# wrangle_report

August 23, 2022

## 0.1 Reporting: wragle_report

- Create a **300-600 word written report** called "wrangle_report.pdf" or "wrangle_report.html" that briefly describes your wrangling efforts. This is to be framed as an internal document.

## 0.2 Introduction

In this project, I perform data wrangling, on the dataset obtained from Twitter user WeRateDogs. WeRateDogs is a Twitter account that gives ratings of dogs, usually with a humorous comment about the dog. For this project, I worked on three datasets. An enhanced Twitter archive data – This is from the WeRateDogs Twitter archive given to, and enhanced by Udacity Additional data via Twitter Api – I queried Twitter Api for additional data for each corresponding tweetID contained in the enhanced archive. Image Prediction File –This table is a classification of dog breeds from a neural network. It contains the top three image predictions with each tweetID, image URL, and the image number that corresponds to the most confident prediction. Tweepy to query Twitter's API for additional data beyond the data included in the WeRateDogs Twitter archive. This additional data will include retweet count and favorite count. ## Data Gathering

The WeRateDogs Twitter archive was provided as a downloadable CSV by Udacity. I downloaded it, and read the data into a pandas DataFrame. The tweet image predictions file was hosted on Udacity's website, and I used Python's request library to programmatically download the file. I queried Twitter API using Tweepy for additional data on the retweet counts, and favorite count of the corresponding tweetID This additional data includes retweet count and favorite count.

## 0.3 Assessing Data

In the next step of the data wrangling process, I assessed the data, searching for quality and tidiness issues that made it unclean. First, I carried out a visual inspection, where I manually inspected the rows of the three tables. I noticed that some tweets were not original tweets, rather were retweets Next, I carried out a programmatic inspection, where I inspected the number of rows and columns, the presence of missing values, and the datatypes of each column. There were multiple issues with the dataset, that made it unclean, but I handled ten issues for this project, 8 quality issues, and 2 tidiness issues

### 0.3.1 1. Quality issues

1. Missing values (retweeted_status_id, retweeted_status_user_id and retweeted_status_timestamp) contains a lot of missing values

2. some tweets in text are not original tweets, but retweets we only want original tweets, not retweets
3. The dog_breed table have nondescriptive column names
4. rating_denominator has a minimum value of 0.00 and a maximum number of 170.00 this is invalid, as all the rating denominators must be 10
5. timestamp is a string, should be astype datetime
6. name column contains incorrect details ie a, an, such, the
7. replace None for Nan in Doggo, Floofer, Pupper, puppo columns
8. source column contains the html tag

#### 0.3.2 Tidiness issues

1. doggo, puppo, flufffer needs to be merged into one column
2. the additional data table need to be joined with the data from twitter archive

### 0.4 Cleaning

To clean the datasets, I carried out the following; * Made a copy of the datasets before working * Dropped columns that contained retweets, and after that, dropped the assciated Columns (retweeted_status_id, retweeted_status_user_id and retweeted_status_timestamp) * Changed p1, p2, p3 to prediction1, prediction2, prediction3 respectivels. These are more descriptive * Standardied the denominators of the rating columns, setting all values = 10 * Converted timestamp clumn to a datetime object * Removed invalid dog names * Obatined a new column dog_stage from the tweet text, and dropped Doggo, Floofer, Pupper, puppo Columns * Removed the html tag of the surce code * Merged all three dataset into one master table

```
In [ ]:
```