Wrangle Report

> Data Source: WeRateDogs Twitter data

Gathering Data

1. For the first twitter archive `twitter-archive-enhanced.csv`, I simply downloaded the csv file and used the `pd.read_csv('twitter-archive-enhanced.csv')` function to load the dataset;
2. For the second file `image_predictions.tsv`, I used `requests.get(url)` function, filling with the url provided by *Udacity.com* to load the dataset;
3. For last dataset `tweet_count`: I applied for a Twitter developer account at first and set my account information in the Jupyter Notebook; using the tweet ids in the `twitter-archive-enhanced.csv` dataset as a reference, I acquired each tweet's status by the `for loop` function; as the instruction stated, I add a `try-except` block and a `code timer` block to print the time for retrieving each tweet status and "error" if the data was not retrived successfully; the status of each tweet is recorded line by line in `tweet_json.txt`; as I only need the data about `tweet_id`, `favorite_count`, and `retweet_count` of each tweet, I used another `for loop` function to retrieve the data and store the data into a new dataset `tweet_count` using the `pd.DataFrame()` function

Assessing Data

Quality

`archive` table

- Useless retweeted info (we only want the original ratings according to instructions)
- inconsistency with rating numerator and denominator columns
- Rating denominator equals to "0" for *tweet_id:835246439529840640*
- Redundant information in the "text" column (repeated rating and short link)
- Ambiguous information in the "source" column
- Missing values (2297 instead of 2356) and repeated info in "expanded-url" column (e.g.url repeated three times for *tweet_id: 835152434251116546*, detected by visual analysis)

`tweet_count` table

- Missing values (2325 instead of 2356)

(Before accessing the `image` table, I first cleaned the useless retweeted tweet to get the accurate number of rating that we need for this data wrangling)

`image` table

- Missing values (2075 instead of 2097)
- Unnecessary columns except "tweet_id" and "jpg.url"

Tidiness

- One variable (dog_stage) in four columns in the `archive` table; some tweets contain two dog stages*(detected by visual analysis: some of them mentions two stages for one dog; others have two dogs presented in the image)*.
- Three sub-tables should be merged into one table as tweet_id duplicated in all datasets

*Note: concerning the missing values in `tweet_count` and `image` tables, I chose the minimum number (2075, same as the total number in `image` table) since we target to the **original ratings that have pictures**.*

Cleaning Data

Note: the tweet (id:835246439529840640), whose `rating_denominator' is 0, was retweeted and hence was cleaned in the previous section.

Define

`archive` table

1. Drop the retweeted rows and the "in_reply_to_status_id", "in_reply_to_user_id", "retweeted_status_id","retweeted_status_user_id", "retweeted_status_timestamp" columns
2. Add a new column "rating" with the calculated rating (formula: rating_numerator/rating_denominator)
3. Move the short link into a new created column "short_url" and drop the long and redundant "expanded_urls" column (which also has some missing values - 2094 instead of 2097)
4. Delete the short link and keep the original text as that in tweets in the "text" column
5. Only keep the essential information in the middle in the "source" column(iPhone, Vine, Web, and TweetDeck)
6. Convert the datatype of "timestamp" column to datetime
7. Create a new column "dog_stage" containing the information in the "doggo","floofer","pupper", and "puppo" columns and drop the intermediate columns. (*Note: . Use "/" to label values have two dog stages)*
8. Arrange the order of columns

Other processes

- Only keep "tweet_id" and "jpg.url" columns in the `image` table (as other information will be used for simulating the confusion matrix)
- Merge the `archive` , `tweet_count` , and `image` tables into one final table using `inner` join (for keeping data completeness at maximum).

Code & Test

- See the detailed process in the `wranle_act.ipynb`
  (Note: the final dataset `twitter_archive_master.csv` contains only 1961 values after the merge)