Tarea 1: Demographic Profile & Migration Landscape

IELE756 – Preparación y Análisis de Datos

Leo Ferres, PhD

March 23, 2026

Tarea 1: Demographic Profile & Migration Landscape

Points: 10
Released: Thursday, March 12, 2026
Due: Thursday, March 26, 2026 (before class)
Submission: Canvas – PDF export of your notebook + link to your GitHub repo

---

Goal

Build a demographic and migration portrait of your assigned comunas using the Census 2024 microdata. In Tarea 0 you proved you could open and inspect each dataset. Now you go deeper into the Census: joining tables, computing indicators, building visualizations, and producing a **comuna-level summary table** that you will reuse in Tarea 3.

By the end of this assignment you should be comfortable with:

- Joining hierarchical tables (vivienda, hogar, persona) on their linking keys
- Computing demographic indicators (age distributions, dependency ratios, education, employment)
- Working with migration variables in the Census
- Building age pyramids and choropleth maps
- Comparing your comunas to national or regional averages

Part 0: Data Loading & Joins (2 pts)

0.1 Load the three Census tables

Load `viviendas_censo2024.parquet`, `hogares_censo2024.parquet`, and `personas_censo2024.parquet`. For each table, select only the columns you will need (see below). **Do not load every column.**

Suggested columns:

- **vivienda**: `id_vivienda`, `region`, `codigo_comuna`, `nombre_comuna`, `materialidad`
- **hogar**: `id_vivienda`, `id_hogar`, `hacinamiento`, `tenencia`
- **persona**: `id_vivienda`, `id_hogar`, `id_persona`, `sexo`, `edad`, `p27_nacionalidad_esp`, `p25_lug_nacimiento_rec`, `p24_lug_resid5`, `p25_lug_nacimiento`, `p26_llegada_periodo`, `escolaridad`, `cine11`, `sit_fuerza_trabajo`, `cod_ciuo`, `cod_caenes`, `p45_medio_transporte`

0.2 Filter to your comunas

Filter the **vivienda** table to your assigned comunas using `codigo_comuna`. Then use this filtered set to drive the joins:

```
viv = vivienda[vivienda["codigo_comuna"].isin(MY_COMUNAS)]
hog = hogar[hogar["id_vivienda"].isin(viv["id_vivienda"])]
per = persona[persona["id_vivienda"].isin(viv["id_vivienda"])]
```

Report the number of rows in each filtered table.

0.3 Join the tables

Join persona to hogar on `[id_vivienda, id_hogar]`, then join the result to vivienda on `id_vivienda`. Verify that the number of rows after joining equals the number of persona rows (a left join should not add or lose rows here).

```
df = per.merge(hog, on=["id_vivienda", "id_hogar"], how="left") \
        .merge(viv, on="id_vivienda", how="left")
assert len(df) == len(per), "Row count mismatch after join!"
```

Report `df.shape` and `df.info()`.

Part 1: Demographic Profile (3 pts)

1.1 Age pyramid by sex (1 pt)

Create a population pyramid (horizontal bar chart) for your comunas combined, with 5-year age bins (`0-4`, `5-9`, ..., `80+`). Males should extend to the left, females to the right.

Use `pd.cut()` to create age bins:

```
bins = list(range(0, 81, 5)) + [200]
labels = [f"{i}-{i+4}" for i in range(0, 80, 5)] + ["80+"]
df["age_group"] = pd.cut(df["edad"], bins=bins, labels=labels,
                         right=False)
```

Overlay: on the same pyramid (or side-by-side), distinguish **Chilean-born** vs. **foreign-born** residents using `p25_lug_nacimiento_rec`. Use different colors or hatching so the reader can see the immigrant age structure at a glance.

Remember: `-99` in `edad` means missing. Filter those out before plotting.

1.2 Dependency ratio (0.5 pts)

Compute the **age dependency ratio** for each of your comunas:

$$\text{Dependency ratio} = \frac{\text{population aged 0-14} + \text{population aged 65+}}{\text{population aged 15-64}}$$

Present the result as a small table with one row per comuna. Briefly comment: is the ratio higher or lower than you expected?

1.3 Household size distribution (0.5 pts)

Using the joined table, compute the number of persons per household (`id_hogar`). Plot the distribution as a bar chart (x-axis: household size 1, 2, 3, ..., 8+; y-axis: count or proportion).

Break it down by nationality group (Chilean vs. Foreign head of household, or all members). Comment on any visible differences.

1.4 Education and employment (1 pt)

For each of your comunas, compute:

- **Mean years of schooling** (`escolaridad`) for the population aged 25+, separately for Chilean-born and foreign-born.
- **Employment rate** among the population aged 15-64: the share with `sit_fuerza_trabajo` indicating employment (consult the variable dictionary for the code).

Present these as a grouped bar chart (one group per comuna, bars for Chilean and Foreign). Comment on any differences.

Part 2: Migration Landscape (3 pts)

2.1 Percentage foreign-born by comuna (0.5 pts)

For each of your assigned comunas, compute the percentage of foreign-born residents. Present as a simple table and as a bar chart.

2.2 Top nationalities (1 pt)

Using `p27_nacionalidad_esp` (the detailed country-of-nationality variable), show the **top 10 nationalities** among foreign-born residents in your comunas. Present as a horizontal bar chart.

Consult the variable dictionary to decode the numeric codes into country names. If the dictionary uses numeric codes, create a mapping dictionary in your notebook.

2.3 Migration status: residence 5 years ago (1 pt)

The variable `p24_lug_resid5` captures where the person lived 5 years before the census. Using the variable dictionary, decode this variable and compute the following for your comunas combined:

- % who lived in the **same comuna**
- % who lived in a **different comuna, same region**
- % who lived in a **different region**
- % who lived **abroad**

Present as a stacked bar chart broken down by comuna. Filter out missing values and people younger than 5 (who were not alive 5 years ago).

2.4 Arrival period of immigrants (0.5 pts)

Among foreign-born residents, use `p26_llegada_periodo` to show **when they arrived** in Chile. Plot the distribution as a bar chart. Comment: is immigration to your comunas a recent phenomenon or does it have a longer history?

Part 3: Spatial Visualization (1 pt)

3.1 Choropleth map: population by comuna

Using `geopandas` and a shapefile of Chilean comunas, create a choropleth map showing total population for each of your assigned comunas.

```
import geopandas as gpd
```

```
comunas_gdf = gpd.read_file("materials/shapefiles/comunas.shp")
# merge your comuna-level data onto the geodataframe
```

If the shapefile path differs, check the `materials/` folder for available shapefiles.

[[Pasted image 20260320115642.png|452]] ## 3.2 Choropleth map: % foreign-born by comuna

Create a second choropleth map showing the **percentage of foreign-born residents** by comuna. Use a sequential color palette (e.g., `YlOrRd` or `Blues`).

Both maps should include a legend, a title, and comuna labels if legible.

Part 4: Comuna-Level Summary Table (1 pt)

This is the most important output of Tarea 1. Build a summary table at the **comuna level** with the following columns, computed separately for **Chilean-born** and **foreign-born** where indicated:

Column	Description
`codigo_comuna`	Numeric comuna code
`nombre_comuna`	Comuna name
`pop_total`	Total population
`pop_chilean`	Chilean-born population
`pop_foreign`	Foreign-born population
`pct_foreign`	% foreign-born
`median_age_chilean`	Median age, Chilean-born
`median_age_foreign`	Median age, foreign-born
`mean_schooling_chilean`	Mean years of schooling (age 25+), Chilean-born
`mean_schooling_foreign`	Mean years of schooling (age 25+), foreign-born
`emp_rate_chilean`	Employment rate (age 15-64), Chilean-born
`emp_rate_foreign`	Employment rate (age 15-64), foreign-born
`dependency_ratio`	Age dependency ratio (overall)

```
summary = df.groupby("codigo_comuna").apply(build_summary)
summary.to_csv("output/tarea1_comuna_summary.csv", index=False)
```

You will define the `build_summary` function yourself. Display the resulting table in your notebook. Also save it as a CSV file in your repository (you will need it in Tarea 3).

Tip: handle `-99` (missing) values by excluding them from calculations, not by treating them as valid numbers.

Deliverables

Submit on **Canvas** before class on Thursday, March 26:

1. **PDF export** of your Colab notebook (File > Print > Save as PDF, or File > Download > Download .pdf).
2. **Link to your GitHub repository** (the notebook and the summary CSV should be committed).

Your notebook must include:

- Markdown cells explaining each step and interpreting results
- All code cells executed with visible output
- The following visualizations:
  - Age pyramid with Chilean/Foreign overlay (Part 1.1)
  - Household size distribution (Part 1.3)
  - Education/employment grouped bar chart (Part 1.4)
  - Top nationalities bar chart (Part 2.2)
  - Migration status stacked bar chart (Part 2.3)
  - Arrival period bar chart (Part 2.4)
  - Two choropleth maps (Part 3)
- The comuna-level summary table (Part 4), both displayed and saved as CSV

Tips and Common Pitfalls

- **Missing values**: the Census uses `-99` to encode missing data. Always filter these out before computing means, medians, or rates. Do not use `dropna()` blindly; check whether missingness is coded as `-99` or as actual `NaN`.
- **Memory**: if your machine struggles with the full national tables, filter to your comunas as early as possible (ideally at load time using pyarrow filters).
- **Variable dictionary**: keep `diccionario_variables_censo2024.xlsx` open while you work. Many variables are numeric codes that need decoding.
- **Consistent born-in-Chile grouping**: throughout this assignment, use `p25_lug_nacimiento_rec` to split Chilean-born vs. foreign-born. Use `p27_nacionalidad_esp` only when you need detailed nationalities (Part 2.2).
- **Age filters**: some indicators only make sense for specific age groups (schooling for 25+, employment for 15-64). Always state your filter explicitly.
- **Ecological note**: everything in this assignment is at the individual or comuna level within a single dataset (Census). Cross-dataset ecological analysis comes in Tarea 3.

Grading Breakdown

Part	Points
Part 0: Data loading & joins	2
Part 1: Demographic profile	3
Part 2: Migration landscape	3
Part 3: Spatial visualization	1
Part 4: Comuna-level summary table	1
Total	**10**

Half of each part's score comes from correct code and output; the other half comes from clear Markdown explanations and thoughtful interpretation of your results.

Typeset with: `pandoc assignments/Tarea1.md -o assignments/Tarea1.pdf --pdf-engine=pdflatex && evince assignments/Tarea1.pdf &`