In this document I will be summarizing my wrangling steps from gathering data from several different sources to assessing data and finally to clean the datasets and store them in a master database (db format and CSV) format.

1. Gathering data:

There were a number of sources to gather the data we required for this project. Below is a list of those sources:

* 1. Twitter\_archive\_enhanced.csv file. This file was provided by Udacity. This dataset contained data for about 2300 entries which were twitter statuses.
  2. Gathering extra data from Twitter API. We used our own consumer key and token to access the tweet statuses and download them. I called this dataset “tweet\_df”.
  3. Downloading programmatically a TSV file containing AI predictions based on the images of the dogs.

1. Assessing data:
   1. I first visually assessed all datasets and found some issues within them. I have listed those issues in the Jupyter Notebook file. There are 2 types of issues with the dirty data that we had in this project:
      1. Tidiness issues
      2. Quality issues
   2. Second, we assessed the data programmatically, using great functions in Pandas.

After a comprehensive assessment of the data, I created a list of 9 quality issues and 8 tidiness issues in total. I have addressed all these issues in my Notebook following the steps of: (Issue, Define, Code, Test).

I have also analysed the data after I cleaned the data frames and stored them in new databases (SQLite databases and CSV files).