# wrangle_report

June 28, 2022

## 0.1 Reporting: wragle_report

## 0.2 Project Overview

- 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. WeRateDogs has over 4 million followers and has received international media coverage. WeRateDogs downloaded their Twitter archive and made it available for this project. This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017.

```
In [ ]:
```

## 0.3 Data Gathering

### 0.3.1 About the data

- There are 3 separate datasets available in this project. The Twitter archive dataset, the Image prediction dataset and an additional data from twitter API.

**Data Gathering Process**

- Twitter Archive Data
  - The twitter archive data was provided in a csv file format, made available for download by Udacity via the link below twitter_archive_enhancement and saved in a dataframe named `twitter_archive_df`. This file was manually downloaded from the above url onto my machine and further uploaded into my Udacity workspace.
- Image Url Prediction
  - This file (image_predictions.tsv) is present in each tweet and hosted on Udacity's servers. This is downloadable programmatically by using the python Request library through the following URL: image_predictions.tsv. This dataset was subsequently loaded as `img_url_df`

  *NB*: This file is in a .tsv format and should be treated as one to avoid errors.
- Data Obtained from Twitter API

  - Gather each tweet's retweet count and favorite ("like") count at the minimum and any additional data you find interesting. Using the tweet IDs in the WeRateDogs Twitter archive, query the Twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON data in a file called tweet_json.txt file. However, for simplicity, Udacity has provided this file in a json file which was manually downloaded and loaded into my workspace using the python json library into a dataframe called `tweet_json_df`

### 0.3.2 Data Assessment

- Assessing data can be done in 2 ways. The visual assessement and programmatic assessment. These 2 techniques were employed in my assessment stages.

#### Visual Assessment (Quality Issues)

- After loading the data into a pandas dataframe, a quick glance and skim was enough to instantly spot a few issues. Some of these include;
  - In the `twitter_archive_df`, the source column contains html anchor tags. This column contains both the useful text and unwanted tags and attributes
  - From the project description, our data doesn't require any retweet, this is a good key to quickly find records that have @RT(Retweets) records in the text column. These records have to be removed.

#### Programmatic Assessment (Quality Issues)

- By programmatic assessment, data quality issues that were hidden in the 3 datasets were revealed. These quality issues would have rather been difficult to spot visually. Pandas have a number of built in functions that were used to programmatically assess the datasets. The .info() function was called to quickly see the datatypes and any missing values if any. The .value_counts() was also used to take a quick look at unique value counts of series data. A quick glance at the datasets using the .head() and .sample() function. These were enough to reveal the following quality issues;

  - The `timestamp` column is an object datatype instead of a datetime datatype. Also, the `source` column is best converted to a categorical datatype rather than the object datatype
  - Some dog names start with articles such as `a, an, actually, one, all`. Most of these names starting with lower cases are most likely quality issues. The `one` is believed to be a None
  - In the `img_url_df`, each image has 3 values of dog breed and of which only one is valid. This has to be addressed.
  - Some `tweet_id`(s) of the `tweet_archive_df` are missing in the `img_url_df`
  - Some ratings have abnormal values in either the `rating_numerator` or `rating_denomenator` columns
  - Some unnecesary columnn such as `retweeted_status_id`, `retweeted_status_user_id`, `retweeted_status_timestamp` and `expanded_urls` from the `tweeter_archive_df` and `img_ur_df`

**Visual Assessment (Tidiness Issues)**

- The 3 datasets have a common unique column in `tweet_id` in all datasets, this a an indication that, there is a possibility to merge all datasets. Further investigation showed that, all datasets need to be combined together into a `twitter_archive_master.csv` dataset. The 2 tidiness issues that were observed in the datasets are;
  - Merge the `twitter_archive_df` and `img_url_df` and `tweet_json` tables
  - doggo, floofer, pupper and puppo columns in `twitter_archive_df` table should be merged into one column.

### 0.3.3 Data Cleaning

- Data cleaning was carried out immediately after the assessment of the data. The cleaning process included addressing the quality and tidiness issues that were discovered. A copy of each of the original dataset were created to prevent tampering with the original datasets in the process of cleaning. To start with, the issues is addressed using the Define, Code and Test template.

### 0.3.4 Data Storage

- Storing the cleaned data was important for the analysis and reproducability. This is done by saving the combined datasets into a `twitter_archive_master.csv` file by calling the `.to_csv()` function on the appropriate combined dataset.

```
In [ ]:
```