Wrangle Report

This is a report about WeRateDogs Tweets Data Wrangling. There are 3 main steps already done here, which are :

1. Data Gathering
2. Data Assessing
3. Data Cleaning

Data Gathering

In this section, there are 3 ways to import the data from Udacity server and WeRateDogs tweets.

1. Import Downloaded CSV File from Udacity
   The file is successfully imported, with a name file of "twitter-archive-enhanced.csv" and will be called **df_csv**. This file will be used throughout the whole report and future analysis.
2. Import TSV File through HTTP Requests
   The file is successfully imported through udacity server and will be called **df_tsv**. This file will be used throughout the whole report and future analysis.
3. Import JSON File using Tweepy API to scrape WeRateDogs tweets
   The file is successfully imported including the use of Tweepy API and will be called **df_json**. Took 20-30 minutes to get all the needed data. This data however, won't be used for wrangling and analysis, as this is actually an uncleaned **df_csv**.

Data Assessing

In this section, there are 2 main steps to assess data, which are assess Tidiness and Quality. Here are the summary of the whole assessment :

1. Tidiness
   a. **df_csv** columns of **doggo**, **floofer**, **pupper**, **puppo** can be simplified by making them into one categorical column
   b. Inner joining **df_csv** and **df_tsv** by **tweet_id**
2. Quality
   a. **df_merged** rows that contains **retweeted_status_id** as not null can be removed
   b. **df_merged** rows that contains **in_reply_to_status_id** as not null can be removed
   c. columns of **df_merged** that the total values is less than 10% of total rows can be removed
   d. in **df_merged** column, change "a" name to NaN
   e. Change datatypes of **tweet_id** to object and **timestamp** to datetime
   f. Fix **rating_denominator** column of **df_merged**
   g. Remove **source** column
   h. Change every "None" value into NaN (real none value)

Data Cleaning

All Data Cleaning steps according to Data Assessment has been successfully done. However, I found that I didn’t do the assessment and cleaning properly. If you take a look at **name** column, there are names called ChNonerlie, LugNonean, DNonerlNone, and many more with “None” substring. After taking a look at the tweet, the name should be Charlie, Lugan, and Darla. Therefore, I found that the “None” substring should be replaced into “a”. This can be done using pandas.Series.str.replace(“None”,”a”).

Conclusion

Assessing data take a lot of times, especially when you need to create the documentation. But the effort is worth it for future analysis. Data Wrangling I did isn't actually done yet, and always can be continued from this point. Data Wrangling is iterative, so doing from the begining to make sure everything is tidy and clean is urged.

I didn't clean the `name` column which could reflect to the final analysis. Always remember to do Data Wrangling before doing analysis!