Marcus Holmgren
Data Analyst Nanodegree
10 April 2019

Project 5: Wrangle and Analyze Data

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 13 quality data issues are documented and addressed in the Jupiter notebook:

- 2 in the image_predictions.tsv
- 9 in the twitter_archive_enhanced.csv
- 2 in the tweet_json.txt

Three tidiness issues:

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

twitter_archive_enhanced.csv

The following issues was discovered in the data source columns.

Quality data issues

1. timestamp - DataFrame stored as object, convert to date time
2. name -
   A. None as null value
   B. Not only names in columns
3. text - non dogs pictures contains phrases like “We only rate dogs”, “Pls stop sending…”, “Only send dogs”
4. rating_numerator - suffers from outliers, max value in data is 1776. Will not be addressed.
5. rating_denominator - suffers from outliers, max value in data is 170. Will not be addressed.
6. doggo - None as null value
7. floofer - None as null value
8. pupper - None as null value
9. puppo - None as null value

Tidiness issues

The four columns doggo, floofer, pupper, and puppo are WeRateDogs language of DoggoLingo

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. p1, p2, p3 - some predicted categories are not dogs. Will not be addressed.
2. p1_dog, p2_dog, p3_dog - some predictions are not dogs. Will not be addressed

Tidiness issues

The top three predictions from the classification model used are stored in three identical columns. This tidiness issue will be address by melting the columns into row values.

1. Prediction 1, 2, & 3:

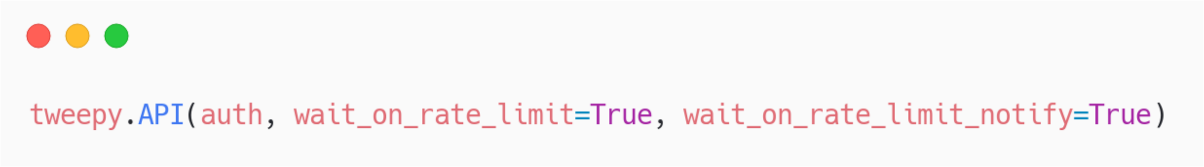
C. p1, p2, p3 - the predicted category of the image

D. p1_dog, p2_dog, p3_dog - boolean value if True if prediction classified image as dog

E. p1_conf, p2_conf, p3_conf - the confidence score of the prediction of the image

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

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

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.