THE DATA WRANGLING PROCESS

Gathering:
I gathered data from three sources:

• The Twitter archive #WeRateDogs, a csv file that contains the tweet, rating, dog name, and dog 'stage' in life (such as puppy).

• An 'image prediction' file, or what breed of dog is in each tweet, according to a neural network. I downloaded the image prediction file programmatically from Udacity's servers using the requests library.

• Twitter's API to gather retweet count and favorite count, two columns missing in the Twitter archive. I queried the API for each tweet's JSON data using Python's Tweepy library.

Assessing:
I assessed the data based on quality and tidiness.

• Low quality (‘dirty’) data has content issues such as missing, invalid, inaccurate, and inconsistent data. I assessed removing unnecessary columns, converting data types, making dog names title case, removing entries that were not really dogs, and removing outliers.

• Untidy (‘messy’) data has structural issues. I assessed gathering dog stages from multiple columns into one, creating a ‘prediction’ column (Dog, Maybe Dog, Not Dog), and combining the three datasets into one.

Cleaning:
I converted each assessment observations into action items. Each item was defined, coded, and tested to ensure the problem was fixed. I first addressed missing data, then structural issues, then quality issues. For example, there were lots of ‘rating’ outliers. Dogs were supposed to be rated from 1-10, but many were well above 10. I converted the ones I could to a ‘10’ scale, such as ‘165 out of 150’ became an 11 rating.

Conclusion:
I analyzed the information in the clean, combined dataframe and created initial visuals using Matplotlib in Python.