WeRateDogs Twitter Archive - Wrangle Report

Project by Chizaram Emenyonu

In this report I'll explain the wrangling activities I carried out to assemble and clean the data required for analysis of the WeRateDogs Twitter Archive.

Data Gathering

I gathered data from 3 different sources:

- WeRateDogs Twitter Enhanced archive, manually downloaded from the Udacity servers.
- The image predictions file, programmatically downloaded from the Udacity servers.
- The JSON data for each tweet, downloaded by querying the Twitter API using the Tweepy library.

I loaded the 3 raw data files into separate dataframes: twitter_archive, image_predictions and extra_data.

Assessment & Cleaning

The assessment started by first looking at the twitter_archive dataset to understand the information presented in it, then I identified several quality and tidiness issues.

I checked to see if there were any tweet duplicates, I didn't see any. All rows containing non-null values in the retweeted_status_id , retweeted_status_user_id and retweeted_status_timestamp , and also in the in_reply_to_status_id and in_reply_to_user_id columns were dropped, according to the requirements. These columns were then also dropped as they were not needed.

The html strings in the source column were replaced with the display portion of itself. The rating_numerator and rating_denominator columns were checked for value ranges; Tweets with large numerators were dropped, as the text didn’t 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.

I noticed some descrepancies in the dog names column. There were some words that obviously weren't names but captured as names. I also noticed that these words started with lower-case letters. I replaced all names starting with lower case in the name column ‘none’. After that, I dropped all roles with 'none' values in its name column.

The timestamp column was converted to datetime data type. The 4 dog stage columns were merged into the stage column; tweets without stages were set to ‘none’. Several had 2 stages set, so I kept only the one with the lower overall count.

Tweets with missing values in expanded_urls , (not retweets or replies) were actually missing the urls from the text itself. These tweets were dropped, and then the column itself. The image_predictions table itself was not cleaned. The extra_data table was also not cleaned. The retweet_count and favorite_count columns were merged into the archive table The remaining cleaned columns in the archive table were reordered, then the table was saved to the new “twitter_archive_master.csv” file.