Wrangle Report

In this report, I document the wrangling processes I did for the project WeRateDog.

The wrangling process consists of three parts:

1. Gathering
2. Assessing
3. Cleaning

1. Gathering

The first step of wrangling process is to gather three different pieces of data from three different sources:

1. The WeRateDogs Twitter archive: I downloaded this file twitter_archive_enhanced.csv manually by clicking the link on the project detail page and upload the file to Jupyter Notebook.

2. The tweet image predictions: I downloaded the image_predictions.tsv file programmatically by using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv .

3. Tweet Data from Twitter API: I used the tweet IDs in the WeRateDogs Twitter archive, query the Twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON data in a file called tweet_json.txt file like on the project detail page requested. In addition I create two lists for tweet_ids to print the existing and non-existing tweet-ids for my own reference. The time module is used in the function to return the elapsed-time, because it turns out pretty useful while the function is running. In my case, it took like almost 50 minutes to download the data. After that I read this file line by line into a pandas DataFrame, I applied all the variables to the data frame at first and dropped irrelevant information later in the process.

2. Assessing

The second step of wrangling process is to assess the data visually and programmatically for quality and tidiness issues. I made a copy of all three files before I do the following steps. I started with the Twitter archive file and found some quality issues listed in the following: there are Erroneous datatypes like in_reply_to_status_id, in_reply_to_user_id, timestamp, tweet_id etc. The retweet columns are included in the dataset. The dog types have None instead of NaN. There are erroneous names in the dog name column. Source column is in HTML code and rating_numerators and denominators are not always correct in comparison to the rating in the column text. In the image_predictions file, I noticed that not all ratings have an image only 2075 from 2356 and there are the same URL for different tweet_ids.

For the tidiness issues: the three tables should be merged into one. The nine columns for predictions and confidence level need to be converted in 2 categorical variables. Also, the four columns: doggo, floofer, pupper, puppo, can be converted into one column as dog stage.

3. Cleaning

In the cleaning process, I use the define, code, and test steps of the cleaning process for each detected issue. I removed the retweet and other irrelevant columns to make it easier for the analyzing process. After that, I corrected the erroneous data types, also the dog names that were wrong. A part of my cleaning process is extracting the source from source-column and converting dog stage and predictions columns to categorical variables. I also removed the tweets with double ratings for the cleanness. Since I do not address all the issues, there are still some quality and tidiness issues remained. To generate visualizations I also extract the year, month, day etc. from the timestamp.

I stored the result as a twitter_archive_master.csv file.