General Motivation

The key to getting better at data science and furthering your learning as an aspiring data scientist is– Practice, Practice, and Practice. After learning the basic data science skills. One of the most important ways to develop your data science skills and improve your employability as a data scientist is to work on real-world data science projects. The first step is to find an interesting dataset to work with.But what about gathering data from many sources try to standardize it according to many specs

Introduction

Real-world data rarely comes clean. Using Python and its libraries, you will gather data from a variety of sources and in a variety of formats, assess its quality and tidiness, then clean it. This is called data wrangling. You will document your wrangling efforts in a Jupyter Notebook, plus showcase them through analyses and visualizations using Python.

The dataset that you will be wrangling (and analyzing and visualizing) is the tweet archive of Twitter user dog rates, also known as WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog

Project Motivation

Context

Your 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.

Software and libraries

- You need to be able to work in a Jupyter Notebook.
- The following packages (libraries) need to be installed. You can install these packages via conda or pip.
    - Numpy
    - Pandas
    - matplotlib
    - Requests
    - Tweepy
    - JSON

Tasks in this project are as follows:

- 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

The Data Gathering process

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 "enhanced." Of the 5000+ tweets, I have filtered for tweets with ratings only (there are 2356).

Additional Data via the Twitter API

Back to the basic-ness of Twitter archives: retweet count and favorite count are two of the notable column omissions. Fortunately, this additional data can be gathered by anyone from Twitter's API. Well, "anyone" who has access to data for the 3000 most recent tweets, at least. But you, because you have the WeRateDogs Twitter archive and specifically the tweet IDs within it, can gather this data for all 5000+. And guess what? You're going to query Twitter's API to gather this valuable data.

Image Predictions File

Which contains images for dogs as an output or neural network that predict style for each dog

Assessing and cleaning the data

Archive Data frame

1. Timestamp is object but it should be a datetime as it represent time on which tweet uploaded

2. retweeted_status_timestamp is object but it should be a datetime as it represent time on which tweet is retweeted

3. Incorrect dog names like a /an /the should be removed and replace none with nan

4. Name column should be convert to dog_name as it may confuse reader as it is dog name or owner name

5. the most approprciate dtype for the rating_numerator column should be float

6. The ID fields, like tweet_id, in_reply_to_status_id etc. should be objects, not integers or floats because they are not numeric and aren't intended to perform calculations.

7. Doggo, floofer, pupper, puppo columns contain 'None' value where NaN should be used.

8. There are 9 dogs with non-English names so i will drop them as in this analysis i interested in english dogs only

9. Missing values in 'dog_name'showing as 'None'

10. Retweets and Favorite Count: retweet_count and favorite_count should be integers, not floats.

11. The rating_denominator column is acceptable as an integer but is preferred as float since there is nothing stopping future dog ratings from having a number with a decimal in the denominator.

image_preds Dataframe

1. change columns name to something informative for example p1 to first_predection / p1_confd to first_prediction_confidence

Tideness

1. img_num column in image_preds contain number for images in tweet i think this is un-useful info so i will drop this column
2. doggo, floofer, pupper, puppo columns are redundant as they describe same thing
3. we need to join dataframes to make clear and complete picture for the entire information needed for the next step