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:

1. Gathering data
2. Assessing data
3. 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.

• 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 interesting.

Assessing Data

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

Quality Issues

archive:

• Completeness:

1.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

2. tweet_id is an int (applies to all tables)

• Validity:

1.dog names: some dogs have 'None' as a name, or 'a', or 'an.'

2.Dataset have duplicated data (as a result, these columns will be empty: retweeted_status_id,

retweeted_status_user_id and retweeted_status_timestamp).

• Accuracy:

1. timestamp is an object.
2. retweeted_status_timestamp is also an object.
3. rating_numerator goes up to 1776.

• Consistency:

1. rating_denominator should be a standard 10, but there are a multitude of other values.
2. the source column still has the HTML tags.

Image

- **Validity:**

    1. p1, p2 and p3 columns have invalid data

- **Consistency:**

    1. 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.

    2. in p1, p2 and p3 columns there is an underscore for multi-word dog breeds

Twitter Counts

- **Completeness:**

    1. missing some data

Tidiness Issues

archive:

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

images:

1. This data set is part of the same observational table show all basic information about the dog ratings

Twitter Counts :

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

Cleaning Data

Define, Code and Test

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. Remove columns source, img_num.

4. Change tweet_id from an integer to a string.

5. Change the timestamp to correct datetime format.

6. Correct naming issues.

7. Calculate standardize dog ratings.

8. Creating a new dog_breed column using the image prediction data