Wrangle Report

Marc Campmany

March 2021

This report documents the steps followed to complete the “WeRateDogs” wrangling and analysing data project from Udacity’s “Data Analyst Nanodegree Program.

The four parts of the Data Wrangling process are:

- Gathering
- Assessing
- Cleaning Data
- Storing, Analyzing and Visualizing

1. **Gathering:** the three pieces of data in the Jupyter Notebook titled wrangle_act.ipynb are:
   1.1 The WeRateDogs Twitter archive. This file is already provided: **twitter_archive_enhanced.csv**

   1.2 The tweet image predictions, i.e., what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network. This file (**image_predictions.tsv**) is hosted on Udacity's servers and has been downloaded programmatically using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv

   1.3 Each tweet's retweet count and favorite count using the Twitter API and tweepy library (**tweet_json.txt**).

2. **Assessing:** the assessment was done first visually inspecting the dataframes and later programmatically with python’s numpy and pandas libraries.

   Quality (content)

   1. Twitter archive enhanced
      - HTML tags in source text.
      - name columns contains words that are not names (ex. 'a').
      - The missing data in "name" or type of dog columns is encoded as 'None'.
      - Incorrect datatypes (*_id, *timestamp)
      - retweet information not needed for this project
      - some rating numerator and denominator have extreme max/min values
      - some tweets contain no information

2. Image prediction file:
    - Dog breed predictions with "(_)" separator.
    - Some predictions are not dog breeds (ex. paper_towel or orange)
    - Incorrect datatypes (*_id)
    - Missing information as "tw_archive" file has 2356 rows and this one 2075. Nothing can be done as the predictions file is given and the neural net is not accessible.

Tidiness (structural)

1. Twiter archive enhanced:
    - There are four columns indicating the type of dog in the tweets: doggo, floofer, pupper, and puppo. These columns can actually be melted into a single column.

2. Image prediction file:
    - "True" dog breed predictions should be merged into the "tw_archive" to consolidate all the info together.

3. Tweepy extra info:
    - retweet_count and favorite_count should be part of the consolidated dataset.

3. **Cleaning:** define and perform the cleaning of the issues (both quality or tidiness) assessed previously:

Tidiness Issue 1 = 'There are four columns: doggo, floofer, pupper, and puppo'
Define: Melt the 4 different columns of the dog types into one single column showing the dog type per row.

Tidiness Issue 2: 'tweet_image_pred: correct predictions should be combined with tweet_data_archived'
Define: Extract the most likely "True" prediction from "tweet_image_pred" and merge them to the "tw_archive" file.

Quality Issue 1: 'Incorrect datatypes (*_id, *timestamp)'
Define: Convert _id columns into strings and timestamp columns into datetime format

Quality Issue 2: 'retweet information not needed for this project'
Define: Remove all rows with retweeted_status_id and in_reply_to_status_id then drop the columns related to retweet fields:

Quality Issue 3: 'remove tweets without image'
Define: Delete tweets where jpg_url is NaN

Quality Issue 4: 'name columns contains words that are not names (ex. 'a')'
Define: Review all dog "names" and modify the ones that are incorrect to "None".

Quality Issue 5: 'tweepy dataframe should be merged into the "tw_archive" to consolidate all the info together for completeness.'
Define: Merge retweet and likes info into the tw_archive dataframe.

Quality Issue 6: 'html tags in source text'
Define: From source column extract the content from = '> + content + <'

Quality Issue 7: 'The missing data is encoded as "None"'
Define: Replace the "None" and "Unknown" for NaNs

Quality Issue 8: 'some rating numerator and denominator have extreme max/min values'
Define: Analyze the extreme values and decide if need cleaning, deleting or modifying.

4. **Storing:** (Analyzing and visualizing will be reported in act_report.pdf)

   Final dataframe saved in:

   - 'twitter_archive_master.csv'
   - in a **sqlite database** named “WeRateDogs.db”.