Dataset Selection and Framing

Auteurs :		*Enseignant :*
Vincent	MORIN	Mme. MOKDAD ANAIS
Antonin	DOAT	
Augustin	MOUTON	
Maxime	LAMBERT	
Maxime	MAEDER	
Mathias	ROBERT	

PARTIE I. Context

This project aims to analyze the local finances of French territorial authorities (communes, departments, and regions) using public data from the Directorate General of Public Finances (DGFiP).

The overall objective is to better understand the distribution of financial resources, territorial disparities, and budgetary trends across different administrative levels.

The data comes from three main datasets:

- **Communes**: detailed budgetary information by commune (revenues, expenditures, population, etc.)
- **Departments**: aggregated data by department
- **Regions**: consolidated regional budgets

This analysis provides insights into local governance, territorial equity, and public financial planning.

PARTIE II. Dataset description

A. Régions

Rows : 22 933

Columns : 31

Variable types : object, int64, float64

B. Départements

Rows : 290 281

Columns : 35

Variable types : object, int64, float64

C. Communes

Rows : 3 887 010

Columns : 43

Variable types : object, int64, float64

PARTIE III. Key variables

A. Régions

Code Insee 2024 Région, Nom 2024 Région: identifiers and names of the regions.

Montant / Montant en millions / Montant en € per capita: main financial indicators.

Population totale: used for relative calculations and comparisons.

Type de budget: principal or supplementary budget.

Exercice: to track temporal trends.

B. Départements

Code Insee 2024 Département, Nom 2024 Département: identifiers and names of the departments.

Montant / Montant en millions / Montant en € per capita: main financial indicators.

Population totale: used to normalize and compare budgets.

Type de budget: principal or supplementary budget.

Exercice: to analyze budgetary changes over time.

C. Communes

Code Insee 2024 Commune: unique identifier for each commune.

Nom 2024 Commune: name of the commune for readability and visualizations.

Code Siren 2024 EPCI: identifier of the EPCI, used to group communes.

Montant / Montant en millions: total amounts of revenues or expenditures.

Montant en € per capita: relative indicator to compare communes of different sizes.

Population totale: used to calculate ratios and analyze disparities.

Type de budget: principal or supplementary budget, to differentiate revenues and expenditures.

Rural / Urban, Mountain commune, Tourist commune, QPV presence: socio-territorial characteristics used to analyze disparities.

Exercice / annee_join: to study temporal trends

PARTIE IV. Analytical objectives

Objective 1: Examine the distribution of financial resources among local authorities (communes, departments, regions) over multiple fiscal years, analyzing both total amounts and per capita indicators.

Objective 2: Study territorial disparities based on population size, degree of rurality or urbanity, and the socio-economic characteristics of communes.

Objective 3: Identify major trends and detect potential anomalies in public budgets (revenues and expenditures), including atypical year-to-year variations.

Objective 4: Compare financial performance according to budget type (principal vs. supplementary) and across different administrative structures (communes, EPCIs, departments, regions).

Objective 5: Provide actionable insights for public decision-making, particularly regarding the allocation of resources and budgetary planning.

Objective 6: Assess the impact of specific territorial characteristics — such as QPV presence, tourism activity, or mountain commune status — on revenues and expenditures, to better understand the factors driving financial variations between local authorities.

PARTIE V. Business questions

Question 1: Which local authorities show the highest levels of expenditures and revenues per capita?
→ Related to Objective 1

Question 2: How have public budgets evolved between 2020 and 2024 across different types of local authorities (communes, departments, regions, EPCIs)?
→ Related to Objective 3

Question 3: Are there significant differences in local finances between rural and urban communes?
→ Related to Objective 2

Question 4: Which EPCIs or communes exhibit atypical or inconsistent budgetary variations from one year to another?
→ Related to Objective 3

Question 5: Which key financial indicators (KPIs) can effectively assess the budgetary health of a territory?
→ Related to Objectives 4 and 5

PARTIE VI. Technical constraints

A. Régions

Some columns in the dataset contain missing values. The `ordre_analyse*_section*` columns, which have a large number of nulls, are not relevant for the analysis and will be removed. For key columns such as budget, population, or region identifiers, no significant missing values are present, so they will be used directly in the analysis.

All columns retained for analysis have consistent data types and do not require modification.

Finally, no duplicate rows were detected in the dataset.

B. Départements

The situation is similar to that of the regions. The `ordre_analyse*_section*` columns will be removed, and no significant missing values are present for the key columns.

All columns retained for analysis have consistent data types and do not require modification.

No duplicate rows were detected in the dataset.

C. Commune

The situation is still similar to that of the regions and departments, except for certain specifics related to data types. The `ordre_analyse*_section*` columns will also be removed. For the key columns, missing values will either be replaced with appropriate values if they represent important indicators, or ignored if their absence does not affect the analysis.

Some columns have data type issues that will need to be corrected to ensure dataset consistency during cleaning:

- **Code Siren 2024 EPCI (float64)**: EPCI identifier; will need to be converted to integer or string to ensure uniqueness and facilitate grouping.

- **Population totale (float64)**: used to calculate ratios; type is correct but may be converted to integer for readability.

- **Exercice / annee_join (int64 / float64)**: fiscal years; the float type may be converted to integer during cleaning.

Finally, no duplicate rows were detected in the dataset.

PARTIE VII. Tools used

For the first phase of data analysis and preparation, we used Python and its main libraries:

- **pandas**: data manipulation and cleaning
- **numpy**: numerical calculations and vectorized operations
- **matplotlib** and seaborn: statistical visualizations and exploratory charts

- **scikit-learn**: advanced analyses and machine learning methods for data exploration

For the interactive visualization phase and dashboard creation, we used Power BI.