wrangle_report

November 30, 2020

1 WeRateDogs Wrangling Report

This document sets out the steps taken to gather, assess, and clean the data relating to the WeRateDogs Twitter account in preparation for analysis.
Gathering First we gathered the date from three different sources:

1) The WeRateDogs Twitter archive, which was a file on hand,

2) The classification of the dogs in the photos. For each tweet there is a photo of the dog being rated and this was run by another team through a machine learning algorithm to try to work out the particulary breed. This was downloaded programmatically from a url.

3) Extra data from the Twitter API. Ideally this would have been done directly but I wasn't able to get access so used the data provided.

1.1 Assessing

The data was loaded into dataframes using the pandas library. Various techniques were then used on the three data sources to detect any missing or incorrect data and to ensure the data is tidy.

The key findings were: - ratings needed to be cleaned up as the technique for pulling them from the tweets wasn't perfect and seemed to rely on finding a "/" - ratings were sometimes for multiple dogs, e.g. 110/100 for ten dogs in a picture - the breed data was badly organised and contained many images where something other than a dog had been identified

The full list of issues is below.

1.1.1 Twitter Archive

Data Quality

- Contains Retweets `in_reply_to_status_id` != `Nan`
- Contains columns with data about retweets `retweeted_status_id`, `retweeted_status_user_id`, and `retweeted_status_timestamp`
- The entires in the `source` columns are obscured inside an `<a>` tag
- The end of the entry in the `expanded_urls` column is often the tweet id
- Ratings are not always pulled out of the tweets correctly and where they are they sometimes refer to multiple dogs so are not comparable
- Some of the names are not correctly extracted from the tweet (relying on "this is"?)
- `text` column contains the tweet url which is already available in expanded form in `expanded_url`

- `tweet_id` should be a `str` (`object`)
- There are 59 `expanded_url` entries missing
- Source, once stripped out, should be a `category` and `NaN` where not available
- `timestamp` should be dtype `datatime`

Tidiness

- The last four columns are `doggo`, `floofer`, `pupper`, and `puppo` are a single category spread over multiple columns

1.1.2 Image Labelling

Data Quality

- `tweet_id` should be a `str` (`object`)
- The breeds themselves are inconsistently capitalised and have underscores

Tidiness

- Breed classification is spread out over several columns and the data is not tidy

1.1.3 Tweet Meta Data

Data Quality

- `id` should be a `str` (`object`) and is the key column that corresponds to `tweet_id`

Tidiness

- `favorite_count` and `retweet_count` should be retained but the others dropped
- For the sake of tidiness it also seems sensible to amalgamate all this data into a single table since each row will be an observation about a tweet.

1.2 Cleaning

Various decisions were made in the cleaning that could have a bearing on the analysis and a summary is below: - Retweets and replies were dropped since there are not simple ratings of a picture of a dog - Ratings were corrected where something other than a rating was picked up - Ratings for multiple dogs were normailsed to that of a single dog by dividing by the number of dogs - Dog breed identification was chosen from the results given as the most likely option

With the above steps completed, it was then time (finally!) to analyse the data.