During my wrangling effort, I first started off by joining all of the individual datasets together to be able to get a better understanding of all the information I had access to for analyzing the twitter archives. From there, I was able to dig into each of the columns and identify some cleaning that needed to be done, along with the tidiness and quality issues. These issues included: needing to remove columns that didn't have useful information, removing name errors and converting them to a None value, identifying a large outlier in the rating_numerator field, removing rows that didn't have images attached to them, and removing rows that were not original tweets. I also had to change the data type on a few of the columns to better reflect the values that they actually contained.

At first, it was a bit hard to decide what to do with a few of the columns such as the predicted breed and confidence of the prediction columns since there were multiple of them. After further examination and thought about it, I decided that combining all of the information from the different rows into 1 analyzable row would be best. To do this, I just pulled out the rows and columns for when the algorithm was able to correctly predict the breed of the dog.

Another issue I ran into is how to interpret the rating_numerator and rating_denominator rows. They meant nothing to me separate. So in order to actually be able to use the rating score, I took the columns and divided them so the outcome would be 1 analyzable rating for each dog.

Personally, I decided to not do any investigation or wrangling with the floofer, doggo, pupper and puppo columns. I don't believe they added any value to the investigation as by use of the dogtionary, the terms mostly related to the size of dog. I think you can mostly deduct this information from the breed type, and therefore it wasn't included in my investigation.

The wrangling effort for me was mostly a struggle of where to start once I merged all 3 sources of data together. It was easy to get rid of the rows that weren't value-adding, but from there it's always hard for me to choose where to start. I did a lot of looking over of the dataset by hand and then once I started to notice anomalies or missed names and such, I would dig deeper into that column.