Wrangle report

I started cleaning the data with 3 separate files a csv file with twitter data from WeRateDogs, a tsv file with neural net predictions of dogs, and scraped data from twitter. To keep from scraping twitter several times I saved all the data in a text file and converted in to a python data frame.

Initially much of the data was dirty. The twitter data from WeRateDogs was particularly troubling. I first noticed everything with a retweet status was a retweet, and therefore a duplicate. I removed all those tweets. Then I noticed several columns were unnecessary - the retweet and in_reply_to columns, so I removed them. I also noticed the text had picture and ratings. I removed this information, so I just had the text. I created a new column for hashstags that grabbed # from the text. I then used beautiful soup to remove the html from source.

For the neural net data, I renamed multiple columns, so they were more reader friendly. For example, I changed p1_dog to top_prediction_is_dog.

Once, I cleaned those two data sets, I created two separate tables of organized data. The first table contains all information from tweets. This information includes retweets, favorites, timestamp, source, text, expanded_url, hashtags, and jpg_url. The second table contains all information about dogs. This table includes names, rating numerator, rating denominator, top predictions, and doggo, floofer, pupper, and puppo.

I was not able to clean every error I discovered. In the future, I would like to fix names. Many names are a or something similar. It appears the csv file identifies name as the first word after is in the text. I would also like to convert all the doggo similar columns from doggo or None to true or false. I would do this for doggo, fluffer, pupper, and puppo.