Udacity Data Analyst Nanodegree
Data Wrangling Project Report – Data Wrangling Steps

By: Amanda Hanway, 12/2/2023

Overview

This report will describe the data wrangling efforts applied throughout completion of the project.

Step 1: Gathering Data

The data used throughout this project were acquired from the following data sources using the methods noted:

1. WeRateDogs Twitter archive
    a. File: twitter-archive-enhanced.csv
    b. I downloaded the file from the "Step 1: Gathering Data" page in the Udacity course.
2. Tweet image predictions
    a. File: image_predictions.tsv
    b. I downloaded the file programmatically in the Jupyter notebook using the python Requests library and this url.
3. Additional data from the Twitter API
    a. File: tweet-json.txt
    b. In place of using the Twitter API, I downloaded this file from the "Additional Resource: Twitter API" page in the Udacity course.

After acquiring the files, I read each into a separate pandas dataframe. Because the image predictions and twitter API files were in .tsv and .txt formats, I also saved these dataframes to .csv files for easier review in Excel.

Step 2: Assessing Data

Each of the three datasets were first assessed programmatically using the pandas dataframe.info() function. This function displays the column names, non-null value counts, data types and total entries for the dataframe.

The datasets were then assessed visually by printing each dataframe to the screen, as well as reviewed in detail within Excel. The following quality and tidiness issues were identified by assessing the data.

Quality issues for twitter-archive-enhanced.csv:

1. The data includes replies and retweets that should be filtered out of the dataset so only original ratings are included. Include only rows where in_reply_to_status_id is null and where retweeted_status_id is null.

2. Data includes tweets without an image that should be filtered out. Include only rows where expanded_urls is not null. (Ex. 785515384317313025)
3. The numerator and denominator fields may be wrong if the text included numbers. Since most rows appear to follow this pattern, the text after "https://" should be removed. The denominator can be extracted as the last number in the string. The numerator can be extracted as the last number before the "/" between the numerator and denominator. (Ex: ID 716439118184652801, 50/50 instead of 11/10).
4. The numerator field may be wrong if the rating was a decimal. This is fixed with the update from item 3. However, a new issue presented itself where if the numerator has a "." immediately before it, it is being considered a decimal number. Fix this by looking for values where "." is the first character and removing the ".". (Ex. ID 883482846933004288, 13.5/10 shows as 5/10, ID 772114945936949249, Rating_numerator_new = .10)
5. Rating_numerator and rating_denominator are in int64 format and do not allow decimals as entered. Convert this to float64.
6. The name field value is "None" if the dog's name could not be identified. This should be replaced with a more usable label or made null.
7. The timestamp column is an object data type. Convert this to datetime64.
8. Some denominators are made up and do not fit the standard rating system. The invalid denominators should be standardized by setting them to 10, or removed.

Tidiness issues for twitter-archive-enhanced.csv:

1. The columns doggo, floofer, pupper, and puppo can be combined into one column. The new column will be a variable called "dog_stage".
2. The source column is in html format. Standardize this field into the following categories:
    a. Twitter for iPhone
    b. Twitter Web Client
    c. Vine - Make a Scene
    d. TweetDeck
3. All three datasets should be merged into one master dataset. Use the twitter-archive-enhanced.csv as the main dataset and merge the image_predictions.tsv and tweet-json.txt into this data on the tweet_id.

Step 3: Cleaning Data

Each of the data quality and tidiness issues were corrected through cleaning. The defined issue was coded, then tested, and reviewed for validation.

Step 4: Storing Data

The final master dataframe was saved to the file twitter_archive_master.csv.

Limitations

The data wrangling steps completed in this project did not solve all issues in the datasets. Given additional time for review, the text fields in the twitter-archive-enhanced.csv file should be scrubbed further to correct for discrepancies in the extracted fields. The image_predictions.tsv file should be reviewed in depth to standardize fields and make use of the output in the analysis. More recent data could be extracted from the Twitter API to make the analysis up to date. These areas have been identified for future work on this project.