Wrangling Report

The WeRateDogs™ (@dog_rates) twitter account receives and collects photos of photogenic pooches, and rates them. A tweet archive was received from the account owner, consisting of 2356 tweets, retweets and replies. This archive had some preliminary data harvesting performed on it to add names, ratings and doggo categories.

As we're interested in analyzing this dataset, additional information, including favorites and likes, was desirable and collected.

For any meaningful numerical analysis, solid numbers are required. Ensuring a consistent rating system was of primary importance. Various denominators were noticed during assessment. These non-standard denominators existed due to the following causes:

- Multiple dogs are handled by WeRateDogs with multiplicative ratings - i.e. 8 dogs with an 11/10 rating would be rated as 88/80.
- String extraction for some tweets collected the wrong value, where another fraction resembling a rating was in the text of the tweet
- String extraction failed on values like 11.76, grabbing values after the period.
- Dates were grabbed where no actual rating existed

Repairing these deficiencies was relatively easy, but highlight the difficulty in parsing language programmatically for the purposes of harvesting data. It's difficult to apply a single solution to a large dataset, and requires care in checking the outputs of any programmatic gathering.

False positives in terms of names were also common in the dataset. Dog names were harvested from tweets programmatically, and frequently grabbed random words where a name did not exist. This was only noticed on more detailed review of dog names, and could easily have been missed until analysis.

As a final insight into the wrangling process, the requirement for an iterative approach to wrangling needs to be taken in to account. During wrangling of the supplied categories of dog-type, it appeared that the only issue requiring cleaning was merging multiple columns into one column. Only after this task was complete did it become apparent that some tweets had multiple categories of dogs assigned to it, requiring additional cleaning for tidy data.

Iteration appears to be key to highlight issues that are hidden

Data-type wrangling was kept until the last part of the wrangling process. While the wrangling code was being written, data-type changes would occasionally revert or be changed to different types, depending on the operations and methods selected for use. While these issues are surmountable, the argument can be made for leaving these items to the end - unless certain datatypes are required for whatever cleaning process is required.