Joe Lalgee 10/10/2020

WeRateDogs Tweets Wrangle Report

In this project we gathered, assessed and cleaned data on tweets by WeRateDogs from three sources:

- A given csv file containing an archive of tweets to be analyzed.
- A remotely hosted tsv file containing predictions of dog breed pictured per tweet.
- Data on tweet retweets and favorites obtained via Twitter's API.

Wrangling activities for this project were limited to assessing and cleaning eight quality issues and two tidiness issues.

Gathering

The csv file was read directly into a Pandas dataframe.

The remotely hosted tsv file was downloaded using the `requests` library, and then read into a Pandas dataframe via the `io` library's `StringIO` decoder.

The data on tweet retweets and favorites was accessed using the `tweepy` library. The JSON data was extracted and saved to a text file. It was then read back line by line into a list using the `json` library, and passed into a Pandas dataframe.

Assessing and cleaning

After gathering the data into the three dataframes, the following issues were discovered and cleaned:

Tidiness

- The data was split across three dataframes but they all related to one type of observational unit. The dataframes were merged.
- The `doggo`, `pupper`, and `puppo` columns all related to a single, categorical `stage` variable (`floofer` relates to fluffiness rather than stage so was regarded as a separate variable). This information was collated into a single `stage` column.

Quality

- Data for 297 tweets was missing from one or more dataframes. These tweets were removed.
- 12 tweets contained references to multiple stages i.e. `doggo`, `pupper`, `puppo`. These tweets were removed.
- The tweets weren't all original tweets - 70 were retweets and 22 were replies. These tweets were removed.
- Names were not always detected and this often indicated a tweet that didn't rate a single dog. These tweets were removed.

- Detected names were not always accurate - sometimes being picked up as `'a'`, `'an'` or `'the'`. This sometimes indicated a tweet that did not rate an individual dog. These tweets were removed.
- Rating denominators were not always accurate. Ratings were corrected where tweets were valid for the analysis, and otherwise were removed.
- Rating numerator were not always accurate due to decimals used in the tweet. Ratings were rounded to the nearest integer.
- One outlier numerator of 1776 was present. This was reduced to 15, preserving the dog's number one rank within the data.
- `timestamp` was a string rather than a datetime. It was converted.
- `stage` was a string rather than a category. It was converted.
- `floofer` was a string rather than a boolean. It was converted.
- The columns `source`, `expanded_urls`, `jpg_url` and `rating_denominator` were not needed for analysis. These columns were removed.