WRANGLE REPORT

“WE RATE DOGS” TWITTER ARCHIVE

GATHERING

Regarding this file, first of all I had to download the file, since this was a file on hand.

ASSESSING AND CLEANING

This was the file with the highest number of quality and tidiness issues.

Regarding the **structure issues**, it was important from the beginning to identify that the WeRateDogs table contained:

- The data of a variable in 4 columns: doggo, floofer, pupper and puppo. One of the actions I had to do was convert all these columns into a single dog_stage call.
- There are 2 types of information in this table:
    - About the dog itself: tweet_id, name, text, rating_numerator, rating_denominator, dog_stage.
    - About the tweet: tweet_id, timestamp, source, expanded_urls.

Apart from that, I had to deal with a lot of **quality issues**:

- There were a lot of missing values related to replies and retweeted tweets. Indeed, we do not want to analyze nor the replies neither the retweets, so this information tells us which records we want to drop.
- There were some missing expanded_urls too. I completed them all using a method of joining the main url with the tweet_id.
- There were a lot of data type errors that I had to fix:
    - tweet_id was an integer instead of string. Though it is a number, we do not need it that way because it is not made to make calculations.
    - Timestamp was not datetime format.
    - I also made dog_stage a category variable.
- Though there are some rating numerators much higher than others, I explored them but they were not errors at all, but the result of an atypical rating system. Indeed, not every denominator was 10. When I created the insights, I added a new column with a proportional numerator so that each rating would be calculated out of 10.

TWEET IMAGE PREDICTIONS TABLE

GATHERING

This file was hosted on Udacity's servers and I had to download it programmatically using the Requests library and the url provided.

ASSESSING AND CLEANING

This was the table with the fewest quality errors. In this case, I only highlighted two:

- Data format: tweet_id and img_num were integer instead of string. Though they are a number, we do not need them that way because we are not going to make calculations.
- Missing records: we only have the image predictions for tweets until August 1st, 2017.

API TABLE

GATHERING

As we wanted additional data from the Twitter API, I needed to gather each tweet's retweet count and favorite count, extracting only that information from the API table

ASSESSING AND CLEANING

We had one **tidiness issue** in this table:

- API table contains information about the tweet and that information should be with the WeRateDogs Twitter archive table. That is why I merge them (after I splitted the WeRateDogs Twitter archive in 2 tables: one with the dog information and one with the tweets information).

Finally, the last quality issues I found and fix:

- Column name: id column was renamed as tweet_id for consistency with the other tables.
- Data format: id is integer was converted to a string.
- Missing records.

In the end, I joined all the tables in a master dataframe and eliminating the records from which we lacked information (except for the dog_stage).