## Data Wrangling Project

This data wrangling project is based on “WeRateDogs Twitter page that include dog pictures and videos as well as kindly ratings on the dogs with humorous comments. The ratings are out of ten and with most of ratings actually above ten.

The data provided include an archive of Twitter data for WeRateDogs’ tweets as a CSV file. Based on the tweet content, each dog would be also categorized into “doggo”, “floofer”, “pupper”, and “puppo”. I also need to gather the dog predictions file, which includes the predictions of dog breeds present in each picture and the prediction confidence. The additional tweet information for each twitter in the page, such as retweet number and likes number, are also need to be acquired from Twitter.

As most data wrangling project, I followed the three steps, gathering, assessing, and cleaning. During data gathering, the image prediction file was retrieved from using Python's requests library. The additional tweets “likes” data and “retweet” data information was downloaded using the Twitter API (all APIs would be replaced with xxxxxxx for privacy reasons). During data assessing, I inspected the three data frames for quality and tidiness issues. Some quality issues include missing values, duplicate columns, useless information, and wrong marked values. The two tidiness issues I found is that dog types have four different columns to represent the same variable and “twitter_id” the same variable has been duplicated in all three files. Then, in the cleaning step, I define each quality issue with code and test on these three data frames based on the correct order. For example, I did all the rename and drop entries work first before removing the duplicate or useless columns. Also, the last step is to merge the three cleaned data frames into a master one based on “twitter_id” and thus keep only one column for “twitter_id”.

All the cleaned data frames (the original three data files) and the master one are saved as csv. format also.