Wrangle report

Prepared by Huu Tri Nguyen

1. Gathering

In the gathering step, 3 different tables are provided: twitter-archive-enhanced.csv: this file is obtained by downloading it manually from the Udacity site. 2. image-predictions.tsv: this file is obtained by downloading it programatically using the python's requests library and lastly. tweet_json.txt: this file is obtained by calling the twitter API for each tweet.

After that, all three files are made sure to be in the same directory so that we can use pandas to read the data from them.

2. Assessing

2.1 Quality

tweet_api_data findings:

- Some columns are unrelated such as user, favorited, retweeted as these are based on the personal user information
- Some columns contain mostly null data: contributors, coordinates, geo, place, quoted_status, quoted_status_id, quoted_status_id_str, quoted_status_permalink
- Some columns seem to contain the whole HTML tags
- both possibly_sensitive and possibly_sensitive_appealable columns only have "False" as value

image_predictions findings

It doesn't seem like to have any issue. To me this is a clean data set

twitter_achives findings

- the dog name may not be accurate, some names are a or an.
- the dog name have None string that should be a null instead.
- Column name doggo, floofer, puppo, and pupper has value either None or its column name.
- timestamp should be in datetime instead of string
- some rating wasn't capture correctly. For example:

  "@jonnysun @Lin_Manuel ok jomny I know you're excited but 960/00 isn't a valid rating, 13/10 is tho*. " 960 and 00 were recorded as numerator and denominator respectively instead of 13 and 10

2.2 Tidiness

- According to this StackOverFlow link: https://stackoverflow.com/questions/18869688/twitter-api-check-if-a-tweet-is-a-retweet If the tweet is a retweet then the column will contains some data, else it will be blank. From observing the table, we can see that some tweets in the tweet_api_data are retweets.

- created_at / timestamp, source, in_reply_to_status_id, in_reply_to_user_id are duplicated in tweet_api_data table and twitter_archive table

- tweet_api_data and image_predictions should be part of twitter_archive table
- Entities data seems to contain image information which are already contained in the twitter archive data, like the image_url and extended url

- Extended entities column contains duplicate information of the entities column

3. Cleaning

tweet_api_data

The first thing to do is to remove the retweets as these do not contain the data from WeRateDog since they come from a different sources. To clean it, first I need to get the indexes of the rows that are retweets (those that have non-null data in retweeted_status). Then I used .drop method to remove the rows of those indexes. As a result, 165 rows were removed

twitter_archives table

removing duplicated columns:

Since tweet_api_data and twitter_archives tables contain some similar columns including
created_at / timestamp
source
in_reply_to_status_id
in_reply_to_user_id are dupli- cated

These columns will be removed by .drop method.

tweet_api_data and image_predictions should be part of twitter_archive table

All I need to do is to an inner join of two tables together on tweet_id and id. And we get a joined table dataframe twitter_archives_clean

The resulting dataframe twitter_archives_clean_final has a lot of columns with null or small number of non-null data. Including both possibly_sensitive and possibly_sensitive_appealable as these only have "False" as value. All of these columns will be removed to and we are left with the final tables with 1987 rows each

Column name doggo, floofer, puppo, and pupper has value either None or its column name

Changing these value to True (its column name) and False (None) respectively by using the replace function. Then we do a value_counts check for each columns to ensure that only True and False are in there

Name of the dog are marked as None instead of nan in twitter_archives_clean table

For this, we used replace function to change None to NaN data

Some dog have name a, an, and the in twitter_archives_clean table

Similar to what we just did. We changed these invalid names to NaN data

Some rows have invalid rating

By looking at the value_counts of the rating_denominator and rating_numerator, it should be safe to assume that valid numerator should be 14 or less and the denominator should be 10 as these has the most value counts.

We sorted this issue by removed the rows with these invalid ratings.