# Report on data wrangling steps for the WeRateDogs Project

By Swathi Munikoti

WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog.

The objective of this project is

1. Wrangle data by
    a. Gathering Data
    b. Assessing Data
    c. Cleaning Data
        i. Define
        ii. Code
        iii. Test
2. Storing, analysing, and visualizing the wrangled data
3. Reporting on
    a. data wrangling efforts and
    b. data analysis and visualizations

Wrangle Data:

**Gathering Data:**

Gather three pieces of data as specified in Udacity's Project Details Page:

1. The WeRateDogs Twitter archive. This archive was in a file twitter_archive_enhanced.csv and was given by Udacity.
2. The tweet image predictions, i.e., what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network. This file (image_predictions.tsv) is hosted on Udacity's servers and had to be downloaded programmatically using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv

3. 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. Each tweet's JSON data should be written to its own line. Then read this .txt file line by line into a panda DataFrame with (at minimum) tweet ID, retweet count, and favorite count.

**Assessing Data:**

**Quality Issues -**

1. **Twitter_archive:**
    a. **Completeness:**

- Remove the columns in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, expanded_urls due to the missing data and are not needed for the analysis.

    b. **Validity:**

- Delete retweets, Select only rows that have null values in retweet related columns, retweeted_status_id, retweeted_status_user_id, and retweeted_status_timestamp columns.

- Select invalid dog names, which most probably starts with lower case letters and "none" to np.nan.

    c. **Accuracy:**

- Change the timestamp from object data type to correct datetime format.

2. **Image_Prediction:**
    a. **Completeness:**
- As the column img_num is not needed for the analysis, hence removed it.

    b. **Validity:**

- Some image prediction data, p1 and p2 and p3 prediction s have invalid dog names like nail, snail, desktop_computer. Remove images that are not dogs.

c. **Accuracy:**

- Delete duplicated url.

3. **Tweet_Details:**
   a. **Completeness:**

- Remove the column retweeted_status as it was not need ed for the analysis.

b. **Validity:**

- There are 162 retweets. Keep only original tweets.

c. **Consistency:**

- Change tweet_id, retweet_count, favorite_count from Object data type to int 64.

**Tidiness Issues -**

1. Twitter_archive - There are four columns namely doggo, floofer,puppo, pupper for the stages of a particular dog. We combine all four columns into single column with name dog_stage.
2. I merged all the three datasets collected from different sources into a single dataset and save it in a file called twitter_archive_master.csv

**Cleaning Data:**

I cleaned all the above issues specified in the assessing stage following the practise for each issue by defining the issue, making changes to the code and testing if the issue was cleaned.