WeRateDogs - WrangleReport

Kathy Mirzaei

In this report I outline the wrangling efforts to assemble and clean the data required for analysis of the WeRateDogs Twitter Archive.

Data Gathering Part:

I gathered data from 3 different sources, each stored in a separate file:

1. WeRateDogs Twitter Enhanced archive, this was manually downloaded from the Udacity resource page.
2. The image predictions file downloaded from Udacity.
3. The entire set of each tweets' JSON data, downloaded by querying the Twitter API using the Tweepy library. The `favourite_count` and `retweet_count` were extracted programmatically from this file.

In the next steps, I loaded the 3 raw data files into separate data frames: archive, predictions and json_data.

Data Assessment and Data Cleaning Parts:

I began the assessment by viewing the information on the archive table first, identifying several quality and tidiness issues.

All rows and columns containing non-null values in the retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, in_reply_to_status_id, in_reply_to_status_id, and in_reply+to_user_id were dropped as per the requirements.

The `timestamp` column was converted to datetime data type from str value.

The 4 dog stage columns were combined into one column named `` `stage` ``; tweets without stages were set to 'none'. Several had 2 stages set, I kept only the one with the lower overall count.

The html strings in the `source` column were replaced with the display portion of itself. Instead of a long html, it then only shows the short form of the actual source name.

The `rating_numerator` and `rating_denominator` columns were checked for value ranges; I decided to keep only tweets with single ratings. Tweets with large numerators were dropped, as the text did not contain a valid rating (# out of 10). After the ratings were fixed, I dropped the `rating_denominator` column (it contained only '10's) and renamed the `rating_numerator` column to `rating`.

The odd words in the `name` column were replaced with 'none'. (e.g. "a", "the")

Tweets with missing values in `expanded_urls`, (not retweets or replies) were actually missing the urls from the text itself. These tweets/columns were dropped

The predictions table itself was not clean. There were many tweets with no dog breed predicted, these were left as is. The best prediction for `breed` and associated `confidence` level were extracted and merged into the archive table.

The json_data table itself was a stand-alone table that I merged with the archive data frame using the tweet Id.

Therefore, the `retweet_count` and `favorite_count` columns were merged into the archive table, and the data type reset to int.

The remaining cleaned columns in the archive table were reordered for the ease of use, then the table was saved to the new “twitter_archive_master.csv” file.