Wrangle Report

With the three data frames, there was specific information in each of them I needed for my questions and analysis. From the twitter archive enhanced data frame, I needed the rating numerator and denominator along with the text. From the new twitter data, I needed the retweet count and the favorite count. Finally, from the image prediction data frame, I needed at least one of the three predictions and its confidence level and validation column. Before merging the three data frames, I filtered the image predictions data frame to only include rows where each prediction was True. This was done for data validity. If I had only done this for prediction one, the analysis would have included extraneous things (such as giant panda, and the algorithm still said this was in fact a dog breed). After this, I merged the data frames on "tweet_id". Then I went ahead and dropped the other two prediction sections because I only wanted to focus on one section for analysis. Prediction one had the most data for confidence levels above 50%. After filtering for confidence levels, I cleaned the dog breed names so there were no underscores and no uppercase. Then, I wanted to ensure the rating system was accurate because this was part of my analysis. I applied my own regex to the text and extracted a column of ratings. I separated the ratings column by the "/" sign and had two columns: numerator and denominator. There were NaN values which were interfering with a for loop in future code. After I fixed the NaN values, I was able to loop through both columns and for every numerator above 15, I changed it to 15, and for every denominator above 10, I changed it to 10. This wrangling effort gave me a cleaned master data frame which only had tweets with actual pictures of dogs (no retweets, etc.). I had each tweets favorite and retweet count, and a clean scoring system.