WeRateDogs Wrangling report

This wrangling effort aims to gather information about the WeRateDogs Twitter account, analyze the data, and see what insights can be gleaned from it.

The data comes from three separate files. The first is a file with comma-separated values provided by Udacity called twitter_archive_enhanced. This file was manually downloaded from the Udacity site and contained information extracted from the tweets. The second file is a tab-separated file named image_predictions. It has been downloaded programmatically using the Requests library and includes predictions from a neural network of each dog's breed. Lastly, additional data was gathered by querying Twitter's API to download the JSON data using Python's Tweepy library and storing the entire JSON data in a text file called tweet_json.

The wrangling began by inspecting each created data frame. There were some inconsistencies in the data types. For instance, the timestamp in weratedogs is more useful when changed to a DateTime format, and twitter_id did not need to be an integer because it identifies the tweet and is not valuable mathematically. A few rows in weratedogs contained incorrect information, such as many dogs named 'a'. After reading through some tweets, it became apparent that the name was in the text and required extraction. Regex aided in the extraction of dog names from the full text of the tweet. Those names were added to a column called dog_name and checked against the existing name and the full text. Many unnecessary columns were dropped, such as language and others deemed superfluous. Breeds in the image_predictions data frame were inconsistent with their capitalization and were standardized in all lowercase.

These data frames also required a bit of tidying up. For example, the weratedogs data frame had four separate columns for the different stages of dogs. Instead of having redundant information, one column, created from the combination of the four dog stages and a newly created unknown column, takes the place of the four columns. Finally, the merged data is saved into one master data frame, and all remaining extraneous columns are dropped. With the wrangling process now complete, the data can now be analyzed.