# Report for the WeRateDogs data wrangling project

## Introduction:

In this document, I will briefly describe how I went through the process of data wrangling in the udacity project "WeRateDogs data wrangling".

This was done in three steps:

- Gathering the data.
- Assessing it.
- Cleaning it.

Each step will be detailed in the following document.

## Gathering data:

In order to practice what was learned during the online course, there was three different ways of gathering data:

### Manually:

This basically means downloading manually a file (csv format) by clicking on a link and then open in it on the Jupyter notebook.

### Programmatically:

- By making a request (using the Requests library) in python to download a file (tsv format) from an url.
- By accessing twitter API using Tweepy library. The downloaded data was in JSON format.

## Assessing data:

After gathering the pieces of data, we needed to assess them visually and programmatically for quality and tidiness issues.

In order to do that, we following functions of pandas library were used :

For visual assessing:

- df.head : to have a quick insight of the first rows of data
- df.tail : to have a quick insight of the last rows of data
- df.sample : to have a quick look at randomly selected rows of data

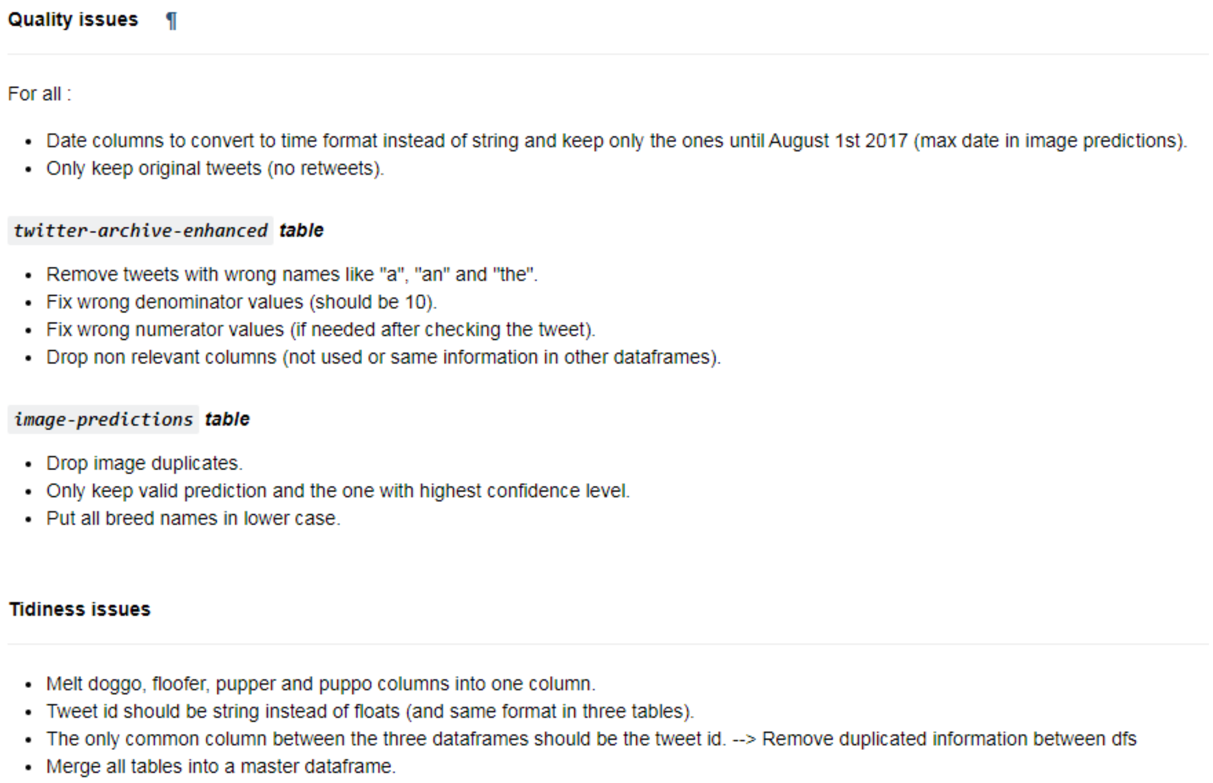
For programmatic assessing:

- df.info : to have formats and null values
- df['column_name'].unique() : to have the array of possibilities of each column
- df['column_name'].value_counts() : to count the number of values of each unique value in the column
- df['column_name'].duplicated() : to check for duplicates
- df['column_name'].is_null : to check null values

## Cleaning data:

After the assessment, some quality and tidiness issues were listed as our homework for the cleaning part.

Here is the list:



### Quality issues ¶

For all :

- Date columns to convert to time format instead of string and keep only the ones until August 1st 2017 (max date in image predictions).
- Only keep original tweets (no retweets).

***`twitter-archive-enhanced` table***

- Remove tweets with wrong names like "a", "an" and "the".
- Fix wrong denominator values (should be 10).
- Fix wrong numerator values (if needed after checking the tweet).
- Drop non relevant columns (not used or same information in other dataframes).

***`image-predictions` table***

- Drop image duplicates.
- Only keep valid prediction and the one with highest confidence level.
- Put all breed names in lower case.

### Tidiness issues

- Melt doggo, floofer, pupper and puppo columns into one column.
- Tweet id should be string instead of floats (and same format in three tables).
- The only common column between the three dataframes should be the tweet id. --> Remove duplicated information between dfs
- Merge all tables into a master dataframe.

We started by making a copy of each dataframe to work on. This cleaned version was the one we will be keeping and storing at the end for visualization purposes.

Then for each action, the plan was to define the task to do, code it and testing right after accordingly to what was done in class.

This iterative process was very useful to advance through small steps towards the final goal.

The most challenging parts were working on the numerators and melting the dog stages in on column.

What was quiet interesting also was that we do not necessarily see all the issues at first but as we start cleaning, we start seeing new things to work on.

## Conclusion:

I personally found the gathering part the most interesting as assessing and cleaning was already discussed in the “Introduction to Data Analysis” part. It was super nice to work on the web scrapping part as it could be a huge source of data. Going to well-known API’s can be helpful and time saving as the gathered data in already not so messy and dirty.