Wrangling Efforts for the WeAreDogs Channel's Tweet and Image Analysis

My data wrangling efforts follow the gathering, assessing and cleaning scheme. In the gathering, I gathered data from three different sources; tweet data as JSON objects in a text file, a tab separated file and a comma separated file. I should note that Twitter didn't approve my application for the developer account, so I downloaded the "tweet-json.txt" file from the Udacity page. In assessing and cleaning, I had three steps in my strategy, dealing with missing values, fixing tidiness issues of the data and dealing with other data quality issues such as fixing the values of the variables etc. I came up with this order due to the relative importance of each issue in the data wrangling process.

In the first step, the data had a few issues. Some NaN values were not represented correctly, however I consider this as part of the third step. Because, they involved changing the representation of a specific cell value. In the Twitter archive dataset, some rows didn't have the url information about the original tweet. I considered this as an indication of deletion and removed these rows from the dataset. Most of the tweets also didn't have information about the retweets or in reply to values. I considered only tweets who didn't have values for retweets and replies as I want to analyze the original tweets to answer my questions. Retweets and replies are correlated with the original tweets which violates the i.i.d assumptions that we usually have about records in our datasets. Last, I also dropped the columns relevant to replies and retweets as they are all NaN values after the former consideration. After these cleaning steps, I still had 2094 rows in my datasets. This meant keeping approximately 90% of all the records.

In the second step, I dealt with tidiness issues. In the Twitter archive dataset, there were three columns in the dataset that represent actually one variable, age. Furthermore, there were too many tables for one observational unit, eventually one supposed table for the actual tweet text and metadata; which includes the derived features about dog properties such as their name, foofer, age etc. I created a new age column and merged all three tables accordingly. This helped my following cleaning step.

In the third step, I dealt with the quality issues of the data. This included changing the NaN representation for the missing values, extracting the HTML element for the source column, changing the classification column names, changing the class labels, removing the statistics of favorite and retweet of the replying and retweeted tweets, removing the incorrect dog names; and, changing the date column's data type. These were general data quality issues; and, fixing them made my analysis more accurate and easier to interpret.