WeRateDogs:
Description of Data Wrangling
Summer Cook

Gather

- Programmatic download of online image prediction tsv file
- Programmatic extraction of Twitter data using API and given keys and tokens provided by Twitter
- Provided archive data in csv format
- Data was read to three pandas dataframes

Assess

Once the data was collected, it was assessed for its quality and tidiness.

- **Visual assessment**: Assessment was done visually both in a spreadsheet and in the Jupyter Notebook. I uploaded the archive data to Google sheets and scrolled through it to see its general structure and to help identify any obvious problems. In Jupyter Notebook, I printed the tables and used head() and tail() to look at parts of the dataframe.

- **Programmatic assessment:** Assessment was also done programmatically. This helped me get a better understanding of the size and structure as well as pinpoint erroneous values, as well as identify where differences are found between parts of the data. Common methods are:
  - info()
  - value_counts()
  - shape
  - duplicated()

- **Quality**: There were mistakes in the collected data. For example, retweets were included in the data, and we are not interested in these. Value _counts() showed that many of the values did not describe what they were meant to, for example, values for names such as "a" and 75 should be 9.75. Values such as 420 and 1776 in the numerator of ratings are not incorrect, but they will make analysis difficult.

- **Tidiness:** The structure of the data did not lend itself to easy analysis. Data was contained in three different sources and these needed to be combined into one table. Furthermore, there should be one column for each variable. The data for three variables, rating, stage and breed, was stored among nine different columns.

Clean

The dataframe from the csv file served as the center point of the data. My first goal was to remove unwanted data and fix major tidiness problems. However, certain quality issues needed to be handled in order to address tidiness, such as changing data types in tweet_id.

I decided to remove tweets that had null or multiple values for rating, retweet count, and retweet count, but keep tweets that had null values for name, stage and breed. The number of observations was 2051 at the time of analysis.

The actions taken were:

- Added missing columns to the main table (retweets and favorite tweets) newly extracted from twitter.
- Removed retweets, reply-to's and tweets without photos. Then removed irrelevant columns.
- Corrected 'stage' column. The stage columns were aggregated into one.
- Re-extracted ratings from the text and made corrections. I removed the denominator as it is not necessary for the exercise of comparing. Outliers such as 1776 were removed. Although they are not incorrect, they are not relative measurements.
- Created single breed column from the image table and added it to the main table. The easiest way for this was to separate the tables by whether the most likely value is the first, second or third prediction. I excluded rows that do not have a prediction that is a dog breed. I then combined the remaining breed columns into one, where the most likely dog breed is selected for each row.
- I noticed that many of the missing are preceded by the word "named" in front of them in the text, so I extracted these into a new column. I corrected the names that do not appear to be real names such as "quite" and "a". I examined the text visually to see if I could find additional names. I made a new names column and dropped the old.
- Fixed data types to aid analysis.