# **PROJECT: WeRateDogs\_Tweets Data Wrangling and Analysis**[¶](http://localhost:8888/notebooks/Documents/Result%20of%20Analysis/UDACITY/WeRateDogs_wrangle_act.ipynb#Project:--WeRateDogs_Tweets-Data-Wrangling-and-Analysis)

**WRANGLING REPORT**

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 numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. Why? Because “they’re good dogs Brent.” WeRateDogs has over 4 million followers and has received international media coverage. The following datasets were used for this analysis:

Enhanced Twitter Archive:

The WeRateDogs Twitter archive contains basic tweet data for all 5000+ of their tweets, but not everything. One column the archive does contain though: each tweet’s text, which was used to extract rating, dog name, and dog “stage” (i.e. doggo, floofer, pupper, and puppo) to make this Twitter archive “enhanced.” Of the 5000+ tweets, only tweets with rating was filtered.

Additional Data via the Twitter API:

Back to the basic-ness of Twitter archives: retweet count and favorite count are two of the notable column omissions. Fortunately, this additional data can be gathered by anyone from Twitter’s API. Well, “anyone” who has access to data for the 3000 most recent tweets, at least. But you, because you have the WeRateDogs Twitter archive and specifically the tweet IDs within it, can gather this data for all 5000+. The resulting data will be read line by line into the pandas Dataframe with (at minimum) tweet ID, retweet\_count and favorite count.

Image Predictions File:

Every image in the WeRateDogs Twitter archive was ran through a neural network that can classify breeds of dogs. The resulting data is a table full of image predictions (the top three only) alongside each tweet ID, image URL, and the image number that corresponded to the most confident prediction (numbered 1 to 4 since tweets can have up to four images).

This Report documents the steps taken in gathering, accessing, cleaning and storing the datasets.

**Gathering the Datasets:**

Three source was given for the data gathering exercise, the first source was a CSV file provided by Udacity and downloaded from website, and it was read into the pandas. The second was a TSV file extracted from a page from Udacity website using requests library and also writing the data into a file that was loaded into pandas. The third source was to be gotten from twitter API, but due to some inconveniences, I resulted to using the TXT file provided by Udacity, it was obtained using requests, loaded with JSON and was loaded into the pandas DataFrame.

**Accessing the Datasets:**

The datasets were accessed programmatically and visually, and some issues were identified. They include;

**Quality Issues**

**df\_twitter\_archive:**

1. Retweeted\_status\_id, retweeted\_status\_user\_id and retweeted\_status\_timestamp should be dropped since we are only interested in tweets.

2. In\_reply\_status\_id, in\_reply\_to\_user\_id and expanded\_urls columns contain missing values and will be dropped since they are not needed for the analysis.

3. Tweet\_id should be a string ('object') and not an integer ('int64').

4. Timestamp should be 'datetime' not 'string'(object).

5. Some values in rating\_denominator column are not '10'(Standard denominator for WeRateDogs).

6. Some values in rating\_numerator column are equal to Zero and some are less than 10.

7. Null values are represented with none in name column.

**df\_images:**

8. Tweet\_id should be 'string' not 'int'.

9. Names in p1, p2 and p3 columns have some starting with upper letters and others with lower letters.

**df\_tweets:**

10. Id column name should be 'tweet\_id' not 'id'

11. The to-be-named tweet\_id should be 'string' not 'int'.

**Tidiness issues**

**df\_twitter\_archive:**

1. Doggo, floofer, pupper and puppo should be in one column not four.

**df\_images:**

2. Columns with 'conf' should be renamed with 'confidence'.

**df\_twitter\_archive`, `df\_images` & `df\_tweets:**

3. All three dataframes should be combined into one dataframe.

**Cleaning the Dataset:**

After the above issues were cleaned, the datasets were merged into a single dataset called ‘df\_weratedogs\_archive’ and was ready for analysis.

**Storing the Data:**

The dataset was stored in a CSV file called ‘weratedogs\_archive.csv’