# Wrangling Report

## Gathering

The data for the WeRateDogs twitter account had to be gathered from a few different sources. The first, and most convenient source, was the downloaded archive of about 2,500 tweets. This was available as a .csv file and was easily read into a pandas dataframe. However, this dataset lacked a crucial component: favorite and retweet count. Acquiring this required a twitter API call. This also allowed us to determine which tweets had been deleted since the archive was downloaded. Finally, a dataset of images and dog breed predictions (produced using an image identification program) was available online and had to be downloaded via the requests library in python.

The twitter API (tweepy) grabs the tweet data as a json object which can be straightforwardly converted into a python dictionary via the json library. In order to do this for each tweet in the archive, we looped through tweet ids and wrote each json object to a text file, one tweet per line. Then we read that back in and created a list of dictionaries with the relevant info: tweet id, favorite count, retweet count, and whether or not the tweet had been deleted. This was then converted into a pandas dataframe.

## Assessing

The primary issues with the data were quality issues, mainly columns being the incorrect datatype. However, there were two tidiness issues: three dataframes when we only needed two and the dog categories (pupper, doggo, etc.) being distinct columns and not one column. Other quality issues include some of the tweets not having favorite/retweet count (they had been deleted) and invalid entries for numerator and denominator (0 and a value not equal to 10, respectively).

## Cleaning

The first issue to address was data being of the incorrect type. First we changed the tweet ids to strings and the favorite/retweet count to integers. Finally, we set the timestamp data to datetime. Then we were able to address the tidiness concerns. We merged the favorite and retweet data with the archived dataframe on tweet id. Then we used pandas melt method to pivot the columns with dog type into one, and merged this with the data. This left us with one column (instead of four) that listed which dog type that tweet consisted of (pupper, doggo, puppo, floofer, or None). This was then changed to a categorical variable.

The rest of the quality issues were then addressed. Tweets that had been deleted were dropped from the dataset (since the data was unrecoverable) and tweets with invalid entries were also dropped (since the analysis could be performed without them and the reasons they were invalid varied immensely).

Finally we had two, mostly clean datasets. The first contained the tweet info: id, favorite/retweet count, dog type, and timestamp. The second contained the image info: tweet id, and the algorithms best guess as to breed of dog.