Wrangle Report

This project involved the gathering of data from different sources and analysing to generate insights. In this project real-world data from a popular twitter account was gathered, cleaned and analysed.

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 importing the twitter archive file, I noticed there were some retweets (about 181 retweets) which I dropped because the project required the absence of retweets in the analysis.

The dog stages were presented in the wide format with each dog stage representing a column. I cleaned this by melting the entire dataframe and converting the stages into the long format while also removing duplicates too.

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 were present in the twitter archive file that had no corresponding image in the images dataframe. I filter the dataframe to include only tweets that had their corresponding dog image and I also added a corresponding breed for each tweet_id which was generated from the image predicted by the neural network.