# Wrangle Report

## Gather data

The following files were gathered for the analysis:

- **WeRateDogs Twitter Archive:** This file (twitter-archive-enhanced.csv) was downloaded manually and it contains over 2300 tweets from WeRateDogs twitter account.
- **The Tweet Image Predictions:** This file (image_predictions.tsv) was downloaded using the Requests library and is present in each tweet according to a neural network.
- **The Twitter API:** This file (json_file.txt) was created by querying each tweet's JSON data using Tweetpy library, and then reading each tweet's retweet count and favorite count from the JSON file into Pandas Dataframe.

## Assess data

The three gathered files were loaded into individual Pandas data frames for the assessment phase. The assessment was done both visually and programmatically. The following quality and tidiness issues were detected during the assessment.

### 1. Quality issues

Twitter Archive

- Keep only original ratings (no retweets) that have images and delete retweets.
- There are several records that contains a value != 10 in the rating_denominator column that needs to be removed.
- Rename some of the columns with more clear and appropriate names, which are timestamp, text, rating_numerator, name.
- Keep only the string in the Source column and remove the HTML tag.
- Remove hyperlinks in the text column to keep only the string.

- Fix the data type of the tweet_id column from int into string and the timestamp column to datetime.
- Delete unusfull column that does not give any usful information of the data.
- Remove tweets that are missing images after the merging.

Image Predictions

- Fix the data type of the tweet_id from int into string.

Twitter API

- Fix the data type of the tweet_id from int into string.

### 2. Tidiness issues

Twitter Archive

- Merge doggo, floofer, pupper and puppo columns into one dog_stage column.

Image Predictions

- Merge Image Predictions dataset into Twitter Archive dataset.

Twitter API

- Merge Twitter API dataset into Twitter Archive dataset as the tweets retweet_count and favorite_count should belong the the Twitter Archive dataset.

## Clean and storing data

The quality and tidiness issues were cleaned on copies of the original Dataframe. The define-code-test framework was used during the cleaning phase. Finally, the high-quality and tidy data was stored in a file called (twitter_archive_master.csv).