The Wrangling Process¶

By Nikhil Kumar

WeRateDogs Twitter Data

July 2020

The wrangling process of this project included gathering, assessing, and cleaning the data from the famous WeRateDogs Twitter account. After all the data was successfully gathered from multiple sources, it was assessed for any quality and tidiness issues that could potentially inhibit thorough analysis for insights.

Gathering the Data¶

The only file available on hand was the twitter archive data stored in *twitter-archive-enhanced.csv*. This file contained useful columns such as each tweet's id in the **tweet_id** column, indication if the tweet was a reply in the **in_reply_to_status**, indication if the tweet was a retweet in the **retweeted_status_id**, and the **rating_numerator** and **rating_denominator** for the rated dog in each tweet.

The next set of data collected was a dataset that contained a prediction for what breed the dog in each tweet was. This dataset was a result of each image in the WeRateDogs twitter account being run through a neural network that would predict what the images in each tweet contained. This data was hosted on the servers at Udacity and was retrieved using the corresponding URL to be stored in the *images_predictions.tsv* file.

While the archive data did provide valuable information, it was missing two very important pieces of information: the number of favorites for each tweet and the number of retweets for each tweet. This data was retrieved through Twitter's API: tweepy. The JSON data for each tweet was pulled using the API and stored in the *tweet_json.txt* file. This file was programmatically read through to pick out the number of favorites and retweets for each tweet and ultimately store all of this data into a data frame saved as *api_data.csv*.

Assessing and Cleaning the Data¶

There were several issues that needed fixing in order to prepare the gathered data for analysis. For the archive data, there were many quality issues such as the fact that there were tweets in the file that were either retweets or replies. These tweets were not original tweets made

by WeRateDogs that included a dog rating. All of such tweets were picked out and deleted from the table. Another notable quality issue was that there were certain outliers when it came to ratings which were far too large. After looking into the details of these tweets, it was found that these numerators of these ratings were odd and extremely large as they were part of a joke and not legitimate ratings for a dog.

The only notable quality issue for the image prediction data was that 543 of the best predictions for each tweet were not dogs. Due to the fact that only the best predictions were to be included in the analysis, all tweets with this condition were deleted as they would be of no use. This data was then tidied up by getting rid of the second and third predictions for each tweet and renaming the columns to be clearer and more descriptive.

The API data did not have any quality issues; however, its id columns was named **id** rather than **tweet_id** so this had to be changed to properly merge this data with the rest. All of this data was merged together and saved into the *twitter_archive_master.csv* file. All other observed quality and tidiness issues that were assessed and cleaned are listed out in the attached notebook.