wrangle_report

April 13, 2018

0.1 Introduction

In this paper we will describe our wrangling effort made in the section of wrangling weRateDog project

Data wrangling consists of:

- Gathering data
- Assessing data
- Cleaning data

0.2 Gathering Data

We need to gather 3 datasets in this project. I have named them as below per my convenience.

- WeRateDogs
- TweetImage
- Tweets

The WeRateDogs dataframe is gathered directly from the existing csv file twitter-archive-enhanced.csv The TweetImage dataframs is downloaded programmatically from the Udacity's servers The Tweets dataframs is gathered from querying Twitter using Tweepy for the retweets and favorates.

0.3 Acessing Data

Data accessiung can be done in two ways. - Programatically - Manually

I have used the python libraries to figure out the Quality and Tidiness issues in the provided dataset, instead of doing manually. We could detect and document the following quality issues and tidiness issues.

0.3.1 Quality

`WeRateDogs` **dataset**

- `name` is sometimes not an actual name
- wrong data types (`in_reply_to_status_id`, `in_reply_to_user_id`, `retweeted_status_id`, `retweeted_status_user_id`, `timestamp` and `retweeted_status_timestamp`)
- missing some `expanded_urls`
- 343rd entry is not a dog rating

- some entries should be classified as puppers (missing data)
- some entries are retweets
- Extra characters after '&'

`TweetImage` **dataset**

- `p1`, `p2`, `p3` inconsistent capitalization (sometimes first letter is capital)
- missing data (only has 2075 entries instead of 2356)

`Tweets` **dataset**

- missing data (only has 2345 entries instead of 2356)

0.3.2 Tidiness

- Three data frames `WeRateDogs`, `TweetImage`, and `tweets` should be one (combined table) since all tables' entries are each describing one tweet

`WeRateDogs` **dataset**

- one variable in four columns (`doggo`, `floofer`, `pupper`, and `puppo`)
- We may want to add a gender column from the text columns in archives dataset
- Get rid of image prediction columns
- Delete unneccessary columns and rename few

0.4 Cleaning

After accessing data and find out the the quality and tidiness issues programatically/maually, the next step is cleaning the issues we find out.

The cleaning step also involves can be done by wither manually or programatically. If the issue is related to one record/column, which might not require and programming efforts as we can clean the issue manually. But if the issue is bigger than it looks and the dataset contains more number of records, then we better do programatic clean up.

Our process was Define, Code and Test and we were always making a copy of tha dataset even we made the copy in file to test the change before applying to the main dataset. We didn't spot all the quality and tidiness assessments at the assessing data section, so we have been iterating and revisiting assessing to add these assessments to our notes.

0.5 Conclusion

Data wrangling indeed is a core skill that everyone who works with data should be familiar with since so much of the world's data isn't clean. If we analyze, visualize, or model our data before we wrangle it, our consequences could be making mistakes, missing out on cool insights, and wasting time. We couldn't be able to make some of the visualization without wrangling (i.e dog gender partition) So best practices say wrangle. Always.