Project: Wrangle and Analyze Data

Wrangling Review

By: Andrea Bredesen

In wrangling the data, here are the steps I took, why, and additional details regarding my wrangling efforts:

Data Gathering:

First, I obtained the twitterAE (twitter-archive-enhanced.csv) file as presented and saved. The imageP (image_predictions.tsv) was obtained via the link. The twitter Json file was to be obtained through the Twitter/X api. Unfortunately, Twitter/X, connection has changed. It seems to be able to pull tweet information, I'm now required to have a basic subscription which costs $100/month.

Assessing Data:

While assessing the data, I identified the following quality issues to be corrected:

Quality issues

1. imageP: Identify probability of "Not a dog" or Breed of dog based on P1, P2, and P3 column values
2. imageP: Format Breed names properly (Caplitalize and spaces not "_")
3. imageP: Remove unneeded columns
4. twitterAE: Single out the original tweets vs retweets
5. twitterAE: Review text about dogs to identify other "Not a dog" or "we rate dogs" to remove other "not a dog" row.
6. twitterAE: Per notes, denominators should almost always be a 10. Review.
7. twitterAE: Review extreme outliners of numerators. Review.
8. twitterAE: Review dog names ie: "a", "an", "not", "just", "mad" and others.

Data Cleaning:

The first 3 issues were addressed together in code. Addressing the probability of a certain breed and the name formatting at the same time prevented duplicate iteration through the df. The drop of unnecessary columns at this time also seemed appropriate.

Singling out the original vs retweets was completed by finding columns with values in the “retweeted_status_user_id” and “in_reply_to_status_id”, and removing those rows.

Issue 5, initially, I was going to work on removing all tweets that referred to “we rate dogs”, “not a dog”, and similar references to indicate a tweet was not a canine. During this process, I found quite a few had high confidence levels of it being a certain breed of dog. With further research, I found that these comments were more of a “Joke” within the post. These tweets were not removed, however, copies of the dataframes were made and the steps needed to review are within the Notebook. Again, these are not coded to be removed from the main copies of the dataframes. Be sure to review the cuteness of some of these “Not a Dog!”s

While reviewing issue 6, to correct the denominator, I also corrected some numerators. Most of the numerator/denominator variations were due to multiple dogs in the photos.

Issue 7, reviewing the other numerators, one dog has a 1776 rating and another is rated 420. Photos of each dog is in the notebook for review. My review is to keep these ratings in the query. However, I have provided coding within the Notebook should you chose to readjust. For “task1776”, indicate “drop” or “update”. If you’re updating, indicate the rate to update with “rate1776”

Example: task1776 = update rate1776 = 10

Similarly for the “Dogg” rated 420.

Example: task420 = drop rate420 = [dropping, you do not need to address]

Issue 8 was reviewing the dog names that didn’t split correctly. Names such as "a", "not", "just", "my", "incredibly" and similar were reverted back to NaN. Per suggestion of the last review, I amended my query to search for lower case text to rever to NaN. I then did a search through the text for “named”, and finallaly updated the column with the newly found names.

The following Tidiness issues were addressed:

Tidiness issues

1. twitterAE: Combine Dogtionary columns into 1
2. twitterAE: Merge 3 dataframes into 2 for saving.

Issue 1, the Dogtionary, which refers to age variations of the dogs, were listed in 4 different columns. I collapsed these columns into 1. Not all dogs had a Dogtionary listed those are “None”. I did an additional text search for these keywords. Per previous review suggestion, I checked for values in multiple columns and changed that value to “multiple Dogtionary”.

Issue 2 with the merge put all the dataframes together for final review and saving.

Analyzing and Visualizing Data

During the analyzing and visualizing of data, I was hopeful to remove “not a dog”. Unfortunately, the confidence level was not sufficient to identify “Not a Dog” as many “Not a Dog” were, in fact, dogs (again, these were found by all P#_Dog being false). Searching tweet text identified more “Not a Dogs” as a funny joke reference and not an actual identification of “not a dog”. Low numerator ratings with retweet and favorite averages also didn’t seem to give an accurate “not a dog” indications, as seen in the last bar chart. Photo Sampling of the “Not a Dog” are included.