Wrangle Report: WeRateDogs Twitter Archive

May 15, 2022

The wrangling process was done in the following stages: Data gathering, Data assessment and Data cleaning.

0.1.1 Data Gathering

The data used for this project was gathered from 3 different sources:

1. `twitter-archive-enhanced.csv` was provided by Udacity

2. `image-predictions.tsv` was downloaded from the WeRateDogs archive using requests library.
3. `tweet-json.txt` was downloaded from Twitter API using tweepy.

The files were loaded into 3 dataframes: `archive`, `image_pred` and `tweet_count`

0.1.2 Data Assessment

The data was assessed both visually and programmatically. Some quality and tidiness issues were noted.

Quality Issues

`archive table`

• Rows contain unoriginal tweets in `retweeted_status_id`
• Rows contain unoriginal tweets in `in_reply_status_id`
• Erroneous datatype for `timestamp`
• Improper dog names in `name` column
• Erroneous values for `rating_denominator` column
• Erroneous values for `rating_numerator` column
• Nan or null values in columns
• Missing images in some of the tweets in `expanded_url` column
• Non-descriptive column headers for `image predictions`
• HTML string in `source` column should be trimmed down.

Tidiness Issues

• Columns `doggo, floofer, pupper, puppo` should be in a single column

• The 3 datasets should be merged into a single dataframe using `tweet_id`

0.1.3 Data Cleaning

The dataframes were cleaned with these steps:

• The different columns `doggo, floofer, puppo` and `pupper` for dog stages were melted into one column. Rows with multiple `dog_stages` were split to have 2 values. The value *None* was replaced with an empty space, missing values were replaced with null. The individual dog stages columns were dropped.

• `expanded_urls` column had 59 missing rows, which means that there are 59 rows without images. These rows were dropped.

• Rows with non-null values for `retweeted_status_id` and `in_reply_to_status_id` were also dropped as they imply that the tweets are not original tweets, we only need the data for original tweets.

• Value_counts() done for `rating_denominator` column showed some erroneous values. These values were later set to 10, the rows were not dropped.

• Some `rating_numerator` values were decimals, other values were too large. These were set to be lesser than 14, the rows were not dropped.

• The datatype for `timestamp` was shown to be a string, this was converted to datetime.

• There were values incorrectly placed as dog names in `name` column. These words were assessed by using the str.contains() function. The values were then replaced with null.

• Column headers for `image_pred` were renamed to be descriptive.

• Source column values were replaced with a more explicit string.

• `retweeted_status_id`, `retweeted_status_user_id`, `retweeted_status_timestamp, in_reply_to_status_id` and `in_reply_to_user_id` columns were dropped as they contained too many null values and were therefore insignificant.

• No cleaning process was done in `tweet_clean` dataframe. This dataset was later merged with `archive_clean` and `image_pred_clean` to create another file `twitter_archive_master.csv`