wrangle_report

June 21, 2022

WeRateDogs Twitter Archive

Christian Chimezie

Data Analytics Nanodegree 2nd Project
June 2022

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

Data wrangling procedure has been completed on the datasets obtained from "WeRateDogs".

There are three steps:

1. Data gathering
2. Data assessment
3. Data cleaning.

1. Data Gathering

- I imported the required libraries.
- I gathered data from 3 sources, stored in separate files:
    - WeRateDogs Twitter Enhanced archive, manually downloaded from the Udacity servers.
    - The image predictions file, programmatically downloaded from the Udacity servers.
    - The entire set of each tweets' JSON data, downloaded by querying the Twitter API using the Tweepy library. The favourite_count and retweet_count was extracted programmatically from this file.
- I loaded the 3 raw data files into separate tables: twitter_archive, image_predictions and tweets_df.

2. Data Assessment

I began the assessment by viewing the information visually and programmatically, identifying several quality and tidiness issues.

2.1 Quality issues:

twitter_archive table:

- Keep original ratings (no retweets).
- Drop columns not needed for our analysis (in_reply_to_status_id, in_reply_to_user_id, expanded_urls).
- Incorrect tweet_id datatype.
- +0000 in timestamp is redundant.
- Incorrect timestamp datatype.
- Twitter source unreadable.
- URL in the end of 'text'

- Dogs without names, but given names of “a” or “an” instead of “None.”
- The rating numerator and denominator has some incorrect values.
- Erroneous datatypes of ‘rating numerator’ and ‘rating denominator’.
- Consider rating_denominator to be 10 and drop column.
- rating_numerator maximum possible value is 15.

image_predictions table:

- Incorrect tweet_id datatype.
- Drop column not needed for our analysis (img_num).

tweets_df table:

- Incorrect tweet_id datatype.

2.2 Tidiness issues:

- There are 3 three dataframes.
- There are 4 columns for dog stages (doggo, floofer, pupper, puppo) in twitter_archive table.
- There are 3 columns for breed predictions (p1, p2, p3) in image_predictions table.

3. Data Cleaning

- Made a copy of the original data before cleaning
- Removed retweets.There are 181 retweets (retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp).
- Drop columns not needed for analysis with pandas drop function.
- Correct datatypes with pandas astype function.
- The timestamp column was converted to datetime data type.
- Remove tags from source by using string extract method.
- Melt doggo, floofer, pupper, and puppo into one column (dog_stage).
- The best prediction for breed and associated confidence level were extracted using conditional columns with Numpy select ().
- The cleaned dataframes were merged on tweet_id.
- The merged data was saved to the new “twitter_archive_master.csv” file.