Wrangle Report

Wrangle and Analyze Data Project Submission
Data Analyst Nanodegree Program
Jessica Ertel

Data Gathering

There were three elements of data gathering involved in this project. The WeRateDogs Twitter archive was provided by Udacity and downloaded manually. Using the tweet IDs from the WeRateDogs Twitter archive, I queried the Twitter API for each tweet's JSON data and wrote the data to a text file. This was the most time-consuming portion of the data gathering phase. In order to use the scraped tweets in the text file, I imported them into a Pandas DataFrame. Finally, I programmatically downloaded tweet image predictions according to a neural network from Udacity's servers using the Requests library.

Data Assessment

After gathering the data, I assess them visually and programmatically for quality and tidiness issues. While there are many areas that can be cleaned, I limit this project to addressing 2 tidiness issues and 8 quality issues, including those that satisfy the Project Motivation.

Data Cleaning

Tidiness Issues

I decided to merge the posts I scraped from Twitter with the WeRateDogs twitter archive on 'tweet_id' since this is what I used to query the twitter API. I decided to leave the image predictions as a separate DataFrame since the data was not directly related to the information in the tweet. The second issue I addressed was the high number of null values appearing in the 'in_reply_to_status_id', 'in_reply_to_user_id', 'retweeted_status_id', 'retweeted_status_user_id', and 'retweeted_status_timestamp' columns. These columns contained over 90% null values and I determined they would not be relevant to this analysis. I also dropped the 'expanded_url' and 'source' columns since they did not provide useful information for the analysis.

Quality Issues

The first step was to ensure the twitter posts DataFrame only contained posts that had been fed through the neural network to generate an image prediction. I used the twitter IDs in the the `image_preds` DataFrame to filter the twitter posts. Addressing incorrect data types was an important data quality issue. I changed 'tweet_id' to a string and updated the two timestamp columns to a pandas datetime object. After comparing the timestamp columns to ensure consistency, I dropped the duplicate.

The original data contained four different columns documenting the dog stage as either doggo, pupper, puppo or floofer. Some dogs fell into multiple categories. I decided to consolidate this information into one column. Next I addressed several inaccurate dog names that had been

identified from the ‘text’ column. I changed names like 'just', 'a', 'the', 'an', 'by', 'his', 'O', 'my', 'all' to 'None'.

As described in the Project Motivation, the fact that the rating numerators are greater than the denominators does not need to be cleaned. However, there were several records with denominators over 10. After closer inspection I realized that some of the scores were incorrectly pulled from the ‘text’ column. I updated the scores where appropriate.

Next I moved to the ‘image_preds’ DataFrame. It was important to change the strings in all the columns containing neural net predictions to lower case, so that golden_retriever wouldn’t be classified separately from Golden_Retriever. Finally, I changed some of the column names in the DataFrame to make them more intuitive.