Data Wrangling WeRateDogs

In this project, data wrangling divided into 3 sections: Gather, Access, and Clean. Step by step wrangling data can be found in ***wrangle_report.html*** or in (jupyter) ***notebook*** part “Gathering The Data”, “Accessing The Data”, and “Cleaning The Data”.

1. Gathering The Data

   In this project, I gathering 3 data from different sources. First data from offline file (csv) which has been given by udacity. Second data from udacity server, and third data from twitter.

   It easier to get first data, it just needs *pandas* to read_csv and we can save it into pandas DataFrame. For second data, a little tricky because we need another library to help that is *requests*. The trickier is third data because the data from twitter, so I need to use twitter API and save the data into JSON. After that, the data can be saved into DataFrame with library *json*.
2. Accessing The Data

   In accessing the data, I just make sure that the data can be accessed properly. I do some task, they are Check length of data, Check the type of data, Check the value of data, Check missing value of data, and Check stat describe data. From that task, I found some dirty and untidiness data. The issues are:

   quality issues:

   1. Axist not original tweet
   2. tweet_id format in the third data doesn't like first data so maybe it can make some problems if we join the two tables
   3. tweet_id position in the third table not same like the other table, so we can't easily see the id
   4. timestamp in the first table not in datetime format
   5. Missing value was not uniformly, sometime NaN but some other None
   6. There are exist columns that have >90% missing value
   7. Cols retwitted and favorited have same value in all row
   8. Cols source have html format
   9. Cols expanded_urls and jpg_urls have duplicated value

   tidiness issues:

   1. Stage of dog must be 1 cols instead of 4 cols
   2. Join all data is needed to make easier for analysis

3. Cleaning The data

   In the cleaning section, I try to solving issues which declare in section accessing the data. After solve the issues, I save each table into csv and I also save csv which contain merge from all tables.