“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 tweet ids were called “tweet_id” in the archive and image dataframes, but ‘id’ in the full_tweets dataframe.
   Solution: I changed the ‘id’ column name to ‘tweet_id’ in full_tweets.

2. Several columns were repeated in archive and full_tweets.
   Specifically: in_reply_to_status_id, in_reply_to_user_id, source, tweet_id, and text (archive) / full_text (full_tweets) and timestamp (archive) / created_at (full_tweets)
   Solution: Except for tweet_id, those columns are redundant. I used .drop() to remove them from the smaller dataframe (archive).

3. 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”).]

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

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 confirm that all 109 of them were not names, but normal words that the algorithm mistook for names. I replaced all lowercase “names” with “” (ie: nothing) using .replace()

2. The time_stamp column (archive) wasn’t in the datetime data type format.
   Solution: I deleted that column from archive anyway (see step #2 in tidiness).

3. The values in the retweeted_status_timestamp column (archive) were objects, not timestamps (wrong data type).
   Solution: I changed retweeted_status_timestamp column’s values to timestamp using .todatetime()

4. Some of the column names in archive were repeated but with different names in full_tweets.
   Specifically: text (archive) / full_text (full_tweets) and timestamp (archive) / created_at (full_tweets)
   Solution: We deleted those two columns from the archive dataframe (see step 2 in tidiness).

5. 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)

6. Tweet_id, retweeted_status_id and retweeted_status_user_id were stored as integers.
   Solution: I change those data types to strings with .astype(str)

1. 14 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.strip

2. Some of the column names in archive are repeated but with different names in full_tweets.
   Specifically: text (archive) / (full_tweets) and timestamp (archive) / created_at (full_tweets)
   Solution: We already deleted those two columns from the archive dataframe (see step 2 in tidiness).

3. 31 of the tweets in full_tweets are quotes. Quotes aren't supposed to be included in the data.
   Solution: I deleted the quotes.