wrangle_report

June 27, 2022

0.1 INSTRUCTION: (Reporting: wragle_report)

- Create a **300-600 word written report** called “wrangle_report.pdf” or “wrangle_report.html” that briefly describes your wrangling efforts. This is to be framed as an internal document.

0.1.1 STUDENT INFOMATION

- **Full Name:** Angel Newportright
- **Email:** angeltucker007@gmail.com
- **Country:** Nigeria

1 WeRateDogs Analysis Report

Note: WeRateDogs is a Twitter account

Wrangling the data for the WeRateDogs Analysis involved three(3) steps 1. Gathering the Data 2. Assessing the Data and 3. Cleaning the Data

1.1 1. Data was gathered from three(3) sources

1st Source- The WeRateDogs Twitter Archive * This was a csv file which contained basic tweet data for all 5000+ of their tweets, but not everything. * WeRateDogs downloaded their Twitter archive and sent it to Udacity via email exclusively for use in this project. This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017.

2nd Source - The Image Predition Date * This was provisioned as a http link provided by udacity conatining a `.tsv file`. The link was accessed programatically * This data contains the tweet image predictions present in each tweet according to a neural network that can classify breeds of dogs

3rd Source - Additional data from the Twitter API * This data conatined each tweet’s retweet count and favorite (“like”) count * This data was to be accessed using Python’s Tweepy library * Note that, while i am still awaiting for the tweeter developer access to be granted, a `Json` file was provided by udacity to help facilitate this wrangling process

1.2 2. Data Assessment was done to pinpoint two type of issues

1. Data Quality Problems and

2. Data Tidyness Problems

While Data quality assessment invloved identifying problems such as data types, incomplete data, inaccurate data, inconsistent data, Data Tidyness assessment invloved indentifying problem that are structural such as deleting columns that are irrelevant to our analysis.

1.2.1 2i. Data Quality Problems

Tweeter Archive Data

1. The `timestamp` column has both date and time, `+0000` is not relevant
2. The `source` column has HTML tags with 4 unique tags, each indicating the source (either from iPhone, Web Client,Vine, Tweet Deck)
3. The ratings almost always have a denominator of 10 but `rating_denominator` column has values greater than ten (10)
4. The `timestamp` column is in object data type should be converted to datetime format

5A. The `retweeted_status_id` column has 78 values

5B. The `retweeted_status_user_id` column has 78 values

6A. The `in_reply_to_status_id` column has 181 values

6B. The `in_reply_to_user_id` column has 181 values

6C. The `retweeted_status_timestamp`column has 181 values

7. The `expanded_urls` has 59 missing values. A Good number of the links seems to have expired
8. The name column has `668` values of `None` and `55` values `a`
9. The name columns has some strings in all lower case and some in title case

Image Prediction Data

10. Not all image predictions are dogs. some are book cases,shopping cart etc (p1_dog,p2_dog == False) with varying confidence level

1.2.2 2ii. Data Tidyness Problems

Tweeter Archive Data

1. `doggo`, `floofer` , `pupper` , `puppo` are all classification. Observation fall into just one of these category.

Image Prediction Data

2. The image url column from the Image prediction table should be part of the Tweeter archive data

1.3 3. Data cleaning was done to make our data ready for analysis and visualization

Data cleaning was done with a simple approach.

For each quality and tidyness issue highlighted above;

STEP 1 - The problem was defined

STEP 2 - The code that would solve the problem was defined

STEP 3 - The code was executed

STEP 4 - The result was tested to see if the problem was solved

It is worthy to note that after the required cleaning was done, the file was saved to the local machine