DATA WRANGLING REPORT

Author: IRAKLIS KALAMAS

This is the report of the data wrangling process required for the project “Wrangle and Analyze Data” of the Data Analyst Nanodegree program by UDACITY.

For the purposes of the project I’ll use a dataset available by Twitter that rate dogs with a humorous comment about the dog (also known as WeRateDogs). WeRateDogs has over 4 million followers and has received international media coverage.

The goals of the project, as depicted by the project’s details page, are:

- Data wrangling, which consists of:
    - Gathering data from three different sources
    - Assessing data
    - Cleaning data
- Storing, analyzing, and visualizing your wrangled data
- Reporting on your data wrangling efforts and your data analyses and visualizations

Gathering data

The three sources that I use in order to gather data about the WeRateDogs project are:

1. The WeRateDogs Twitter archive. The twitter_archive_enhanced.csv file was provided to Udacity students. The archive contains basic tweet datafor all 5000+ of their tweets as they stood on August 1, 2017. This source will be gathered using the read_csv method of Pandas library.
2. The tweet image predictions file. This file was again provided to Udacity students and is the result of running a Neural Network algorithm that can classify breeds of dogs. The resulting file is a table of image predictions that corresponded to the most confident prediction about most rated dog breed. This file will be accessed using the get method of the requests library.
3. Data from Twitter API and Python's Tweepy library to gather each tweet's retweet count and favorite ("like") count at minimum, and any additional data not included in the previous datasets. I’ll query the API using the tweepy library and I’ll store its contents in a json file using the json library.

Assessing data

After the data was gathered (I end up with three datasets), I assessed their quality and tidiness issues.

First dataset: twitter-archive-enhanced.csv

- **Quality issues**
  - Completeness:
    - missing data in the following columns: in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, expanded_urls
    - tweet_id is an int (same for all datasets)
  - Validity:
    - dog names: some dogs have 'None' as a name, or 'a', or 'an'
    - his data-set includes retweets, which means there is duplicated data
  - Accuracy:
    - timestamp is an object
    - retweeted_status_timestamp is also an object
  - Consistency:
    - the source column still has the HTML tags
    - rating_denominator should be 10 always, but there are other values as well
- **Tidiness issues**
    - the last four columns (dogoo, floofer, pupper, puppo) all relate to the same variable

Second dataset: image_predictions.csv

- **Quality issues**
  - Validity:
    - p1, p2 and p3 columns have invalid data naming a dog photo as a starfish, boathouse or something else
  - Consistency:
    - P1, p2 and p3 columns have naming (underscores between a multi word dog breed) and capitalization (some breed names start with a capital letter some don't) issues
- **Tidiness issues**
    - the dataset is part of the same observational unit as the data in the 'twitter-archive-enhanced.csv', i.e. same basic information regarding dogs

Third dataset: tweet_json.txt

- **Quality issues**
  - Completeness:
    - missing data in various columns
- **Tidiness issues**
    - the dataset is part of the same observational unit as the data in the 'twitter-archive-enhanced.csv', i.e. same basic information regarding dogs

Cleaning data

For the final part I used the **Define** (identifying and documenting an issue), **Code** (implementing code for fixing the issue) and **Test** (checking if the code produces the correct results):

Quality issues addressed

Quality issues addressed

- Change tweet_id from an integer to a string
- Change the timestamp to correct datetime format
- Delete retweets (the duplicates of the originals)
- Standardize dog ratings
- Correct naming issues (wrong names and lowercase issues)
- Replace undersores with spaces in dog names
- Creating a new dog_breed column using the image prediction data
- Remove columns no longer needed: in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, and retweeted_status_timestamp

Tidiness issues addressed

- Merge the three datasets
- Create one column for the various dog stages: doggo, floofer, pupper, puppo

(The details of the coding for the above cleaning process is available in the file "wrangle_act.ipynb")

After that my final dataset is produced and I'll move to Analysis of the data. In general maybe there is the need to iterate through the whole process if the analysis results are not satisfying, but for this project I'll continue with the final dataset as it is.