Issues | Solutions |
|---|---|
| Dog stages data are separated into 4 columns | Melt all the columns into one column called dog_stage |
| Three dataset are separated into 3 tables | Merge all the tables |

## Storing Cleaned Data:

Now the data set is clean and ready for analysis. I saved the cleaned data to twitter_archive_master.csv.