Udacity Data Wrangle Module
Project – Wrangling Twitter Data (WeRateDogs)

Data Wrangling

May 22, 2020
Author: Jiayi Zhang

Motivation: The aim of this assignment was to gather raw and messy data from a variety of sources (namely available data, programmatically downloading data, and web scraping via the twitter API) and to assess, clean and make valuable analysis.

Part 1: Data Gathering

Three sets of data were gathered from 3 separate sources. The initial WRD archive was provided for by the assignment as a csv file that could be read directly. A predictions file was also provided but had to be programmatically downloaded into the appropriate directory before being read. Lastly, the most difficult section, was to programmatically scrape additional Twitter data from Twitter using an API. I was able to achieve this by gathering a list of tweet_ids from the given twitter archive, then using a for loop in conjunction with Tweepy, search for each individual tweet_id off twitter and download its json file (which contains various additional information regarding each tweet). Lastly, the file containing all jsons was read individually, and information related to each tweets' "number of likes" and "retweet count" was extracted to form our third and final dataset.

Part 2: Data Assessment

This section visually and programmatically inspects each dataset to identify and document tidiness and quality issues, which will be used during the data cleaning stage. For my three datasets, the list of issues is documented as follows:

Tidiness Issues:

1. *doggo*, *floofer*, *pupper*, *puppo* should be one column as opposed to four
2. all three dataframes *archive*, *predictions*, *tweet_data* should be merged into a single dataframe

Quality Issues:

1. Wrong datatypes for *tweet_id* in all three tables (*int* should be *str*) as well as the *timestamp* column in *archive* (*str* should be *datetime*)
2. dropping columns and rows that are not particularly useful to our analysis
3. some predictions in the *predictions* dataset were not images of dogs
4. some *rating_denominator* in *archive* were not 10
5. **missing data!** Inconsistent number of entries between the three dataframes

6. predictions of dog breed (*p1*) in *predictions* are separated by underscores (_), would be more visually appealing if spaces were used
7. inconsistencies in the capitalization of dog breeds (*p1*) in *predictions*
8. 55 dogs named "*a"*, 8 dogs named "*the*" and 7 dogs named "*an*" in archive
9. p1, p1_conf, *jpg_url* and *expanded_urls* should be changed to *predicted_dog_breed, prediction_confidence, image_url* and *tweet_url* for clarity
10. since all non-dog tweets were dropped, the column *p1_dog* can also be dropped
11. minor fix: reorder columns

Part 3: Data Cleaning

Most of the cleaning work was quite straightforward, however, I will use this section to justify a few of my decisions:

1. Because we only want to keep original tweets from WeRateDogs, tweets that were comments or retweets from other accounts were dropped from our dataset
2. Tweets that were predicted to not be dogs (approx. 532 tweets) were dropped
3. The number of tweets between all three datasets are inconsistent for various reasons. One complication which we must first address is figuring out which dataset to use as the model for the merge. We previously filtered all tweets that were retweets and comments from *archive*, but these still exist in the predictions and *tweet_data* datasets. On the other hand, we filter out all tweets that weren't dogs from *predictions*, however these still exist within *archive* and *tweet_data*. The simplest method that maintains the most data would be to remove all "nondog" data from *archive*, and to merge *predictions* and *tweet_data* onto *archive*.