MOHIT SINDHANI

Objective:

Wrangle WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations.

Tasks

Tasks in this project are as follows:

- Data wrangling, which consists of:
  - Gathering data (downloadable file in the Resources tab in the left most panel of your classroom and linked in step 1 below).
  - Assessing data
  - Cleaning data
  - Storing, analyzing, and visualizing your wrangled data
- Reporting on 1) data wrangling efforts and 2) data analyses and visualizations

Dataset

The dataset is a 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.

Gathering

1. The WeRateDogs Twitter archive: twitter_archive_enhanced.csv
2. The tweet image predictions, i.e., what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network. (image_predictions.tsv)
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, we have queried 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

Accessing

After accessing the data, we found various quality and tidiness issues:

1. **twitter_archive_enhanced.csv**

Issue Quality:

-- Name column has "None", "a", "the", and "an" names that don't seem to be real names

-- timestamp date is wrong format(unecessary +0000 in date). The timestamp knowledge sort is wrong; object ought to be modified to timedata sort of tweet_id ought to be string rather than object

-- there ar fifty nine null entries within the expanded_urls column

-- rating_denominator column ought to solely have one divisor worth - ten

-- rating_numerator have values under ten that ought to be removed.

-- some ratings is extracted from the text of the tweet and be wont to fill within the rating_numerator column

Issue Tidiness:

-- in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, and retweeted_status_timestamp columns have null values in them. this can be a difficulty as a result of we wish solely the first tweets. we'll drop these columns later within the project.

2. image_predictions.tsv

Issues Quality:

some names of dogs area unit capitalized and a few don't seem to be within the p1, p2, and p3 columns

Issues Tidiness:

archive, image , and df_json may be combined into one table since all of them describe one tweet

Cleaning Data

- Merge the clean versions of archive, images, and twitter_counts_df information frames
- Correct the dog type
- now not needed: in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, and retweeted_status_timestamp.
- Delete retweets.
- take away columns now not required.
- amendment to tweet_id from associate degree whole number to a string.
- amendment to the timestamp to correct datetime format.
- Correct naming problems and Standardize dog ratings.
- making a brand new dog_breed column victimization the image prediction information.