Data Wrangling Report

In this project, data of WeRateDogs Twitter were gathered, assessed, cleaned, and stored. Analysis based on the cleaned data was conducted and four insights were concluded.

Gathering Data

First, three pieces of data were gathered to Jupyter Notebook and imported to three pandas DataFrames. 'twitter_archive_enhanced.csv' file was uploaded to Jupyter Notebook; 'image_predictions.tsv' was downloaded from the given URL; '`tweet_json.txt`' was uploaded to Jupyter Notebook and read line by line to a pandas DataFrame. The three dataframes are: *twitter_archive_enhanced*, *twitter_image_predictions*, and *tweet_count*.

Assessing Data

By pandas functions such as head(), info(), value_counts(), duplicated(), and nunique(), etc., the three DataFrames were investigated and several quality and tininess issues are listed as the following:

Quality Issues

twitter_archive_enhanced table

1. Column 'tweet_id' data type is int64, should be corrected to string
2. Too much missing data in column 'in_reply_to_status_id' and column 'in_reply_to_user_id', which should be deleted
3. Retweets and tweets without images should be deleted. Three columns about retweet status should be deleted
4. Some rating denominators are not 10 and some numerators are abnormal large numbers. Rating data should be re-extracted from 'text' column
5. Column 'name' has missing data and mistake names such as, 'a', 'an', 'the', etc. All names with problems initialed with lowercase letter. Those problem names would be set as NaN
6. Most dog stage data are missing and some dogs have two stages. Dog stage data should be re-extracted from 'text' column

twitter_image_predictions table

7. Duplications in column 'jpg_url', which should be deleted
8. Data type of column 'tweet_id' is int64, should be corrected to string

tweet_count table

No issue

Tidiness Issues

9. In '*twitter_archive_enhanced*' table, variables as doggo, floofer, pupper,and puppo are values of one variable. They should be included in the variable 'dog_stage'
10. All three tables could be merged on 'tweet_id' for analysis, as all variables belong to one observational unit - dog rating

Cleaning Data

Based on items listed in the assessing data section, cleaning process was conducted. Due to time limitations, only 8 quality issues and 2 tidiness issues were cleaned according the project requirements and motivations. Therefore, the data set still has much space to be improved.

The worst and most time consuming part of data is dog rating, which is listed as item 4 in the assessing list. Some rates have abnormal high numerators; while denominators of some rates are not 10; some tweet has more than one rates extracted from 'text' column. This could be improved by reading and understanding the content of text and then modified manually. However, as the portion of tweets with problems is small, we set their numerators and denominators as NaN for simplicity.

The final DataFrame was merged from the three pieces, which was stored to a CSV file 'twitter_archive_master.csv'.