Wrangle Report

The dataset wrangled in the project is 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.

The WeRateDogs Twitter project goals included:

- Wrangling the twitter data through the following processes:
    - Gathering data
    - Assessing data
    - Cleaning data
- Storing, analyzing, and visualizing your wrangled data
- Reporting on the data wrangling efforts and data analyses and visualizations

Gathering Data

My wrangling efforts for the WeRateDogs Twitter project included gathering data from the following sources:

- The WeRateDogs Twitter archive. The twitter_archive_enhanced.csv file was provided to Udacity students. WeRateDogs downloaded their Twitter archive and sent it to Udacity via email exclusively for you to use in this project. This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017.
- The tweet image predictions, i.e., what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network. This file was provided to Udacity students.
- Twitter API and Python's Tweepy library to gather each tweet's retweet count and favorite ("like") count at minimum, and any additional data I find interesting.

Assessing Data

Once the data was gathered, I began to assess the data on both quality and tidiness issues.

<u>Quality Issues</u>

archive:

- Completeness:
    - missing data in the following columns: in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, expanded_urls
    - tweet_id is an int (applies to all tables)
- Validity:

    - dog names: some dogs have 'None' as a name, or 'a', or 'an.'
    - this dataset includes retweets, which means there is duplicated data (as a result, these columns will be empty: retweeted_status_id, retweeted_status_user_id and retweeted_status_timestamp)
- Accuracy:
    - timestamp is an object
    - retweeted_status_timestamp is also an object (the other retweeted statuses are floats)
    - rating_numerator goes up to 1776
- Consistency:
    - rating_denominator should be a standard 10, but there are a multitude of other values
    - the source column still has the HTML tags

images:

- Validity:
    - p1, p2 and p3 columns have invalid data...why would the algorithm labeled a dog photo as a starfish, boathouse, or mailbox (among other things)?
- Consistency:
    - p1, p2 and p3 columns aren't consistent when it comes to capitalization: sometimes the dog breed listed is all lowercase, sometimes it is written in Sentence Case.
    - in p1, p2 and p3 columns there is an underscore for multi-word dog breeds

twitter_counts_df:

- Completeness:
    - missing some data

<u>Tidiness Issues</u>

archive:

- The last four columns all relate to the same variable (dogoo, floofer, pupper, puppo)

images:

- this data set is part of the same observational unit as the data in the archive - one table with all basic information about the dog ratings

twitter_counts_df:

- this data set is also part of the same observational unit - one table with all basic information about the dog ratings

Cleaning Data

After the assessment, I cleaned the data through the following means:

<u>Define, Code and Test</u>

1. Merge the clean versions of archive, images, and twitter_counts_df dataframes Correct the dog types

2. Create one column for the various dog types: doggo, floofer, pupper, puppo 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
3. Delete retweets
4. Remove columns no longer needed
5. Change tweet_id from an integer to a string
6. Change the timestamp to correct datetime format
7. Correct naming issues
8. Standardize dog ratings
9. Creating a new dog_breed column using the image prediction data