wrangle_report

January 1, 2023

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.

Reported by Mahshid Kalantari

1. Data Gathering

WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. The account was started in 2015. The project aims to gather API and tweet data, to create an analysis of the tweets and the predicted dog's breed. It includes three dataframes, first one is the twitter-archive-enhanced.csv file which was downloaded through the course resources of Udacity.

The second dataframe is the image_predictions dataframe which contained mainly the top 3 predictions for the corresponding dog. This dataframe was obtained with the help of requests library to download the TSV of the URL that was specified in the function. To read the TSV we needed to specify the "sep" parameter to be .⁀

The third dataframe was obtained with the Twitter API tweepy by obtaining the authentication with the API tokens, using the tweepy package, and storing it as text_json.txt for. The dataframe contained the retweeted counts and the favorite counts.

2. Data Assessing

This step was performed visually and programmatically for three dataframes. I assessed the piece of gathered data displayed in the Jupyter Notebook for visual assessment purposes. Pandas' functions methods are used to assess the data programmatically assessing.

1) Checking the Datatypes with info()

2) The duplicated rows were assessed with duplicated()
3) With isnull().sum() the null values in some columns were found.
4) With head() tail(), and sample() I found the information about the columns.

3. Data Cleaning

There were 9 quality issues I fixed during the data cleaning process:

1. There were a lot of null data in columns in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp. We did not need these columns and I could drop them.

2. There were some rows without an image where the “jpg_url” column was not null. Therefore, I dropped them.

3. The images which did not display dogs were the ones without a dog breed identification in the df_api and were dropped from the data.

4. The type of tweet_id was an integer. It could be changed to str type.

5. The type of timestamp was not datetime. I changed the type of timestamp to datetime type.

6. There were some rows containing two or more rating numbers. I removed these rows to have clean data.

7. There were some null values in two columns. I removed the null values from "reteet_count" and "favorite_count".

8. I wanted to be sure I had only tweets beyond August 1st, 2017.

9. Replacing the wrong names of the name column to nan.

There were 5 Tidiness issues I fixed during the data cleaning process:

1. For merging df.json to the other data I needed to change the name of id to tweet_id.

2. For better analysis we needed to merge the tables. I merged the df data to df.json by tweet_id column and merged the df.prediction table to the merged table.

3. For the columns with "None" replaced "None" with "" in each column and add the four string columns together to create one column called dog_type and remove three columns.

4. For selecting the breed with the highest confidence, drop the p2 and p3, and p1 to make a column as breed.

5. Splitting timestamp column into two columns for date and time and dropping timestamp.

After these steps, I saved the file as a CSV file, and the process of Analysing and Visualizing Data was performed.