wrangle_report

January 28, 2023

1 Data wragling report of tweet archive of Twitter user WeRateDogs (@dog_rates).

1.1 Introduction

The dataset that we are wrangling (and analyzing and visualizing) is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. Why? Because "they're good dogs Brent." WeRateDogs has over 4 million followers and has received international media coverage.

1.2 Data Gathering

We are using three datasets with different file formats and gathering techniques.

1. The WeRateDogs Twitter archive (twitter_archive_enhanced.csv) is downloaded from udacity resources and uploaded directly to the workspace environment
2. The tweet image predictions (image_predictions.tsv) is downloaded programatically using the Requests library and the given URL
3. Additional data from the Twitter API (tweet_json.txt) is downloaded either by using the Tweepy library or reading the already given file from udacity

1.3 Data Assessing

All the three datasets were assessed using:

1.3.1 1. Visual assessment:

Where we can explore the entire dataset

1.3.2 2. Programmatic assessment:

Some pandas functions were used to help assessing the datasets: [.head(), .tail(), .sample(5), .info(), .describe(), .duplicated(), .value_counts(), .sort_values(), .isnull()]

After assessing the datasets, some quality and tidiness issues were found:

1.3.3 Quality issues

Completeness, validity, accuracy, consistency (content issues)

- The original tweets only are needed, no retweets
- Many un-needed columns
- ['rating_numerator', 'rating_denominator'] columns type is int
- duplicated jpg_url
- naming issues
- timestamp type is object instead of being datetime
- the extracted numerators did not contain the decimals values
- missing [rating] column that devids the numerator by the denominator
- the tweet_id type is int instead of being object or string

1.3.4 Tidiness issues

rows, columns, and tables (structural issues)

- Many datatypes (doggo, floofer, pupper and puppo columns)
- day, month, year are in one timestamp column
- many dog type and confidence level columns
- twitter_archive_clean, image_predictions_clean, and twitter_api_clean must be combined

1.4 Data Cleaning

Here, all the discovered issues were solved.

1.4.1 Quality Issues

- Delete retweets by filtering the null values of retweeted_status_user_id to keep the original tweets only.
- Delete the unneeded columns
- Convert ['rating_numerator', 'rating_denominator'] columns type from int to float
- Delete the duplicated jpg_url
- Correct naming issues
- Convert timestamp to datetime
- Correct decimal values in the numerator of the rating
- Create a new [rating] column by deviding the numerator by the denominator
- convert tweet_id type to string

1.4.2 Tidiness Issues

- Melt [doggo, floofer, pupper, puppo] columns to dogs and dogs_stage column
- Extract [year, month and day] values to new columns
- Keep only the dog type with the max confidence level
- merge twitter_archive_clean and image_predictions_clean
- merge df_twitter1 and twitter_api_clean

1.4.3 Storing Data

Save the gathered, assessed, and cleaned master dataset to a CSV file named "twitter_archive_master.csv".