Wrangle report

Written by: Felipe Godoy

There are 3 main difficulties when analyzing a dataset and fixing it in such a way that it is useful to generate reports. The first difficulty is given by what everyone already knows; a dataset (either obtained from a database or created by oneself) usually comes with very low quality, by this I mean that the data comes with errors or with wrong formats. The second difficulty (and that one is able to visualize it even when doing a project of this type) is the facility to get lost and confused among so much created code; it usually happens that among all the steps to perform the sorting, the person in charge gets confused having to modify the data several times, since each modification usually brings other subsequences with which one must face later. Finally, the third error is given by the difficulty of the programming language itself, which often requires external libraries to facilitate the work, and in more extreme cases, various sources of external help that can facilitate things (stackoverflow is one of the main ones).

This work was not absent of these three problems: the need to learn to use a new element such as the twitter API made that much of the code was consulted to external people (in my case, I had knowledge given a previous course), in addition, the fact of being a page that does not usually take data collection very seriously, usually makes the task even more difficult. In addition, the absence of data because they are obtained programmatically does not make things any easier.

I think the best recommendation (both for myself and for others who read this document) is that order is a key factor in data wrangling. An orderly code (easy to recognize variables), clear and understandable documentation and a programmed plan of how to work the database facilitates the collection and cleaning of data considerably, in addition, it will greatly reduce the possible humans that may have to work with data.