As a prerequisite for the completion of the ALX-T Data Analyst Nanodegree, I was tasked with wrangling and analyzing data from the WeRateDogs Twitter account. WeRateDogs is a Twitter account that rates people's dogs with humorous comments about the dog. 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.

The data wrangling process consists of 3 steps: Gathering data, Assessing data and Cleaning data.

Gathering data:

The first data set was gathered by the udacity team and downloaded manually from here.

The second data set was downloaded programmatically from this link. The data in this file consists of image predictions according to a neural network hosted on Udacity's servers.

The third data set was queried from the Twitter API.

Assessing data:

The data gathered was then assessed for quality and tidiness issues using both visual and programmatic methods. The following issues were detected:

Quality issues

Twitter archive data

1. The in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, and retweeted_status_timestamp, doggo, floofer, pupper and puppo columns contain too many null values to be used for analysis
2. The rating_numerator and rating_denominator columns should be float instead of int
3. Incorrect values in rows 55, 313, 516, 695, 763,1712 and 1165
4. The timestamp has the object data type instead of datetime
5. The tweet_id column present in all dataframes should be an object instead of an int
6. incorrect dog name values
7. Some of the tweets are retweets and not original tweets by the account
8. The text column contains the image url which should be a seperate variable
9. Not all the tweets have images

Image predictions data

10. inconsistent number of rows with archive data
11. incomprehensible column names
12. underscores between text in p1, p2 and p3
13. different case style for text in p1, p2 and p3
14. some predictions aren't dogs

Tidiness issues

Twitter archive data

15. the dog stage is one variable and hence should form single column. But this variable is spread across 4 columns - doggo, floofer, pupper, puppo.

16. Information about one type of observational unit (tweets) is spread across three different files/data frames.

Cleaning data:

After assessing the three data frames for issues, the next step is to resolve them so they don't hinder the analysis. The cleaning process is documented in the jupyter notebook.

After cleaning, the data sets are merged and saved in a csv file that is included in this upload