The goal of this project was to gather, assess, and clean data. The dataset that you I was wrangling 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.

Gathering

First, I have gathered some data from 3 different sources:

1. The Twitter archive was given to me. It consisted of basic info about each tweet.
2. Then I programmatically downloaded file image predictions which utilize a neural network to identify which breed of dog is on each picture. The file is hosted on Udacity's servers and was downloaded using Request library and provided URL information.
3. By using tweet ID provided from the WeRateDogs Twitter archive I queried Twitter API for each tweet's JSON data using python library *tweepy*. Then I stored this data set in the text file. The data set was read line by line into a pandas data frame with tweet ID, favorite count, retweet count, followers count, friends count, source, retweeted status and URL.

Assessing

Second, I assessed data:

1. Visually I used Excel and SubLime Text to go through the data sets looking for some tidiness and quality issues.
2. Then programmatically using functions like value counts, duplicated etc. I assessed the data set getting a good understanding of the tidiness and quality issues.

Then I pointed out and documented issues which I thought was the most urgent to be cleaned or fixed.I divided them into tidiness and quality issues.

Cleaning

Third, I cleaned the data:

I have tackled each issue in the following manner: I defined the issue, then code and finally test the code. I also created the copy for each file to isolate gathered data from clean data.

I cleaned the numerator and denominator ranking. Got rid of retweeted tweets which should not be taken into account. Fixed those records where dog names were recorded wrong by the neural network.

Then I merged all the data frames into one and made some analysis on the data.