Udacity Data Analysis Fundamentals

Project 4: Wrangling Data

Wrangling Report

Iván Muñoz Nunez

About this Project:

For me, it was the most difficult project, by far. Gathering data from three different sources is a challenge for someone who is "starting" in this activity. This combined with the inherent difficulties of getting information from web (linking through Twitter API) made this project the most profitable project I have ever had in data analysis, ever.

The process included an important effort to correct data types (e.g., datetimes), clean useless (for this project) data, merge data from different sources and filtering to get relevant analysis.

I had problems with authentication as Twitter developer profile (after 10 days, I'm still waiting for an answer…), hence I had to use the alternative proposed in the Udacity' page, and "replicate" those code cells.

My most useful tool was testing and commenting code, for remembering what I wanted to get (because is an exhausting work for the memory, sometimes). Establishing different data frames, although is not so convenient, is a good friend to group data faster.

In Wrangling Data' Udacity modules, the tutors commented it's common to come back to cleaning after you started assessing… and they were bloody right (hahan't). It takes a lot of time, resources and moral as well. At least, making graphs is funny for me!

For this reason, some columns still appear with missing or incorrect information; there is a lot of things to improve quality issues, but tidiness issues have been almost completely solved.

I know I'm late with the delivery of this project, but I just started in a new job, and these have been chaotical days. I just want to thank my classmates for helping me, and Luis, for his enormous patience with us!