Data Scheme

Technical Test for Marco Faggian

Combining Datasets

Two datasets can be comined if they have at least a variable (column) in common.
This variable can be matched in order to make a connection (join operation) between the information in the first dataset's rows and those in the second dataset.

Name	Surname	Date of birth	Commune
Marco	Dupont	18/11/1996	Paris
Jean	Delacroix	02/02/1944	Caen

Commune	Depart.	Region
Toulouse	Haute-Garonne	Occitanie
Caen	Calvados	Normandie
Paris	Paris	Ile de France

Join

Name	Surname	Date of birth	Commune	Depart.	Region
Marco	Dupont	18/11/1996	Paris	Paris	Ile de France
Jean	Delacroix	02/02/1944	Caen	Calvados	Normandie

The Saint-Denis problem: Duplicates in lieux.csv

The Places file ‘Lieux.csv’ contains a duplicate row in the column « Commune », corresponding to the entry ‘**Saint-Denis**’, which corresponds both to the place in **La Reunion** region and the one in the **Ile de France** region.

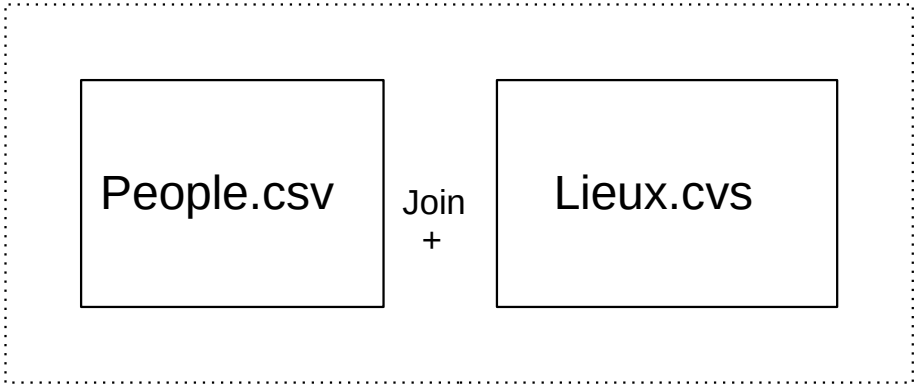
Due to a lack of provided data, we cannot distinguish the people coming from Saint-Denis in La Reunion from those of Saint-Denis in Ile de France.

For this reason, I decided to create a new row that includes all the people living in the two regions. This row has ‘Seine Saint-Denis/La Reunion’ as department and ‘Ile de France/La Reunion’ as region.

Removing the Saint-Denis rows would not be a wise choice, as their department and region might be determined in a second time thanks to the use of another new dataset.

The Algorithm

The algorithm proposed in the Git repository updates a database (initialised to empty) by adding the information about the people crossed via a join operation with the information about the places.



Add

Database

Name	Surname	Date of birth	Commune	Depart.	Region
……	……..	………..	……	……	…….

Writer : counting people per department/region