Wrangled write up

Gathering

My first step in this data wrangling project was to gather all the data. For the .csv file this was pretty straight forward because I just had to download it onto my local computer and then could just use pandas to read it in. For the .tsv file, I had to use the requests library to download the file through its URL, whilst this way a bit more complicated than the .csv file, I was able to quickly get it working by rewatching the Udacity lesson on how to use the requests library. Lastly, I had to use the Twitter API, Tweepy, the gather data about tweets such as retweet count and favourite count from various tweets, and save them into a .json file. I struggled with this for quite a while but was able to find a blog [1] of someone going through how to do it. I have linked the blog in the Jupyter Notebook, and below in the resources section.

Assessing

For assessing the data, I used mainly programmatic assessment such as using the .info() function to get a brief overview of the DataFrames. I also used visual assessment however, such as when I noticed that the URL in the 'source' column of the twitter_archive.csv contained HTML. The majority of the quality issues which I found included removing tweets that were retweets or responses, as we only wanted original tweets for this analysis. By using the .info() function, I was also able to notice that some columns such as the 'expanded_urls' contained missing rows, however not many, so I was able to remove these without affecting the integrity of the dataset. I also checked for duplicate data and could not find any.

Cleaning

For the cleaning portion of this project, I made sure to copy all of the DataFrames to avoid messing up the original DataFrame. My main steps included dropping all the rows which were not original tweets through various drop methods. I also dropped all rows where the rating_denominator was not 10. Considering the majority of the rows had a denominator of 10, I felt that removing these not 10 rows would improve the integrity of the data. I also removed all rows with a numerator above 14, as the admin of the Twitter account said themselves that they had not rated a dog above 14, therefore, the few ratings which were above 14 were deemed 'joke' rows and were removed. I also removed the last 6 characters of the timestamp column and then converted it into a DateTime type, this was done to remove the microseconds as they were always 0.Resources

[1] Tweepy tutorial: **http://blog.impiyush.com/2015/03/data-analysis-using-twitter-api-and.html**
[2] Remove HTML regex: **https://stackoverflow.com/questions/45999415/removing-html-tags-in-pandas#comment89865881_45999467**