INF442 : projet informatique 8
Accidents corporels

Adrien Ehrhardt

January 27, 2020

1 Software

Contrary to the TDs, you are allowed to use any software you'd like. It could even be a good idea to mix them, *e.g.* by performing the ETL (Extract, Transform and Load) tasks with a high level language like R or Python (Dataiku would also be suitable for this task; there is a free version), and the computationally heavier stuff in C++. Again, it's all up to you, as long as it is an equivalent of 500 lines of C++ code (at your appreciation).

Bonus points if you're able to mix them in the same file / package / library (see *e.g.* `reticulate` and `Rcpp` for R, or `rpy` and `Cython`).

2 Scenario

You are a Data Scientist at the *Ministère de la Transition écologique et solidaire, chargé des Transports*. Your mission, should you choose to accept it, is to provide policy makers with data insights on traffic accidents. Although at school, you often end a lab / course / project with *only* these insights, *e.g.* "we show that we are able to predict very accurately the probability of survival on the *Titanic*", these are useless without actionable levers. In real life, you will have to suggest actions based on these insights. Regarding traffic accidents for example, you might be asked by your boss, who has a fixed budget, if (s)he should rather strengthen the controls of alcohol levels (which means employing more police officers) or build a median made of concrete on roads not yet divided to avoid front crashes.

3 Problem description

First, you will have to transform the raw data to a format suitable to your subsequent analysis. **Example:** there are several databases that need to be joined to access the vehicles' information.

Second, you will reuse and / or implement any algorithm seen during the course that is suited to the analysis you want to perform on the sanitized data.

Third, you will present your analysis: do not focus too much about the technical aspects of your method(s), bring the overall reasoning, the results, and possible actions forward.

The subject is purposely open-ended: do not get lost into details, *e.g.* a perfect formatting of the data, do not hesitate to make hypotheses / simplifications (in which case you must state them clearly).

4 Data at hand

Your primary data source should be `https://www.data.gouv.fr/fr/datasets/base-de-donnees-accidents-corporels-de-la-circulation/#_`. In particular the first file, description-des-bases-de-donnees-onisr-annees-2005-a-2018.pdf explains what all other files contain and is a good entry point to the problem.

For this project, you should at least use the 2018 files. Depending on your analyses and at your appreciation, you might need / want to look into previous years' data.

Note that at the end of the page, you will find the *Contributions communautaires* which are mostly already transformed / cleaned data. While you might not need / want to use these directly, the explanations given by their authors might be interesting and applicable to your analyses, as well as the *Réutilisations* section, which might give you ideas of analyses and / or data visualizations.

5 Bonus

As you can see from the link provided above, the data is available freely. Although open-sourcing your work might not be part of your daily job (highly job-dependent), it might be a good idea to "give back" to the community by making the result of your *Projet Informatique* publicly available, *e.g.* by posting on the original page a *Contribution communautaire* and / or a *Réutilisation*, by sharing your analysis *via* a blog post, a Github repository, etc.