WeRateDogs Data Wrangling Report

The Purpose of this Analysis was to take dog breeds from posts on the WeRateDogs twitter handle and compare them with their ratings, number of retweets and number of favorites. To achieve that goal, quality issues and tidiness issues needed to be cleaned.

Tidyness Issues

- We needed to first merge the extra tweet information (favorites and retweets) to the main dataframe.
- Then we needed to add the dog breed predictions as well.

Quality Issues

- Tweet_id on all the files were made into a string object instead of an integer.
- The timestamp column in the master dataframe was changed to datetime format and the rating numerator data type was changed to a float to incorporate decimal ratings.
- In the image prediction dataframe only the 1st breed predictions were chosen and the rest were dropped.
- Rows that were retweets or replies were dropped because only original tweets were used for the analysis.
- The errors in dog names were fixed by changing any names that did not start with a capital letter to “None”.
- Ratings were put in a different format to be used in the analysis, since they almost all have a denominator of 10, the denominator column was dropped.
- Fixed any rating errors that were above 14 by manually changing them in the rows.
- Dog "Stage" was dropped because there were not enough rows to do a meaningful analysis and it would be very difficult to extract that data from the text because there are many words to describe dog "stages".

Limitations

After dropping rows, and cleaning the data, we were only left with 1,947 rows to work with. This analysis could potentially be improved by getting a bigger sample and getting all of the tweets directly using Tweepy. Further, someone with better text extracting capabilities may be able to write scripts to get more rating data, dog name data, and even possibly more dog "stage" data that could be used in an analysis. Lastly, a neural network with more accuracy predicting dog breeds could also possibly give us better results.