# Data Wrangling Report

Overall, the wrangling process for this dataset was not complicated but there are some difficulties faced in this project that will be mentioned in this report.

## Gathering:

The gathering process went smooth for the CSV file (archive) and the TSV file (image prediction) using the read_csv and requests library to import the files.

For the tweets interaction file, I was very motivated to work query the Twitter API and start working on the background of an application I really liked . But Unfortunately, I could not obtain access to Twitter API, since the process was taking too long I decided to use the file given by Udacity. I have read the tweepy code provided by Udacity and included it in my submission.

Even after using Udacity’s code, reading the Json file and creating a retweet and favorite count dataframe was one of the most challenging parts of this project.

## Assessing:

Assessing was the easiest part of this project, there were many obvious issues in the quality and tidiness of the dataframes. After visually assessing the dataframes, it was obvious that the entries that had reply and retweet ids are the to be omitted since the project criteria specified that we only want original tweets. No duplicated entries was found in any of the dataframes.

Some quality issues were documented although they could not be fixed, like the missing and wrong dog names. This issues do not affect the quality of the analysis at the end, so they were just ignored. Other identified quality issues were later found to not be an issue at all, like the values for the rating numerator and denominator. As mentioned in the code documentation, the account ratings are not fixed and changes to fit the joke of the tweets.

## Cleaning:

Cleaning process was the most time consuming of the three wrangling phases, however a lot was learnt for me during this process. First copies were made for each one of the dataframes to compare the results of the cleaning process to the original dataframes.

Some columns were created during the cleaning process to make merging easier at the end, like the column dog breed which shows the correct based on prediction values. Using .loc as an if condition inside the dataframes is one of valuable lessons learnt during this project.

Quality issues were more challenging to clean than tidiness for this project, although merging the four stages columns into one needed some research to do.