WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10 and the numerators that could go as high as 1776. Rating score is only a small portion of this Twitter account. Twitter's programming interface allows programmers to extract all the metadata attached to tweets that this user posted. To eventually analyze all the interesting info on these tweets, it was up to me to ensure that data are formatted, typed correctly and that particular data that I wanted are extracted accurately from tweet text.

Since tweet data were initially fragmented in three different files, my initial task was to read the datasets separately into three separate dataframes and then assess the issues associated with the integrity of data, or lack thereof. Once data assessment was finished, I went on to clean the dataset, mostly 'twitter-archive-enhanced.csv' file that contained a substantial portion of interesting information.

In terms of data cleaning, the lowest hanging fruit was fixing wrong data format and types. For example, I changed dog stage info into categories, timestamp into datatime, rating scores into float, and tweetID into string. I also removed retweets, which are not the tweets authored by WeRateDogs account owner and hence should not be a part of our data for analysis. Once I was done with these, I looked at some columns that had empty strings and broken encoding, which I corrected by replacing them with either NAs or correct string information. Our original dataset also had unreadable html tags for 'source' column, which contains information on from which device or platform the user wrote his/her tweets. I also extracted text by removing HTML tags from the column.

The most intensive data cleaning involved dog ratings, stage information, and dog names. Although the dataset contained dog rating scores and stage information, I extracted them once again since the original collector admitted that he/she might have made mistakes while parsing the dataset. Both of these tasks involved extracting numbers and distributing them into several columns, and hence I had to drop old (hence wrong) columns and concatenate new columns into the existing dataframe. We had one dog with a rating numerator of whopping 1776 and another one with zero denominator. Since the presence of such outliers could distort our analysis, I decided to drop those two dogs from our master dataset. I then went on to extract correct stage info and dog names from tweet text. Once I finished these tasks, I finally merged cleaned datasets into a master dataframe I labeled as 'df'.