Data Wrangle Report

Gathering Data for this Project

For this project, data was gathered from the three sources indicated below. Different methods of data collection were utilized for each of the data sources, including:

Importing data via csv

Using requests to download data off internet

Scrape data from an API

Three data sources:

- Enhanced Twitter Archive - The Udacity-provided WeRateDogs Twitter archive. This only includes some of the basic tweet information for all 5000+ of their tweets. I personally downloaded this file by visiting the `twitter_archive_enhanced.csv` URL.
- Image Predictions File - The tweet image predictions, i.e., what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network. This file (`image_predictions.tsv`) is hosted on Udacity's servers and should be downloaded programmatically using the Requests library and the following URL: image_predictions.tsv
- Data via the Twitter API - The number of favorites (or "likes") and retweets each tweet has received, as well as any other information you think is relevant. Using the tweet IDs from the WeRateDogs Twitter archive, use Python's Tweepy package to query the Twitter API for each tweet's JSON data, and then store the whole set of JSON information for each tweet in a file called tweet json.txt. JSON data for each tweet should be written to a separate line. Then, each by line, read this.txt file into a pandas DataFrame that contains (at a minimum) the tweet ID, retweets, and favorites.

Data Assessing

After gathering the datasets, I then proceed to assess the datasets. Assessment of the dataset was done visually and programmatically.

In this section, I focused on the Quality and Tidiness issues of the datasets.

At the end of the data assessments, I was able to come up with these issues:

- The retweeted_status_timestamp, timestamp data types should be datetime instead of object (string).
- Incorrect numerators with decimals

- Some values in `numerator_rating` and `denominator_rating` seem to be in error or suspicious outliers. Incorrect names or missing names in name column such as, a, an, the... - all are written with lower case letters
- Tweet_id data type is wrong
- p1, p2, and p3 should be categorical data type.

Data Cleaning

To the best of my ability, I was able to overcome the aforementioned problems using my understanding of Python and online resources like Google, Stack Overflow, etc. There was a lot of trial and error involved when using regular expressions in challenging situations, but other tasks, like removing the unwanted columns, removing duplicates were rather simple.

Overall, I learned a lot about how to use python effectively and efficiently to clean data and store it.