The first step of this wrangling was to gather all the data. Once the provided archive of WeRateDogs tweets and the image predictions were gathered the Tweety API was used to find extended data for all the tweets in the archive dataset. In total there were 2356 tweets to start with in the archive dating between November of 2015 and August of 2017. With data gathered the assessment could begin. The goals for this project were not to necessarily clean up every item, but to clean up those necessary for understanding the data and looking for trends.

When pulling in the extended data, eleven tweets ids were not found in the Tweepy API. Best guess is these tweets were deleted since the archive was created though I can't be positive that there wasn't an issue with the provided tweet id. Either way, without the ability to get the extended data these eleven tweets were removed from the archive. I also was most interested in the tweets created solely by WeRateDogs, so tweets that they retweeted or replied to were removed.

In the tweets WeRateDogs rates the dogs. In most cases this was on an any number to ten score (part of their charm is a dog can get, for example, a 13/10 score). There were 23 cases where the scale was not ten. It made sense to normalize the scores to ten, by dividing ten by the tweet's denominator and multiplying by the numerator. This allowed the numerator to serve as the rating and the denominator to be deleted.

There were a few columns to clean excess off the content. In the source column for the archive and extended data source code was around the course so this was split off. In the text column of the archive both the link to the photo and rating were included. This is already in other fields, so I stripped that information out.

The date/time field came in as objects, so I changed them to datetime so they could be used later for visualizations. The dog name field had 'None' for many of the dogs. For cleanliness I changed this to NaN so in counts 'None' wouldn't rank high in popularity.

Finally for tidiness I converted the four stages of dogs to a single column and combined the three datasets. From the extended data, the most useful information that wasn't in the archive was the retweet and favorite counts. I merged this into the archive. The image predictions I chose to merge the top rated dog breed for each tweet rather than all the information including non-dog matches.

While there is other cleaning that could be done to do a more complete evaluation of this data, this was sufficient for the creation of a few visualizations that are a good intro to WeRateDogs.