WeRateDogs Twitter Archive
– Wrangle Report

Data Gathering

There are three files we need to collect into our Jupyter Notebook, one is twitter archive csv file, this file is given by Udacity, so we just use pandas library to read it. The second one is image prediction file, it needs to extract from the url of the Udacity server. So I need to learn how to use requests library to download and store it. The third one is Json file. We need to apply the authority to use the Twitter API and also learn the tweepy library to download the Json file.

Assessment

Here is the problem I found in these three files:

Twitter Archive File:

1. The format of id is wrong and it should be changed to integer.
2. The rating denominator could only be 10 and other values are invalid.
3. When we explore the type of two timestamp columns, one is string, the other one is float. The format of timestamp should be changed to datetime.
4. There are 109 invalid names not starting with a capitalized alphabet.
5. In four columns of dog's stage, the "None" value should be changed to "NaN" in these four columns.
6. As we do not want the duplicated information, so we would clear away the rows of retweet based on retweet id.
7. Change the value for source column.

Tweet Image Predictions TSV:

- There are 2075 rows in prediction file, 2354 rows in JSON data. Based on Twitter Archive file (2356 rows), we know some rows are not matching.

- There are 66 jpg_url which are duplicated.

JSON File:

- No quality issue for the json file.

Cleaning

For cleaning the file, I would first merge it into one. Then clean the tidiness issue like drop the columns we do not use, so we do not have to solve the quality issue from these columns, it saves our time. Then we could start to clean our quality issue that we found on the second part.

In the process of cleaning data, my challenge part is if I am not familiar with the dataset, I need to scroll the page up and down to figure out what is each variable for. So, before assessing the data, we need to make clear the purpose for each column so that we could better clean it.