# Wrangle Report

## Introduction:

This project is about applying concepts of Data Wrangling, on a dataset of archived tweets of the account: @dog_rates, popularly known as WeRateDogs. This twitter account rates dogs as the name suggests with wholesome comments to compliment the ratings, with a denominator of 10 and numerators always greater than 10 (12/10, 15/10 etc.). Because...well we all love dogs!!

## Project Details:

The data wrangling steps taken for this project were:

1. Gathering Data
2. Assessing Data
3. Cleaning Data

## Gathering Data:

Three different datasets were gathered for this project, as follows:

1. Twitter Archive : The WeRateDogs twitter archive which was provided in Udacity Classroom resources and was downloaded manually.
2. The tweet image predictions : That is, what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network. This file (image_predictions.tsv) was 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

3. Twitter API & JSON : The tweet IDs from the twitter archive of WeRateDogs were used to query the Twitter API for JSON data of every tweet using Python library Tweepy, then the complete set of JSON data of each tweet was stored in tweet_json.txt file. After that this .txt file was read into a Pandas DataFrame with tweet ID, favourite count, retweet count, followers count, friends count, source, retweeted status and URL.

## **Assessing Data:**

The data was assessed in the following manner:

1. Visually : by printing the entire DataFrames to look for any discrepancies in the data.
2. Programmatically : by using the built-in Python functions like .info(), .duplicated(), value_counts() etc.

The result of our assessment was divided into Quality and Tidiness, with more detailed assessment under the names of different datasets.

## **Cleaning Data:**

For cleaning the data, the process was divided into three parts, namely : Define, code, and test.

The first step in the cleaning process was creating copies of the three datasets, in order to not mess up the original datasets if the cleaning process resulted in errors or unexpected results.

For Twitter archive dataset -
retweets were deleted as only the original tweets were required, unwanted columns were dropped, multiple columns were melt into a single column, datetime was used to split timestamp into three new columns, numerators and denominator values were corrected.

For image prediction dataset -
Duplicates were deleted, new columns for image prediction and confidence level were created, unwanted columns were removed.

For tweet JSON dataset -
All retweets were deleted to retain only the original tweets.

Then lastly, to resolve tidiness issues the datatype of the tweet_id was changed to merge the tables.

## Conclusion:

By completing this project I feel my data wrangling skills have been refined, I gained more confidence in handling Python libraries like Numpy, Pandas and learned about working with Requests, Tweepy and JSON which were completely new for me before this Nanodegree program.
I learnt the useful skill of gathering data through Web Scraping and API's, assessing data through built-in Python functions and cleaning data manually as well as programmatically via coding.