(Data wrangling)

- Gathering data:The first step in gathering data I open the file and read it. I look at the data information and see the names of the columns and all the information that will help me in the next step. Then I request to read the URL file that contains the image predictions. The final step I open JSON files
- Assessing data: In this section, I notice that there is a lot of missing data that I will delete the missing data in the cleaning section. In the image files, the name columns had wrong names.
- Cleaning data: In this section, I had to find Quality Issues and Tidiness Issues.

in Quality Issues, I find 8 issues:

1/drop a missing data in this two-column (in_reply_to_status_id ,in_reply_to_user_id).

2/convert timestamp to DateTime.

3/replace the (<a href=)from source column.

3/replace incorrect data in the column name.

4/convert tweet_id to string.

6/convert retweeted_status_id to string because it was floating with scientific number.

7/for rating column rating_denominator we just keep the data that is 10 denominator because ratings almost always have a denominator of 10.

8/replace the '_' with space to make it cleaner.

in Tidiness Issues, I find 2issues:

1/merged the two dataframes:df_clean and dfimage.

2/change, the column name to make it more clear.

Final step: Store the clean DataFrame(s) in a CSV file with the main one named twitter_archive_master.csv.