Wrangle_report by Matt Brendel
Jan 11, 2019

For this project, we gathered, assessed, and cleaned 3 datasets relating to the WeRateDogs Twitter site. The first dataset was the enhanced archive, which had information on the tweet, including the timestamp, url, tweet text, parsed-out dog names and scores, etc.

This dataset involved the most cleaning since some of the columns had been created by attempting to parse information from the tweet text. That programmatic parsing wasn't completely correct, so we had to conduct some manual cleaning. For instance, many of the dog names were tagged as 'the', which is obviously not correct. We updated those entries to be NaN. There were similar issues I caught using visual assessment with the rating columns. Most of the denominators were 10, so I investigated a few that were not. One, for instance, was 7, and looking at the tweet text, the parsing had mistakenly drawn out the user talking about 'something happening 24/7' instead of the rating intended for the pup.

The archive dataset also had a few columns at the end for different dog types (ex: doggo, floofer, etc). These columns either reflected the column name of dog type, if the text listed that type of dog, or 'None'. This structure would have made it difficult to aggregate or groupby particular dog types, in a Series-sense, so we condensed the various dog type columns into one. We initially explored using the melt function that was discussed in the class, but as some tweets had multiple dog types identified across the different dog type columns, the output from this was pretty confusing to keep track of in terms of having to drop various duplicates. I solutioned a more straightforward approach using the apply method and lambda function to concatenate the original dog type columns into one, then dropping the originals. If multiple dog types were listed, I recategorized those as dog type= 'multiple'.

There wasn't as much to clean with the other two datasets. The second included data relating to the tweet image and predictions as to what type of dog breed was pictured. The third we created from querying the Twitter API to gather tweet retweet and favorite counts. Because all three Data Frames contained the one base observational unit, tweet_id, we were able to merge on that column, keeping only those rows which had an image-url (as per project instructions).