Quality Issues to be Wrangled

The ‘tweet-json.txt’ requires two issues to be wrangled. One is that the ‘id’ column needs to be renamed ‘tweet-id’ in order to match the name of the primary keys of the other two tables. All ‘tweet-id’ columns, including the tweet-json.txt ‘tweet-id,’ need matching datatypes. The datatype of ‘tweet-id’ has been changed to strings in all three tables.

The main issue to be wrangled in the ‘image predictions’ table is to remove the predictions from p1 and p2 that are not dogs. This means removing the rows where p1_dog or p2_dog values are false.

The table with the most cleaning needed to fit the needs of the data analysis is the ‘twitter-archive-enhanced’ table. Because we are getting retweet count and favorite count information from tweet-json.txt, we need to remove anything that is not an original tweet from ‘twitter-archive-enhanced.’ This means that we need to remove rows that are not blank or not null in in_reply_to_status_id and in_reply_to_user_id, and we also need to remove rows that are not blank or not null in retweeted_status_id, retweeted_status_user_id, and retweeted_status_timestamp. After the rows are removed, these five columns can be removed from the table.

We only have full data up to August 1, 2017, so that means that any dates in the timestamp column in the ‘twitter-archive-enhanced’ table that are after August 1, 2017 can be removed.

Also, some of the ratings have not been copied correctly to the rating_numerator and rating_denominator columns correctly. This will be handled by removing rows that have ratings that are outliers, which in this case will be numerator/denominator ratios greater than 1.4.

Tidiness Issues to be Wrangled

In the ‘twitter-archive-enhanced,’ tidiness rules are violated with the four dog stages, 'doggo,' 'floofer,' 'puppo,' and 'pupper,' in four columns. To meet the tidiness rules, the four columns can be combined into one column called dog_stage.

The other tidiness issue to be wrangled is a merge of the three tables, ‘twitter-archive-enhanced,’ ‘image predictions’, and ‘tweet-json.txt’, into one large table.