Who's who in civic tech?
A data analysis project

Tatiana de Feraudy

Ironhack Paris
DAFT 0410

June 2023

Table of Contents

Introduction

The word 'civic tech' was first used in France in 2013, when the Knight foundation released its report on the investment and political potential of these technologies[1]. From that date until the Open Government Summit held in Paris at the end of 2016[2], activists, entrepreneurs, investors and public policy-makers participated in creating a "civic tech" ecosystem in the country and giving meaning to the word. The great diversity of actors involved in the ecosystem (NGOs, companies, researchers, public agencies, local and national governments…) is what makes it interesting, but also what makes it difficult to define what 'civic tech' means.

For my PhD research in political science and sociology, I study the people, organizations and networks that have contributed to defining civic tech in France and making digital citizen participation a public problem[3]. This research is conducted using mainly qualitative methods, i.e. semi-structured interview and participant observation, as well as through the collection of "grey" literature (reports, information and communication material produced by governmental entities, NGOs, companies, etc.) and media on the topic.

The goal of the project presented in this report was therefore to use data analytics to enrich my research and better understand the civic tech ecosystem in France. This general goal was divided into two more precise objectives:

1. Use quantitative methods to 'decenter' my approach and gain a broader, more holistic understanding of civic tech. Since data analysis methods require a large amount of data, this also implied thinking about what "big data" could be collected on this topic, and how it could be analyzed.
2. Develop a framework that could be shared and replicated with other researchers or analysts. From data collection and cleaning to database construction and data analysis, I made an effort to think about how the approach could be replicated. This would namely allow draw comparisons with other related ecosystems (for instance the edtech, health-tech, fintech, and more generally the different start-up and tech for good ecosystems in France).

[1] Patel Y. et. al., *The Emergence of Civic Tech : Investments in a Growing Field*, Knight Foundation, December 2013, URL : https://knightfoundation.org/features/civictech/.

[2] Barack Obama's administration was the first to use the terms 'open government' to describe a program of governmental actions aimed at making public action more transparent, collaborative (with civil society, including the private sector) and more participative (i.e. involving citizens). This program is detailed in Obama B., *Transparency and Open Government. Memorandum for the heads of executive departments and agencies*, January 21st, 2009, The White House, URL : https://obamawhitehouse.archives.gov/the-press-office/transparency-and-open-government.

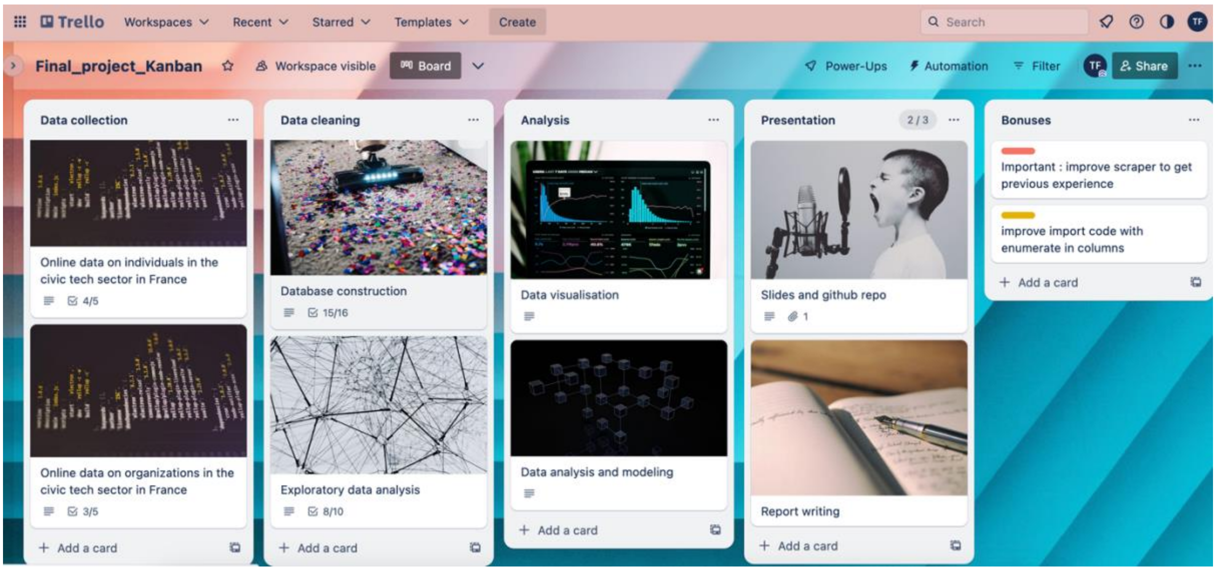
The Open Government Partnership is an international organization that promotes this program around the world. Founded in 2011, it now includes more than 70 countries, whose governments commit to voluntarily improving their practices. Objectives are stated in multi-year plans which are reviewed by independent auditors. In December 2016, the French government, as co-chair of the initiative at the time, hosted the OGP international summit in Paris.

[3] A 'public problem' is an issue that is recognized as requiring 'public' action (as opposed to an individual or a 'private' problem). Unemployment, water pollution, childcare or citizen participation can be defined as public problems. The concept is interesting because it invites us to see the problems that public policies address not as issues that get a response because they are more important than others, but as "constructions". When studying public problems, we focus on the work done by different actors (activists, companies, lobbies, individual and collective groups in government agencies, researchers…) to define the problem and acceptable solutions, identify who is responsible, and convince the public sector to take action. For more detail, cf. Erik Neveu, « L'analyse des problèmes publics. Un champ d'étude interdisciplinaire au cœur des enjeux sociaux présents. », *Idées économiques et sociales* 2017/4 n°190, p.6-19.

Project planning

Building a database and analyzing data on the civic tech ecosystem in France required different types of tasks. These can be divided into data collection, data cleaning, data analysis, data visualization and data modeling activities. Although we have separated them in this planning and in the report, they were not conducted in a strict chronological manner.

In my opinion, this is one of the key lessons learned through this project (and more broadly during the bootcamp). Data collection is informed by the expected design of the database, data cleaning is conducted in several steps based on the requirements for data analysis and modeling, and data visualization happens at every stage of the project, although it takes different shapes.



Data collection
- Online data on individuals in the civic tech sector in France 4/5
- Online data on organizations in the civic tech sector in France 3/5
- + Add a card

Data cleaning
- Database construction 15/16
- Exploratory data analysis 8/10
- + Add a card

Analysis
- Data visualisation
- Data analysis and modeling
- + Add a card

Presentation 2 / 3
- Slides and github repo 1
- Report writing
- + Add a card

Bonuses
- Important : improve scraper to get previous experience
- improve import code with enumerate in columns
- + Add a card

Figure 1. Screenshot of the kanban board for the Who's who in civic tech? Ironhack projet, (4 june 2023).

Data collection – 2 days

Data collection was conducted in several steps:

1. Identifying relevant and available data sources.
2. Scraping LinkedIn profiles and Pappers' websites to produce individual files (per person, per company, per NGO).
3. Producing a framework and tools to increase the dataset and automate data collection. While we started with a limited amount of data, we progressively increased the size of our dataset, and now have the tools to keep doing it if necessary.

Data cleaning – 3 days

Data cleaning is by far the most time-consuming step of data analysis. As we mentioned earlier, cleaning data implies thinking ahead and imagining the possible uses of the data. To summarize, different types of activities were conducted to improve different elements of the data.

1. Structure: scraped data is oddly structured, with nested dictionaries and dataframes, repeated information, and some source-specific information that is not useful for analysis. Cleaning it is necessary to identify the ideal structure to keep as much detail as possible while achieving a clear database structure.

2. Readability: data from websites is produced in order to be read by humans. However, we need the data to be as raw as possible to be able to conduct analyses on it. We need to remove special characters, harmonize data types, and more generally make data readable for non-humans.

Data analysis and visualization – 3 days

These two tasks are described together here as they often go hand in hand. While our analysis code is often divided into "data discovery", "data cleaning", and "data analysis", these often happen at the same time. For instance, data cleaning often implies visualizing the distribution of data, the weight of different categories, or the importance of correlations between features. Our data analysis and visualization steps aimed at answering two main questions.

1. Who are the people in the French civic tech ecosystem?
2. What characterizes the organizations associated with this ecosystem?

The data analysis we conducted is based on the data collected, so it has its limitations. For people, we worked on their education, their latest professional experiences, the languages they speak, and how they define themselves. For organizations, we mainly analyzed information on creation date, city, founders/ directors, whether they are still active, their field of activity and their financial performance (sales, revenue…).

Data analysis and visualization was done through three different tools:

- Jupyter notebooks coded in Python, for table construction, exploratory analysis, basic plotting of relationships and distributions and reformatting/ exporting new datasets (libraries: pandas, json, numpy, matplotlib, seaborn, and sqlalchemy)
- MySQL Workbench, which allowed to organize the tables and their relationships in a database format and to formulate queries which gave new insights on the data.
- Tableau (desktop version) to produce visual renditions of data analysis.

Project tracking, documentation, presentation and archiving – 2 days

Some of the tasks to produce data analysis projects are hidden behind our Trello board. These include the project tracking activities (creating plans, making to-do lists), discussions with colleagues and teachers to decide on the best options[4], documentation of activities (keeping readme and gitignore files updated for the github repository, reviewing and organizing notebooks for readability), organizing files and repositories, and of course preparing presentation materials (reports and slides). Stating them here seemed important as part of this work is aimed at sharing the process with other researchers who may be interested in these tools.

[4] Many thanks to Thomas, Andy and Elnara for their advice and patient support during the bootcamp, to Lucien, Robin, Katrina, Axèle, Adèle, Matthieu, Ismael, Basat, Huyen, Romain, Louis, Yuchen and Hye-jin for the group work, park lunches and laughs, to Emma and Choco for keeping me sane.

Data collection

Challenge 1. How do we collect data when there is no data?

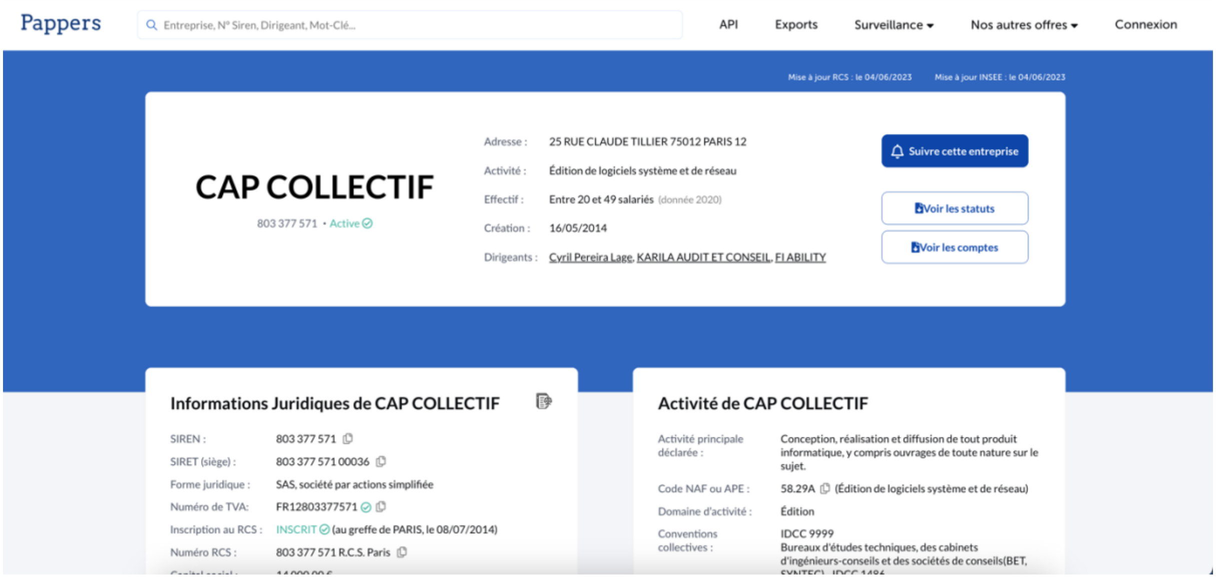
One of the reasons why I started by PhD a few years ago is precisely that there is very little data about the civic tech ecosystem in France. The information about the topic is mostly produced by the actors themselves, and is mostly promotional material, whether it is in press articles, blogs or conferences. Moreover, since companies in the field are relatively young and unstable, they are not very keen on disclosing information, and neither are individuals.

The qualitative data collected during my PhD research provides a very good understanding of individual trajectories and professions, but only for the 50 individuals with which I was able to conduct semi-directive interviews. I also studied organizations and networks based on participant observation during three years in the field. However, I was often faced with questions that I didn't have an answer to. One of them was: "what is the best technology?" For this one, my answer was usually that it depends on what you want to do with it.

Another question was: how do we know which company will survive? Although I am still not convinced this is something that can be predicted (otherwise, no company would every fail), or that it is something that should be predicted (cf. self-fulfilling prophecies), I thought it would be interesting to see if we could model whether civic tech (or more general, innovative "for good" startups) would survive or not based on the data we have. We can also ask the question in terms of whether people will stay for a long time in the field, or not.

Although the modeling part of the project is yet to be done, the data collection was informed by this objective, as well as by the availability of data. There were three sources of data:

- my PhD research provided me with a list of people and organizations that I could focus on. These lists were namely constructed by analyzing press articles and extracting the people who talked about civic tech in the press and the organizations that were mentioned;
- the LinkedIn[5] profiles of individuals related to the civic tech ecosystem;
- the information on companies and NGOs provided by the Pappers[6] website.



Mise à jour RCS : le 04/06/2023 Mise à jour INSEE : le 04/06/2023

CAP COLLECTIF

803 377 571 • Active ✓

Adresse : 25 RUE CLAUDE TILLIER 75012 PARIS 12
Activité : Édition de logiciels système et de réseau
Effectif : Entre 20 et 49 salariés (donnée 2020)
Création : 16/05/2014
Dirigeants : Cyril Pereira Lage, KARILA AUDIT ET CONSEIL, FI ABILITY

Informations Juridiques de CAP COLLECTIF

SIREN : 803 377 571
SIRET (siège) : 803 377 571 00036
Forme juridique : SAS, société par actions simplifiée
Numéro de TVA: FR12803377571 ✓
Inscription au RCS : INSCRIT ✓ (au greffe de PARIS, le 08/07/2014)
Numéro RCS : 803 377 571 R.C.S. Paris

Activité de CAP COLLECTIF

Activité principale déclarée : Conception, réalisation et diffusion de tout produit informatique, y compris ouvrages de toute nature sur le sujet.
Code NAF ou APE : 58.29A (Édition de logiciels système et de réseau)
Domaine d'activité : Édition
Conventions collectives : IDCC 9999
Bureaux d'études techniques, des cabinets d'ingénieurs-conseils et des sociétés de conseils(BET,

Figure 2. Screenshot of the Pappers page concerning the civic tech company Cap Collectif (june 4th, 2023)

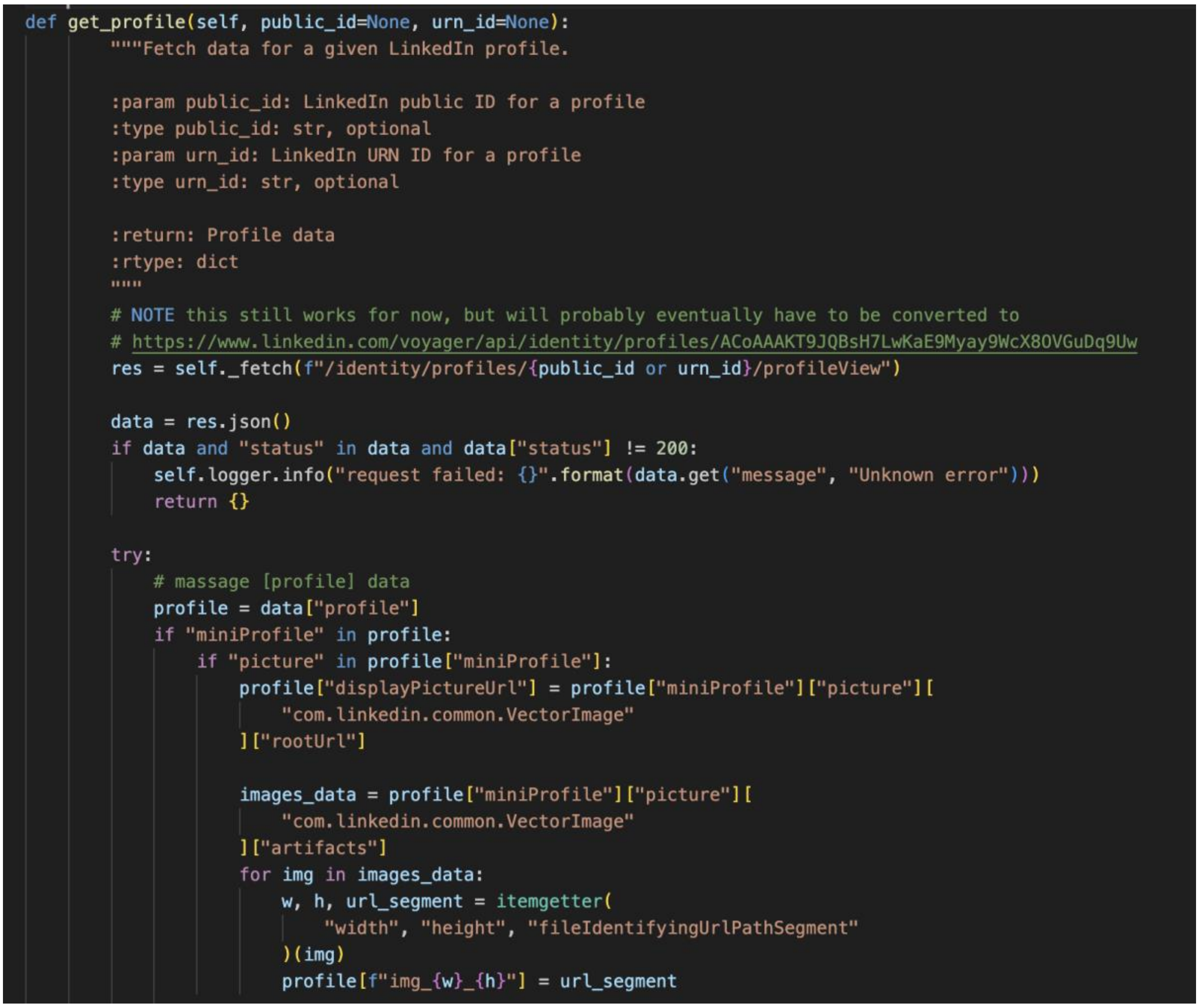
[5] LinkedIn is an online social network focused on professional relationships (cf. https://www.linkedin.com)

[6] Pappers is a French company that makes open public data on businesses available on an online platform (https://www.pappers.fr/a-propos). Downloading the data, however, requires paying to use the API.

Challenge 2. How do we collect data when people don't want us to?

The two online data sources identified don't allow to download data for free. In addition, the requests library does not function for these websites (it returns a 403 forbidden error), even for single page requests. Fortunately, I had the chance to work on web scraping with research colleagues a while ago, using the R language. For this project, I adapted several different scripts:

- Two scripts written namely by Constantin Brissaud and Aurélien Goutmesdt. The first script, in R, uses the Selenium library on R to automate the process of researching a name on Google and saving the urls that the search returns (for the two first result pages). It produces a csv file with a column of names and a column of urls. The second script (in Python) reads through the csv file and extracts the urls including linkedin.
- The LinkedIn scraper developed by Tom Quirk (https://github.com/tomquirk/linkedin-api). Although some of the endpoints have changed and some of the code required adaptation, this repository (recommended by Constantin) provides Python code to extract information from LinkedIn. However, LinkedIn blocked me after a certain number of requests. The code is now adapted to include sleep time, but for the moment my account is still blocked.
- For the Pappers website, I chose to manually save the source pages of the organizations I wanted to study (the code to automate this is under development).

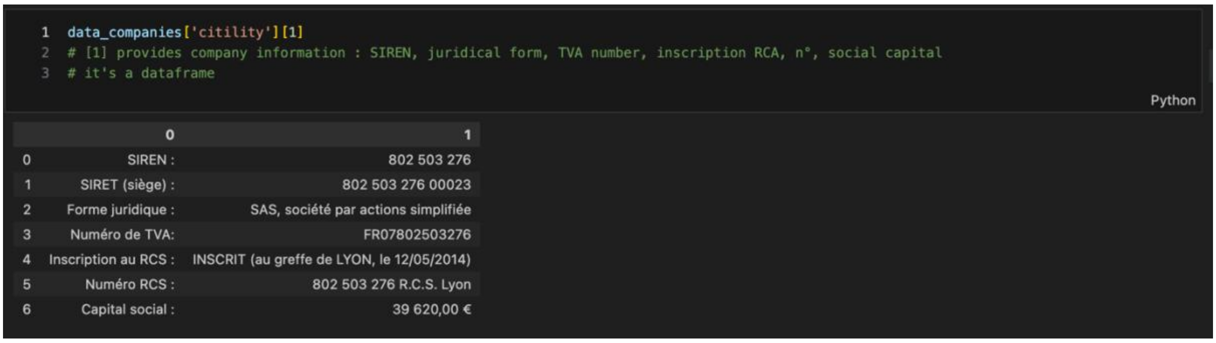


```python
def get_profile(self, public_id=None, urn_id=None):
        """Fetch data for a given LinkedIn profile.

        :param public_id: LinkedIn public ID for a profile
        :type public_id: str, optional
        :param urn_id: LinkedIn URN ID for a profile
        :type urn_id: str, optional

        :return: Profile data
        :rtype: dict
        """
        # NOTE this still works for now, but will probably eventually have to be converted to
        # https://www.linkedin.com/voyager/api/identity/profiles/ACoAAAKT9JQBsH7LwKaE9Myay9WcX8OVGuDq9Uw
        res = self._fetch(f"/identity/profiles/{public_id or urn_id}/profileView")

        data = res.json()
        if data and "status" in data and data["status"] != 200:
            self.logger.info("request failed: {}".format(data.get("message", "Unknown error")))
            return {}

        try:
            # massage [profile] data
            profile = data["profile"]
            if "miniProfile" in profile:
                if "picture" in profile["miniProfile"]:
                    profile["displayPictureUrl"] = profile["miniProfile"]["picture"][
                        "com.linkedin.common.VectorImage"
                    ]["rootUrl"]

                    images_data = profile["miniProfile"]["picture"][
                        "com.linkedin.common.VectorImage"
                    ]["artifacts"]
                    for img in images_data:
                        w, h, url_segment = itemgetter(
                            "width", "height", "fileIdentifyingUrlPathSegment"
                        )(img)
                        profile[f"img_{w}_{h}"] = url_segment
```

Figure 3. Code snippet from Tom Quirk's LinkedIn scrapper for profile scraping

Challenge 3. How do we fix the structure in scraped files?

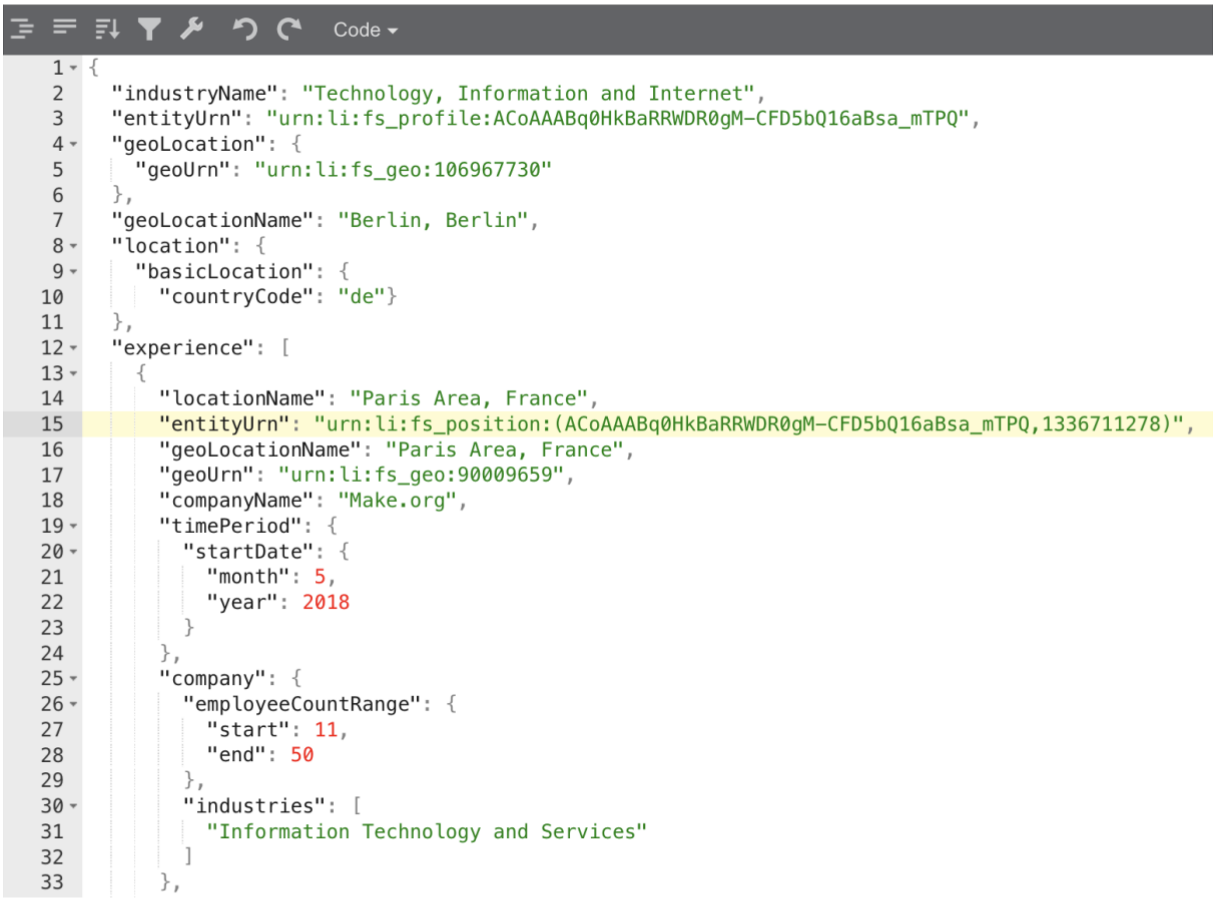
The LinkedIn scraper returns json files, while the collection of Pappers' pages returns html files. Both formats need to be transformed to create tables of information that can be used in a database. Our main challenge, in both cases, was to handle nested information (nested dictionaries and lists in json files and nested dataframes in the html files).



```python
data_companies['citility'][1]
# [1] provides company information : SIREN, juridical form, TVA number, inscription RCA, n°, social capital
# it's a dataframe
```

	0	1
0	SIREN :	802 503 276
1	SIRET (siège) :	802 503 276 00023
2	Forme juridique :	SAS, société par actions simplifiée
3	Numéro de TVA:	FR07802503276
4	Inscription au RCS :	INSCRIT (au greffe de LYON, le 12/05/2014)
5	Numéro RCS :	802 503 276 R.C.S. Lyon
6	Capital social :	39 620,00 €

Figure 4. Example of a nested dataframe in the dictionary created from the html files.

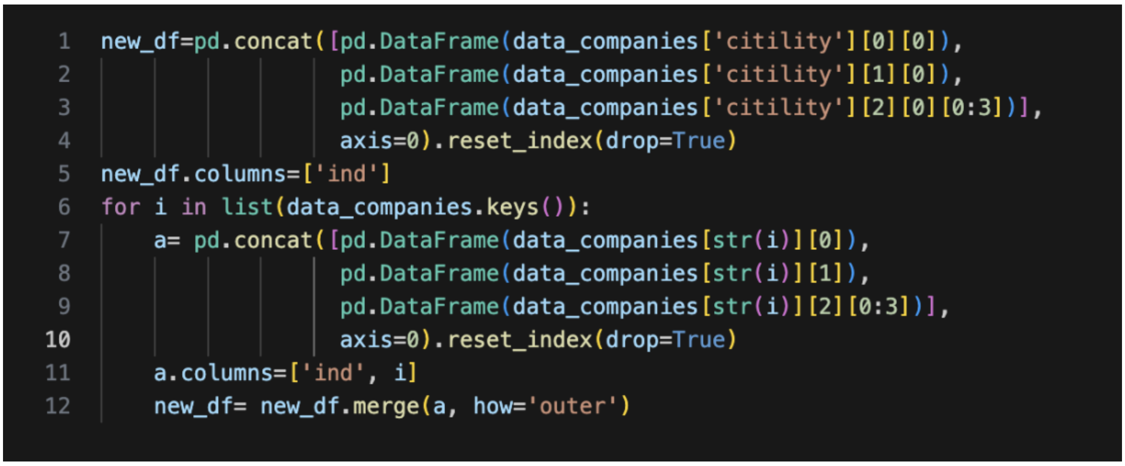


```json
{
  "industryName": "Technology, Information and Internet",
  "entityUrn": "urn:li:fs_profile:ACoAAABq0HkBaRRWDR0gM-CFD5bQ16aBsa_mTPQ",
  "geoLocation": {
    "geoUrn": "urn:li:fs_geo:106967730"
  },
  "geoLocationName": "Berlin, Berlin",
  "location": {
    "basicLocation": {
      "countryCode": "de"}
  },
  "experience": [
    {
      "locationName": "Paris Area, France",
      "entityUrn": "urn:li:fs_position:(ACoAAABq0HkBaRRWDR0gM-CFD5bQ16aBsa_mTPQ,1336711278)",
      "geoLocationName": "Paris Area, France",
      "geoUrn": "urn:li:fs_geo:90009659",
      "companyName": "Make.org",
      "timePeriod": {
        "startDate": {
          "month": 5,
          "year": 2018
        }
      },
      "company": {
        "employeeCountRange": {
          "start": 11,
          "end": 50
        },
        "industries": [
          "Information Technology and Services"
        ]
      },
```

Figure 5. Excerpt from a json file produced by the LinkedIn scraper (produced with https://codebeautify.org/jsonviewer)

Handling the html structure of nested dataframes

For the dictionary of companies' information, the re-structuring was done in an iterative manner. After exploring the layers and sub-layers of the dictionary (cf. figure 4), I used the pandas library to select specific elements in the dictionary and build a new dataframe through a mix of concatenating and merging options (see figure 6).



```python
new_df=pd.concat([pd.DataFrame(data_companies['citility'][0][0]),
                  pd.DataFrame(data_companies['citility'][1][0]),
                  pd.DataFrame(data_companies['citility'][2][0][0:3])],
                  axis=0).reset_index(drop=True)
new_df.columns=['ind']
for i in list(data_companies.keys()):
    a= pd.concat([pd.DataFrame(data_companies[str(i)][0]),
                  pd.DataFrame(data_companies[str(i)][1]),
                  pd.DataFrame(data_companies[str(i)][2][0:3])],
                  axis=0).reset_index(drop=True)
    a.columns=['ind', i]
    new_df= new_df.merge(a, how='outer')
```

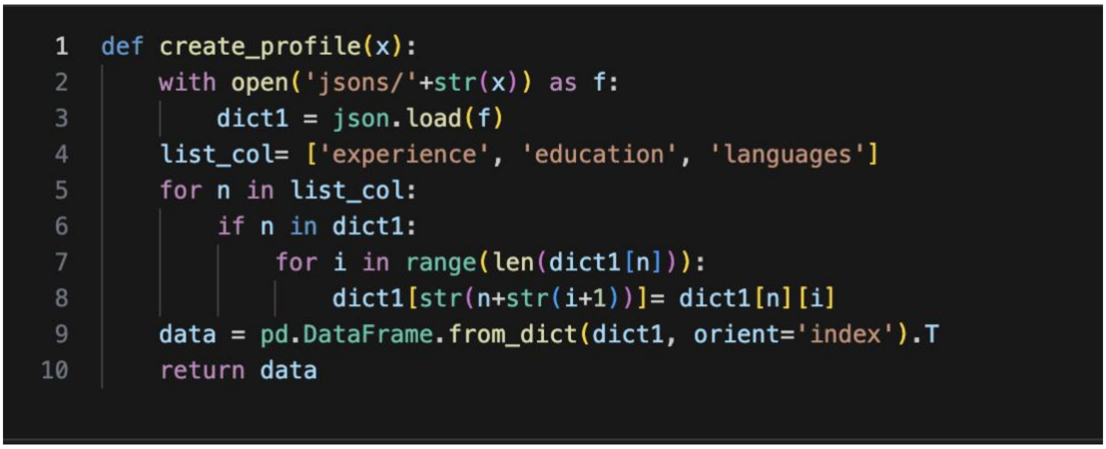
Figure 6. Creating a new dataframe from elements in nested dataframes in a dictionary

Handling the nested dictionaries in the json files

Json files have the advantage of having a dictionary structure, which makes it easy to navigate in it. However, when the dictionary has a complex structure with many sub-dictionaries, it is quickly easy to lose ourselves in it. In addition, we wanted to be able to read the data to think about how to structure it and organize it.

Our main issue was that for each individual, all his or her work experiences were lumped together in a dictionary that included, for each work experience, the name of the company, the number of employees in the company (in a sub-dictionary), the job title, the start date (with a sub-dictionary of month and year), etc… This was also the case for the education, languages, honors received, publications, certifications, volunteering, and projects sections of the profiles.

After trying to explode all these sections into individual data points, I noticed many ended up producing columns with a lot of missing values, and with little explaining power. Although this detail could be useful later, for specific queries, for this project I chose to focus on education and experience.

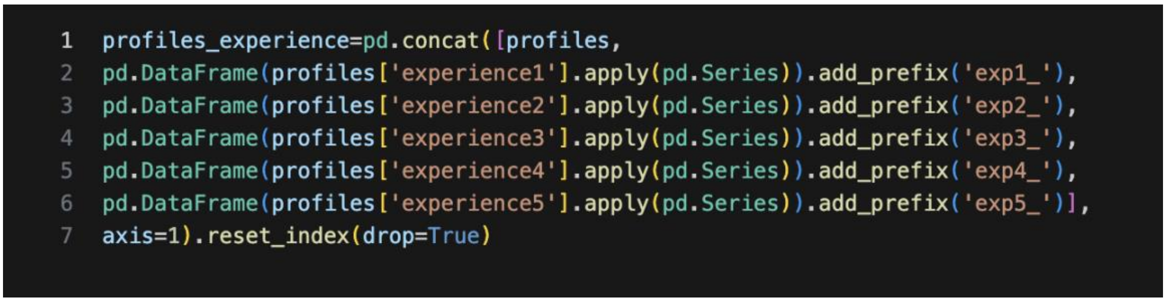


```python
def create_profile(x):
    with open('jsons/'+str(x)) as f:
        dict1 = json.load(f)
    list_col= ['experience', 'education', 'languages']
    for n in list_col:
        if n in dict1:
            for i in range(len(dict1[n])):
                dict1[str(n+str(i+1))]= dict1[n][i]
    data = pd.DataFrame.from_dict(dict1, orient='index').T
    return data
```

Figure 7. Function to import json files in a dataframe while restructuring data organization

I decide to include the data structuring directly in the json import function, for the education, experience, and language sections, as shown in figure 7. For the other sections, I only created a new column that states whether or not the section was filled. For languages, I chose to also keep the information on whether the person had included more than 3 languages, as this seemed to be an important element to study international companies and highly educated individuals.

Later on, I chose to further separate the information (exploding each of the work experiences and of the education programs, for instance), but using the panda Series function, as shown for example in figure 8. More detail of the column construction for each of the scraped sources is available in the commented jupyter notebook files named companies_data_scraping and jsons_to_dataframe_cleaning.



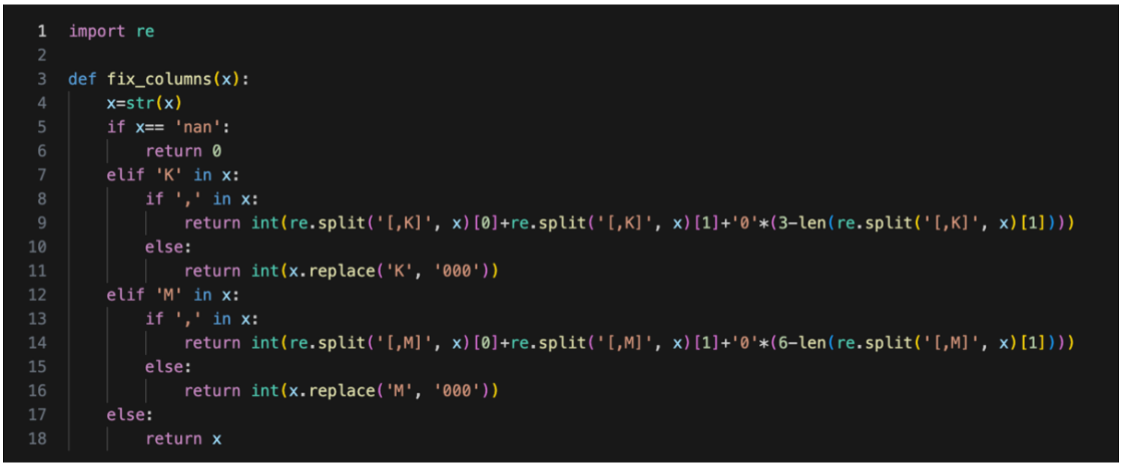
```
profiles_experience=pd.concat([profiles,
pd.DataFrame(profiles['experience1'].apply(pd.Series)).add_prefix('exp1_'),
pd.DataFrame(profiles['experience2'].apply(pd.Series)).add_prefix('exp2_'),
pd.DataFrame(profiles['experience3'].apply(pd.Series)).add_prefix('exp3_'),
pd.DataFrame(profiles['experience4'].apply(pd.Series)).add_prefix('exp4_'),
pd.DataFrame(profiles['experience5'].apply(pd.Series)).add_prefix('exp5_')],
axis=1).reset_index(drop=True)
```

Figure 8. Creating new dataframe columns with the pd.Series method

Data cleaning

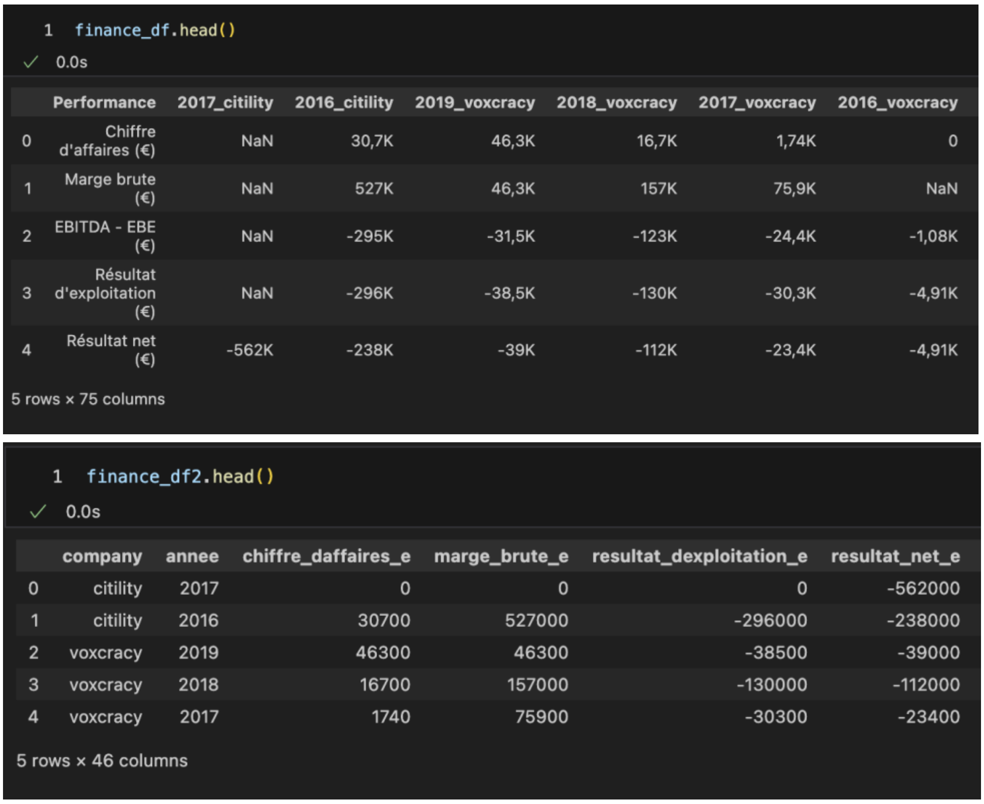
In addition to selecting relevant information and structuring it into tables, data cleaning was necessary to simplify and make the datasets more readable. I am here separating two types of data cleaning operations.

The first part of data cleaning is purely functional. It includes dropping duplicates and columns that don't have a significance for us (e.g. identifiers from source website), renaming columns to avoid spaces and special characters, managing the index, and dealing with data types. The figure 9, below, provides an example of this type of data cleaning, here replacing numbers formatted as text by only numeric characters, in order to be able to assign a numeric data type to the column and analyze it as such. Figure 10 provides a before/ after screenshot of a table that goes through simple data cleaning.



```python
import re

def fix_columns(x):
    x=str(x)
    if x== 'nan':
        return 0
    elif 'K' in x:
        if ',' in x:
            return int(re.split('[,K]', x)[0]+re.split('[,K]', x)[1]+'0'*(3-len(re.split('[,K]', x)[1])))
        else:
            return int(x.replace('K', '000'))
    elif 'M' in x:
        if ',' in x:
            return int(re.split('[,M]', x)[0]+re.split('[,M]', x)[1]+'0'*(6-len(re.split('[,M]', x)[1])))
        else:
            return int(x.replace('M', '000'))
    else:
        return x
```

Figure 9. Function to deal with numbers stored as text.



```python
finance_df.head()
```

✓ 0.0s

	Performance	2017_citility	2016_citility	2019_voxcracy	2018_voxcracy	2017_voxcracy	2016_voxcracy
0	Chiffre d'affaires (€)	NaN	30,7K	46,3K	16,7K	1,74K	0
1	Marge brute (€)	NaN	527K	46,3K	157K	75,9K	NaN
2	EBITDA - EBE (€)	NaN	-295K	-31,5K	-123K	-24,4K	-1,08K
3	Résultat d'exploitation (€)	NaN	-296K	-38,5K	-130K	-30,3K	-4,91K
4	Résultat net (€)	-562K	-238K	-39K	-112K	-23,4K	-4,91K

5 rows × 75 columns

```python
finance_df2.head()
```

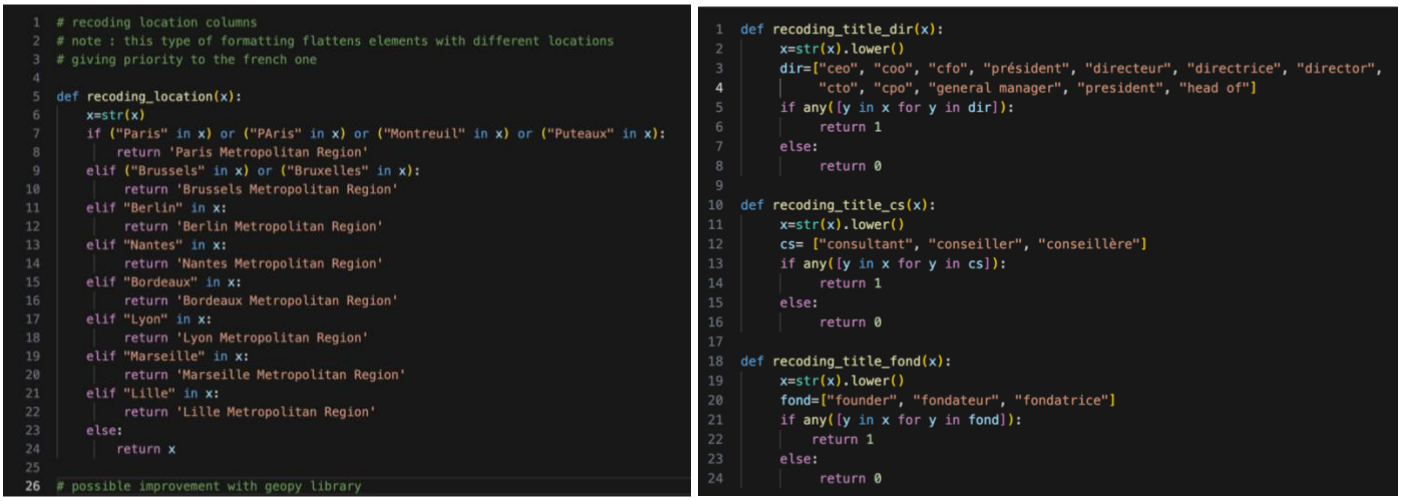
✓ 0.0s

	company	annee	chiffre_daffaires_e	marge_brute_e	resultat_dexploitation_e	resultat_net_e
0	citility	2017	0	0	0	-562000
1	citility	2016	30700	527000	-296000	-238000
2	voxcracy	2019	46300	46300	-38500	-39000
3	voxcracy	2018	16700	157000	-130000	-112000
4	voxcracy	2017	1740	75900	-30300	-23400

5 rows × 46 columns

Figure 10. Before and after: functional data cleaning

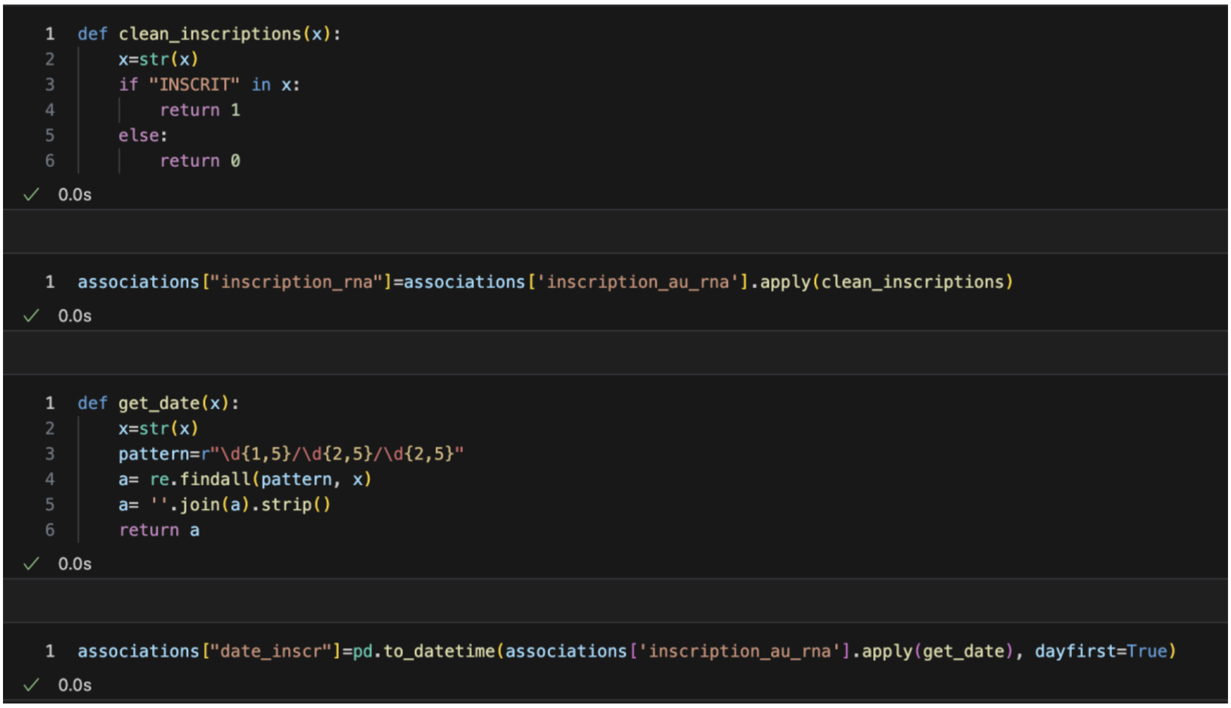
A second type of data cleaning operations could be defined as more analytical: they imply recoding some of the information, creating new categorical variables, in order to make it possible to analyze a large amount of information. One characteristic of our dataset is that it is mainly text, and although that will be precious when doing NLP treatments, it is not so practical for initial data exploration and database construction. Below are some examples of the recoding operations I performed to have a dataset that would be appropriate for EDA, visualization and direct querying.



```python
# recoding location columns
# note : this type of formatting flattens elements with different locations
# giving priority to the french one

def recoding_location(x):
    x=str(x)
    if ("Paris" in x) or ("PAris" in x) or ("Montreuil" in x) or ("Puteaux" in x):
       return 'Paris Metropolitan Region'
    elif ("Brussels" in x) or ("Bruxelles" in x):
        return 'Brussels Metropolitan Region'
    elif "Berlin" in x:
        return 'Berlin Metropolitan Region'
    elif "Nantes" in x:
        return 'Nantes Metropolitan Region'
    elif "Bordeaux" in x:
        return 'Bordeaux Metropolitan Region'
    elif "Lyon" in x:
        return 'Lyon Metropolitan Region'
    elif "Marseille" in x:
        return 'Marseille Metropolitan Region'
    elif "Lille" in x:
        return 'Lille Metropolitan Region'
    else:
       return x

# possible improvement with geopy library
```

```python
def recoding_title_dir(x):
    x=str(x).lower()
    dir=["ceo", "coo", "cfo", "président", "directeur", "directrice", "director",
        "cto", "cpo", "general manager", "president", "head of"]
    if any([y in x for y in dir]):
        return 1
    else:
        return 0

def recoding_title_cs(x):
    x=str(x).lower()
    cs= ["consultant", "conseiller", "conseillère"]
    if any([y in x for y in cs]):
        return 1
    else:
        return 0

def recoding_title_fond(x):
    x=str(x).lower()
    fond=["founder", "fondateur", "fondatrice"]
    if any([y in x for y in fond]):
       return 1
    else:
        return 0
```

Figure 11. Recoding to have a limited number of categories (location) or to create new binary columns that qualify a position (title)



```python
def clean_inscriptions(x):
    x=str(x)
    if "INSCRIT" in x:
        return 1
    else:
        return 0
```

```
✓ 0.0s
```

```python
associations["inscription_rna"]=associations['inscription_au_rna'].apply(clean_inscriptions)
```

```
✓ 0.0s
```

```python
def get_date(x):
    x=str(x)
    pattern=r"\d{1,5}/\d{2,5}/\d{2,5}"
    a= re.findall(pattern, x)
    a= ''.join(a).strip()
    return a
```

```
✓ 0.0s
```

```python
associations["date_inscr"]=pd.to_datetime(associations['inscription_au_rna'].apply(get_date), dayfirst=True)
```

```
✓ 0.0s
```

Figure 12. Using regular expressions (regex) to extract meaningful information from columns and create new categories (date, registered or not) of single data type.

Database construction

NoSQL vs. SQL

SQL and NoSQL are two different types of database management systems with different characteristics. Here are some major differences between the two:

1. Data Model: SQL databases use a tabular data model (data is stored in tables consisting of rows and columns) while NoSQL databases use a broader range of data models (document-oriented, key-value pairs, graph, and column-family).
2. Schema: SQL databases requires the structure to be defined before data can be stored. NoSQL databases are more flexible: data can be stored without a predefined schema structure.
3. Scalability: SQL databases are vertically scalable: they handle an increasing amount of data by increasing the resources of a single server. NoSQL databases are horizontally scalable, i.e. they handle an increasing amount of data by adding more servers to a distributed system.
4. Transactions: SQL databases have built-in support for transactions that ensure data consistency and integrity. Depending on the database technology used, NoSQL databases sometimes don't support transactions.
5. Querying: SQL databases support complex querying through Structured Query Language. NoSQL databases have specific query languages (these can be simpler but are often focused on specific data models).
6. Cost: SQL databases tend to be more expensive due to their need for specialized hardware and software. NoSQL databases are often more affordable because they can run on commodity hardware and open-source software.

For this project, as I have pre-structured tables and a pre-defined database structure, as well as the objective to add more data, SQL is a good fit.

Building separate tables

In order to be able to analyze our data, we want each data point to be unique, and to have a unique identifier. In order to achieve this, I split my datasets into 7 different tables.

The "names" table includes only first and last name as well as individual ID. I decided to separate this information to allow for a minimum of anonymization. This is not enough for a larger scale project dealing with personal data, but for the scope of this experimentation, at least it allows to envision how to separate personal data, specify different authorizations for this part of the database, and respect GDPR requirements.

The "people" table provides basic information about the individuals. It contains a lot of the original information as text, as well as some basic information recoded as binary variables, as explained in the sections above.

The "people_experience" table repeats the individual IDs for each professional experience of the person (n=5), while also allowing to have an id specific to each experience (or each row).

The "people_education" considers each education experience as an instance (row), and again relates it to the individual ID so we can connect it to the people table.

The "companies_info" table provides basic information on companies (commercial, for profit) in the civic tech ecosystem, including whether or not they are still active.

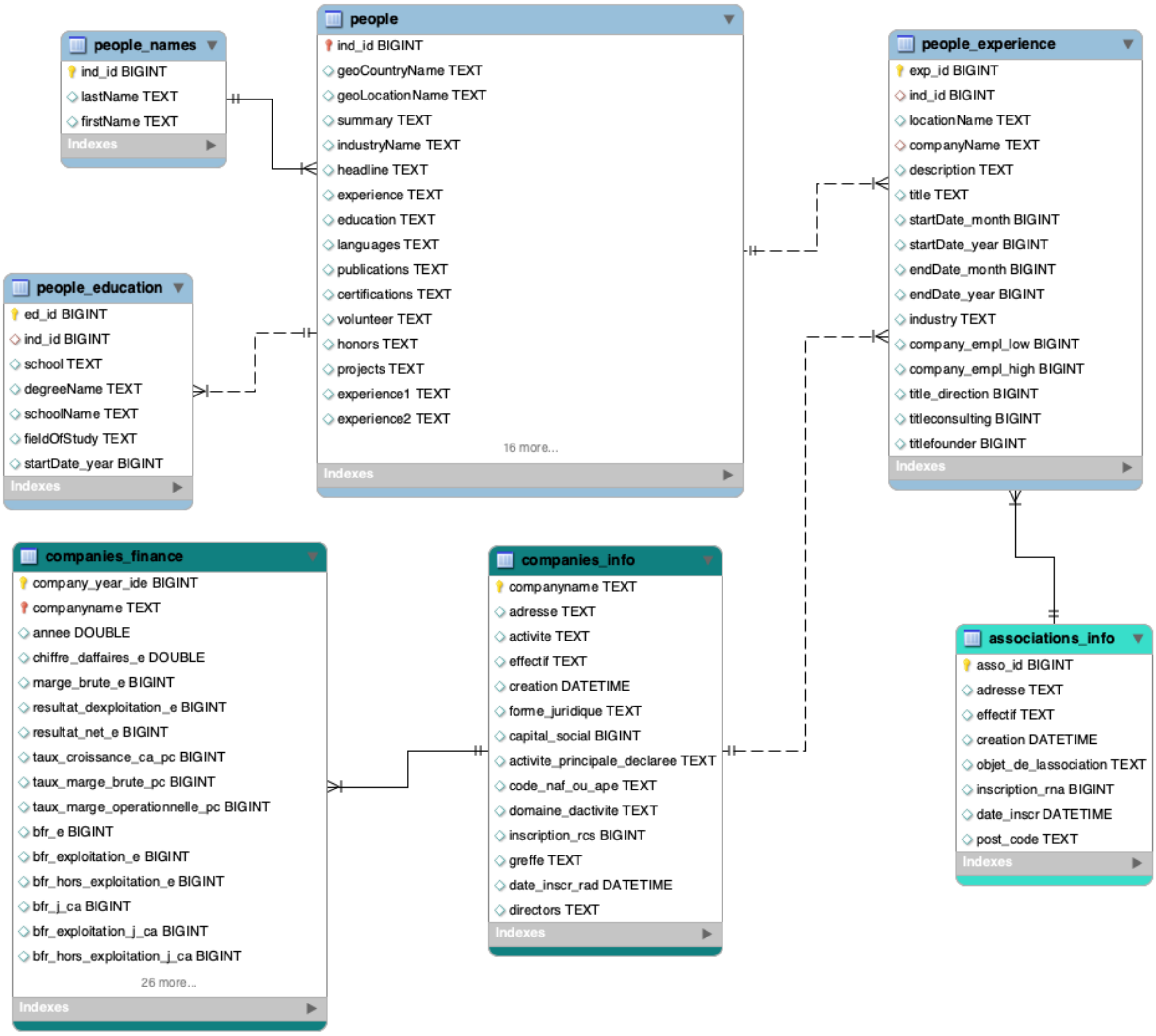
The "companies_finance" table contains the financial information about the companies in the previous table, with a distinct row for each year for which there is information.

The "associations_info" table was produced from a subset of the companies_info table : some of the organizations identified are not for profit. Since these organizations have different data points than the companies, I chose to have a separate table. However, the association name serves as an identifier to relate these organizations to the people_experience table, as some people have identified experiences in NGOs.

Annex 1 provides screenshots of the first rows of these 7 different tables. The detail of the tables construction can be found in the jupyter notebook titled creating_sql_tables.

Entity relationship diagram

Below is the diagram that shows the relationships between the tables in our database.



Data analysis and visualization

EDA (Exploratory Data Analysis) with Python

I used Python to conduct exploratory data analysis mainly on the financial data table. This table has many columns concerning companies' financial health. Before we can analyze those, we need to see whether some of them were correlated between them. The figure below shows that there were some highly correlated columns ('marge_brute_e', 'valeur_ajoutee_sur_ca_pc', 'resultat_net_e', 'bfr_hors_exploitation_j_ca', 'capacite_dautofinancement_e', 'marge_nette_pc'). We dropped most of them, except "net revenue", which we kept because it is the best documented column in our database. To remove the correlation issue, we dropped the other column (resultat d'exploitation) that was highly correlated with net revenue.

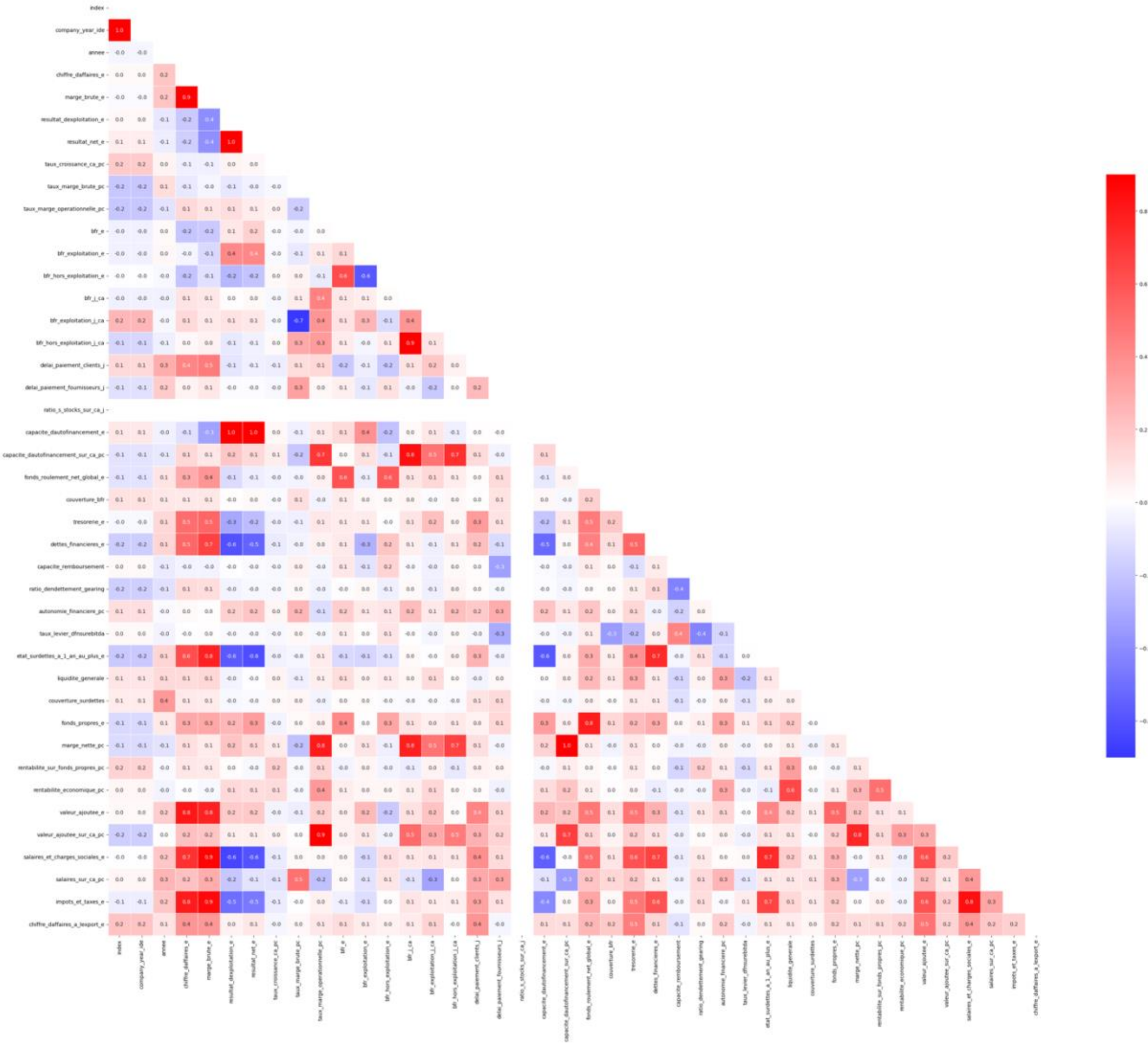


Figure 13. Correlations between columns in the financial dataset (scores range from -1 to 1)

We also explored the table by looking at the most common measure of a company's success, sales ('chiffre d'affaires'). We first calculated the mean of sales for each company, as shown in the first figure below. Noticing that several of the columns had many outliers, however, we chose to check whether sales per year were coherent for each of the companies. What we see, by focusing on the six main companies, for which we have detailed data, is that sales vary a lot per year. Therefore, calculating the mean is probably not the best idea to evaluate companies (as we will discuss again later).

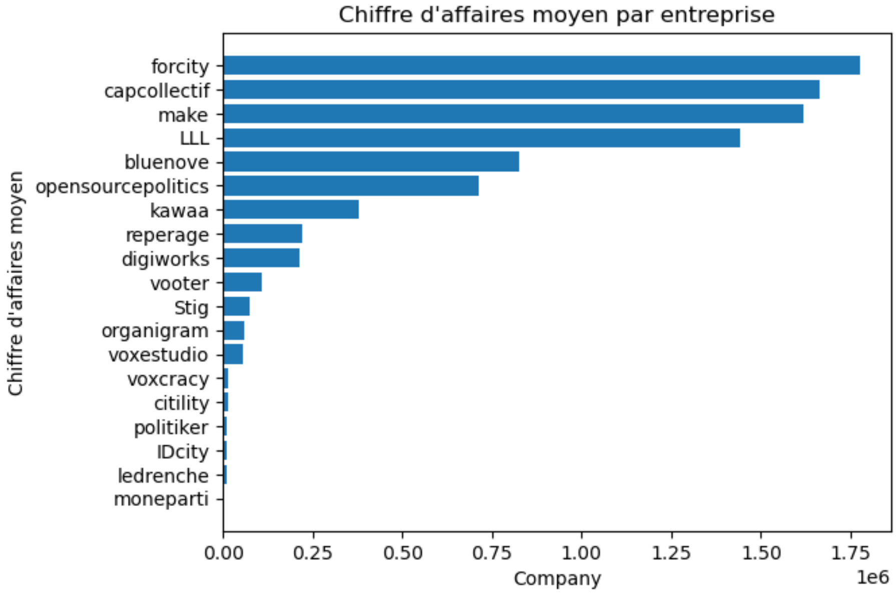


Figure 14. Mean sales per company (for all documented years)

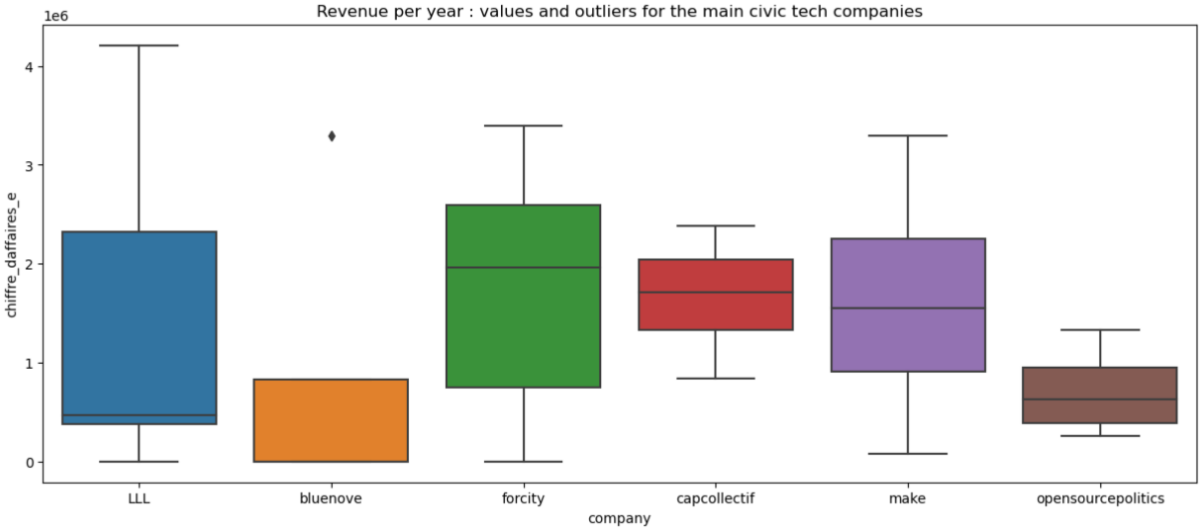


Figure 15. Distribution of sales amounts per company per year (boxplot with whiskers and outliers).

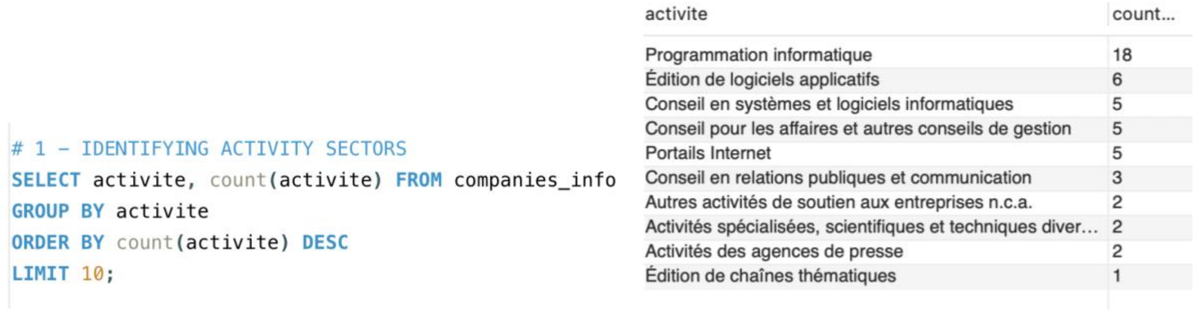
Querying the database with MySQL Workbench

The advantage of SQL is the ability to cross tables to produce analyses. The detail of the queries produced is available in the file final_project_db_queries.sql. Here we highlight five queries that produced interesting insights on our dataset.

Defining civic tech companies.

The queries below were used to get basic information about civic tech companies. We notice that companies are mainly registered in the "programmation informatique" activity (i.e. computer programming), and that most of them have no employees.

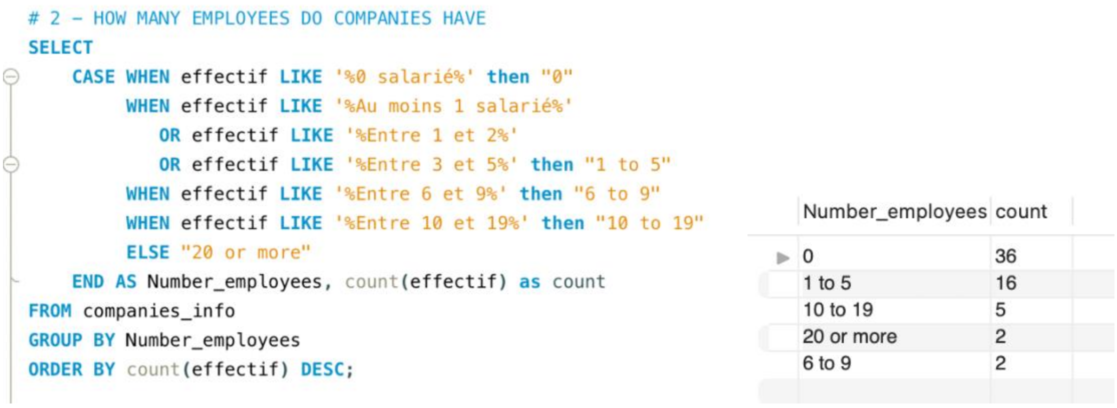
This does not necessarily mean that nobody works for these companies, but it does mean that only founders or associates are working there, and not being hired as salaried staff. We can also see that 7 companies have 10 employees or more, so there are also some companies that have already concentrated most of the business in the sector.



```sql
# 1 - IDENTIFYING ACTIVITY SECTORS
SELECT activite, count(activite) FROM companies_info
GROUP BY activite
ORDER BY count(activite) DESC
LIMIT 10;
```

activite	count...
Programmation informatique	18
Édition de logiciels applicatifs	6
Conseil en systèmes et logiciels informatiques	5
Conseil pour les affaires et autres conseils de gestion	5
Portails Internet	5
Conseil en relations publiques et communication	3
Autres activités de soutien aux entreprises n.c.a.	2
Activités spécialisées, scientifiques et techniques diver...	2
Activités des agences de presse	2
Édition de chaînes thématiques	1

Figure 16. SQL Query 1 (activity sectors) for "who's who in civic tech?" and results table (count per activity)

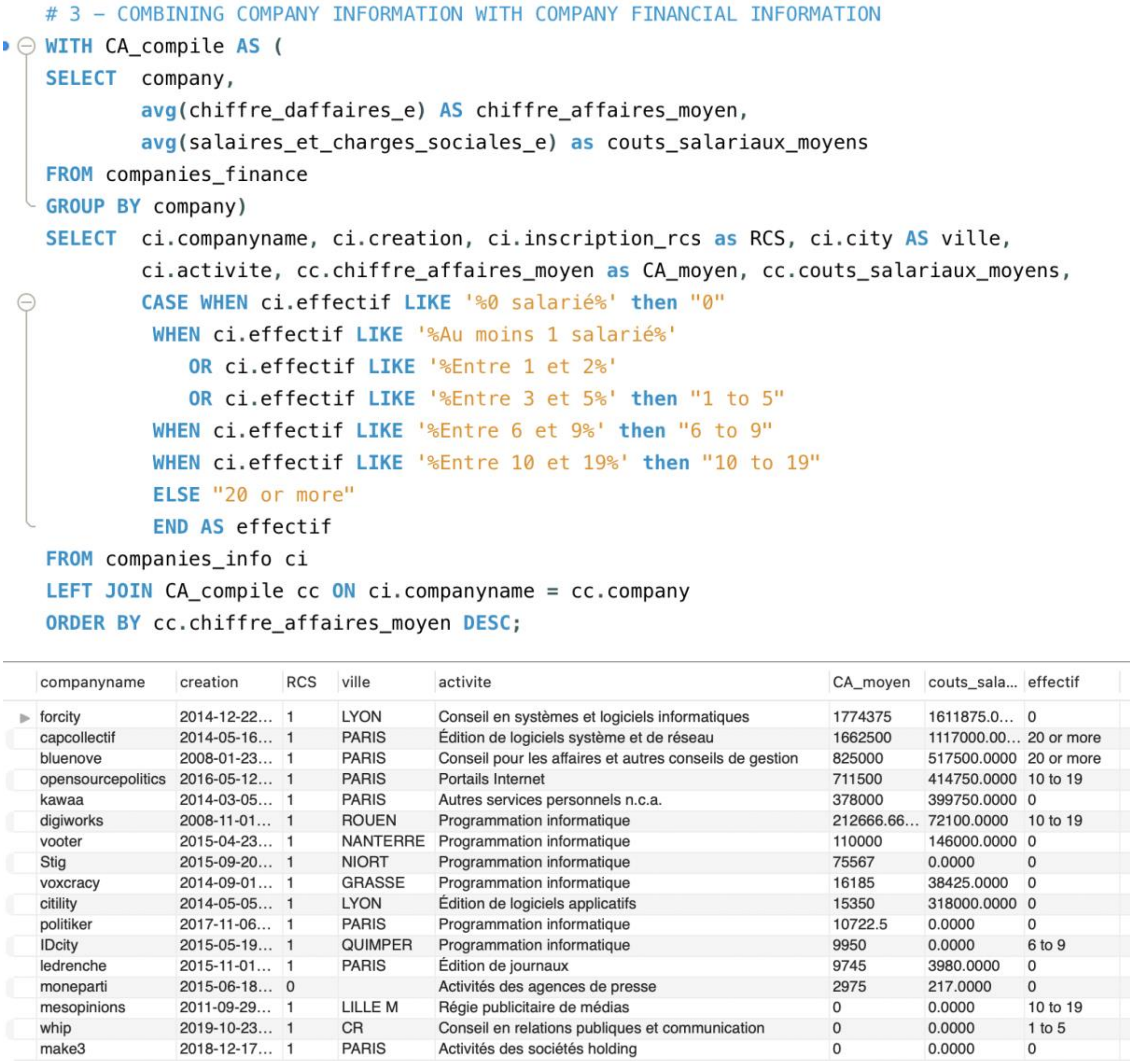


```sql
# 2 - HOW MANY EMPLOYEES DO COMPANIES HAVE
SELECT
    CASE WHEN effectif LIKE '%0 salarié%' then "0"
         WHEN effectif LIKE '%Au moins 1 salarié%'
            OR effectif LIKE '%Entre 1 et 2%'
            OR effectif LIKE '%Entre 3 et 5%' then "1 to 5"
         WHEN effectif LIKE '%Entre 6 et 9%' then "6 to 9"
         WHEN effectif LIKE '%Entre 10 et 19%' then "10 to 19"
         ELSE "20 or more"
    END AS Number_employees, count(effectif) as count
FROM companies_info
GROUP BY Number_employees
ORDER BY count(effectif) DESC;
```

Number_employees	count
0	36
1 to 5	16
10 to 19	5
20 or more	2
6 to 9	2

Figure 17. SQL Query 2 (employees) for "who's who in civic tech?" and results table (count of companies per category of employee numbers)

We also see, if we use SQL to join the company information and the company financial information tables (cf. query 3), that the companies with the highest average revenue are the ones with the highest number of employees (Cap Collectif, Bluenove, Open Source Politics, Digiworks). However, there are also a lot of companies with an employee count of 0 in the top rows of this list: this is related to the dataset. Using the average revenue over the different years can skew the results because for some companies, we only have revenue information for one year (cf. EDA).



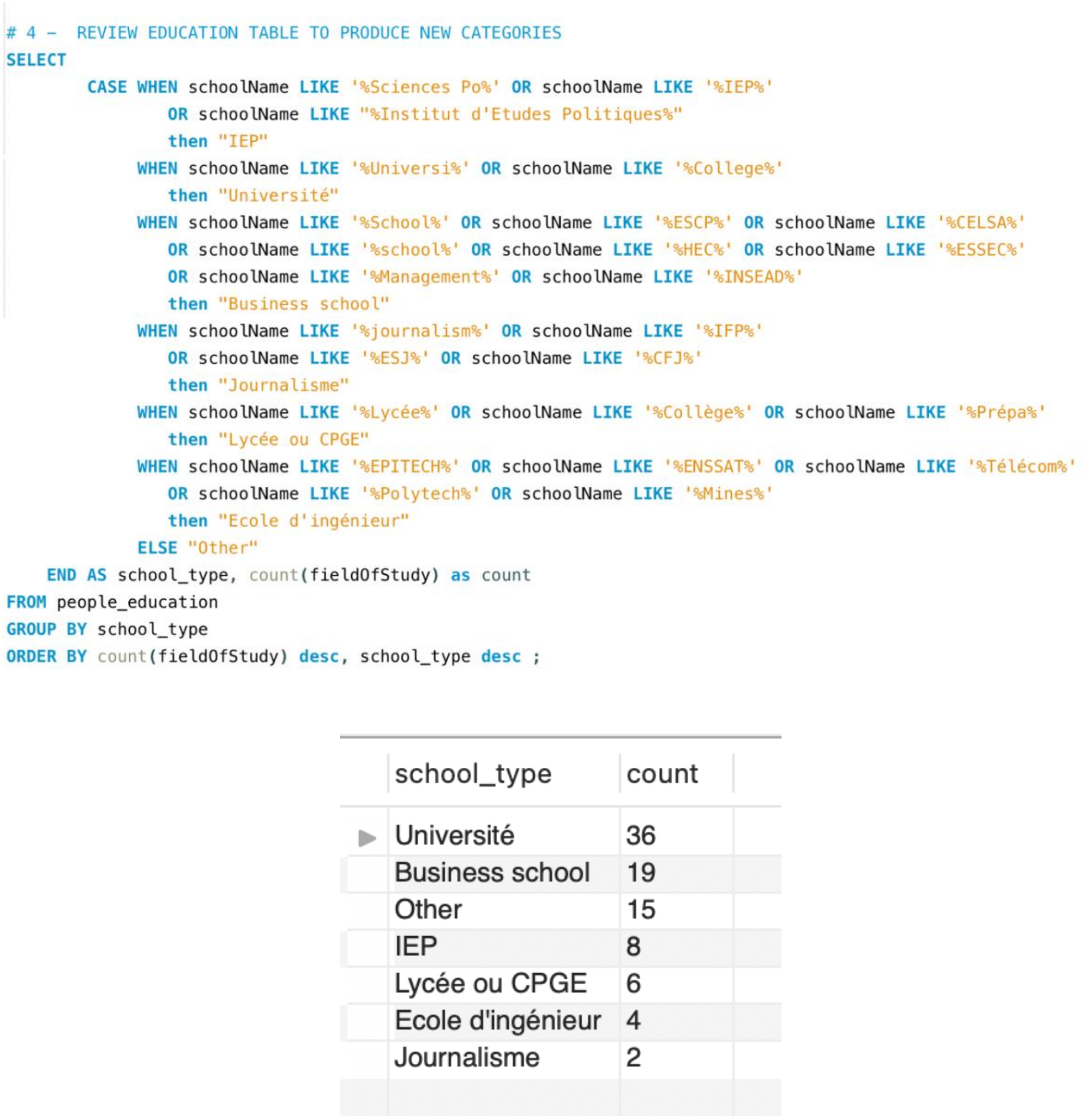
```
# 3 - COMBINING COMPANY INFORMATION WITH COMPANY FINANCIAL INFORMATION
WITH CA_compile AS (
SELECT  company,
        avg(chiffre_daffaires_e) AS chiffre_affaires_moyen,
        avg(salaires_et_charges_sociales_e) as couts_salariaux_moyens
FROM companies_finance
GROUP BY company)
SELECT  ci.companyname, ci.creation, ci.inscription_rcs as RCS, ci.city AS ville,
        ci.activite, cc.chiffre_affaires_moyen as CA_moyen, cc.couts_salariaux_moyens,
        CASE WHEN ci.effectif LIKE '%0 salarié%' then "0"
         WHEN ci.effectif LIKE '%Au moins 1 salarié%'
            OR ci.effectif LIKE '%Entre 1 et 2%'
            OR ci.effectif LIKE '%Entre 3 et 5%' then "1 to 5"
         WHEN ci.effectif LIKE '%Entre 6 et 9%' then "6 to 9"
         WHEN ci.effectif LIKE '%Entre 10 et 19%' then "10 to 19"
         ELSE "20 or more"
         END AS effectif
FROM companies_info ci
LEFT JOIN CA_compile cc ON ci.companyname = cc.company
ORDER BY cc.chiffre_affaires_moyen DESC;
```

companyname	creation	RCS	ville	activite	CA_moyen	couts_sala...	effectif
forcity	2014-12-22...	1	LYON	Conseil en systèmes et logiciels informatiques	1774375	1611875.0...	0
capcollectif	2014-05-16...	1	PARIS	Édition de logiciels système et de réseau	1662500	1117000.00...	20 or more
bluenove	2008-01-23...	1	PARIS	Conseil pour les affaires et autres conseils de gestion	825000	517500.0000	20 or more
opensourcepolitics	2016-05-12...	1	PARIS	Portails Internet	711500	414750.0000	10 to 19
kawaa	2014-03-05...	1	PARIS	Autres services personnels n.c.a.	378000	399750.0000	0
digiworks	2008-11-01...	1	ROUEN	Programmation informatique	212666.66...	72100.0000	10 to 19
vooter	2015-04-23...	1	NANTERRE	Programmation informatique	110000	146000.0000	0
Stig	2015-09-20...	1	NIORT	Programmation informatique	75567	0.0000	0
voxcracy	2014-09-01...	1	GRASSE	Programmation informatique	16185	38425.0000	0
citility	2014-05-05...	1	LYON	Édition de logiciels applicatifs	15350	318000.0000	0
politiker	2017-11-06...	1	PARIS	Programmation informatique	10722.5	0.0000	0
IDcity	2015-05-19...	1	QUIMPER	Programmation informatique	9950	0.0000	6 to 9
ledrenche	2015-11-01...	1	PARIS	Édition de journaux	9745	3980.0000	0
moneparti	2015-06-18...	0		Activités des agences de presse	2975	217.0000	0
mesopinions	2011-09-29...	1	LILLE M	Régie publicitaire de médias	0	0.0000	10 to 19
whip	2019-10-23...	1	CR	Conseil en relations publiques et communication	0	0.0000	1 to 5
make3	2018-12-17...	1	PARIS	Activités des sociétés holding	0	0.0000	0

Figure 18. SQL Query 3 (revenue) for "Who's who in civic tech ?" and results table (information per company)

Who's who in civic tech? From higher education analysis to understanding companies' cultures.

Our fourth query was focused on people, and more precisely on their higher education. Here I used SQL to explore the data and build new categories. I could have done it in Python, but since I am expecting to add additional data, I am still working on the categories for education. This results in a relatively complex query, but it produces results that will be interesting for our following query.

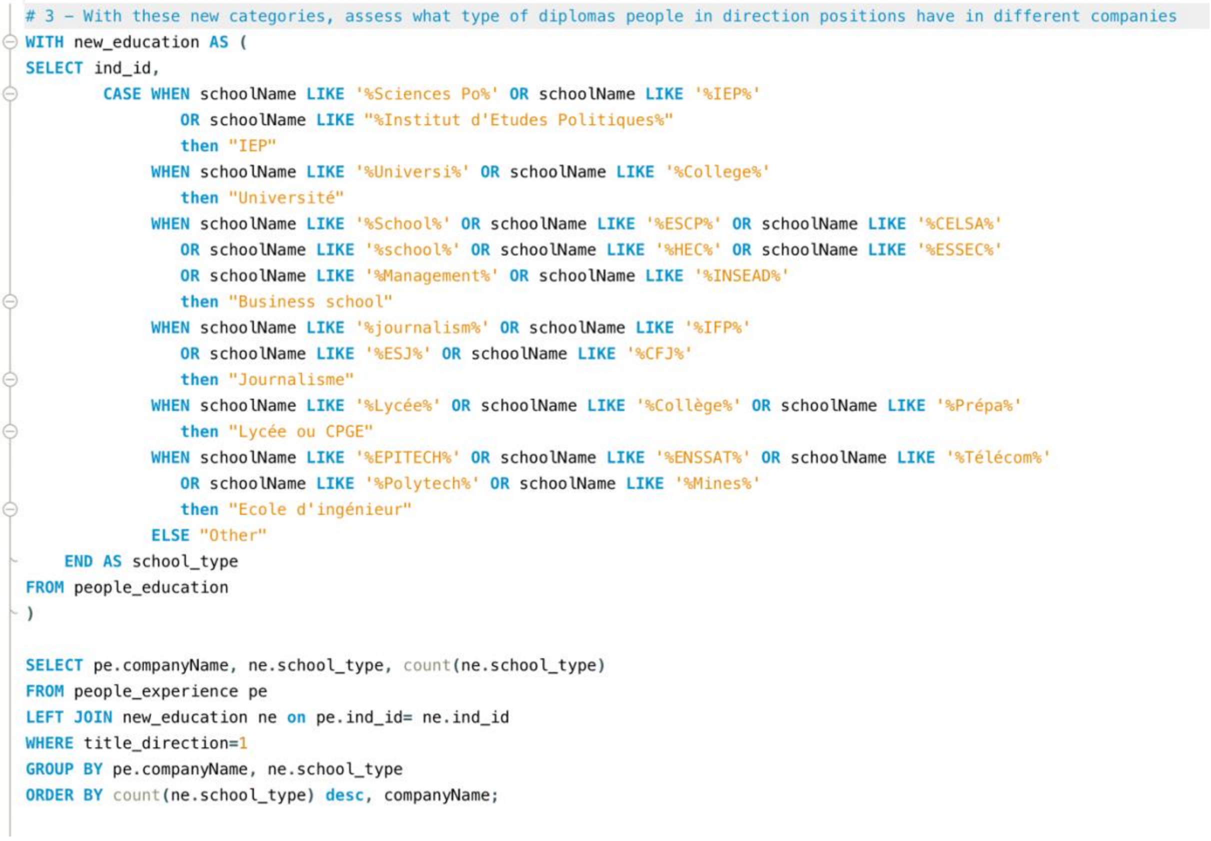


```
# 4 -  REVIEW EDUCATION TABLE TO PRODUCE NEW CATEGORIES
SELECT
        CASE WHEN schoolName LIKE '%Sciences Po%' OR schoolName LIKE '%IEP%'
                OR schoolName LIKE "%Institut d'Etudes Politiques%"
                then "IEP"
             WHEN schoolName LIKE '%Universi%' OR schoolName LIKE '%College%'
                then "Université"
             WHEN schoolName LIKE '%School%' OR schoolName LIKE '%ESCP%' OR schoolName LIKE '%CELSA%'
                OR schoolName LIKE '%school%' OR schoolName LIKE '%HEC%' OR schoolName LIKE '%ESSEC%'
                OR schoolName LIKE '%Management%' OR schoolName LIKE '%INSEAD%'
                then "Business school"
             WHEN schoolName LIKE '%journalism%' OR schoolName LIKE '%IFP%'
                OR schoolName LIKE '%ESJ%' OR schoolName LIKE '%CFJ%'
                then "Journalisme"
             WHEN schoolName LIKE '%Lycée%' OR schoolName LIKE '%Collège%' OR schoolName LIKE '%Prépa%'
                then "Lycée ou CPGE"
             WHEN schoolName LIKE '%EPITECH%' OR schoolName LIKE '%ENSSAT%' OR schoolName LIKE '%Télécom%'
                OR schoolName LIKE '%Polytech%' OR schoolName LIKE '%Mines%'
                then "Ecole d'ingénieur"
             ELSE "Other"
    END AS school_type, count(fieldOfStudy) as count
FROM people_education
GROUP BY school_type
ORDER BY count(fieldOfStudy) desc, school_type desc ;
```

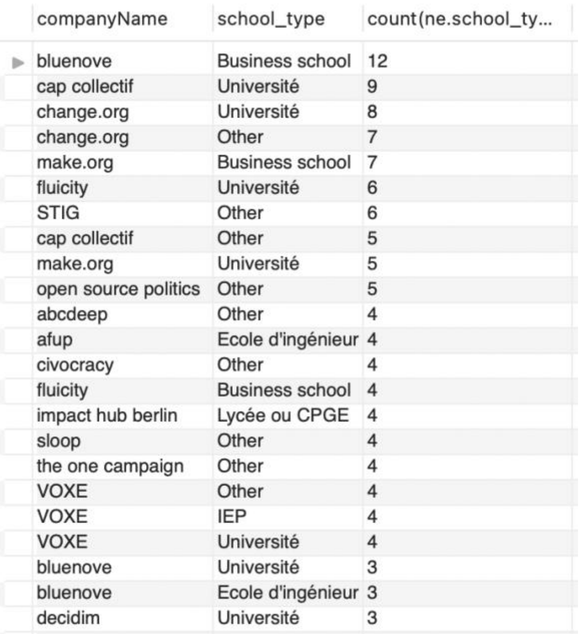
school_type	count
Université	36
Business school	19
Other	15
IEP	8
Lycée ou CPGE	6
Ecole d'ingénieur	4
Journalisme	2

Figure 19. SQL Query 4 (higher education) for "Who's who in civic tech?" and results table (count of people per higher education type)

Finally, we used the results of the previous query to better understand if there were differences between companies when it comes to the higher education of the people that work there, and more specifically, the people that have decision-making positions (i.e. CEOs, directors, CFOs, etc., as coded during the data cleaning step). Here we use the previous query as a CTE (common table expression) to join on the company information table and add a request to consider only rows where the individual has a direction role.



```
# 3 - With these new categories, assess what type of diplomas people in direction positions have in different companies
WITH new_education AS (
SELECT ind_id,
        CASE WHEN schoolName LIKE '%Sciences Po%' OR schoolName LIKE '%IEP%'
                OR schoolName LIKE "%Institut d'Etudes Politiques%"
                then "IEP"
             WHEN schoolName LIKE '%Universi%' OR schoolName LIKE '%College%'
                then "Université"
             WHEN schoolName LIKE '%School%' OR schoolName LIKE '%ESCP%' OR schoolName LIKE '%CELSA%'
                OR schoolName LIKE '%school%' OR schoolName LIKE '%HEC%' OR schoolName LIKE '%ESSEC%'
                OR schoolName LIKE '%Management%' OR schoolName LIKE '%INSEAD%'
                then "Business school"
             WHEN schoolName LIKE '%journalism%' OR schoolName LIKE '%IFP%'
                OR schoolName LIKE '%ESJ%' OR schoolName LIKE '%CFJ%'
                then "Journalisme"
             WHEN schoolName LIKE '%Lycée%' OR schoolName LIKE '%Collège%' OR schoolName LIKE '%Prépa%'
                then "Lycée ou CPGE"
             WHEN schoolName LIKE '%EPITECH%' OR schoolName LIKE '%ENSSAT%' OR schoolName LIKE '%Télécom%'
                OR schoolName LIKE '%Polytech%' OR schoolName LIKE '%Mines%'
                then "Ecole d'ingénieur"
             ELSE "Other"
    END AS school_type
FROM people_education
)

SELECT pe.companyName, ne.school_type, count(ne.school_type)
FROM people_experience pe
LEFT JOIN new_education ne on pe.ind_id= ne.ind_id
WHERE title_direction=1
GROUP BY pe.companyName, ne.school_type
ORDER BY count(ne.school_type) desc, companyName;
```



companyName	school_type	count(ne.school_ty...
bluenove	Business school	12
cap collectif	Université	9
change.org	Université	8
change.org	Other	7
make.org	Business school	7
fluicity	Université	6
STIG	Other	6
cap collectif	Other	5
make.org	Université	5
open source politics	Other	5
abcdeep	Other	4
afup	Ecole d'ingénieur	4
civocracy	Other	4
fluicity	Business school	4
impact hub berlin	Lycée ou CPGE	4
sloop	Other	4
the one campaign	Other	4
VOXE	Other	4
VOXE	IEP	4
VOXE	Université	4
bluenove	Université	3
bluenove	Ecole d'ingénieur	3
decidim	Université	3

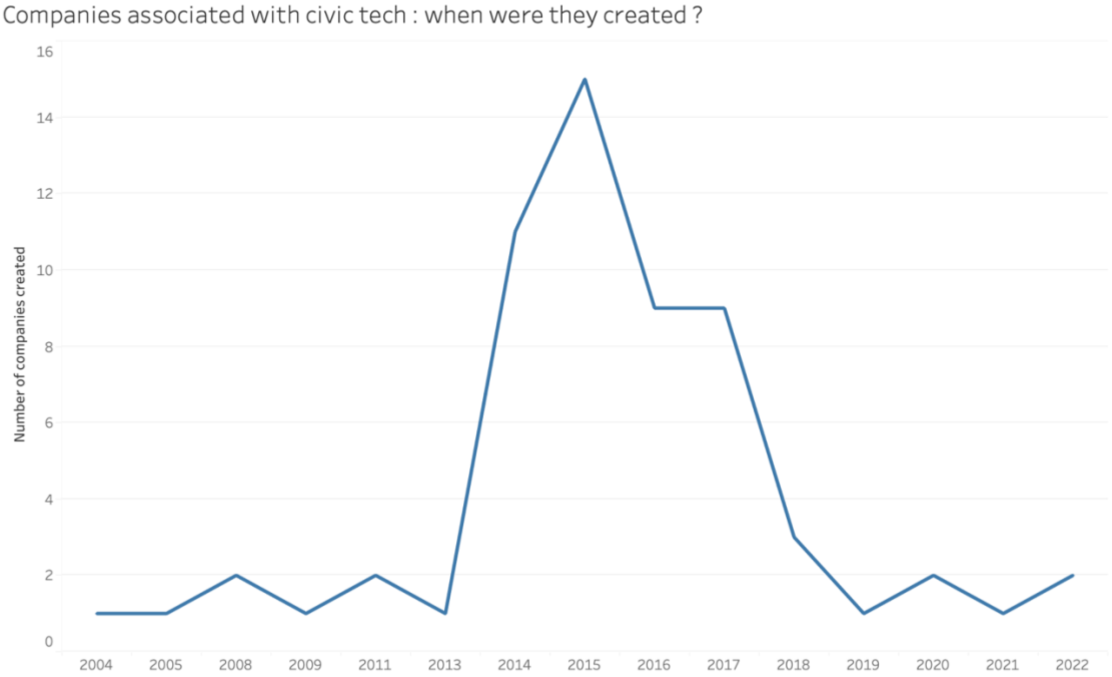
Although the results could be improved, they point towards the idea that companies have different types of recruitment strategies. A next step would be to combine this information with the fields of study. In fact, there are two French universities that deliver diplomas focused on citizen participation. It would be interesting to see which companies recruit people that have been trained on this topic.

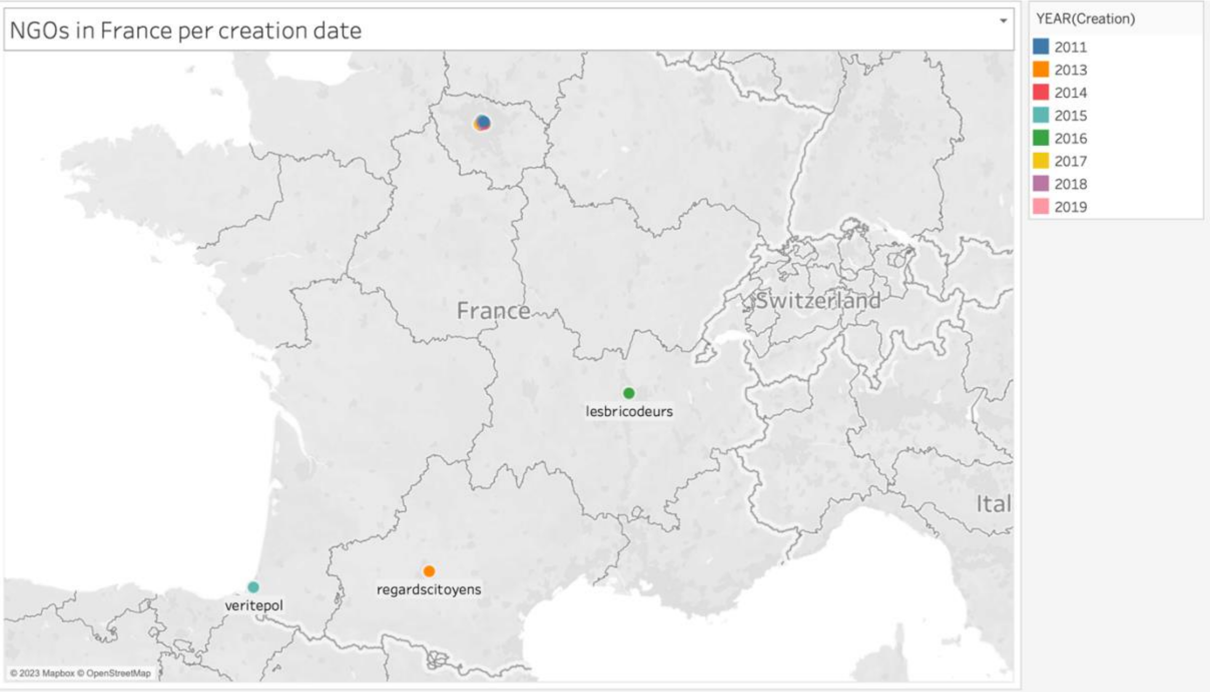
Figure 20. SQL Query 5 (company directors) for "Who's who in civic tech?" and results table (count of people in direction positions per education type in each company)

Tableau visualizations

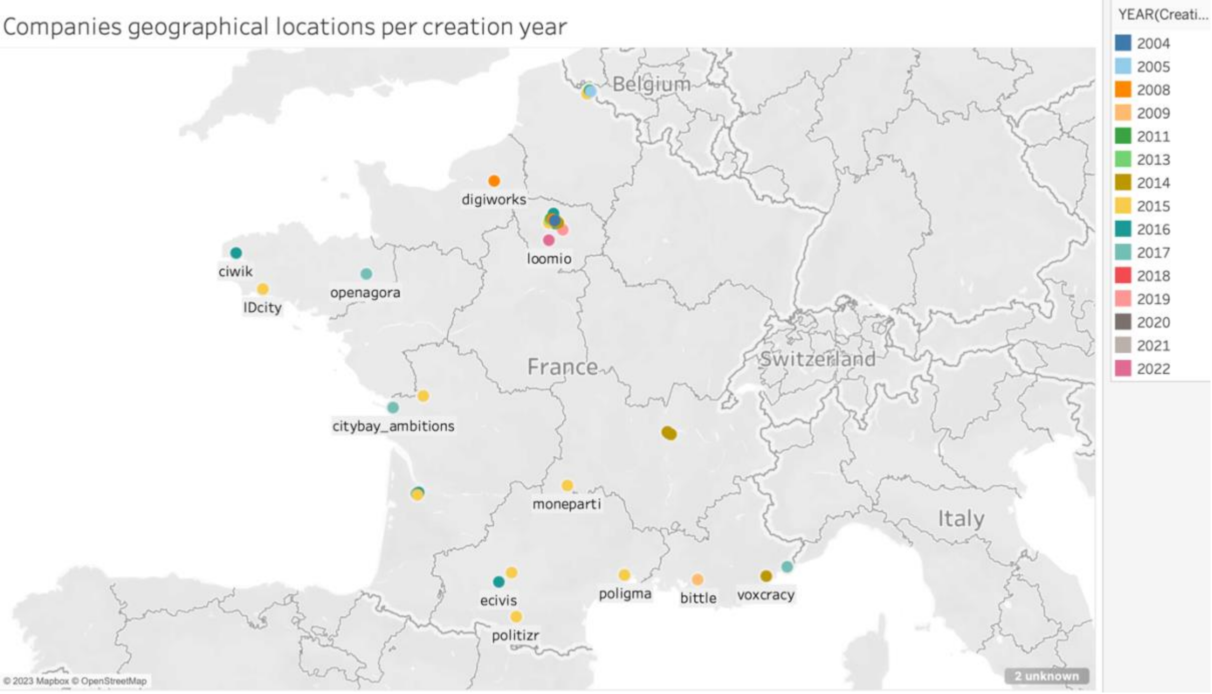
Geographical distribution

Using tableau to analyze the geographical distribution of companies and NGOs is interesting because it allows for more visual insights into our date. We can see that civic tech companies are concentrated in cities, and especially in Paris (this may explain that a lot of the companies that have closed are also in Paris). Of course, we also notice a lot of the companies were created between 2014 and 2016.

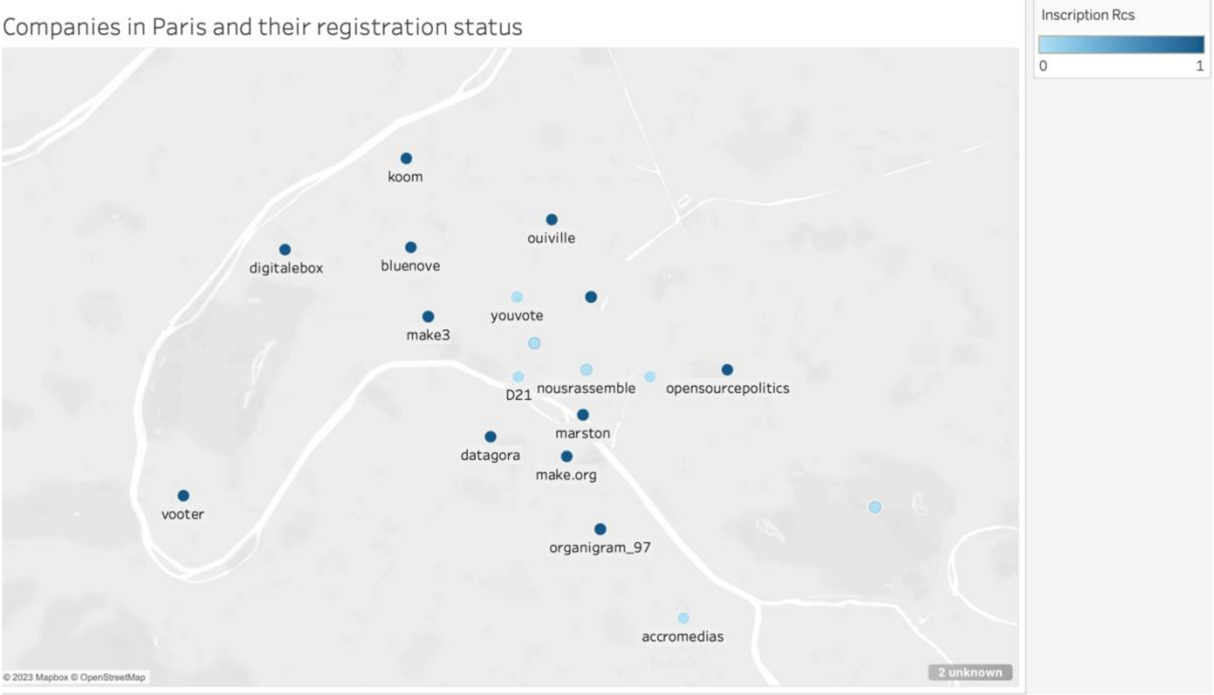




Companies geographical locations per creation year

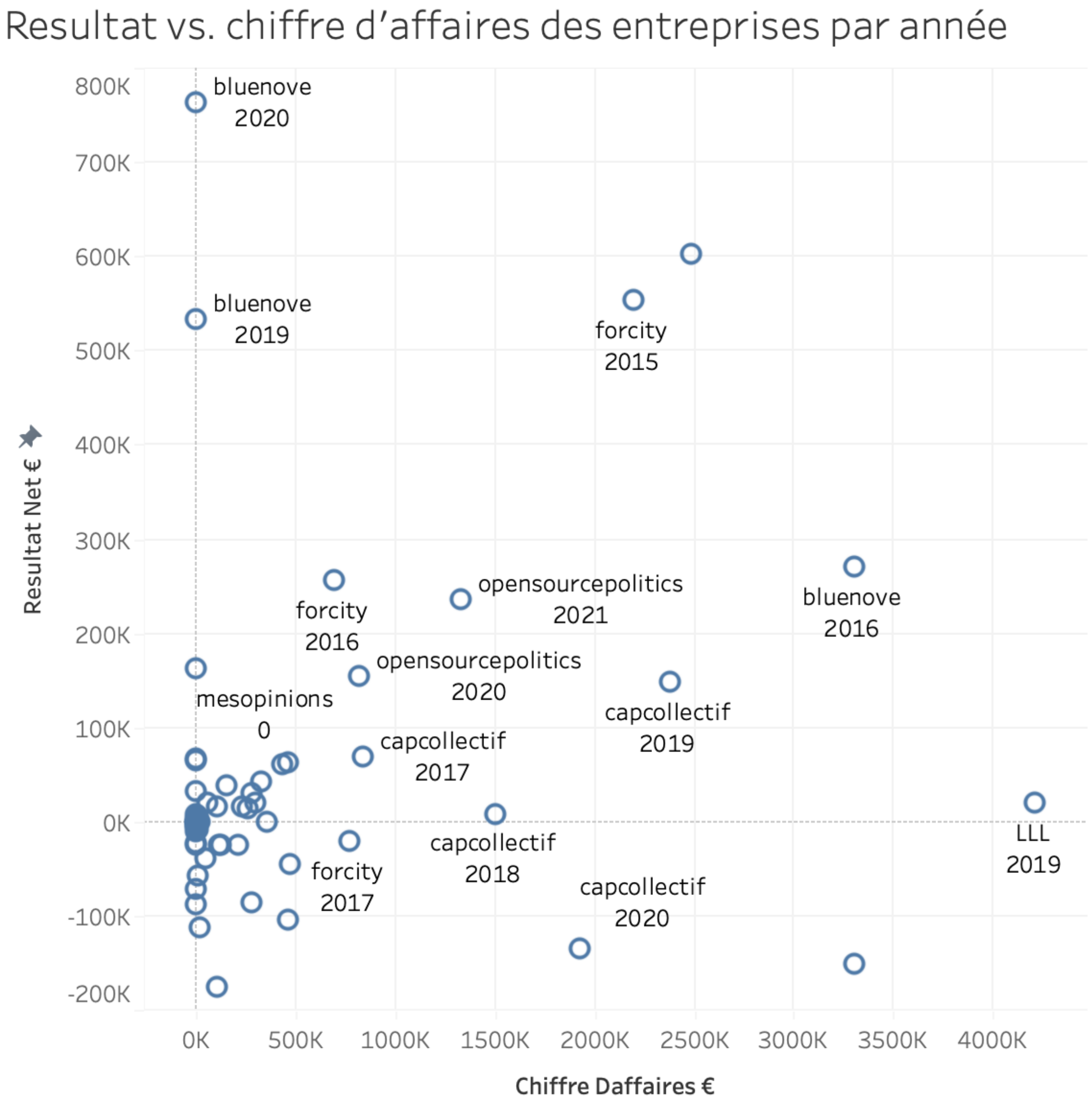


Companies in Paris and their registration status



Companies' revenues

As we have mentioned earlier, it is difficult to analyze revenue based on the average of all the yearly values, because these can vary a lot. Drawing a scatterplot with tableau allows to show the difference between the revenue and the net gains, as well as see how companies' financial results evolve over the years. We also see that a lot of companies lose money.



Summary of key insights on the data and next steps

In the introduction of this report, we stated that the goal of this project was to better understand the civic tech ecosystem in France by using data analysis tools. Through the collection of data from LinkedIn profiles and from the Pappers' website, we started putting together a dataset on this topic. Part of the work was to define the structure for the database we are building, based on the data collected but also on the methods we wanted to use for analysis.

Regarding my first objective, which was to gain a broader understanding of civic tech in France through quantitative approaches, the data analysis returned some interesting insights. First of all, the people in civic tech, their professional experiences and their education trajectories are interesting to study. Understanding the people is necessary to better understand the skills and competencies which will define the civic technology professions and how digital citizen participation tools are produced.

Although most of the companies are registered as computer programming companies, we see that the people who are visible in the field mostly have a general university education and have attended business or political science sciences schools. Very few have a background in computer programming. To improve our analysis, we need to find additional data sources (more companies, more employees, but also a count of employees per year. In fact, we see that most of the companies don't have employees, and only 7 have more than 10 employees. This could also be explained by the instability of the market, but could also be contextual.

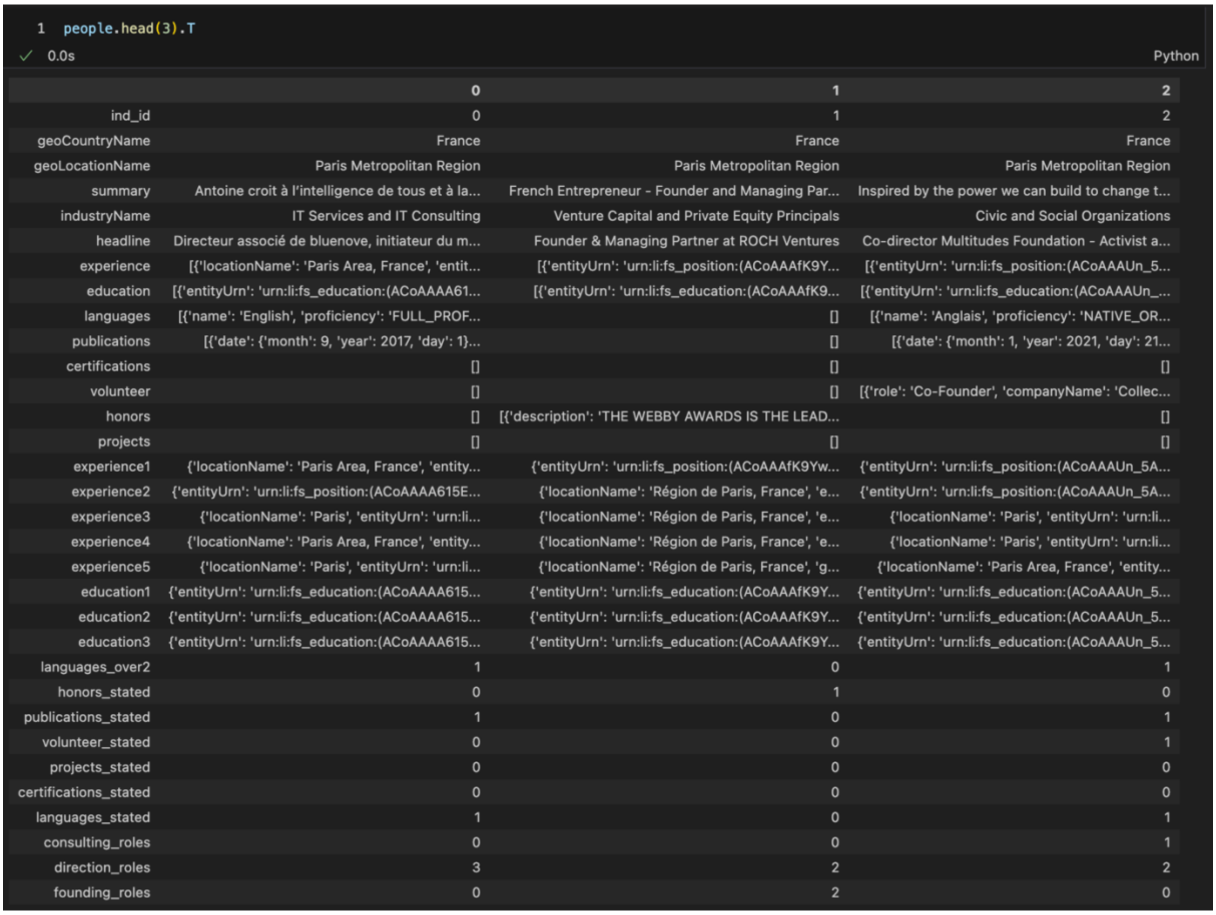
In this project, I focused especially on qualifying and understanding data points to analyze the organizations, i.e. companies and NGOs, and how to relate them to the people who work in the field. Focusing on building quantitative data and categorical variables was also important in preparation for the machine learning processing part of the project. Some of the data was not surprising. For instance, we see that most companies and NGOs are located in the parisian region. We also notice that there was a peak in civic tech startup and NGO creations between 2014 and 2017.

Regarding the market, having more precise data on companies' revenue per year allowed to see that the ecosystem is still quite unstable: there is an important variation in companies' revenues from one year to another. The market seems to be split in two groups: a lot of small start-ups which have very low sales amounts, and a few important companies with important sales amounts. However, even these companies are not 'safe': Cap Collectif, the company that provided the platform for the National Grand Débat in 2019, saw its sales decrease in 2020, as well as its net revenue.

Overall, the project was especially interesting to achieve my second objective, which was to develop a framework for data collection and cleaning, database construction and analysis for "young" ecosystems, where data is not yet available. The next steps will be to keep populating the database and to improve the code, namely to further automate the actions and to improve functions to account for variation in data. In parallel, it will be interesting to work with other researchers who want to test the framework to see if it is applicable to other ecosystems, and compare our results.

Annex 1 – Screenshots of database tables (first rows)

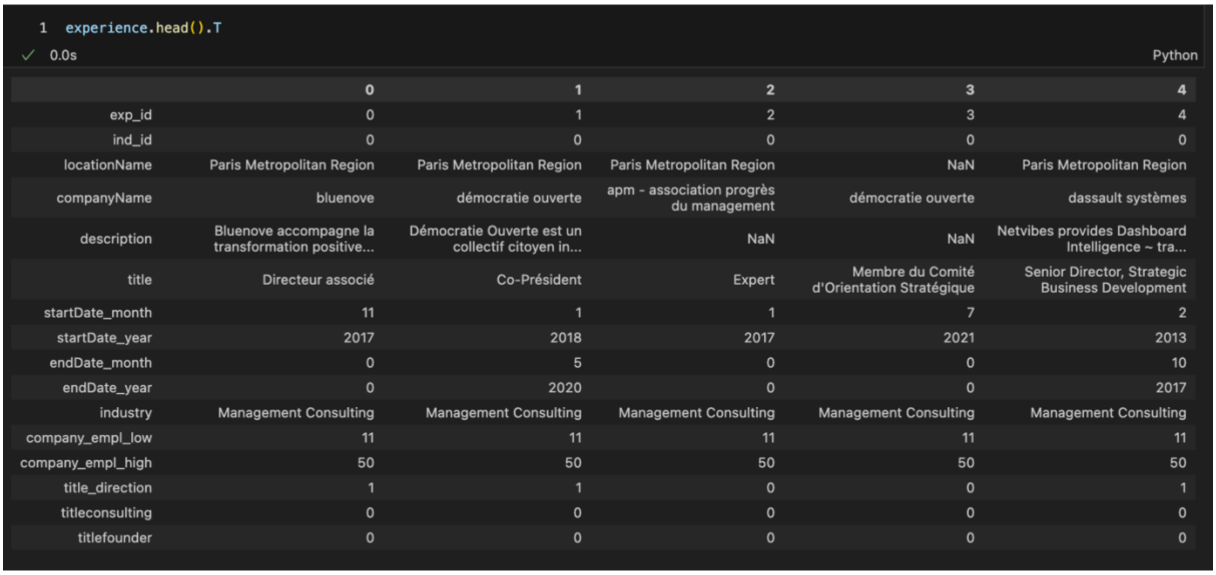
1. People (transposed)



```
people.head(3).T
```

	0	1	2
ind_id	0	1	2
geoCountryName	France	France	France
geoLocationName	Paris Metropolitan Region	Paris Metropolitan Region	Paris Metropolitan Region
summary	Antoine croit à l'intelligence de tous et à la...	French Entrepreneur - Founder and Managing Par...	Inspired by the power we can build to change t...
industryName	IT Services and IT Consulting	Venture Capital and Private Equity Principals	Civic and Social Organizations
headline	Directeur associé de bluenove, initiateur du m...	Founder & Managing Partner at ROCH Ventures	Co-director Multitudes Foundation - Activist a...
experience	[{'locationName': 'Paris Area, France', 'entit...	[{'entityUrn': 'urn:li:fs_position:(ACoAAAfK9Y...	[{'entityUrn': 'urn:li:fs_position:(ACoAAAUn_5...
education	[{'entityUrn': 'urn:li:fs_education:(ACoAAAA61...	[{'entityUrn': 'urn:li:fs_education:(ACoAAAfK9...	[{'entityUrn': 'urn:li:fs_education:(ACoAAAUn_...
languages	[{'name': 'English', 'proficiency': 'FULL_PROF...	[]	[{'name': 'Anglais', 'proficiency': 'NATIVE_OR...
publications	[{'date': {'month': 9, 'year': 2017, 'day': 1}...	[]	[{'date': {'month': 1, 'year': 2021, 'day': 21...
certifications	[]	[]	[]
volunteer	[]	[]	[{'role': 'Co-Founder', 'companyName': 'Collec...
honors	[]	[{'description': 'THE WEBBY AWARDS IS THE LEAD...	[]
projects	[]	[]	[]
experience1	{'locationName': 'Paris Area, France', 'entity...	{'entityUrn': 'urn:li:fs_position:(ACoAAAfK9Yw...	{'entityUrn': 'urn:li:fs_position:(ACoAAAUn_5A...
experience2	{'entityUrn': 'urn:li:fs_position:(ACoAAAA615E...	{'locationName': 'Région de Paris, France', 'e...	{'entityUrn': 'urn:li:fs_position:(ACoAAAUn_5A...
experience3	{'locationName': 'Paris', 'entityUrn': 'urn:li...	{'locationName': 'Région de Paris, France', 'e...	{'locationName': 'Paris', 'entityUrn': 'urn:li...
experience4	{'locationName': 'Paris Area, France', 'entity...	{'locationName': 'Région de Paris, France', 'e...	{'locationName': 'Paris', 'entityUrn': 'urn:li...
experience5	{'locationName': 'Paris', 'entityUrn': 'urn:li...	{'locationName': 'Région de Paris, France', 'g...	{'locationName': 'Paris Area, France', 'entity...
education1	{'entityUrn': 'urn:li:fs_education:(ACoAAAA615...	{'entityUrn': 'urn:li:fs_education:(ACoAAAfK9Y...	{'entityUrn': 'urn:li:fs_education:(ACoAAAUn_5...
education2	{'entityUrn': 'urn:li:fs_education:(ACoAAAA615...	{'entityUrn': 'urn:li:fs_education:(ACoAAAfK9Y...	{'entityUrn': 'urn:li:fs_education:(ACoAAAUn_5...
education3	{'entityUrn': 'urn:li:fs_education:(ACoAAAA615...	{'entityUrn': 'urn:li:fs_education:(ACoAAAfK9Y...	{'entityUrn': 'urn:li:fs_education:(ACoAAAUn_5...
languages_over2	1	0	1
honors_stated	0	1	0
publications_stated	1	0	1
volunteer_stated	0	0	1
projects_stated	0	0	0
certifications_stated	0	0	0
languages_stated	1	0	1
consulting_roles	0	0	1
direction_roles	3	2	2
founding_roles	0	2	0

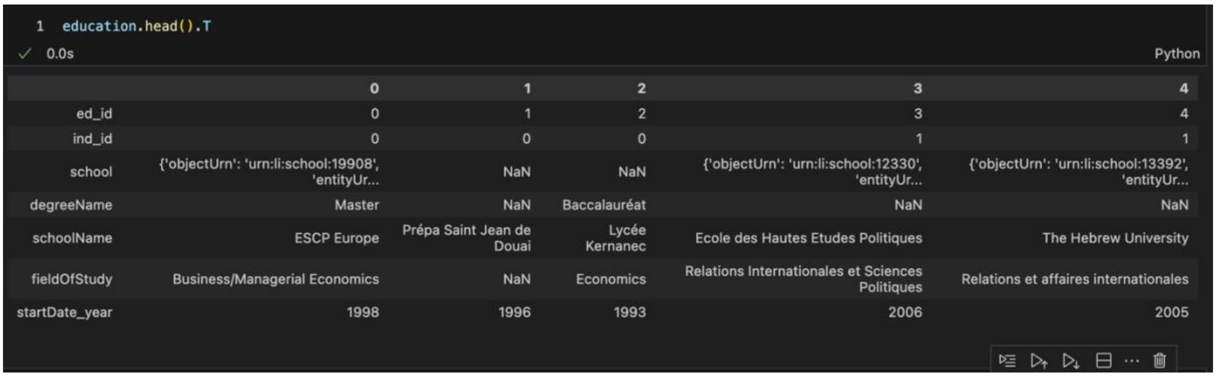
2. Experience (transposed)



```
experience.head().T
```

	0	1	2	3	4
exp_id	0	1	2	3	4
ind_id	0	0	0	0	0
locationName	Paris Metropolitan Region	Paris Metropolitan Region	Paris Metropolitan Region	NaN	Paris Metropolitan Region
companyName	bluenove	démocratie ouverte	apm - association progrès du management	démocratie ouverte	dassault systèmes
description	Bluenove accompagne la transformation positive...	Démocratie Ouverte est un collectif citoyen in...	NaN	NaN	Netvibes provides Dashboard Intelligence ~ tra...
title	Directeur associé	Co-Président	Expert	Membre du Comité d'Orientation Stratégique	Senior Director, Strategic Business Development
startDate_month	11	1	1	7	2
startDate_year	2017	2018	2017	2021	2013
endDate_month	0	5	0	0	10
endDate_year	0	2020	0	0	2017
industry	Management Consulting	Management Consulting	Management Consulting	Management Consulting	Management Consulting
company_empl_low	11	11	11	11	11
company_empl_high	50	50	50	50	50
title_direction	1	1	0	0	1
titleconsulting	0	0	0	0	0
titlefounder	0	0	0	0	0

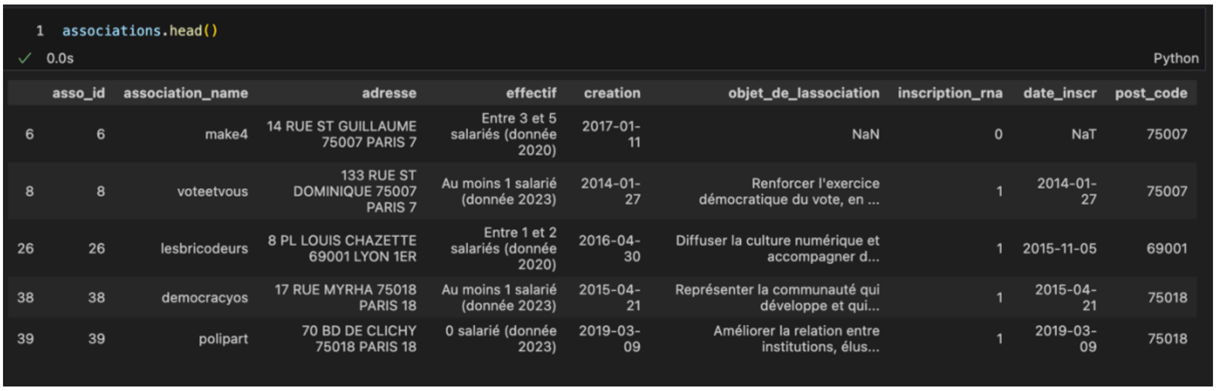
3. Education (transposed)



```python
education.head().T
```

	0	1	2	3	4
ed_id	0	1	2	3	4
ind_id	0	0	0	1	1
school	{'objectUrn': 'urn:li:school:19908', 'entityUr...	NaN	NaN	{'objectUrn': 'urn:li:school:12330', 'entityUr...	{'objectUrn': 'urn:li:school:13392', 'entityUr...
degreeName	Master	NaN	Baccalauréat	NaN	NaN
schoolName	ESCP Europe	Prépa Saint Jean de Douai	Lycée Kernanec	Ecole des Hautes Etudes Politiques	The Hebrew University
fieldOfStudy	Business/Managerial Economics	NaN	Economics	Relations Internationales et Sciences Politiques	Relations et affaires internationales
startDate_year	1998	1996	1993	2006	2005

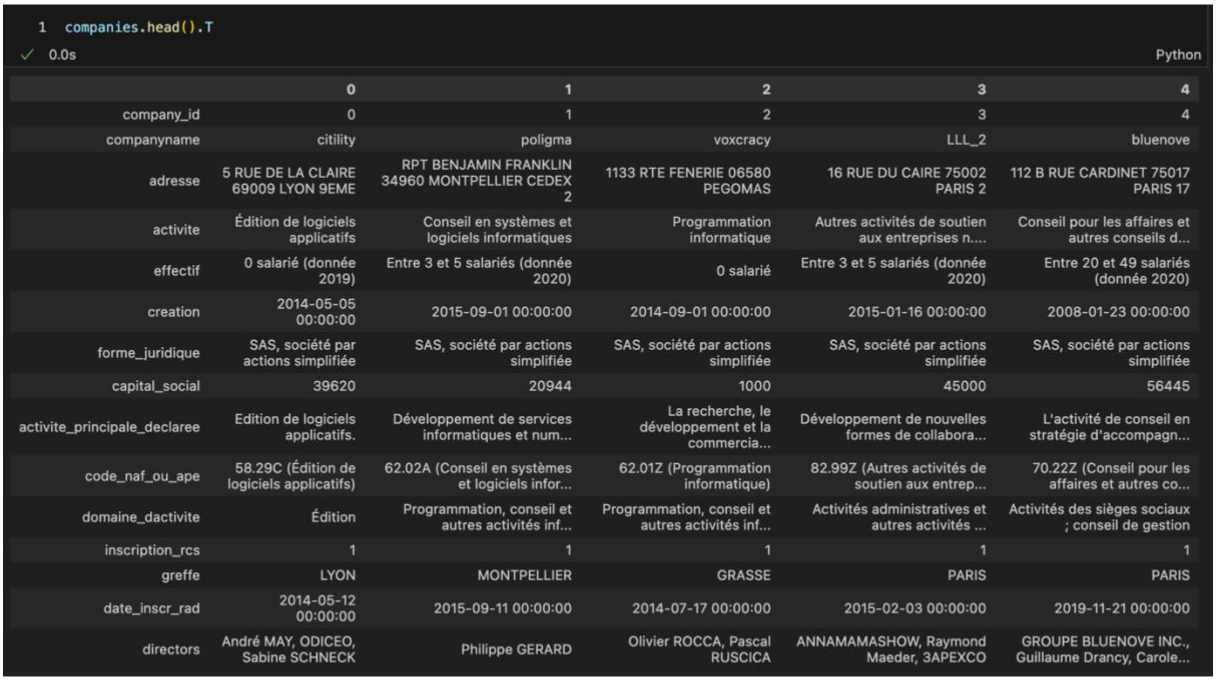
4. Associations



```python
associations.head()
```

	asso_id	association_name	adresse	effectif	creation	objet_de_lassociation	inscription_rna	date_inscr	post_code
6	6	make4	14 RUE ST GUILLAUME 75007 PARIS 7	Entre 3 et 5 salariés (donnée 2020)	2017-01-11	NaN	0	NaT	75007
8	8	voteetvous	133 RUE ST DOMINIQUE 75007 PARIS 7	Au moins 1 salarié (donnée 2023)	2014-01-27	Renforcer l'exercice démocratique du vote, en ...	1	2014-01-27	75007
26	26	lesbricodeurs	8 PL LOUIS CHAZETTE 69001 LYON 1ER	Entre 1 et 2 salariés (donnée 2020)	2016-04-30	Diffuser la culture numérique et accompagner d...	1	2015-11-05	69001
38	38	democracyos	17 RUE MYRHA 75018 PARIS 18	Au moins 1 salarié (donnée 2023)	2015-04-21	Représenter la communauté qui développe et qui...	1	2015-04-21	75018
39	39	polipart	70 BD DE CLICHY 75018 PARIS 18	0 salarié (donnée 2023)	2019-03-09	Améliorer la relation entre institutions, élus...	1	2019-03-09	75018

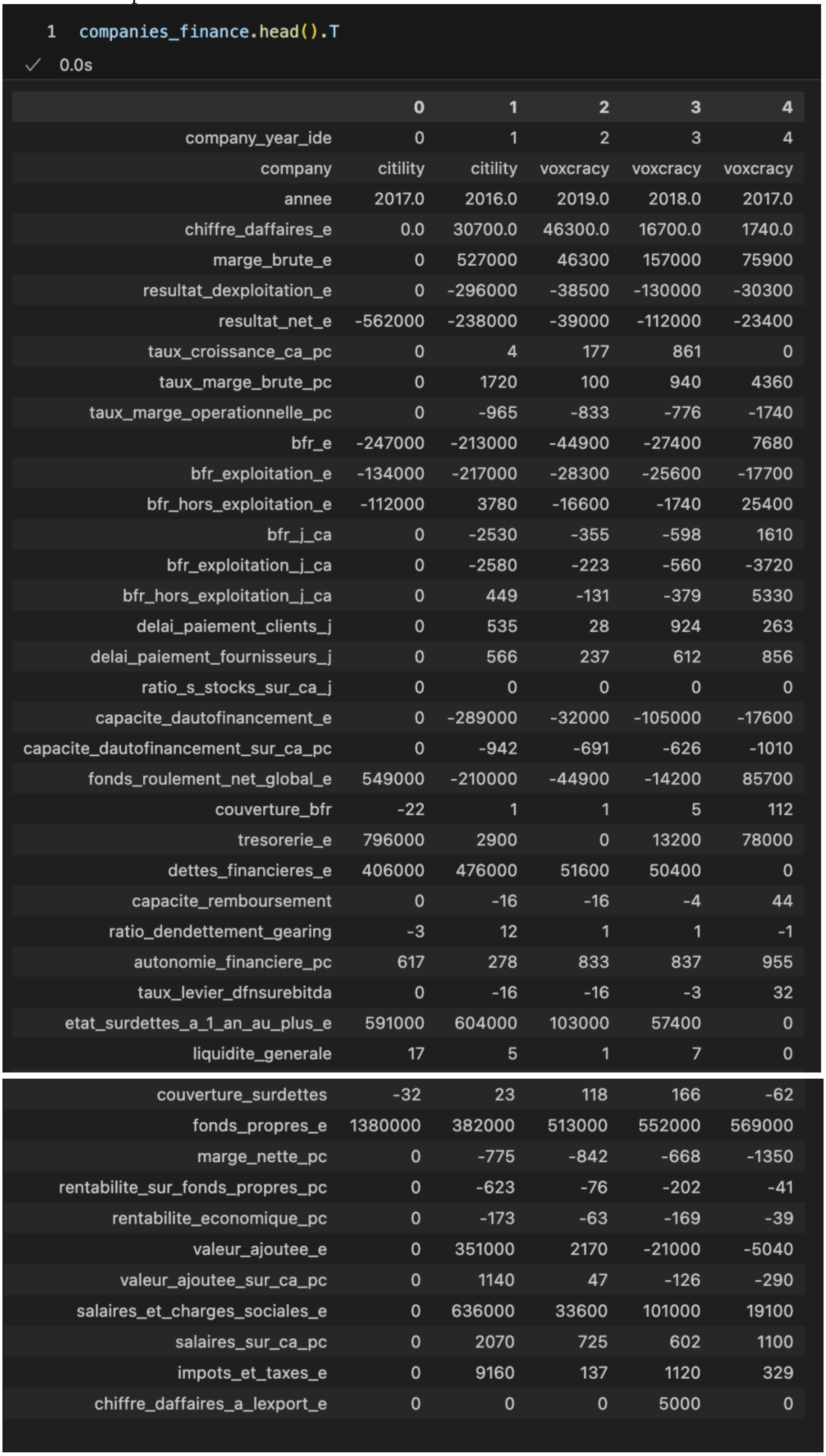
5. Companies information (transposed)



```python
companies.head().T
```

	0	1	2	3	4
company_id	0	1	2	3	4
companyname	citility	poligma	voxcracy	LLL_2	bluenove
adresse	5 RUE DE LA CLAIRE 69009 LYON 9EME	RPT BENJAMIN FRANKLIN 34960 MONTPELLIER CEDEX 2	1133 RTE FENERIE 06580 PEGOMAS	16 RUE DU CAIRE 75002 PARIS 2	112 B RUE CARDINET 75017 PARIS 17
activite	Édition de logiciels applicatifs	Conseil en systèmes et logiciels informatiques	Programmation informatique	Autres activités de soutien aux entreprises n....	Conseil pour les affaires et autres conseils d...
effectif	0 salarié (donnée 2019)	Entre 3 et 5 salariés (donnée 2020)	0 salarié	Entre 3 et 5 salariés (donnée 2020)	Entre 20 et 49 salariés (donnée 2020)
creation	2014-05-05 00:00:00	2015-09-01 00:00:00	2014-09-01 00:00:00	2015-01-16 00:00:00	2008-01-23 00:00:00
forme_juridique	SAS, société par actions simplifiée	SAS, société par actions simplifiée	SAS, société par actions simplifiée	SAS, société par actions simplifiée	SAS, société par actions simplifiée
capital_social	39620	20944	1000	45000	56445
activite_principale_declaree	Edition de logiciels applicatifs.	Développement de services informatiques et num...	La recherche, le développement et la commercia...	Développement de nouvelles formes de collabora...	L'activité de conseil en stratégie d'accompagn...
code_naf_ou_ape	58.29C (Édition de logiciels applicatifs)	62.02A (Conseil en systèmes et logiciels infor...	62.01Z (Programmation informatique)	82.99Z (Autres activités de soutien aux entrep...	70.22Z (Conseil pour les affaires et autres co...
domaine_dactivite	Édition	Programmation, conseil et autres activités inf...	Programmation, conseil et autres activités inf...	Activités administratives et autres activités ...	Activités des sièges sociaux ; conseil de gestion
inscription_rcs	1	1	1	1	1
greffe	LYON	MONTPELLIER	GRASSE	PARIS	PARIS
date_inscr_rad	2014-05-12 00:00:00	2015-09-11 00:00:00	2014-07-17 00:00:00	2015-02-03 00:00:00	2019-11-21 00:00:00
directors	André MAY, ODICEO, Sabine SCHNECK	Philippe GERARD	Olivier ROCCA, Pascal RUSCICA	ANNAMAMASHOW, Raymond Maeder, 3APEXCO	GROUPE BLUENOVE INC., Guillaume Drancy, Carole...

6. Companies financial information



```
companies_finance.head().T
```

```
✓ 0.0s
```

	0	1	2	3	4
company_year_ide	0	1	2	3	4
company	citility	citility	voxcracy	voxcracy	voxcracy
annee	2017.0	2016.0	2019.0	2018.0	2017.0
chiffre_daffaires_e	0.0	30700.0	46300.0	16700.0	1740.0
marge_brute_e	0	527000	46300	157000	75900
resultat_dexploitation_e	0	-296000	-38500	-130000	-30300
resultat_net_e	-562000	-238000	-39000	-112000	-23400
taux_croissance_ca_pc	0	4	177	861	0
taux_marge_brute_pc	0	1720	100	940	4360
taux_marge_operationnelle_pc	0	-965	-833	-776	-1740
bfr_e	-247000	-213000	-44900	-27400	7680
bfr_exploitation_e	-134000	-217000	-28300	-25600	-17700
bfr_hors_exploitation_e	-112000	3780	-16600	-1740	25400
bfr_j_ca	0	-2530	-355	-598	1610
bfr_exploitation_j_ca	0	-2580	-223	-560	-3720
bfr_hors_exploitation_j_ca	0	449	-131	-379	5330
delai_paiement_clients_j	0	535	28	924	263
delai_paiement_fournisseurs_j	0	566	237	612	856
ratio_s_stocks_sur_ca_j	0	0	0	0	0
capacite_dautofinancement_e	0	-289000	-32000	-105000	-17600
capacite_dautofinancement_sur_ca_pc	0	-942	-691	-626	-1010
fonds_roulement_net_global_e	549000	-210000	-44900	-14200	85700
couverture_bfr	-22	1	1	5	112
tresorerie_e	796000	2900	0	13200	78000
dettes_financieres_e	406000	476000	51600	50400	0
capacite_remboursement	0	-16	-16	-4	44
ratio_dendettement_gearing	-3	12	1	1	-1
autonomie_financiere_pc	617	278	833	837	955
taux_levier_dfnsurebitda	0	-16	-16	-3	32
etat_surdettes_a_1_an_au_plus_e	591000	604000	103000	57400	0
liquidite_generale	17	5	1	7	0
couverture_surdettes	-32	23	118	166	-62
fonds_propres_e	1380000	382000	513000	552000	569000
marge_nette_pc	0	-775	-842	-668	-1350
rentabilite_sur_fonds_propres_pc	0	-623	-76	-202	-41
rentabilite_economique_pc	0	-173	-63	-169	-39
valeur_ajoutee_e	0	351000	2170	-21000	-5040
valeur_ajoutee_sur_ca_pc	0	1140	47	-126	-290
salaires_et_charges_sociales_e	0	636000	33600	101000	19100
salaires_sur_ca_pc	0	2070	725	602	1100
impots_et_taxes_e	0	9160	137	1120	329
chiffre_daffaires_a_lexport_e	0	0	0	5000	0