wrangle_report

June 25, 2022

Robert Ombati
project 2
June, 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.

In this report I outline the wrangling efforts to assemble and clean the data required for analysis of the WeRateDogs Twitter Archive

DATA GATHERING

I gathered data from 3 sources, stored in separate files:

1. WeRateDogs Twitter Enhanced archive, manually downloaded from the Udacity servers.
2. The image predictions file, programmatically downloaded from the Udacity servers.
3. The entire set of each tweets' JSON data, downloaded by querying the Twitter API using the Tweepy library. The favourite_count and retweet_count were extracted programmatically from this file.

I loaded the 3 raw data files into separate tables: df_tae, df_prediction and df_tweet.

Assessment & Cleaning

I began the assessment by viewing the information on the archive table first, identifying several quality and tidiness issues.

All rows containing non-null values in the retweeted_status_id , retweeted_status_user_id and retweeted_status_timestamp , and also in the in_reply_to_status_id and in_reply_to_user_id columns were dropped, as per the requirements. These columns were then also dropped. The timestamp column was converted to datetime data type.

The 4 dog stage columns were melted into the stage column; tweets without stages were set to 'none'. Several had 2 stages set, so I kept only the one with the lower overall count.

The html strings in the source column were replaced with the display portion of itself.

The rating_numerator and rating_denominator columns were checked for value ranges; I decided to keep only tweets with single ratings.

Tweets with large numerators were dropped, as the text didn't contain a valid rating (# out of 10). After the ratings were fixed, I dropped the rating_denominator column (it contained only '10's) and renamed the rating_numerator column to rating .

The odd words in the name column were replaced with ‘none’.

Tweets with missing values in expanded_urls , (not retweets or replies) were actually missing the urls from the text itself. These tweets were dropped, and then the column itself.

The df_prediction table itself was not cleaned. There were many tweets with no dog breed predicted, these were left as is. The best prediction for breed and associated confidence level were extracted and merged into the df_tae table.

The df_tweet table itself was not cleaned. The retweet_count and favorite_count columns were merged into the df_tae table, and the data type reset to int. One tweet was missing both counts so was dropped. The remaining cleaned columns in the df_tae table were reordered, then the table was saved to the new “twitter_archive_master.csv” file. The df_prediction and df_tweet tables had not been cleaned, so were not saved

```
In [ ]:
```