Wrangling efforts

This report was not too painful to collect and clean as a whole, but there were a few pain points. The given csv and querying the images archive was fairly straightforward and I had done it several times in the past. However, getting the tweet information from Twitter was quite challenging. After getting the developer account it took me a long time to get the json file downloaded correctly and then it was very painful to parse, I think in part to how I downloaded it. I did find a way to parse the retweet and favorite counts after much difficulty but I do not know if the method I used is sustainable for future use on different json files.

Once all the data was populated it was time to assess and clean the data. For the most part this was fairly straightforward. The most difficult part was programmatically extracting the correct rating values as there was no set format in the text cell. The program that parses out the text to get the ratings and name is clearly very flawed. Many of the cells that contained the name ‘a’ as mentioned in the report had a name associated with the picture but because of the rules it used in parsing it seemed to be just grabbing the first word after is, as in “This is NAME”. Regardless, the name column was not used for analysis, so they were able to be updated to NONE for the name safely.

The joining of the two data sets and reducing all three dfs to contain only the same set of tweet_ids was difficult as well. Originally when I extracted the Tweepy data I used the tweet ids from the images data set and I was only successfully returning 2034 records(less than the ratings df) and this was very curious. When I tried to join the two sets they did not contain the same records and I could not understand why for some time. However, when I realized I should be using the ratings df tweet ids it was much easier.

The other changes I made were purely for my own use in visualization and assessment. I made two categorical variables, one that aggregated prediction of the image being of a dog or not and another that consolidated the dog stage into a single column. This was done by simply grouping and replacing values then cleaning up the old unnecessary rows.

Overall, this was a very useful and interesting assignment. I learned a lot about the problems that can happen due to improper data types or outlier values.