My data wrangling effort of the WeRateDogs Twitter datasets was achieved through 3 steps: Gathering, Assessing and finally Cleaning Data.

Regarding the first step, Gathering Data, the request library was utilized to obtain the image_predictions dataset by downloading it from the given url. After having successfully downloaded and stored it to a tsv file, I continued with trying to get the last dataset from Twitter API. Since I encountered some problems with setting up a Twitter developer account, the tweet_json.txt was taken directly from Udacity course page instead so as not to hamper with our project's progression. After that, another library named json was used to extract data from the .txt file, with 3 columns being chosen to include in the final tweet_data dataset: id, retweet_count and favorite_count.

Once all three datasets were finally ready and stored as Pandas Dataframe Objects, I began to move on to the next step: Assessing Data. After having carefully studied the three datasets, both by visual and programmatic assessment via various Python methods such as .info(), sort_values(), describe() and sample(), I found numerous existing quality and tidiness issues that would have to be taken care of.

Firstly, these datasets could be interpreted as separate pieces of one common larger dataset, one that should be joined together by the tweets' id numbers. Additionally, the dog stages' information in twitter_archive table seems to be highly messy, as it was spread out among four columns and therefore was deemed to be improvable. With regard to image-prediction table, using duplicated() method, I found a number of duplicated rows with similar information, which turned out to be retweets of original ones in the same dataframe.

Secondly, twelve quality issues among the datasets were pointed out in this project of mine. These ranges from rows which are in fact retweets to invalid values found in the 'Name' column of the twitter_archive dataframe. Wrong datatypes were detected for many columns, which included timestamp, in_reply_to_user_id, in_reply_to_status_id, tweet_id, rating_numerator, rating_denominator doggo, floofer, pupper and puppo, all of which belonged to twitter_archive table. One especially intriguing error found via the describe() method was the rating denominator value of 0, which was obviously and mathematically not correct. Finally, again in twitter_archive table, extra parts that were not needed, namely the html tags and +0000 text, in source and timestamp column respectively, would also have to be removed.

Last but certainly not least, when all issues had already been identified from the previous step, it was time to sort them out altogether. The first two tidiness issues were pretty straight-forward to fix by making use of duplicated() and merge() method. The remaining one about the dog stage columns, on the other hand, was a little bit more tricky and required multiple steps to clean it. Moving on to quality issues, we pay particular attention to these most tricky two: invalid names in "Name" column and the tweet with the rating denominator of zero. From the Assessing Data section, I noticed that with the exception of "None", all the invalid names tended not to be capitalized. This was confirmed by a simple line of code to retrieve all the non-capitalized

names, which was also used as the list of the invalid names for cleaning purpose. Finally, we could be certain that the rating denominator of 0 was indeed a false value by looking at its tweet link, which was obtained by combining https://twitter.com/dog_rates/status/ with tweet id from twitter_archive table.