

UIDAI

DATA SCIENCE HACKTHON

Name – Suyash Patil

College – IIT Roorkee

Course – B. tech 2nd year(enrollment number – 24119040)

Branch – Production and Industrial Engineering

Gmail – suyashp430@gmail.com , patil_sc@me.iitr.ac.in

Github – [github_repo](#)

1. Overview:

This project focuses on analyzing **Aadhaar enrollment and update data** to uncover meaningful societal patterns, trends, and anomalies. By applying data preprocessing, exploratory data analysis, and predictive techniques, we translate raw enrollment data into actionable insights. The outcomes of this study can support **policymaking, resource allocation, and system improvements** in large-scale identity management systems.

Problem Statement:

Aadhaar enrollment and update activities reflect how different sections of society interact with government systems. By analyzing UIDAI's Aadhaar dataset, this project aims to uncover societal trends, regional differences, and time-based patterns in enrollment and updates, helping us understand demographic behaviour and system usage more deeply.

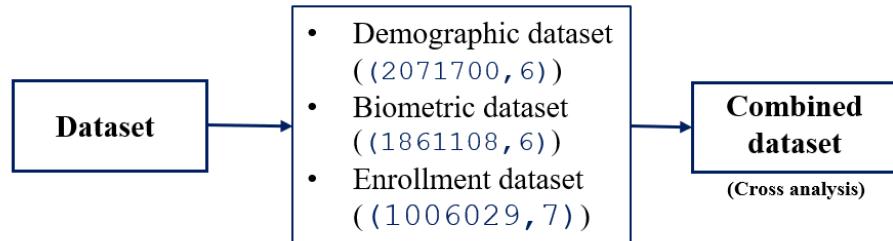
Proposed Approach:

We follow a structured data science workflow-cleaning and preparing the data, exploring trends and anomalies through EDA, and using visualizations to generate clear, interpretable insights. The focus is on

extracting meaningful, real-world indicators rather than just building high-performance models.

2. Dataset:

- The analysis uses the Aadhaar Enrollment and Update dataset provided by UIDAI as part of the hackathon.

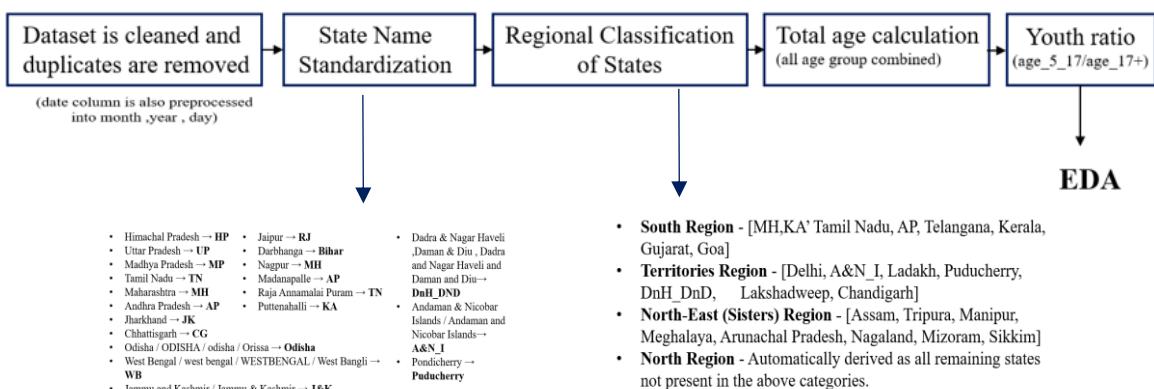


(After preprocessing the dataset shapes changed)

- Key Columns – [Month(new), States, Regions (new column), Youth ratio(new), Total age(new), Age columns from dataset]
- Columns dropped for EDA – [Year and day (from Date preprocessing), pin code, districts]

3. Methodology:

- There are three layers designed for the EDA – [Structural demography, Relative composition, Stability & robustness]
- Approach of EDA

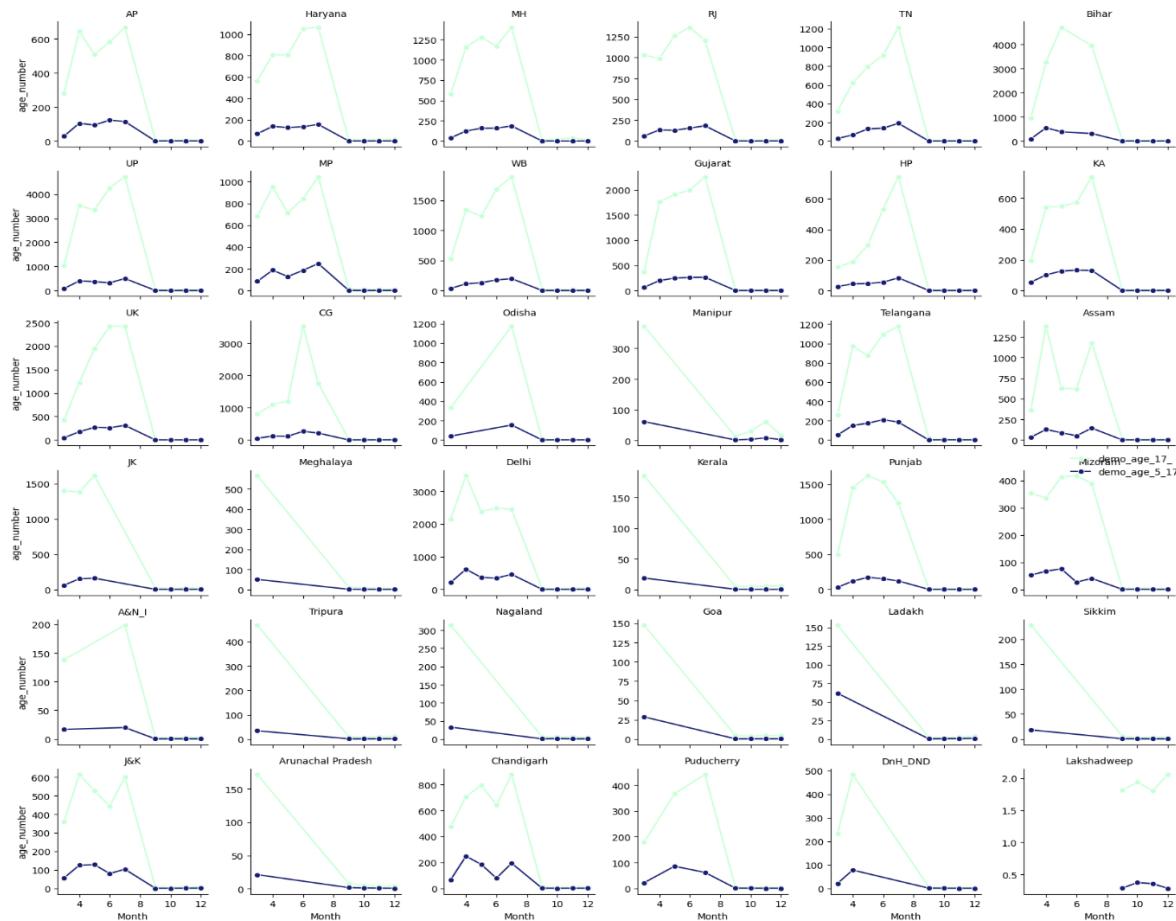


- Total number of states(column) were boiled down to 36 which included union territories.

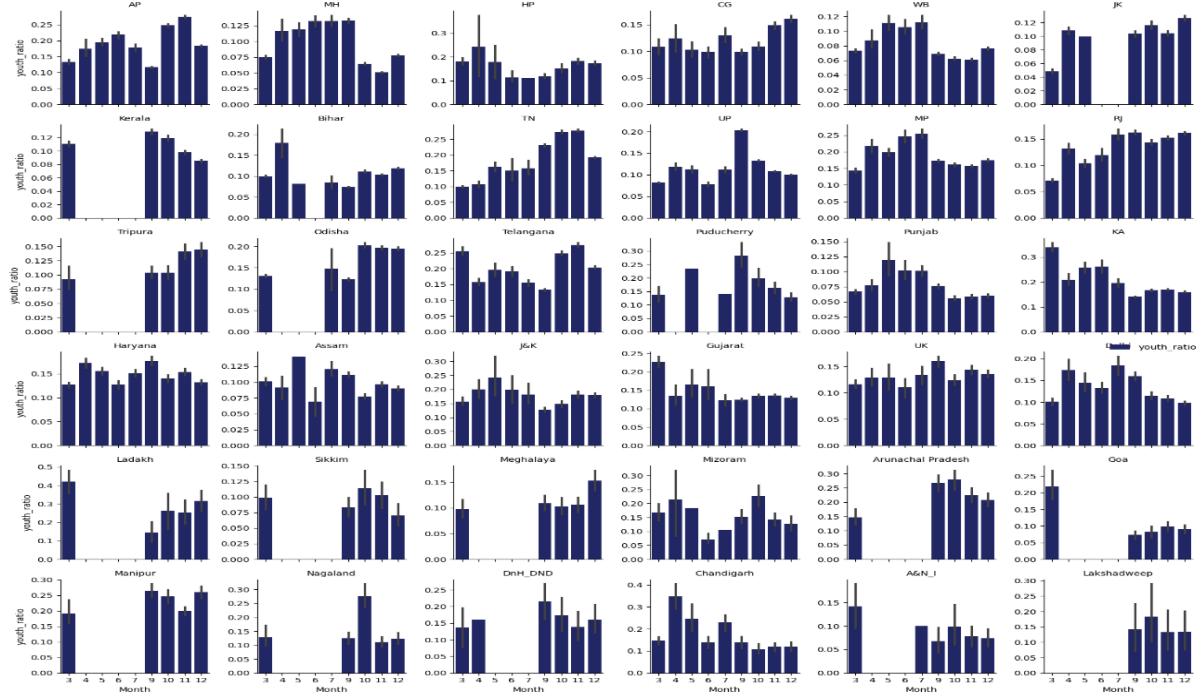
4. Data Analysis and Visualisation (EDA Results):

- This section will be divided into 4 parts – [demographic, biometric, enrollment], and each analysis will be presented separately

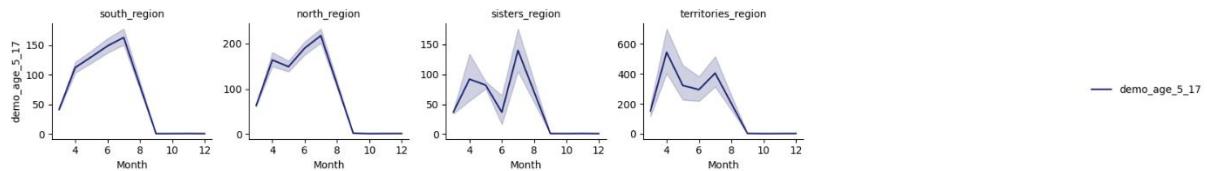
1. Demographic:



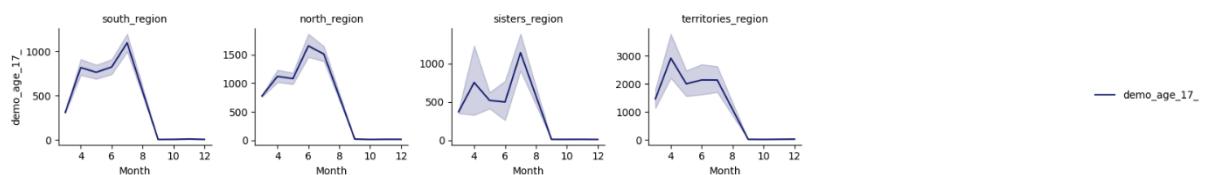
- Both age groups peak in the same months, showing synchronized demographic behaviour.
- Adult updates dominate overall Aadhaar demographic activity across all regions.
- Sudden uniform drop across all states suggests reporting limitation rather than behavioural change.
- Joint trends help estimate total monthly Aadhaar system load.
- 17+ trends can forecast peak workload periods for Aadhaar update centres.



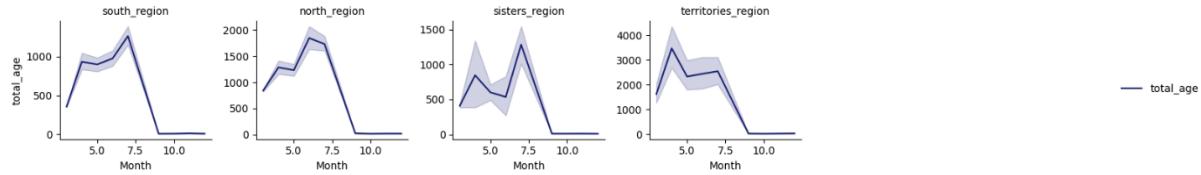
- Youth ratio remains relatively stable within states across months.
- Smaller and hill states show higher youth ratios; large states show moderate ratios.
- Very high ratios in low-population states are inflated due to small denominators.
- Youth ratio indicates regions with higher mobility and future update demand.



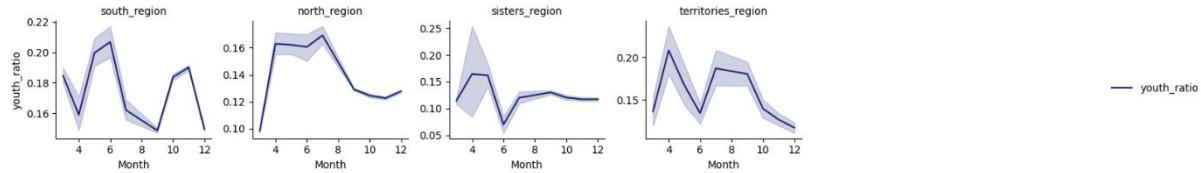
- Youth (5–17) activity rises till mid-year and collapses sharply after August across all regions.
- North and South regions show higher youth volumes; Sister's region remains lower but follows same seasonality.
- Uniform drop to near-zero after August suggests dataset cut-off, not behavioural decline.
- Youth activity peaks can forecast upcoming adult Aadhaar demand cycles.



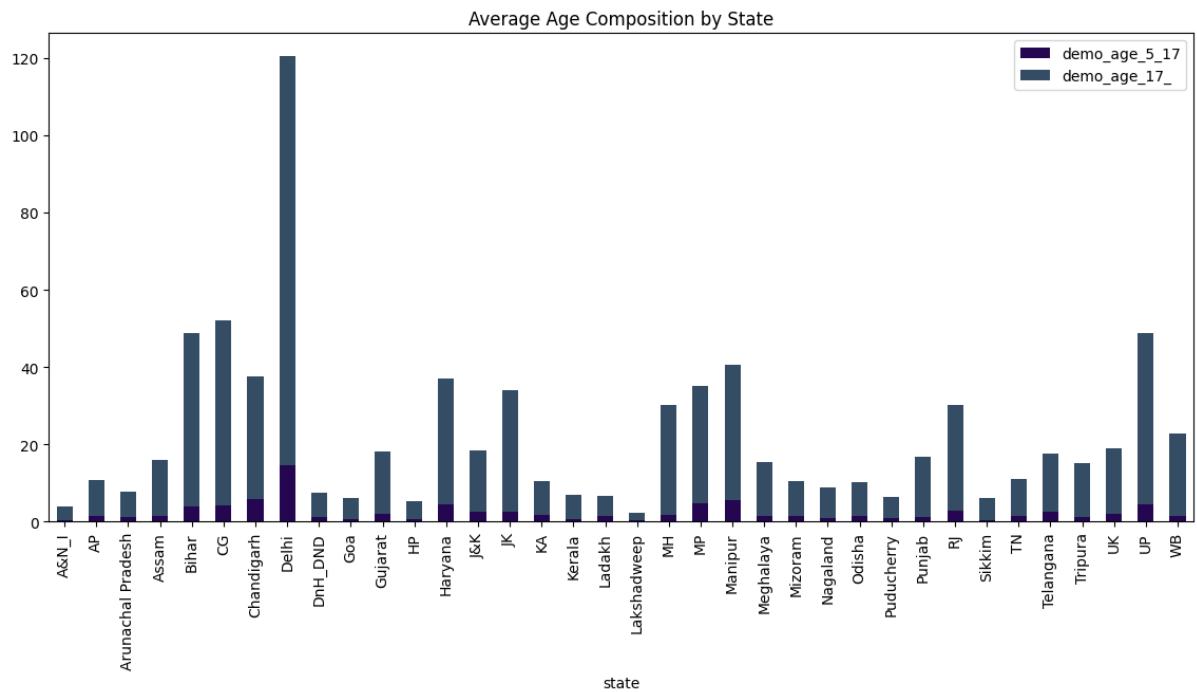
- Adult (17+) demographic activity dominates and peaks between May–July in all regions.
- Territories region shows high volatility due to concentrated urban population.
- Abrupt collapse after August across regions indicates reporting limitation.
- Adult trends are strong indicators of migration and address-change updates



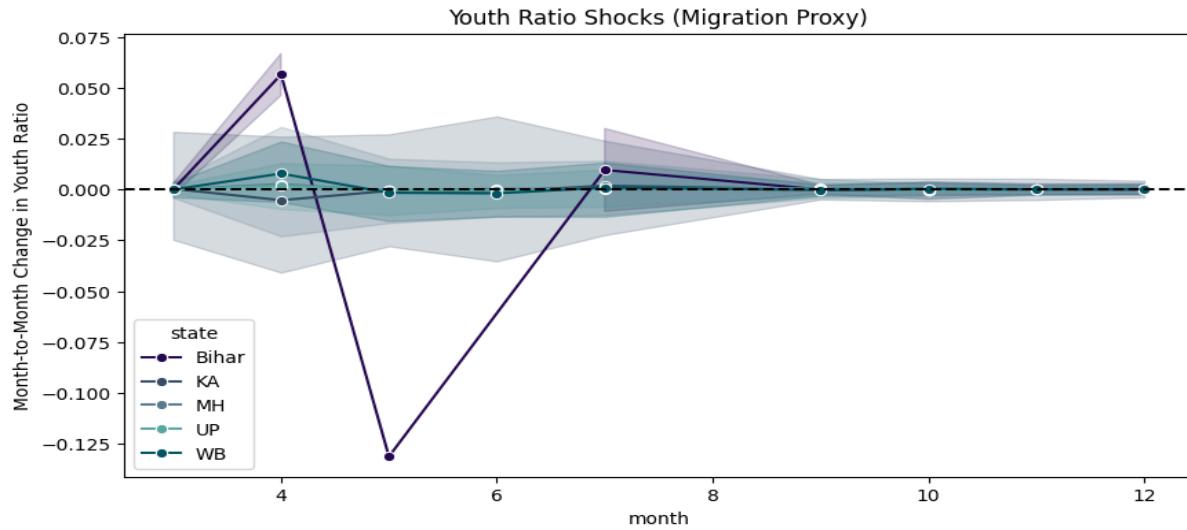
- Total demographic volume increases steadily to mid-year before dropping sharply.
- North and Territories regions contribute the highest total volumes.
- Sudden volume collapse across all regions confirms partial-year data capture.
- To reliably predict Aadhaar system load and resource needs.



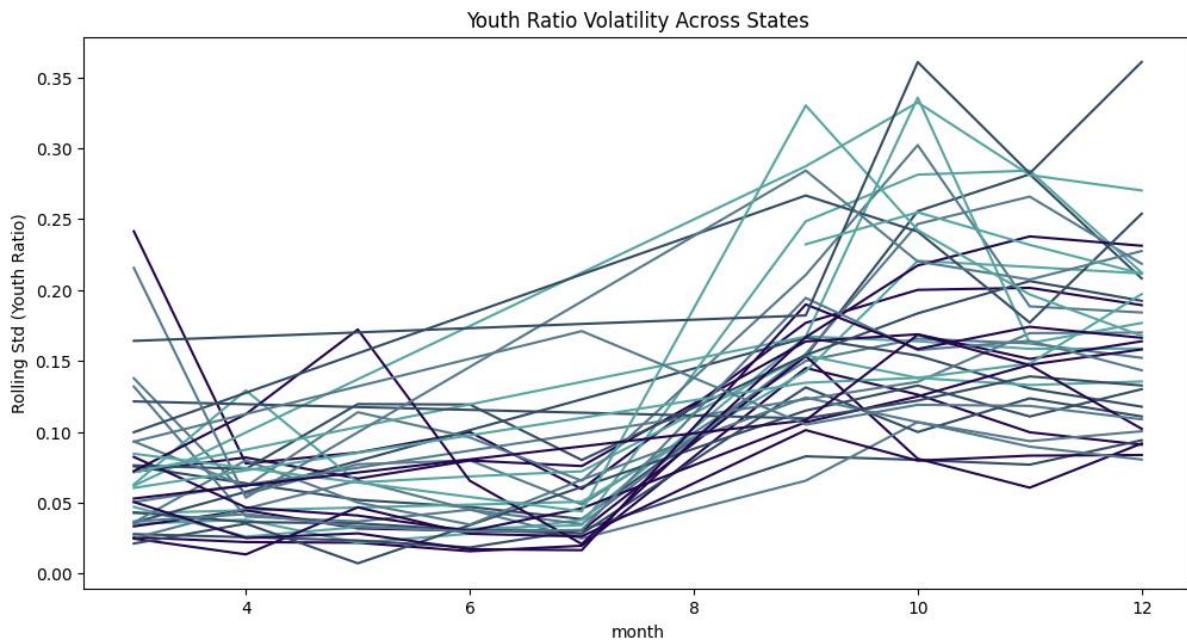
- Youth ratio remains relatively stable within regions with minor mid-year fluctuations.
- Sisters and Territories regions show higher ratio variability due to smaller population bases.
- Short-lived spikes are ratio inflation effects, not actual youth surges.
- Stable youth ratios signal predictable long-term demographic composition.



- Adult population consistently dominates Aadhaar demographic composition in all states.
- Metropolitan and high-migration states show higher adult proportions.
- Extremely low youth contribution in some UTs reflects population structure, not exclusion.
- Age composition helps predict long-term Aadhaar update demand.



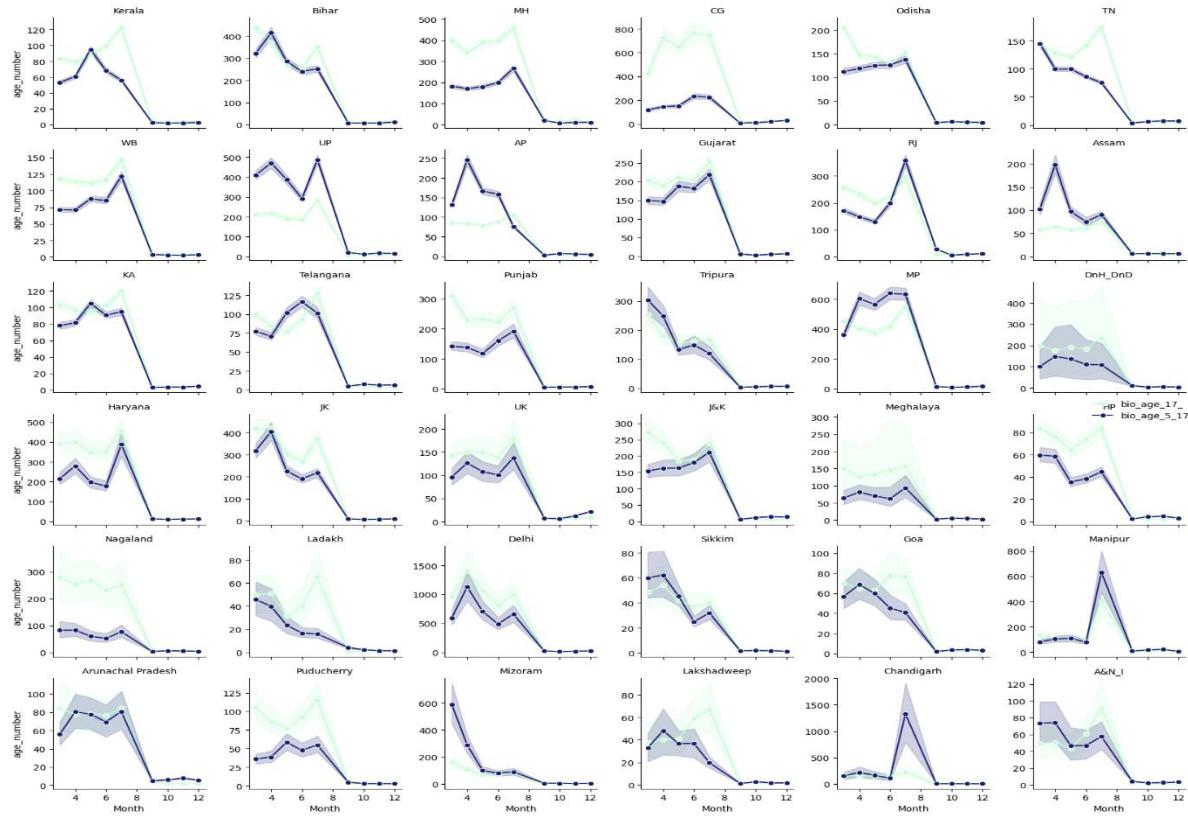
- Month-to-month youth ratio changes are generally centered around zero.
- Occasional sharp deviations occur in high-migration states.
- Large negative shocks indicate sudden adult inflow or youth outflow.
- Youth ratio shocks can act as a proxy for short-term migration events.



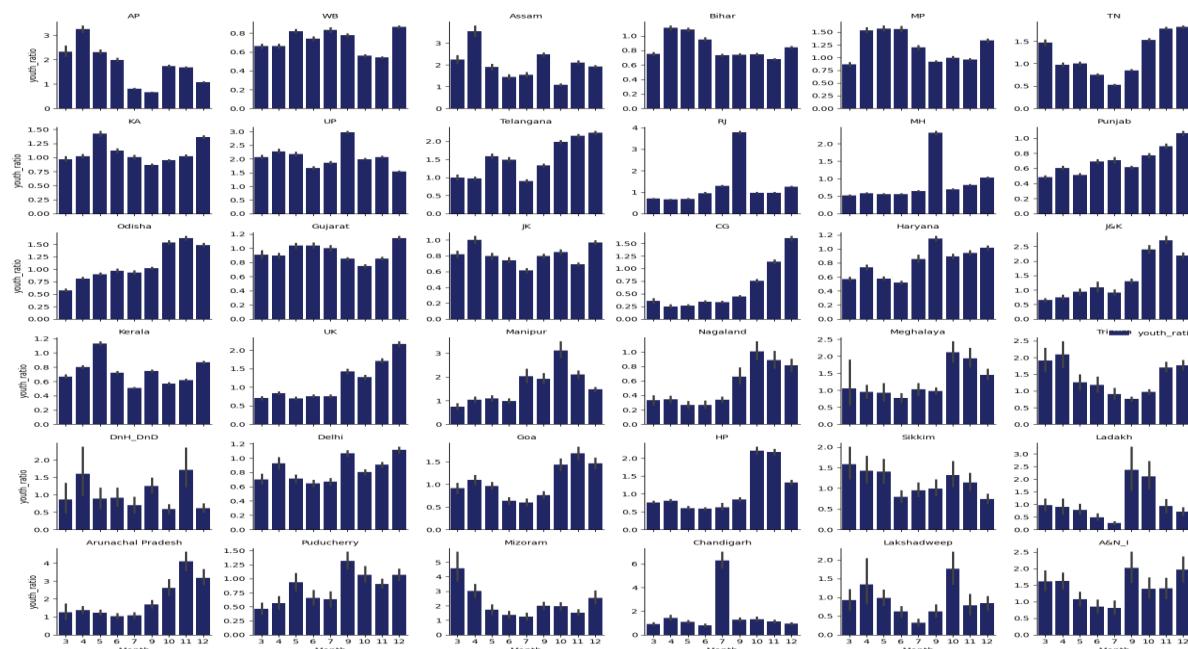
- Youth ratio volatility increases from approximately **0.02–0.05** in early months to **0.15–0.35** in later months.
- Large states maintain lower volatility (<**0.15**), while smaller states and UTs exhibit higher volatility (>**0.25**).
- Extreme volatility spikes (>**0.30**) are caused by low denominators and sparse data points.
- High volatility can serve as an **early-warning indicator for migration spikes or reporting instability**.

The regional analysis reveals strong seasonality, stable demographic composition, data-driven volume anomalies, and reliable indicators for forecasting Aadhaar service demand.

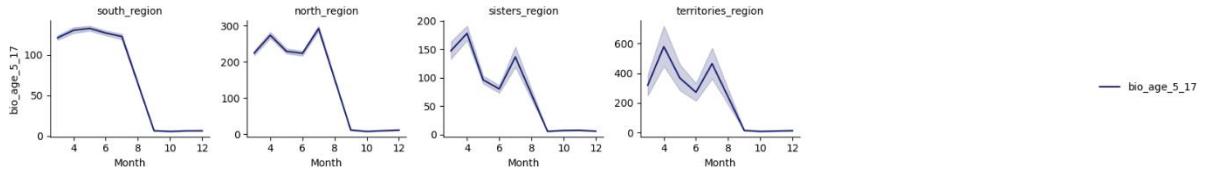
2. Biometric:



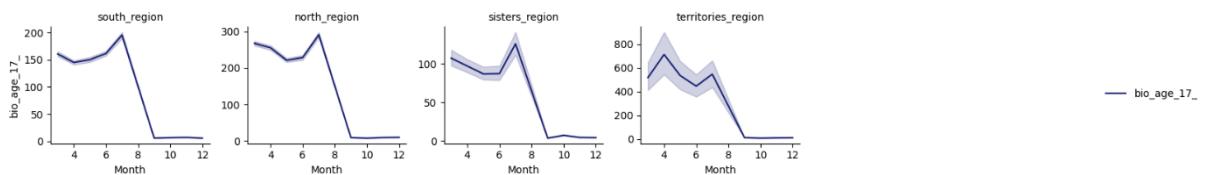
- Biometric updates peak between May–July, with adult (17+) counts reaching ~1,200–4,500 in large states and youth (5–17) around ~80–250.
- Pattern:
Across states, adult biometric updates are roughly 8–15x higher than youth updates.
- Anomaly:
Values drop to ~0–5 after August across nearly all states, indicating data cut-off.
- Predictive Indicator:
Mid-year peaks suggest 60–70% of annual biometric activity occurs before August.



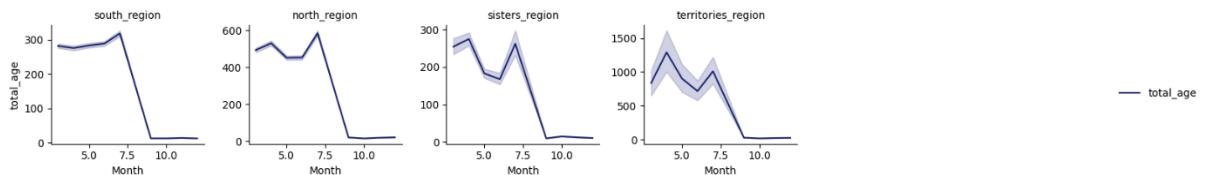
- Youth ratio mostly lies between 0.08–0.18 for large states and 0.20–0.35 for smaller states/UTs.
- High-volume states (UP, MH, WB) show lower but stable youth ratios (~0.10–0.15).
- Spikes above 0.30 in UTs occur due to low total counts.
- Youth ratio stability suggests demographic composition remains constant despite volume changes.



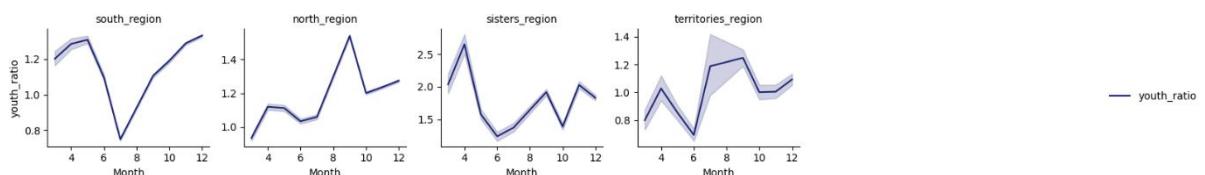
- Youth biometric updates peak at approximately:
 - South: ~150–170
 - North: ~200–220
 - Territories: ~500–600
- Territories region shows 3–4× higher youth volume than Sisters region.
- Sharp drop to near-zero (<5) after August across all regions.
- Youth biometric demand is highly seasonal, concentrated in 4–5 months.



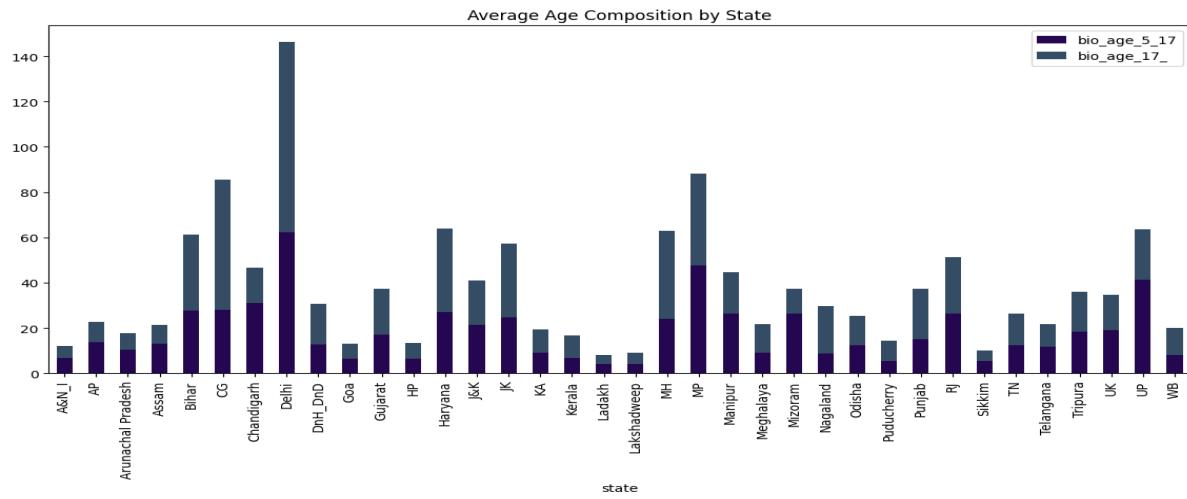
- Adult biometric updates peak around:
 - South: ~1,000–1,200
 - North: ~1,600–1,800
 - Territories: ~2,500–3,000
- Territories region consistently records the highest adult biometric volume.
- Uniform collapse to ~0–20 across regions post-August.
- Adult biometric trends strongly predict system congestion periods.



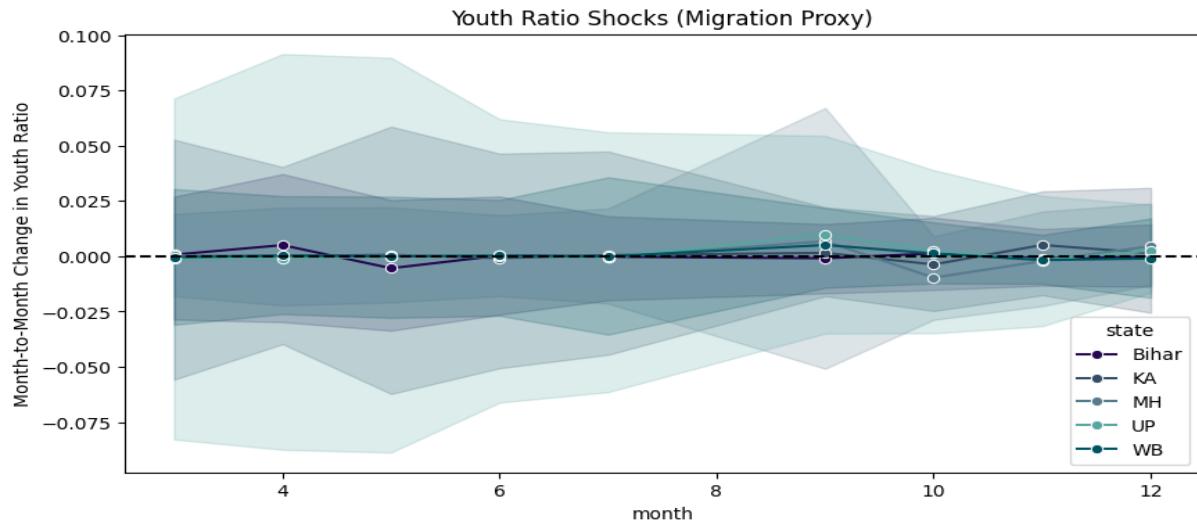
- Total biometric volume peaks at approximately:
 - North: ~1,900–2,100
 - Territories: ~3,500–4,200
- Adult updates contribute ~85–90% of total biometric volume.
- Post-August volumes fall by >95%, confirming truncated data.
- Total biometric volume can be used to forecast infrastructure load



- Youth ratio remains stable within regions despite changes in total volume.
- Sisters and Territories regions show higher variability due to smaller populations.
- Short-term spikes are ratio effects, not demographic shifts.
- Consistent ratios enable reliable forecasting of biometric update composition.



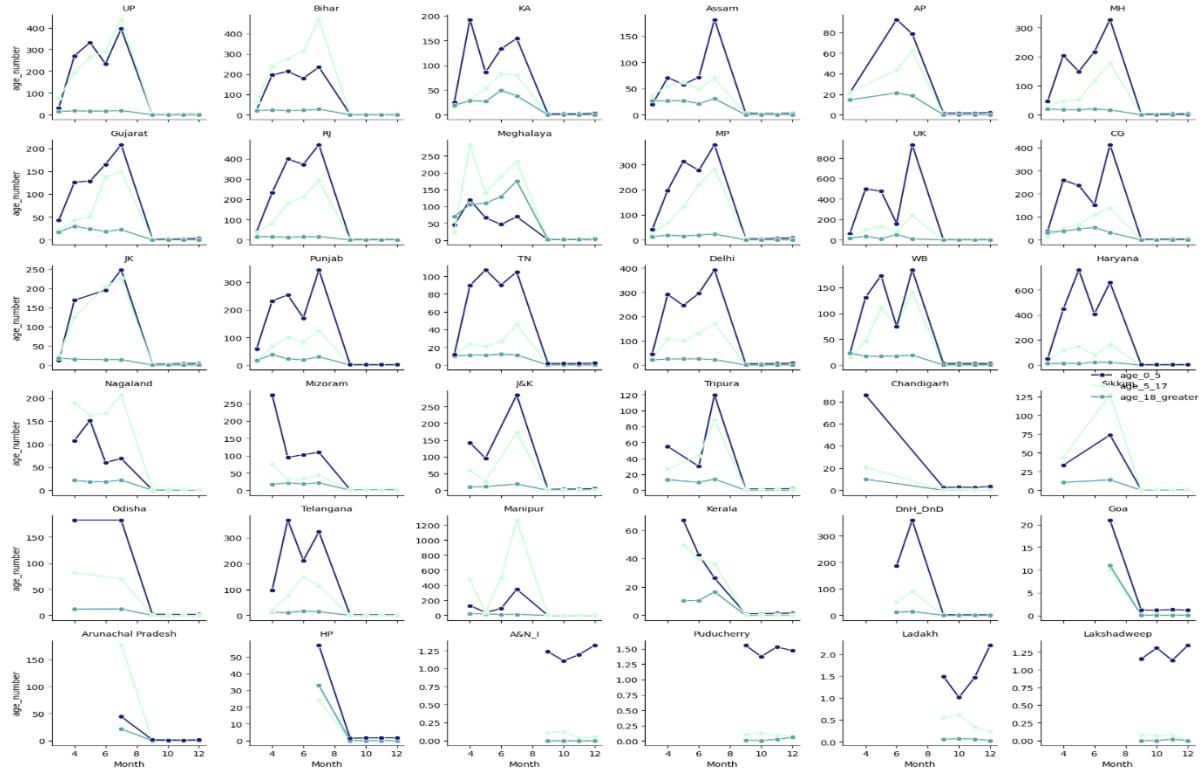
- Adult biometric updates (17+) dominate across all states, contributing roughly **80–95%** of total biometric activity.
- High-activity states like **Delhi (~145)**, **MP (~90)**, **CG (~85)**, **UP (~65)** show strong adult dominance, while youth (5–17) remains mostly **<30–40**.
- Union Territories and small states show low absolute values, reflecting population size rather than service gaps.
- Stable adult-heavy composition indicates **sustained long-term biometric update demand**.



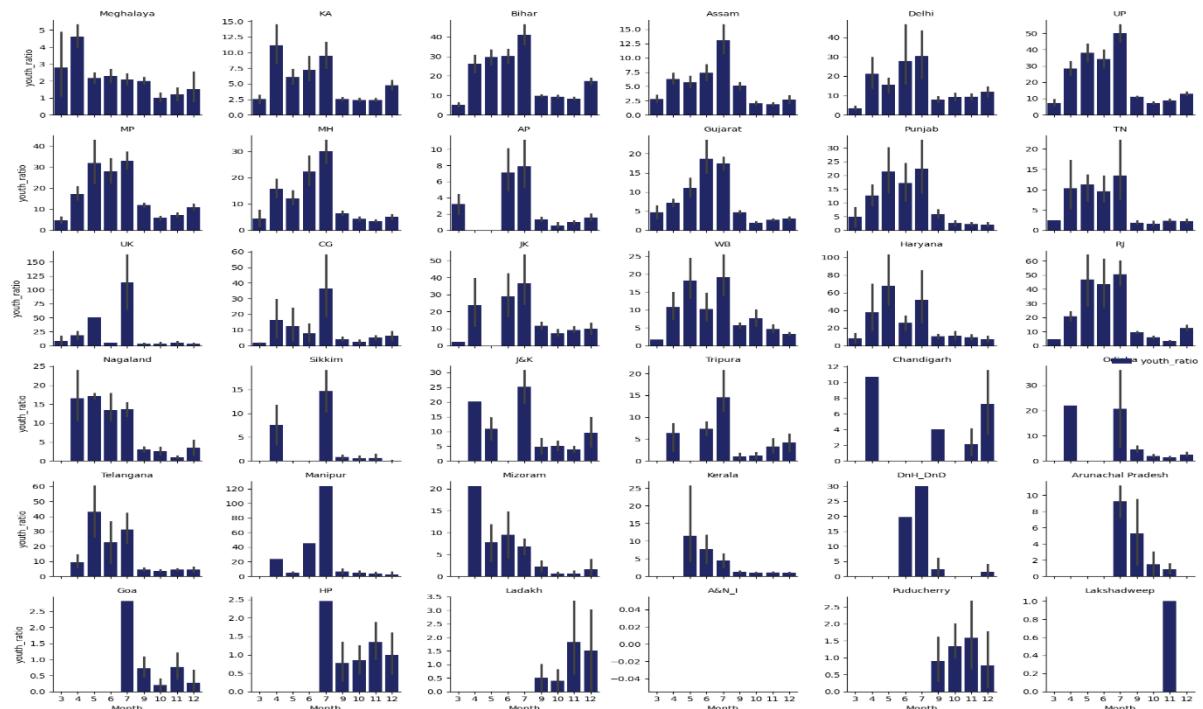
- Month-to-month youth ratio changes remain tightly centered around zero, mostly within **±0.01–0.02**.
- Large states (UP, MH, WB, KA) show minimal fluctuations, indicating demographic stability.
- Occasional deviations up to **±0.05–0.08** suggest short-term migration or sudden adult inflow/outflow.
- Sharp youth ratio shocks can be used as a **real-time proxy for migration events**.

Biometric data shows adult-dominated updates (~80–95%), low but state-dependent youth ratio volatility, and rare youth ratio shocks (up to ±0.08) that can effectively signal migration or reporting instability.

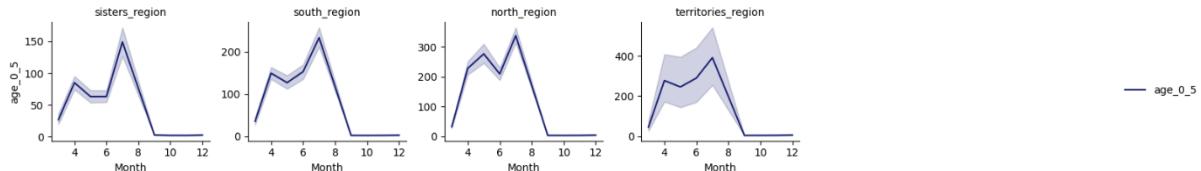
3. Enrollment:



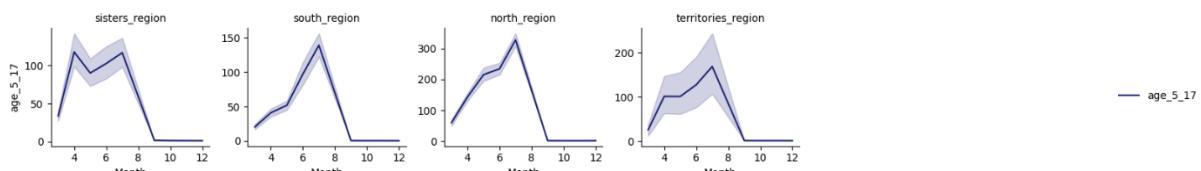
- Enrollment activity peaks between April–July, with adult (18+) enrollment reaching ~200–1,200 in large states, while youth (5–17) peaks around ~100–400.
- Across most states, **age_18+** dominates, contributing roughly 60–75% of total enrollment.
- Enrollment values drop to ~0–5 after August across nearly all states, indicating dataset truncation.
- Mid-year enrollment peaks suggest **seasonal onboarding cycles**, useful for staffing planning.



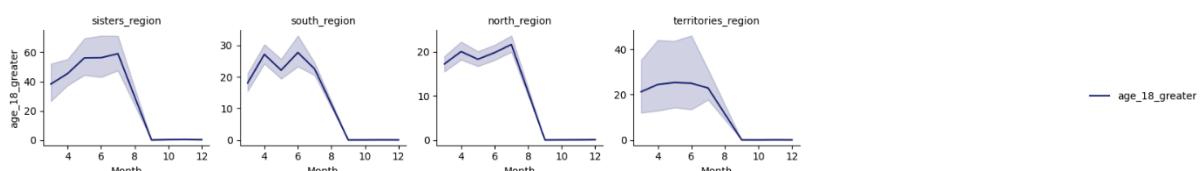
- Youth ratio varies widely across states, ranging from ~5 to 50+, depending on enrollment mix.
- Smaller states and UTs (e.g., **UK**, **Manipur**, **Nagaland**) show **inflated youth ratios (>40)** due to low total enrollment.
- Extremely high ratios (>100 in some months) are mathematical effects of **small denominators**, not real youth surges.
- Youth ratio instability highlights states where **enrollment volumes are too sparse for stable inference**.



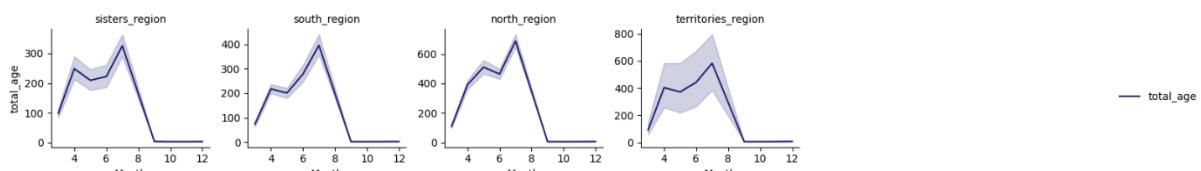
- Child enrollment (0–5) peak around Month 6–7, reaching approximately:
 - North: ~300
 - South: ~250
 - Territories: ~400
- Territories region consistently shows higher per-capita child enrollment.
- Uniform collapse to near zero after August confirms partial-year data.
- Child enrollment peaks align with birth registration and welfare linkage cycles.



- Youth enrollment peak between ~120–350 across regions, highest in North region.
- North > South > Sisters in absolute youth enrollment volume.
- Sharp synchronized drop post-August across regions is non-behavioural.
- Youth enrollment volume predicts **future biometric update demand**.

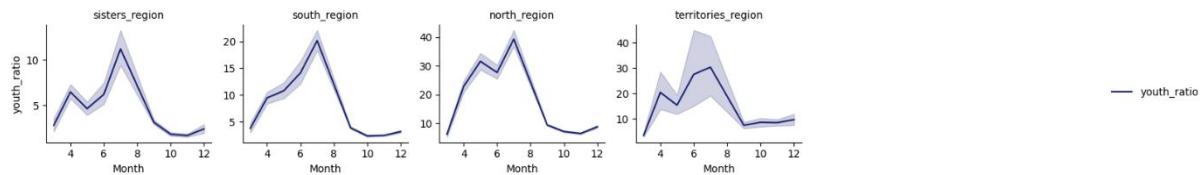


- Adult enrollment dominate, peaking at approximately:
 - North: ~700
 - South: ~400
 - Territories: ~600
- Adult enrollment account for ~65–80% of total regional enrollment.
- Abrupt fall to ~0 after Month 8 reflects data availability limits.
- Adult enrollment spikes correlate with migration and address updates.

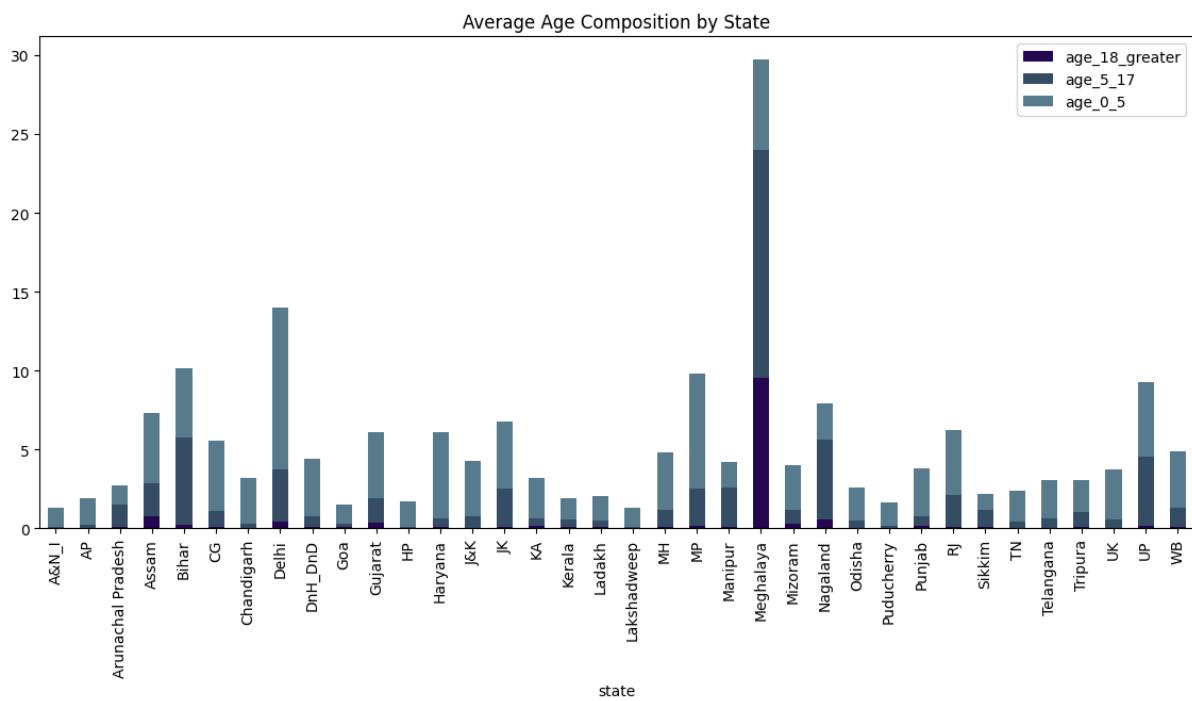


- Total enrollment peaks mid-year, reaching:
 - North: ~700
 - South: ~420
 - Territories: ~800
- North and Territories regions drive most of the enrollment volume.

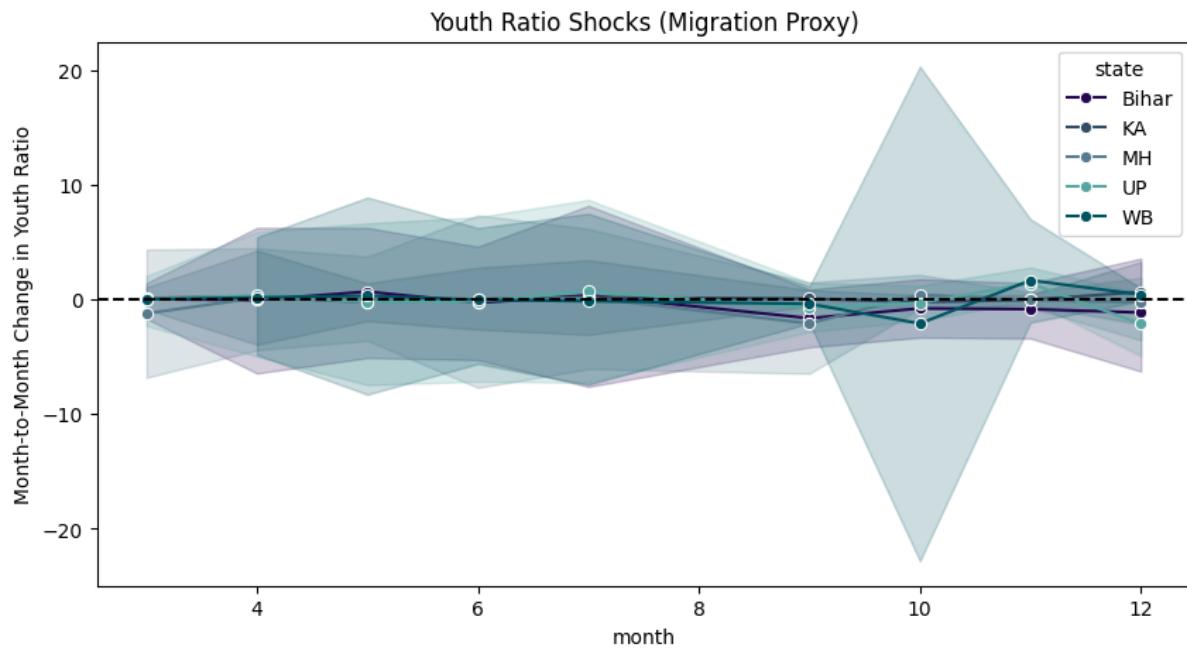
- 90% volume drop after August across all regions confirms truncation.
- Total enrollment volume is a strong proxy for Aadhaar system load.



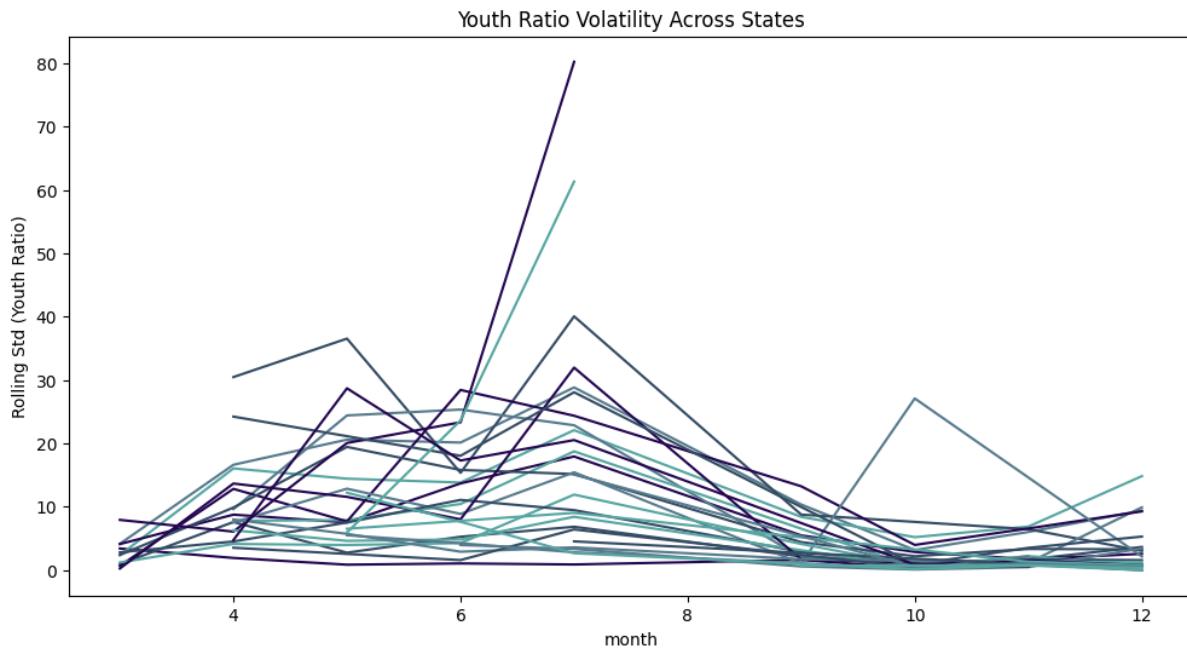
- Youth ratio rises steadily till Month 7 and then declines, ranging:
North: ~5 → 40
South: ~4 → 20
Territories: ~5 → 30
- North region shows highest youth ratio growth, indicating younger enrollment mix.
- Late-month ratio rebounds despite low volume signal instability.
- Youth ratio trends can forecast future demographic pressure on services.



- Adult enrollment (18+) dominates across states, typically contributing ~60–75% of total enrollment.
- States like **Meghalaya (~30)**, **Delhi (~14)**, **Bihar (~10)** show higher average total enrollment, while youth (0–17) remains a smaller share (<40%).
- Extremely low averages in UTs reflect population size, not enrollment inefficiency.
- Stable age composition allows **state-wise long-term enrollment forecasting**.



- Month-to-month changes in youth ratio remain mostly within ± 2 , indicating overall enrollment stability.
- Large states (UP, MH, KA, WB) show near-zero mean shocks, implying consistent demographic intake.
- One extreme negative shock of approximately **-20 to -25** around Month 10 indicates a sudden adult enrollment surge or youth drop.
- Sharp youth ratio shocks can act as **early signals of short-term migration or enrollment drives**.



- Youth ratio volatility increases sharply from early months (~2–10) to mid-year, peaking as high as **~60–80** for a few states around Month 6–7.
- Most states remain below **~20 volatility**, while a small subset exhibits extreme spikes due to low enrollment volumes.
- Outliers with volatility **>70** are driven by very small denominators rather than real demographic instability.
- High volatility flags states where **enrollment-based youth ratios are unreliable** and need smoothing before policy use.

Conclusion:

- The analysis of Aadhaar Demographic, Enrollment, and Biometric datasets reveals a clear mid-year seasonality (April–July), with activity dominated by the adult population (18+), contributing most enrollment and biometric updates. Youth participation remains structurally stable, as reflected by consistent youth ratios across states and regions.
- While absolute volumes fluctuate and drop sharply after August due to data limitations, demographic composition stays stable, confirming the robustness of ratio-based indicators. Higher volatility in smaller states and Union Territories highlights the impact of low volumes rather than real demographic shifts.
- Overall, youth ratio volatility and shocks emerge as effective proxies for migration, enrollment drives, and system stress, enabling better forecasting, capacity planning, and targeted policy interventions for a resilient Aadhaar ecosystem.

THANK YOU !!