WeRateDogs Twitter Data Wrangling Documentation

By: Jason Kim

As part of my data wrangling efforts, three files that we need for this project (Twitter-archive-enhanced, image-predictions, and tweet_json) were gathered through programmatic download. Specifically, one required the use of an API directly from Twitter to download the necessary data regarding the number of retweets and favorites by tweet ID onto a txt file. A loop on the tweet ID of the twitter archive file was required to query the necessary data. All three files are of a different file type (csv, tsv, json).

After gathering these files, they were programmatically and visually opened to inspect the contents. As part of the assessing phase, the structure was set up so that all the quality and tidiness issues would be listed first. The quality and tidiness are subdivided into the different tables where the issues arose. Below this, a brief exploration of the data was done broken up by table to find any issues. Any issues found during the exploration would be documented in the beginning of the assessment phase.

After assessment, cleaning was done to each problem listed out in the assessment phase. The cleaning was done by issue type (quality/tidiness) and then by table. Each issue was resolved through a standard format of first explaining the issue (explained in the assessment phase), followed by the plan of action (define), the code to fix it, and the testing to ensure the problem has been resolved.

Once all issues have been resolved, the three tables were merged by tweet ID to create a singular master table. A left merge was used to ensure the base table (i.e. twitter-archive -enhanced table) keys are intact. The resulting master table was stored as a csv file. Note that there were several quality issues that were present after the merge. These issues included altered data types, missing values, columns unhelpful in analysis.

Lastly, please note that due to the extent and volume of wrangling needed for these files, only a subset of all issues has been cleaned.