wrangle_report

December 25, 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.

Next are shown the quality and tidiness issues that were observed in the asses process and solved in the cleaning process and for the different files:

`twitter_archive_enhanced.csv`: This file has the most quality and tidiness issues, some of the are describe below. 1. Many dogs named 'a', solved as extracting some of the name from some of them from the 'text' column and for the others were replaced to 'None'. 2. Eliminate the retweeted rows and replies. By eliminating the rows that have different values than 'NaN' in the rows 'in_reply_to_status_id' and 'retweeted_status_id' from the file. 3. Timestamp of the tweet was in text format, solved quickly by changing the data type to datetime. 4. After all this, the file has columns with only 'NaN' values, all of which were eliminated. 5. Merged the dog type 'doggo', 'floofer', 'pupper' and 'puppo' into one column. 6. Create a rating column as float type with the 'rating_numerator' and 'rating_denominator'.

`Image predictions` file: This file has only lesser quality issues. 1. Rows with inconsistent dog breed. This was solved by eliminating the rows with no possible consistent dog breed and also made a column with the most probable dog breed for the image. 2. Mixed lower and upper cases in the names of the dog breeds in the image predictor file, solved by lowering the cases of all dog breeds.

`Tweetpy api`: The file obtained was analyzed to obtained the precise data with lesser quality issues. 1. Eliminate the retweeted rows by eliminating the tweets marked as retweeted in the text obtained by the api.

Finally, the files were merged into a single table or dataframe with the 'inner' method to have all the information from the different tables joined in this one. Also, the columns with reiterative information, like jpg_url and the different probable columns were droped.

The final file was saved as `twitter_archive_master.csv`.