Data Wrangling Report

Todd Allen

Overall this project was both challenging and rewarding. I was a little concerned at the beginning on how much time I would have to spend to complete this project. However, with the great content in the videos and the additional links provided helped a tremendous amount.

The gathering of the data was very challenging as I have not worked with APIs before. The most challenging was the reading and writing to json files then converting it to a data frame. Once I figured that out, I thought I would try to just download only the columns information I thought I would use. After downloading just few columns, I thought it would be best to get all the columns. This allowed me to look over all the data and see if there was anything else interesting to analyze.

The assessing was the most difficult for me as I am not to picking about minor details like column headings and underscores or words not being capitalized. However, I do see the benefit of consistent data and more readable and understandable headings. This was a good section for me as it really got me looking at the data and seeing how I could make it better for analysis. Some of the issues that took some time to find and resolve, included:

1. Fixed fractions in rating_numerators
2. Removal of tweet_id 810984652412424192 as there was no longer information associated with this tweet_id.
3. Manually update ratings for tweet_id 835246439529840640 and tweet_id 666287406224695296. These both had a second set of numbers that were mistaken as ratings.
4. Removed 181 retweets
5. Fix inaccurate dog names
6. Split timestamp column into date, time, month, hour, and day for more analysis by time.
7. Timestamp had wrong data type.
8. Change tweet_id to string (object) datatype

Again, this section was the most difficult but also very rewarding as it makes the cleaning section a little easier. I did have to come back and add to this list, once I started to produce reports it was clear that more cleaning was needed. Also, thanks to the reviewer for pointing me in the right direction for the fraction issues as I completely missed it the first time around.

Cleaning section I enjoyed the most. I enjoy the programming aspects as well as the documentation piece. The documentation makes it much easier to follow for someone else and for myself after stepping away from it for a couple of months. After cleaning my first round which included the removing of retweets, I went back and removed the retweeted columns in the archived dataset as those were no longer needed to determine the status of tweets as retweets. I did not rename the id column in the tweet api download and realized I needed to match columns names in order to merge on that column. I went back and rename id to tweet_id so columns headings would match in the data frames I would be merging as the tweet_id column was the merge key.   I also went back to separate the date column to included columns month, day and hour as I thought it may be interesting to see when these dog lovers do most of their tweeting.