# Wrangle Report

The report briefly describes the data wrangle efforts exerted in the project.

The dataset that was wrangled, analyzed and visualized is the tweet archive of twitter user @dogrates, also known as we WeRateDogs. This is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. The numerators, though?

Almost always greater than 10. 11/10, 12/10, 13/10, etc. Why? Because "they're good dogs Brent". WeRateDogs has over 4 million followers and has received international media coverage.

The Wrangling process is divided into three steps:

1. Gathering the data
2. Accessing the data
3. Cleaning the data

Each step was explained below:

**1. Gathering Data**

The data used was gathered from three different sources

a. Enhanced Twitter Archive

Contains data extracted programmatically from tweet data sent by WeRateDogs. The data include the

rating, dog name, and dog stage and some other related information.

**b. Image Prediction File**

Produced by running every image in the WeRateDogs Twitter archive through a neural network that classifies breed of dogs. This process resulted in a table full of image predictions (the top 3 only) alongside each tweet_id, image URL and the image number that corresponded to the most confident prediction (numbered 1 to 4 since tweet can have up to 4 images).

**c. Additional Data through Twitter API**

Obtained by querying Twitter's API then stored in a txt called tweet-json.

Gathering this data requires a Twitter developer account

The ready made version was used in this work and was read line by line into a pandas DataFrame with tweet ID, retweet count, and favorite count , and was later saved to a 'tweet_data.csv' file for future purpose.

2. **Accessing Data**

After gathering each of the above pieces of data, they were assessed visually and programmatically for quality and tidiness issues.

**a. Tidiness:**

T1. Dog stage data is separated into 4 columns

T2. All data is related but divided into 3 separates dataframes

**b. Quality**

**i. Enhanced Twitter Archive**

Q1. There are 181 retweets as indicated by tweeted _status_id

Q2. Some dog names are invalid (None, a, an, & the missed of name)

Q3. Invalid tweet_id data type (integer instead of string)

Q4. Invalid timestamp data type (string not datetime)

**ii. Tweet Image Predictions**

Q1. Missing photos for some IDs (2075 rows instead of 2354)

Q2. using underscores are used in multi-word names in columns p1, p2, and p3 instead of space

Q3. Some p names start with an uppercase letter while other starts with lowercase

**iii. Tweet Data From Twitter API**

Q1. Missing entries (Only 2354 entries instead of 2356)