Wrangle Report

James Franchino

The purpose of this report is to put into practice what I have learned in the Masterschool data wrangling unit. In this unit we used data from the Twitter user @dog_rates, better known as WeRateDogs. WeRateDogs is a humorous Twitter account that rates pictures of peoples dogs, almost always, with a denominator of 10 and a numerator over 10 (i.e. 14/10).

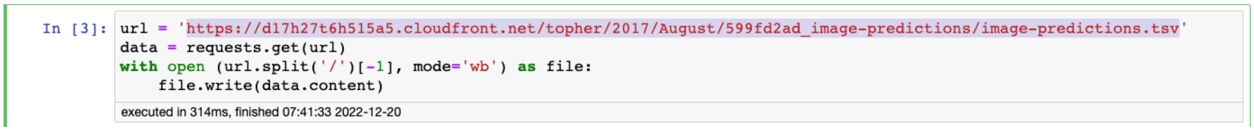
This project consisted of three major parts:

- **Gathering data**
- **Assessing Data**
- **Cleaning and visualizing data**

Gathering Data:

The data for this project came in three distinct files.

- **twitter-archive-enhanced.csv** was downloaded directly from Udacity.
- **image-predictions.tsv** is the second file in this data set and came from Udacity's servers, downloaded programmatically using pythons' requests library. The full URL for the file is https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv



```
In [3]: url = 'https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv'
        data = requests.get(url)
        with open (url.split('/')[-1], mode='wb') as file:
            file.write(data.content)
        executed in 314ms, finished 07:41:33 2022-12-20
```

- **tweet-json.txt** was obtained by querying the Twitter, API, and matching the tweet_id column using the tweepy library. Despite my best efforts, Twitter would not allow me access to their enhanced API, so I downloaded a archive of the data from Udacity's servers.

Assessing Data:

I assessed the downloaded data in two ways:

- I began visually, opening the files in Microsoft Excel and looking them over before loading them into three separate pandas DataFrames in a Jupyter Notebook.
- I then assessed the data programmatically using pandas and python. I used methods such as .info(), .head(), .tail(), and .value_counts().

I documented issues in the data that I found into two separate categories, tidiness issues, and quality issues. After documenting these issues, I copied and combined the three datasets into one to make the next step easier.

Cleaning Data

I then begin working through the issues that I found programmatically. These issues consisted of things like columns with wrong data types, combining multiple columns into one column, separating a text string from a URL, removing unneeded columns, removing columns that were primarily NaN or null values, and removing HTML tags from a text column.

With each of these issues, I first defined the problem, I then wrote code to fix the issue, and then I tested my code to make sure that it was successful and effective. Each of these steps were performed in a Jupyter Notebook, each segregated into sections:

- Define - I defined the issue that I am fixing
- Code – I wrote the python code to fix the issue
- Test – I tested my solution to make sure that it was successful and effective

Some of these problems were quite challenging but gave me excellent practice for what I may be doing in the real world of data analysis.
After all the cleaning steps were completed, I saved my work one last time into a new file, **twitter_archive_master.csv** before exploring, and visualizing the combined and cleaned data.