Wrangle Report
Jean Simmons-Delpit

This project represents a student example of completing an analysis with emphasis on the data wrangling step. The team at Udacity has provided a data set across a few different data sources from the Twitter user @dog_rates. Wrangling was broken up into three (3) main steps; data gathering, data assessment, and data cleaning.

During the gathering process data was gathered from a previously prepared .csv file downloaded directly from the Udacity Student Classroom and read into Jupyter Notebooks using the Pandas library. A .tsv file was also read in but was sourced from a website using the Requests and IO libraries. Lastly a .json file was read in using the Pandas and JSON libraries. An attempt was made to get the json data from Twitter using the Tweepy library, but Twitter's API doesn't allow users with a standard developer account to grab archived data, so I had to rely on the .json file provided by Udacity.

During the data assessment process data was reviewed for quality issues and tidiness issues. Tidiness issues revolved around searching for structural issues and 2 issues were identified. The first issue was that there was actually no need for all three data sets to be in separate data frames since together they all represent one observational unit. The other tidiness issue revolved around a categorical feature being spread across multiple columns when in fact this feature needed only be represented by one (1) column as tidiness calls for each variable to form its own column. Quality issues were also explored and identified with eight (8) issues being identified as required of the project. In particular assessment focused on searching out missing data points, duplicate data, and incorrect data as well is incorrect data types.

Finally, during the cleaning process, the issues identified and documented in the assessment process were addressed starting with the easiest perceived task to the hardest. Cleanup efforts included removing records that were outside the scope of work (in particular retweets and replies), updating features with incorrect data types, removing data with little significance, converting ambiguous data into human readable data, and correcting incorrect values. Tidiness issues were also corrected; the split data frame was combined into one (1) data frame and the feature split across multiple columns was combined into one (1) column.