Data Wrangling

- **Gathering**

We first read the database from the CSV file with pandas that we stored in a data frame. Then we used the requests library to read the predictions file in TSV format. And finally to retrieve the API data we read the json file.

- **Assessing**

This process was carried out from dataset by dataset. In each of the datasets, it was a question of identifying quality and order problems to later find a solution in the cleaning section. In our first dataset, the one relating to the twitter archives, we were able to identify several flaws such as, for example, errors related to the extraction of dog names or notes. in the second dataframe, we saw that it was important to rename the columns for a better understanding of the information they convey and on the last dataset we saw that there was non-conformity of the dimensions with the first dataset

- **Cleaning**

To make cleaning, we merge “outer” the three datasets before applying any change. In the new dataset, we firstly make an info() function to observe if we have missing values. Some missing values are filled by fixed values and the others are removed. We drop unnecessary columns like the retweeted status and user id because they have no impact on our study. After this, we replace the timestamp data types from object to datetime. With the observation of name columns, we saw that there are many non-dog that have been recorded. To fix it, we go over our text columns to better extract the names. Also this strategy is being used for bad ratings problems. In our source columns, we observe html code and replace them by better specifying the source with comprehensive values. Also, we make dog classification into one column and rename confused column names for better comprehension of the values. After this, we remove duplicate image urls