Data Wrangling Report

Documentation for data wrangling steps: gather, assess, and clean
Monday, June 17, 2019
Jaekyung Lee

Data Gathering

Three datasets are given:

- ‘twitter_archive_enhanced.csv’, the WeRateDogs Twitter archive. Downloaded manually.
- ‘image_predictions.tsv’, contains 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. Downloaded programmatically using the Requests library and the given URL.
- ‘tweet_json.txt’, using the tweet IDs in the WeRateDogs Twitter archive, I query the Twitter API for each tweet's JSON data using Python's Tweety library and store each tweet's entire set of JSON data in ‘tweet_json.txt’

I gathered each of them in a Jupyter Notebook titled ‘wangle_act.ipynb’ by using following python packages(libraries), such as pandas, NumPy, requests, tweepy and json.

To obtain Twitter API keys, I sign up for the Twitter Developer’s account and received all accesses.

Data Assessing

Tidiness

1. Dog stages: ‘doggo', 'floofer', 'pupper', 'puppo' are values, not variables.
2. Given Three DataFrames should be merged into one DataFrame.

Quality

1. There are many rows(tweets) without images
2. Remove retweets & reply since we only want original ratings
3. Some dogs names are improper (even though it can be anything)
4. There are many missing values in 'name' column and dog stages
5. Change 'source' column more readable.
6. Remove columns that will not be used for analysis
7. Delete tweets without ratings.
8. Need to change wrong datatypes (date, tweet_id)

Data Cleaning

The two tidiness were easy to fix and creating a new ‘dog_stage’ column was interesting because I decided to use lambda function (not melt) to challenge my coding skills.

Dealing with the quality issues was not only challenging but also interesting since was were no instruction. However, things that I have learned in the Udacity Data Wrangling Course were really practical and helped me go through all data cleaning process.