Data Analytics

Match Project

Finding partners, identifying companies

Javier PEYRIERE

April, 2024

1 Table of Contents

2 Introduction

When you are planning to open a business, when you are already running a business and growing (or not) it is never straightforward and easy to gather information about:

- Your market, who are the players?
- Your competition, who is already there running the show?
- Your potential partners, would it be possible to join forces?
- How good the others are doing
- If you need to expand, with whom to talk to?

In particular for entrepreneurs and SMEs. You always need to rely on experts and go through consultancy. Similarly, for bigger actors, companies that need to expand externally, it is always a challenge to identify a suitable possible target. In this case also we rely on experts and professionals to do the work for us.

The idea of the project is to make a tool to help, for possible market analysis, to get first answers and directions in identifying companies depending on the activities and size. Identifying key companies so you get a better grasp of the market you want to work in , or for possible partnerships.

From multisource data (Infogreffe, Cap-Fi, pappers, BnF, Insee…) and combining them, the idea is to be able to narrow down a starting point for more investigation and… more data gathering.

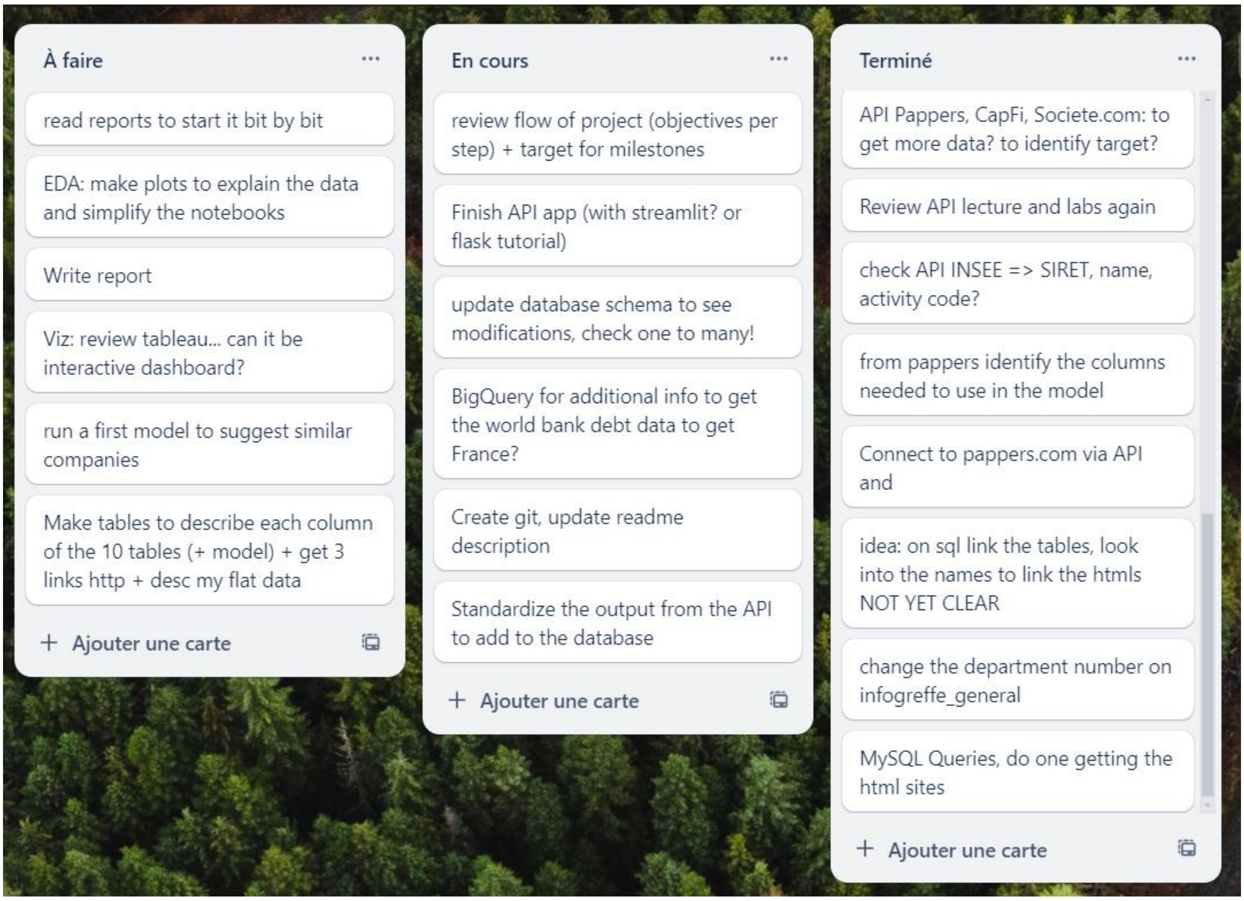
We are at the beginning of defining a tool, and many things can be done from the data gathered. For the sake of the exercise, we will look more in detail the codes APE 62.02A (Conseil en systèmes et logiciels informatiques) and APE 70.22Z (Conseil pour les affaires et autres conseils de gestion) in Paris. Can we understand a bit better the market from the data gathered? Could we infer some suggestions?

This reports presents the preliminary phases to build a prototype tool:

- After the introduction and brief project management methodology, a first part describes the gathering of the data from different sources while selecting possibly useful information.
- The next part presents the cleaning of the different data and adapting them to useful formats.
- Followed by a part about the Exploratory Data Analysis to have some data insights.
- The next part presents the creation of the data base.
- Continue by a part on how some data is shared via a preliminary API (more to come).
- The next part is the beginning of work on a Machine Learning model to see if there are patterns in our available data.
- Last but not least we shall share our conclusions and way forward.

3 Project Management

A Kanban board was used to manage daily project tasks (Trello).



À faire

- read reports to start it bit by bit
- EDA: make plots to explain the data and simplify the notebooks
- Write report
- Viz: review tableau... can it be interactive dashboard?
- run a first model to suggest similar companies
- Make tables to describe each column of the 10 tables (+ model) + get 3 links http + desc my flat data

En cours

- review flow of project (objectives per step) + target for milestones
- Finish API app (with streamlit? or flask tutorial)
- update database schema to see modifications, check one to many!
- BigQuery for additional info to get the world bank debt data to get France?
- Create git, update readme description
- Standardize the output from the API to add to the database

Terminé

- API Pappers, CapFi, Societe.com: to get more data? to identify target?
- Review API lecture and labs again
- check API INSEE => SIRET, name, activity code?
- from pappers identify the columns needed to use in the model
- Connect to pappers.com via API and
- idea: on sql link the tables, look into the names to link the htmls NOT YET CLEAR
- change the department number on infogreffe_general
- MySQL Queries, do one getting the html sites

Fig.: Snapshot of the Trello during the project (20th of April 2024)

From a daily brainstorm, different tasks were posted on the To Do list column ('A faire' in the figure). When started, they would move to the middle and to the very left when done.

On a daily basis the priority of each task would be checked and the task would be moved towards the top in the ongoing work.

4 GDPR : General Data Protection Regulation

The names of managers or owners of companies are mentioned in the database. The information comes from data published about the companies and can be found when connecting to different professional websites. No personal information is used.

However, in order to ensure full compliance, the table containing this information (management) is on purpose not share on the git.repository,

5 Data and data sources

The data comes from principally six different sources:

- the French National Library (Bibliothèque nationale de France) : BnF,
- the French Institute for statistics and economical studies (Institut national de la statistique et des études économiques) : Insee
- the French register of companies : Infogreffe,
- two professional websites tracking the financial data of companies: Cap-Fi, pappers,

- and the Big Query World Bank for debt data.

It was possible to join all the tables, sometimes via created junction tables, except for the World Bank data that was too high level to connect.

5.1 Web Scrapping

The BnF has a specific entity, PRISME, focusing on business oriented information. It is from this entity that we could check on some insights regarding possible web links .

On the following link, https://bnf.libguides.com/signets_prisme/sectoriels , were gathered, per BnF activity domains websites/links of corporations and associations linked with different lines of work or domains of companies.

By scrapping such data we could create, later on, a link to websites to look for information about a domain of activity starting from a company id number (siren) or a code 'APE'.

In a python code, using the *requests* and *bs4* (for BeautifulSoup) librairies to gather the data of the table in a dataframe format we were able to scrap the data from the link mentioned above.

```
import requests
from bs4 import BeautifulSoup
```

A first dataframe was generated, which was then exploded into a second dataframe with one line per website.

bnf Table (45 x 4)

#	columns	Description
1	bnf_codes	Internal codes to BnF, *unique key*
2	bnf_names	Description of the BnF code
3	web_sites_description	list of brief descriptions
4	web_sites	list of links web sites

Description of the columns of bnf Table, 45 rows and 4 columns

bnf_exploded Table (1962 x 4)

#	bnf_exploded	Description
1	bnf_codes	Internal codes to BnF
2	bnf_names	Description of the BnF code
3	web_sites_description	Brief description of the web site
4	web_sites	link to the web site, *unique key*

Description of the columns of bnf_exploded Table, 1962 rows and 4 columns

Of the resulting two tables (see above), the 'exploded' one was exported to the data base using pandas and sqalchemy librairies, as well as a csv backup in case of need.

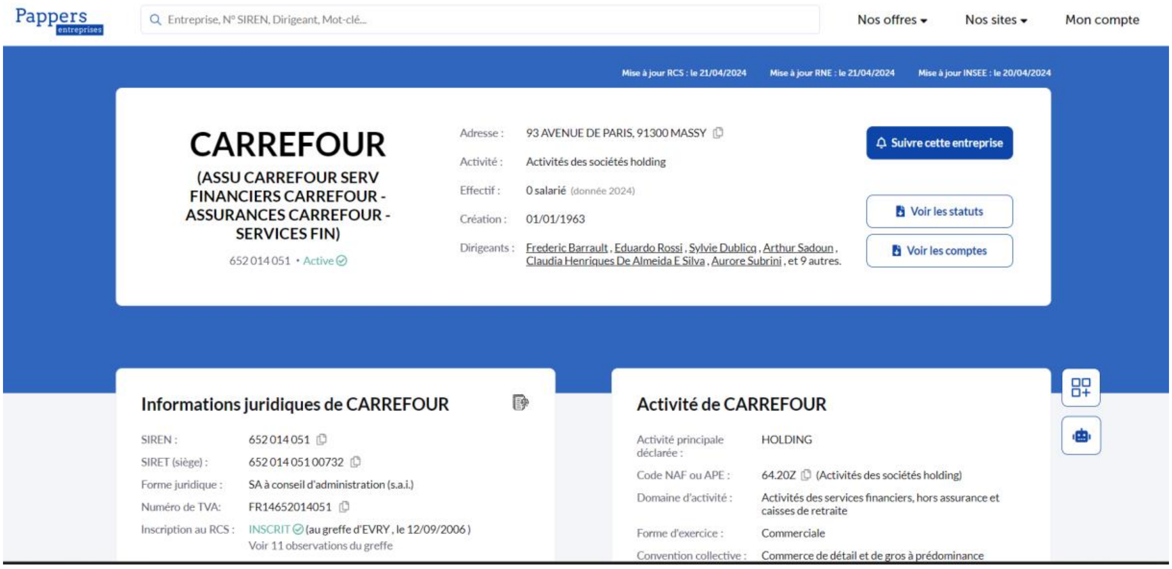
```
df_exploded.to_sql('bnf_exploded',engine, 'match_project', if_exists=’replace'
, index=False) # 'pushing' the data

df.to_csv("bnf_links.csv")
df_exploded.to_csv("bnf_links_exploded.csv")
```

A conversion table to link it with other data will be presented in the following chapter when mentioning importing the codes ape information.

5.2 API

The website pappers (https://www.pappers.fr/), allows to collect all legal and financial information on the companies of your choice (statutes, social accounts, brands, managers, shareholders).



Mise à jour RCS : le 21/04/2024 Mise à jour RNE : le 21/04/2024 Mise à jour INSEE : le 20/04/2024

CARFOUR

CARREFOUR

(ASSU CARREFOUR SERV FINANCIERS CARREFOUR - ASSURANCES CARREFOUR - SERVICES FIN)

652 014 051 • Active

Adresse : 93 AVENUE DE PARIS, 91300 MASSY

Activité : Activités des sociétés holding

Effectif : 0 salarié (donnée 2024)

Création : 01/01/1963

Dirigeants : Frederic Barrault , Eduardo Rossi , Sylvie Dublicq , Arthur Sadoun , Claudia Henriques De Almeida E Silva , Aurore Subrini , et 9 autres.

Informations juridiques de CARREFOUR

SIREN : 652 014 051

SIRET (siège) : 652 014 051 00732

Forme juridique : SA à conseil d'administration (s.a.i.)

Numéro de TVA: FR14652014051

Inscription au RCS : INSCRIT (au greffe d'EVRY , le 12/09/2006)
Voir 11 observations du greffe

Activité de CARREFOUR

Activité principale déclarée : HOLDING

Code NAF ou APE : 64.20Z (Activités des sociétés holding)

Domaine d'activité : Activités des services financiers, hors assurance et caisses de retraite

Forme d'exercice : Commerciale

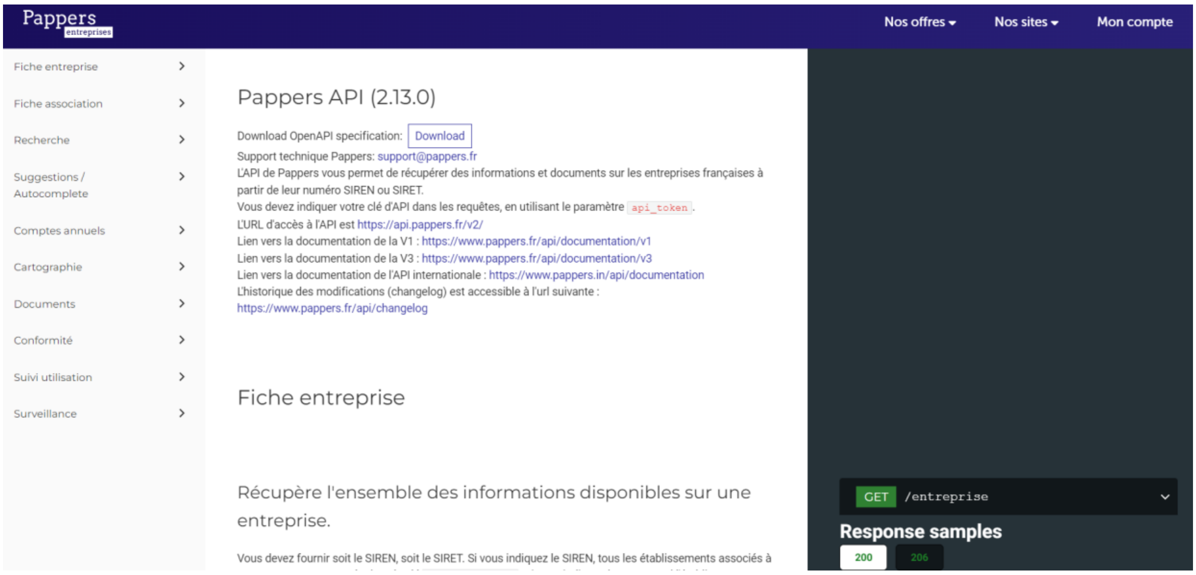
Convention collective : Commerce de détail et de gros à prédominance

Fig.: Screenshot example of the website, focused on a company

Our original intention was to get as much data as possible via the API, however the free connection only allows for 100 credits per month.

Even though, it could not be used as the main source of data, it still is worth connecting via the API in order to:

- download the information on the “future identified companies”
- to create a possible entry point for the machine learning
- to check on the data to see if it could used as additional input in the database (filling the data base with updated or missing information



Pappers API (2.13.0)

Download OpenAPI specification: Download
Support technique Pappers: support@pappers.fr
L'API de Pappers vous permet de récupérer des informations et documents sur les entreprises françaises à partir de leur numéro SIREN ou SIRET.
Vous devez indiquer votre clé d'API dans les requêtes, en utilisant le paramètre `api_token`.
L'URL d'accès à l'API est https://api.pappers.fr/v2/
Lien vers la documentation de la V1 : https://www.pappers.fr/api/documentation/v1
Lien vers la documentation de la V3 : https://www.pappers.fr/api/documentation/v3
Lien vers la documentation de l'API internationale : https://www.pappers.in/api/documentation
L'historique des modifications (changelog) est accessible à l'url suivante :
https://www.pappers.fr/api/changelog

Fiche entreprise

Récupère l'ensemble des informations disponibles sur une entreprise.

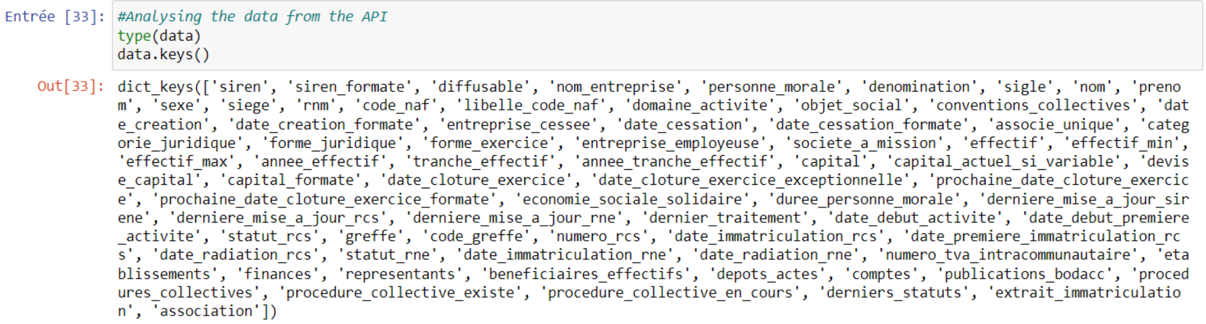
Vous devez fournir soit le SIREN, soit le SIRET. Si vous indiquez le SIREN, tous les établissements associés à

Fig.: screenshot of the API documentation of the pappers.fr website

When connecting to https://api.pappers.fr/v2/entreprise , using the appropriate token, specific data can be downloaded when providing the company Id (Siren number).

The data covers the full information about the company, from financials to managers, used in a dictionary type of format (json).

A very time consuming exercise was to identify the potentially useful data to keep from the importation of the data. The information obtained via the API is very complete and the data as a dictionary has more than 70 keys, with values that could be themselves other dictionaries.

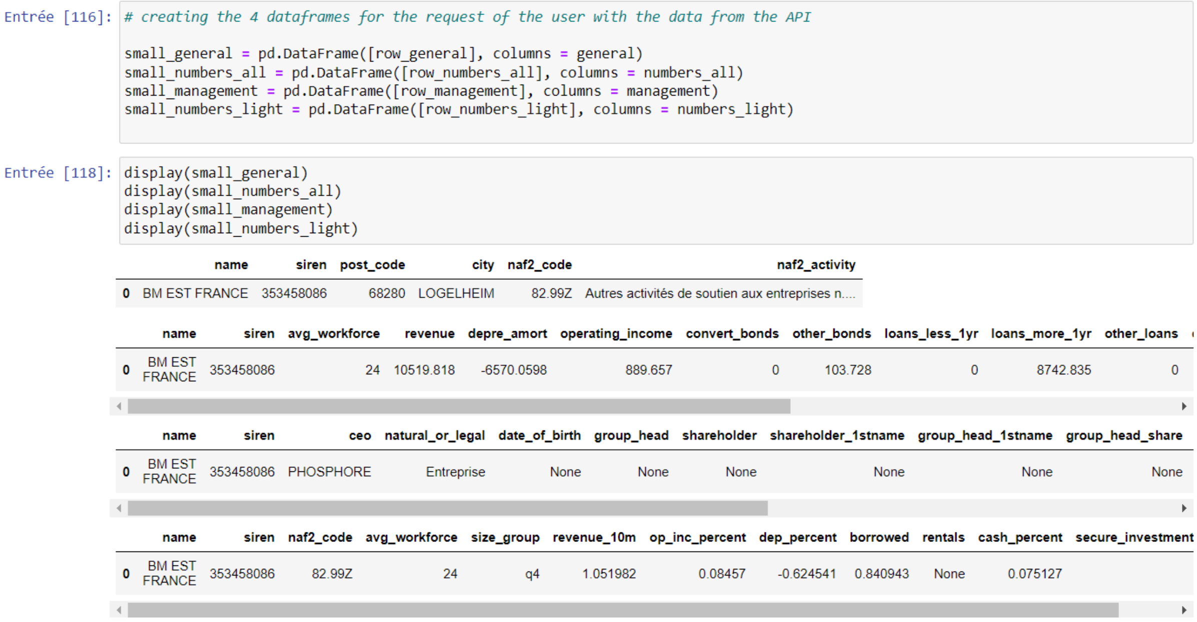


```
Entrée [33]: #Analysing the data from the API
             type(data)
             data.keys()

   Out[33]: dict_keys(['siren', 'siren_formate', 'diffusable', 'nom_entreprise', 'personne_morale', 'denomination', 'sigle', 'nom', 'preno
             m', 'sexe', 'siege', 'rnm', 'code_naf', 'libelle_code_naf', 'domaine_activite', 'objet_social', 'conventions_collectives', 'dat
             e_creation', 'date_creation_formate', 'entreprise_cessee', 'date_cessation', 'date_cessation_formate', 'associe_unique', 'categ
             orie_juridique', 'forme_juridique', 'forme_exercice', 'entreprise_employeuse', 'societe_a_mission', 'effectif', 'effectif_min',
             'effectif_max', 'annee_effectif', 'tranche_effectif', 'annee_tranche_effectif', 'capital', 'capital_actuel_si_variable', 'devis
             e_capital', 'capital_formate', 'date_cloture_exercice', 'date_cloture_exercice_exceptionnelle', 'prochaine_date_cloture_exercic
             e', 'prochaine_date_cloture_exercice_formate', 'economie_sociale_solidaire', 'duree_personne_morale', 'derniere_mise_a_jour_sir
             ene', 'derniere_mise_a_jour_rcs', 'derniere_mise_a_jour_rne', 'dernier_traitement', 'date_debut_activite', 'date_debut_premiere
             _activite', 'statut_rcs', 'greffe', 'code_greffe', 'numero_rcs', 'date_immatriculation_rcs', 'date_premiere_immatriculation_rc
             s', 'date_radiation_rcs', 'statut_rne', 'date_immatriculation_rne', 'date_radiation_rne', 'numero_tva_intracommunautaire', 'eta
             blissements', 'finances', 'representants', 'beneficiaires_effectifs', 'depots_actes', 'comptes', 'publications_bodacc', 'proced
             ures_collectives', 'procedure_collective_existe', 'procedure_collective_en_cours', 'derniers_statuts', 'extrait_immatriculatio
             n', 'association'])
```

Example: list of the keys of the dictionary obtain

In order to be able to use the imported data, four(4) DataFrames (“small_general”, “small_management”, “small_numbers_light”, “small_numbers_all”) of 1 line are created to mimic the format of some of the existing flat flies presented in 4.3.



```
Entrée [116]: # creating the 4 dataframes for the request of the user with the data from the API

              small_general = pd.DataFrame([row_general], columns = general)
              small_numbers_all = pd.DataFrame([row_numbers_all], columns = numbers_all)
              small_management = pd.DataFrame([row_management], columns = management)
              small_numbers_light = pd.DataFrame([row_numbers_light], columns = numbers_light)
```

```
Entrée [118]: display(small_general)
              display(small_numbers_all)
              display(small_management)
              display(small_numbers_light)
```

	name	siren	post_code	city	naf2_code	naf2_activity
0	BM EST FRANCE	353458086	68280	LOGELHEIM	82.99Z	Autres activités de soutien aux entreprises n....

	name	siren	avg_workforce	revenue	depre_amort	operating_income	convert_bonds	other_bonds	loans_less_1yr	loans_more_1yr	other_loans
0	BM EST FRANCE	353458086	24	10519.818	-6570.0598	889.657	0	103.728	0	8742.835	0

	name	siren	ceo	natural_or_legal	date_of_birth	group_head	shareholder	shareholder_1stname	group_head_1stname	group_head_share
0	BM EST FRANCE	353458086	PHOSPHORE	Entreprise	None	None	None	None	None	None

	name	siren	naf2_code	avg_workforce	size_group	revenue_10m	op_inc_percent	dep_percent	borrowed	rentals	cash_percent	secure_investment
0	BM EST FRANCE	353458086	82.99Z	24	q4	1.051982	0.08457	-0.624541	0.840943	None	0.075127	

Example of the resulting dataframes for the company ‘BM EST France’ obtained via the API

The specific columns are described in 4.3. Some columns had to be calculated from the finance part of the dictionary, some others were directly matched.

Note that some information is empty as missing in the website. However some additional information exist and with time the format and information from *pappers* could be used as main reference.

5.3 Flat Files and Open Sources

5.3.1 Infogreffe

On the Infogreffe web site opensource data is available. Amongst other information, the key figures of commercial companies having filed their annual accounts for the 2022 financial years, enriched with the years 2020 and 2021 are available.

The 2022 version of the data (latest complete update) was retrieved as a csv file from the following link:

https://opendata.datainfogreffe.fr/explore/dataset/chiffres-cles-2022/information/

The original file is a table of 624,154 x 41.

After wrangling and cleaning it has been split into two tables infogreffe_general and infogreffe_numbers to be imported in our database, keeping only data that could be of use for an in depth analysis or possible machine learning.

The description of the columns of the tables is as follow:

- infogreffe_general = 624,042 x 17
- infogreffe_numbers = 624,042 x 18

infogreffe_general Table (624,042x 17)			infogreffe_numbers Table (624,042 x 18)		
#	Column	Description	#	Column	Description
1	denomination	Name of entity	1	denomination	Name of entity
2	siren	Company id number, 9 digits	2	siren	Company id number, 9 digits
3	nic	5 digits specific to each establishment	3	code_postal	Post code
4	forme_juridique	Legal form company	4	millesime_1	Latest refered year (year 1)
5	code_ape	Main activity code	5	duree_1	Active months year 1
6	libelle_ape	Main activity description	6	ca_1	Revenue year 1
7	adresse	Adress	7	resultat_1	Income year 1
8	code_postal	Post code	8	millesime_2	Year before year 1 (year 2)
9	ville	City	9	duree_2	Active months year 2
10	num_dept	French department number	10	ca_2	Revenue year 2
11	departement	French department name	11	resultat_2	Income year 2
12	region	French region	12	millesime_3	Year before year 2 (year 3)
13	code_greffe	Commercial court id	13	duree_3	Active months year 3
14	greffe	Commercial court name	14	ca_3	Revenue year 3
15	date_immatriculation	Date of registration	15	resultat_3	Income year 3
16	date_de_publication	Last data update date	16	tranche_ca_millesime_1	Revenue category for year 1
17	name_in_capitals	Entity name in capital letters	17	tranche_ca_millesime_2	Revenue category for year 2
			18	tranche_ca_millesime_3	Revenue category for year 3

5.3.2 Insee

On the Insee websites, the description of the activity code and nomenclatures are described. They are necessary for a better grasp of the code_ape mentioned on the infogreffe tables, but also on the ones to come.

Codes APE were extracted as csv files from https://www.insee.fr/fr/information/2028155

In a nutshell, the NAF2 Classification system gathers the information at several level of granularity.

Hence there is a ‘tree’ that presents Classifications per activity:

- 21 sections divided in
    - 88 divisions divided in
        - 272 groups divided in
            - 615 classes and ending with
                - 732 sub-classes

What is commonly used are these 732 sub classes that depending on the databases could be called ‘code_ape’, ‘naf2’, …

A first table is used to summarize the 732 sub classes, conversion codes (French, , Old French, European, International correspondances).

Originally naf2 was 732 x 23 and was reduced to 732 x 10, here are the resulting columns:

naf2 Table (732 x 10)		
#	Column	Description
1	sub classes	code_ape or naf2_code)
2	description	description of the sub_class
3	class	class code
4	group	group code
5	division	division code
6	A129	A129 classification code
7	A64	A64 classification code
8	A38	A38 classification code
9	A17	A17 classification code
10	A10	A10 classification code

A second table with the 88 divisions description: naf2_agreg = 88 x 4, here are the resulting columns:

naf2_agreg Table (88 x 4)		
#	Column	Description
1	section	section id
2	section_description	section description
3	code_division	division id
4	Description_division	Division description

These two tables link all the classifications to all the company data. However it still needs a further step to link it to the BnF data mentioned in 4.1.

Creating a junction table:

From bnf_exploded we got the bnf_names (description of the codes) and compared them to the description_division from the naf2_agreg file.

For such an operation, we used the fuzzywuzzy library that compares and matches values of two columns. The similarity score was not perfect, however it is very similar to the manual one I had started (but much faster). This is the agreg_to_bnf table:

agreg_to_bnf Table (45 x 3)		
#	Column	Description
1	bnf_names	Description of the BnF code
2	Description_division	Division description
3	similarity_score	Score of how much the two other columns related to each other in words

5.3.3 Personnal – Cap Fi

Another batch of data comes from personal 2022 excel files we had, extracted most likely from the Cap-Fi database

It gathers, similar data as in the website of *pappers*, with more details in terms of information regarding the balance sheet (hence both sources could possibly be used with time).

The four excel spreadsheets gathered the 2022 information for companies in the Ile-de-France region ("départements: 75, 77, 78, 91, 92, 93, 94, 95") from 2.8 million € and above (reaching the 55 billion € for some entities).

These excel tables where respectively of: (65,535 x 39) ; (21,216 x 39) ; (33,084 x 39) and (15,471 x 39) lines x columns, creating a table of (16,932 x 39).

From the four excel spreadsheets, the data was cleaned and wrangled (see section 5 Data cleaning) resulting in four tables gathering the information as follows.

general Table (16,932 x 6)

#	Column	Description
1	denomination	Name of entity
2	siren	Company id number, 9 digits
3	post_code	post_code
4	city	city
5	naf2_code	code ape or sous_classe
6	naf2_activity	description of the naf2_code

This table shows the basic general data for each company of the Ile de France dataset.

numbers light Table (16,932 x 12)

#	Column	Description
1	denomination	Name of entity
2	siren	Company id number, 9 digits
3	naf2_code	code ape or sous_classe
4	avg_workforce	calculated from the min and max of the class size
5	size_group	decile of the distribution (for the companies with the information)
6	revenue_10m	revenue in 2022, in 10Millions
7	op_inc_percent	operating income divided by the revenue
8	dep_percent	depreciation divided by revenue
9	borrowed	all debts summed divided by revenue
10	rentals	all rentals yearly divided by revenue
11	cash_percent	cash divided by revenue
12	secure_investment_percent	amount of investment divided by revenue

This table gathers the numbers that could possibly be used for the prediction models.

management Table (16,932 x 14)

#	Column	Description
1	denomination	Name of entity
2	siren	Company id number, 9 digits
3	ceo	Name of the top manager/company
4	natural_or_legal	natural (personne physique) or legal (personne morale) entity

5	date_of_birth	year of birth of the manager
6	group_head	Owner (main shareholder)
7	shareholder	Name of the next shareholder
8	shareholder_1stname	first name of the shareholder
9	group_head_1stname	first name of the owner
10	group_head_share	percentage of shares
11	shareholder_share	percentage of shares
12	reference_shareholder	reference for contacts
13	reference_shareholder_1stname	first name of reference
14	reference_shareholder_share	percentage of shares of reference

This table gathers information suitable to get in contact with the companies, if there is an interest.

numbers_all Table (16,932 x 17)

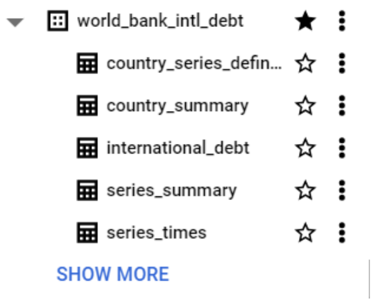
#	Column	Description
1	denomination	Name of entity
2	siren	Company id number, 9 digits
3	avg_workforce	calculated from the min and max of the class size
4	revenue	revenue in 2022 in K€
5	depre_amort	depreciation in 2022 in K€
6	operating_income	operating income in 2022 in K€
7	convert_bonds	convertible bonds in 2022 in K€
8	other_bonds	other bonds in 2022 in K€
9	loans_less_1yr	loans less than 1year 2022 in K€
10	loans_more_1yr	loans more than 1year 2022 in K€
11	other_loans	other loans in 2022 in K€
12	debts_to_group	debts to group in 2022 in K€
13	commit_mov_prop	rentals on movables in 2022 in K€
14	commit_immov_prop	rentals on immovables in 2022 in K€
15	bills_exch	bills echangeable 2022 in K€
16	cash	cash in bank end 2022 in K€
17	market_security	external investments 2022 in K€

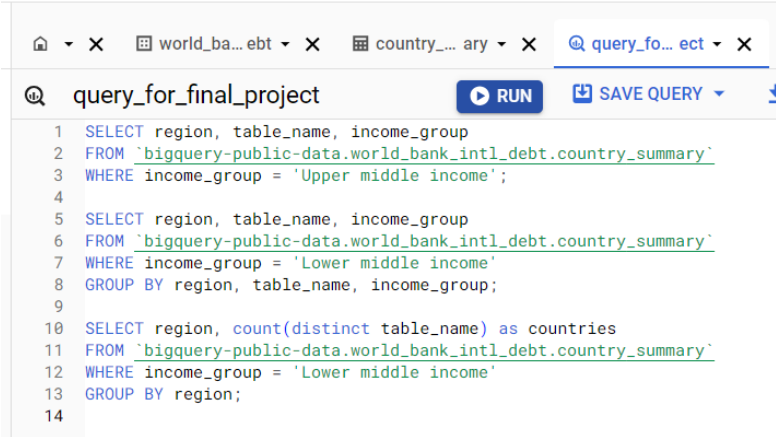
This is the table that gathers the numerical data in its original format in case other calculations should be done.

5.4 Big Query

After reviewing the data available, no data really matched with the topic of this report. However, we went into Big Query to run some queries in order to see if we could get some debt information about France.

The data was empty, so we looked into the eventual classification of countries with debt and ran some basic queries:

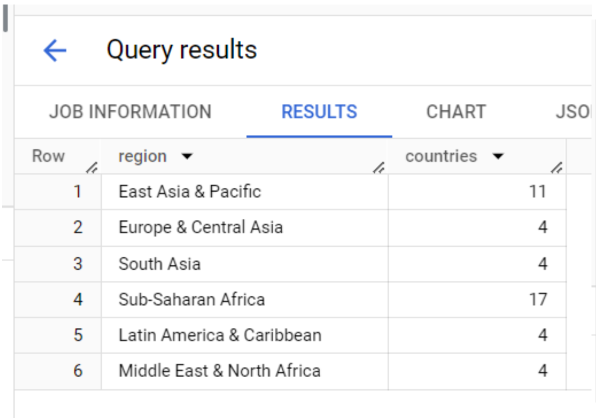




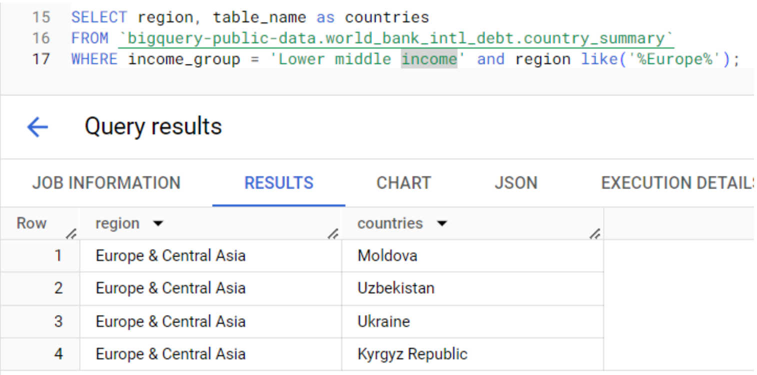
```
SELECT region, table_name, income_group
FROM `bigquery-public-data.world_bank_intl_debt.country_summary`
WHERE income_group = 'Upper middle income';

SELECT region, table_name, income_group
FROM `bigquery-public-data.world_bank_intl_debt.country_summary`
WHERE income_group = 'Lower middle income'
GROUP BY region, table_name, income_group;

SELECT region, count(distinct table_name) as countries
FROM `bigquery-public-data.world_bank_intl_debt.country_summary`
WHERE income_group = 'Lower middle income'
GROUP BY region;
```



Row	region	countries
1	East Asia & Pacific	11
2	Europe & Central Asia	4
3	South Asia	4
4	Sub-Saharan Africa	17
5	Latin America & Caribbean	4
6	Middle East & North Africa	4



```
SELECT region, table_name as countries
FROM `bigquery-public-data.world_bank_intl_debt.country_summary`
WHERE income_group = 'Lower middle income' and region like('%Europe%');
```

Row	region	countries
1	Europe & Central Asia	Moldova
2	Europe & Central Asia	Uzbekistan
3	Europe & Central Asia	Ukraine
4	Europe & Central Asia	Kyrgyz Republic

And finally zooming to see the countries in this category in Europe & Central Asia:

6 Data cleaning & wrangling

Data cleaning was relatively straightforward, data wrangling was more challenging as many columns had to be reviewed to be usable later.

The main steps were:

- checking for size, shape,
- checking for duplicates,
- column names for renaming and string formatting
- column selection understanding them (the longest),
- handling null values (keeping them or changing the values),
- filtering values,
- dropping columns and rows ,
- format checking and datatypes (int, float, strings, dates)
- creating columns
- concatenation, merging or splitting
- indexing

6.1 Infogreffe

The data had empty values, but not in the main columns, and the formats of the data were relatively clean.

```
# changing the names of the columns:
def clean_my_columns_titles(df):
    df.columns = df.columns.str.replace("
","_").str.replace(".","").str.replace("é","e").str.lower().str.strip()
    return df
clean_my_columns_titles(infogreffe)
```

Please find below, in bullet point format, which data were kept to be used in the two tables ***infogreffe_general*** & ***infogreffe_numbers***:

- denomination, siren, forme_juridique , nic (to make siret) as only 116 missing for this last

- code_ape & libelle_ape (only missing 13K)
- adresse, code_postal,ville, num_dept, departement, region (even if for the last two columns 6550 are missing, as it is ‘reconstructible’ if needed)
- code_greffe, greffe, date immatriculation, statut, date_de_publication, with no changes
- for years 1-2-3: millesime, date de cloture exercice, duree, ca, resultat, with no changes
- _ca_millesime 1-2-3 : good to filter (finally this categorization was not that handy)

Dropped:

- date radiation, geolocalisation, id (all empty)
- date_de cloture_exercice 1-2-3 (not using it),
- effectif 1-2-3 (too much missing info, more data in the other file)

6.2 bnf and Insee

bnf

The ***bnf*** data had no empty values. All the columns were kept (acquired via web scrapping).

Some clean up was done regarding the scrapped strings, and additional lines were generated in the exploded dataframe to be able to access more easily to the 1900 websites.

Insee

There were some empty columns as well as some non-usable old classification codes for the categories. Some renaming of columns was done, to identify them more easily. The formats were usable, the ‘.’ or spaces were removed for easy joining in the database

```
col_to_drop = ['Unnamed: 10','Unnamed: 11', 'Unnamed: 12', 'Unnamed: 13',
'Unnamed: 14', 'Unnamed: 15', 'Unnamed: 16',
               'Unnamed: 17', 'Unnamed: 18', 'Unnamed: 19', 'Unnamed: 20',
'Unnamed: 21', 'Unnamed: 22']
naf2.drop(col_to_drop, axis = 1, inplace = True)
```

6.3 Ile de France data

These are the tables that required the most work for the data to be prepared for the database.

- Imported the 4 excel files and converted them in pandas dataframes.
- Checked for duplicated
- Dropped the lines with no Company names, renamed and cleaned the columns of each dataframe
- Prepared the 4 dataframes for the concatenation (eg. dropping columns not to be used ["Nom du CAC", "Nom de la banque", "Filiales - Nom", "Filiales - Mnémo", "Filiales - NACE - Code principal"]…) & concatenation
- Translating the names of the columns in English

- Splitting numerical and categorical for the next changes
- Checking for null values, replacing them by ‘-1’
- Setting a date format

Then the numerical columns had to be prepared for use, find below the quick notes on how to simplify the columns and compare companies (to create the numbers_light table):

- avg_workforce: at least 25% are -1 (no data), to analyse without -1, so we get a better value. To re-run without, # hunch that it is right skewed, with high outliers affecting the mean (mean > median)
- revenue : min at 2800K€ (2.8 million) max at 55,500,000K€ (55.5 billion)
  - median at 9.8 million, mean at 6.5 million, 50% of the companies are between 5.5M€ and 23M€
  - we are in thousands K€ to million K€ (and to use 10M for numbers_light)
  - depreciation: negative values? to remove the negative outliers or to set to -1; then rerun the data without -1 , divide by revenue
  - operating income: check if there are many -1; to remove from the analysis, the others make sense, divide by revenue
- convert_bonds, other bonds: check if there are many -1; many at 0 to remove from the analysis, the others make sense
- loans less than 1 y, more than 1 y, other loans: negative values? to remove the negative outliers or to set to -1; then rerun the data without -1
- to add all the borrowed money: 'borrowed' = "convert_bonds", "other_bonds", "loans_less_1yr", "loans_more_1yr", "other_loans", "debts_to_group"; divide by revenue
- cash: check if there are many -1; to remove from the analysis, the others make sense, divide by revenue
- "commit_mov_prop", "commit_immov_prop": without the -1?: "commit_mov_prop", "commit_immov_prop" as 'rentals', divide by revenue
- "bills_exch"=> to double check the use
- "market_security" => to rename 'securities_investment', divide by revenue

7 EDA

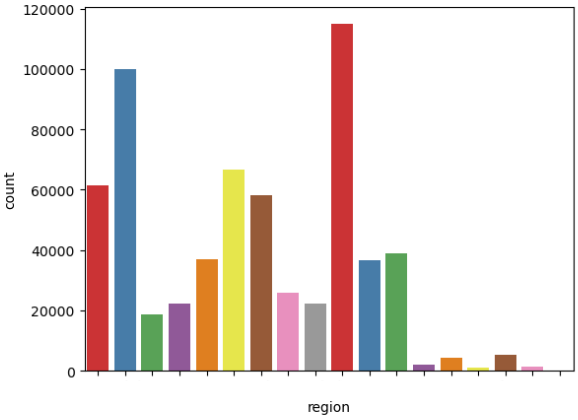
7.1 Looking at the Infogreffe Data

Categorical values: running a function for some selected columns that have less unique values than the others. The address, denominations and the like are not relevant to check in terms of distribution or statistics.

- region,

```
****** Brief analysis of region *****
region
Ile-de-France                           114873
Auvergne-Rhône-Alpes                    100003
Languedoc-Roussillon-Midi-Pyrénées       66774
Provence-Alpes-Côte d'Azur               61478
Aquitaine-Limousin-Poitou-Charentes      58281
Pays-de-la-Loire                         39066
Alsace-Champagne-Ardenne-Lorraine        37000
Nord-Pas-de-Calais-Picardie              36534
Bretagne                                 25725
Normandie                                22357
Bourgogne-Franche-Comté                  22236
Centre-Val de Loire                      18666
La Réunion                                5170
Corse                                     4378
Guadeloupe                                2171
Martinique                                1402
Guyane                                    1159
Mayotte                                    220
```

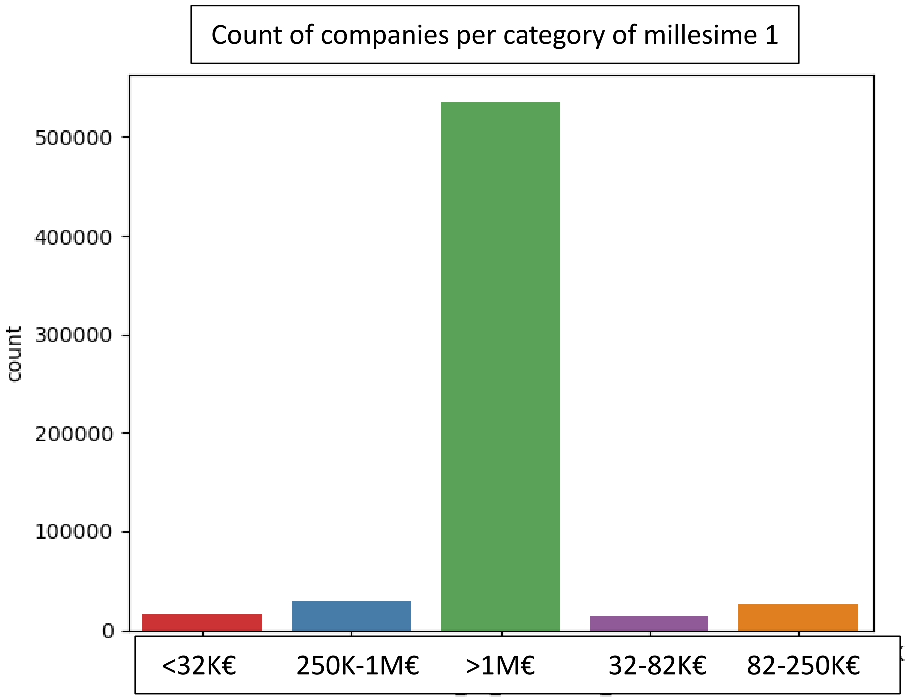
Count of companies per French region



Without surprise the most active regions in France are more represented.

- tranche_ca_millesime_1 (the data for the other millesime 2 and 3 are very similar)

```
****** Brief analysis of tranche_ca_millesime_1 *****
tranche_ca_millesime_1
E + d 1M                536448
D entre 250K et 1M       29673
C entre 82K et 250K      27421
A - de 32K               15474
B entre 32K et 82K       15026
Name: count, dtype: int64
tranche_ca_millesime_1
E + d 1M                0.859634
D entre 250K et 1M      0.047550
C entre 82K et 250K     0.043941
A - de 32K              0.024796
B entre 32K et 82K      0.024079
Name: proportion, dtype: float64
'mode = 0    E + d 1M\nName: tranche_ca_millesime_1, dtype: object'
```

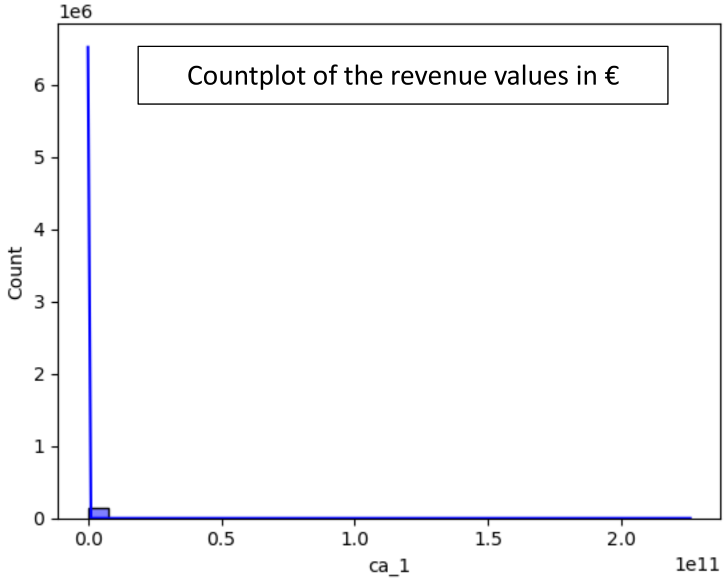


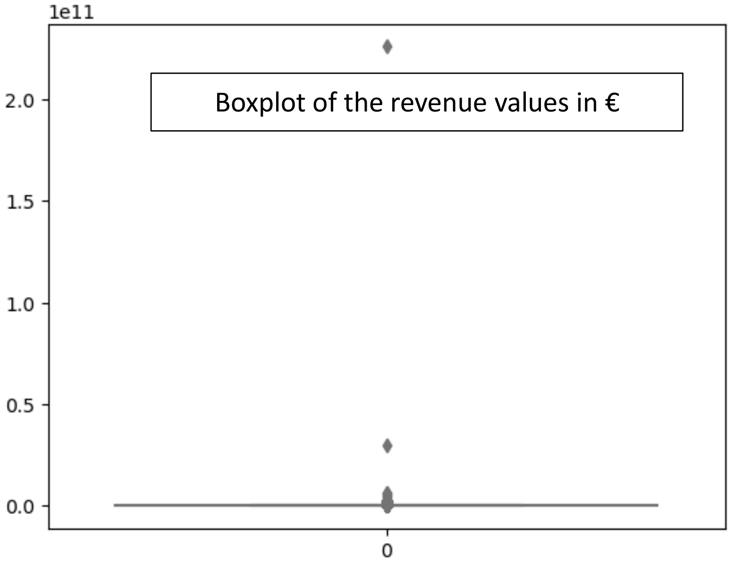
It is not the best tool to use as: it is not compulsory for small companies to declare their revenue. So the missing values are the small revenues.

It presents a very unlikely distribution with most of the numerical data in the 1M€ and above range for the revenue.

Numerical values: running a function displaying the histogram and boxplot for each column ('ca_1', 'resultat_1', 'ca_2', 'resultat_2', 'ca_3', 'resultat_3')

```
****** Brief Analysis of ca_1 *****
mean= 7675483.0, median= 400628.5, mode= 0.0
var = 3.820382135883424e+17, std_dev = 618092398.91, min = -19132705.0, max = 22592
8206000.0, range = 225947338705.0
quantiles :
0.25       97440.75
0.50      400628.50
0.75     1970318.25
Name: ca_1, dtype: float64
```



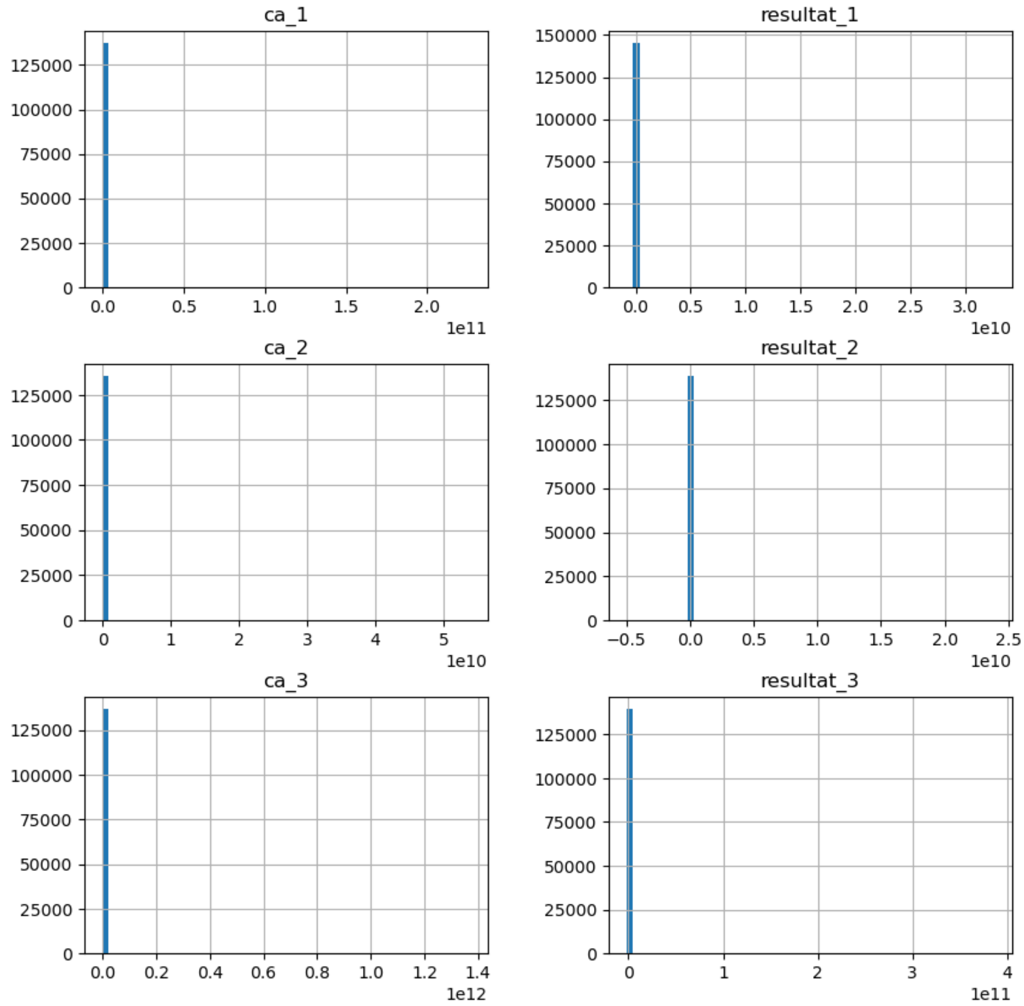


It is more complex as the data is extremely right skewed and the visualization needs a lot of cleaning of the 'right hand side' outliers. The distribution is a spike on the left hand-side, the boxplot is the squeezed line close to 0.0.

We do not want to remove the outliers as they bring important information, but we need a different approach to look at the data.

As mentioned, all the numerical values (see plots following) present a similar distribution close to the zero value, the values are not at zero, but the high outliers impact the scale that is too high.

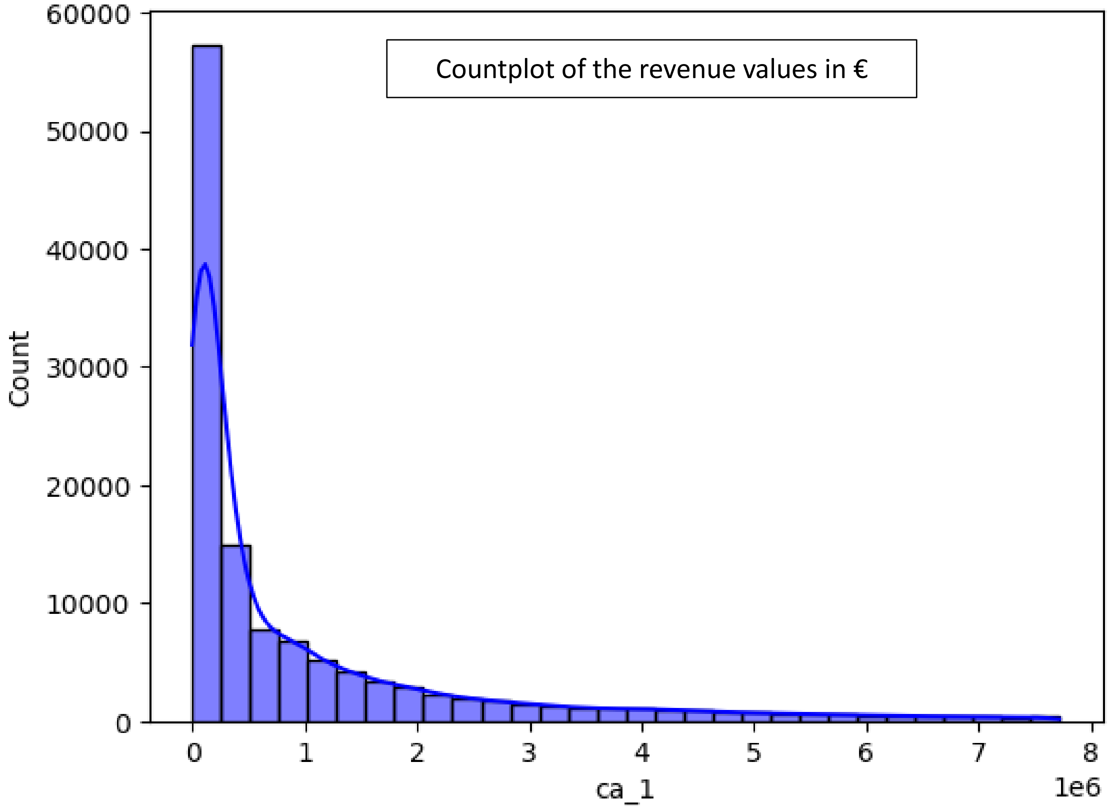
Distributions of the numerical data, infogreffe_numbers Table

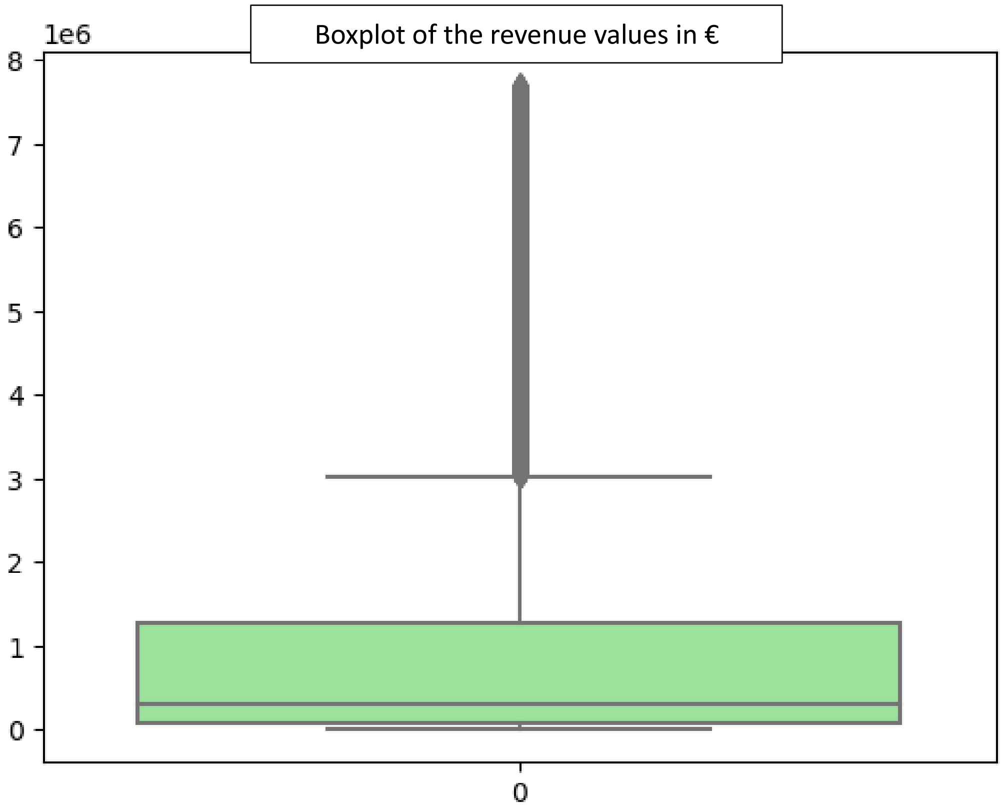


Numerical values second EDA, in order to visualize (and confirm our hunch) we have:

removed the < 0 values for the revenue or income if any, and used qcut for deciles. Then we have removed the top 10% and reached a better level of visualization.

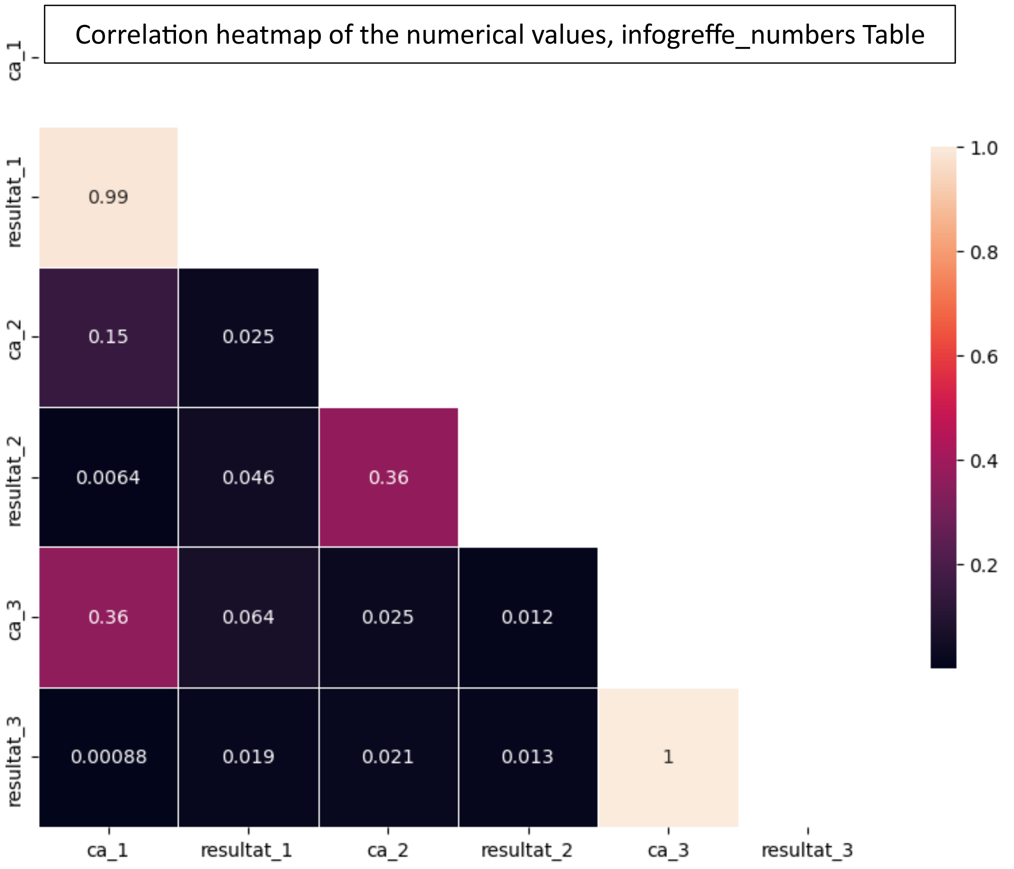
```
****** Brief Analysis of ca_1 *****
mean= 1009669.0, median= 303435.5, mode= 120000.0
var = 2286068721084.31, std_dev = 1511975.11, min = 1.0, max = 7713997.0, range = 7
713996.0
quantiles :
0.25      87567.75
0.50     303435.50
0.75    1263285.00
Name: ca_1, dtype: float64
```





And this confirms the hunch it was right skewed, with the outliers ‘squeezing’ the data. Even filtering out the top 10% of the values there still are numerous outliers

Last and not least, through the heatmap, we can confirm there are some correlations between revenue and income.


Correlation heatmap of the numerical values, infogreffe_numbers Table

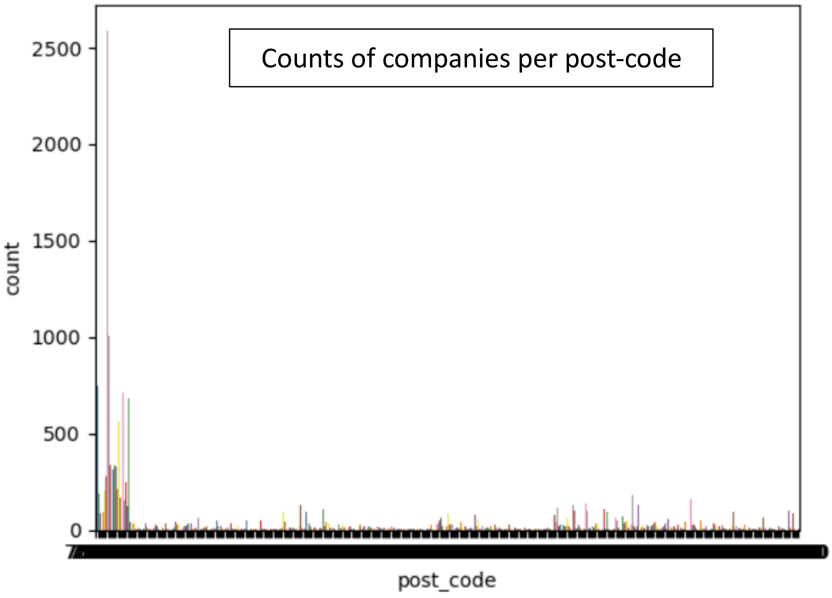
7.2 Looking at the Ile de France Data

Categorical values: running a function for some selected columns that have less unique values than the others. The address, denominations and the like are not relevant to check in terms of distribution or statistics.

Looking at post_code, city: there is Paris and the rest.

```
****** Brief analysis of post_code*******

post_code
75008    2592
75009    1005
75002     745
75017     708
75116     684
         ...
78780       1
77880       1
77138       1
78330       1
94480       1
```

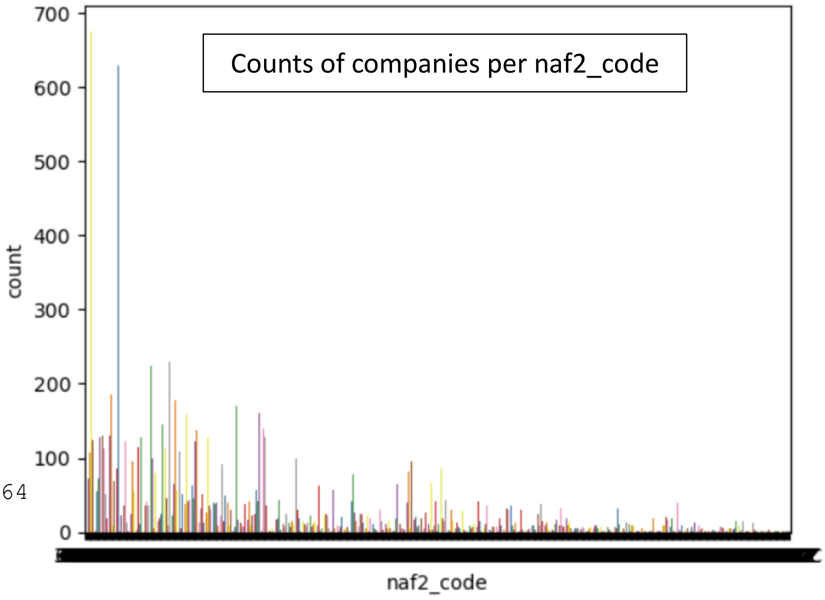

Counts of companies per post-code

```
Name: count, Length: 424, dtype: int64
post_code
75008    0.153083
75009    0.059355
75002    0.044000
75017    0.041814
75116    0.040397
           ...
78780    0.000059
77880    0.000059
77138    0.000059
78330    0.000059
94480    0.000059
Name: proportion, Length: 424, dtype: float64
'mode = 0    75008\nName: post_code, dtype: int32'
```

Naf2-code

A bit less unbalanced, the most represented are less than 4% of the total categories

```
****** Brief analysis of naf2_code *****
naf2_code
6820B    676
7022Z    630
6202A    416
7112B    380
4669B    338
        ...
1041A      1
3522Z      1
2432Z      1
2365Z      1
8690C      1
Name: count, Length: 598, dtype: int64
naf2_code
6820B    0.039924
7022Z    0.037208
6202A    0.024569
7112B    0.022443
4669B    0.019962
           ...
1041A    0.000059
3522Z    0.000059
2432Z    0.000059
2365Z    0.000059
8690C    0.000059
```



Counts of companies per naf2_code

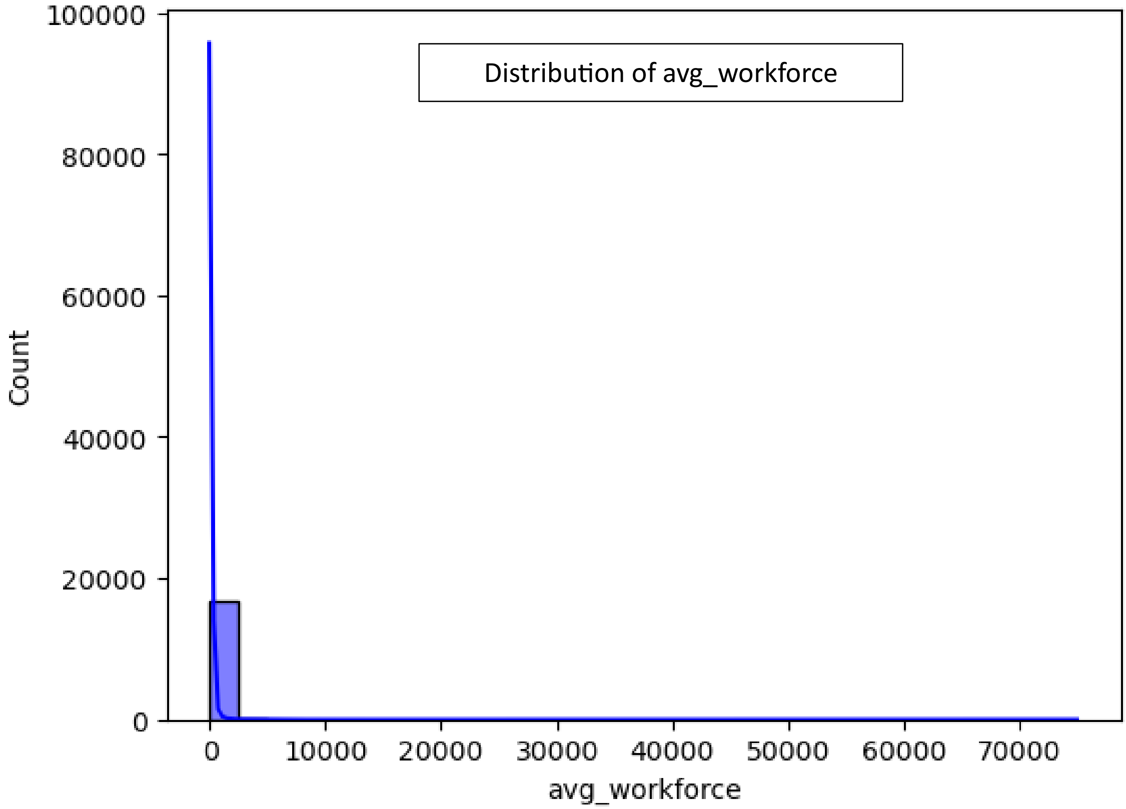
Numerical values: running a function displaying the histogram and boxplot for each column

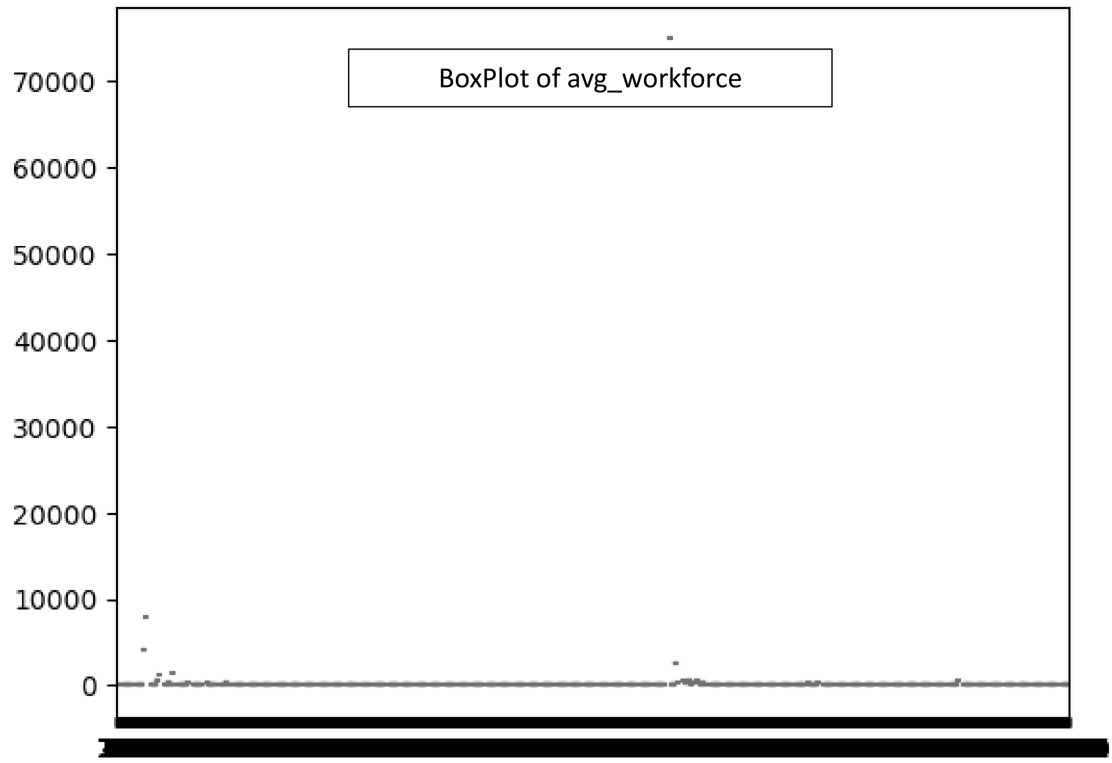
It is more complex as the data is extremely right skewed and the visualization needs a lot of cleaning of the ‘right hand side’ outliers. Also the -1 values should be removed from the analysis

- avg_workforce: at least 25% are -1 (no data), to analyse without -1, so we get a better value.

  It is rigth skewed, with high outliers affecting the mean (mean > median)

```
****** Brief Analysis of avg_workforce *****
mean= 100.0, median= 6.0, mode= -1.0
var = 1247958.15, std_dev = 1117.12, min = -1.0, max = 74900.0, range = 749
01.0
quantiles :
0.25    -1.0
0.50     6.0
0.75    44.0
Name: avg_workforce, dtype: float64
```



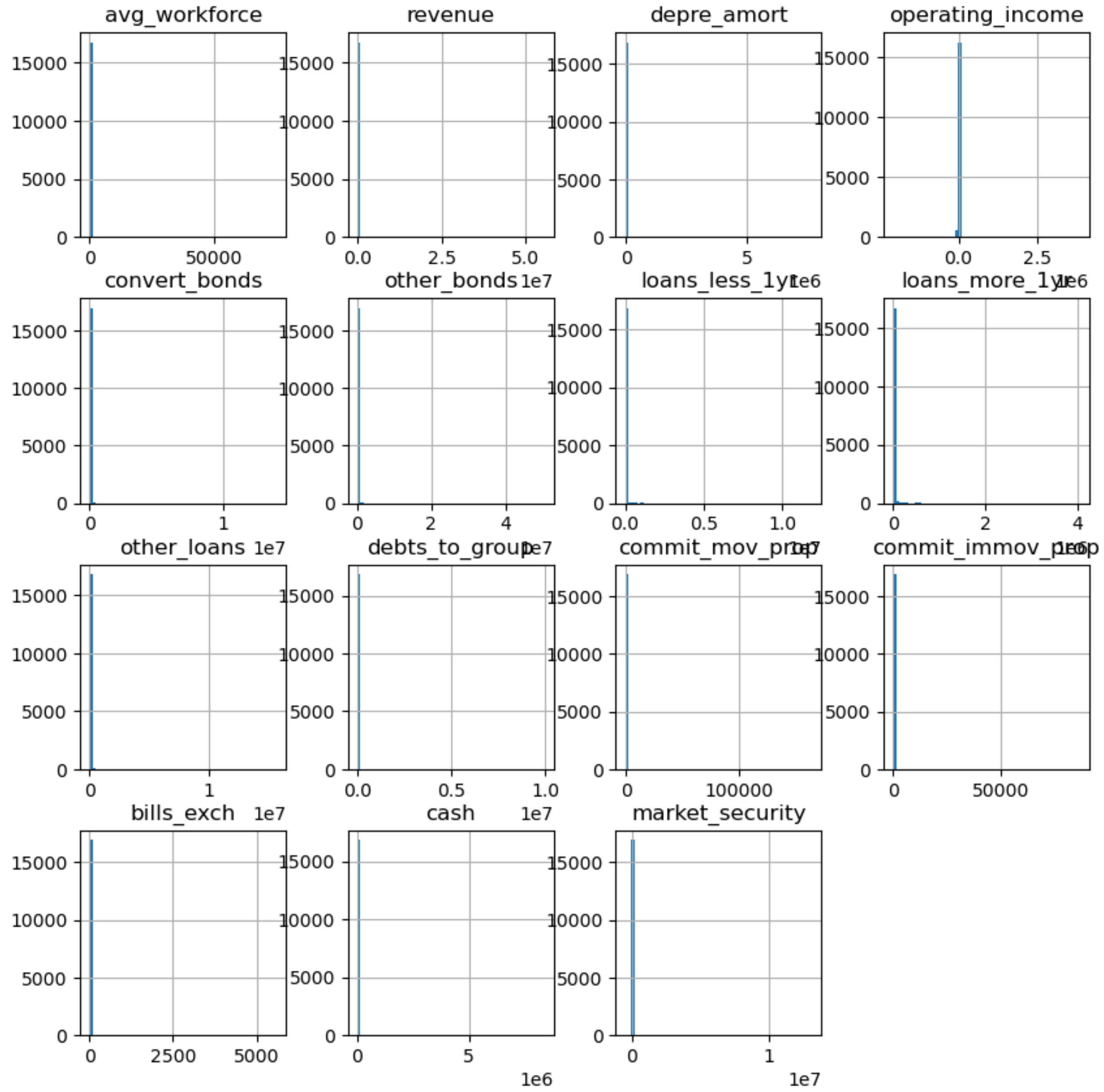


The boxplots do not show anything (too flat with outliers) without previous preparation of the data.

- Revenue and others: min at 2800K€ (2.8 million) max at 55,500,000K€ (55.5 billion), median at 9.8 million, mean at 6.5 million, 50% of the companies are between 5.5M€ and 23M€

All other numerical present a very similar aspect, with an extreme effect of the right outliers affecting the display scale and compressing the data on the right (or in the center for the operating income

Distributions of the numerical data, numbers_all Table



We do not want to remove the outliers as they bring important information, but we need a different approach to look at the data.

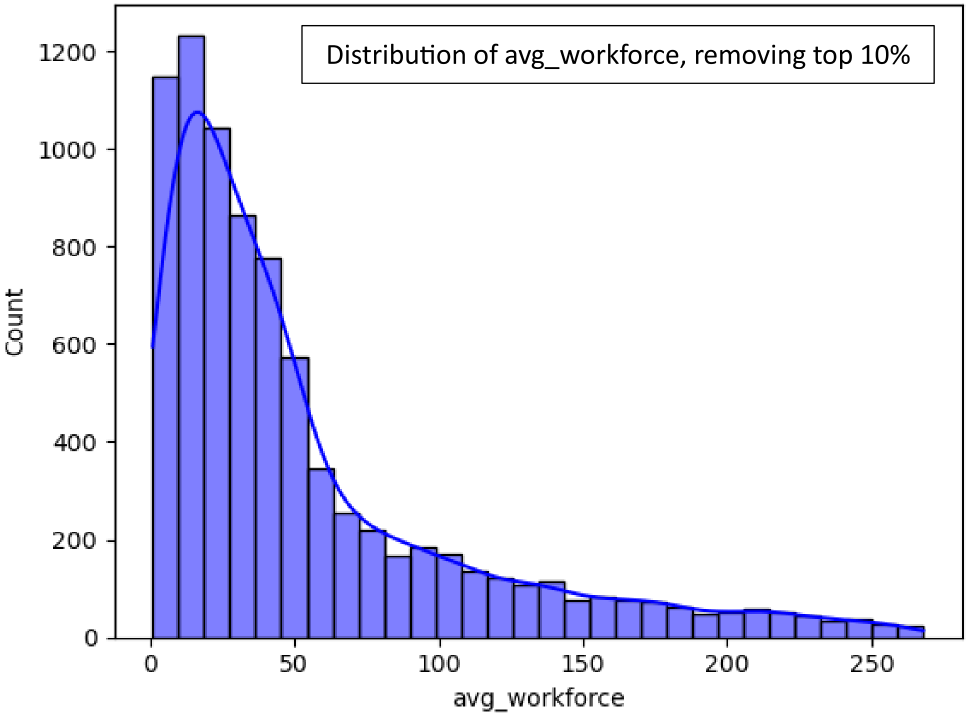
Numerical values second EDA, in order to visualize (and confirm our hunch) we have:

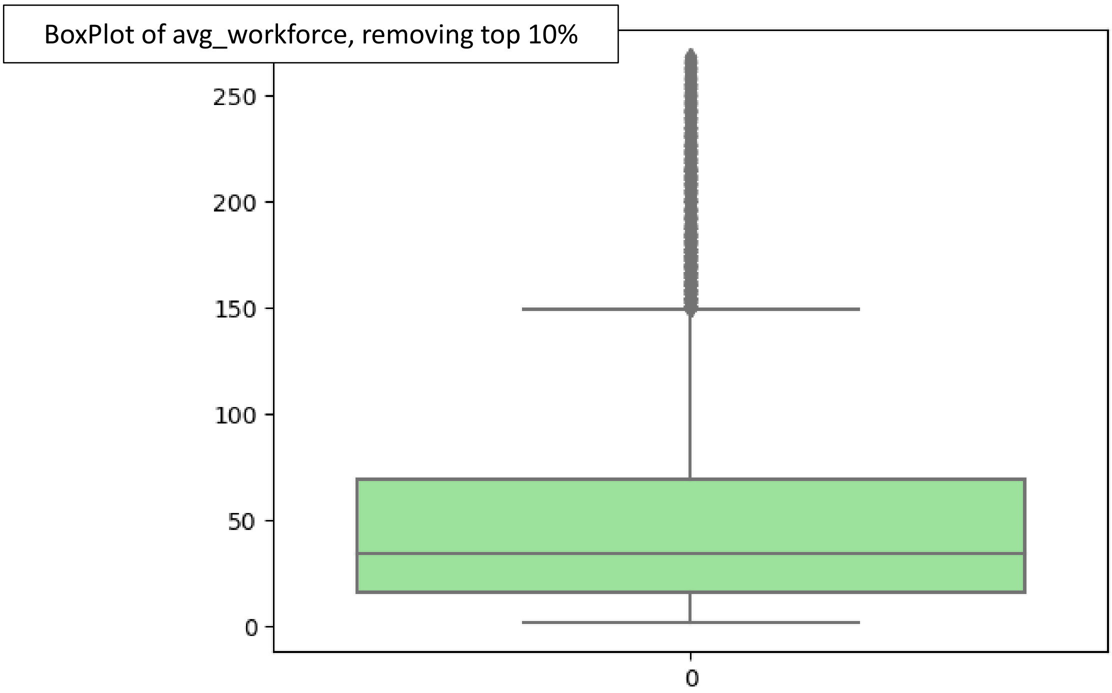
removed the -1 values for avg_workforce or the < 0 values for the revenue (the other numbers could be negative) and used qcut for deciles. Then we have removed the top 10% and reached a better level of visualization.

- avg_workforce:

```
****** Brief Analysis of avg_workforce *****
mean= 54.0, median= 34.0, mode= 7.0
var = 3063.8, std_dev = 55.35, min = 1.0, max = 268.0, range = 267.0
quantiles :
0.25    16.00
0.50    34.00
0.75    69.25
```

```
Name: avg_workforce, dtype: float64
```


Distribution of avg_workforce, removing top 10%

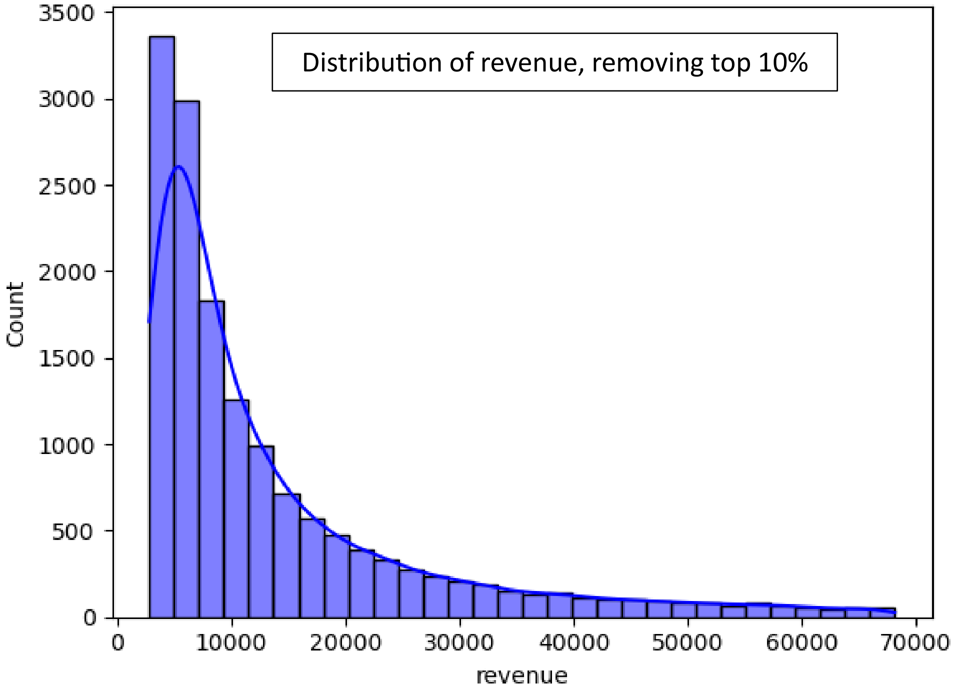

BoxPlot of avg_workforce, removing top 10%

If we remove the top 10% (>286 employees), the median is at 34 and the average at 54 employees, with still a sizeable number of outliers above 150 employees. The distribution is right skewed.

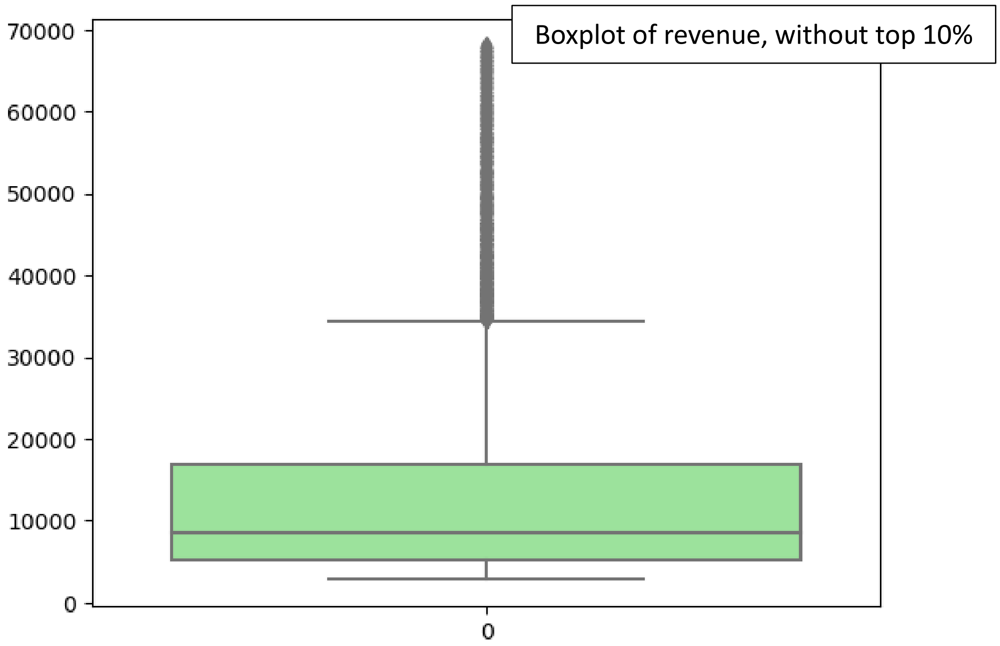
- revenue: with a similar approach we can visualize and also confirm the right skew

```
****** Brief Analysis of revenue *****
mean= 13701.0, median= 8546.2235, mode= 4072.297
```

```
var = 169951108.78, std_dev = 13036.53, min = 2800.036, max = 68102.193, range = 65
302.157
quantiles :
0.25      5224.1950
0.50      8546.2235
0.75     16919.4180
Name: revenue, dtype: float64
```



Distribution of revenue, removing top 10%

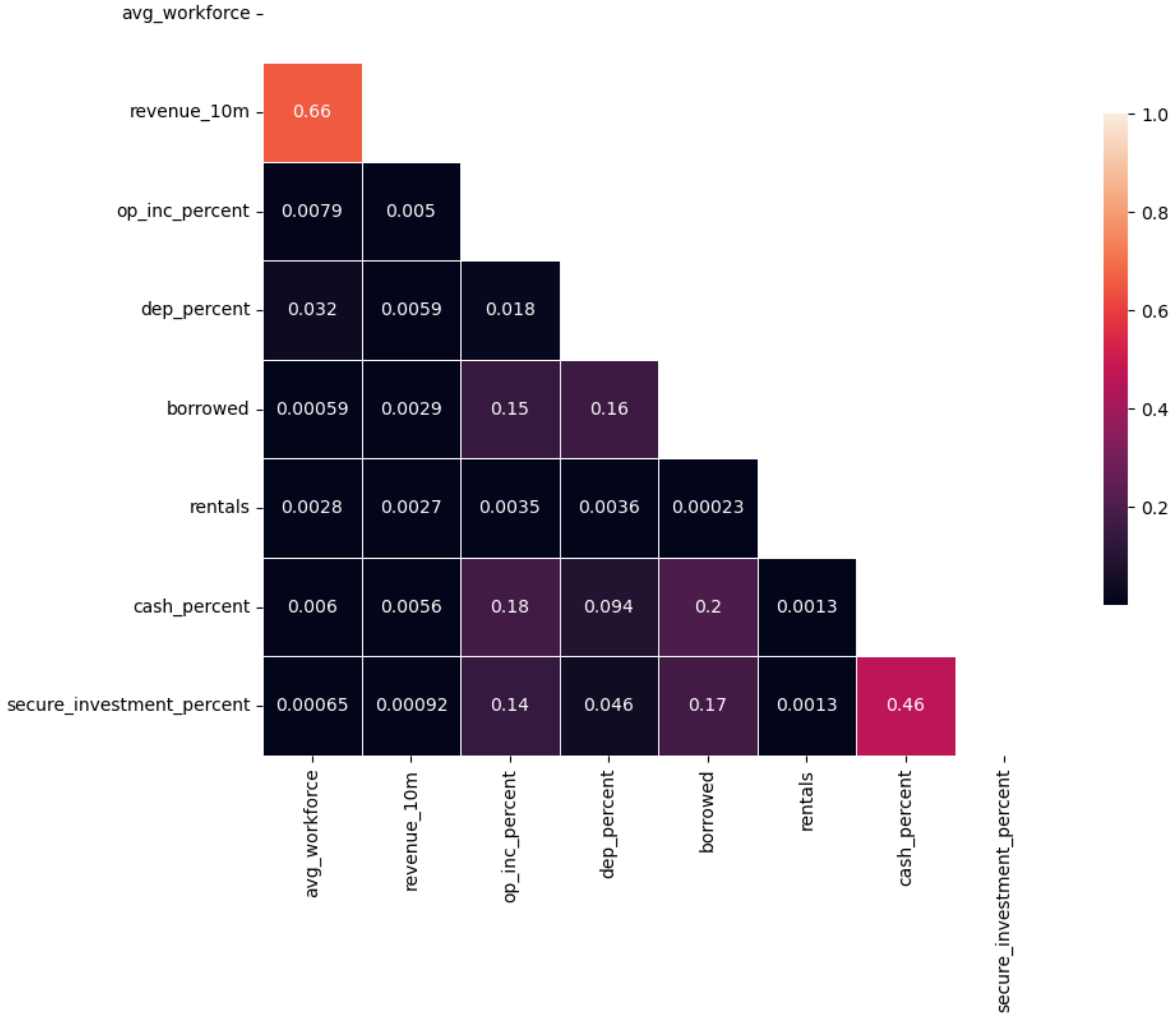


Boxplot of revenue, without top 10%

Similar plots and analysis was done for the other numerical columns

Correlation heatmap: in order to see if there are relationships between the numerical data (for further use), a map was run on the data removing the -1 values for the avg_workforce:

Correlation heatmap of the numerical values, numbers_light Table



It does not come as a surprise that there might be a correlation between the total revenue and number of employees of a company.

Similarly it is likely that when there is more cash there are more external investments…

7.3 1st global visualisations

For codes 62.02A and 70.22Z, comparisons for revenue and income for 2020-21-22:

- code 62.02A vs all the codes in France,
- code 62.02A vs all the codes in Paris,

Keep in mind that in general growth is expected between 2020 and 2021 as 2020 has a great impact of the Covid-19 pandemia.

a/ For the code APE 62.02A (Conseil en systèmes et logiciels informatiques) :

From the line plots presented on the following page, we can see that:

Revenue

- the *code 62.02A* median revenue is lower than *all the codes*, at the scale of France, but higher than *all the codes* at the scale of Paris. The companies in Paris do not follow the national level.

- the *code 62.02A* median revenue grows from 2020 to 2022 but not as fast as for *all the other codes*.

Income :

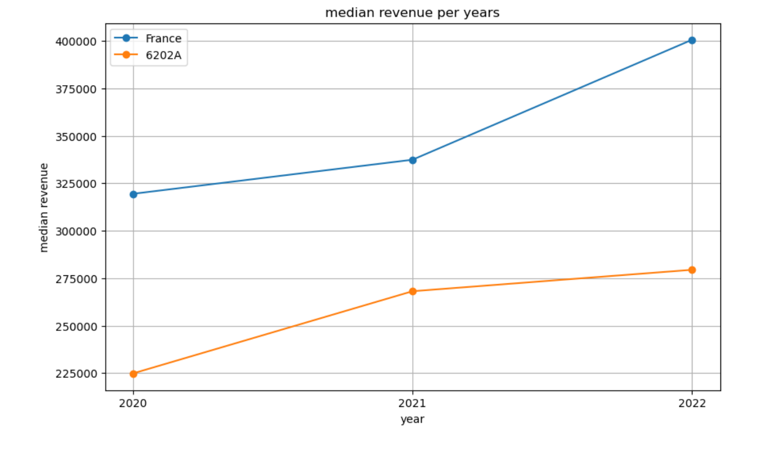
Median income of the code 62.02A does not follow the same trends as the Median revenue .

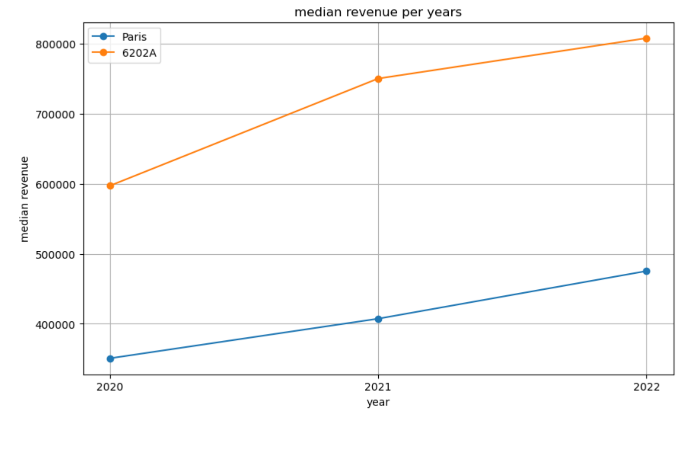
- the *code 62.02A* median income is lower at the scale of France, but it is higher at the scale of Paris.

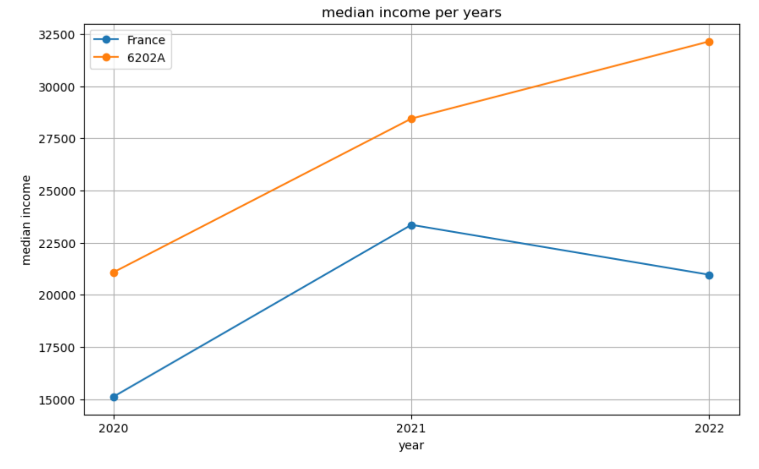
- the *code 62.02A* median income grows from 2020 to 2022 at the French level, while the median income of *all the codes* spikes in 2021 and decreases in France

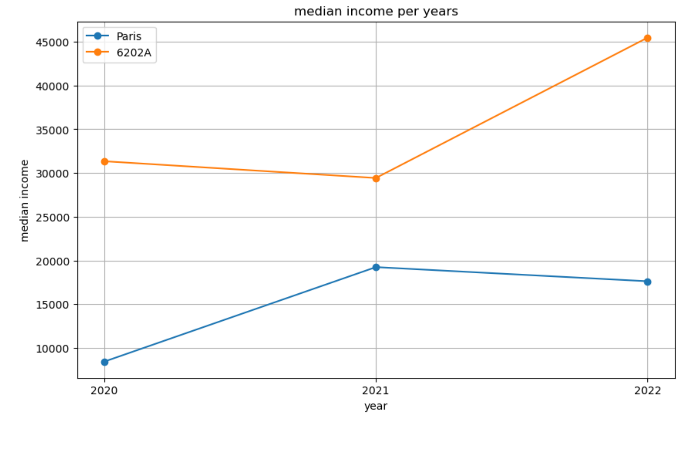
- in Paris the *code 62.02A* median income is stable in 2020-21 and grows in 2022, when all the median income of *all the codes* grow from 2020-21 and stabilize to decline in 2021

The values in Paris are higher than compared to the whole France values.





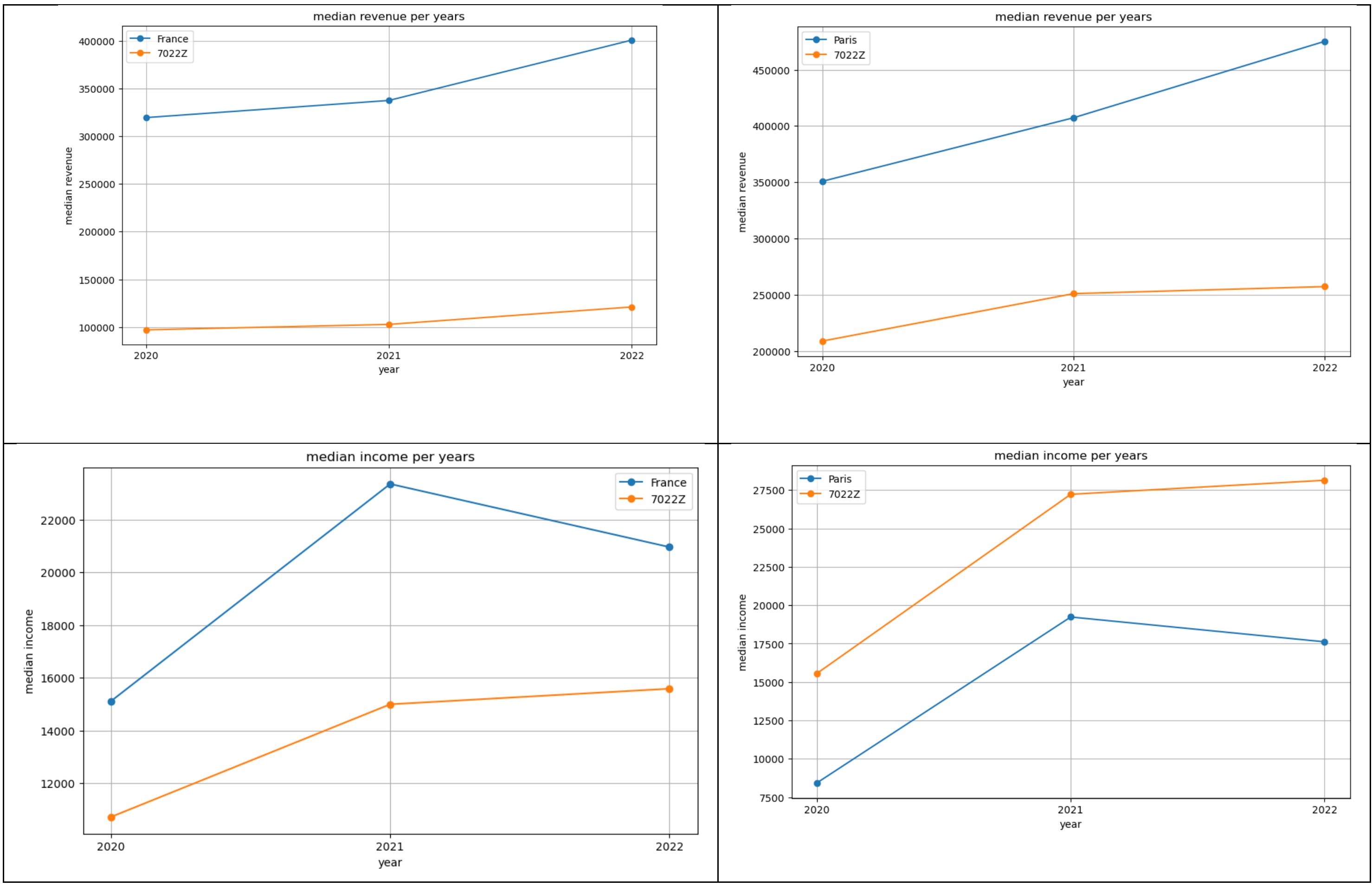




In other words, in 2022 the companies in Paris with the *code 62.02A have* a higher median revenue and median income than the French global numbers (all codes together). The median value is about twice as much as the France value (of the companies that have declared their numbers).

Code 62.02A shows growing trends, solid results in Paris in particular but overall above the French global numbers.

- code 70.22Z vs all the codes in France,
- code 70.22Z vs all the codes in Paris



For the code APE 70.22Z (Conseil pour les affaires et autres conseils de gestion) compared to all the other codes:

- the median revenue is lower at the scale of France, as well as at the scale of Paris

- the median revenue grows from 2020 to 2022 but not as fast as for all the other codes.

- the median income is lower at the scale of France, but it is higher at the scale of Paris

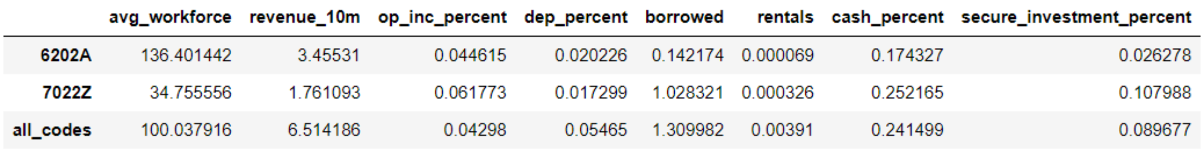
- the median income grows from 2020 to 2022 stabilising a bit in 2021-22, but it spikes in 2021 and decreases for all the other codes in France or in Paris.

The values in Paris are higher than compared to the whole France values.

It seems, in 2022, that for *code 7022Z* it is more interesting to work in Paris where the median income is higher than for France (even if it is not the case for the revenue), and still increasing.

The code is not negatively affected by a drop of median income like global France.

Finally, we are looking at the averages for the two codes for the numbers_light data:



	avg_workforce	revenue_10m	op_inc_percent	dep_percent	borrowed	rentals	cash_percent	secure_investment_percent
6202A	136.401442	3.45531	0.044615	0.020226	0.142174	0.000069	0.174327	0.026278
7022Z	34.755556	1.761093	0.061773	0.017299	1.028321	0.000326	0.252165	0.107988
all_codes	100.037916	6.514186	0.04298	0.05465	1.309982	0.00391	0.241499	0.089677

Snapshot of the dataframe generated to compare the 2 activities.

They are more cryptical and difficult to read. It shows however different behaviours. On average, in Ile de France:

The companies whose activity is *7022Z* have smaller workforce (35) than the average of all activities (100), while the ones whose activity is *6202A* are bigger(136) than the average of all activities.

Regarding average revenue, activity *7022Z* (17.6M€) is lower than *6202A* (34.5M€) but both are lower than the all_codes (65.1M€). They are smaller companies than the average.

Operating income rates are better for 7022Z and comparable for the others.

Depreciation values is more important for the all_codes (more equipment?) than for our two reference activities.

The code 6202A presents very little debt (borrowed) compared to the others.

Rentals are negligible.

Cash is comparable for 7022Z and all_codes, and lower for 6202A.

7022Z and all_codes have more investments than 6202A

8 Database Creation

8.1 Choice of database

We have structured tables that can be connected via siren or code_ape principally (with the notable exception of the BnF data that has to go through a made junction table).

SQL type of database is suitable for this project (using MySQL via MySQL Workbench). Additionally it is well adapted to accept addition of data from similar or even different sources and could evolve easily.

Typically the main advantages and disadvantages of each database type are presented below:

SQL databases

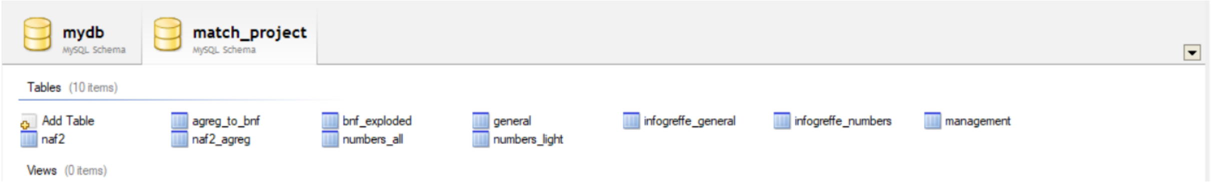
- are relational
- use structured query language and have a predefined schema
- are vertically scalable,
- are table-based,
- are better for multi-row transactions

NoSQL databases

- are non-relational,

- have dynamic schemas for unstructured data
- are horizontally scalable
- are document, key-value, graph, or wide-column stores
- are better for unstructured data like documents or JSON

We have created a MySQL database using python to export the cleaned and wrangled dataframes.



agreg_to_bnf, bnf_exploded, general, infogreffe_general, infogreffe_numbers, management, naf2, naf2_agreg, numbers_all, numbers_light

8.2 SQL queries

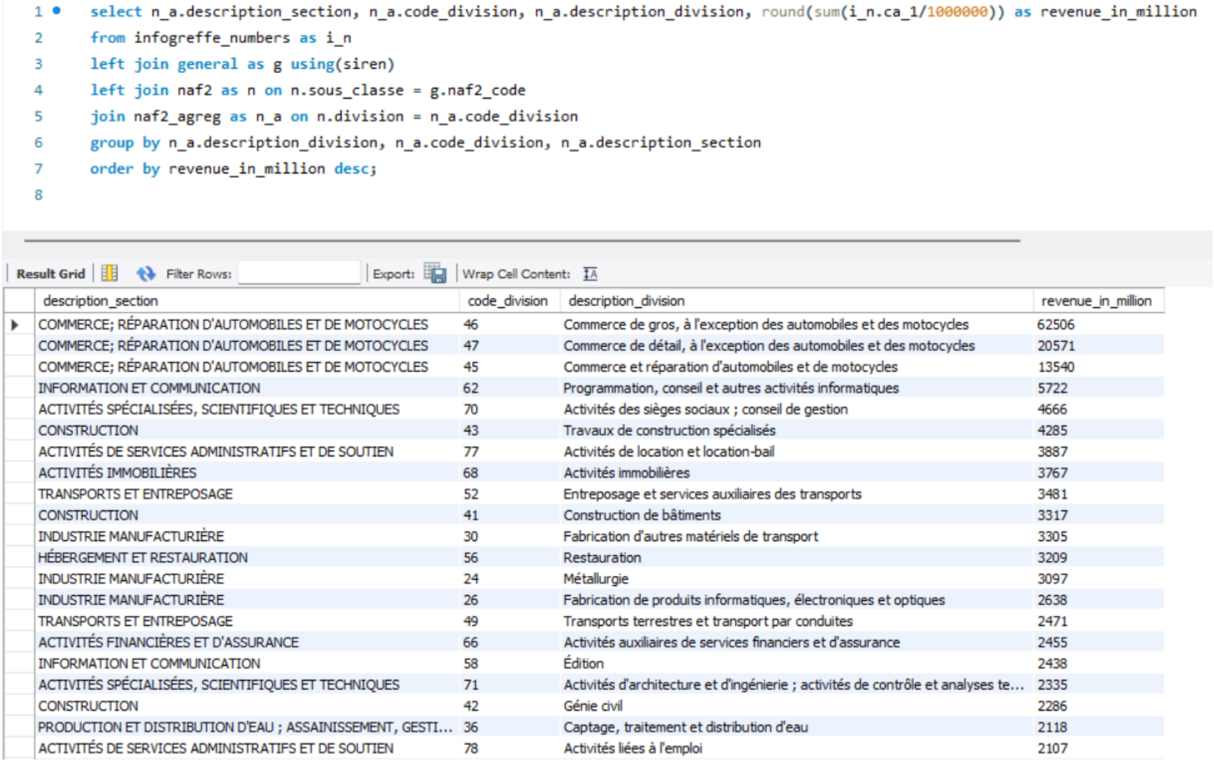
Please find below some examples of the queries run:

a) **Query 1:** The total of **declared revenue** of the divisions, in million, ordered from largest to smallest, in 2022.

Keeping in mind that declaring the revenue is not compulsory for small companies, it is not the total French revenue.

We can see that the top 3 are divisions 46, 47, 45 (linked with vehicule maintenance and repairs).

Next comes division 62, linked with IT programming and division 70, linked with head-quarters and management. These are the divisions of the activities we are looking at (6202A and 7022Z).



```sql
select n_a.description_section, n_a.code_division, n_a.description_division, round(sum(i_n.ca_1/1000000)) as revenue_in_million
from infogreffe_numbers as i_n
left join general as g using(siren)
left join naf2 as n on n.sous_classe = g.naf2_code
join naf2_agreg as n_a on n.division = n_a.code_division
group by n_a.description_division, n_a.code_division, n_a.description_section
order by revenue_in_million desc;
```

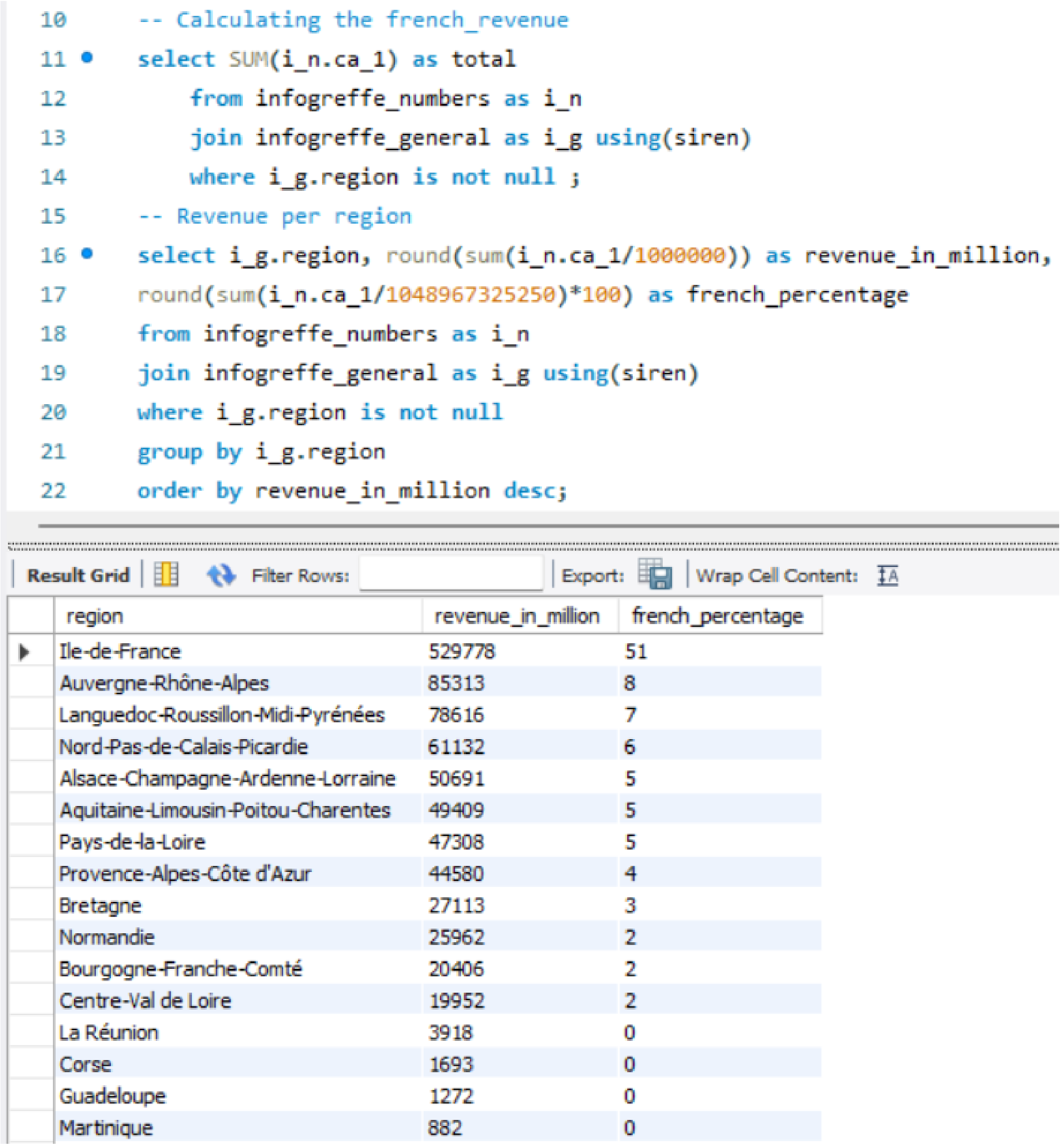
description_section	code_division	description_division	revenue_in_million
COMMERCE; RÉPARATION D'AUTOMOBILES ET DE MOTOCYCLES	46	Commerce de gros, à l'exception des automobiles et des motocycles	62506
COMMERCE; RÉPARATION D'AUTOMOBILES ET DE MOTOCYCLES	47	Commerce de détail, à l'exception des automobiles et des motocycles	20571
COMMERCE; RÉPARATION D'AUTOMOBILES ET DE MOTOCYCLES	45	Commerce et réparation d'automobiles et de motocycles	13540
INFORMATION ET COMMUNICATION	62	Programmation, conseil et autres activités informatiques	5722
ACTIVITÉS SPÉCIALISÉES, SCIENTIFIQUES ET TECHNIQUES	70	Activités des sièges sociaux ; conseil de gestion	4666
CONSTRUCTION	43	Travaux de construction spécialisés	4285
ACTIVITÉS DE SERVICES ADMINISTRATIFS ET DE SOUTIEN	77	Activités de location et location-bail	3887
ACTIVITÉS IMMOBILIÈRES	68	Activités immobilières	3767
TRANSPORTS ET ENTREPOSAGE	52	Entreposage et services auxiliaires des transports	3481
CONSTRUCTION	41	Construction de bâtiments	3317
INDUSTRIE MANUFACTURIÈRE	30	Fabrication d'autres matériels de transport	3305
HÉBERGEMENT ET RESTAURATION	56	Restauration	3209
INDUSTRIE MANUFACTURIÈRE	24	Métallurgie	3097
INDUSTRIE MANUFACTURIÈRE	26	Fabrication de produits informatiques, électroniques et optiques	2638
TRANSPORTS ET ENTREPOSAGE	49	Transports terrestres et transport par conduites	2471
ACTIVITÉS FINANCIÈRES ET D'ASSURANCE	66	Activités auxiliaires de services financiers et d'assurance	2455
INFORMATION ET COMMUNICATION	58	Édition	2438
ACTIVITÉS SPÉCIALISÉES, SCIENTIFIQUES ET TECHNIQUES	71	Activités d'architecture et d'ingénierie ; activités de contrôle et analyses te...	2335
CONSTRUCTION	42	Génie civil	2286
PRODUCTION ET DISTRIBUTION D'EAU ; ASSAINISSEMENT, GESTI...	36	Captage, traitement et distribution d'eau	2118
ACTIVITÉS DE SERVICES ADMINISTRATIFS ET DE SOUTIEN	78	Activités liées à l'emploi	2107

Fig.: Snapshot of the query and principal results.

b) **Query 2 and 3:** The total of **declared revenue** of the regions, in million, ordered from largest to smallest, in 2022. Calculating first the French Total Revenue

Keeping in mind that declaring the revenue is not compulsory for small companies, it is not the total French revenue.

It is no surprise to see that the top 3 regions are: Ile de France (51%!), Auvergne_Rhône-Alpes, Languedoc-Roussillon-Midi-Pyrénées.



```
-- Calculating the french_revenue
select SUM(i_n.ca_1) as total
    from infogreffe_numbers as i_n
    join infogreffe_general as i_g using(siren)
    where i_g.region is not null ;
-- Revenue per region
select i_g.region, round(sum(i_n.ca_1/1000000)) as revenue_in_million,
round(sum(i_n.ca_1/1048967325250)*100) as french_percentage
from infogreffe_numbers as i_n
join infogreffe_general as i_g using(siren)
where i_g.region is not null
group by i_g.region
order by revenue_in_million desc;
```

region	revenue_in_million	french_percentage
Ile-de-France	529778	51
Auvergne-Rhône-Alpes	85313	8
Languedoc-Roussillon-Midi-Pyrénées	78616	7
Nord-Pas-de-Calais-Picardie	61132	6
Alsace-Champagne-Ardenne-Lorraine	50691	5
Aquitaine-Limousin-Poitou-Charentes	49409	5
Pays-de-la-Loire	47308	5
Provence-Alpes-Côte d'Azur	44580	4
Bretagne	27113	3
Normandie	25962	2
Bourgogne-Franche-Comté	20406	2
Centre-Val de Loire	19952	2
La Réunion	3918	0
Corse	1693	0
Guadeloupe	1272	0
Martinique	882	0

Fig.: Snapshot of the query and principal results.

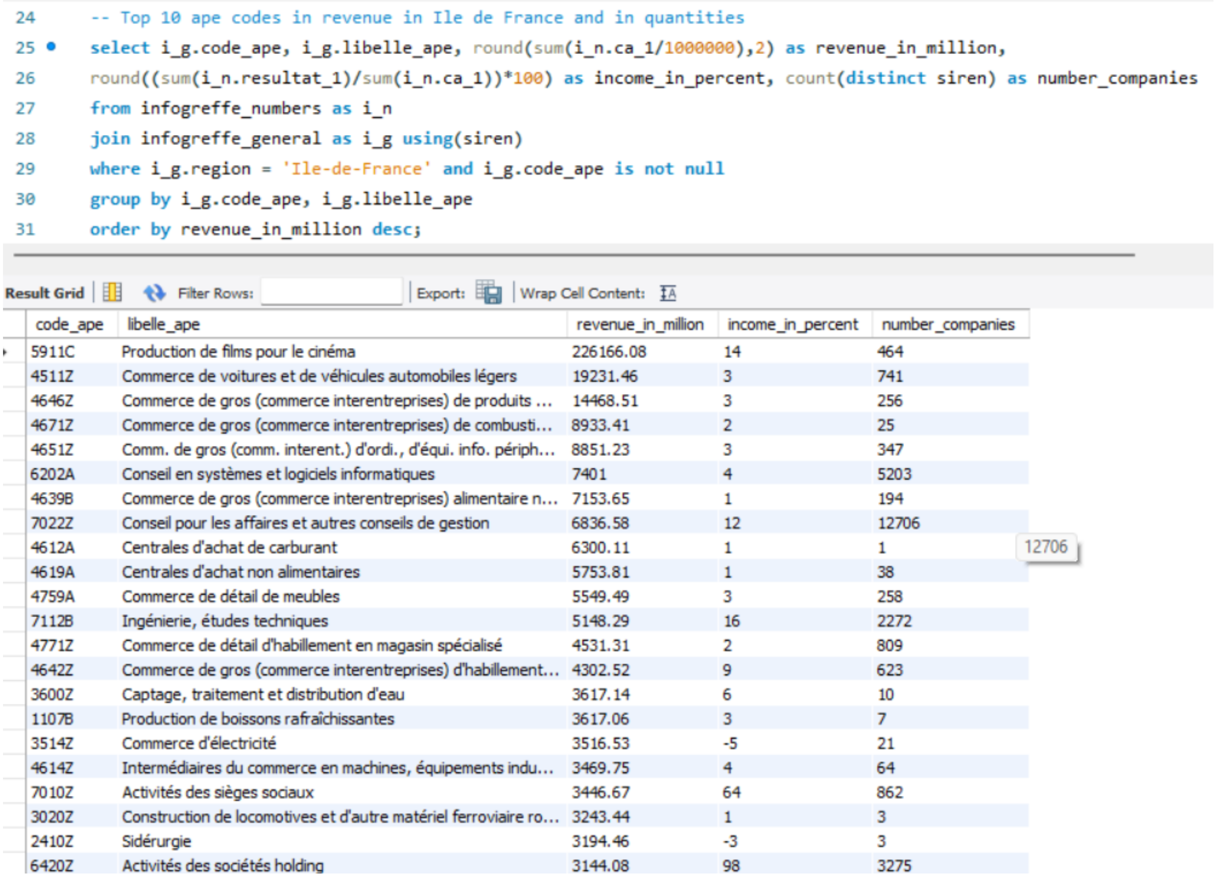
c) **Query 4:** The **declared revenue** of the code_ape in the region Ile-de-France, ordered from largest to smallest, with income percentage and the number of companies in each code, in 2022

Keeping in mind that declaring the revenue is not compulsory for small companies, it is not the total French revenue.

Code 59.11C, Cinema production, comes first in revenue, with a good income at 14% for only 464 companies in Ile de France.

We find again the 45 and 46 sections (46.46Z, 45.11Z, 46.71Z, 46.51Z, 46.39B), but with much more companies and a lower incomes (in the range of 2-3%).

IT is still present with 62.02A, many companies (5203, and an income in the range of 4%. Interesting to note, 7022Z generates a healthy income (12%) for many more companies (12706!).



```
-- Top 10 ape codes in revenue in Ile de France and in quantities
select i_g.code_ape, i_g.libelle_ape, round(sum(i_n.ca_1/1000000),2) as revenue_in_million,
round((sum(i_n.resultat_1)/sum(i_n.ca_1))*100) as income_in_percent, count(distinct siren) as number_companies
from infogreffe_numbers as i_n
join infogreffe_general as i_g using(siren)
where i_g.region = 'Ile-de-France' and i_g.code_ape is not null
group by i_g.code_ape, i_g.libelle_ape
order by revenue_in_million desc;
```

code_ape	libelle_ape	revenue_in_million	income_in_percent	number_companies
5911C	Production de films pour le cinéma	226166.08	14	464
4511Z	Commerce de voitures et de véhicules automobiles légers	19231.46	3	741
4646Z	Commerce de gros (commerce interentreprises) de produits ...	14468.51	3	256
4671Z	Commerce de gros (commerce interentreprises) de combusti...	8933.41	2	25
4651Z	Comm. de gros (comm. interent.) d'ordi., d'équi. info. périph...	8851.23	3	347
6202A	Conseil en systèmes et logiciels informatiques	7401	4	5203
4639B	Commerce de gros (commerce interentreprises) alimentaire n...	7153.65	1	194
7022Z	Conseil pour les affaires et autres conseils de gestion	6836.58	12	12706
4612A	Centrales d'achat de carburant	6300.11	1	1
4619A	Centrales d'achat non alimentaires	5753.81	1	38
4759A	Commerce de détail de meubles	5549.49	3	258
7112B	Ingénierie, études techniques	5148.29	16	2272
4771Z	Commerce de détail d'habillement en magasin spécialisé	4531.31	2	809
4642Z	Commerce de gros (commerce interentreprises) d'habillement...	4302.52	9	623
3600Z	Captage, traitement et distribution d'eau	3617.14	6	10
1107B	Production de boissons rafraîchissantes	3617.06	3	7
3514Z	Commerce d'électricité	3516.53	-5	21
4614Z	Intermédiaires du commerce en machines, équipements indu...	3469.75	4	64
7010Z	Activités des sièges sociaux	3446.67	64	862
3020Z	Construction de locomotives et d'autre matériel ferroviaire ro...	3243.44	1	3
2410Z	Sidérurgie	3194.46	-3	3
6420Z	Activités des sociétés holding	3144.08	98	3275

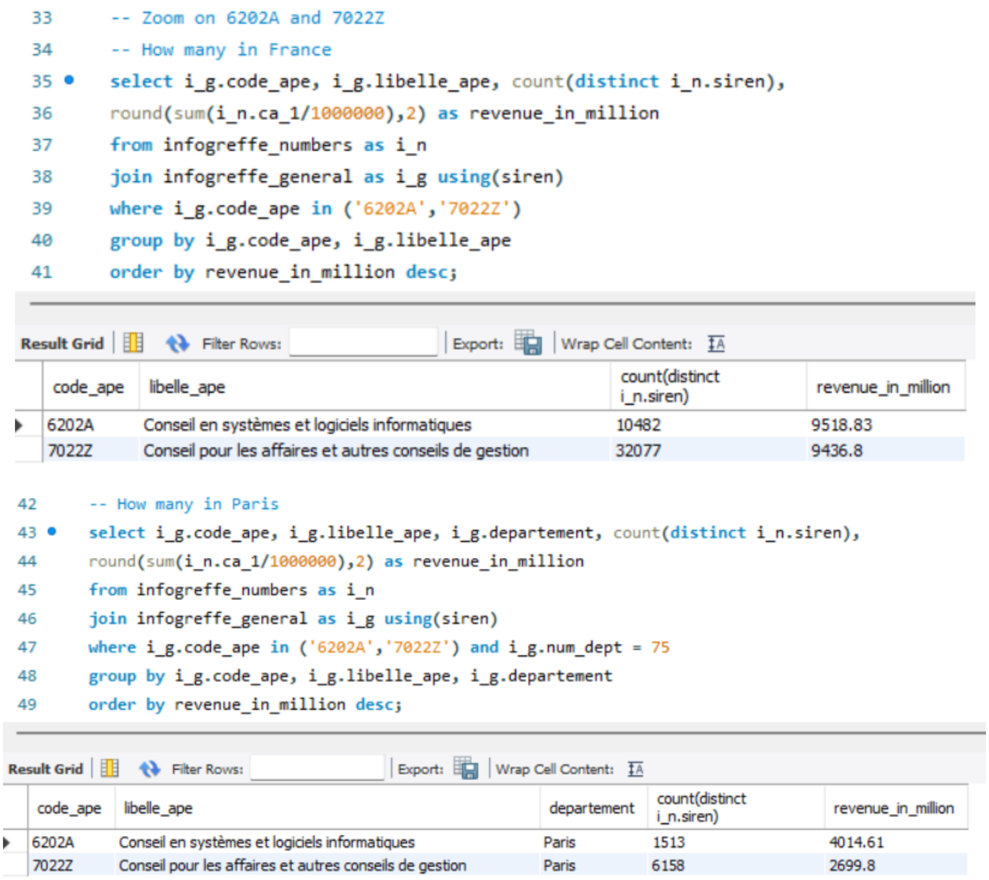
Fig.: Snapshot of the query and principal results.

d) **Query 5:** Zoom on code ape 62.02A and 70.22Z, in Paris more than 2.8 million € of revenue

Keeping in mind that declaring the revenue is not compulsory for small companies, it is not the total French revenue.

For code 62.02A, ‘Conseil en systèmes et logiciels informatiques’, in 2022 in Paris, there are 10,482 companies more than 2.8M€ yearly revenue. This volume represents 9.5 Billion €. The average ‘size’ (revenue) of such companies is about 906K€/year. However there are ‘elephants’ such as Accenture with several thousands of employees, and only about 15 that have more than 60 as average workforce.

For code 70.22Z, ‘Conseil pour les affaires et autres conseils de gestion’, in 2022 in Paris, there are 32,077 companies more than 2.8M€ yearly revenue. This volume represents 9.4 Billion €. The average ‘size’ (revenue) of such companies is about 293K€/year. The biggest company in terms of employees is just below 330 on average, and only 16 companies had more than 50 on average workforce.



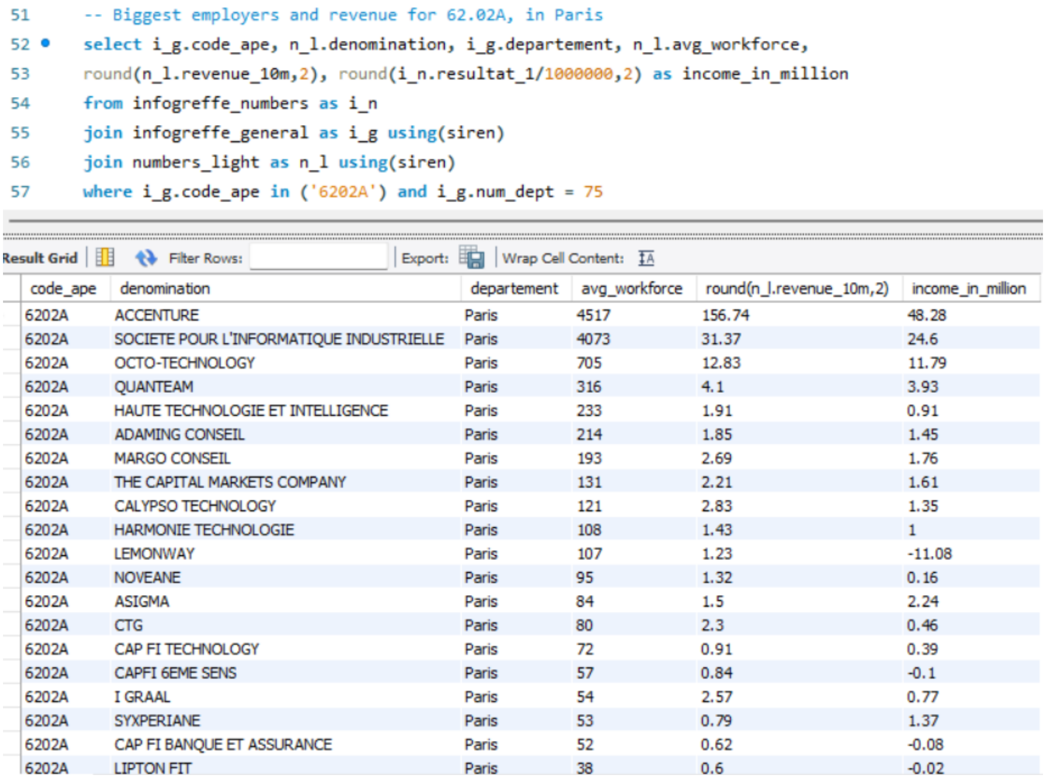
```sql
-- Zoom on 6202A and 7022Z
-- How many in France
select i_g.code_ape, i_g.libelle_ape, count(distinct i_n.siren),
round(sum(i_n.ca_1/1000000),2) as revenue_in_million
from infogreffe_numbers as i_n
join infogreffe_general as i_g using(siren)
where i_g.code_ape in ('6202A','7022Z')
group by i_g.code_ape, i_g.libelle_ape
order by revenue_in_million desc;
```

code_ape	libelle_ape	count(distinct i_n.siren)	revenue_in_million
6202A	Conseil en systèmes et logiciels informatiques	10482	9518.83
7022Z	Conseil pour les affaires et autres conseils de gestion	32077	9436.8

```sql
-- How many in Paris
select i_g.code_ape, i_g.libelle_ape, i_g.departement, count(distinct i_n.siren),
round(sum(i_n.ca_1/1000000),2) as revenue_in_million
from infogreffe_numbers as i_n
join infogreffe_general as i_g using(siren)
where i_g.code_ape in ('6202A','7022Z') and i_g.num_dept = 75
group by i_g.code_ape, i_g.libelle_ape, i_g.departement
order by revenue_in_million desc;
```

code_ape	libelle_ape	departement	count(distinct i_n.siren)	revenue_in_million
6202A	Conseil en systèmes et logiciels informatiques	Paris	1513	4014.61
7022Z	Conseil pour les affaires et autres conseils de gestion	Paris	6158	2699.8

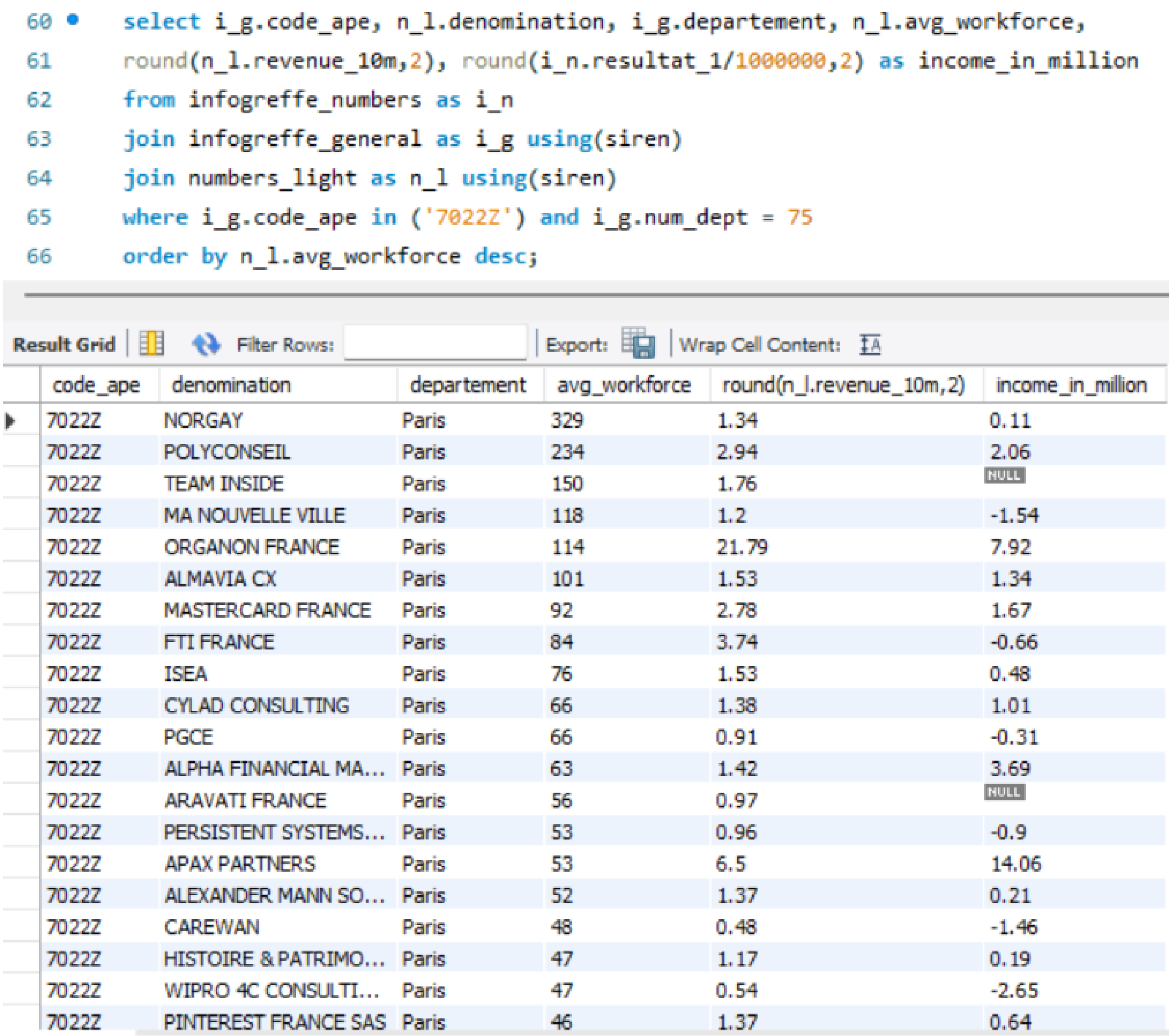
Fig.: Snapshot of the query on count and revenue for 62.02A and 70.22Z



```sql
-- Biggest employers and revenue for 62.02A, in Paris
select i_g.code_ape, n_l.denomination, i_g.departement, n_l.avg_workforce,
round(n_l.revenue_10m,2), round(i_n.resultat_1/1000000,2) as income_in_million
from infogreffe_numbers as i_n
join infogreffe_general as i_g using(siren)
join numbers_light as n_l using(siren)
where i_g.code_ape in ('6202A') and i_g.num_dept = 75
```

code_ape	denomination	departement	avg_workforce	round(n_l.revenue_10m,2)	income_in_million
6202A	ACCENTURE	Paris	4517	156.74	48.28
6202A	SOCIETE POUR L'INFORMATIQUE INDUSTRIELLE	Paris	4073	31.37	24.6
6202A	OCTO-TECHNOLOGY	Paris	705	12.83	11.79
6202A	QUANTEAM	Paris	316	4.1	3.93
6202A	HAUTE TECHNOLOGIE ET INTELLIGENCE	Paris	233	1.91	0.91
6202A	ADAMING CONSEIL	Paris	214	1.85	1.45
6202A	MARGO CONSEIL	Paris	193	2.69	1.76
6202A	THE CAPITAL MARKETS COMPANY	Paris	131	2.21	1.61
6202A	CALYPSO TECHNOLOGY	Paris	121	2.83	1.35
6202A	HARMONIE TECHNOLOGIE	Paris	108	1.43	1
6202A	LEMONWAY	Paris	107	1.23	-11.08
6202A	NOVEANE	Paris	95	1.32	0.16
6202A	ASIGMA	Paris	84	1.5	2.24
6202A	CTG	Paris	80	2.3	0.46
6202A	CAP FI TECHNOLOGY	Paris	72	0.91	0.39
6202A	CAPFI 6EME SENS	Paris	57	0.84	-0.1
6202A	I GRAAL	Paris	54	2.57	0.77
6202A	SYXPERIANE	Paris	53	0.79	1.37
6202A	CAP FI BANQUE ET ASSURANCE	Paris	52	0.62	-0.08
6202A	LIPTON FIT	Paris	38	0.6	-0.02

Fig.: Snapshot of the query and principal results.

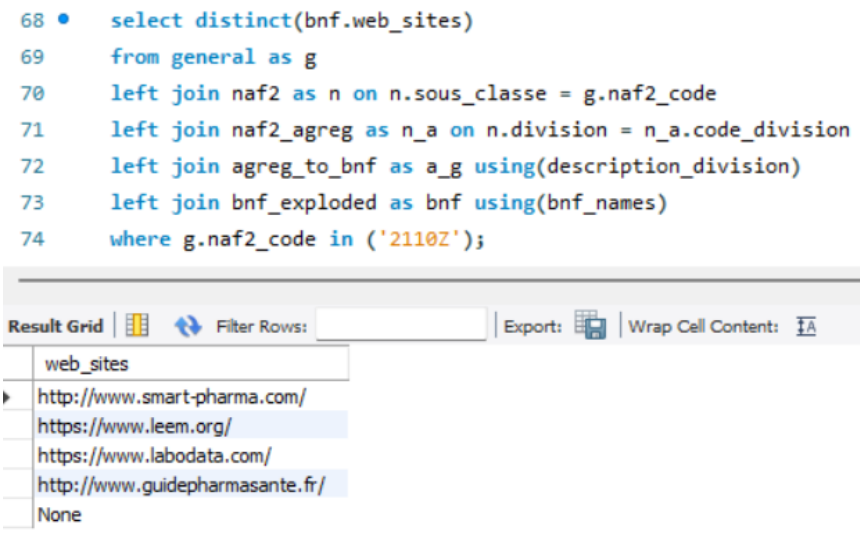


```sql
select i_g.code_ape, n_l.denomination, i_g.departement, n_l.avg_workforce,
round(n_l.revenue_10m,2), round(i_n.resultat_1/1000000,2) as income_in_million
from infogreffe_numbers as i_n
join infogreffe_general as i_g using(siren)
join numbers_light as n_l using(siren)
where i_g.code_ape in ('7022Z') and i_g.num_dept = 75
order by n_l.avg_workforce desc;
```

code_ape	denomination	departement	avg_workforce	round(n_l.revenue_10m,2)	income_in_million
7022Z	NORGAY	Paris	329	1.34	0.11
7022Z	POLYCONSEIL	Paris	234	2.94	2.06
7022Z	TEAM INSIDE	Paris	150	1.76	NULL
7022Z	MA NOUVELLE VILLE	Paris	118	1.2	-1.54
7022Z	ORGANON FRANCE	Paris	114	21.79	7.92
7022Z	ALMAVIA CX	Paris	101	1.53	1.34
7022Z	MASTERCARD FRANCE	Paris	92	2.78	1.67
7022Z	FTI FRANCE	Paris	84	3.74	-0.66
7022Z	ISEA	Paris	76	1.53	0.48
7022Z	CYLAD CONSULTING	Paris	66	1.38	1.01
7022Z	PGCE	Paris	66	0.91	-0.31
7022Z	ALPHA FINANCIAL MA...	Paris	63	1.42	3.69
7022Z	ARAVATI FRANCE	Paris	56	0.97	NULL
7022Z	PERSISTENT SYSTEMS...	Paris	53	0.96	-0.9
7022Z	APAX PARTNERS	Paris	53	6.5	14.06
7022Z	ALEXANDER MANN SO...	Paris	52	1.37	0.21
7022Z	CAREWAN	Paris	48	0.48	-1.46
7022Z	HISTOIRE & PATRIMO...	Paris	47	1.17	0.19
7022Z	WIPRO 4C CONSULTI...	Paris	47	0.54	-2.65
7022Z	PINTEREST FRANCE SAS	Paris	46	1.37	0.64

Fig.: Snapshot of the query and principal results.

e) **Query 6:** And last query to see if we can access to some websites, checking for code 21.10Z (Fabrication de produits pharmaceutiques de base) – the bnf tables had no connection with our usual codes 62.02A & 70.22Z-.



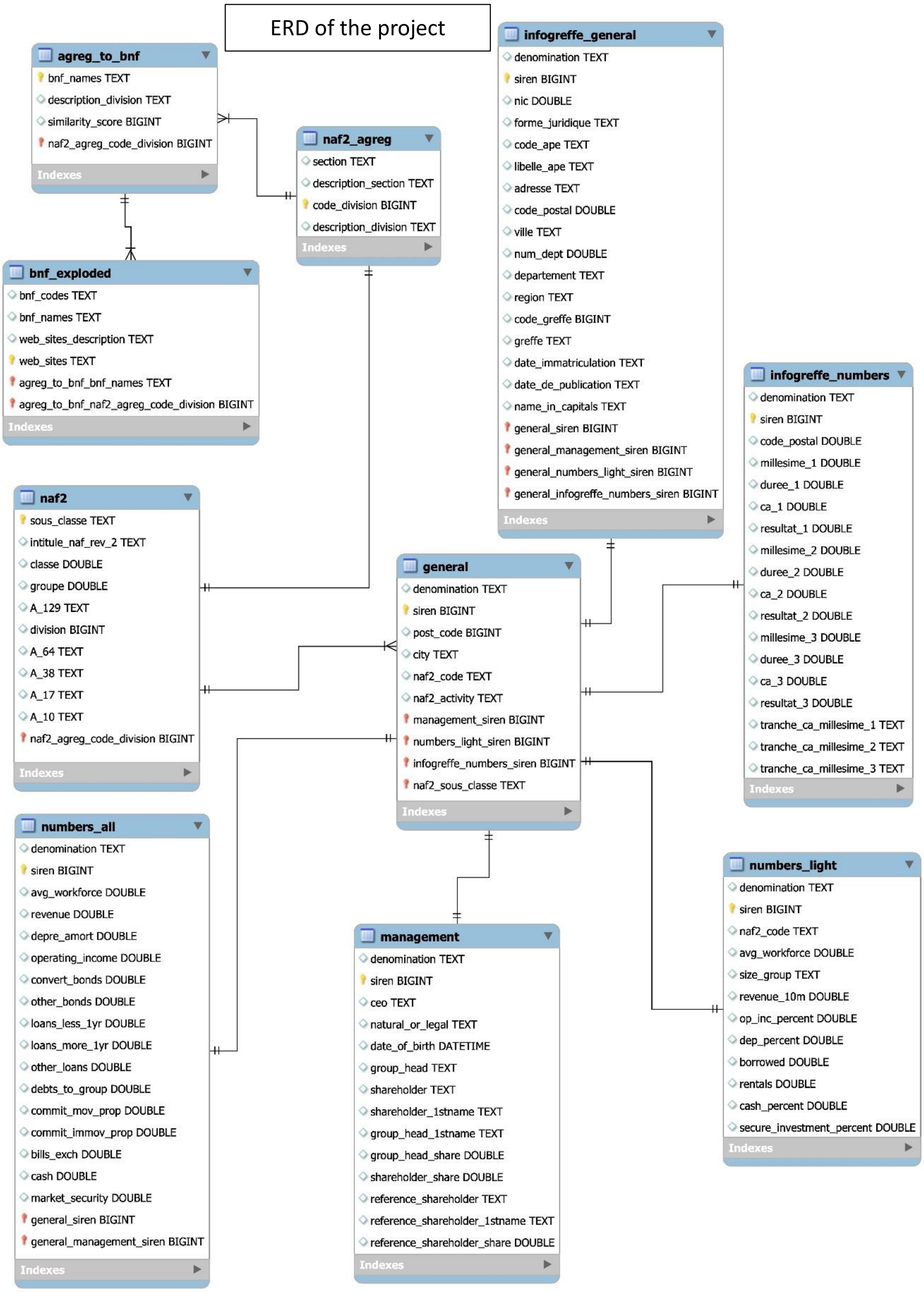
```sql
select distinct(bnf.web_sites)
from general as g
left join naf2 as n on n.sous_classe = g.naf2_code
left join naf2_agreg as n_a on n.division = n_a.code_division
left join agreg_to_bnf as a_g using(description_division)
left join bnf_exploded as bnf using(bnf_names)
where g.naf2_code in ('2110Z');
```

web_sites
http://www.smart-pharma.com/
https://www.leem.org/
https://www.labodata.com/
http://www.guidepharmasante.fr/
None

8.3 ERD

Find as follows the ERD done on the 22/04/2024. The tables ***general*** and ***naf2*** are central to the base.

The tables ***infogreffe_numbers***, ***infogreffe_general***, ***general***, ***numbers_light***, ***numbers_all*** and ***managemen***t can all be joined using their common primary key: **siren** (the company id). If is a one to one relationship.



ERD of the project

Note that ***infogreffe_general*** or ***general*** could be both used as junction to ***naf2***. They connect to ***naf2*** via **sous-classe** connecting to **naf2_code** (or **code_ape** for ***infogreffe_general***). It is a *one-to-many* relationship (one in ***naf2***)

The tables ***infogreffe_general*** and ***numbers_light*** share also this column that can be used for direct joins.

Finally the link from ***naf2*** to ***bnf_exploded*** goes through two junction tables (***naf2_agreg*** and ***agreg_to_bnf***). The table ***naf2_agreg*** connects to ***naf2*** in a *one-to-many* types of relationship (**code_division** to **division**). The table ***naf2_agreg*** connects to ***agreg_to_bnf*** in a *one-to-many* types of relationship (on **description_division**). Finally the table ***agreg_to_bnf*** connects to ***bnf_exploded*** in a *one-to-many* types of relationship (on **bnf_names**).

9 Flask API Creation

9.1 Creation

An API was developed using a .py file and the flask python library. Through our conda terminal to run the app, and on our web-browser, we run the app (to be able check the works of the API).

On the conda terminal, once in the folder where the app1 is, we run the following:

```
flask --app app1 run --port 8080 --debug
```

The checks are then done on the web browser on : http://localhost:8080/docs/

(or http://localhost:8080/company?name=conseil , http://localhost:8080/code_ape?ape_number=7022Z&department=75 etc…)

Three (3) routes were created:

1/ Route /company : to get the code_ape and siren number from 'pieces of company names', presenting more than one result.

2/Route / code_ape: to be able to access to the companies corresponding to a code_ape, with a filter on the department.

3/ Route /siren to be able to access to the company information, through the siren number

The documentation is accessible on /docs, via the openapi created through swagger UI, using a yazzle format.

Here is the introductory text:

This API exposes the Match Project dataset. It is a mix of four data sources:

* The infogreffe opendata ; https://opendata.datainfogreffe.fr/explore/dataset/chiffres-cles-2022/.

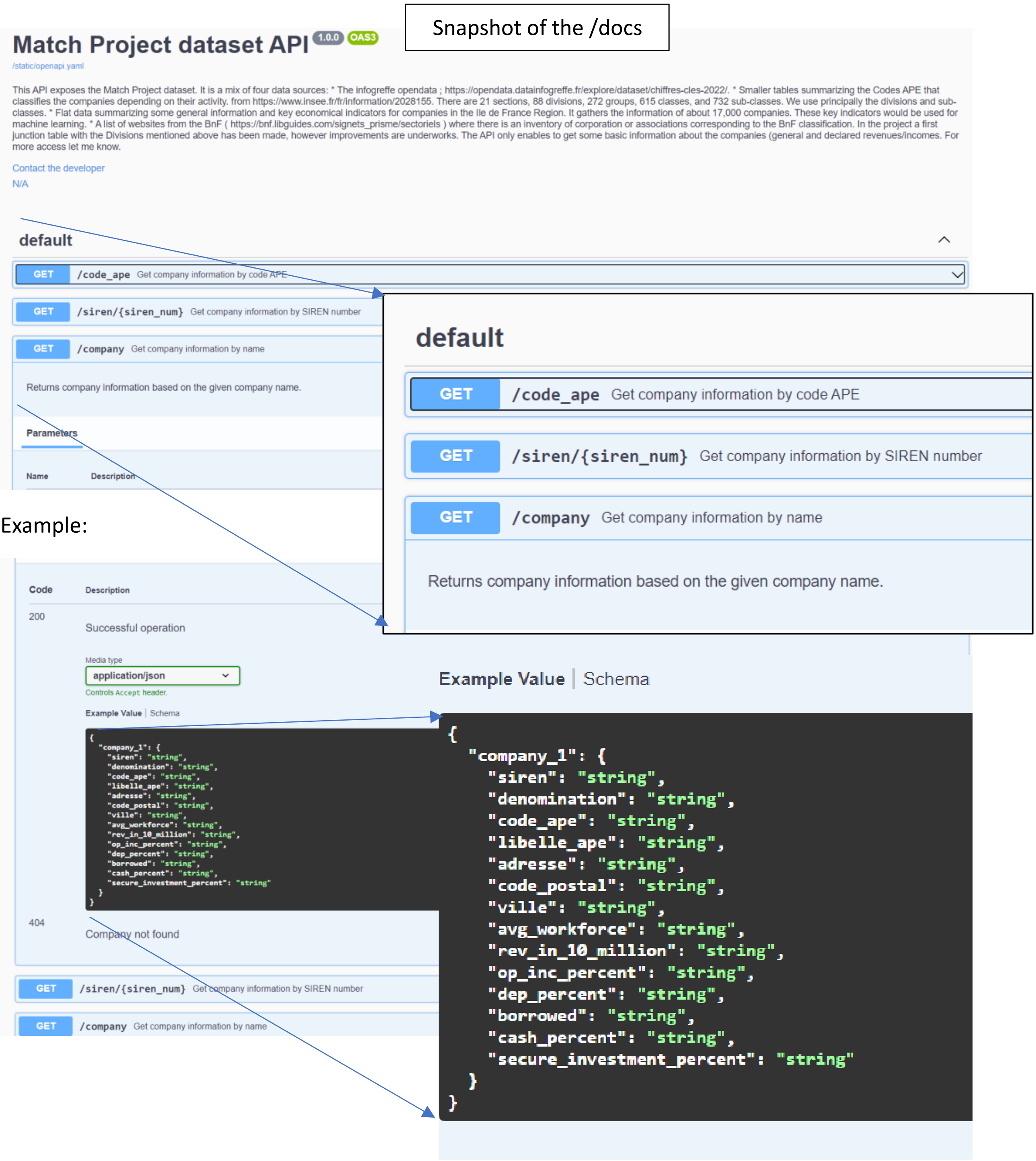
* Smaller tables summarizing the Codes APE that classifies the companies depending on their activity. from https://www.insee.fr/fr/information/2028155. There are 21 sections, 88 divisions, 272 groups, 615 classes, and 732 sub-classes. We use principally the divisions and sub-classes.

* Flat data summarizing some general information and key economical indicators for companies in the Ile de France Region. It gathers the information of about 17,000 companies. These key indicators would be used for machine learning.

* A list of websites from the BnF (https://bnf.libguides.com/signets_prisme/sectoriels) where there is an inventory of corporation or associations corresponding to the BnF classification. In the project a first junction table with the Divisions mentioned above has been made, however improvements are underworks.

The API only enables to get some basic information about the companies (general and declared revenues/incomes. For more access let me know

9.2 Exposing data



Snapshot of the /docs

Match Project dataset API 1.0.0 OAS3

/static/openapi.yaml

This API exposes the Match Project dataset. It is a mix of four data sources: * The infogreffe opendata ; https://opendata.datainfogreffe.fr/explore/dataset/chiffres-cles-2022/. * Smaller tables summarizing the Codes APE that classifies the companies depending on their activity. from https://www.insee.fr/fr/information/2028155. There are 21 sections, 88 divisions, 272 groups, 615 classes, and 732 sub-classes. We use principally the divisions and sub-classes. * Flat data summarizing some general information and key economical indicators for companies in the Ile de France Region. It gathers the information of about 17,000 companies. These key indicators would be used for machine learning. * A list of websites from the BnF (https://bnf.libguides.com/signets_prisme/sectoriels) where there is an inventory of corporation or associations corresponding to the BnF classification. In the project a first junction table with the Divisions mentioned above has been made, however improvements are underworks. The API only enables to get some basic information about the companies (general and declared revenues/incomes. For more access let me know.

Contact the developer

N/A

default

GET /code_ape Get company information by code APE

GET /siren/{siren_num} Get company information by SIREN number

GET /company Get company information by name

Returns company information based on the given company name.

Parameters

Name Description

Example:

Code	Description
200	Successful operation Media type application/json Controls Accept header. Example Value | Schema
404	Company not found

```
{
  "company_1": {
    "siren": "string",
    "denomination": "string",
    "code_ape": "string",
    "libelle_ape": "string",
    "adresse": "string",
    "code_postal": "string",
    "ville": "string",
    "avg_workforce": "string",
    "rev_in_10_million": "string",
    "op_inc_percent": "string",
    "dep_percent": "string",
    "borrowed": "string",
    "cash_percent": "string",
    "secure_investment_percent": "string"
  }
}
```

GET /siren/{siren_num} Get company information by SIREN number

GET /company Get company information by name

10 Preliminary Machine Learning

10.1 Considerations

The machine learning exercise is based on the features from ***numbers_light*** and ***general tables***

It is going to be mainly using KMeans as the clustering technique around different financial information (revenue, operating income, depreciation etc…) together with the average workforce.*

It is an attempt to identify patterns in the way companies might be managed.

10.2 1st model: not scaled KMeans, all numerical values of numbers_light

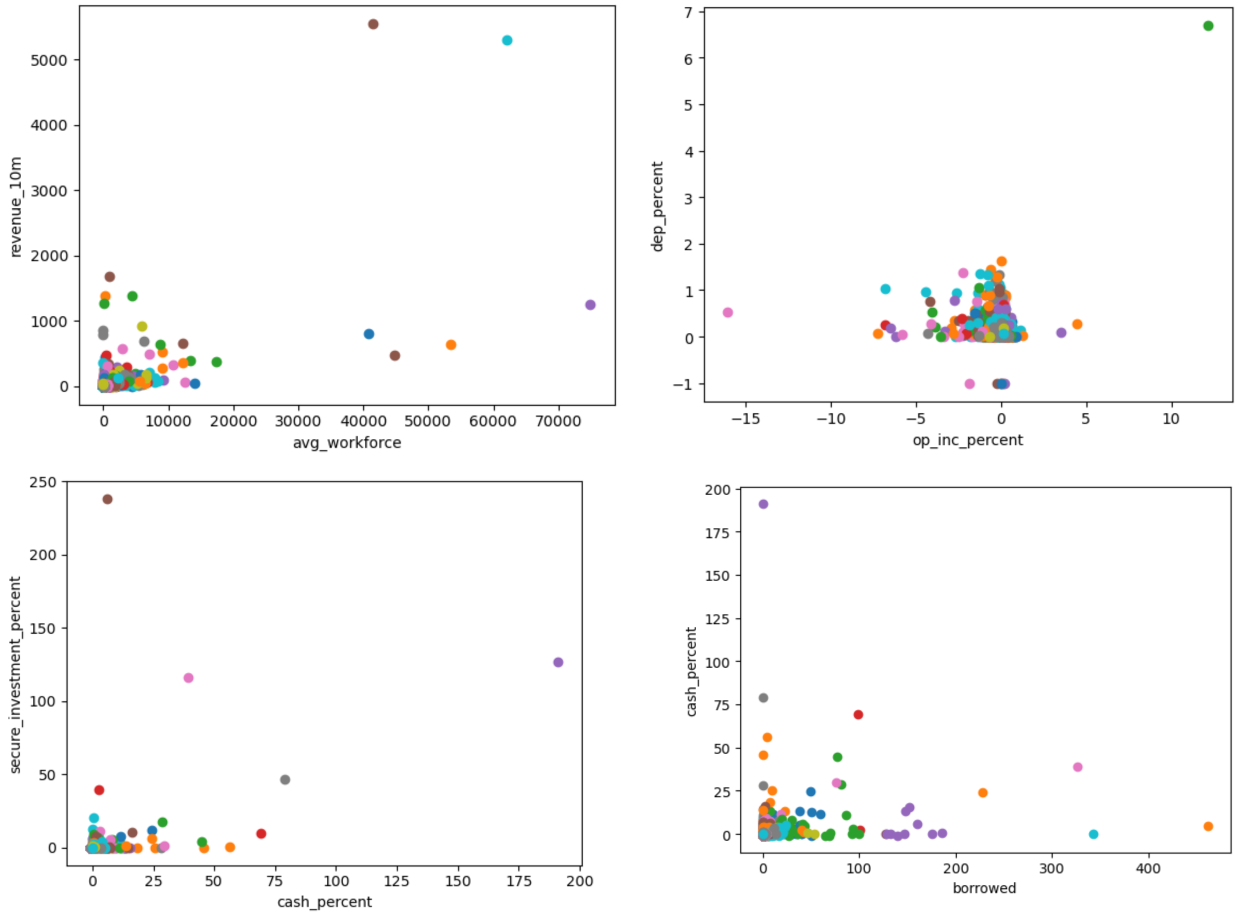
features_model_1_on, : 8 dimensions, 200 clusters

avg_workforce, revenue_10m, op_inc_percent, dep_percent, borrowed, rentals, cash_percent, secure_investment_percent

The idea was to make clusters with a more important weight on revenue (ranging from 2.8million€ to 55000 million €) by using the 10 million scale.

Find below some crossplots of the cluster combining two dimensions

Example of four crossplots comparing some dimensions with each other, for all the clusters (colours)



The bottom right crossplots seems to show clustering, we guess groups of the same colours. To compare the clusters, I generated 10 random examples of each category size of the avg_workforce.

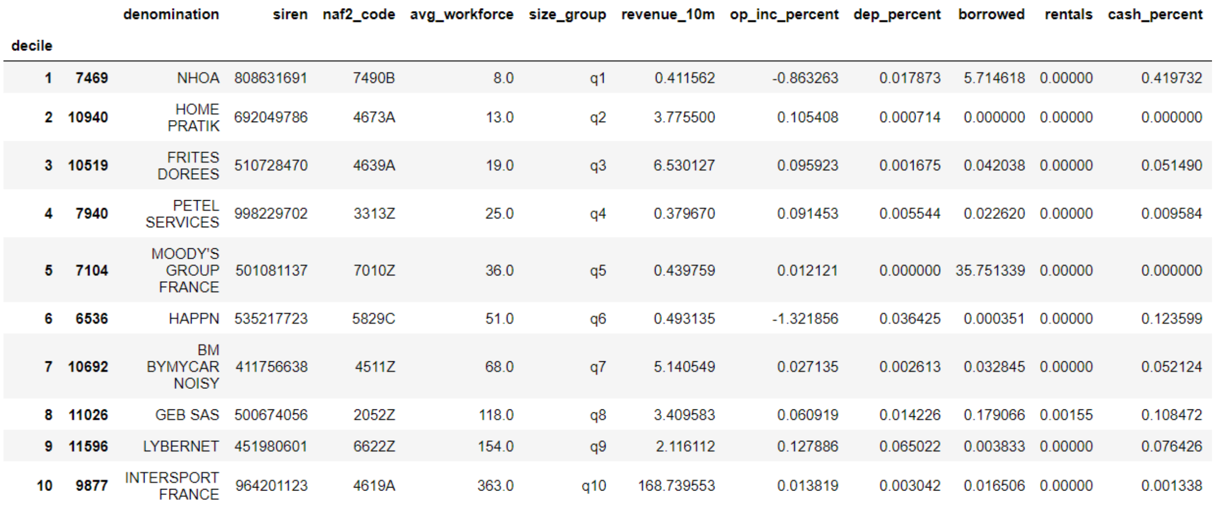
```
# selecting 10 companies to check, 1 per decile excluding the -1
# Defining a mapping from string values to numerical values representing
deciles
decile_map = {'q1': 1, 'q2': 2, 'q3': 3, 'q4': 4, 'q5': 5, 'q6': 6, 'q7': 7,
'q8': 8, 'q9': 9, 'q10': 10}
df_numbers_light_for_random = df_numbers_light.copy()

# Mapping the string values to numerical values
df_numbers_light_for_random['decile'] =
df_numbers_light_for_random['size_group'].map(decile_map)

def select_random_row(group):
    """function to select one random row from each group"""
    return group.sample(n=1)

# Grouping by decile and apply the function to select one random row from each
group
random_rows =
df_numbers_light_for_random.groupby('decile').apply(select_random_row)

random_rows
```



decile		denomination	siren	naf2_code	avg_workforce	size_group	revenue_10m	op_inc_percent	dep_percent	borrowed	rentals	cash_percent
1	7469	NHOA	808631691	7490B	8.0	q1	0.411562	-0.863263	0.017873	5.714618	0.00000	0.419732
2	10940	HOME PRATIK	692049786	4673A	13.0	q2	3.775500	0.105408	0.000714	0.000000	0.00000	0.000000
3	10519	FRITES DOREES	510728470	4639A	19.0	q3	6.530127	0.095923	0.001675	0.042038	0.00000	0.051490
4	7940	PETEL SERVICES	998229702	3313Z	25.0	q4	0.379670	0.091453	0.005544	0.022620	0.00000	0.009584
5	7104	MOODY'S GROUP FRANCE	501081137	7010Z	36.0	q5	0.439759	0.012121	0.000000	35.751339	0.00000	0.000000
6	6536	HAPPN	535217723	5829C	51.0	q6	0.493135	-1.321856	0.036425	0.000351	0.00000	0.123599
7	10692	BM BYMYCAR NOISY	411756638	4511Z	68.0	q7	5.140549	0.027135	0.002613	0.032845	0.00000	0.052124
8	11026	GEB SAS	500674056	2052Z	118.0	q8	3.409583	0.060919	0.014226	0.179066	0.00155	0.108472
9	11596	LYBERNET	451980601	6622Z	154.0	q9	2.116112	0.127886	0.065022	0.003833	0.00000	0.076426
10	9877	INTERSPORT FRANCE	964201123	4619A	363.0	q10	168.739553	0.013819	0.003042	0.016506	0.00000	0.001338

Then I visualized the 10 clusters of these examples.

```
def cluster_check(num):
    """getting the table of the cluster from the siren number"""
    cluster = companies_clustered_df[companies_clustered_df['siren'] ==
num]['cluster_km200'].iloc[0]
    return companies_clustered_df[companies_clustered_df['cluster_km200'] ==
cluster]
```

Looking at the clusters… it basically groups the data according to the avg_workforce as the data is not scaled for the worforce and it takes too much importance. This model 1 is not conclusive.

10.3 2nd model: scaled KMeans, all numerical values of numbers_light

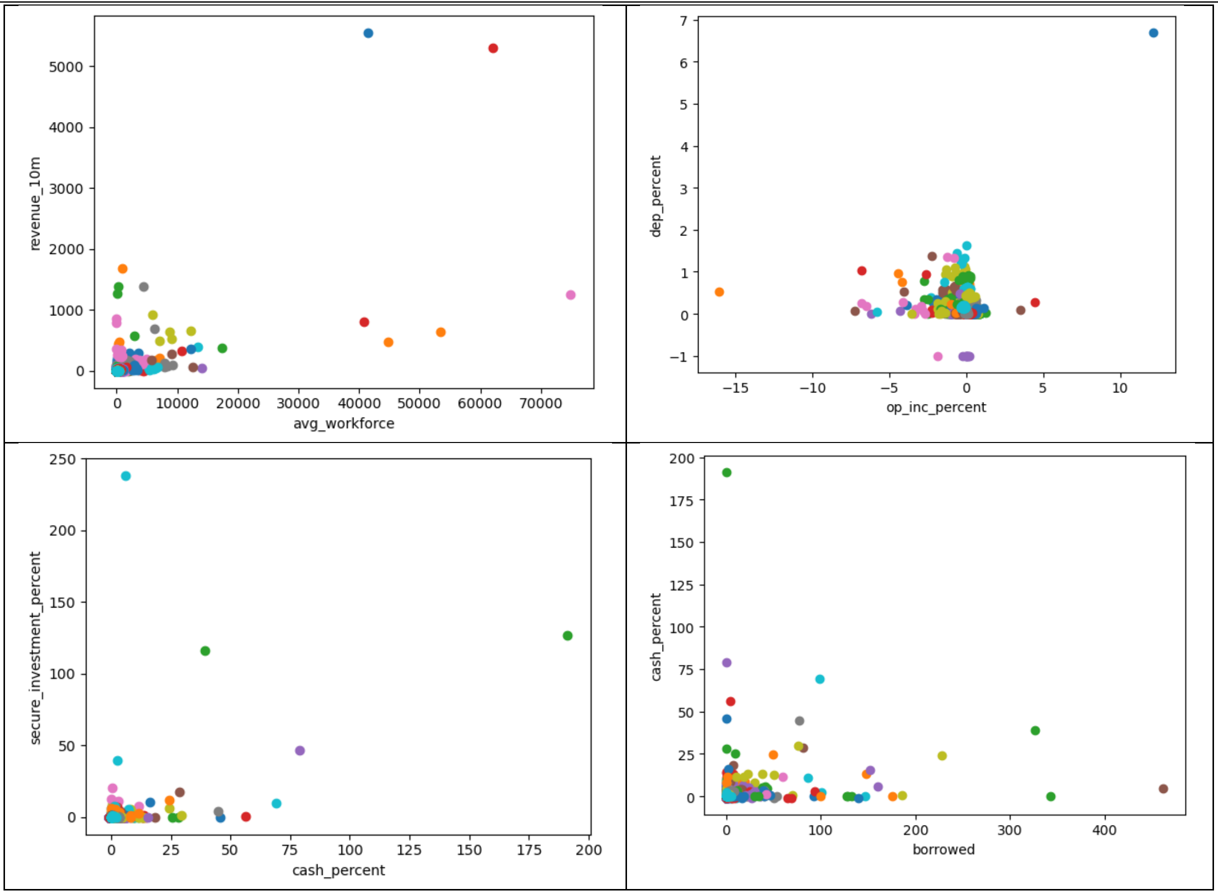
We had time to run another model, model 2, this time scaling the data using MinMax Scaler. The other features are unchanged.

features_model_2_on : 8 dimensions, 200 clusters

avg_workforce, revenue_10m, op_inc_percent, dep_percent, borrowed, rentals, cash_percent, secure_investment_percent

Find below some crossplots of the cluster combining two dimensions:

Example of four crossplots comparing some dimensions with each other, for all the clusters (colours)



The visualization, although colorful does not show obvious clustering. It is difficult to estimate from the plots if the model is clustering properly the different companies.

About 10,000 companies were used to generate the model (after filtering on missing data). As a quick check, we ran through the size of the 200 clusters (see the snapshot below of the counts per cluster). Many clusters only have 1 company in it, the two biggest gather about the third of the population and the top 5 is close to half.

```
cluster_km200
0       1807
131     1310
184      686
174      509
134      443
         ...
```

```
46         1161       1
45         1
171        1
133        1
Name: count, Length: 200, dtype: int64
```

Count per clusters (extract from jupyter notebook)

This alone would make the clustering and model 2 inconclusive.

10.4 Way forward

The next attempts should reduce the number of dimensions, possibly removing workforce and/or revenue (they are possibly not independent and hence tilting the modelling) and use them more as a filter within the clusters.

Also after acquiring more data from pappers, there would be enough information to start another model using different features available on the data to gather.

This is work for the next weeks to come!

11 Conclusions

The idea of the project was to make a tool to help, for possible market analysis and help in identifying partner companies.

The tool is onworks, but still shows promising initial results. It is a very interesting learning experience.

Through the gathering of data, and its combination we have now a database that allows us to get some insights. Using the codes APE 62.02A (Conseil en systèmes et logiciels informatiques) and APE 70.22Z (Conseil pour les affaires et autres conseils de gestion) in Paris, we could see the players of this sectors, evaluate how the activity is doing compared to the rest of activities or how some company perform vs the overall group. In 2022, these two codes performed better and differently than the French global trend. The financial leaders could be identified (via the ratio of income). Additionally it gives some insights of the market. For instance for 70.22Z, in Paris, there are many companies of a relatively ‘small size’ compared to the code 62.02A.

Last and not least, for the most complete information, a one to one comparison could be made on the more specific features depending on the goals (looking for partners, looking for merging, pure market analysis).

The main challenges were identifying suitable data available on official web or other sources, and understanding how to use the available data. Creating links between datasets is also not always straightforward. Time is the essence, so focusing on part of the data (70.22Z for instance) will allow to define a tool and a model that can then be implemented to others on a larger scale.

Many improvements can be made, the visualization could be generalized to create a company id with some activity comparisons. More web scrapping of the identified web sites to be able to create secondary codes to identify markets Also the machine learning preliminary phases show it is not straightforward to cluster companies from only some financial features.

Last and not list, the quest for information is linked with the quest for data, and to improve the tool, updating the data with time. There are already more data on infogreffe (from previous years) and some data could be added from pappers. Looking forward to undertake these new tasks.

12 References

Flat files:

- https://opendata.datainfogreffe.fr/explore/dataset/chiffres-cles-2022/information/
- https://opendata.datainfogreffe.fr/explore/dataset/chiffres-cles-2022/
- https://www.insee.fr/fr/information/2028155
- general, numbers_all, management from personal source extracted most likely from the Cap-Fi database

Web scrapping:

- https://bnf.libguides.com/signets_prisme/sectoriels

API:

- https://www.pappers.fr/
- https://api.pappers.fr/v2/entreprise

In progress repository:

- https://github.com/javierpeyriere/match_and_find_project