Wrangling Efforts Report

At the beginning, I started by giving a look at the data in general to evaluate the kind of data I was working on. After gathering all the data from the three different sources (a tsv file, a csv file and a txt file), I started assessing it and looking for some issues visually. The dataset presented itself with lots of missing values but not all of them were essential for the analysis. There were a couple of erroneous datatypes, such as *'tweet_id'* that was formatted as an integer instead of a string and *timestamp* was formatted as object but it should have been considered as datetime so that we can run some analysis with the time variable. Programmatically I found some values that were probably wrong because they implicated a mathematical operation that is impossible. Then I started my cleaning process with the selection of the tweets that were relevant. To exclude the ones that weren't reviews I used the columns with a *retweeted_status_id* or a *in_reply_status_id* because the values in that columns confirmed that the tweets from WRD were replies or retweets of other users. After that I started cleaning the data in the *rating_numerator* and *rating_denominator* columns that were erroneously gathered by extracting the new values from the tweet's text. Then I proceeded to clean the name column because there were some names that were objects that clearly aren't of any help for our analysis. So I started by eliminating all the tweets that had a lowercase letter as the initial because visually you can see that all the names we accurately written always with a capital letter. The next step was the reorganisation of the *dog_stage* columns. At the beginning I replaced all the "None" values with blank spaces to have a clear view of what to keep, then I melted all the values of the four columns into one and joined this new column with the main dataframe. The last process was to select the dog breeds that were gathered with the machine learning algorithm. I chose to keep only the breeds that had a confidence level over the confidence level mean to have a higher probability that the breeds were correctly recognised but also to have a decent number of values to use for my analysis.