Data Wrangling

The purpose of the Wrangle and Analyze Data project was to gather, assess, and clean data from the WeRateDogs Twitter archive. Gathering data consisted of using Pandas to read a provided archive of tweets, using Tweepy to download detailed tweet information using the tweet IDs provided in the archive, and downloading an image prediction data set using Requests. While assessing, I identified various issues with data integrity and thought about how I would structure data for analysis. Finally, in cleaning data, I performed the tasks that I outlined during my assessment. I bounced back and forth between these stages frequently as I would discover additional items to clean or assess as I dug further in to the data. Additionally, I looked at a lot of dog photos.

Gather

Opening the Twitter archive provided was completely straightforward considering I had done similar work in prior assignments. Downloading the image predictions was similarly simple because the library is uncomplicated. Tweepy was different. Where I found documentation for Tweepy's features, it was both inaccurate and incomplete. For instance, though the existence of the *parser* argument was documented for the API function, valid parsers did not seem to be in the documentation. This could also be due to Twitter's moving target API. Do to the loosened restriction on tweet length, tweets longer than 140 characters are truncated, and the full text is available in a separate entity, requiring the use of tweet_mode='extended'. While using this option is easy in the single-tweet-downloading get_status function, it is not available in statuses_lookup. This meant that getting full text, or other extended entities, required hundreds of individual API calls instead of 30, and downloading the tweet details therefore took 30 minutes instead of seconds. The worst part is that I didn't end up using extended entities.

Assess

Assessing the data was relatively uncomplicated. I chose to reduce the data set significantly, by removing unnecessary columns, to simplify my final analysis. There were several items that I reviewed for accuracy. I looked through ratings to determine if they had been captured accurately from tweets and noted where they were not. I reviewed dog names to see what non-names had been included and noted them for removal. I also reviewed the dog breeds, how image predictions were stored, and where the most confident predictions were.

Clean

While cleaning the data, I started by working on areas I had noted during my assessment. This included removing columns I considered extraneous, removing a select few tweets that did not benefit the analysis, and removing tweets that were retweets or that had no photos. Next, I combined the multiple columns containing Doggo Lingo into a single column of values. I corrected dog names and dog ratings, and I normalized the ratings for ratings that contained multiple dogs. I reduced the prediction data set to the most confident prediction that contained a dog, cleaned the dog breed names, and joined the tables together in a reduced set, which I saved as a CSV file.