Wrangle Report

Introduction:

This Project aims to use a learned data analysis lesson to develop good insights about WeRateDogs Twitter account.
Using Python and its libraries, you will gather data from a variety of sources and in a variety of formats, assess its quality and tidiness, then clean it

Project Details:

This Project is divided into:

- Gathering data
- Assessing data
- Cleaning data
- Visualizing Data

1. Gathering data:

Get data from different sources such as

- **Enhanced Twitter Archive**

  The WeRateDogs Twitter archive contains basic tweet data for all 5000+ of their tweets, but not everything. One column the archive does contain though: each tweet's text, which I used to extract rating, dog name, and dog "stage" (i.e. doggo, floofer, pupper, and puppo) to make this Twitter archive "enhanced." Of the 5000+ tweets, I have filtered for tweets with ratings only (there are 2356).

- **Image Predictions File**

  a neural network that can classify breeds of dogs*. The results: a table full of image predictions (the top three only) alongside each tweet ID, image URL, and the image number that corresponded to the most confident prediction (numbered 1 to 4 since tweets can have up to four images).

- **Twitter API & JSON:**

  Back to the basic-ness of Twitter archives: retweet count and favorite count are two of the notable column omissions. Fortunately, this additional data can be gathered by anyone from Twitter's API.

2. Assessing data:

Using methods such as nfo, value_counts, sample, duplicated, and groupby.

3. Cleaning data:

Quality:

1.filter out the retweets and replies:

2.remove unnecessary columns

3.timestamp: from string to DateTime

4.Correct Numerators and Denominators¶

5.Clean text

6.remove NaN url rows

7.Remove img_num col from img_pred_clean

8.remove duplicates in img_pred

9.Create image breed prediction and confidence columns

10.remove unnecessary cols from t_json_clean

Tidiness

1.merge ('doggo', 'floofer', 'pupper', 'puppo') into 'dog_stage'

2.remove unnecessary cols from t_json_clean

3.Create image breed prediction and confidence columns

4.compress all columns into one sheet

4. Visualizing Data:

For making insights easier to read.