Marcus Holmgren
Data Analyst Nanodegree
13 April 2019

Project 5: Wrangle and Analyze Data

Tweets 9,960
Following 12
Followers 7.9M
Likes 141K
Moments 4

WeRateDogs™
@dog_rates

Your Only Source For Professional Dog Ratings Instagram and Facebook ➪ WeRateDogs partnerships@weratedogs.com

「 DM YOUR DOGS 」
weratedogs.com
Joined November 2015

Pinned Tweet
WeRateDogs™ @dog_rates · Mar 15
This is Gabanna. Her human Jose rushed past firefighters into his burning home to rescue her. She suffered burns on her nose and paws, but will be just fine. The family lost everything else. You can help them rebuild below. 14/10 for both (via @KUSINews)

gofundme.com/stand-by-jose-…

My wrangling efforts started off by investigating what data files existed that I should download. Two files where provided as a starting point and was accessible though a public URL. **image_predictions.tsv** and **twitter_archive_enhanced.csv.** The third file **tweet_json.txt** is supposed to be created by downloading WeRateDogs tweets.

At the top of the Jupiter Notebook wrangle_act.ipynb I have collected a handful of helper functions that are used later in the steps for gather, assess, and clean.

A total of 8 quality data issues are documented and addressed in the Jupiter notebook:

- 1 in the image_predictions.tsv
- 5 in the twitter_archive_enhanced.csv
- 2 in the tweet_json.txt

Two tidiness issues:

- 1 in the twitter_archive_enhanced.csv - 4 columns that should be combined into one.
- 1 in the tweet_json.txt

twitter_archive_enhanced.csv

The rating is a cumulative sum of the number of dogs rated in a picture. Because of this there are outliers that gives the rating_numerator and rating_denominator a right skewed. This was not addressed in the cleaning process.

The following issues was discovered in the data source columns.

Quality data issues

1. tweet_id - convert to string
2. timestamp - DataFrame stored as object, convert to date time
3. name -
   A. None as null value
   B. Not only names in columns
4. text - non dogs pictures contains phrases like “We only rate dogs”, “Pls stop sending…”, “Only send dogs”
5. None as null value
   A. doggo - None as null value
   B. floofer - None as null value
   C. pupper - None as null value
   D. puppo - None as null value

Tidiness issues

The four columns doggo, floofer, pupper, and puppo are WeRateDogs language of DoggoLingo
The column text and source is the same as full_text from the tweet_json.txt file.

image_predictions.tsv

This dataset consist of 12 columns with each row the top three classification results from analysis of dog images. Each of the three prediction contains a descriptive category name, a confidence score and a field that is true if the classification was a dog.

Quality data issues

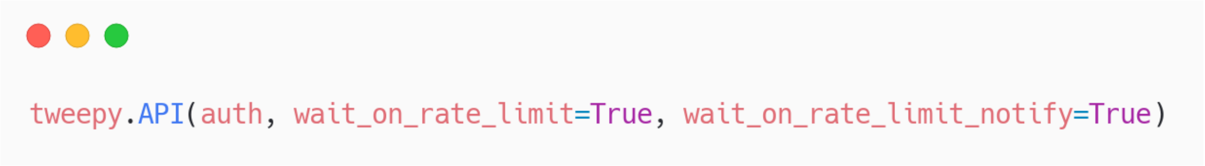
1. tweet_id - convert to string

Tidiness issues

No tidiness issues discovered in the dataset.

tweet_json.txt

To generate the dataset I needed to create a developer API key with Twitter so I could retrieve tweet messages with the Python package tweepy. Because the twitter_archive_enhanced.csv contains a tweet_id it was straight forward to query status for a specific tweet. What did take some trail and error to find the correct initialisation of the tweepy API, because the Twitter rate limit would be enforced and not allow querying and getting a response for up to 10 minutes. But after setting rate limit properties all tweets could be downloaded into the file in about 30 minutes.



```
tweepy.API(auth, wait_on_rate_limit=True, wait_on_rate_limit_notify=True)
```

The dataset consist of 32 columns and there are several columns that have missing values so they will be dropped.

Quality data issues

Id column convert to string.
full_text - non dogs pictures contains phrases like “We only rate dogs”, “Pls stop sending…”, “Only send dogs”. This is the same issues as file twitter_archive_enhanced.csv

Tidiness issues

The column full_text and source is the same as text from file twitter_archive_enhanced.csv

Several columns are JavaScript objects and contains several properties that could be split into their own columns in the Dataframe. This will not be done since I plan on dropping most of the columns for the master dataset.

twitter_archive_master.csv

All three collected Dataframe was combined int to the final master dataset. I choose to only have 16 columns with what I think is the the most relevant data for further exploration and data analysis.