# Data Wrangling Report on “WeRateDogs” Tweets

This report is on the data wrangling procedures for the datasets from the WeRateDogs Twitter account. This project had three major stages: data gathering, data assessment, and data cleaning.

Data were acquired from three sources; (1) tweet archive data from WeRateDogs, (2) a downloaded file through the internet using the provided URL (tweet image predictions), and (3) the tweepy API (retweet and favorite counts). The datasets were loaded as; **archive**, **prediction**, and **tweets_api** and evaluated for quality and tidiness issues and cleaning.

For data assessment, standard python methods and functions like .head(), .value_countes(), .info(), .isnull(), .duplicated() and others were used and the following problems were identified:

## Quality issues

### *archive*

1. *only original tweets are needed for this project, these columns retweeted_status_id, retweeted_status_user_id, and retweeted_status_timestamp have some entries therefore are retweets*
2. *retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp variables have a lot of missing data and, need to be dropped*
3. *the name variable has many non-dog names such as 'None’, ‘a’, ‘all’, ‘an’, ‘by’, ‘his’, ‘just’, ‘my’, ‘not’, ‘such’, ‘the’, ‘this’, ‘very’*
4. *the '+0000' in the timestamp column is redundant information*
5. *the timestamp should be converted to datetime*
6. *the URLs in the source column contain the tweet sources embedded in them, and are to be extracted*
7. *some rating denominator entries are greater or less than the usual value of 10*
8. *the retweet_count and favorite_count columns have a few NaN values*

### *prediction*

9. *the values in the p1, p2 and p3 columns characters like '-', '_'*
10. *breed names in columns p1, p2 and p3 begin with lowercase letters*
11. *duplicate values in column jpg_url*

**Tidiness issues**

1. *Variables doggo, floofer, pupper and puppo will be combined into a new column 'stage' and then the redundant variables - doggo, floofer, pupper and puppo will be dropped*
2. ***For the purpose of this project, I would create a two new columns dog_type and confidence_level, which will be filled with dog breed names from predictions that were true for p1, p2 and p3***
3. *Information about one type of observational unit (tweets) is spread across three different files/dataframes. Datasets spread across the three datasets "archive", "prediction" and "tweets_api" into one*

Before the cleaning activities, copies of the original data were made as archive_original, prediction_original and tweets_api_original. During the cleaning stage for each issue detected, it was defined and a code was developed to implement the needed action, and a test was used to confirm the implemented action.

In the archive data frame, some variables such as retweeted_status_id, retweeted_status_user_id, and retweeted_status_timestamp had some entries and as such considered retweets. For the purpose of this project, retweets are not needed and these entries were dropped.

The non-dog names such as 'None', ‘a’, ‘all’, ‘an’, ‘by’, ‘his’, ‘just’, ‘my’, ‘not’, ‘such’, ‘the’, ‘this’, ‘very’ in the name variable were treated as missing entries and changed to NaN.

The timestamp variable was converted to DateTime. The tweet sources were embedded in the URLs in the ‘source’ column and were extracted into four categories (Twitter for iPhone, Vine, Twitter Web Client, TweetDeck).

A few rating_denominator values were either higher or lower than 10 and 10 is the general denominator, therefore, any higher or less value was equated to 10.

In the prediction data frame, new columns were created: dog_type and confidence_level. A function that adds values to the lists dog_type and confidence_level when P1 dog, P2 dog, and P3 dog are all True.

The three datasets "archive", "prediction" and "tweets_api" were merged as ‘tweeter_archive_master’. Some observations in the new dataset had missing values (NaN) in the retweet_count and favorite_count columns and were then dropped to make all columns that will be analysed to equal observations.

Finally, the dataset was stored as a .csv file: **twitter_archive_master.csv**.