Luke DeMaster-Smith
https://github.com/ldemastersmith/GoodDoggo

Project Motivation

The primary purpose of this project was to demonstrate and practice the data wrangling process. In order to do so, multiple pieces and types of Twitter Data related to the WeRateDogs Twitter account (@dog_rates) were collected. This project collected data from the following sources:

- a (provided) Twitter archive
  - .csv format
- the results of a neural network analysis of images associated with WeRateDogs Twitter account
  - .tsv format
- the Twitter API, accessed using Tweepy inside of a Jupyter notebook
  - JSON and .txt formats

Process

After the tweet archive was downloaded from local storage and the tweet image analysis data was downloaded from a URL, additional tweet data was gathered via a Tweepy API object. While numerous pieces of data could have been gathered with this last method, it was decided to only capture some basic information including 'tweet_id', 'favorite_count', and 'retweet_count'.

After the data was gathered, it was assessed for Quality and Tidiness issues. Several completeness, validity, consistency, and tidiness issues were subsequently defined, corrected (via code), and then checked (also via code). Since the objective of this project was to *demonstrate* and *practice* the data wrangling process, the data was not assessed for every possible or likely issue. There are still known (and probably unknown) issues with the data. For instance, no notable accuracy issues were identified, but some probably do exist.

Ultimately, the following issues were defined, cleaned, and verification tested:

- **Quality:**
  - Different number of records in the (initial) dataframes
  - Incorrectly high numerator values
  - Numerators with decimal points were rounded up
  - Tweets with incorrect numerator and denominator pairs were corrected
    - i.e., tweets with multiple fractions in the tweet text and had the wrong numerator and denominator values assigned when ratings were parsed
  - Numerators and denominators with zero values

    - ‘timestamp’ and 'retweeted_status_timestamp' changed to type ‘datetime'
    - changing the following columns to type 'int64':
        - 'in_reply_to_status_id'
        - 'in_reply_to_user_id',
        - 'retweeted_status_id'
        - 'retweeted_status_user_id'
    - Removed retweets
    - Converted ‘tweet_id’ to type int64 for consistency across dataframes
        - *NOTE*: converting all ‘tweet_id’ values to type string would have also been valid
    - Added a 'none' column to df_archive for tweets that do not have a dog stage
- **Tidiness:**
    - dropped columns 'retweeted_status_id', 'retweeted_status_user_id', and ‘retweeted_status_timestamp'
    - combined ‘doggo', 'floofer', 'pupper', and 'puppo' columns into a single column
        - (since they represent different categories of the same variable)
    - extracted url’s from the ‘text’ field
    - merged the ‘df_archive’ and ‘df_tweetInfo’ (i.e., the dataframe with JSON data) dataframes into a single dataframe
    - In ‘df_master’ and ‘df_images’, kept only those tweet_ids that were common to both df’s

After the above quality and tidiness issues were defined / corrected / tested, some light feature engineering was performed to enable more insightful analysis.

Potential Future Work

This analysis could become vastly more interesting by performing more extensive feature engineering. Many of the variables contain a lot of information (for instance, ‘text’, ’timestamp’, or ‘name’) or could be confined with each other to create new information.

Last, but not least, some potential data issues may still exist, such as:

- There maybe some entries that are not pictures of dogs
- In ‘df_master’, invalid dog names, such as ‘a’, ‘an’, ‘the’, etc.
- In ‘df_images’, checking whether the # of tweets with a .jpg_url matches the total # of tweets in the df
- the existence of records that include a picture and a dog name but do not actually have a picture of a dog