Wrangle Report

Gathering data:
For the first file, twitter-archive-enhanced.csv, I had to download it manually.

For the second file, I had to use the requests library in order to download the data using a link. After that, I saved the information using Python's manipulation files.

For the third file, I used tweepy, the twitter's API, in order to download the information about each tweet. To do so, I run a try-except code using the tweet-ids that I already had from the previous files. The problem is that some of those publications were deleted, because of that sometimes the except line was called. After that, I used the json library to transform the data and extract relevant information from each json.

Assessing data:

- **Quality:**

I focused on the general points for each table, but I didn't find any mistake in the second neither in the third. In the first table I found five incorrect datatypes: in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id and timestamp. Specifically, there could be repeated tweets in the retweet section, so I checked that out.

Then I focused on the null values, most of them were in the columns name, doggo, floofer, pupper and puppo; but it was registered as 'None' instead of NaN. After this I focused my attention in some mistakes that could occur, for example, by seeing some tweets from @WeRateDogs, I noticed that the denominator was more likely going to be a multiple of ten, so I took that into consideration. I also notice that the score is always more than the denominator, so I started eliminating those rows. Most of them were related to the fact that a decimal point was involved or a date was placed in the tweet.

Finally, I took into consideration that the project asked specifically for tweets before August 1st, 2017.

- **Tidyness:**

The main focus in the tidyness part is the merge of the three tables. But aside from that, we needed to merge the cathegories of doggo, pupper, fluffer and puppo, since they were categorizing each dog and were mutually exclusive.

Cleaning data:
In this section I focused on executing all the observations that have been made in the quality section and tidyness section, by using pandas' and numpy's functions.