wrangle_report

March 22, 2021

For this project, I was tasked with working with the WeRateDogs twitter data. The wrangling efforts consisted of 3 parts: gathering, assessing and cleaning.

To start, the gathering efforts required different methods of data gathering as there were 3 different sources. First, a .csv file was provided, which was easily accessed in the notebook with the pandas module. The second set of data was located on the udacity servers and we first used the python requests library access the data, followed by simple os library to write it into a file within our workspace. Similar to the first data source, we completed the gathering using pandas to read the .tsv file. For the last data source, we were required to pull data directly from Twitter. Fortunately, multiple APIs exist to accomplish this task. In our case, we used Tweepy. With Tweepy, we iterated over each of the tweet IDs in our first data source (.csv file) and populated a json file. We then wrote line by line from the json file into a .txt file. Finally, from the .txt file, we populated our dataframe by extracting columns of interest (tweet_id, retweet counts and favorite counts). This led us with 3 dataframes to use in our next step.

The next step in the wrangling process is assessing. In our first dataframe, the WeRateDogs dataframe, we found significant number of quality issues. These are: - some rows are replies/retweets to tweets (not original tweets) - extra not useful columns (retweeted_status_id etc.) - timestamp should be datetype - tweet_id is int64 but should be string - incorrect dog names ('very', 'the' etc.) - ratings are int64 but should be floats to preserve decimals

For our image dataframe sourced from the udacity servers, some quality issues were: - tweet_id is int64 but should be string - extra not useful columns - underscores in dog names that are 2 words

And lastly, in our tweets dataframe from Twitter, we had: - tweet_id is int64 but should be string - retweets and favorite counts are objects (strings)

Tidiness issues were less, with 2 main issues. The first one was that the dog types (pupper, floofer etc.) were in individual columns. As we cannot have values as column headers, we will have to find a way to combine into a single column. The second tidiness issue was that the rating numerator and denominator should be combined into one rating to be used to compare across rows.

The last step is the cleaning process. Addressing all the issues in the assessing step, we used pandas module and simple python to fix all the issues. Once done, we joined the 3 dataframes into 1 on the tweet_id column, and proceeded to filter out rows with missing values. We finally exported the data to a .csv file.