wrangle_report

June 27, 2022

1 Project Report

- WeRateDogs is a Twitter account that rates people's dogs with a humorous content about the dog
- This report documents the steps taken to gather, clean and store the data.

1.1 Gathering

- The project includes three datasets, the first dataset was a csv ,`twitter-archive-enhanced.csv`, provided by Udacity that required manual downloading, the second dataset was a tab separated value text hosted on Udacty;s servers and was downloaded programmatically using the `Requests` library. The final was additional data from the twitter API, using the `tweet_ids` in the first dataset I queried the API for each tweets JSON data using the python's `Tweepy` library and stored each entire set of json data in a file called `tweet_json.txt` file

1.2 Accessing the data

- The data was programmatically and visually accessed and the following issues were identified:

1.2.1 Quality issues

1. Retweets and replies were identified and dropped
2. Ratings were not properly extracted and even after extraction some had invaluid values
3. The `rating_denominator` and `expanded_url` were dropped as they will not impact our analysis. The `rating_nnumerator` was renamed to `rating`
4. The `timestamp` column is in the wrong data type and was converted to datetime
5. The source column had text in the anchor tag and we needed to extract only the text
6. Some rows contain non dog ratings and were dropped
7. Invalid names were renamed to "None"
8. Outliers in the third dataset were dropped

1.2.2 Tidiness

1. The columns doggo, floofer, pupper an puppo will be combined into one
2. The columns p1, p1_conf, p1_dog, p2, p2_conf, p2_dog, p3, p3_conf, p3_dog will be combined into two columns; breed and confidence 3 The three datasets will be combined as a one-combinied dataset

1.3 Cleaning

- The datatypes of the columns `img_num`, `retweet_count` and `likes` in the combined dataset were converted to `int64`

1.4 Storing

- The data was stored in a sfile called `twitter_archive_master.csv`