A Data Wrangling Report

Data Gathering

I worked with three datasets in this project, two of which were provided by Udacity. The first dataset, which I refer to as the **twitter_archive_data** was downloaded manually. It contained over 2000 tweets about dogs by WeRateDogs between 2015 and 2017 but missed two important columns which ultimately formed the third dataset I had to gather. The second dataset contained predictions of dog breeds in each tweet. I had to download it by writing some code. Finally, I employed **tweepy**, a python library, in querying Twitter's API to gather the third dataset, containing namely, the number of retweets, and the number of likes for each tweet.

Using tweet IDs from **twitter_archive_data**, I obtained the associated tweets iteratively. I wrote each tweet's json data line by line into a text file. Each json object was then loaded into a list, using python's json library. To conclude this phase, I grabbed the tweet IDs, retweet_count, as well as the favorite_count and loaded them into a dataframe.

Data Assessment

I began to assess each dataset visually and programmatically. I first spotted the pressing need to merge the `twitter_archive_data` with the newly collected twitter_data as they did not need to exist separately. Further, I noticed four columns that went against the rule of **tidy data**. These were slang names given to the dogs, namely *doggo, puppo, pupper*, and *floofer* and needed to be represented by just one column.

The data also contained retweets that were not needed for this project. Further, some columns had incorrect data types, others were a mix of relevant data that needed to be separated from the irrelevant part, some contained missing values, others contained invalid data, and some were not particularly helpful.

I also noticed some discrepancies with the ratings, as some denominators were above–while others were below–10; moreover, there were some unusually high ratings such as 420, 1776, and more. With some research I learnt some of these instances resulted from several dogs being rated at once in the same tweet, which also explained why some denominators were multiples of 10. I realized it was best not to address this until it might be required during analysis.

There were no major issues with the image predictions dataset.

Data Cleaning

Addressing data tidiness issues first, a copy of `twitter_archive_enhanced` was merged with the newly collected twitter data based on the IDs common to both of them. This effectively reduced the size of the dataset as some tweets had been deleted. I also combined the four columns relating to the dog stage slang into one and replaced the tweets which had no dog slang with proper null values.

Next, I identified retweets and dropped them, renamed some columns that weren't explanatory enough, cast the time of a tweet as a datetime object and addressed missing data appropriately. I also dropped some columns, including "in_reply_to_status_id" and "in_reply_to_user_id" as they contained mostly missing data and were not helpful. Finally, I replaced some incorrect dog names with null values. I must point out here that not all the tweets revealed the names of the dogs being rated, but neither was the algorithm that was used to extract the dog names perfect also. Hence, the need to use null values in these instances.

The cleaned tweet data was finally merged with the image predictions dataset and stored under the name “twitter_archive_master.csv”.