In this project the Data Wrangling steps were done: Gathering, Assessing, Cleaning.

Where in the Gathering step I have:

1- Read the twitter-archive-enhanced.csv file through pandas library, and stored it in twitter_archive_enhanced.
2- Downloaded programmatically the image-predictions.tsv file, and stored it in image_prediction.
3- Reading the tweet_json file and extracting only the 'tweet_id', 'favorite_count', 'retweet_count' columns.

While on the Assessing step, I have assessed the data based on:

1- Displaying the data frames.
2- Investigating their information.
3- Checking for maximum and minimum values, in order to see if there are any outliers.
4- Checking for duplicated entries.
5- Creating 'image_url' column by extracting it from the 'text' column.
6- Investigating invalid values for 'rating_numerator' and 'rating_denominator' columns.

However, the Cleaning process was very interesting and intriguing, as I have started off by cleaning each data frame copy – to preserve the integrity of the original data frames – on it's own before combining them together. Where my sequence was to firstly define the quality & tidiness points that I'll solve, and then work on them with the same order I have put.

Quality issues to be cleaned:

- **twitter_archive_enhanced:**
  1- retweeted_status_id has some non-null values, which indicate the presence retweets.
  2- Drop columns related to any retweets info.
  3- Delete unnecessary columns.
  4- name column has some entries that are not names, ex: a, the, an,....; so we will replace them with Nan.
  5- Correct the values in rating_denominator&rating_denominator.
  6- convert timestamp into datetime type.
  7- Converting ID column to object dtype.

- **tweet_json:**

  1- It has 25 less entries than twitter_archive_enhanced.

- **image_prediction:**

  1- tweet_id type to be changed to object.

  2- Dropping the 66 duplicated entries in the jpg_url column.

Tidiness issues to be handled:

1- Merge the columns (doggo, floofer, pupper, and puppo) into one column 'dog_stage'.

2- Merging all data frames together.

After finishing the cleaning process I stored the cleaned combined data frame into a csv file called twitter_archive_master.

Last but not least, I have reached the last part of the project which is the Data Analysis and Visualization, where I have extracted insights and made some visualization of the dataset.