This data wrangling project is divided in three steps: **gathering**, **assessing** and **cleaning** data.

In the first stage of the process I downloaded manually the csv file 'twitter-archive-enhanced.csv', then I used the requests library to get the tsv file 'image-predictions.tsv' from the web. To gather these two files and transform them into DataFrames was quite straightforward. The third DataFrame I had to create myself by gathering data through an Application Programming Interface (API) instead of using a flat text file. This later task was more challenging, and I started by creating an API in twitter's webpage. Then, I extracted the tweet ids from the csv file as a key to obtain the tweets information in a JSON format, which I wrote in a txt file called 'tweet_json.txt'. I run this application for about half hour until I got all the information concerning the tweet ids. After this, I read this .txt and filter the data by using the regular expressions method to create a dictionary that served as the basis for my third DataFrame.

In the second stage I assessed the data of all DataFrames both visually and programmatically. I checked a few rows visually to understand the content of the datasets, and used methods such as .info() to identify, for example, the number of null values compared to the number of observations (rows). Besides I have also checked if there would be any duplicated rows (e.g. duplicated tweet ids). I found out many quality issues and a few tidiness issues, which I have reported in the jupyter notebook.

The last stage, cleaning the dataset, required also some reassessment of the data because I started finding more quality issues while I was investigating the implementation of the data cleaning process. I reported the cleaning for each of the remarks made in the assessment stage. For example, I removed data that would not contribute to my final analysis, and I compared data between different datasets to decide which ones were correct to keep.