Modern workflows in data science

Assignment 3

Alexandru Cernat

In this assignment you are going to produce an analytic notebook looking at the change in time in the number of covid cases. An analytically notebook is slightly different from what you submitted last time. It typically includes all the code and results and walks the reader through all the steps of your analysis. These are very useful for work in progress and for communicating with colleagues who know programming/stats.

You are going to use the data from the first assignment with as spin on it. You will use a local Spark server for some of the activities and then write your report.

Your submission will again be a GitHub repo that will include the README and the analytical report.

1. Set up a local Spark server and add the two datasets you used in the first assignment from their GitHub repository. **(20%)**
2. In Spark, merge the two datasets, make a smaller version that includes only: Germany, China, Japan, United Kingdom, US, Brazil, Mexico and calculate the number of cases and rate of cases (cases/population) by country and day. Do two graphs and interpret them: change in the number of cases and change in rate by country. **(25%)**
3. Run a `ml_linear_regression` explaining the log of number of cases using: country, population size and day since the start of the pandemic. Interpret the results. **(25%)**
4. Write up everything in an analytic notebook (pdf Rmarkdown) that shows all the syntax you used, the results and walk the reader through the steps of your analysis. **(20%)**
5. Presentation, presentation, presentation: README gives overview of the project and has session info, report text is easy to follow, graphs are easy to understand and properly formatted. **(10%)**

Top tips:

- for setting up a Spark server check out chapter 2 from the reading for this unit
- before moving the two dataset in Spark I recommend to make some changes to the data while still in R *(this is because it's easier to do these things using all the tidyverse commands and it might take too long to figure out how to do in Spark)*: make the data with the number of cases in long format, define the time variable as a date and also make another variable that says the number of days since the start of the data collection *(this is because you can't include date variables in `ml_linear_regression` but you can use relative time as a numerical variable)*. **Please do all the data manipulation in Spark otherwise.**
- the `broom` package can help you get coefficients out from the regression object. The `texreg` package can make a nice table. Please note that these don't always work as expected with the Spark objects and how they work depends on the version of R, sparklyR and Spark used