Wrangle_Report

11th December 2018

Gather

Data were collected from three different sources. First data was collected from the “twitter-archive-enhanced.csv” file which was in the same directory in which project notebook was located. The csv file was imported into pandas dataframe. The dataframe was named “twitter_archive”.

Second data was extracted programmatically from a URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv.

Python’s request library was used to extract data from URL. The URL was split using “/” as the separator and the last value was the file name. This file was written in the content of our request. Then this file was imported as a dataframe in pandas using tab as the separator. The dataframe was named “img_predictions”.

The third data was extracted from Twitter API using python’s tweepy library. I needed to extract the favourites and retweet counts for each tweet. This data was then saved as a JSON file using UTF-8 encoding.

The images dataframe, the JSON file and the archive data were merged into a single dataframe. A copy of this merged data was saved in CSV format.

Assess

Quality

1. Several columns have empty values, like in_reply_to_status, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp.
2. The name column has many entries which do not look like names. The most frequent entry in name column is "a", which is not a name.
3. The numerator and denominator columns have unusual values.
4. The timestamp column is an object. It has to be a datetime object.
5. There are 2075 rows in the images dataframe and 2356 rows in the archive dataframe.
6. In several columns, null values are not treated as null values.
7. Incorrect Dog Names
8. Erroneous datatypes (timestamp, source, dog stages, tweet_id, in_reply_to_status_id, in_reply_to_user_id)
9. Dataset contains retweets

Tidyness

1. Dog "stage" variable in four columns: doggo, floofer, pupper, puppo
2. The columns 'retweeted_status_id' 'retweeted_status_user_id' and 'retweeted_status_timestamp' are not useful after we get rid of retweets.
3. Join 'tweet_info' and 'img_predictions' to 'twitter_archive'

Clean

- Delete retweets and observations without ID, delete columns: 'retweeted_status_id','retweeted_status_user_id','retweeted_status_timestamp'
- Delete observations without image and merge img_predictions_clean with twitter_archive_clean
- Create dog stage variable and remove individual dog stage columns.
- Condensing dog breed predictions.
- Add tweet_info to twitter_archive table.
- Convert timestamp to datetime data type.
- convert in_reply_to_status_id,in_reply_to_user_id to string data type. Query data from API
- Set the value wrong names to 'None' and replace 'None' with np.nan
- Change the rating_numerator and rating_denominator for oberservations with wrong value
- Oberservations with tweet_id '810984652412424192'doesn't have a valid rating, so drop this row.
- Create new column rating=rating_numerator/rating_denominator. Drop rating_numerator and rating_denominator.
- Drop oberservations with extreme ratings.
- After cleaning, the data was exported to a CSV file named “twitter_archive_master.csv”