Wrangle Report

We have used tweet archive of Twitter user @dog_rates as the project data set. @dog_rates or "WeRateDogs" is a Twitter account which rates people's dogs with a humorous comment about the dog.

The WeRateDogs (data wrangling) Twitter project goals covered the following:

- Wrangling the twitter data through the following processes:
    - Gathering data
    - Assessing data
    - Cleaning data
- Storing, analyzing, and visualizing the wrangled data
- Reporting of the data wrangling efforts with supportive visualizations

Gathering Data

Data were obtained from the following various diversified 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 us to use in this project. This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets .
- The tweet image predictions, i.e., what breed of dog is present in each tweet , as per the neural network predictions. 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 that I find interesting.

Assessing Data

Once the data gathering stuffs were completed, I began to assess the data on both quality and tidiness issues.

Quality Issues

Twitter Archive:

- Completeness:
    - Twitter Archive data frame has missing data in the following columns: in_reply_to_status_id (78 not null, out of 2356), in_reply_to_user_id (78 not null, out of 2356), retweeted_status_id (181 not null, out of 2356), retweeted_status_user_id (181 not null, out of 2356), retweeted_status_timestamp (181 not null, out of 2356), expanded_urls (2297 not null, out of 2356).
    - tweet_id is an integer. It is too big to be an integer.

- Validity:
    - dog names have invalid data such as 'none', 'a' and 'an'.
    - The data frame contains columns like retweeted_status_id, retweeted_status_user_id and retweeted_status_timestamp. These are unnecessary columns for this data frame.
- Accuracy:
    - timestamp is an object there
    - retweeted_status_timestamp is also an object (the other retweeted statuses, like retweeted_status_id, retweeted_status_user_id are floats)
    - The value for rating_numerator column, as we can see from the statement executed above, is going till 1776. This doesn't seem accurate.
- Consistency:
    - rating_denominator column is having values other than 10 as well. This is inconsistent.
    - source column still contains html tags. This is also inconsistent.

Twitter Images:

- Validity:
    - data frame's p1 column contains various invalid values, for example, 'seat_belt', 'envelope', 'radio_telescope', 'mailbox' etc.
    - Similarly, p2 also contains various invalid values such as 'purse', 'boathouse' etc.
    - Same applies for p3 column as well.
- Consistency:
    - The cases for p1 or p2 or p3 column values are not consistent. Sometimes, those are starting with capital letter, sometimes they are all in lowercase letters. While sometimes, those are written in sentence case.
    - p1, p2 and p3 column values contain an underscore for multi-word dog breeds.

ReTweet Favourite Collection:

- Completeness:
    - The data frame contains missing data

Tidiness Issues

Twitter Archive:

- The last four columns ('doggo', 'floofer', 'pupper', 'puppo') relate to the same variable.

Cleaning Data

After the assessment, cleaning process was done as the following:

Define, Code and Test

1. Merging of the three (clean) data frames, that is archive_clean, images_clean and twitter_counts_df_clean
2. Creating one column for the various dog types: doggo, floofer, pupper, puppo
3. Deleting retweets

4. Removing unnecessary columns like 'in_reply_to_status_id', 'in_reply_to_user_id', 'source', 'img_num'
5. Changing tweet_id from an integer to a string
6. Removing time zone from timestamp column and convert it to datetime
7. Maintaining common casing for name
8. Standardizing dog ratings
9. Creating a new column named dog_breed using the image prediction data