Final Project: One Anomaly, Defended

IELE756 – Preparación y Análisis de Datos

Leo Ferres, PhD

May 9, 2026

Final Project: One Anomaly, Defended

Points: 25
Released: Friday, May 9, 2026
Anomaly proposal due: Friday, May 15, 2026 (gate, ungraded)
Repo + video due: Friday, May 29, 2026 (Canvas)
In-person session (handwritten memo): Friday, June 5, 2026
Submission: Canvas (PDF of `final_anomaly.ipynb`, link to your GitHub repo, link to your video)

Why this format

Tarea 3 already asked you to integrate three datasets at the comuna level and write 800 words of synthesis. A second, broader summary would mostly repeat that work. The final project does the opposite: it asks you to **shrink the scope and grow the depth**.

By the end of Tarea 3 your pipeline has produced thousands of numbers: counts, rates, coefficients, residuals, predicted values. Most of them are uninteresting. A few are not. Your job for the final project is to pick **one** finding from your pipeline that genuinely surprised you, and explain it well enough that an outsider believes you have understood it.

This is also an exercise in critical thinking *without leaning on AI*. The repo and the video are async and you may use any tools you like to produce them, including AI assistants, with appropriate disclosure in the README. But the third component is taken in person, on paper, and it will tell us whether you actually own what your repo claims.

Components at a glance

Component	Mode	Pts
1. GitHub repo	Async, pair	10
2. Video	Async, pair (8 to 10 min)	7
3. Handwritten in-person memo	Sync, **individual** (45 min)	8
Total		**25**

Same pairs as the Tareas. Both members co-own the repo and the video; the memo grades each member individually even though the project is paired.

What is an "anomaly"?

An anomaly is a single, concrete, *defensible* finding from your pipeline that you did not expect. Examples that would qualify:

- A comuna whose ENO notification rate is more than two standard deviations from the regional mean for a specific disease, after adjusting for population.
- A coefficient whose sign flips between Poisson and Negative Binomial in your Tarea 3 model, with material consequences for interpretation.
- A subgroup (a nationality, an age band, an occupation) whose hospitalization length-of-stay distribution has a heavy right tail not visible at the comuna level.
- An apparent migration corridor (place of residence 5 years ago vs. current comuna) that is much larger or smaller than you would guess from population alone.
- A choropleth where the spatial pattern is the *opposite* of what the demographic correlation predicts.

Examples that would **not** qualify:

- "Foreign-born residents have a higher TB notification rate." (Already documented; not a surprise.)
- "Maipú has more discharges than San Ramón." (Population scale, not an anomaly.)
- A finding that is fully explained by the ecological-fallacy discussion you already wrote in Tarea 3 Part 4. Pick something *new*.

If in doubt, propose more than one in the gate (see below) and we will pick.

Step 0: Anomaly proposal (Fri May 15, ungraded gate)

Submit on Canvas, one paragraph (about 150 words):

1. **What is the anomaly**, in one sentence?
2. **Where does it live**: which comuna(s), which variable(s), which figure or table cell from your Tarea 1, 2, or 3 notebooks?
3. **Why is it surprising**: what would you have predicted, and how far off is the observed value?
4. **First-pass alternative explanations** you already suspect (you are not committing to them yet).

You will hear back by **Mon May 18** with one of: `approved`, `pick another`, or `narrow this`. No grade attaches to the proposal, but the gate is required: a missed proposal blocks the rest of the project.

Component 1: GitHub repo (10 pts, async)

You already have a working repo from Tareas 0 to 3. For the final project you will extend it with a focused notebook and a polished README built around your chosen anomaly.

Required structure

```
your-repo/
  README.md
  requirements.txt              (or environment.yml)
  notebooks/
    tarea0.ipynb                (carry forward, no edits required)
    tarea1.ipynb
    tarea2.ipynb
    tarea3.ipynb
    final_anomaly.ipynb         (NEW)
  data/                         (paths or download script; do
                                 NOT commit large parquet/zip)
  figs/
    headline.png                (the figure used in your video)
    ...                         (diagnostic / supporting figures)
```

What `final_anomaly.ipynb` must contain

- A **section header** (Markdown) stating the anomaly in one paragraph, identical wording to your video opener.
- The **headline figure**: a single, well-crafted plot or table that shows the anomaly and is referenced from the video.
- The **isolation code**: the minimal data slice (filters, joins, aggregations) that produces the headline number. Do not re-run the entire pipeline; load the pre-computed master tables and slice.

- At least **two alternative-explanation checks**, each as a small code cell with a short Markdown comment on what it would have shown if true and what it actually showed.
- A **closing Markdown cell** with the three or four sentences that match the conclusion of your video.

README requirements

Anyone (including a TA who has never seen your repo) should be able to read the README and, in under two minutes, know:

- What your two or three comunas are.
- What your anomaly is.
- Which notebook produces the headline figure and approximately how long it takes to run.
- How to install dependencies and run the notebook.
- An **AI-use disclosure**: a short paragraph stating which AI tools you used during the project and for what (writing, code, figures, none). We are not penalising honest AI use; we *are* penalising undisclosed use.

Repo rubric (10 pts)

Criterion	Pts	Check
Reproducibility	3	`git clone` then follow the README, the headline figure regenerates without manual fixes
Pipeline correctness	3	Joins, offsets, units verified by spot-check against documentation; no silent imputation
`final_anomaly.ipynb` quality	2	Anomaly clearly isolated, alternatives explored in code, not just in prose
README + organization + AI disclosure	2	Two-minute read, present and complete

Component 2: Video (7 pts, async, 8 to 10 minutes)

A focused defense of one finding, not a tour of three datasets.

Suggested structure

Section	Time	Content
1. The anomaly	60 to 90 sec	One sentence. One figure. No throat-clearing.
2. Demographic context	60 to 90 sec	The minimum from Tareas 1 and 2 needed to make the anomaly legible.
3. Depth dive	3 to 5 min	Alternative explanations. Evidence for each. Model checks. Ecological fallacy framing where it applies.
4. So what	60 sec	What this suggests for public health or migration policy AND what it cannot say.
5. Limits and next steps	30 to 60 sec	What you would do with one more week.

Both partners must speak. Voice quality matters less than clarity: record in a quiet room, use a real microphone if you can, otherwise your laptop microphone is fine if you keep close to it.

Video rubric (7 pts)

Criterion	Pts	Check
Anomaly framing	2	Specific, falsifiable, grounded in numbers, not vibes
Depth of explanation	3	Alternatives are *ruled out with evidence*, not asserted away

Criterion	Pts	Check
Visual + delivery quality	2	Figures legible at 1080p, both partners speak, paced for an outside viewer

---

Component 3: Handwritten in-person memo (8 pts, INDIVIDUAL, 45 min)

This is the *only* integrity lever in the project, so it carries real weight. The session is **in person**. No laptops, no phones, no notes, no AI. You write by hand on paper provided by the staff. Each member of a pair writes their own memo; the two memos are graded independently.

On the day, the staff will pick **three** prompts from the bank below. You answer all three, in roughly 15 minutes each. Bring a pen.

Prompt bank

1. State your team's anomaly in one paragraph. Reference the specific comuna(s), variable(s), and the approximate magnitude of the effect.
2. Which notebook and which cell produces the headline number for your anomaly? Sketch the data path that gets you there: file to filter to join to aggregation.
3. Name two alternative explanations for your anomaly that you considered. For each, write the quickest check that would have falsified it, and what your check actually showed.
4. Why does (or does not) your anomaly survive an ecological fallacy critique? Use your own data.
5. If you had one more week, what is the single check you would run, and what would you expect to find?
6. A skeptical reader claims your anomaly is an artefact of anonymisation in the ENO data. Respond.
7. Pick one comuna *not* assigned to your team and predict, with reasoning, whether your anomaly should appear there too.

Memo rubric (8 pts, per student)

Criterion	Pts	Check
Specificity	3	Cites real numbers, real comunas, real code paths from memory
Critical reasoning	3	Alternatives, falsification, limits, not assertion
Coherence with team artefacts	2	What you wrote is what your repo and video say

If one partner produces a clearly thin memo while the other does not, only the thin partner loses points. The grade is individual.

AI policy for the final project

Async components (repo and video): you may use AI tools. You must **disclose them** in the README, and you remain fully accountable for every line of code and every claim. "The AI wrote it" is not a defence if a number is wrong or a claim is unsupported.

In-person component (memo): no AI, no laptops, no notes. This is the contract: the async work is allowed to lean on tools because the in-person work will check that you actually understood it. If you cannot defend your repo on paper, the repo's score will not save you.

Deliverables checklist

By **Friday, May 15** (anomaly proposal):

- ☐ One paragraph on Canvas, structured as in Step 0.

By **Friday, May 29** (async submission):

- ☐ PDF export of `final_anomaly.ipynb` on Canvas.
- ☐ Link to GitHub repo on Canvas (must include README, AI-use disclosure, `requirements.txt`, the four Tarea notebooks, and `final_anomaly.ipynb`).
- ☐ Link to the video (YouTube unlisted, Vimeo, Drive, whichever is easier; it must play without a login).

On **Friday, June 5** (in person):

- ☐ Bring a pen. We provide paper.
- ☐ Bring nothing else.

What we are looking for

Same as Tarea 3, only more so. A confident statement of an anomaly, two or three plausible alternative explanations honestly considered, and an answer that is grounded in your own numbers. We would rather see a small, modest anomaly defended carefully than a sweeping claim asserted loudly.

A perfect repo and a polished video do not by themselves earn an A. The in-person memo is 8 of 25 points and it is designed to distinguish students who genuinely worked through their data from students who outsourced the thinking.

Tips and common pitfalls

- **Don't pick something you already explained in Tarea 3 Part 4 or Part 6.** That work is graded. Pick something new.
- **Don't over-claim.** An anomaly that you *think* is real but cannot rule out as an artefact is fine to present, as long as you say so.
- **Pre-mortem the memo.** Sit across from your partner and quiz each other with the prompt bank above, on paper, with no tools open. If a prompt surprises you, your README is missing something.
- **Treat the AI disclosure as a feature, not a confession.** An honest, specific disclosure ("we used Claude to draft the README and to debug the geopandas projection") is fine. A vague "we used AI" or, worse, an undisclosed use that surfaces when a memo prompt asks where a number came from, is not.
- **Headline figure first.** Build the one figure you would defend first; everything else (notebook, video, README) is downstream of that figure.

Grading breakdown

Component	Pts
GitHub repo	10
Video	7
Handwritten memo (individual)	8

Component	Pts
Total	**25**

Typeset with: `pandoc assignments/FinalProject.md -o assignments/FinalProject.pdf --pdf-engine=pdflatex && evince assignments/FinalProject.pdf &`