Wrangle report

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.
This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017. More on this soon.

Project Motivation

Context

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 analyses and visualizations.

Enhanced Twitter Archive

The WeRateDogs Twitter archive contains basic tweet data for all 5000+ of their tweets, but not everything. One column the archive does contain though: each tweet's text, which I used to extract rating, dog name, and dog "stage" (i.e. doggo, floofer, pupper, and puppo) to make this Twitter archive. Of the 5000+ tweets, I have filtered for tweets with ratings only (there are 2356).

Additional Data via the Twitter API

query Twitter's API to gather WeRateDogs valuable data.

Image Predictions File

One more cool thing: I ran every image in the WeRateDogs Twitter archive through a neural network that can classify breeds of dogs*. The results: a table full of image predictions (the top three only) alongside each tweet ID, image URL, and the image number that corresponded to the most confident prediction (numbered 1 to 4 since tweets can have up to four images).

Project Details

Data wrangling, which consists of:

- Gathering data:
- Assessing data
- Cleaning data
- Storing, analyzing, and visualizing your wrangled data

- Reporting on 1) your data wrangling efforts and 2) your data analyses and visualizations

Gathering data:

1. Twitter-Archive-Enhanced: Available on the project page. Downloadable
2. image_predictions.tsv: Available on the project page downloadable
3. Twitter API:

Extract data from Twitter using the following code:

```
import tweepy

consumer_key = 'YOUR CONSUMER KEY'
consumer_secret = 'YOUR CONSUMER SECRET'
access_token = 'YOUR ACCESS TOKEN'
access_secret = 'YOUR ACCESS SECRET'

auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_secret)

api = tweepy.API(auth)
```

Assessing Data for this Project

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.

Cleaning Data for this Project

Clean each of the issues documented while assessing. The result should be a high quality and tidy master pandas DataFrame.

Storing, Analyzing, and Visualizing Data for this Project

Store the clean DataFrame(s) in a CSV file.

Analyze and visualize your wrangled data

At least three (3) insights and one (1) visualization must be produced.