Subject: Wrangle and Analyze Data

Date: Sunday, September 25, 2022

Author: Adaobi Muoemenam

1.0 Introduction.

As a student of Udacity currently enrolled in the ALX-T Data Analyst Nanodegree Programme, this project is one of several requirements for successful completion of the course. In this project, we will wrangle WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations. The primary tools used are Python programming language in a Jupyter notebook titled wrangle_act.ipynb.

2.0 The Data.

I worked with three datasets:

Enhanced Twitter Archive:

The WeRateDogs Twitter archive contains basic tweet data for all 5000+ of their tweets, but not everything. One column the archive does contain though: each tweet's text, which I used to extract rating, dog name, and dog "stage" (doggo, floofer, pupper, and puppo) to make this Twitter archive "enhanced."

Additional Data via the Twitter API

This contains the retweet count and favorite count of the tweets in the archive by querying Twitter's API.

Image Predictions File

This contains image predictions (the top three only) alongside each tweet ID, image URL, and the image number that corresponded to the most confident prediction (numbered 1 to 4 since tweets can have up to four images).

3.0 Project Activities

In this project, I was tasked with the following activities:

1. Data Wrangling. This is grouped into three steps viz:
   - Gathering data.
   - Assessing data.
   - Cleaning data.
2. Store, analyze and create visualizations of the wrangled data.
3. Write a report.

Gathering data

Below is a summary of how the data in section 2 above was gathered:

Twitter Archive: Udacity extracted this data programmatically and made it available for the students to use. The file name is twitter_archive_enhanced.csv

Image Predictions: This file (image_predictions.tsv) is hosted on Udacity's servers and was downloaded programmatically using the Python's Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv

Additional Data: This can be accessed via two ways. One is programmatic and involves using the tweet IDs in the twitter archive to query the twitter API for each tweet's JSON data using python's Tweepy library. The relevant data is then extracted from each JSON and stored in a tweet_json.txt file. Alternatively, this file was provided for those with reservations to the afore mentioned approach.

4.0 Accessing Data

I used both visual and programmatic approaches to access the three datasets. This stage involves identifying and taking note of the issues in the datasets which fall under two categories of Tidiness and Quality issues.

5.0 Cleaning Data

Before I began this stage of the wrangling process, I made copies of each dataset I was working with. Then, each of the issues identified in the Access Data step above were then cleaned using the Define-Code-Test framework.

For the twitter archive, I fixed the issues of wrong datatypes, purged records with wrong dog names, and dropped all retweets (we were only interested in original tweets) among other issues.

The issue of joining dog breeds with underscore was fixed in the image prediction's dataset. I also handled the case of incorrect and inconsistent casing for dog breeds.

Retweets were also dropped from the tweet json dataset.

6.0 Storing Data

The three datasets were merged into one and saved in a csv file named twitter_archive_master.csv

7.0 Conclusion

Data wrangling is necessary before analysis can be carried out on real-world data. The processes involved are also iterative and can be revisited as many times as the need arises in the data analysis process.