Wrangle Report

The first part of the Wrangle Act project was to do just that - wrangle some data. The objective was to load in and later combine three separate data files, each related to a tweet from the WeRateDogs twitter accounts.

The frist data file, twitter-archive-enhanced, was provided as a csv and was the easiest to work with. I simply imported it into the Jupyter Notebook as a pandas dataframe.
The second file, image-predictions, contained neural network predictions of what the breed of the dog was based on the image included in the tweet. This file was also relatively easy to load in using the requests package. A more difficult process was using Python's Tweepy library to access and store each tweet's entire set of JSON data. For one, this was the first time I've used an API to gather data. I am also not an active Twitter user, so setting up an app to generate access and consumer token/ secret was something entirely new to me.

In my data cleaning efforts, I focused on eliminating fields and rows that were non pertinent to my findings. For example, I was not interested in retweeted posts since I wanted to get the favorite count and retweet counts of the originals. I also dropped fields which contained extensive information about each tweet, such as “place” as they didn't include a lot of usable information and rather cluttered the data frame.

When joining the different data frames created from the aforementioned files into a single one, I opted for using and inner join. An inner join finds and returns matching data from tables, or in this case, pandas data frames. This ensures that the result of the join contains no empty cells. While an outer join finds and returns matching data and some dissimilar data from tables.
Something I would be interested in doing in further analysis, is also converting the timestamp to a datetime format.