“We Rate Dogs” Data Cleaning

I worked with three data frames:

1. archive was given to Udacity by “We Rate Dogs”. It contains tweets, and programmatically harvested info (dog name, and dog stage [‘puppo’, ‘doggo’, etc.])
2. full_tweets I downloaded from “We Rate Dogs” using Tweepy. It contains additional information about the tweets in archive. For example, full_tweets includes the number of times a tweet was retweeted or favorited.
3. image was created by Udacity. They used image processing software to label the dogs in the pictures (top three guesses of the image processing software.

Tidiness Issues

1. The information about dog stages should be in one column, not spread out in four columns (since dog stage is one variable (one piece of data)).
   Solution: I combined the four dog stage columns into one column called dog_stage.
   [There are some rows that have more than one dog stage.
   In Quality step #6 I went through all 14 rows with more than one dog status. After that process, there were still some rows that retain either both “doggo” and “pupper” or “doggo” and “puppo”, but no other combinations (ex: no rows with “doggo” and “floofer”).]

2. Both archive and image contained descriptive information about the dogs in the tweets.
   Solution: I joined the archive and image dataframe using merge.

   Note:
   Udacity recommends combining the full_tweets data frame with the archive and image data frames because they “have the same observational units”.

   I don’t think so.

The archive and image dataframes’ most salient feature is that they have descriptions about the dogs in the tweets. The unique columns in the full_tweets data frame is information about how people reacted to the tweets.

Tidiness requires each variable to form a column, each observation a row, and each type of observation a table. I feel like extracted info about the dogs (archive and images) and recorded info about what people did with the tweets (full_tweets) are two different categories of information. Therefore, I don’t think we should join full_tweets to the other two dataframes.

However, to compromise with the Udacity graders, and since I am going to combine all three dataframes for the analysis step anyway, as the final step of the cleaning process, I will combine the data frames into one large data frame.

Quality Issues

1. Some of the names (archive) were clearly wrong. Like “a”, “an”, “the” ...
   Solution: I used .str.islower() with .loc to examine all the lowercase “names” and confirmed that all 109 of them were not names, but normal words that the algorithm mistook for names.
   Solution: I replaced all lowercase “names” with “” (ie: nothing) using .replace()

2. The timestamp and retweeted_status_timestamp columns (both archive) should be in the datetime data type format, but were not.
   Solution: I changed both column’s data types using .todatetime()

3. Also, the time_stamp and retweeted_status_timestamp columns had an extra +0000 at the end, while the created_at columns (full_tweets) are written like this 2017-08-01 16:23:56 (without the +0000).
   Solution: This was automatically corrected when I applied the .todatetime() in the last step (the two +0000s were removed).

4. Some of the dog breeds (images) were capitalized and others weren’t.
   Solution: I made all dog breeds lowercase with .apply(lambda x: x.str.lower(), axis=1)

5. Tweet_id, retweeted_status_id and retweeted_status_user_id were stored as integers (archive and images).
   Solution: I changed those data types to strings with .astype(str)

6. Fourteen of the rows (archive) were labelled with more than one dog stage. I went through them and found that six of those double-labels should be a single label.

   This was discovered with the following code:

   ```
   mask=archive[['doggo', 'puppo', 'pupper', 'floofer']].apply(lambda x: x.replace("None", np.nan)).dropna(thresh=2).index
   archive.loc[mask,['tweet_id','doggo', 'puppo', 'pupper', 'floofer']]
   ```

   There were 12 doggo, pupper rows
   And one doggo, puppo row
   And one doggo, floofer row

   Solution: I changed each with a change list, a mask, and .loc, along with .str.replace and .str.stri

7. In the full_tweets_clean dataframe, id, id_str, in_reply_to_status_id, in_reply_to_status_id_str, in_reply_to_user_id, in_reply_to_user_id_str, quoted_status_id, and quoted_status_id_str should all be in the string data type.
   Solution: we will change the datatypes with .astype(str)

8. 31 of the tweets in full_tweets are quotes. Quotes aren’t supposed to be included in the data.
   Solution: We will delete the quotes.