Introduction

I will present in this report an application of what I have learned in data wrangling data section from Udacity Data Analysis Nanodegree program. The dataset that is wrangled is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs.

Project details

- Gathering data
- Assessing data
- Cleaning data

Gathering data

The data for this project consist on three different dataset that were obtained as

following:

- Twitter archive file: the twitter_archive_enhanced.csv was provided by Udacity and downloaded manually.

- The tweet image predictions, i.e., what breed of is present in each tweet according to a neural network. This file (image_predictions.tsv) is hosted on Udacity's servers and was downloaded programmatically using the Requests library and URL information.

- Twitter API & JSON: by using the tweet IDs in the WeRateDogs Twitter archive, I queried the Twitter API for each tweet's JSON data using Python's Tweepy library and stored each tweet's entire

set of JSON data in a file called tweet_json.txt file. I read this .txt file line by line into a pandas dataframe with tweet ID, favorite count, retweet count, followers count, friends count, source, retweeted status and url.

Quality issues

- Tweets with no images
- Dataset contains retweets
- Contents of 'text' cutoff
- Incorrect dog names
- Missing values in 'name' and dog stages showing as 'None'
- Rating numerators with decimals not showing full float
- Tweet ID# 810984652412424192 doesn't contain a rating
- Extra characters after '&'
- Sources difficult to read
- Erroneous datatypes (timestamp, source, dog stages, tweet_id, in_reply_to_status_id, in_reply_to_user_id)

Cleaning data

- Create dog stage variable and remove individual dog stage columns.
- Add tweet_info and image_predictions to twitter_archive table.
- Remove rows where there are no images (expanded_urls).
- Remove retweets where 'retweeted_status_id' is null.
- Change incorrect dog names.
- Change missing values in 'name' from 'None' to NaN (dog stages already covered).

- Fix rating numerator and denominators that are not actually ratings.
- Fix rating numerator that have decimals.
- Remove tweet without rating.
- Change sources to more readable categories.
- Change datatypes of timestamp to datetime, dog_stage to categorical, and tweet_id , in_reply_to_status_id, and in_reply_to_user_id to strings.