Review of Data Wrangling

Data cleaning steps 7 and 8 are imperfect in that they do not consider some special cases. For example, one particular tweet shows a rating of 20/16 in the 'text' column. Perhaps ratings like this could be converted to have a denominator of 10 but that would require code that considers a wide variety of cases. Another alternative might be to simply eliminate certain rows where *both* the numerator and the denominator are outside of a normal range.

Cleaning item number 9 came to mind only after I realized that the 3 DataFrames could not be joined or merged unless there was consistency in the 'tweet_id' column. It is important to note that the extract method used is good but it did not capture all tweet_ids since there are rows in which the 'expanded_urls' column did not have the tweet_id in it. This means that certain data was excluded from our master_dataset but it was a relatively limited number of rows that were lost.

For Cleaning item number 10 I ran into certain difficulties in extracting the dog's name from the text column in cases where it was not present in the 'name' column. Also, I should mention that there three or four cases where the dog's name appears as a noun such as 'not' or 'officially'. These were isolated cases so they did not have an impact on Analysis item number 2, which dealt with dog names.

For the data tidiness section, I had intended to use the "melt" method but the presence of multiple labels for the same dog in about 15 cases made it difficult to do so. The methods I used required more code but provided a clearer solution and kept the data complete.

The last item to note is that there were a number of cases where I found it difficult to use vectorization so I opted for a for-loop instead. I understand that this dataset was relatively small so this was not a problem. With a larger dataset then vectorization would be preferable.