I started my data wrangling process by gathering the necessary data. I loaded in the twitter-archive-enhanced.csv file using Pandas' read_csv method. Then, I downloaded the image-predictions file from the internet link using the requests library and saved it to a new file. Finally, to get the favorite_count and retweet_count data, I used my API keys, tokens, and secrets and accessed the Twitter API. I iterated through the tweet_id column of the twitter-archive-enhanced dataset and used a try-except block to save all the json information on each tweet in the tweet_info list. If the tweet_id was deleted, the code would append that tweet's id to the deleted_tweets list. To keep track of the ids, I decided to print each tweet's id once it was appended to the list. Once I had saved all the json information to the list, I iterated through the list and saved the information to the 'tweet_json.txt' file while moving to a new line after each tweet. I then used Pandas' read_json method to load the information to a Pandas Dataframe.

With the gathering step completed, it was time for assessing the data. I used both visual and programmatic assessment to find quality and tidiness issues, and I documented the issues at the bottom of the assessment section. Visual assessment allowed me to identify issues such as missing data in multiple columns and incorrect data in name column ("a", "an", and "this"). On the other hand, running code that assesses data such as the .info(), .duplicated(), and .describe() methods allowed me to detect additional errors such as erroneous data types and invalid ratings.

After documenting the issues, I began the cleaning process by making copies of my original datasets and cleaning the completeness issues. Then, I cleaned the tidiness issues by melting and merging columns. Finally, I tackled the rest of the issues in the data while also re-iterating the assessment step and finding more issues. I grouped these issues in the assessment documentation under the title, "After fixing tidiness issues, new quality issues arise". The new issues mostly consisted of erroneous data types, missing data, and duplicated data due to merging and melting the original datasets. Once all of the issues were fixed, I stored the final, clean dataset in the file, 'twitter_archive_master.csv'.