Wrangle Report

Linda Xu

May 2021

The dataset that I wrangled (and analyzed and visualized) was the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. There have been three steps:

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

Gathering Data

1. The WeRateDogs Twitter archive. Download this file manually by the Udacity Server.
2. The tweet image predictions. This file can be downloaded manually by the Udacity Server as well.
3. entire set of JSON data in a file called “tweet_json.txt”. And I manually downloaded through the server from Udacity.

I renamed those three files as twitter_archive, img_df, and status_df.

Assessing Data

I found those quality and tininess issues through assessing data. Those are the issues I found :

Quality

1. "twitter_archive_df" has 2356 rows while the "image_predictions_df" has only 2075 rows, probably due to retweets and missing photos.
2. tweet_id is an int (applies to all tables).
3. The timestamp column is in string format.
4. Dog names: some dogs have 'None' as a name, or 'a', or 'an.'
5. This dataset includes retweets, which means there is duplicated data.
6. Strange, unexpected values in rating_denominator.
7. Strange, unexpected values in rating_numerator.
8. p1, p2 and p3 columns have invalid data, and those three columns are not consistent, and I must create two more columns to adjust the issue.

Tidiness

1. All three files have common tweet_id column, which can combine those three to one data frame.
2. 4 different columns (doggo, floofer, pupper, and puppo) on dog stages should be combine in only one column.

3. Drop unnecessary columns that will not use in the analysis.

Cleaning Data

In terms of cleaning data, these are the steps I followed:

1. Merge the clean versions of twitter_archive, img_df, and status_df dataframes Correct the dog types.
2. Create one column for the various dog types: doggo, floofer, pupper, puppo.
3. Delete retweets. Remove columns no longer needed: in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, and retweeted_status_timestamp.
4. Change tweet_id from an integer to a string.
5. Change the timestamp to correct datetime format.
6. Correct dog naming issues.
7. Standardize dog ratings. All values in the rating_denominator column is 10, so we don't need this column any more, and the rating_numerator column can be renamed rating.
8. Creating breed and confidence columns in Twitter data frame.

    i. Create two new columns in predictions called breed and confidence: check each dog breed prediction flag in order (p1_dog, p2_dog, p3_dog), and copy the breed with the highest confidence level into the breed column.
    ii. Copy the associated confidence level into the new confidence column.
    iii. For cases where all 3 prediction flags are False (NOT a valid dog breed), set breed as 'none' and confidence to 0.

For each step cleaning procedure was documented as “Define”, code was developed and tested.

Finally, the new dataset was stored as .csv file as “twitter_archive_master.csv”.

Most importantly, the above cleaning procedures cannot guarantee this is the tidiest or best quality dataset since further quality and tidiness issues need to be discovered in order to have a better picture of this twitter dataset.