# Wrangle Report

For this project I had to gather data from three different sources. WeRateDogs gave Udacity access to their Twitter archive for this project which Udacity provided to its students in the form of a .csv file. This file contained basic tweet data (tweet ID, Timestamp, Text, etc.) for all 5000+ of their tweets from August 1, 2017. The other source of data was a .tsv file that I programmatically had to import in Jupyter Notebook. This data was a database containing each tweet image that was run through a neural network which analyzed the images of dogs and correctly identified their breeds. The third source of data came from the Twitter API. In order to load this data, I had to use Python's Tweepy library.

The data gathering part was the hardest for me. Especially getting familiar with using Tweepy and working with the twitter API. In order to figure it out I looked up a lot on Stack overflow and watched some tutorials on YouTube.

Once I had gathered all the data, I started assessing the data and looked for things I could improve in order to analyze the data and come up with some insights. One important thing I had to take care of was removing the retweets since that isn't original data. I also noticed that some dog names were spelled funny. I took care of this by using the .replace function. After that I got rid of the columns I wouldn't be using for my analysis. I did this by using Python's .drop function. I also separated the timestamp column into three separate columns. Day, Month and year because I might get some useful insights out of that. I also changed the rating_numerator since there were some mistakes there. Some people for example gave their dogs rating in decimals and the rating_numerator wasn't correct in those cases. I corrected this by changing the data type from a integer to a float and corrected all the mistakes with the .replace function.

After using the .info function on all the dataframes I found out that I had to change the datatypes of some of the columns. The last thing I noticed was that the different breeds of dogs were noted funny. Some started with a capital letter and some weren't. In order to make everything the same I used the .str.title() function.

After changing all those things, it was time to work on the tidiness of the datasets. I combined all the different dog stages into one column. I did this by using the .replace function. I also noticed that the tweet ID columns were integers instead of strings so I changed those to strings. After cleaning up the data in the three different datasets I joined them all together and made one huge complete dataset. I used this dataset to look for insights in the data.