Wrangle Report with WeRateDog Data

Author: Mubarak Hamza

September 9, 2022

Purpose: Document data wrangling process including gathering, assessing and cleaning data

Wrangle Report with WeRateDog Data

Introduction of Data Set:

There are three data sets in total for this project: twitter-archive-enhanced.csv, imagepredictions.tsv , tweet-json.txt

The Goal:

The goal of the data wrangle process is to generate a clean data set called twitter_archive_master.csv for data visualization and analysis later.

Data Gathering:
The datasets for this project are from the tweet archive of Twitter user @dog_rates (WeRateDogs).

1. Enhanced Twitter Archive: contains tweet data for all 5000+. Only 2356 records have ratings.
   File name: twitter-archive-enhanced
   Format: csv
   Source: directly download from Udacity website.

2. Image Predictions File: the output from neural network
   File name: image-predictions
   Format: tsv
   Source: get the data from url = 'https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv'

3. Additional Data via the Twitter API
   File name: tweet_json
   Format: txt
   Source: connect Twitter API to download json format text file and use pandas to read into the notebook.

Data Assessing and Cleaning

Data Issues: I found 11 Quality issues and 4 tidiness issues

Tidiness Issues

1. Create dog classifier column and drop individual dog stage columns.

2. merge tables
3. numerator_rating and denominator should be merged in one rating column instead of two column.
4. extract date, time. year, month, and weekday from timestamp

Quality Issues

df_arch

1. Classify all dog stages into one column and drop individual columns ['doggo','pupper', 'floofer', 'puppo']
2. Some of the dog names are not correct (None, an, by, a, …)
3. In timestamp column +0000 is redundant information
4. The data type of the timestamp should be DateTime, not Object
5. Column source content is too long for such source information, shorten it and replace it with a more descriptive one.
6. The columns “in_reply_to_status_id”, “in_reply_to_user_id”, “retweeted_status_id”, “retweeted_status_user_id”, and “retweeted_status_timestamp” have lots of NA values.
7. The data type of tweet_id should be String, not Integer
8. Remove the string starting 'HTTPS' in the text column

image_predictions

9. The prediction p1,p2, and p3 is an uppercase and lowercase mix, also there are “_” in the breed name, also change variable names to a more descriptive name
10. The data type of tweet_id should be string, not Integer

df_tweets

11. The data type of tweet_id should be string, not Integer

Conclusion:

In summary, the most challenging part of the wrangling process for me was cleaning the data in order to prepare it for analysis. All knowledge shared from the classroom were beneficial though I had to source for more knowledge to understand certain concepts and functions better.