# Wragling Report - WeRateDogs Twitter

BY YANG, Linjing
2019/02/10

## Data Gathering

Data was gathered from 3 different sources.

### 1. Twitter Archive

The Twitter archive #WeRateDogs, a csv file that contains various kinds of information, such as dog ratings, stages, tweet ids, posted date, etc.

### 2. Additional Information via the Twitter API

The retweet and favorite count of each tweet gathered from Twitter's API by Tweetpy Library in a Json format.

### 3. Image Prediction File

The prediction of dog images and breeds downloaded programmatically by Requests Library from the neural network.

## Data Accessing

Data was accessed based on both of quality and tidiness issues.

### Quality Issues

#### Dataset 1

Completeness:

- in_reply_to_status_id: 78 out of 2356 is non-null
- in_reply_to_user_id: 78 out of 2356 is non-null
- retweeted_status_id: 181 out of 2356 is non-null
- retweeted_status_user_id: 181 out of 2356 is non-null
- retweeted_status_timestamp: 181 out of 2356 is non-null
- name: string 'None' should be replaced by Null

Validity:

- rating_denominator contains invalid values (e.g. denominator = 0)

- expanded_url just includes incomplete url, which contains "https://twitter.com/dog_rates/status/" plus part of tweet_id

Accuracy:

- rating_numberator contains extremely large values (e.g. 1776 when denominator = 10)

Consistency:

- source: the long name with HTML tag could be shorten.

Data Types:

- tweet_id: integer -> object
- timestamp: object -> datetime
- in_reply_to_status_id: float -> object
- in_reply_to_user_id: float -> object
- retweeted_status_id: float -> object
- retweeted_status_user_id: float -> object

### Dataset 2

Completeness:

- retweet & favourite count: "Nan" should be replaced by Null.
- retweet & favourite count: 16 missing values

Data Type:

- retweet & favourite count: object -> integer

### Dataset 3

Data Type:

- tweet_id: integer -> object

## Tidiness Issues

### Dataset 1

- The dog stages - doggo / floofer / pupper / puppo - could be combined into 1 column as categorical data.

### Dataset 3

- The column names could be more clear.

# Data Cleaning

- The three datasets were combined into one with more clear column names.
- The columns that could not contribute to analysis were dropped.
- The row that contains invalid values was deleted.
- All missing values were presented as Null.
- The data types of all columns were corrected.
- The dog stages were combined into one column as categorical data.
- The HTML tags in the ‘source’ column were deleted.