Data Wrangling

● Gathering

We first read the database from the CSV file ***"twitter-archive-enhanced.csv"*** with pandas that we stored in a data frame. After that we used the requests library to read the predictions file in TSV format file named ***"image-predictions.tsv"***. The last dataset was json data from Twitter API data named ***"`tweet_json.txt`"***.

● Assessing

In assessing steps we opened each dataset for manual visualization for a quick look of each dataset and their content. The following step was to use python libraries like pandas, numpy, etc to identify quality and tidiness problems, and later try to find a solution in the cleaning section. In our datasets, the one relating to the twitter archives, we were able to identify several issues such as:

Quality issues

1. The retweets publication date need be to renamed
2. Some rating_denominator greater than 10 and other less than 10
3. Wrong data type for tweets date
4. Dogs category issue (doggo or floofer or pupper or puppo)
5. The retweets publication date to rename
6. The name value is just a letter (Ex: a)
7. Some tweets are the representation others (Retweets) need to be cleaned
8. image_predictions name does sense anything things (need to be renamed)

Tidiness issues

1. all datasets should be merged in the right order.
2. In the api dataset imported tweet_id is considered as index
3. image_predictions dataset columns are not ordered

● Cleaning

When importing json dataset the tweet_id was used as an index so we reset that index. Secondly we reordered some columns from the prediction dataset and renamed them. Then we create a new category_dog field with all dogs columns

('puppo', 'pupper', 'doggo', 'floofer'). We have made the three datasets merge with the “outer” option before applying any other change. We drop unnecessary columns like the retweeted status and user id because they have no observations in need. 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. After this, we replace the timestamp data types from object to datetime and also renamed columns with timestamp in their names. We also created new columns named ***fraction*** and ***pattern*** for our analysis by extracting pattern from **text**.