Wrangle Report

In this project, a real-world data on dogs from the popular WeRateDogs twitter account was used.

Three datasets were gathered from different sources, cleaned and combined. Two of the datasets were provided while one had to be accessed through the twitter API which formed the tweet.json file. From this file, I pulled the retweet_count and favourite_count for each tweet. While assessing the data generated from the twitter API, about 179 tweets had 0 favourite counts which is quite unlikely because most of these tweets had a number of retweets, coupled with the fact that WeRateDogs is a popular twitter account. Since these data points were less than 10% of the 2354 data points gathered, I simply dropped them and focused on the data points which appeared to be more accurate (those with a favourite count greater than zero).

After I imported the twitter_archive table (main dataset) into pandas, I noticed it contained some retweets (about 181 rows of data), which I removed because the specification of project required that only original tweets were required for analysis.

The dog stages in the twitter_archive_enhanced dataset were represented as a column each which I had to convert to rows and clean appropriately. Also, the name of some dogs was poorly recorded and some were designated as *None* and I cleaned them all designating such dogs as having no name.

The neural network produced image-predictions.tsv also had some accuracy problems. For example, it did not detect a dog image for tweet_id 718454725339934721 whereas there was a dog in the image. Nonetheless, since it was able to fairly detect dogs and their species, I simply focused my analysis on the data points where the neural network predicted a dog and its specie.

Also, some tweet_ids that were present in the twitter-archive-enhanced table were missing in the image predictions file produced by the neural network and since the analysis required that only tweets that had images be used; I simply made use of the tweet_ids that were also found in the .tsv file.