# Wrangle Report

Udacity Data Analyst NanoDegree
Wrangle and Analyze Data Project

The dataset used in this analysis came from the Tweet archive of Twitter's WeRateDogs website. This project only includes original tweets with images, no retweets or replies.

## Gathering:

For this analysis, 3 sources of data were collected. These sources include:

- WeRateDogs original twitter archive. This CSV file was provided to us by Udacity.
- Tweet image predictions data, which was also provided by Udacity.
- Retweet counts and favorite counts obtained from the Twitter API using Tweepy

## Assessing

The data was visually and programmatically assessed for quality and tidiness issues. The issues are as follows:

### Quality Issues

#### Twitter Archive

- Missing values for columns: in_reply_to_status_id,in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp
- Expanded_URL column has duplicate URL's
- Missing values for expanded_URL
- File contains retweets
- Timestamp is an object
- Name column has non ascii characters
- Name column has missing and inaccurate names (a, an, the, etc.)
- Rating numerator goes up to 1776
- Rating denominator goes up to 170
- Tweet_id is an int instead of object

#### Image Predictions

- All three prediction columns have inaccurate data (ie, toaster, vacuum, and toilet tissue)
- All three prediction columns have different formatting
- All three prediction columns have underscores instead of spaces
- Tweet_id is an integer instead of object
- There are duplicate jpg_url images

**Twitter API**

- tweet_id is an integer instead of an object

### Tidiness Issues

**Twitter Archive**

- The last four columns all relate to the same variable (doggo, floofer, pupper, puppo)

**Image Predictions**

- The dataset is separate from the rest of the WeRateDogs data

**Twitter API**

- The dataset is separate from the rest of the WeRateDogs data

## Cleaning

After the initial assessment, I cleaned the data according to the following:

**Twitter Archive**

- Filtered columns with retweets and dropped columns with missing or inaccurate data ( in_reply_to_status_id,in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, expanded_URL)
- Changed timestamp data type from object to timestamp
- Removed the non ascii characters from the name column
- Removed inaccurate names and formatted correct names
- Edited the rating denominator to 10.0
- Standardized the numerator to reflect accurate rating with a denominator of 10.0
- Changed tweet_id from int to object once all datasets were merged
- Combined the four dog types into one column

**Image Predictions**

- Created a new dog_breed column and filtered out the false predictions, which removed the inaccurate data (ie, toaster, vacuum, and toilet tissue)
- Removed underscores and capitalized the first letters of the three prediction columns
- Changed tweet_id from int to object once all datasets were merged
- The duplicate jpg_url images were filtered out when merged

**Twitter API**

- Changed tweet_id from int to object once all datasets were merged

## Analyzing

The analyses that I've conducted include:

- Line chart displaying the number of tweets over time starting in 2015
- Scatter plot of favorite counts vs retweet counts
- A bar chart displaying the most rated dog breeds
- A bar chart displaying the most rated dog names
- A bar chart displaying the most retweeted dog breeds