Data wrangling for WeRateDogs Twitter

I Gathering

I gathered data from three different sources.

1) I downloaded ‘twitter_archive_enhanced.csv’ manually from Udacity website.

2) I downloaded ‘image_predictions.tsv’ programmatically using the Requests library from Udacity website.

3) I obtained the ‘tweet_json.txt’ file by querying the Twitter API with Tweepy library.

II Accessing

1) Visual assessment

I imported the ‘twitter_archive_enhanced.csv’, ‘tweet_json.txt’ and ‘image_predictions.tsv’ into three pandas dataframes for visual assessment purposes. I found a few dogs were named 'a'/'an', which are clearly mistakes when comparing with column 'text'.

2) Programmatic assessment

I found the following problems through programmatic assessment with methods such as ‘info’, ‘value_counts’, ‘isnull’.

Quality problems:

1. The table has retweets but we only want original ratings. Retweets are tweets with non-null retweeted_status_id.
2. Columns retweeted_status_id, retweeted_status_user_id and retweeted_status_timestamp are all null after deleting retweets. So these columns should be removed.
3. Erroneous datatypes (tweet_id, in_reply_to_status_id, in_reply_to_user_id should be string type instead of int/float since their calculation doesn't make any sense)
4. timestamp have a datatype of string, which should be changed to datetime
5. rating_denominators and rating_numerator of tweet_id(835246439529840640, 832088576586297345) are not correct.
6. dog named 'a'/'an'/'the'
7. dog named 'by'
8. repeated urls in column 'expanded_urls'
9. tweet_id in image_predictions table should has a data type of string

Tidiness problems:

1. One variable in four columns in `twitter` table ('doggo', 'floofer', 'pupper','puppo').
2. `twitter_download`, `image_predictions` and `twitter` should be merge into one table.

III Cleaning

I cleaned all the problems found in part II programmatically with the define-code-test working process.

- The retweets were deleted with Boolean selection.
- The datatype change from int to string was achieved with pandas astype method.
- The datatype change from string to datetime was achieved with pandas to_datetime method.
- The dog names of ‘a’/ ‘an’/ ‘the’ were changed to ‘None’ with pandas replace method.
- I defined a function ‘remove_repeated_url’ to remove the repeats in ‘expanded_urls’ column.
- I wrote a for loop to create a column dog_stages, which contains all the information in columns 'doggo', 'floofer', 'pupper', 'puppo'.

IV Storing

The cleaned data was saved to ‘twitter_archive_master.csv’ without index.

V Analysis and visualization

1. How many original tweets were posted in each month?

The quality problems 1 and 4 are the basis of this analysis.

2. How many times each tweet was retweeted on average in each month?

The quality problems 1 & 4, and tidiness problem 2 are the basis of this analysis.

3. How many favourites each tweet got on average in each month?

The quality problems 1, 3 & 4, and tidiness problem 2 are the basis of this analysis.

4. What are the main sources of tweets?

The quality problems 1, 3 & 9, and tidiness problem 2 are the basis of this analysis.

5. When the neural network makes mistakes (believe the image is not a dog), what usually the neural network thinks the image is?

The quality problems 1, 3 & 9, and tidiness problem 2 are the basis of this analysis.

6. What are the most common dog names?

The quality problems 1, 3, 6 & 7, and tidiness problem 2 are the basis of this analysis.

7. What are the distributions of dog stages?

The quality problems 1 and tidiness problem 1 are the basis of this analysis.

8. How did the rating changes over time?

The quality problems 1, 4 &5 and tidiness problem 2 are the basis of this analysis.