Codebook: Thai Provinces Dataset (`thai_provinces.csv`)

Course: POLS 641 — Introductory Political Analysis **Unit of analysis:** Province (changwat), Kingdom of Thailand **Observations:** 77 (76 provinces plus the Bangkok Metropolitan Administration) **Structure:** Cross-sectional, one row per province **Built by:** `build_thai_province_dataset.R` (replication script distributed with this file)

---

1. Data Sources

Source A — Civil registration statistics. Bureau of Registration Administration (BORA), Department of Provincial Administration (DOPA), Ministry of Interior. Official registered population and house counts as of December, downloaded from the BORA statistical service (stat.bora.dopa.go.th). Years used: 2020 (BE 2563) and 2025 (BE 2568).

Source B — Provincial economic accounts. Office of the National Economic and Social Development Council (NESDC), *Gross Regional and Provincial Product*, 2023 preliminary edition ("2023p"). Per capita gross provincial product at current market prices, from the "PER CAPITA" statistical table (nesdc.go.th).

The two sources are merged using a province code-to-name crosswalk included in the replication script. DOPA identifies provinces with standard two-digit codes; NESDC tables use English province names, which were harmonized before merging (see Section 4).

---

2. Variables

Variable	Description	Units / Values	Source
`cc_code`	Standard two-digit province code (matches ISO 3166-2:TH numbering). Use this as the key when merging additional data.	"10"–"96" (text; keep leading zeros)	DOPA
`prov_th`	Province name in Thai	Text	DOPA
`prov_en`	Province name in English (harmonized spelling)	Text	Crosswalk
`region`	Region classification used in the	Bangkok and Vicinity, Central,	Crosswalk

Variable	Description	Units / Values	Source
	crosswalk	East, North, Northeast, South, West	
`northeast`	Indicator: province is in the Northeast (Isan)	1 = yes, 0 = no	Constructed
`pop`	Registered population, December 2025	Persons	DOPA
`sex_ratio`	Registered males per 100 registered females, 2025	Ratio × 100	Constructed from DOPA
`hh_size`	Registered persons per registered house, 2025. A proxy for household size (see limitations).	Persons per house	Constructed from DOPA
`pop_growth_pct`	Percentage change in registered population, 2020–2025	Percent	Constructed from DOPA
`gpp_pc`	Gross provincial product per capita at current market prices, 2023 (preliminary)	Baht per person per year	NESDC
`log_gpp_pc`	Natural log of `gpp_pc`	log(baht)	Constructed

For orientation: `gpp_pc` ranges from roughly 64,000 baht (Narathiwat) to roughly 942,000 baht (Rayong), with Bangkok at about 676,000 baht.

---

3. Variable Construction Notes

- `sex_ratio` = 100 × (registered males / registered females), December 2025.
- `hh_size` = total registered persons / registered houses. DOPA counts *houses* (registered dwellings), not households in the survey sense; a dwelling can be empty or contain multiple family units.
- `pop_growth_pct` = 100 × (pop 2025 / pop 2020 − 1).

- `northeast` = 1 if `region` == "Northeast". The 20 Isan provinces are historically Thailand's lowest-income region, which makes this a substantively meaningful control.
- `log_gpp_pc`: GPP per capita is heavily right-skewed (Bangkok and the Eastern Seaboard industrial provinces). Examine both raw and logged versions before choosing your specification, and justify the choice in your Methodology section.

---

4. Known Issues and Limitations (read before you write your paper)

1. **Registered vs. resident population.** DOPA figures count people where they are *registered*, not where they live. Internal migrants often keep their home-province registration while working in Bangkok, the Eastern Seaboard, or tourist provinces. Registered population therefore undercounts actual residents in migrant-receiving provinces and overcounts them in sending provinces. This affects `pop`, `sex_ratio`, `hh_size`, and especially `pop_growth_pct`.
2. **Denominator mismatch.** NESDC computes `gpp_pc` using its own population estimates, not DOPA registration counts. Do not recompute per capita figures by dividing NESDC totals by DOPA population.
3. **Preliminary economic data.** The 2023 GPP figures are marked "p" (preliminary) and are subject to revision in later NESDC editions.
4. **Timing mismatch.** Demographic variables are measured in 2025 (and growth over 2020–2025), while GPP is 2023. Think about what this implies for any causal story you tell.
5. **GPP measures production, not welfare.** Output is attributed to the province where it is produced. Rayong's petrochemical complexes generate enormous GPP per capita, but much of that income accrues to firms and workers elsewhere. GPP per capita is not household income.
6. **Outliers.** Bangkok and Rayong dominate the GPP distribution. Report how your results change when they are excluded (a robustness check belongs in your Results or Discussion section).
7. **Name spellings.** Government sources romanize province names inconsistently (the 2023 NESDC table itself contains spelling errors, e.g., "Amnat Chareon"). Merge on `cc_code`, never on names, when adding new variables.

---

5. Citing the Data

- Department of Provincial Administration. (2025). *Official statistics of registration systems: Population and house statistics* [Data files]. Bureau of Registration Administration, Ministry of Interior. https://stat.bora.dopa.go.th
- Office of the National Economic and Social Development Council. (2025). *Gross regional and provincial product, 2023 edition* [Statistical tables]. https://www.nesdc.go.th