wrangle_report-Copy1

June 8, 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.

1 DATA GATHERING

I gathered data from 3 sources and stored it in three seperate dataframe.

1. WeRateDogs Enhanced twitter archive, manually downloaded from the udacity servers.
2. The Image predictions file, programmatically downloaded from udacity servers.
3. The JSON_DATA file downoladed programmatically because twitter didn't grant me a developer's account on time.

2 ASSESSMENT AND CLEANING

1. I began the assessment by looking at the archive dataset first. I was able to see several quality and tidiness issues in the dataset.

The timestamp column was converted into datetime data type. All rows containg 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. This was required.

The invalid names in the name column were replaced with 'None'

Tweets with missing values in the expanded_url column were dropped.

The best predictions for breed and the associated confidence level were extracted and merged in the archive dataframe.

The retweet counts and favourite count columns from the Json_data were merged into the archive file.