# Project Wrangle Report

## Introduction

The project is about ‘WeRateDogs’ which is a is a Twitter account that rates people's dogs with a humorous comment about the dog.

## Gathering Data

Data was gathered from 3 different sources, the first dataset was a CSV file - the WeRateDogs Twitter archive that was downloaded manually from the udacity project space through the twitter_archive_enhanced.csv link and was loaded using the pd.read_csv, the second data set was cas a TSV file hosted on Udacity's servers and was downloaded programmatically using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv . The third dataset was supposed to be gotten by querying the twitter API of each tweet using Tweepy library but I used the tweet_json.txt file provided and it was read line by line into a pandas DataFrame with the minimum columns required. All data were loaded using Pandas.

## Assessing Data

The datasets were assessed visually and programmatically, and some quality and tidiness issues were detected.

### Quality Issues

Twitter Archive Dataset

1. The following columns are replies, had a lot of missing values and not needed in the analysis. So, they were dropped
    a. in_reply_to_status_id
    b. in_reply_to_user_id
2. The timestamp column had a wrong data type and was converted to datetime datatype
3. The source column had html tags that were not needed, the real sources were extracted
4. The rating numerators and denominators were correctly extracted from the text column
5. The datatype of correct rating numerators and denominators were changed to integers

6. Drop the original rating numerator and denominator columns
7. Some of the texts were not about dogs, so they were dropped
8. The following columns are retweets, had a lot of missing values and not needed in the analysis. So, they were dropped
    a. retweeted_status_id ,
    b. retweeted_status_user_id
    c. retweeted_status_timestamp
9. The expanded_urls column was not needed and was dropped
10. The rating_numerator and rating_denominator columns were deleted, since another one was correctly extracted from the text
11. The none values in doggo, floofer, pupper, puppo columns were dropped

Tweets Dataset

1. The id column was renamed to tweet_id to match with tweeter archive dataset

**Tidiness Issues**

Twitter Archive Dataset

1. The doggo, floofer, pupper, and puppo columns were combined into one

Image Dataset

1. The p1, p1_conf, p1_dog, p2, p2_conf, p2_dog, p3, p3_conf, and p3_dog columns were all combined into two columns breeds and confidence
2. The p1, p1_conf, p1_dog, p2, p2_conf, p2_dog, p3, p3_conf, and p3_dog columns were all dropped
3. All the 3 datasets were merged using join

## Cleaning Data

Before performing the cleaning, copies of the original datasets were made, during cleaning, the define-code-test framework was used and clearly documented. After cleaning, the individual pieces of datasets were merged into one according to the rules of tidy data.

## Storing Data

The merged dataset was stored store as a DataFrame in a CSV file with the name ‘twitter_archive_master.csv’

## Analzing and Visualising Data

Insights were documented from the merged dataset and visualisations were made