wrangle_report

September 7, 2022

1 Report: Wrangle and Analyze WeRateDogs' Twitter archive

WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog and ratings almost always greater than 10; 11/10, 12/10, 13/10, etc. WeRateDogs has over 4 million followers and has received international media coverage.

1.1 Project Steps

1.1.1 The following steps were taken in this analysis

1. Data was gathered from different sources; direct download, using Python's Request library and scraping Twitter's API. The different data files were then loaded using the Pandas library. The gathered files are:

- The WeRateDogs Twitter archive (manual download).
- The tweet image predictions (Request library).
- Each tweet's entire set of JSON data (scraping Twitter's API).

2. The loaded data was then assessed visually and programatically using different pandas methods.

3. Quality issues in the data were:

   - Archived tweets

   1. Retweets and replies to original tweets are included. Only original tweets are needed.
   2. Extraneous columns: in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, and retweeted_status_timestamp.
   3. Erroneous datatypes in the tweet_id (also needs to be changed in the other two tables) and timestamp columns.
   4. Invalid values in the rating_numerator and rating_denominator columns. The numerator is usually between 10 and 15 while the denominator rating is always 10.
   5. Inaccurate dog names. Examples are 'a', 'an', 'by', 'O'... This is a result of incorrect extraction of dog names from the text column.

   - Image predictions

   1. Some rows contain predictions that are not dog types.

    2. Non-descriptive column headers: p1, p1_conf, p1_dog.
    3. Extraneous columns: jpg_url, p2, p2_conf, p2_dog, p3, p3_conf, p3_dog.
    4. Tweet IDs should be stored as strings/objects.

    - Scraped JSON data

    1. Tweet IDs should be stored as strings/objects.

4. Tidiness issues were

    1. The last four column headers in the archived_tweets table qualify as variable values not variable names and so needs consolidation into a column.
    2. The retweet_count and favorite_count columns of the tweet_json table should be included in the archived_tweets table.
    3. Some columns in the image_predictions table should be part of the archived_tweets table

5. Data cleaning was carried out programatically after enumerating the different quality and tidiness issues in the data.

6. The cleaned data was stored in a master archive appropriately named.

7. Finally, a brief analysis and data visualization was done to glean a few insights into the data.