First, I went through each of the dataframes and examined what needed to be changed. For the twitter archive dataset, there were several useless columns. The first of these was source- right now we aren't interested to see what device was used to post the tweet. The second and third were all the columns relating to replies and retweets. We're only looking at original tweets, so any tweets that were actually replies or retweets should be removed.

In addition, there were many dogs with the name 'None', and the name 'a'. I assumed that the 'a' dogs were named as such because someone was merely filling in the simplest option, and that there are no dogs actually named 'a'. These all needed to be removed.

There were also several type issues, that I noticed both immediately and after beginning cleaning. Immediately, I noticed that the timestamp needed to be in datetime format. After beginning the cleaning process I noticed that all of the tweet_id's needed to be ints in order to properly match with the id's from the other dataframes, and that the numerator and denominator of the ratings would have to be transformed into a string in order to be joined together. Finally, the dog types had to be collected into a single column ('doggo', 'floofer', 'puppo', 'pupper').

In the image prediction dataset, there were again several columns that we couldn't really use- the second and third predictions. The confidence intervals on these predictions were so low that they were basically useless for our purposes. I also combined this dataframe with the twitter archive dataframe.

In the retweets and likes dataframe, these first needed to be transposed. After that, a column header needed to be added for tweet_id, and again, tweet_id (and retweets and likes) transformed into ints. In addition, None values were changed to 0. Finally, this dataframe was also added to the twitter archive dataframe- by combining everything into a single dataframe, it becomes easier to see the data for each tweet.