wrangle_report

May 18, 2020

1 Project 04: Data Wrangling: WeRateDogs Tweets Archive

1.1 Gather Data

Data is gather from 3 resources and saved as 3 dataframes.

1.1.1 Gather data from `twitter-archive-enhanced.csv`

Use `pd.read_csv()` to read data from existing file `twitter-archive-enhanced.csv` and save it as ‘data1.

1.1.2 Gather data by download `image_prediction.tsv` using `Requests` library

1.1.3 Gather data from twitter API using Tweepy library

Using tweepy API to save each tweet’s return JSON as a new line in a `.txt` file.

Read the `txt` file and get `retweet_count` and `favorite_count` to store in data3

1.2 Assessing Data

1.2.1 Overview:

`data1` (Twitter archive) columns:

- **tweet_id**: unique id for each tweet
- **in_reply_to_status_id**: if the represented Tweet is a reply, this field will contain the original Tweet’s ID
- **in_reply_to_user_id**: if the represented Tweet is a reply, this field will contain the integer representation of the original Tweet’s author ID
- **timestamp**: time when this Tweet was created
- **source**: utility used to post the Tweet, as an HTML-formatted string. e.g. Twitter for Android, Twitter for iPhone, Twitter Web Client
- **text**: actual UTF-8 text of the status update

- **retweeted_status_id**: if the represented Tweet is a retweet, this field will contain the integer representation of the original Tweet's ID
- **retweeted_status_user_id**: if the represented Tweet is a retweet, this field will contain the integer representation of the original Tweet's author ID
- **retweeted_status_timestamp**: time of retweet
- **expanded_urls**: tweet URL
- **rating_numerator**: numerator of the rating of a dog. Note: ratings almost always greater than 10
- **rating_denominator**: denominator of the rating of a dog. Note: ratings almost always have a denominator of 10
- **name**: name of the dog
- **doggo**: one of the 4 dog "stage"
- **floofer**: one of the 4 dog "stage"
- **pupper**: one of the 4 dog "stage"
- **puppo**: one of the 4 dog "stage"

`data2` (tweet image predictions) columns:

- **tweet_id**: the unique identifier for each tweet
- **jpg_url**: dog's image URL
- **img_num**: the image number that corresponded to the most confident prediction (numbered 1 to 4 since tweets can have up to four images)
- **p1**: algorithm's #1 prediction for the image in the tweet
- **p1_conf**: how confident the algorithm is in its #1 prediction
- **p1_dog**: whether or not the #1 prediction is a breed of dog
- **p2**: algorithm's #2 prediction for the image in the tweet
- **p2_conf**: how confident the algorithm is in its #2 prediction
- **p2_dog**: whether or not the #2 prediction is a breed of dog
- **p3**: algorithm's #3 prediction for the image in the tweet
- **p3_conf**: how confident the algorithm is in its #3 prediction
- **p3_dog**: whether or not the #3 prediction is a breed of dog

`data3` (tweet status) columns:

- **id**: the unique identifier for each tweet
- **retweet_count**: number of times this Tweet has been retweeted
- **favorite_count**: indicates approximately how many times this Tweet has been liked by Twitter users

1.2.2 Quality

- In data1,the tweet_ID is not the right data type and value.
- Some wrong datatypes and values for in_reply_to_status_id, in_reply_to_user_id
- In data1, we only want original ratings (not the retweets).
- We only want ratings with images. Not all ratings have images.
- In data1, some ratings are wrong.
- In data1, some NOK datatype for timestamp
- In data1, nulls represented as 'None' in columns 'name', 'doggo', 'floofer', 'pupper','puppo'.
- In data1, some dog names are not correct.
- In data2, some predictions are not dogs, there is no column for the most possible breed of a dog.

1.2.3 Tidiness

- In data1, the columns 'retweeted_status_id' 'retweeted_status_user_id' and 'retweeted_status_timestamp' are not useful after removing retweets.
- In data1, the columns 'doggo', 'floofer', 'pupper','puppo' show one variable.
- data3 should be part of data1.
- rating_numerator and denominator should be one variable rating.

1.3 Cleaning Data

1.3.1 Issue 1:

In data1, the columns 'retweeted_status_id' 'retweeted_status_user_id' and 'retweeted_status_timestamp' are not useful after removing retweets.

- Define:

Remove all retweets and observations without ID, the remove columns: 'retweeted_status_id', 'retweeted_status_user_id', 'retweeted_status_timestamp'

1.3.2 Issue 2:

Rating without image must be removed

- Define:

Remove observation without images

1.3.3 Issue 3:

4 Columns 'doggo', 'floofer', 'pupper', 'puppo' mean the same. No value represented as word 'None' in columns 'name', 'doggo', 'floofer', 'pupper','puppo'. - Define:

Create column 'stage' to show dog stage, drop columns 'doggo','floofer','pupper','puppo'. Replace 'None' with np.nan.

1.3.4 Issue 4:

data3 must be part of data1 - Define:

Merge content of data3 into data1, on tweet_id

1.3.5 Issue 5:

Missing retweet_count and favorite_count - Define:

Drop the rows with missing value

1.3.6 Issue 6:

In data1, wrong datatype for timestamp - Define:

Convert timestamp to datetime data type

1.3.7 Issue 7:

NOK datatypes and values for in_reply_to_status_id, in_reply_to_user_id - Define:

convert in_reply_to_status_id, in_reply_to_user_id to string data type.

1.3.8 Issue 8:

Issue with 'name' column: no value as 'None', some values are wrong, not capitalized name are wrong - Define:

Set the value wrong names to 'None' and replace 'None' with np.nan.

1.3.9 Issue 9:

- In data1, some ratings are wrong.
- Rating_numerator and denominator should be one variable rating.
- **Define:**
- Create new column rating = rating_numerator/rating_denominator.
- Drop rating_numerator and rating_denominator.
- Drop oberservations with extreme ratings.

1.3.10 Issue 10:

In data2, some predictions are not dogs, there is no column for the most possible breed of a dog and the confidence. - Define:

Create new columns `predicted_breed` and `predicted_conf` for the most possible breed of a dog and the confidence.

1.4 Store Data

Store the clean `df1_clean` in a CSV file named `twitter_archive_master.csv` and `df2_clean` in additional file `twitter_image_predictions.csv`.

```
[ ]:
```