We have three datasets or three resources of data, let's discuss the issues in each and then the approaches to solve them.

Let's discuss gathering and assessing then go to the resulted quality and tidiness issues:

Gathering:

- Well for the first two datasets Twitter-archive and image predictions I downloaded them manually and the third one using twitter API and storing the data in tweet-json.txt as per the project guide
- I needed to have the ratings with the breeds in one dataset and that required some work as the image predictions dataset was read as one column and the same for tweet json, the separators couldn't be identified.
- So I used regular expressions after taking closer look at them by .head and .info
- I fixed the variation of white spaces as separators in image predictions by using the right regex (one space or more)as delimiter.
- I used read_json for the json file after I changed the extension of it.
- Now the three datasets are ready to be communicated with. I started working on merging the first two datasets using join and matching the tweet_id in both, but that required me to change the column name in one of the data frames.
- Finally, I saved the two data frames of the first two data sets in one .csv file and the third data frame in another .csv file.

Assess:

- I loaded the two new datasets to two data frames and used .info .head .describe to discover them.
- I counted the unique values in the important columns.
- I checked the number of single character in the names column.

→Quality issues:

First: Twitter-archive-enhanced.csv

1- Invalid rows that hold retweets not original tweets.
2- The null values in different columns
3- The duplicate values
4- The extremely out of range ratings nominator
5- The "none's" in the columns of dog dictionary should be recognized as nulls.
6- The invalid dog names that are one letter for example, this should counted either as a name or not.
7- Many rows in the columns of dog dictionary are just nulls.

8- Incompleteness for the columns of the dog's dictionary.

Second: Image-predictions.tsv

1- When all the predictions of one row equal false
2- The unapproximated values of accuracy

→Tidiness issues:

First: Twitter-archive-enhanced.csv

1- The columns of dog dictionary (doggo, floofer, ..) should be one column
2- Information about one type of observational unit (tweets) is spread across three different files/dataframes. So these three dataframes should be merged as they are part of the same observational unit.

Second: Image-predictions.tsv

1-It holds one column which need to be separated due to different delimiters.

The efforts to treat the previous issues:

First: Twitter-archive-enhanced.csv

- For this data set we will start by replacing all "None" with null values
- We will start treating the invalid values of the names column by removing all names with length one.

- We will start treating the rating columns by adjusting them to a reasonable range that will start by separating our data frame to two data frames the first will include the ratings out of value twenty and the other data frame we will add 10 to the denominator the we will divide the rating nominator by the new denominator
- We will drop non needed columns.
- We will merge the doggo dictionary columns to one column.
- We will separate those columns with values from the previous point and analyze them.
- We will fix the format of the tweet_id
- Match the tweet id with that in the other dataset.

Second: Image-predictions.tsv

- For this data set we will start by adjusting the separators to reach each column safely.
- We will remove the null values.
- We will remove those with all predictions = false.