Wrangling Report

The wrangling process was a difficult yet fulfilling and eye opening experience. The first difficult part was getting familiar with using API's. This was my first experience with obtaining authorization to use a company's API and obtaining the necessary keys, tokens, and secrets. Once I obtained authorization, actual use of the API was quite simple. One bump in the road with working with the tweepy API was not at first knowing why I was unable to perform 'get.status' on every tweet ID. I then realized that there were some invalid tweet ID's that did not have any corresponding statuses. In order to programmatically obtain a decent amount of data from a list a tweet ID's compiled from the given CSV, invalid tweet IDs were error handled and skipped over.

Assessment of the data was done visually and programmatically. The visual assessment consisted of looking at sample groups of five data frame observations. The programatic assessment consisted of using the functions .info(), .isnull(), .duplicated(), .shape(), and .value_counts(). The first focus of the assessment process was to look for obvious quality issues such as; missing data, inconsistent value formats (i.e. names that are sometimes lowercase and sometimes capitalized), data that is in the wrong format (i.e. 'None' instead of NaN), hard to read data, etc. The next focus was to focus on the tidiness issues of the data. There were two obvious tidiness issues that were observed and cleaned in the wrangling process. First, the separate 'doggo', 'floofer', 'pupper', and 'puppo' were separate columns that should instead be values of a single column. Second, the data frame with retweet and favorite counts would be more use-full if it was combined with the twitter archive data frame.

Throughout the cleaning I had to make a new function called 'source_val_converter()', use string slicing, directly reassign column values, drop columns, merge data frames, melt columns, etc. Both programatic and visual tests were used to determine the success of each cleaning process.