Wrangling report:

The dataset that is getting be wrangling (and analyzing and visualizing) 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. 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.

Data wrangling consists of three steps:

Gathering data, assessing data then cleaning data.

Data Gathering:

The data is gathered from three different resources:

- twitter_archive_anhanced.csv: this data contains some information about tweets and extracted from dog_rate user
- image-predictions.tsv: this data contains the predication of dog breed generated by neural networrk
- tweet_json.txt: this file contains the json objects that have all information of each tweet extracted by tweeter ap.

Data assessing:

The files consist many data issues:

Quality issues:

#Twitter archive file

- some rows are Retweets

- wrong datatypes of columns: tweet_id, in_reply_to_status_id, in_reply_to_user_id and timestamp

- rating_numerator column has values less than 10 and large numbers such as 1176!

- rating_denominator column has values other than 10

- name column has an invalid dog names! such as: the, a, an, officially, old, one, quite

- text column: some contains url.

- missing values in some columns.

- source column contains tag

#image_predication file:

- wrong datatype of tweet_id column.

- jpg_url colume has some diplicated.

- missing values in some columns.

- p1, p2, p3 coulmns contain underscores of names/labels

- some rows get all false dog breed prediction within p1, p2, p3

#json file:

- wrong datatype of tweet_id column.

Tidiness issue:

- dog stage separates in many columns!

- data separates in three different dataframes

Data cleaning:

At this step, I took each data issue and handle it spaitely. At the end of this step the data became cleaned and combined to one dataframe.