wrangle_report

October 2, 2022

0.1 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.

We have three steps in the data wrangling process. Namely; Data Gathering, Data Accessing and Data Cleaning. The process carried out for each step is detailed below.

1 DATA GATHERING

In this stage, we gathered three different types of data from three different sources. The frist data was "WeRateDogs" twitter archive data that was readily available for download. This data was in a csv format and contained tweets/retweets about the rated dog. This data was easily imported using the pd.read_csv command.

The second data was the "image_predictions" data. This data is a tsv format and contains the prediction confidence of each dog based on the image passed. The dog falls into the category with the highest confidence. The data was read into a jupyter notebook using pd.read_csv and with a tab separator. With a total of 2075 entries and 12 columns.

The third data was the "tweet_json" dataset retireived via Twitter API using consumer_key, consumer_secret, access_token, and access_secret. We retrived the number of liked and retweeted tweet of the id's present in "WeRateDogs" dataset. The data was a total of 2327 entries and three columns with 29 entries missing.

2 DATA ACCESSING

Both visual and programmatic data accessing was done. For programmatic accessment, we used functions like info, isnull, value_counts, and sum to get some details of the data. For visual accessment, we were able to visually spot issues like unessessary html 'a' tag. Data accessing was used to evalute 2 issues. Namely, quality issue and tidiness issue. The issues found under these two categories are listed below:

2.0.1 Quality issues

1. 'WeRateDogs_twitter_archive' table, some columns have duplicates(i.e retweet columns) which is unnecessary: in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id , retweeted_status_user_id, retweeted_status_timestamp"

2. Delete null rows on 'WeRateDogs_twitter_archive' table (expanded_urls).

3. 'Source' column in 'WeRateDogs_twitter_archive' table has html a tag

4. 'WeRateDogs_twitter_archive' table has wrong datatypes: in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id , retweeted_status_user_id (should be integers), retweeted_status_timestamp (should be datetime)

5. The 'rating numerator' column in 'WeRateDogs_twitter_archive' table has some very huge values

6. The 'rating denominator' column in 'WeRateDogs_twitter_archive' table has other values other than '10'

7. Some dog names in 'WeRateDogs_twitter_archive' table have article names such as 'a', 'an', 'the'

8. Not all "WeRateDogs_twitter_archive" have "image predictions"

2.0.2 Tidiness issues

1. 'WeRateDogs_twitter_archive' table has duplicate/unecessary dog "stage" columns (i.e. doggo, floofer, pupper, and puppo). They can all be in a single column

2. Dublicate tables, we can combine 'tw_json' table with 'WeRateDogs_twitter_archive'. Note that 'tw_json' table has 29 rows less, so address this issue

3. Create a 'breed' column in 'image_predictions' table and join it with 'WeRateDogs_twitter_archive' table to avoid duplicates.

3 DATA CLEANING

3.0.1 Quality Cleaning

1. For quality issue 1, we can see that in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id , retweeted_status_user_id, and, retweeted_status_timestamp are almost empty, and they arent useful columns so we can drop them using.drop function.
2. For quality issue 2, only expanded_urls had null entries(59), so we dropped the entries with null.
3. For issue 3, using regex, we extracted the urls from the a tag
4. For issue 4, we easily changed the data type to the appropraite one for timestamp column
5. For issue 5, all numerators grater than 15, were made to be 15, since the according to google search, the current highest rating is 15
6. For issue 6, the only denominator should be 10, so all values in the column were grounded to 10
7. Some dog names were articles, so these names were changed to 'None'
8. For issue 8, not all archived dogs had image prediction, so we deleted those dogs without prediction

3.0.2 Tidiness Cleaning

1. The duplicate dog stages were made to a single column instead of 4 columns
2. We combined tw_json table with archive table to avoid duplicates
3. We only need the 'breed' column from image prediction table, so we created one, and combined it with the tweet arcive

4 DATA STORING

The clean data is stored to a csv file (twitter_archive_master.csv)

5 DATA ANALYSIS AND VISUALIZING

5.0.1 We made 3 insights and visualization on the clean data, regarding the ratings, the names and the breed of the dogs

5.0.2 Insights:

1. From the analysis we notice that 10,11,and 12 are the highest dog ratings
2. From the analysis we notice that while most dog names are missing('None'), some dogs still share the same names. The most common are Tuker, Oliver, Penny, cooper and charlie
3. From the analysis we notice that while most dog breeds were missing, however the most common dog breeds are golden_retriever and labrador_retriever

The visualization can be seen in the "wrangle_act.ipynb" file

```
In [ ]:
```