Wrangle Report for WeRateDogs Tweet Archive

Date

Farnaz Zarian

January 7, 2021

Project Overview

Using Python libraries, I will gather data from a variety of sources and in a variety of formats, I will perform data wrangling on these datasets by assessing their quality and tidiness, then cleaning them.

The dataset that I will be wrangling (and analyzing and visualizing- in another file) is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10 and a numerator of greater than 10. Why? Because "they're good dogs Brent." WeRateDogs has over 8.9 million followers and has received international media coverage.

Data for the Project

Enhanced Twitter Archive

The WeRateDogs Twitter archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017. Udacity provided this data as a csv file.

Additional Data via the Twitter API

Retweet count and favorite count are two of the notable column omissions in the Enhanced Twitter Archive. I am going to query Twitter's API to gather this valuable data.

Image Predictions File

The tweet image predictions, which represent what breed of dog, other subject or animal is present in each tweet (according to a neural network), will be downloaded programmatically.

Data Wrangling Process

I gathered the data from different sources and documented unclean issues first. Then assessed the data files for quality and addressed completeness, validity, accuracy and consistency of the available data. Next, I assessed the data for tidiness to make sure that structural or organizational issues are addressed. For this process, I ensured that each variable forms a column, each observation forms a row, and each type of observational unit forms a table. The assessment process started with visual assessment and was followed by programmatic assessment. Finally, I cleaned the data by fixing the quality and tidiness issues that were identified in the previous step, using Python and Pandas by following a Define ➔ Code ➔ Test process for each quality and tidiness issue.

Below are the issues that raised during the assessment process:

Quality Issues

WeRateDogs_twitter_archive

- *retweeted_status_id* is a float and not an integer
- Multiple formats for *retweeted_status_id*
- *retweeted_status_user_id* is a float and not an integer
- Multiple formats for *retweeted_status_user_id*
- *retweeted_status_timestamp* is a string and not a datetime object
- Missing records in (*in_reply_to_status_id, in_reply_to_user_id, retweeted_user_id, retweeted_status_user_id, retweeted_status_timestamp, expanded_urls* columns)(**can't clean as no additional data available**)
- Odd values for *rating_numerator* and some erroneous values for *rating_denominator*
- There are 4 categorical values in the *source* column. Twitter for iPhone, Vine - Make a Scene, Twitter Web Client, TweetDeck

image_predictions_df

- Lower case *p1* names sometimes, upper case other times
- Lower case *p2* names sometimes, upper case other times
- Lower case *p3* names sometimes, upper case other times
- Erroneous/unrelated information where *p1_dog, p2_dog and p3_dog* are all False
- Missing dog name information for name column

Tidiness Issues

WeRateDogs_twitter_archive

- *doggo, floofer, pupper and puppo columns can be merged into one column (dog_stages) and the data type for the dog_stages needs to be categorical*
- *Two variables in the text column should be split into text and short_urls*
- *The key points in the project description indicate that we are only interested in original tweets and not in retweets. The columns retweeted_status_id, retweeted_status_user_id and retweeted_status_timestamp can be removed to make the table cleaner. Same goes for in_reply_to_status_id and in_reply_to_user_id*