Workshop : manipulating data 2/2

Data and Algorithms for Public Policy

Timothée Gidoin

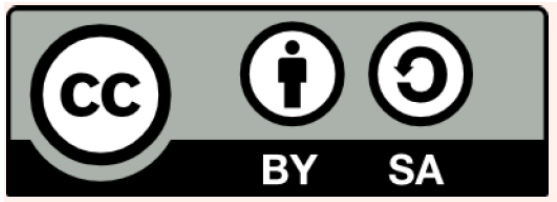
Sciences Po, 2019-10-04

Before we start

Slides : https://gidoin.github.io/sciencespodata/lesson5_dataworkshop.html

Sources : https://github.com/Gidoin/sciencespodata/

The content of this presentation is partly inspired by other presentations made by Datactivist team. I warmly thank them and notably Joël Gombin for his help.



Before we start: reminder

Date	Session #		Teacher
05/09	1	Introduction to data policy	Simon Chignard
12/09	2	Open Data (data as a policy)	Timothée Gidoin
19/09	3	Public sector algorithms: « with great power comes great responsibility »	Simon Chignard
26/09	4	Workshop data manipulation #1: data preparation and analysis	Timothée Gidoin
03/10	5	Workshop data manipulation #2: datavisualization	Timothée Gidoin
10/10	6	Workshop Machine-learning #1 : introduction	Jean-Marie John Mathews
17/10	7	Use Case #1: Predisauvetage (Guest: Antoine Augusti)	Simon Chignard
24/10	8	Explainable algorithms: why ? how ?	Simon Chignard
07/11	9	Use case #2: Predictive policing (Guest: Bilel Benbouzid)	Simon Chignard
14/11	10	Workshop Machine-learning #2 : Fairness	Jean-Marie John Mathews
21/11	11	Workshop Machine-learning #3 : Explicability	Jean-Marie John Mathews
28/11	12	Oral presentations of coursework	Simon Chignard

Before we start: reminder

- **Midterm Exam** :
  - **25% of your total grade**
  - By group of 2 students
  - Data manipulation, analysis visualisation exercise based on open data
  - To be submitted before 25/10 11:59 pm

Instructions :

- You are a policy advisor working in a cabinet (*you select the Ministry and country you want*) and you need data to advise properly your Director of cabinet regarding a new public policy to design
- Due to the emergency of the situation - The Minister has a general policy speech in the Parliament in 5 days where he will detail its policy roadmap, and your policy recommandation may be included - you can't wait to get internal files or data coming from the administration
- Thus, you can only use open data as any other common citizen !

SO EXCITING !

Before we start: reminder

Instructions :

- Find an interesting dataset opened by govt or local authorities or public companies (not necessary a French one). Topics : fiscality, education, R&D...
- Analyze the data through pivot table or some functions we covered
- Cross the data with another open dataset to enrich them and strenghten your analysis (vlookup / index.match).
- Analyze your results and produce a graph or a map (Khartis) displaying the key results
- Bonus : provide a real policy recommandation based on your results (if relevant)
- Explain the limits of the data you worked with (quality, scope, granularity..) and the data you would have needed to improve your work

You will write a 2-pages (graph/map included) note, **1** page of annex **MAX** allowed, for your Head of cabinet with the main result you got by analysing this data, a draft of a public policy based on this data (it doesn't need to be original, just consistent), the limits of your analysis and data needed to strenghten it.
Send attached the CSV/XLS files you worked with

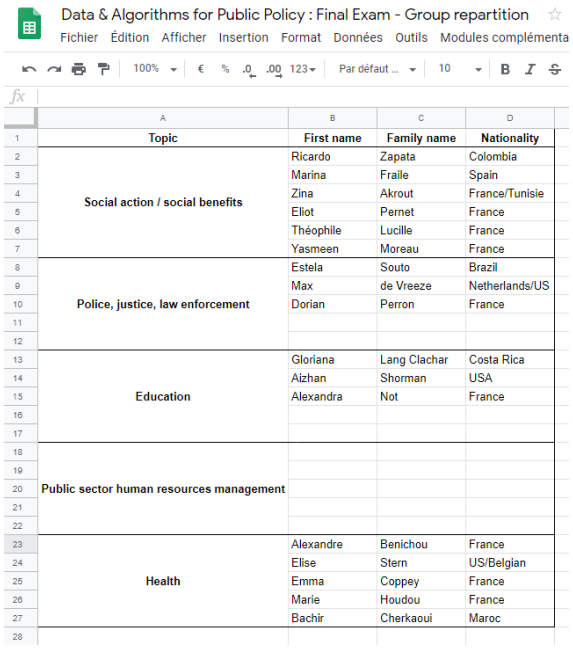
Before we start: reminder

- **Final Exam** :
  - **75% of your total grade**
  - By group of 4 students
  - 10-pages paper on the analysis of 3 uses cases (outside France) in one of the following topics: social benefits, police / justice, education, public sector human resources
  - 1 oral presentation (15 min) during the last session
  - 1 Medium blog post (1,5 pages) to present your findings
  - Evaluation of final exam: 50% quality of the analysis, 25% oral presentation, 25% quality of Medium blogpost
- 5 topics :
  - **Social action / social benefits**
  - **Police, justice, law enforcement**
  - **Education**
  - **Public sector human resources management**
  - **Health**

Before we start: Final exam

You decided last session of the "algorithm" of groups composition and of topics repartition for the final exam :

- You constitute by yourselves your group (5 students, at least 1 French student & 1 foreign student in each group)
- You decide by yourselves the selected topic



Data & Algorithms for Public Policy : Final Exam - Group repartition

Fichier Édition Afficher Insertion Format Données Outils Modules complémenta

	A	B	C	D
1	**Topic**	**First name**	**Family name**	**Nationality**
2	**Social action / social benefits**	Ricardo	Zapata	Colombia
3		Marina	Fraile	Spain
4		Zina	Akrout	France/Tunisie
5		Eliot	Pernet	France
6		Théophile	Lucille	France
7		Yasmeen	Moreau	France
8	**Police, justice, law enforcement**	Estela	Souto	Brazil
9		Max	de Vreeze	Netherlands/US
10		Dorian	Perron	France
11				
12				
13	**Education**	Gloriana	Lang Clachar	Costa Rica
14		Aizhan	Shorman	USA
15		Alexandra	Not	France
16				
17				
18	**Public sector human resources management**			
19				
20				
21				
22				
23	**Health**	Alexandre	Benichou	France
24		Elise	Stern	US/Belgian
25		Emma	Coppey	France
26		Marie	Houdou	France
27		Bachir	Cherkaoui	Maroc
28				

Let's go back to data

Data manipulation through spreadsheet : Vlookup

- ISF data are interesting but there is one key variable missing...

population !

- Find and download a dataset enabling you to get demographic data given city per city
- Look at its structure and select the variables / columns that you want
- Open a now tab in your ISF spreadsheet, copy/paste the columns from your tab "patrimoine" and add a new column where you will collect the number of inhabitants associated to each city that have ISF taxpayers. For this, you need to use the function *rechercheV* / *Vlookup*

Data manipulation through spreadsheet : Index&match

- Interesting, but not sufficient to compare the ISF paid per inhabitants, why ?
- Now you want to create a new column with only the number of adult inhabitants (>= 18) per city
- First create a column that sums the pop > 18 in each city
- Then, instead of vlookup function (also functioning), use **index&match** or index/equiv (*in French*)
- Now you can get the ratio of number of adult inhabitants that were ISF licence payers
- What are the cities/districts with the highest proportion of adult ISF payers ?

Paris 07, Neuilly, Paris 06, Paris 08...

Data manipulation through spreadsheet : Index&match

STXT	Commune	nombre de redevables	patrimoine moyen en €	impôt moyen en €	Pop totale recherche	Pop totale index.equiv	Pop > 18	Ratio ISF	Ratio ISF > 18
75107	PARIS	5 795	5 749 206	28 387	52512	52512	44479,3	13,03%	13,03%
92051	NEUILLY SUR SEINE	5 834	6 026 825	27 701	60580	60580	48299	12,08%	12,08%
75106	PARIS	4 004	4 576 866	23 029	40916	40916	35651	11,23%	11,23%
75108	PARIS	3 148	4 445 890	21 425	36453	36453	29803,7	10,56%	10,56%
75116	PARIS	14 361	4 469 350	21 972	165446	165446	136002	10,56%	10,56%
75101	PARIS	833	3 587 881	19 208	16252	16252	13952,4	5,97%	5,97%
92064	SAINT CLOUD	1 358	3 272 585	14 975	30193	30193	23054	5,89%	5,89%
75104	PARIS	1 333	3 731 819	18 510	27487	27487	23640,2	5,64%	5,64%
75105	PARIS	2 720	3 057 294	13 114	59108	59108	50765,1	5,36%	5,36%
94067	SAINT MANDE	821	2 976 369	13 085	22731	22731	17741	4,63%	4,63%

- How many cities of more than 20 000 inhabitant don't have at least 50 ISF payers ?

More than 100

- What are the biggest French cities (Dom Tom excl) that don't have at least 50 ISF payers ?

Évry-Courcouronnes, Venissieux, Sarcelles..

Data manipulation through spreadsheet : pivot table

- Open a new tab
- Insert a *pivot table* / *tableau croisé dynamique* based on the scope of the data where you have all your previous columns
- Insert in your pivot each variable and split them this way :
- Filters : Pop+20y / Cities / Departments / Pop
- Lines : Region
- Value : Sum of number of ISF taxpayers / Average Property / Average ISF paid / % of adult ISF taxpayer
- Filter by removing Guadeloupe, Réunion, Martinique

Data manipulation through spreadsheet : pivot table

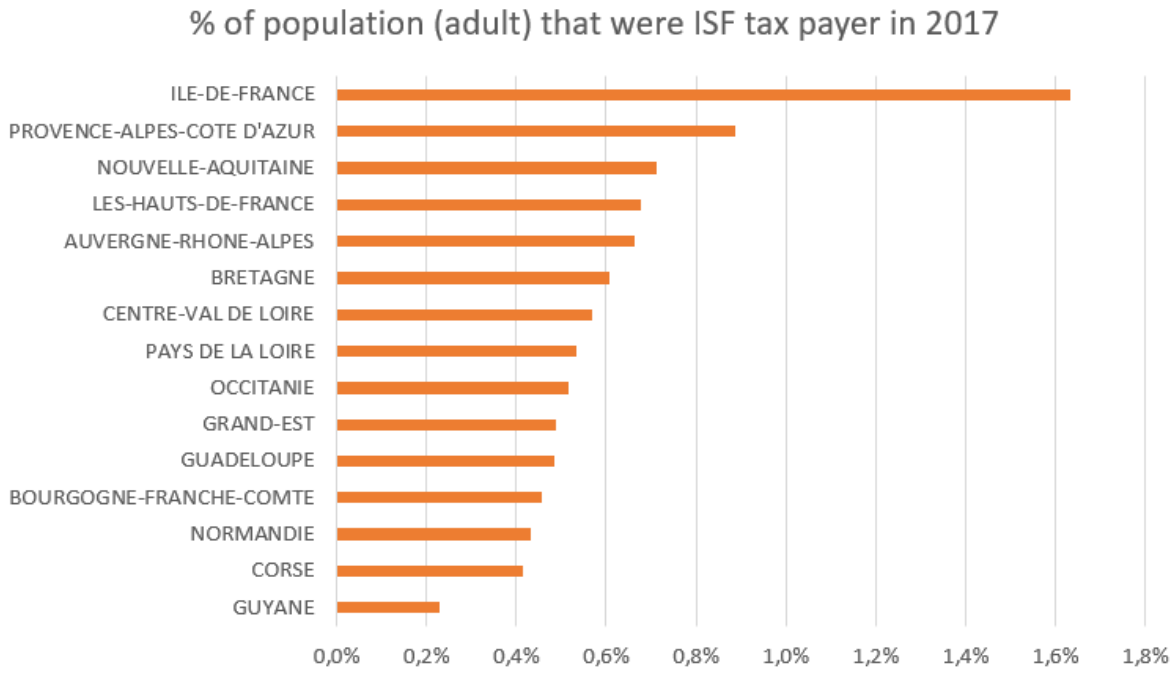
You should get the following pivot table :

Pop > 18	(Tous)
Départements	(Tous)
Commune	(Tous)

Étiquettes de lignes	Somme de nombre de redevables	Moyenne de patrimoine moyen en €	Moyenne de impôt moyen en €	Moyenne de Ratio ISF > 18
AUVERGNE-RHONE-ALPES	12 674	2 532 304	9 438	0,66%
BOURGOGNE-FRANCHE-COMTE	2 100	2 588 279	10 198	0,46%
BRETAGNE	3 200	2 489 823	8 810	0,61%
CENTRE-VAL DE LOIRE	2 538	2 446 726	9 841	0,57%
CORSE	395	2 631 619	11 437	0,41%
GRAND-EST	6 047	2 718 509	10 713	0,49%
GUADELOUPE	101	2 603 757	10 851	0,49%
GUYANE	116	2 729 434	14 089	0,23%
ILE-DE-FRANCE	112 790	2 576 788	10 014	1,63%
LES-HAUTS-DE-FRANCE	6 267	2 721 583	10 867	0,68%
NORMANDIE	2 128	2 423 130	9 249	0,43%
NOUVELLE-AQUITAINE	9 543	2 501 082	9 509	0,71%
OCCITANIE	8 202	2 429 669	9 048	0,52%
PAYS DE LA LOIRE	5 169	2 549 588	9 174	0,53%
PROVENCE-ALPES-COTE D'AZUR	19 477	2 442 852	9 445	0,89%
Total général	**190 747**	**2 553 124**	**9 849**	**1,01%**

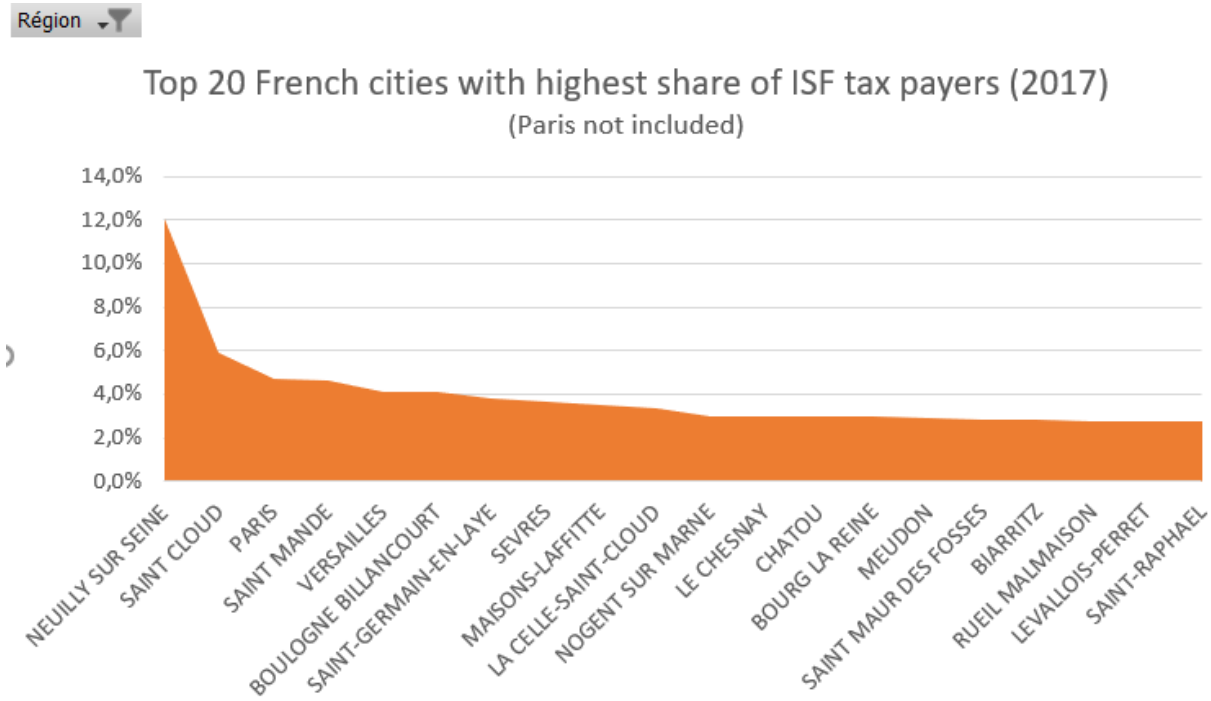
Data manipulation through spreadsheet : pivot table

And the following pivot graph :



Data manipulation through spreadsheet : pivot table

You can create as many graphs as you want, so amazing !



Data manipulation through spreadsheet : pivot table

Coffee break : 5 minutes \o/

Datavisualisation

Missing Migrants Project

- Missing Migrants project, led par the IOM, UN agency, records every known incidents that imply migrants that are missing or that died through their migration
- IOM **agregates in one public database different sources of information** (fishermen, NGOs, media, coastguards...) and add a "source quality" note to assess the reliability of the source
- This work of census by an international organisation allows to give credit to the statistics collected and to shed the light on this issue. This is a way to **raise awareness among the public opinon and to make it political**. Collecting data is often necessary to quantify an issue if we want it to be taken into account in the public debate
- Illegal imigration is not the only topic where it doesn't exist any official statistics. See "**missing number projects**" for more examples (homeless persons, number of suicides in some professions...).

Missing Migrants Projects

Missing Migrants Project

> Être ignoré de la statistique, c'est être exclu de la cité, de la citoyenneté // Being ignored by statistics means being excluded by the society, by the citizenship

Maryse Marpsat, statistician and sociologist (INED)

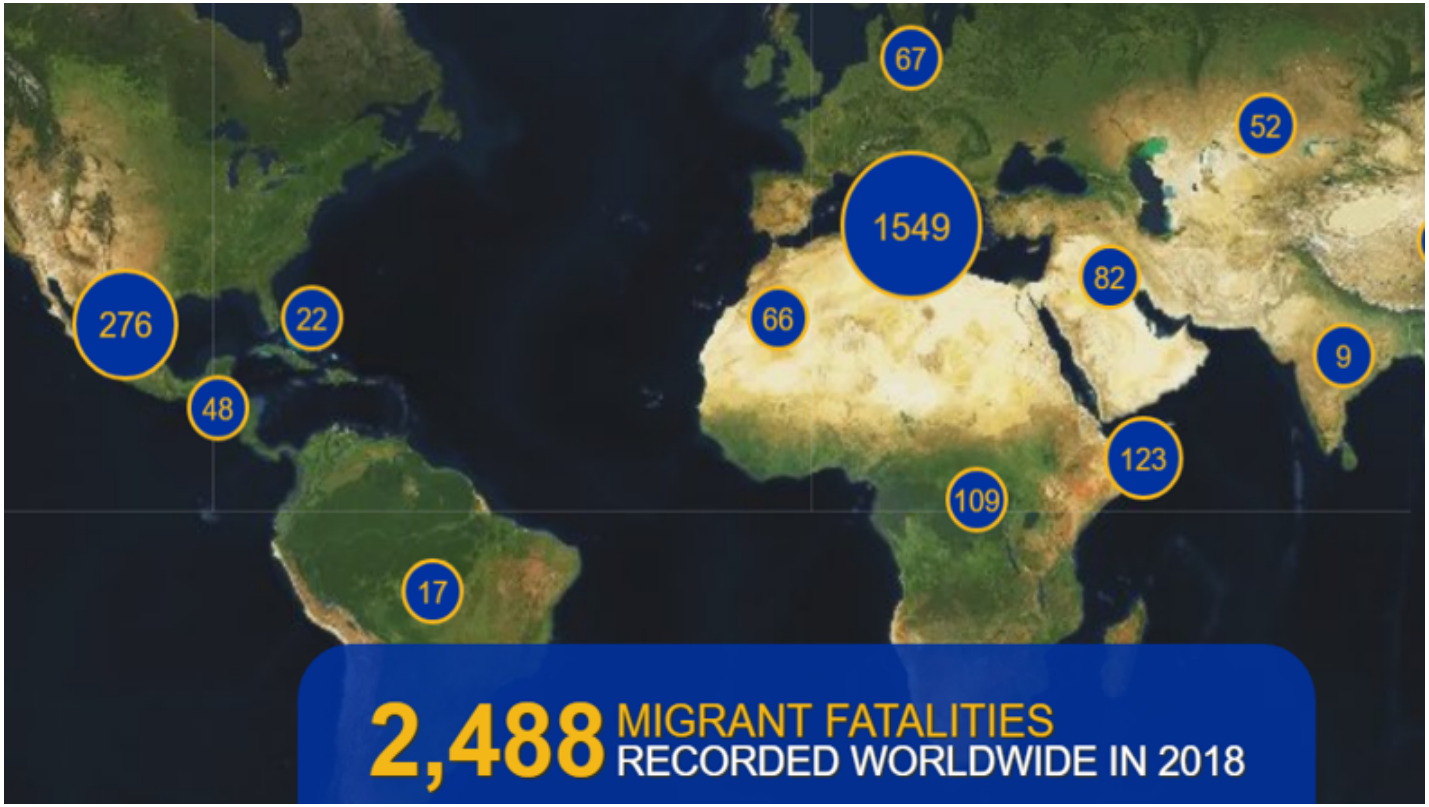
Not collecting data, is a very political decision

> EU state members document thoroughly the situation of their own population and the entrance in their territory of foreign citizens but, in their own statistical process, they don't care about people died by tempting to cross their borders. **By ometting to count those deaths, they condemn them to invisibility.**

Antoine Pécoud, sociologist

But is OIM, by counting missing migrants, such a do-gooder ? Some experts consider that, in fact, this work supports government.. French article on the topic

Missing Migrants Project



Missing Migrants Projects

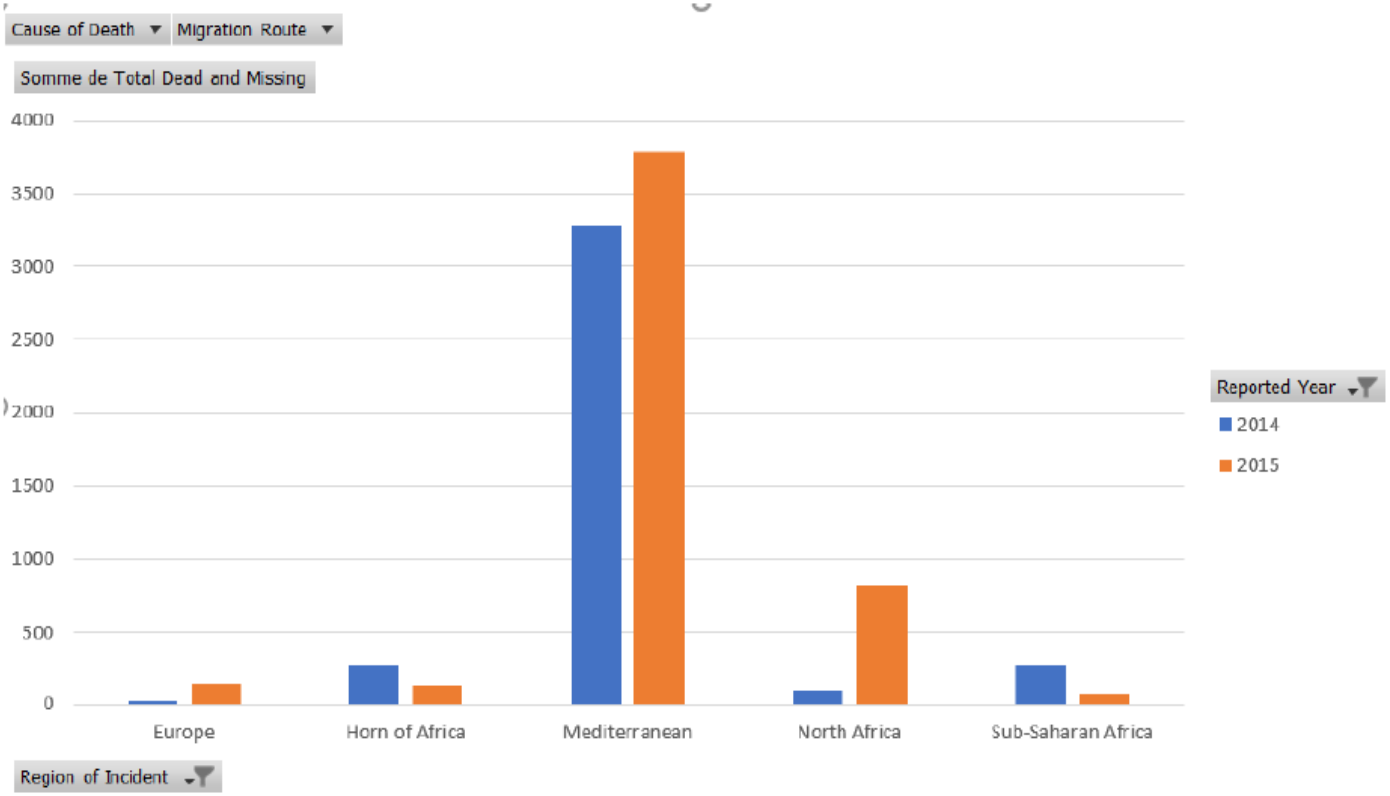
Practical exercice : missing migrants

- Download OIM Missing migrants dataset

	A	B	C	D	E	F
1	Region of Incide	Reported Ye	Total Dead and Missi	Cause of Death	Migration Route	Source Qual
2	Mediterranean	2015	750	Drowning	Central Mediterranean	4
3	Southeast Asia	2014	750	Starvation, Violence, Dehydration		4
4	Mediterranean	2016	550	Drowning	Central Mediterranean	4
5	Mediterranean	2014	500	Drowning	Central Mediterranean	4
6	Mediterranean	2016	459	Drowning	Central Mediterranean	4
7	Mediterranean	2015	400	Drowning	Central Mediterranean	4
8	Mediterranean	2016	339	Drowning	Central Mediterranean	4
9	Mediterranean	2015	307	Drowning	Central Mediterranean	4
10	Mediterranean	2016	288	Drowning	Central Mediterranean	4
11	Mediterranean	2016	255	Drowning	Central Mediterranean	4
12	Mediterranean	2014	251	Drowning	Central Mediterranean	4
13	Sub-Saharan Africa	2014	251	Drowning		4
14	Mediterranean	2016	245	Drowning	Central Mediterranean	4
15	Southeast Asia	2015	243	Mixed		4
16	Mediterranean	2014	240	Drowning	Central Mediterranean	4
17	Mediterranean	2015	222	Drowning	Central Mediterranean	4
18	Mediterranean	2014	217	Drowning	Central Mediterranean	3
19	Mediterranean	2015	202	Drowning	Central Mediterranean	4
20	Mediterranean	2014	200	Drowning	Central Mediterranean	1

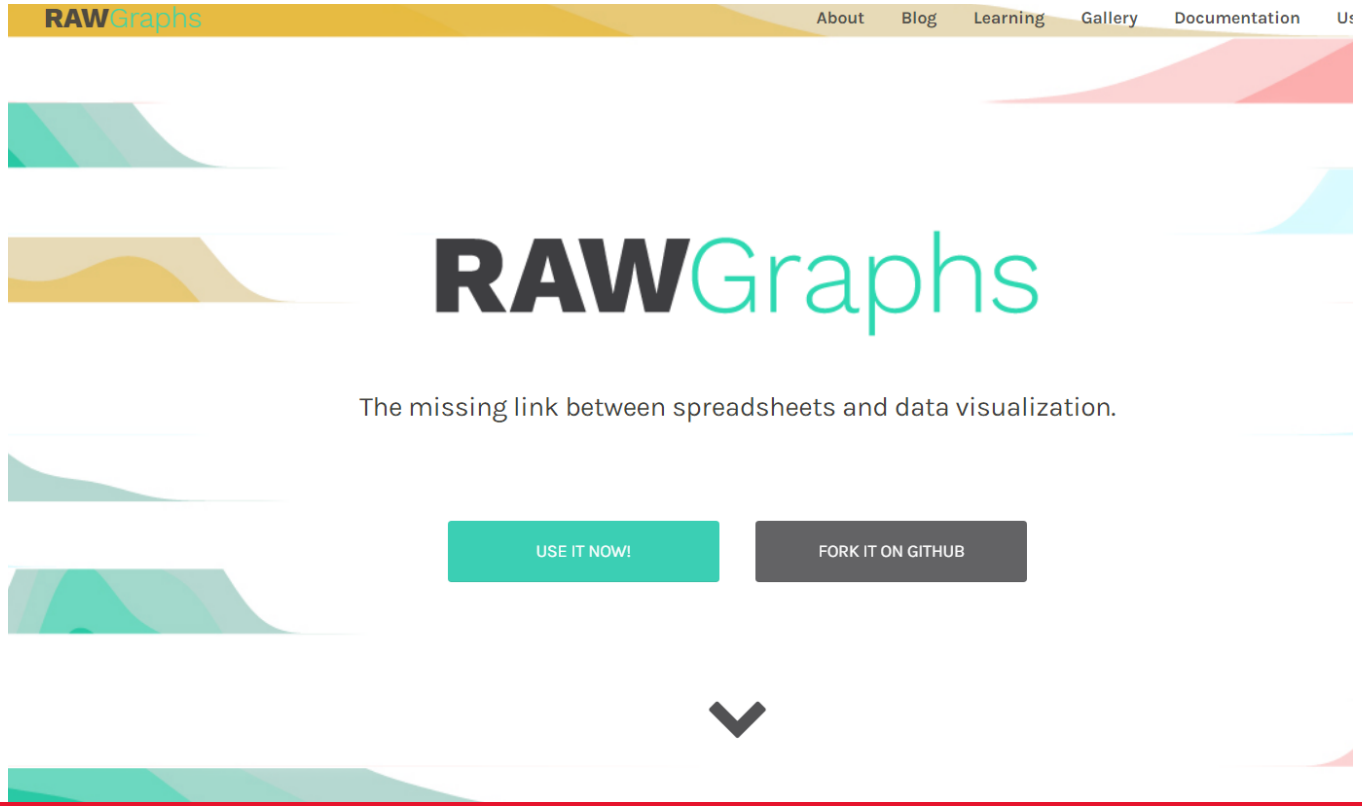
Practical exercice : missing migrants

- Create a pivot table and dynamic pivot graphic

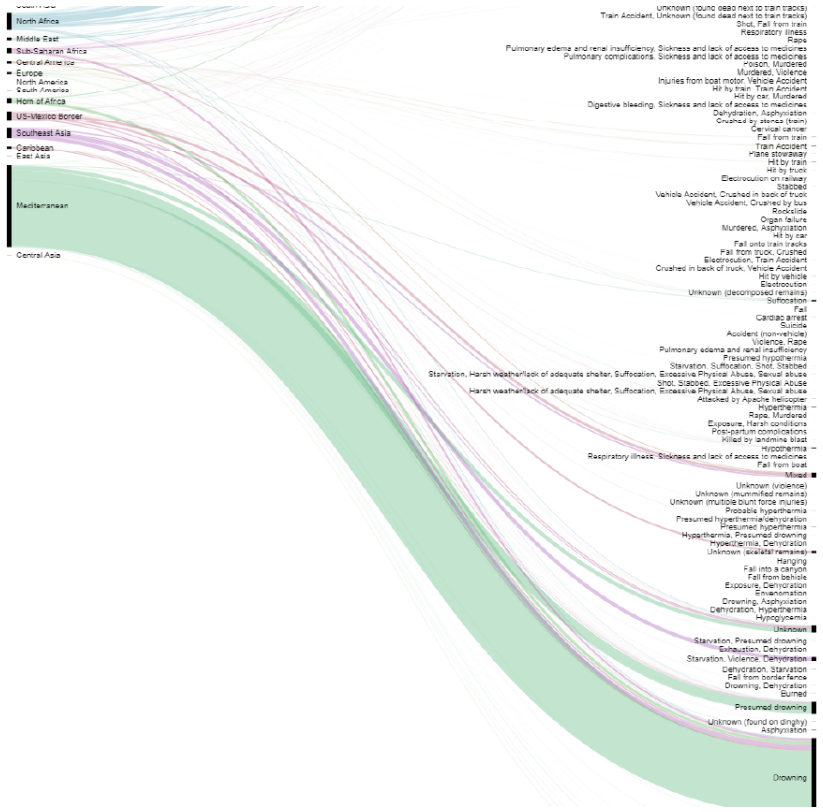


Practical exercice : missing migrants

- Once your data are cleaned, go on RawGraphs



Practical exercice : missing migrants



Missing numbers : a blog dedicated to missing data

Missing numbers

Dataviz : other ressources

Panorama des formats et des outils de data visualisation

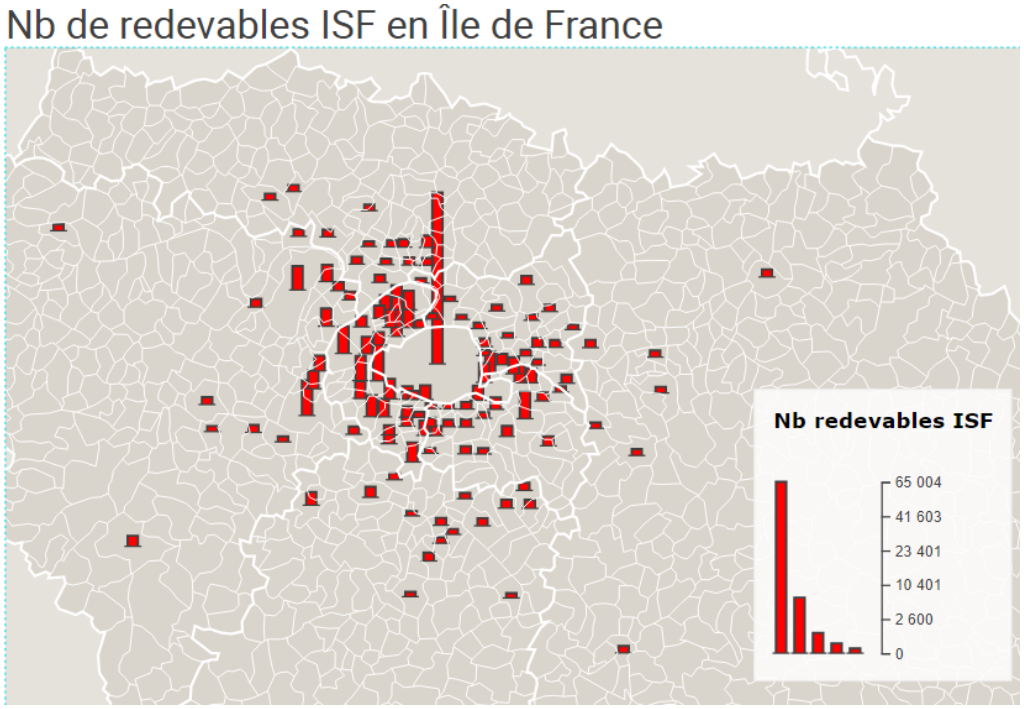
Magalie Dartus, Datactivist

Dataweek, 2019-06-24

Mapping data

Mapping data through Khartis

Represent in Khartis the number of ISF tax payer (per city) in Ile-de-France in 2017



So happyyyyy togetheeeeeer !

Thank you !

Contact : timothee.gidoin@sciencespo.fr