To create interesting and trustworthy analyses and visualizations of WeRateDogs Twitter data, I wrangled multiple pieces of data to create one clean final dataset.

To start, I gathered three separate pieces of data from three different formats: (1) Enhanced Twitter Archive that contains basic Twitter data (.csv), (2) Image Predictions File (.tsv) that predicted dog breed through pictures, and (3) Additional Data Via the Twitter API (.txt) that includes favorite and retweet count. I converted them all to pandas DataFrames.

Next, I assessed each piece of data and recorded down each content or structural issue I found. This took a surprisingly long time, especially because I had to iterate back while cleaning the data later on. I assessed the `tweet_json` dataframe first, and noticed that there were a couple fields that had no data, or only one value, hence should be removed. Using the isnull() and notnull() functions, I found that there were a lot of tweets that were not original, but rather replies or retweets or quotes. After getting familiar with this dataset, it was easier to dig through `twitter_archive`. It had multiple columns that were also only necessary for replies or retweets, and since the final dataset would only contain original tweets, I knew those columns would later be removed. I remembered that Project Motivation mentioned something about the rating numerators and denominators, so I also checked if those values were extracted correctly. After spending a good chunk of time finding the quality issues, I moved on to tidiness issues. There were a few cases where I found that columns would best be combined. For example, the four columns doggo, floofer, pupper, and puppo in `twitter_archive` should be one column. All the prediction-related columns in `image_predictions` should be narrowed down to the main predicted dog breed. Most importantly, all three datasets should be combined into one master dataset.

Finally, I moved onto the cleaning part of the wrangling. Before cleaning up datatypes and fixing any errors and creating new columns, I wanted to remove the irrelevant columns and rows. This would make joining the tables much more easier before moving onto other quality and tidiness issues. I started with `tweet_json` first because it was the dirtiest table, then moved on to `twitter_archive` to take similar steps. Only then, I cleaned up datatype and formatting (image_predictions finally came into action), corrected errors with numerator and denominator, and created a new column for rating ratio. After cleaning up majority of the quality issues, I cleaned up tidiness issues, namely combining multiple columns into a select few and joining the three datasets together by tweet id. At the very end, I cleaned one more quality issue by removing the columns in the master dataset `twitter_master_archive` not needed for analysis.