# wrangle_report

March 31, 2018

### 0.0.1 Overview of Wrangling Done

Each source of data had a very different combination of required wrangling. The first data source, twitter_archive_enhanced.csv, was "on-hand" and available for direct download. The second piece, the image_predictions.tsv file, was hosted on a Udacity server, so the Request library was used to download it. The third piece had to be created. Tweet data was needed that could be found on Twitter itself, so this data was obtained by using the Tweepy library to work with the Twitter API. This information was returned as JSON data, and stored in a text file.

After gathering the data, each file was read into its own data frame and then assessed. The twitter_archive information had a number of data type issues for the columns, which were fairly straightforward, but the biggest hurdle was the rating system. The numerators spanned from 0 to 1776, and the denominator had values from 0 to 170. Upon closer investigation, several issues emerged. The ratings were obtained from the tweet text, but if another piece of the text contained a "#/#" format, such as "24/7" or "9/11", it could get erroneously pulled into the rating. To fix that, I re-pulled the rating from the text, using a more specific regular expression. After that, a number of formerly problematic ratings were resolved. However, there were still many ratings with denominators greater than 10. Specifically, they were multiples of 10. The text for these tweets indicated that the picture was of more than one dog, which lead me to believe that the rating was a rating over 10, multiplied by the number of dogs. So, I created a column to hold the number of dogs, and normalized each rating to have a denominator of 10.

The image_predictions data was quite clean already, but I did update the column names to be more descriptive, and fixed a couple of column data type issues, such as setting the tweet ID to be a string, rather than an integer.

The data obtained from the Twitter API in JSON format took a bit of work. Reading the data into Python from the csv I'd created gave a dictionary object, which I then transformed into a data frame. However, the data frame had a column for the tweet's ID, but then both the favorite count and retweet count were inside the same column, and each tuple was inside of a dictionary inside of a list. After figuring out how to access the desired elements inside that column, I ran a loop that pulled them out into their own 'favorite_count' and 'retweet_count' columns, respectively.

After that, I decided that a single data frame would be the tidiest. I used the twitter_archive table as the base, and then joined the image_predictions and tweet_data

tables via LEFT JOINs ON tweet_id. The resulting table was saved into a csv called 'twitter_archive_master.csv'.