DATA WRANGLING

WeRateDogs

JUNE 11, 2020

JALAL TAREEN

Introduction

My goal: wrangle WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations. The Twitter archive is great, but it only contains very basic tweet information. Additional gathering, then assessing and cleaning is required for "Wow!"-worthy analyses and visualizations.

The dataset that i will be wrangling (and analyzing and visualizing) 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. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. Why? Because "they're good dogs Brent." WeRateDogs has over 4 million followers and has received international media coverage.

WeRateDogs downloaded their Twitter archive and sent it to Udacity via email exclusively for you to use in this project. This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets.

Project Details

Tasks in this project are as follows:

- Data wrangling: which consists of: Gathering data, Assessing data and Cleaning data.
- Storing: which consists of analyzing, and visualizing your wrangled data

Reporting on:

1) Data wrangling efforts
2) Data analyses and visualizations

Gathering Data

Gather each of the three pieces of data as described below in a Jupyter Notebook titled wrangle_act.ipynb:

1) The WeRateDogs Twitter archive.Downloaded this file manually by clicking the following link: twitter_archive_enhanced.csv

2) The tweet image predictions, i.e., what breed of dog (animal, etc.) is present in each tweet according to a neural network. This file (image_predictions.tsv) is hosted on Udacity's servers and should be downloaded programmatically using the Requests library and the following URL: Image-Predictions

3) Each tweet's retweet count and favorite ("like") count at minimum, and any additional data you find interesting. Using the tweet IDs in the WeRateDogs Twitter archive, query the Twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON data in a file called tweet_json.txt file. Each tweet's JSON data should be written to its own line. Then read this .txt file line by line into a pandas DataFrame with (at minimum) tweet ID, retweet

count, and favorite count. Note: do not include your Twitter API keys, secrets, and tokens in your project submission.

Assessing Data

Project Requirment:

After gathering each of the above pieces of data, assess them visually and programmatically for quality and tidiness issues. Detect and document at least eight (8) quality issues and two (2) tidiness issues in your wrangle_act.ipynb Jupyter Notebook. To meet specifications, the issues that satisfy the Project Motivation (see the Key Points below) must be assessed.

Key Points

Key points to keep in mind when data wrangling for this project:

- You only want original ratings (no retweets) that have images. Though there are 5000+ tweets in the dataset, not all are dog ratings and some are retweets.
- Assessing and cleaning the entire dataset completely would require a lot of time, and is not necessary to practice and demonstrate your skills in data wrangling. Therefore, the requirements of this project are only to assess and clean at least 8 quality issues and at least 2 tidiness issues in this dataset.
- Cleaning includes merging individual pieces of data according to the rules of tidy data.
- The fact that the rating numerators are greater than the denominators does not need to be cleaned. This unique rating system is a big part of the popularity of WeRateDogs.

Quality:

- missing and incorrected dog names and the most popular dogname is 'a' which is not a name given by owner.
- Retweeted records removed.
- TimeStamp is in string format.
- Source not extracted from hyper link tag.
- Columns: in_reply_to_status_id, in_reply_to_user_id, retweeted_status_user_id, retweeted_status_id and retweeted_status_timestamp, have a lot of null values.
- The datatype of the tweet_id - columns is integer and should be str.
- Taking care of incorrect ratings such as numerator less than 10 or denominator not equal to 10.
- Dog breeds start with lower case letters and contains '_'

Tidiness

- The dog stage columns in twitter_archive can be arranged into a single column.
- The image predictions could be condensed to show just the most confident dog breed prediction.
- All three dataframes can be combined into one single dataframe.

Clean:

Cleaning process consists of three steps: Define, code & Test. First we define how to tackle the issue. Then, we code to resolve the issue and finally we test our code to see if the issues with the data have been resolved. So, in order to clean these 3 dataframes, I carried out the 3 steps for each of the issues and was finally able to achive a clean dataframe.