Wrangle Report:

Tidiness:

1. After obtaining and visualizing my data, I noticed a couple of problems that needed fixing. For starters on the data obtained via the Tweeter api, Tweepy, we had a completely redundant column for tweet id but as string, id_str. Since our tweet_id in other dataframes were already of int type, I decided to drop the id_str column.
2. Then, I noticed that there were some columns on multiple dataframes that were duplicated besides just the id. To make life easier when merging our df's, I decided to drop them from all but one df.

For quality issues:

1. When performing a visual inspection on a sample from a dataframe, I noticed a weird rating of 24/7. Took a further look and it was extracted from the text as 24/7 because it was a tweet supporting a gofundme. Unfortunately this did not contain a dog or a rating, so it was removed.
2. A couple of tweets weren't detected to be in the English language, so I cleared those out with a simple query.
3. Then I proceeded to remove all the retweets.
4. Next, observing a describe() of our ratings and a quick boxplot, we can tell that the ratings were all over the place. To make our analysis meaningful I wanted to use only the ratings up to 20 (turns out the max ended up being only 17 this way). For denominators, since I was using more standard numerators that basically had eliminated almost all of the outliers, a quick filtering kept only those at value 10.
5. Simple Timestamp to datetime, incase needed to analyze anything over a given period of time
6. While performing a visual inspection of some samples from df2 (obtained with the requests.get()) I noticed a lot of images were predicted false 3 times for dog. After examining a couple it turns out that yes, people were submitting more than just dogs to be rated. A quick filter removed all the images that were not guessed as dogs by the neural network.
7. Some 0 Ratings were still leftover, so I manually checked them, turns out to be a repost and not a dog.
8. To make sure our final dataframe had the complete text, and since through the api we retrieved the tweets as extended, I decided to drop the original text and keep the full_text column from the api data.