# Interactive Information Visualization: Plotwisters Dataset Exploration & Cleaning of French Campus Close-Up

Pablo Mollá Chárlez, Pavlo Poliuha and Junjie Chen January 16, 2025

#### Contents

5	Resources Availability	4
	4.3 Cycling Dataset (Initial Size: 23 MB $\rightarrow$ Final Size: 9.3 MB)	4
	4.2 Rental Affordability (Initial Size: 4.6 MB $\rightarrow$ Final Size: 1.3 MB)	3
	4.1 French Institutions (Initial Size: $0.3 \text{ MB} \rightarrow \text{Final Size}$ : $0.026 \text{ MB}$ )	
4	Data Cleaning	3
3	Modification of Research Problem	2
<b>2</b>	Dataset Exploration	1
1	Introduction: Team & Topic	1

# 1 Introduction: Team & Topic

The following project, titled French Campus Close-Up, is conducted by Team Plotwisters composed of Junjie Chen, Pavlo Poliuha, and Pablo Mollá.

The purpose of this project is to combine three key datasets—French Institutions of Higher Education (Geolocation), Rental Affordability, and Cycling Stations in France—to determine which towns in France meet specific criteria regarding campus proximity to surveilled, private-access cycling stations and affordable rental housing. By focusing on these conditions, we can pinpoint localities that offer both convenient sustainable transport options and student-friendly living costs.

# 2 Dataset Exploration

Below are the three datasets we identified for our study, each accompanied by a brief description:

- 1. French Institutions of Higher Education (Geolocation)
  - Found by: Pavlo
  - Source: French Institutions of Higher Education (Geolocation)
  - **Description:** A listing of universities and higher education institutions in France, containing both geolocation data and institutional attributes.
  - Organization: Published by the Ministère de l'Enseignement supérieur et de la Recherche.

#### 2. Rental Affordability

• Found by: Junjie

• Source: Rental Affordability

- **Description:** Contains data on rental prices per square meter in French municipalities, useful for evaluating housing affordability.
- Organization: Published by the Ministère de la Transition écologique.

#### 3. Cycling Stations in France

• Found by: Pablo

• Source: Cycling Stations in France

- **Description:** Provides detailed information on cycling station locations across France, including capacity, surveillance, and access type.
- Organization: Published by OpenStreetMap contributors.

We chose these three datasets because they collectively provide the necessary information to address our main question about which French towns (1) host universities close to private-access, surveilled cycling stations and (2) offer affordable housing. More specifically:

- French Institutions of Higher Education (Geolocation): Identifies where universities are located and confirms their presence within specific communes. The dataset was created the 24th April 2024 and last modified the 1st January 2025. The publisher is Ministère de l'Enseignement supérieur et de la Recherche. It contains a total of 251 rows and 98 different variables/columns, from which we filtered most of them. The total file size is 0.3mb.
- Rental Affordability: Supplies municipality-level rental price indicators, ensuring that we can assess whether housing is "affordable." The dataset was originally created in 2018 and lastly modified the 23 December 2024. The publisher is Ministère de la Transition écologique. It contains a total of 34960 rows and 12 different variables/columns, from which we filtered some of them. The total file size is 4.6mb.
- Cycling Stations in France: Reveals where cycling stations are placed, allowing us to check if they are private-access, equipped with surveillance, and within two kilometers of the universities. The dataset was originally created the 24 September 2021 and lastly modified the 15 January 2025. The publisher is OpenStreetMap. It contains a total of 117986 rows and 22 different variables/columns, from which we filtered some of them. The total file size is 23mb.

These datasets were each published on official French government platforms (data.gouv.fr and enseignementsup-recherche.gouv.fr) and aggregated from institutional sources. Each dataset is maintained and released by a reputable government entity or open-data organization, ensuring overall trustworthiness and official coverage of data across France.

## 3 Modification of Research Problem

Our original research question additionally required that the town's population be below 300,000 and its international student enrollment exceed the national average for similar-sized university towns. To streamline our study, we removed these criteria and now focus solely on:

Which towns in France meet all the following criteria:

- 1. They host universities with campuses located within 2 kilometers of private-access cycling stations equipped with surveillance.
- 2. They offer affordable rental housing for students.

By omitting population size and international enrollment factors, we concentrate on campus proximity and housing affordability, making our inquiry more direct and manageable.

## 4 Data Cleaning

Each dataset underwent a tailored cleaning process, primarily using OpenRefine, to ensure we could join and compare them effectively.

## 4.1 French Institutions (Initial Size: 0.3 MB → Final Size: 0.026 MB)

- Variables Kept (8/98): identifiant\_interne, libellé, type d'établissement, secteur d'établissement, géolocalisation, Commune, Unité urbaine, Code commune
- Geolocation Split: We split the single géolocalisation field into separate X and Y columns, then removed the original field.
- Non-French Institutions Removed: We excluded 4 records of institutions located outside France.

The cleaning procedures resulted in:

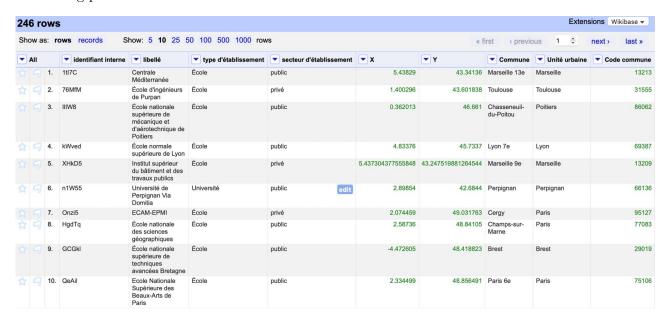


Figure 1: French Institutions of Higher Education: Cleaned Dataset Version

#### 4.2 Rental Affordability (Initial Size: 4.6 MB → Final Size: 1.3 MB)

- Columns Kept (3/12): INSEE\_C, LIBGEO, loypredm2
- Column Renaming:
  - INSEE\_C: This column provides the official numerical identifier (assigned by INSEE, France's national statistics bureau) for each French commune. It is often used for data matching or joining with other administrative datasets. The variable is renamed to code\_commune.
  - LIBGEO: This column holds the textual (human-readable) name of the French municipality. For example: "Paris", "Toulouse", or "Lyon". The variable is renamed to **commune**.
  - loypredm2: This column indicates the predicted or average rent price, in euros, per square meter. It is a key metric for assessing affordability at the municipality level. The variable is renamed to price\_euro\_m2.
- Numeric Conversion: Replaced commas with decimal points and converted strings to numeric values using: toNumber(replace(value,",","."))

The cleaning procedures resulted in:



Figure 2: Rent Affordability in France

## 4.3 Cycling Dataset (Initial Size: 23 MB → Final Size: 9.3 MB)

- Blank Row Removal: We eliminated ~10% of rows in the capacity column that were entirely blank.
- String Cleaning: Used chained replace() functions to remove non-numeric characters (e.g., "arceaux", "6-7", "O", etc.) so that capacity is numeric.
- Columns Removed: id\_osm, coordonessxy, lumiere, url\_info, capacite\_cargo, d\_service, proprietai
  gestionnaire, commentaires, source, date\_maj, type\_accroche, protection (mostly empty or duplicated info).
- Data Type Conversion: Converted X, Y, code\_com, and capacity to numeric data types.

The cleaning procedures resulted in:

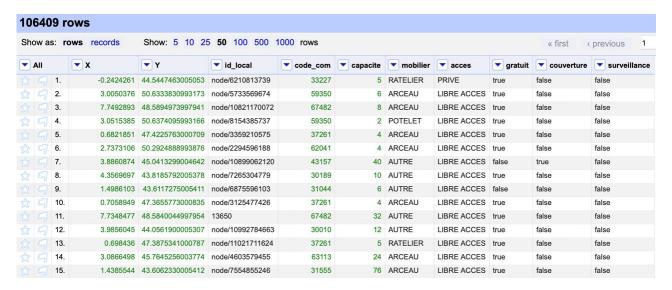


Figure 3: Rent Affordability in France

# 5 Resources Availability

All documentation can be found in our shared folder here: IIV Project (Google Drive Link) where we distinguish between the Original Datasets, Modified Datasets and Data Cleaning Operations.