Wrangle report

By Jingjun He

Introduction:

This wrangle report is part of the Wrangle and Analyze Data project in order to document the wrangling efforts of the project. The dataset used in this project is the tweet archive of Twitter user@dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with humorous comment about the dog. The wrangle report documents the three steps of data wrangling: gathering data, assessing data and cleaning data.

1. Gathering Data

In this project, I need to gather data from several sources and different of formats.

1. The WeRateDogs Twitter archive. The file was provided by the project and can be downloaded directly from Udacity website.
2. The tweet image predictions. The file is hosted on Udacity's servers. I downloaded this file programmatically by using the *Requests* library in Python.
3. Get retweets count and favorite count information missing from the Twitter archive from another file. I chose to download the tweet JSON file programmatically by using the Requests library since I don't have a Twitter account.

2. Assessing data

After gathering the data, I assessed the data both visually and programmatically to identify any data quality and tidiness issues. Quality relates to content while tidiness relates to data structure. Tidy data requirements: each variable forms a column, each observation forms a row, each type of observational unit forms a table. Since the datasets were not large, I was able to open them in EXCEL and scrolled through the data and spot any obvious issues. I also used code in Jupyter Notebook to view specific portions and summaries of the data, for example, pandas' `info`, `head`, `sample`, `value_counts`, `duplicated`, `query`, and `describe` methods. I've made notes of the observations while assessing the data so that I can fix the issues later in the cleaning step.

Quality Issues

twitter_archive table:

- datatypes for in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id are float, should all be int
- only need original ratings with pictures, retweets and replies should be removed, related columns should be removed too. The picture part will be fixed later.

- timestamp is str, should be datetime, remove +0000 in timestamp
- abnormal values in rating_denominator, e.g., 170, 150, 130, etc. The rating_denominator is almost always 10
- abnormal values in rating_numerator, e.g., 1776, 960, 666, 204, 165, etc. make no sense.
- source info redundant, not easy to read

image_prediction table:
- inconsistent capitalization in p1, p2 and p3 columns
- jpg url duplicates
- many entries are not dogs, e.g., jaguar, mailbox, peacock, cloak, etc.

tweet_json table:
- missing data probably due to retweets in twitter_archive

Tidiness Issues

- twitter_archive: doggo, floofer, pupper, puppo are all stages of dog, should be in one column
- The three tables should be combined into one since they're all related to the same type of observational unit according to tidy data requirements.

Cleaning data

I cleaned each of the issues documented while assessing. Although there are many issues with the entire dataset, it would be very time-consuming to clean all of them. So, I just focused on those related to my analysis. The programmatic data cleaning process consists three steps: define, code and test. It's also important to make copies of original pieces of data before cleaning. It's important to clean issues in a logical order because some of the issues will disappear as we clean the issue one by one. For example, some of the abnormal rating numerators and denominators we discovered earlier disappeared after we fixing the issue that many images in the image_prediction are not dogs. I removed entries that are not dogs, and some of the abnormal ratings were automatically gone. The datatype issues with in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id is another example. Those variables are all typed as float, while they should be int. Because the project requires only original tweets, any retweets and replies information will be removed. This datatype issue was again not an issue anymore after deleting those columns. Most of the cleanings were done using programmatic tools but some are done manually, for example, correcting the abnormal rating numerators and denominators. I had to filter abnormal ratings and read though the text to find the correct ratings.

After fixing all the issues, reassess the dataset and iterate if necessary. Then store clean data in a csv file.