In wrangling for this project, I pulled three different file types in various ways. Firstly, I read in a simple .csv file. Then, I used requests to read in a .tsv file from a URL that was provided to me. Lastly, we pulled extra data to analyze from the Twitter API using tweepy. After gathering all of the data I needed. I then had to assess all the data to find any issues.

I first assessed the data visually using my favorite method for assessing data, .sample(). I used this method to sample the dataframe several times, which gave me a random sample of 10 instances across the data that allowed me to pinpoint some issues just by doing that. I also pulled the twitter-enhanced file into excel to also look more issues and to also be able to use the simple find function. After assessing the data, I found 13 quality issues and 2 tidiness issues that I wanted to fix, in order to analyze certain aspects.

A lot of the quality issues were related to the data not being extracted correctly, such as the dog names and some of the ratings. For these issues, I was able to use regular expressions, which I practiced during the lessons. Some issues listed actually took care of themselves after merging the databases and dropping some related columns. The most difficulty I had was dropping the unneeded prediction columns, since I only wanted the most confidently predicted dog breed, as well as, dropping the names in the dataframe that weren't actually names. For both of those, I wrote functions that could be applied to the dataframe in order to append these in a new column or to drop them. As for the other tidiness issue, outside of merging tables, I melted the dog description columns into one column that listed one of the following: doggo, pupper, puppo, floofer, or NaN, if there wasn't a type listed in the text taken from Twitter. After cleaning these issues and some more I was able to perform my analysis.