wrangle_report

June 24, 2022

0.1 WeRateDogs Twitter Archive : wrangle_report

In this report, the processes involved in the wrangling and cleaning the data for analyses of the WeRateDogs Twitter Archive will be outlined

0.1.1 Data Gathering

Each of the three datasets was gathered from distinct Sources

1. WeRateDogs Twitter Archive Enhanced, a csv file which was downloaded directly from the Udacity server
2. Image predictions, a Tsv file which is present in a neural network hosted on Udacity's servers and was downloaded programmatically using the Requests library from udacity
3. Using the tweet IDs in the WeRateDogs Twitter archive, Twitter API was queried for each tweet's JSON data using Python's Tweepy library.

Each data set was loaded into separate dataframe for assessment and cleaning.

0.1.2 Assessment

Each of the different datasets was assessed both visually and programmatically using python pandas library. Pandas library functions used for the programmatic assessement includes; info, describe, sample, value_counts, duplicate, head, etc. Several quality and tidiness issues were noted during the assessment which was further cleaned.

0.1.3 Cleaning

All rows containing non-null values were dropped as well as columns containing informations on retweets and reply as only original tweets data was the requirement for this analyses.

The timestamp and tweet_id column in the twitter archive enhanced was converted to datatime and string respectively.

Both the rating_denominator and rating_numerator were cleaned to meet the standard rating of 10 and rating between 10 - 15 respectively. The rating denominator was dropped to achieve one rating column for the analyses.

Tweets with missing values in expanded_urls was dropped alongside the expanded_urls column.

Odd words found in the name column were replaced with 'none'.

The retweet_count and favorite_count columns in the json file generated from twitter API was extracted and merged with the twitter Archive. Dog_breed and Confidence was extracted from image predictions and merged to Twitter Archive Enhanced file.

The gathered, assessed, and cleaned master dataset was saved to a CSV file named "twitter_archive_master.csv" which was further analysed for insights into the dataset.