Wrangle Report

By: Sofyan Sahrom
Date: 8th February 2018

Project Mission: Show proficiency in data wrangling and analyses through 3 tasks: 1) Gathering Data, 2) Assessing Data and 3) Cleaning Data on a Twitter archive (WeRateDogs).

1) Gathering Data

Data that was to be wrangled and analysed was available from 3 different sources. These data were either given (twitter-archive-enhanced.csv) or requested from different sources. These datas were then loaded as 3 different dataframes

1) Given (Existing file): - A csv file that was given directly. This csv file titled (twitter-archive-enhanced.csv) is the main @dog _rates twitter archive. The data it contained are (for, e.g.) tweet id and the tweet texts.
    a. This file is loaded as df_archive.
2) Request from an external source: - a tsv file titled (image_predictions.tsv) that contained the dog breed predictions from a neural network. This data was available from an external source (hosted on Udacity servers - here). To gather this data, request methods had to be used to gather this data.
    a. This file is loaded as df_image
3) Request (Query) from an external source + written to json format: - tweets were extracted through the utilisation of twitter's API and then written to a text file (tweet_json.txt) in JavaScript Object Notation (json) format.
    a. This data is loaded as df_tweet.
    b. A second attempt was made for failed queries, by themselves, the successful ones were then added to the dataframe.

All three data were then loaded as 3 different data frames.
For the query request from twitter, there were some failed queries. A 2nd attempt was made with the failed queries only. 9 of these queries were successful and appended to the original json file. This was then uploaded as df_tweet.

2) Assessing the Data

Assessing at this level is at a macro level. The main objectives is to look at the quality and tidiness of the data collected and understand the data. To do so, several pandas functions was used. These included, .describe(), .info(), .sample(), .value_counts().

Data was assessed at 2 levels. The first level is for quality (content) issues (C) and tidiness (structural) issues (T). We identified at least 8 possible issues at this levels. Issues are given tags (eg. C1, T2, T3 etc.) for two purposes. First is to help monitor the issues and second is to help organise the chronological order of the C&C (Cleaning and Correction).
At the next level, data issues (D) are identified of which at least 3 have been identified. Some of these issues overlap with tidiness issues, therefore these will addressed at the same time.

Quality - Content Issues

C1) 8x values for (666020888022790149). Considering the context, this means that it has been duplicated 8 times

C2) Missing values

- Image Dataset (2075 instead of 2356) - Missing Pictures = 281
- Name? Is this Dog Names? - C6

C3) Confounding Information - Not sure data useful, especially since they are very few (low in counts0.

- 'in_reply_to_status_id', 'in_reply_to_user_id'

C4) Null objects are non-null (None to NaN)

C5) 4 categories should be integers instead of float

- in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id
  - However might not be necessary as this columns might be deleted as they are not useful.
- Possibility of using strings instead of integer - Since there id. Might be better to prevent possible operations

C6) Column (Name) have invalid entries.

- This will be fixed with a column dog_name.

C7) Datetime vs Timestamp: retweeted_status_timestamp, timestamp should be datetime instead of object (string)

C8) Ratings - Derived from rating_numerator and rating_denominator

- Nonsensical values exist (eg. 17775 out of 10).

Tidiness - Strutural issues

T1) 3 tables instead of 1. Should be working on one main table/dataframe. This will be rectified.

T2) Information overload (extra columns providing unecessary information).

- Image Dataset (2075 instead of 2356)
- 'in_reply_to_status_id', 'in_reply_to_user_id', 'user_favourites'
- retweet columns are not necessary

T3) Maturation? - This exist as several columns and should be categories and merge into 1.

Data Issues - Data Issues

As mentioned, basic anthropometric values are missing, however based on visual analysis of the df_master, it might be possible to extract 3 categories:

D1) Name (from text e.g "This is X")

D2) Gender (from text e.g him, his, etc)

D3) Maturation? (from several column eg. 'doggo', 'floofer', 'pupper', 'puppo). This is linked to T3. While this is D3, thus 3rd in the order, since it is linked to T3, it might and most probably be addressed first.

3) Cleaning the Data

During the assessment part, several quality issues that was identified above was rectified. The first was to rectify T1, which is to merge the 3 tables into 1 dataframe. Next was to eliminate duplicates and so on.

4) Storing the data

At several points, the data was exported to .csv format for two purposes. The first is as a backup and the second is to allow the rest of the code to be analysed at a later time.

5) Storing the data

Once the data has been exported. We can begin the analysis.

Regards
Sofyan Sahrom