Data Wrangling for WeRateDogs Datasets

Author: Xueyun Zhou

This report summarizes the processes of data wrangling for the three WeRateDogs Datasets.

1. Gathering data

Three pieces of data as follows are gathered:

1) the WeRateDogs Twitter archive, saved as df_enhanced;
2) the tweet image predictions, saved as df_image;
3) each tweet's retweet count and favorite ("like") count from the Twitter API, saved as df_json.

2. Assessing data

2.1 Quality issues

2.1.1 df_enhanced

(1) Some "rating_denominator"s are not 10.
(2) The values for "doggo", "floofer", "pupper", and "puppo" are not the Boolean type but showed as the column names themselves.
(3) There are observations whose "retweeted_status_id" are not NaN, so they are retweets.
(4) The data type of "tweet_id", "rating_numerator", and "rating_denominator" is int64 and that of "timestamp" is str.
(5) Some values of "rating_numerator" are unusual (420, 666, 960, 1776, etc.).
(6) There are zeros in "rating_denominator".
(7) There are too much missing data for the "doggo", "floofer", "pupper", and "puppo" columns.

2.1.2 df_image

(8) The values of "img_num" for some observations do not represent the images with the highest confidence.
(9) The data type of "tweet_id" is int64.
(10) The "img_num" of some observations is 4, while there is no data for the 4th image.
(11) There are 2356 observations in df_enhanced, but 2075 in df_image, so there are tweets in df_enhanced which do not have images.

2.1.3 df_json

(12) The data type of "id" is int64.

2.2 Tidiness issues

(13) The column name of the tweet ID in df_json is different from that in df_enhanced and df_image.
(14) There are too many unrelated columns collected in df_json.
(15) The "source" column is in a html formatting.
(16) The columns "in_reply_to_status_id" and "in_reply_to_user_id" in df_enhanced are useless. The columns "retweeted_status_id", "retweeted_status_user_id", and "retweeted_status_timestamp" are useless after the retweets are deleted.
(17) The three datasets should be merged.

3. Cleaning data

3.1 Issue (13), (14)

Remove all columns for df_json_copy except "id", "favorite_count", and "retweet_count".
Change the column name of "id" to "tweet_id".

3.2 Issue (3)

Delete retweets.

3.3 Issue (4), (9), (12)

Change the data type of "tweet_id" in three datasets to string.
Change the data type of "timestamp" to pandas datetime.
Change the data type of "rating_numerator" and "rating_denominator" to float.

3.4 Issue (15)

Delete the html contents in "source".

3.5 Issue (16)

Delete the columns "in_reply_to_status_id", "in_reply_to_user_id", "retweeted_status_id", "retweeted_status_user_id", and "retweeted_status_timestamp".

3.6 Issue (2), (7)

Gather information of the stage from "text".
Reassign the stage to "doggo", "floofer", "pupper", "puppo" as Boolean type.

3.7 Issue (1), (5), (6)

Check the unusual rating_denominator and rating_numerator, and update based on "text".

3.8 Issue (8), (10)

Create a column in df_image_copy for the dog type based on confidence columns and "is dog" columns.
Delete useless columns.

3.9 Issue (11), (17)

Merge 3 datasets. The df_enhanced_copy and df_image_copy should be merged by inner to remove tweets without images.

After cleaning and merging, the clean dataset is ready for storing, analysing, and visualizing.