wrangle_report

September 23, 2022

0.1 Reporting: wragle_report pdf

0.2 Gathering

The data for this project was gathered from three different sources. The first dataframe was loaded from the downloaded file 'twitter-archive-enhanced.csv'. The second dataframe was programmatically requested from 'https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv'. The third dataframe was gathered from the .json data accessed from the Twitter API through Tweepy.

0.3 Assessing

Assessed the dataframes visually and programmatically and found lots of Quality and Tidiness issues but did minimum of ten(10) quality and four(4) tidiness issues for cleaning as shownn below.

0.3.1 Quality issues(Completeness, accuracy, consistency, validity)

0.3.2 archive

1.Some name column entries are not names such as 'a' which should be corrected, and change lowercase names to None as they are wrong.

2.Remove the string starting 'https' in text column.

3.retweeted_status_timestamp, timestamp should be datetime instead of string

4.Create 3 columns day,month ,year and change their datatype

5.Drop timestamp column

6.Remove retweets by deleting rows with non-null values in retweeted_status_id column

0.3.3 image_preds

7.The column names such as p1,p2 are not descriptive, lets make it understandable

8.The dog breed name values in the p1, p2, and p3 columns are not consistent and capitalize first letters of dog predictions p1, p2, p3

9.Remove duplicate jpg_url entries as not to affect our analysis

10.False predictions, predictions contain other than dog animals.

0.3.4 json_tweet

11.Delete columns that won't be used for analysis

12.Change favorite_count datatype to be integer

13.Convert data type of tweet_id in all tables to object string data type for merging

14.Drop unusual columns we will not use in analysis

0.4 Cleaning

After assessing three dataframes i make a copy of the original data before cleaning. I named it clean_arch, clean_image, and clean_json. So any changes i make won't affect original copies.

0.4.1 Quality issues

The name column had one hundred and nine(109) values that were not names and these were changed to 'None' like all others that did not have a name, and all lowercases were capitalized. Removed the string starting 'https' in text column and created a column called correcttext were stored the cleaned text.All id columns were changed to the object dtype to reflect their non-numeric nature. Removed the rows that have null retweet status id so as to keep original tweets only'

The names of dog breeds in the 'p1', 'p2', and 'p3' columns were cleaned to make them more readable and understandable by changing the underscores to spaces, changing them to be prediction_1, prediction_2, prediction_3 and consistently capitalizing the words with capitalize casing. Dropped each row with false predictions and keep rows with entries that have 1st, 2nd, & 3rd_dog values as True only.

0.4.2 Tidiness issues

Some outliers were removed because they had multiple values in the development stage categories when it appeared that each should only have one value. These development stage columns were then combined into a single 'types_of_dog' column and the values from the four stages were recorded here. These four columns, along with four other columns were dropped from the 'clean_arch' because of their redundancy.

These 3 cleaned copies of the dataframes were merged based on the 'tweet_id' columns into the 'master_archive' dataframe before any other analysis and visualization was performed in order to avoid duplicating any fixes.

Some minor wrangling was required for the analysis portion of the project as well. A column named 'division' was created to reflect the ratio between the 'rating_numerator' and 'rating_denominator' values. Lastly, the 'master_archive' dataframe was stored in the file 'twitter_archive_master.csv'.

0.4.3 Conclusion

The largest issue was that the numerator and denominator values of the ratings were incorrectly pulled in several cases. It appears that they were algorithmically pulled from the 'text' values, but that this algorithm did not account for the possibility for multiple ratings or fractions in the text field. Additionally, it did not account for the possibility of decimal places, and float values existing as a part of the rating. Each row was iterated over and the 'rating_numerator' and 'rating_denominator' values were recalculated and stored with these points in mind. All rows with text values that contained multiple forward slashes (′/′) were checked manually to ensure that the rating values were pulled correctly.

```
In [ ]:
```