# Data Wrangling Project – Udacity

During this part of the Data Analyst Nano degree, I understood the importance of data wrangling as part of the data analyses process.

Once I loaded the data in the notebook, as advised I firstly did a quick check just by looking at the sheets from all three data documents available for this project. I immediately noticed some parts that need intervention.

**Twitter-archive-enhanced.csv**

I started with the ***twitter-archive-enhanced.csv*** data file creating a copy of the dataset. At the first phase of data observation, I noticed that in the Timestamp column some irregular presentations showing a "+0000" at the end of every timestamp.
As for the analysis, only original ratings are the ones that count we removed the retweeted ratings by converting the values from columns (retweeted_status_id, retweeted_status_user_id, and retweeted_status_timestamp) into null values and then removing the rows that started with "RT @" from the "text" column.
In the next step, I checked the data file using code where I found out that some columns were missing data which then I decided to drop them for the reason that in some of the columns data was missing in most of the rows. I also noticed that some of the columns had the wrong type so I converted them into the right one (tweet_id to string, timestamp to DateTime rating numerator, and rating_denominator to float). Last but not least I removed the Html tags from the source column to make it easier to read.

Tidiness: Only one tidiness issue was found in this data file. Columns "doggo", "floofer", "pupper", and "puppo" represent dog stages and logically a dog can be in only one of them so from the 3 columns I created one named "stage".

**Image-predictions.tsv**

The first quality problem I noticed where the column names which were unclear and not very informative. I changed the column names into more informative names where you could easily know what they represent (p1:dog_breed_1, p2: dog_breed_2, p3: dog_breed_3 and img_num: number_of_images). Just as in the last datafile the tweet_id was not stored in string format, so I converted it into string data type.

**Tweet-json.txt**

This data file was not easy to be loaded as it was a ***JSON*** text file. However, it did not have any big issues and we didn't have many matters to correct. The only thing that we had to correct in this data set was the data type of the column *tweet_id* which we change from object to string.

**Twitter_archive_master.csv**

After we did all these steps it was time to create a new data set that was supposed to be clean. We created the data set based on the tweet_id column which made it possible to match rows based on their ids. This was the other tidiness issues found on this project. After creating a clean data step the next step was to start data analyses and visualization. At the end the file was stored as **twitter_archive_master.csv.**