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

Pablo Mollá Chárlez, Pavlo Poliuha and Junjie Chen

January 16, 2025

Contents

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:

246 rows Extensions: Wikibase

Show as: **rows** records Show: 5 **10** 25 50 100 500 1000 rows « first ‹ previous 1 next › last »

All		identifiant interne	libellé	type d'établissement	secteur d'établissement	X	Y	Commune	Unité urbaine	Code commune
1.	1tl7C	Centrale Méditerranée	École	public		5.43829	43.34136	Marseille 13e	Marseille	13213
2.	76MfM	École d'ingénieurs de Purpan	École	privé		1.400296	43.601838	Toulouse	Toulouse	31555
3.	IIIW8	École nationale supérieure de mécanique et d'aérotechnique de Poitiers	École	public		0.362013	46.661	Chasseneuil-du-Poitou	Poitiers	86062
4.	kWved	École normale supérieure de Lyon	École	public		4.83376	45.7337	Lyon 7e	Lyon	69387
5.	XHkD5	Institut supérieur du bâtiment et des travaux publics	École	privé		5.437304377555848	43.247519881264544	Marseille 9e	Marseille	13209
6.	n1W55	Université de Perpignan Via Domitia	Université	public	edit	2.89854	42.6844	Perpignan	Perpignan	66136
7.	Onzi5	ECAM-EPMI	École	privé		2.074459	49.031763	Cergy	Paris	95127
8.	HgdTq	École nationale des sciences géographiques	École	public		2.58736	48.84105	Champs-sur-Marne	Paris	77083
9.	GCGkl	École nationale supérieure de techniques avancées Bretagne	École	public		-4.472605	48.418823	Brest	Brest	29019
10.	QeAil	Ecole Nationale Supérieure des Beaux-Arts de Paris	École	public		2.334499	48.856491	Paris 6e	Paris	75106

Figure 1: French Insititutions 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:

34960 rows

Show as: **rows** records Show: 5 **10** 25 50 100 500 1000 rows

All	code_commune	commune	price_euro_m2
1.	31529	Samouillan	10.8814528339144
2.	31343	Mirambeau	10.8814528339144
3.	32185	Lalanne-Arqué	10.8814528339144
4.	31008	Anan	10.8814528339144
5.	32410	Samatan	12.2633693497271
6.	31477	Saint-Élix-Séglan	10.8814528339144
7.	31001	Agassac	10.881452833 edit
8.	32186	Lamaguère	10.8814528339144
9.	32438	Tachoires	10.8814528339144
10.	31347	Molas	10.8814528339144

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:

106409 rows

Show as: **rows** records Show: 5 10 25 **50** 100 500 1000 rows « first ‹ previous 1

All	X	Y	id_local	code_com	capacite	mobilier	acces	gratuit	couverture	surveillance
1.	-0.2424261	44.5447463005053	node/6210813739	33227	5	RATELIER	PRIVE	true	false	false
2.	3.0050376	50.6333830993173	node/5733569674	59350	6	ARCEAU	LIBRE ACCES	true	false	false
3.	7.7492893	48.5894973997941	node/10821170072	67482	8	ARCEAU	LIBRE ACCES	true	false	false
4.	3.0515385	50.6374095993166	node/8154385737	59350	2	POTELET	LIBRE ACCES	true	false	false
5.	0.6821851	47.4225763000709	node/3359210575	37261	4	ARCEAU	LIBRE ACCES	true	false	false
6.	2.7373106	50.2924888993876	node/2294596188	62041	4	ARCEAU	LIBRE ACCES	true	false	false
7.	3.8860874	45.0413299004642	node/10899062120	43157	40	AUTRE	LIBRE ACCES	false	true	false
8.	4.3569697	43.8185792005378	node/7265304779	30189	10	AUTRE	LIBRE ACCES	true	false	false
9.	1.4986103	43.6117275005411	node/6875596103	31044	6	AUTRE	LIBRE ACCES	false	false	false
10.	0.7058949	47.3655773000835	node/3125477426	37261	4	ARCEAU	LIBRE ACCES	true	false	false
11.	7.7348477	48.5840044997954	13650	67482	32	AUTRE	LIBRE ACCES	true	false	false
12.	3.9856045	44.0561900005307	node/10992784663	30010	12	AUTRE	LIBRE ACCES	true	false	false
13.	0.698436	47.3875341000787	node/11021711624	37261	5	RATELIER	LIBRE ACCES	true	false	false
14.	3.0866498	45.7645256003774	node/4603579455	63113	24	ARCEAU	LIBRE ACCES	true	false	false
15.	1.4385544	43.6062330005412	node/7554855246	31555	76	ARCEAU	LIBRE ACCES	true	false	false

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**.