First and foremost, appropriate libraries were imported. As the wrangling goes on, other libraries were still imported.

The data wrangling involves the gathering, assessing and cleaning of the @weratedogs twitter handle. There were 3 datasets to be used and in the end of the data wrangling, the three datasets are merged into one dataset and saved as a csv file. Each dataset had its peculiarity of gathering it. Each also had their uses.

The first dataset to be downloaded was a csv file that was sent to Udacity by Twitter and it was already worked on though not fully. The gathering process here was straightforward. Using the pandas function ‘read_csv’ to download the file into the Jupyter dashboard.

The second dataset was to download image predictions using the requests library. Since it was a tsv file. It was read into a dataframe using \t as the separator.

The third dataset was to query twitters api using the tweepy library. I first had sign up to a twitter account and apply for developer’s account. After Twitter had known my reasons for applying, it granted me access and gave me some secret keys. I used the tweet_id from the first dataset to query the twitter API, writing the content into a text file which was converted to a json file and read in a dataframe using the read_json pandas function. This dataset is required because the first dataset did not have 2 important variables, retweet count and favorite count.

Assessing the datasets was pretty straight forward. Visual assessment was used using Microsoft Excel. Pandas functions were also used like, describe, info, head etc to understand the structure of the data. Doing these, some quality and tidiness issues in the dataset came to light.

Quality issues

1. Dataframe 1: Timestamp column is not of the correct datatype.

2. Dataframe 1: Some rating denominators have wrong values.

3. Dataframe 1: Stop words are being wrongly interpreted as dog names.

4. Dataframe 1: Some dogs have 2 dog stages.

5. Dataframe 1: Some observations are retweets and replies. Only tweets are needed.

6. Dataframe 2: Some images are not pictures of dogs.

7. Dataframe 2: There is a lower/upper case consistency issue concerning the p1, p2 and p3 columns.

8. Dataframe 3: The id column should be in sync (same spelling) with the tweet_id column in other dataframes

Tidiness issues

1. Dataframe 1: The doggo, fluffer, pupper, poppo should be on one column.

2. Dataframe 2: The p1,p2 and p3 columns are merged to get the most likely breed of dog.

Copies of the original dataset was created and the copies were worked on(cleaned) with respect to the quality and tidiness issues raised.