wrangle_report

September 14, 2020

The data was loaded into pandas dataframes. I chose to use the tweet data provided by Udacity because I had trouble applying for a Twitter developer account. Visual and programmatic assessment of the three tables revealed the following issues:

Quality Issues

`tweets` **table**

- Source column contained unnecessary html tags
- Some dog's names were incorrectly extracted
- text column values contained links
- timestamp data type was incorrect
- timestamp is in UTC(GMT) time zone.
- table contained tweets retweeted by WeRateDogs.
- contains data on tweets retweeted by WeRateDogs ie not original tweets

`predictions` **table**

- predicted dog breed names contained underscores
- predicted dog breed names are not consistently formatted

`tweet_data` **table**

- unwanted columns present

Tidiness Issues

- dog stage divided into multiple columns instead of 1 in the `tweets` table
- columns in `tweet_data` and `predictions` should be part of the `tweets` table
- too many possible dog breeds in `predictions`

0.0.1 Summary of cleaning steps

Quality

`tweets` **table**

- content extracted from html in source column using BeautifulSoup, then html deleted.

- incorrect names found by parsing through name column to find all-lowercase words. tweets with these names were printed to observe patterns to exploit for name extraction. 2 popular patterns were found and used to extract name by custom function. the rest of the lowercase words were replaced with the placeholder "None".

- http links in text column were removed using a regex substitution with an empty string.

- timestamp column datatype changed using pd.to_datetime method.

- timestamp timezone changed to home location of WeRateDogs creator's area of residence (US/Eastern).

- retweets were removed by deleting rows with non-NaN retweeted_status_id value.

- unwanted columns were dropped from the tweets table.

`predictions` **table**

- underscores were removed from the predicted dog breed names using the replace method.

- predicted dog breed names were changed to titlecase.

`tweet_data` **table**

- unwanted columns removed using df.drop method.

Tidiness

- dog stage extracted using str.extract to create dog_stage column. individual stage columns dropped.

- the first True dog prediction was extracted from the columns, along with the corresponding confidence and prediction rank. the original columns were then dropped.

- renames id column as tweet_id in `tweet_data` table, then merged with `tweets` and `predictions` tables into `combined` table.

the `combined` **master table was then exported to a csv file** `combined.csv`