wrangle_report

June 18, 2022

0.1 Reporting: wrangle_report

After gathering the three datasets and converting them into dataframes I proceeded to assess them.

0.1.1 Assessment

First I used commands as .head() or .sample() to have a look at each column, the data in them, the structure of the datasets. Then I tried assessing the datasets programmatically, by either using methods as .info(), .value_counts() or others that could help me deepen my understanding of the issues.

I divided the issues in two types: a) Tidiness issues b) Quality issues

0.1.2 Tidiness issues

First, I tried to fix two tidiness issues: 1. Dog stages being in four columns instead of one

```
Solution -> To fix the first issue I replaced all the 'None' strings by whitespaces and then cre
```

2. Creating a single dataframe from the three datasets.

   Solution -> In order to merge the three dataframes I used the pd.merge method. I chose to do the merge from the images file side in order to keep only the tweets with images on them.

0.1.3 Quality issues

I identified 9 Quality issues: 1. Some rows were retweets or answers to other tweets instead of original content tweets.

```
Solution -> I filtered out the rows were the columns 'retweet_status_id' and 'in_reply_to_status
```

2. Some columns' ('timestamp', 'dog_stage' or 'img_num' among others) data type is not correct.

   Solution -> I used astype method to correct the datatype of the concerned columns at once.

3. There were wrong values in 'name' column, indicating that extraction from the tweets' text hadn't been done correctly.

   Solution -> I extracted the names all over again. I ran the extraction and compared the result several times, I tweak the extraction 'regex' until I found the expression that got most of the names correctly.

4. 'Source' column didn't add value to the dataframe so I decided to drop it.

5. I realized in the assessment stage that when none of the image predictions corresponded to a dog the image was of something else.

   Solution ->In order to have a clean dataframe that contain only original tweets about dogs, with images of dogs, I filtered out all the rows where none of the three image predictions corresponded to a dog.

6. I identified 66 duplicated images. Nevertheless, they were filtered out in previous cleaning actions.

7. 'Rating_numerator' and 'rating_denominator' contained many wrong values.

   Solution -> I extracted the ratings from the text of the tweets using regex. I run the code and compared the results to the previous existing columns as well as to the text of the tweets themselves to make sure I did the extraction as accurately as possible.

8. Some of the data in columns of the dataframe based in tweets extracted with tweepy had an erroneous format: 'like_count', 'quote_count', 'reply_count', 'retweet_count' and 'tweet_id'.

   Solution -> I changed the data type of the columns with astype method.

9. There were 29 rows with only 'tweet_id' values (NaN in the rest of the columns). This issue was solved in previous steps