# Wrangling Report: WeRateDogs Project

## Jade Samala

## December 9, 2018

For this wrangling project, we gathered data from three different sources: downloaded data manually through a given csv file, downloaded data programmatically using Python's request library and the given url and lastly, we imported data through Twitter's API. I had great difficulty gathering data through Twitter's API and failed numerous times. What I found most difficult during this stage was dealing with the API's syntax as well as the fact that it would take 30-40 minutes to import the data and to find out later that not all records were gathered. After a few days of searching the web and numerous attempts of trial and error, I was able to successful import data from Twitter's API.

During our Assessment stage, I discovered that my data had a great number of quality and tidiness issues that I wished to fix but our project made clear that this would be time consuming and requested that we focus on eight quality issues and two tidiness issues. For our quality issues, I had removed retweets from our dataset since we only want to be analyzing original tweets. I noticed that some values under the text variable would state a specific phrase "we only rate dogs" which led me to discover that this would more than likely indicate our tweet was not of a dog. Other quality issues that I had cleaned up were issues such as incorrect mapping of dog names, incorrect data type, and inconsistency under the breed prediction where some breed names would be uppercase and some would be lowercase. To help with our dataset's tidiness, I deleted unnecessary variables such as those that dealt with retweets, and created an additional table for our source variable to give a cleaner look. After our datasets were all cleaned up, I merged our datasets together to create a master dataset that would consist of all of a tweet's information.

I would definitely consider this project to be one of the most difficult project in our course to date and at the same time, I would also say that this project pushed me beyond my comfort zone and caused me to gain a far greater understanding of the various data analysis tools Python offers. After a great amount of time spent on cleaning our data, I definitely have a great appreciation for datasets that have gone through the wrangling process!