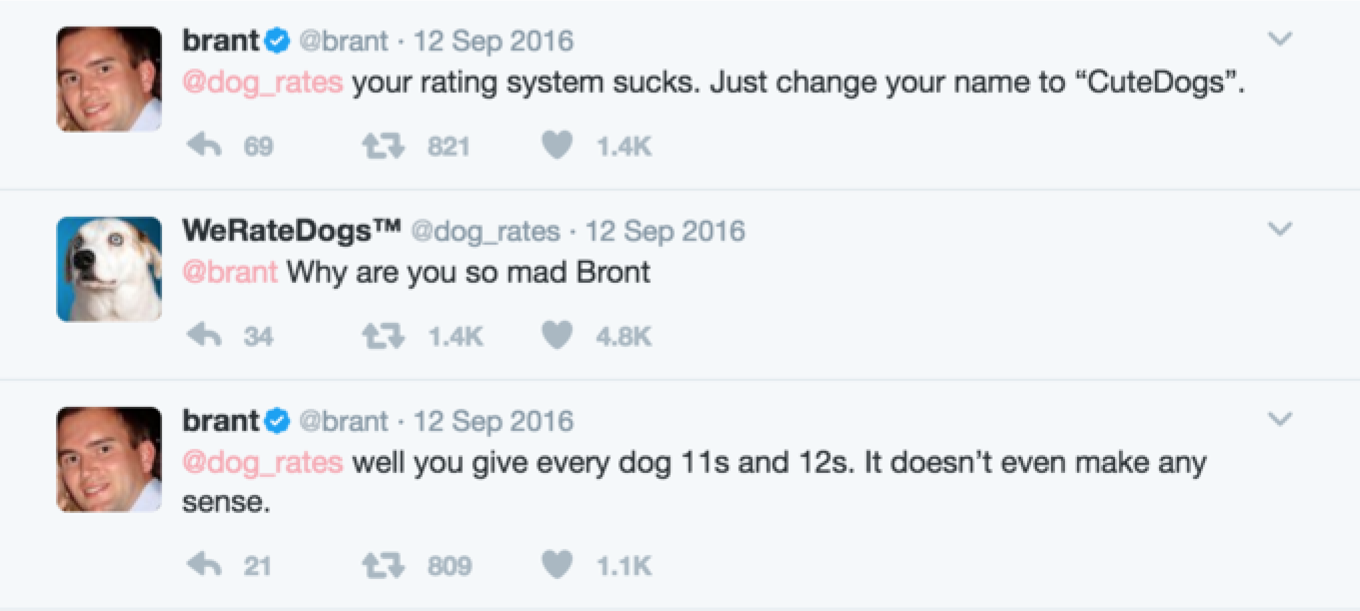
Wrangle Report

Introduction

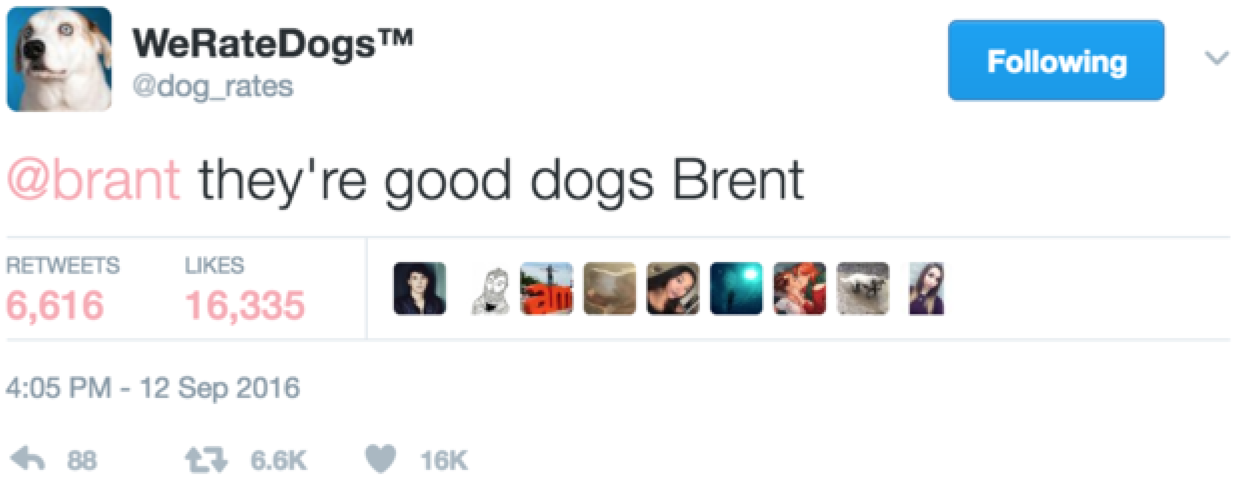
The data set used in wrangling is Twitter archive from WeRateDogs on Twitter. They are (according to their bio on Twitter) the only source for professional dog ratings. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. Why? Because "they're good dogs Brent." WeRateDogs has over 4 million followers and has received international media coverage.



brant @brant · 12 Sep 2016
@dog_rates your rating system sucks. Just change your name to “CuteDogs”.
69 821 1.4K

WeRateDogs™ @dog_rates · 12 Sep 2016
@brant Why are you so mad Bront
34 1.4K 4.8K

brant @brant · 12 Sep 2016
@dog_rates well you give every dog 11s and 12s. It doesn’t even make any sense.
21 809 1.1K



WeRateDogs™
@dog_rates

@brant they're good dogs Brent

RETWEETS 6,616 LIKES 16,335

4:05 PM - 12 Sep 2016

88 6.6K 16K

Project Steps

Gathering

The data used in our projects are from three different sources:

1. **Enhanced Twitter Archive** from WeRateDogs on Twitter which is available to use on this project.

2. **Image Predictions File** that has every image in the WeRateDogs Twitter archive were run through a neural network that classifies dogs' breed.
3. **Additional Data via the Twitter API** to extract retweet count and favorite count.

Assessing

In the assessment procedure, eight quality issues and two tidiness issues were found in the three data sets.

Quality

1. For `df_twitter_archive data`:

- Change timestamp column data type from string to datetime type
- Rating denominator less than 10 (Incorrect values in ratings)
- Remove tweets that are retweets or replies
- Drop retweets and replies columns
- Remove tweets that have no image
- Change wrong dog names to 'None'

2. For `df_image_processing data`:

- Remove IDs that is not in df_twitter_archive data
- Fix dog names in p1, p2 and p3

3. For all three dataframes

- Change tweet_id type to string (object) in the three dataframes

Tidiness

- Merge the `doggo`, `floofer`, `pupper` and puppo in one column df_twitter_archive data
- Join `df_tweet_data` and `df_image_processing` to `df_twitter_archive`

Cleaning

The cleaning process was the most challenging part in the wrangling effort. Using the methodology pattern *define, clean and test*, The quality and cleaning issues were solved.

Udacity's key points were helpful in the cleaning procedure, and they are:

- You only want original ratings (no retweets) that have images. Though there are 5000+ tweets in the dataset, not all are dog ratings and some are retweets.
- Assessing and cleaning the entire dataset completely would require a lot of time, and is not necessary to practice and demonstrate your skills in data wrangling. Therefore, the requirements of this project are only to assess and clean at least 8 quality issues and at least 2 tidiness issues in this dataset.
- Cleaning includes merging individual pieces of data according to the rules of tidy data.
- The fact that the rating numerators are greater than the denominators does not need to be cleaned. This unique rating system is a big part of the popularity of WeRateDogs.

- You do not need to gather the tweets beyond August 1st, 2017. You can, but note that you won't be able to gather the image predictions for these tweets since you don't have access to the algorithm used.