Wrangling Report

Introduction:

The dataset that you will be wrangling, 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.

Wrangling Steps:

Wrangling consists of 3 parts:

- Gathering data
- Assessing data
- Cleaning

Gathering Data:

we must import three data files:

- twitter_archive_anhanced.csv

- image-predictions.tsv

- tweet_json

from the resources section.

Assessing Data:

As we can see the files contain many quality issues. Furthermore, it needs to be tied and cleaned:

Quality issues:

#(Twitter archive) table >> t_ar:

- 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, quiet.

- text column: some contains url.

- missing values in some columns.

- source column contains tag.

#image predication >> img_url:

- wrong datatype of tweet_id column.

- jpg_url column has some duplicated.

- missing values in some columns.

- p1, p2, p3 columns contain underscores of names/labels.

- p1. p2. p3 all get false prediction.

#json fie >> data_df:

- wrong datatype of tweet_id column.

Tidiness:

- dog stage separates in many columns!

- data separates in three different dataframes

Cleaning Data:

consist of three steps:

- defining

- coding to solve the issue

- testing

In the jupyter notebook, we will be solving each issue individually, after which, the three data files will be merged.