**Data Collection**

Data has been collected from three (03) different sources as requested by the project.

* The first table was loaded from a csv file downloaded from the Udacity website.
* The second table was programmatically downloaded using the ‘requests’ library then loaded as a `.tsv` file.
* For the last table, we had to create a Twitter Developer account and retrieve the data using Twitter’s ‘Tweepy’ API. Before interrogating the API, we had to create a list of tweet\_id for the query. That list was created using all distinct tweet\_id already existing in our first two tables. Because of the poor quality of our internet connection, the retrieval of data using the API took several hours for completion. Moreover, some of the tweets couldn’t be retrieved because they were already deleted.

**Data assessing**

The next part of our wrangling process was to assess the collected data. We did both visual and programmatic assessment.

During the visual assessment, we were able to have a quick look at the tables using pandas sample() function and also using an external software (**Numbers** application in our case) : we made some observations for instance the existence of missing information in some columns, or the invalid ratings. Furthermore, we opened some of the urls in the dataset, accessing Twitter website, to understand the rating values for some lines (ratings with a numerator greater than 10). There we were able to understand that ratings with pictures having more than one dog are aggregated values.

The programmatic assessment aimed first at identifying invalid data types and column names. Then we tried looking for duplicated values which we didn’t find in all three datasets.

Another important part of the wrangling process was to understand the relationship behind the 04 different dog stages, we had to make some research on the Internet.

At the end of this process, we were able to gather 12 quality issues and 02 tidiness issues which were resolved during the cleaning process.

**Data cleaning**

The cleaning process started after duplicating the original datasets to avoid unexpected errors. We did manage each group of issues together:

* The data type issues
* The column names issues
* The issue with retweets for which we had to remove the lines that were not relevant for our study
* The issue with expanded urls
* The issue with the ratings
* The issue with dogs’ name, stage, and species

For the extraction of the dogs names using python `regex`, we had to rely on visual observation to identify all possible patterns in the `text` column used to introduce the name of the dogs. For instance 'This is xxx' or 'Meet yyy', etc.

Each issue treated was well explained and detailed then properly defined, coded, and tested at the end.