Monica London
9-12-2018

WeRateDogs Data Wrangling Report

The WeRateDogs twitter data wrangling project consisted of gathering data from three different sources, assessing, merging the data into one data frame, cleaning, and finally analyzing the data.

Gathering

Source #1 twitter_archive_enhanced.csv - I downloaded this data set from the Udacity project website. I used pandas.read_csv() to load the data set into my Jupyter notebook.

Source #2 image_predictions.tsv – This file was hosted on the Udacity server. With the given URL, I downloaded the file using the Requests library. I then read in the file using pandas.read_csv().

Source #3 Using the tweet ids from the twitter_archive_enhanced.csv data set, I queried Twitter's API for each Tweet's JSON data using Python's Tweepy library and stored each tweet's JSON data, written to its own line, in a txt file. I then read the txt file line by line into a pandas data frame with tweet id, retweet count, and favorites count.

Assessing

I visually and programmatically assessed all three data frames, documenting all tidiness and quality issues.

Tidiness Issues

- Favorite_retweet: rename column 'id' to 'tweet_id' in order to match the other dataframes
- Combine all three dataframes
- Each dog stage has a separate column. Create a single column titled 'Dog Stage'.

Quality Issues

- Inaccurate ratings:
    - Inaccurate denominators: 18 instances of rating denominators that aren't 10

        - Delete tweet id 8.109847e+17 that inaccurately uses 24/7 as the rating. There is no actual rating for this tweet.
        - Some rating numerators are extreme outliers
        - Some rating numerators are 0. Check to be sure these are legitimate observations.
- Dataset contains retweets
- Incorrect names
    - Non-names listed in name column:
        - Lower case names are not real names. Replace with correct names.
        - O'Malley is listed incorrectly as 'O'. Correct name.
    - Change 'None' names to NaN
- Erroneous data types (tweet_id, in_reply_to_status_id, and in_reply_to_user_id to string, timestamp to datetime, and dog stage to category)
- Incorrect dog stages:
    - Some tweets have multiple dog stages
    - Change blank dog stage to NaN values
- Source column includes URL and other characters, rendering it difficult to analyze

Merging/Cleaning

I created copies of all three data frames. In preparation to merge the three data frames together, I changed the id column name in one data set to match the id column name in the other two. I then merged all three data sets using the pandas merge function on the tweet id column.

I cleaned all documented tidiness issues followed by quality issues by defining, coding, and testing each issue.

Once the cleaning process was completed, I stored the final data frame and began analysis.