wrangle_report

March 30, 2021

1 Data Analyst Nanodegree

2 Project 4 - Wrangling and Analyze Data

2.1 Wrangling Report

2.1.1 Project

The project is centered around **WeRateDogs** Twitter account.
Their sual tweets are dogs photos with humourous ratings of x/10, where x is often more than 10, eg. 13/10.
The account is popular and therefore has enough data to explore and attempt learning from.

2.1.2 Input

Data used in this project came from a number of different sources: - First file was **downloaded manually** from link provided, as per instructions
`twitter-archive-enhanced.csv` - Second file was **downloaded programatically**
`image-predictions.tsv` - Third file was created with data downloaded directly from **Twitter API**
`tweet_json.txt` - Supporting file was created, listing tweets that couldn't be accessed (ie. were deleted) `tweets_not_found.txt`

Data in the files can be **joined on `tweet_id`**, which is present in each of them and doesn't pose quality issues.
For the purpose of wrangling and later analysis, the files were read into pandas **data frames**.

2.1.3 Cleaning

A few **tidiness** issues and a number of **quality** issues were found during the visual and programmatic **assessment**.
They were later dealt with in **cleaning** phase.
The work here can be **summarised** as below.

Tidiness

1. Dog **stages** were melted from four variable columns into one.
2. Data from various sources was grouped into **separate sets** and eventually separate tables in SQLite database:
   `tweets` for tweet data (excuding likes and retweets),

`dogs` for dog data (including tweet likes and retweets for ease of analysis),
`dog_stages` for mapping dog stage strings to integers,
`image_preds` for image predictions data.

Quality

1. Tweet data with no corresponding image prediction data was removed, as per project demand.
2. Retweets duplicating original tweets in the dataset were removed.
3. Duplicated image data was removed.
4. Less than 0.5% of remaining records had missing likes and retweets.
   For the purpose of analysis only they were filled with means.
5. ID columns with float values were converted to integers.
6. Timestamp columns were converted to datetime.
7. Dog stage column was converted to category first, and later made into separate table mapping strings to integers.
8. Dog names and stages were cleaned and additional data was extracted from tweets' text field.
9. Rating numerators and denominators were reviewed and cleaned.
10. Additional column for normalized rating to support analysis was created.
11. Column containing irrelevant, repeated url removed.
12. Made all string columns into lower case.

As a result of quality fixing, in the final data set used for analysis only 2 columns had null values: `stage` and `name`.
In most cases the reason is that they are not to be found in those tweets' text at all.

2.1.4 Output

In the end of wrangling, the resulting data was written to **csv files**: - `twitter_archive_master.csv` - Tweet data - `dogs_master.csv` - Dawgz data - `image_pred_master.csv` - Image predictions data

as well as into an **SQLite database**, in tables listed in Cleaning / Tidiness section above.