Wrangle Report

This report will briefly describe how the dataset twitter_archive_master.csv was generated.

The dataset consists of three original raw data sources. An archive dataset of WeRateDogs tweets, data from the twitter API that contains retweets and likes for each tweet, and an image prediction dataset, given by the data analytics site Udacity.com, which provided breed predictions on each dog generated by a neural network.

The three datasets are merged using left joins to capture key metrics from the archived dataset while allowing for NULL values on the image predictions and on the retweets/ likes. This join method was chosen because all the textual data from all the columns for richer text analytics is needed. All of the retweet and like data for every ID wasn't available. About 100 out of 2400 were missing. This was possibly due to deleted tweets that the official API no longer held.

Lastly, some of the image predictions are not dog breeds and were left out of the analysis. This solved all data structure issues (tidiness) and the data quality issues having to do with completeness and validity. Next, data quality issues of accuracy and consistency are addressed including accuracy issues having to do with inaccurate names and ratings being extracted imperfectly by the Udacity team. Visual assessment and than programmatic cleaning is used.

Lastly consistency issues are addressed for the textual data. The data needs to be presented in a word cloud and analyzed by sentiment for polarity and subjectivity. To accomplish this, the text data is constrained to follow a format of all lowercase, no punctuation except for rating slashes to separate numerator and denominator, and the removal of all links. This allows for a richer word cloud and better sentiment analysis.

The clean master dataset is available in two files: twitter_archive_master.csv and a sqllite file twitter_archive_master.db. This allows for further analysis and integrity to be maintained in a SQL database while also allowing for further analytics to be performed on the CSV file through pandas.