The wrangling process followed this three steps:

1. Gathering
2. Assessing
3. Cleaning.

First, the data were gathered using three different methods. First file was uploaded using the read_csv function. The second one was downloaded directly from the website. The third one was supposed to be downloaded via Twitter API, however I have a problem with accessing the API. Twitter changed the policy and to access those data I would need an account with extended access, I was unable to receive one, so I followed the guidelines in the project description section and downloaded the file with that data.

Secondly, the data was assessed. Both visual and programmatic assessment types were used. I have also demonstrated the approach which is a combination of those two. For example in assessing the problem of wrong numerator and denominator. Because it will be hard and time consuming to scroll through all more than 2000 rows. The suspicious numerators and denominators were assessed programmatically and rows with those values were filtered out. Then they were visually assessed.

Both quality and tidiness issues were identified. The following issues were identified:

1. In `we_rate_dogs_df`:
- presence of the records which were retweets (not the original tweets) - QUALITY,
- incorrect dogs name - QUALITY,
- presence of the columns with the information not needed for the analysis - QUALITY,
- wrong values in the rating_numerator and rating_denominator - QUALITY,
- wrong data format - QUALITY,
- presence of the records which were tweets that doesn't contain the photos - QUALITY,
- the stages of dogs (doggo, floofer, pupper, puppo) are separate columns - TIDINESS.

2. In `tweet_df`:
- presence of the columns with the information not needed for the analysis - QUALITY

3. In `we_rate_dogs_df`:
- inconsistent names of breeds of dogs - QUALITY,
- columns related to the three predictions are combined into one dataframe instead of three different - TIDINESS.

In total, a required number of issues (8 quality issues and 2 tidiness issues were identified.

After making the copies of all data frames, all of those issues were cleaned programmatically. The various techniques have been used like dropping columns and rows, replacing values, regular expressions, updating specific values and changing data types.

Clean data frames were saved to files.