wrangle_report

August 29, 2022

0.1 Reporting: wragle_report

The wrangling process involved three setps. The first being gathering three datasets, assessing them and then finally, the cleaning step.

0.2 GATHERING

Three datasets were gathered in this step; Twitter Archive, Images and Additional Data Datasets. The Twitter Archive dataset was gathered using pandas read_csv function on a .csv file. The second dataset(images) was gathered programatically using the requests library and the third dataset was gathered from twitter's API using the tweepy library.

0.3 ASSESSING

Assessment was done by two ways; programatically and visually. The following quality and tidiess issues were some of the issues detected.

Quality Issues

1. Chaning all 'None' values to 'NaN' (Twitter Archive Dataset)
2. Timestamp has a data type object instead of a datetime format (Twitter Archive Dataset)
3. Dropping columns with a lot of null values; retweeted_status_id, retweeted_status-user-id, retweeted_status_timestamp, in_reply_to_status_id, in_reply_to_user_id (Twitter Archive Dataset)
4. Removing ratings that are not for dogs; ratings, text (Twitter Archive Dataset)
5. Removing ratings with huge numerators and denominators (Twitter Archive Dataset)
6. Deleting anchor tags and extracting the source
   (Twitter Archive Dataset)
7. Dropping columns that are not important (Twitter Archive Dataset)
8. Most of the animals with 'false' in the p2_dog column were not dogs (Images Dataset)

Tidiness Issues

1. Three datasets are a lot, they need to be merged (Twitter Archive, Images and Additional Data Datasets)

2. Same variable in four columns (Twitter Archive Dataset; doggo, puppo, pupper and floofer columns)

0.4 CLEANING

For the cleaning process, quality issues were addressed first and tidiness issues were addressed after. For the data quality issues, None values were changed to NaN for easy understanding. Datatype for timestamp was changed from object to date time, rows that did not relate to dogs were dropped and those with null values were dropped too. Ratings with huge numerators and denominators were deleted; specifically numerators with figures higher than 50 and denominators that were other figures aside 10. Anchor tags were deleted in order to extract the source of the ratings. For the tidiness issues, the four dog columns(doggo, pupper, puppo and floofer) were merged into one column which was assisgned the name 'four_columns'. Lastly, all the three datasets (Twitter Archive Clean, Images Clean and Additional Data Clean) were combined using the merge function.It was then saved under the name 'Twitter_Archive_Master'

```
In [ ]:
```