Unlocking Societal Trends in Aadhaar Enrollment and Updates

Data Analysis, Visualization, and Anomaly Detection

Team Analysis

January 20, 2026

1 Problem Statement

The Aadhaar system, one of the world's largest biometric identification databases, has matured from primarily new enrollments to a complex ecosystem of demographic updates, biometric refreshes, and residual enrollments. The core challenge is to identify meaningful patterns, trends, and anomalies within massive, fragmented Aadhaar datasets to translate raw data into actionable insights for informed decision-making. Specifically, the analysis addresses:

- Understanding temporal and geographic patterns in enrollment and update activities
- Detecting anomalies in data distribution across age groups and regions
- Investigating unexpected adult enrollment patterns (age 18+) in a saturated market
- Identifying correlation between enrollment activities and geographic factors such as border regions

2 Proposed Technical and Analytical Approach

The analysis employs a three-phase methodology:

2.1 Phase 1: Data Engineering

Establishes a robust pipeline to consolidate and prepare data:

- **Dynamic file detection**: Merges fragmented CSV files into unified master datasets using Pandas
- **De-duplication**: Removes exact duplicates to prevent skewed analysis
- **Entity resolution**: Standardizes state name variations to official nomenclature
- **Date standardization**: Converts all dates to ISO 8601 format for consistent time-series analysis
- **Logical sorting**: Implements nested sorting (date $\rightarrow\rightarrow$ ß$state$ß $\rightarrow$ ß$district$)$using stable merge - sort for time - ordered, geographically standardized data$

2.2 Phase 2: Exploratory Visualization

Generates comprehensive visualizations to establish baseline trends:

- Analyzes biometric updates through temporal trends and age group distributions
- Examines demographic updates across daily patterns
- Investigates enrollment data by age categories and temporal variations
- Performs trilateral analysis using radar charts and normalized heatmaps
- Identifies top-performing states across all three categories (biometric updates, demographic updates, new enrollments)

2.3 Phase 3: Targeted Case Study

Focuses on anomaly investigation using specialized geospatial algorithms:

- **Anomaly detection**: Identifies persistent 3
- **Geospatial analysis**: Develops custom Python module (`indiaenroll.py`) for granular district and pin-code level analysis
- **Hotspot identification**: Isolates geographic concentrations of adult enrollments
- **Correlation analysis**: Investigates relationship between adult enrollments and international border regions (Bangladesh, Nepal, Pakistan borders)

Dataset Description

Data Sources

The analysis utilizes Aadhaar transaction data provided in multiple fragmented CSV files. These files contain records of three primary transaction types:

- **Biometric Updates**: Records of biometric information refreshes
- **Demographic Updates**: Records of demographic information modifications
- **New Enrollments**: Records of new Aadhaar registrations

Dataset Structure

The consolidated dataset contains the following key columns:

Column Name	Data Type	Description
Date	Date	Transaction date in DD-MM-YYYY format (standardized to ISO 8601: YYY
State	String	Name of the state where transaction occurred (standardized to official nomen
District	String	Name of the district where transaction occurred
Pin Code	Integer	Six-digit postal code identifying specific geographic location
Age Group	Categorical	Classification of enrollee/updater age (0-5, 5-17, 18+)
Transaction Type	Categorical	Type of transaction (Biometric Update, Demographic Update, New Enrollme
Count	Integer	Number of transactions for the given combination of attributes

Data Quality and Preprocessing

Data Quality Issues Addressed

The raw datasets exhibited several quality issues that required preprocessing:

- **Fragmentation**: Data scattered across multiple CSV files requiring consolidation
- **Duplicate Records**: Exact duplicate entries that could skew aggregate counts
- **Inconsistent Naming**: State names with variations (e.g., "Orissa" vs. "Odisha", "Uttaranchal" vs. "Uttarakhand")
- **Date Format Variations**: Multiple date formats requiring standardization
- **Missing Values**: Incomplete records requiring validation and handling

Data Volume

- **Total Records Analyzed**: 3,971,882 records after de-duplication
- **Time Period Covered**: 2025-03-01 to 2025-12-31
- **Geographic Coverage**: All 36 states and union territories of India
- **Total Biometric Updates**: 1,766,212
- **Total Demographic Updates**: 1,222,598
- **Total New Enrollments**: 983,072

Data Limitations

- Dataset represents aggregated transaction counts rather than individual-level records
- Age groups are categorical rather than precise ages
- Geographic granularity limited to district and pin code level
- Temporal resolution is daily, not capturing intraday patterns

3 Methodology and Analytical Framework

Our analysis followed three distinct phases:

1. **Data Engineering:** Established a pipeline to merge, clean, and sort fragmented datasets, addressing inconsistencies and ensuring time-ordered, geographically standardized data.
2. **Exploratory Visualization:** Generated comprehensive visualizations to establish baseline trends.

The enrollment age distribution revealed a critical anomaly: a persistent 3% segment of adult enrollments (age 18+) in a saturated market.

3. **Targeted Case Study:** This observation triggered focused investigation using specialized geospatial algorithms to trace adult enrollments, identifying border-region hotspots as primary drivers.

3.1 Data Aggregation

Raw data was provided in multiple fragmented CSV files. We used dynamic file-detection to consolidate them into unified master files using Pandas.

```
import glob, pandas as pd, os

csv_files = glob.glob(os.path.join(input_dir, '*.csv'))
if not csv_files:
    print("No CSV files found!")

dfs = [pd.read_csv(f) for f in csv_files]
df = pd.concat(dfs, ignore_index=True)
```

Listing 1: Dynamic merging of fragmented datasets

3.2 Data Cleaning and Standardization

The merged data contained duplicates, inconsistent state names, and varying date formats. Key transformations included:

- **De-duplication:** Removed exact duplicates to prevent skewed counts
- **Entity resolution:** Mapped state name variations to official titles
- **Date standardization:** Converted all dates to ISO 8601 format

```
STATE_MAPPING = {'orissa': 'Odisha', 'uttaranchal': 'Uttarakhand'}

def standardize_state(state_value):
    state_lower = str(state_value).strip().lower()
    return STATE_MAPPING.get(state_lower, 'INVALID')

df['state'] = df['state'].apply(standardize_state)
df['date'] = pd.to_datetime(df['date'], format='%d-%m-%Y',
                            errors='coerce').strftime('%Y-%m-%d')
```

Listing 2: Standardization and validation logic

3.3 Logical Sorting

For time-series analysis, we applied nested sorting: date → state → district using stable merge-sort.

```
df.sort_values(by=["date", "state", "district"],
               ascending=[True, True, True],
               inplace=True, kind="mergesort")
```

Listing 3: Nested sorting for time-series analysis

4 Data Analysis and Visualization

Analysis was divided into modules exploring individual trends, pairwise relationships, and holistic patterns.

4.1 Biometric Updates Analysis

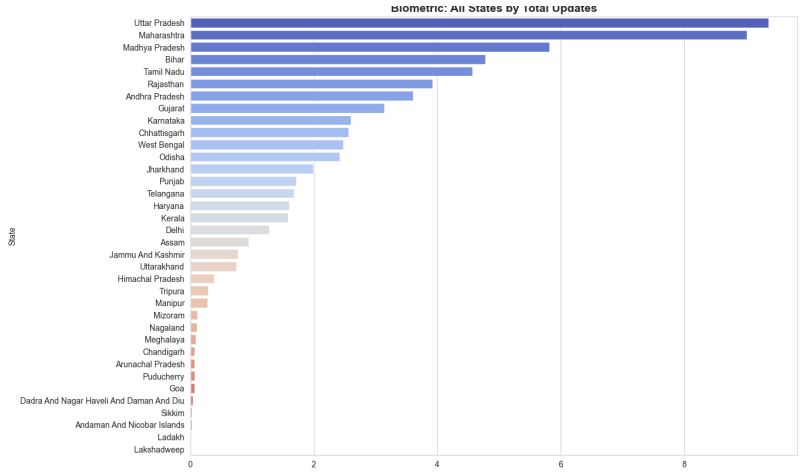


Figure 1: Biometric All states

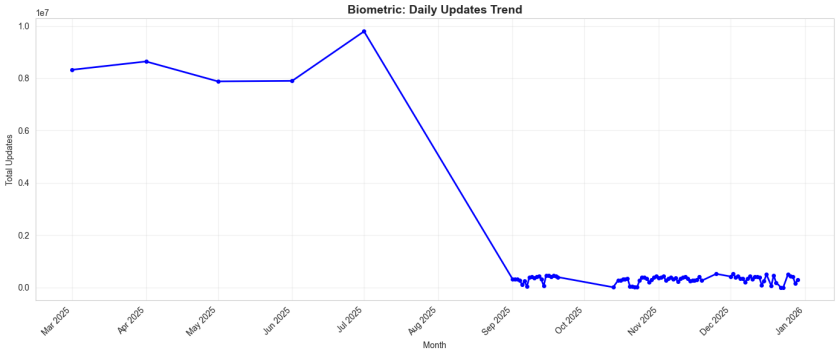


Figure 2: Biometric updates

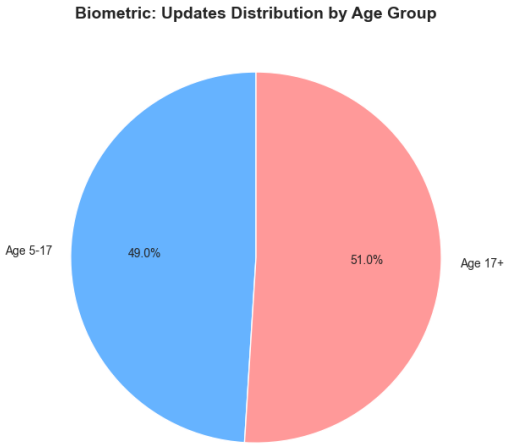


Figure 3: Age group Distribution

4.2 Demographic Updates Analysis

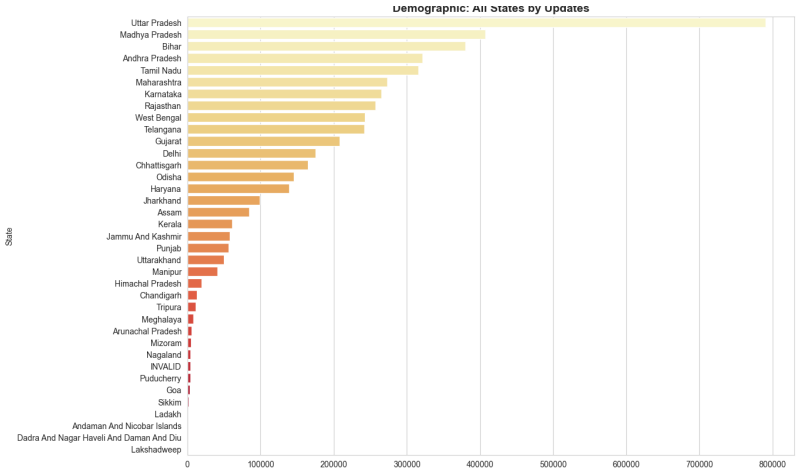


Figure 4: Demographic All states

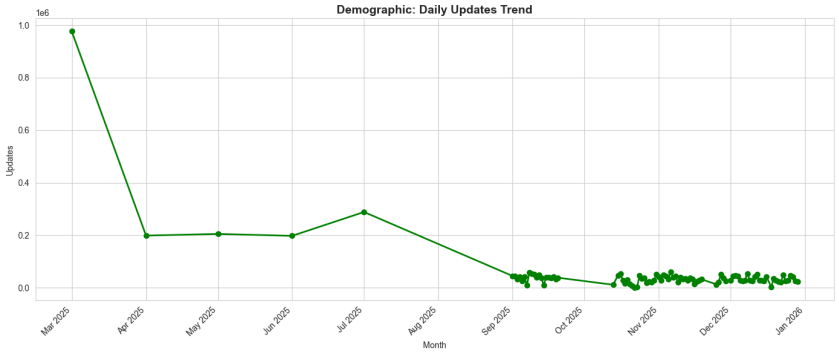


Figure 5: Demographic updates: daily trend

4.3 Enrollment Analysis

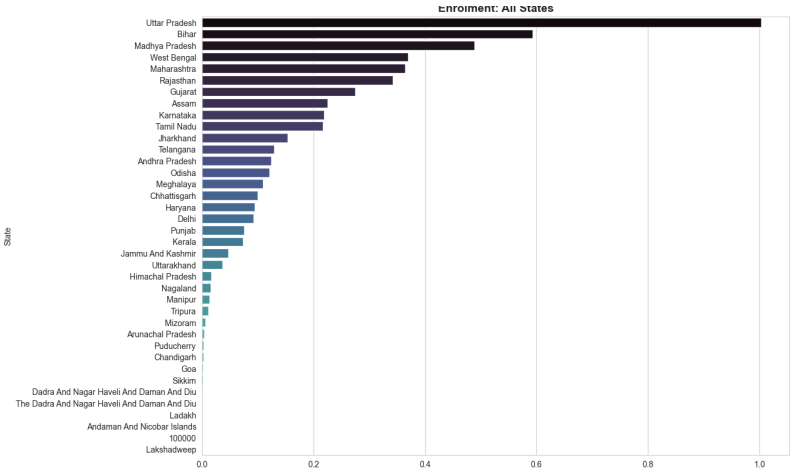
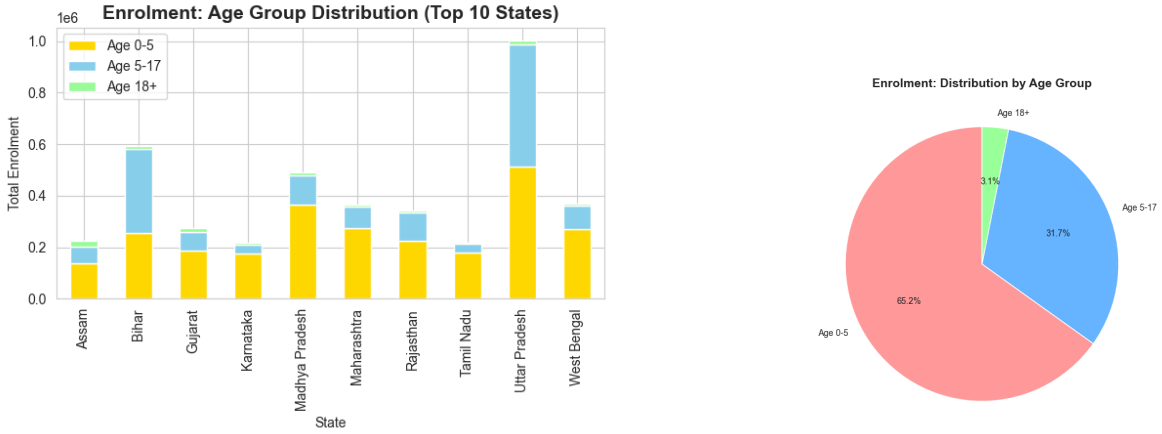


Figure 6: Enrollment All States



(a) Daily trends by age (b) Age distribution

Figure 7: Enrollment: age and temporal analysis

4.4 Trilateral Analysis

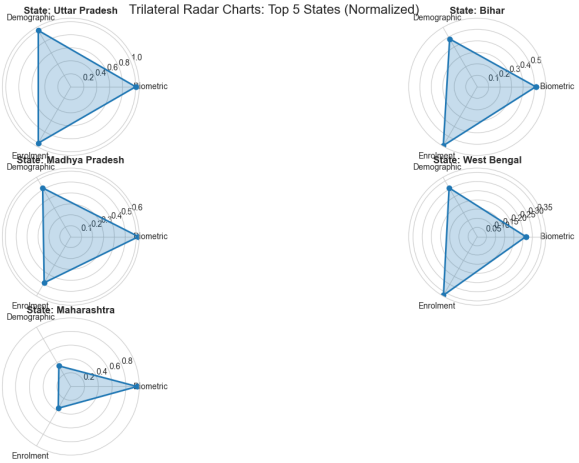


Figure 8: Radar chart (five states)

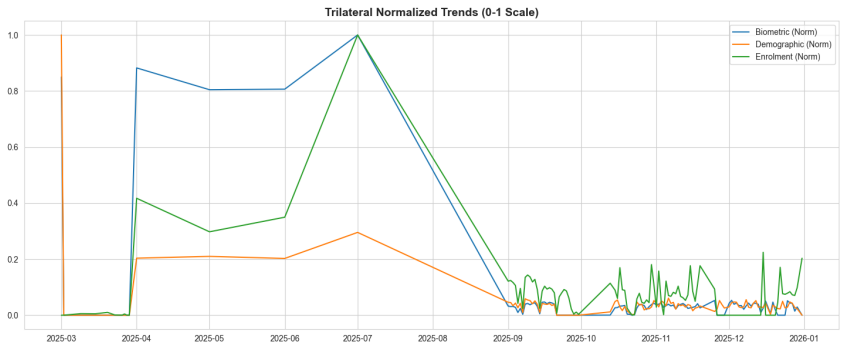


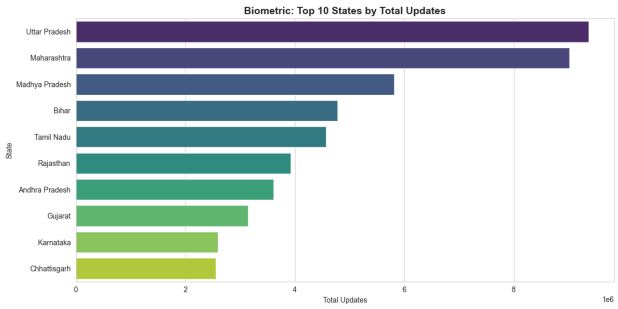
Figure 9: Normalized heatmap

4.5 Key Findings: Top States

Biometric Updates:

Rank	State	Total Updates
1	Uttar Pradesh	9,367,083
2	Maharashtra	9,020,710
3	Madhya Pradesh	5,819,736
4	Bihar	4,778,968
5	Tamil Nadu	4,572,152

(a) Top 5 states by biometric updates



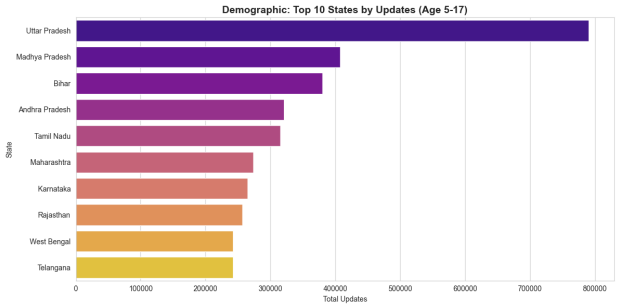
(b) Top 10 states visualization

Figure 10: Biometric updates: top states analysis

Demographic Updates:

Rank	State	Total Updates
1	Uttar Pradesh	8,542,328
2	Maharashtra	5,054,602
3	Bihar	4,814,350
4	West Bengal	3,872,318
5	Madhya Pradesh	2,912,938

(a) Top 5 states by demographic updates



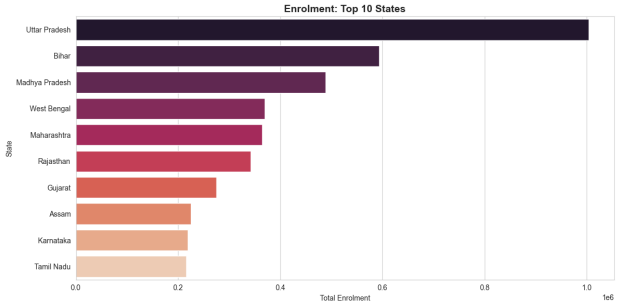
(b) Top 10 states visualization

Figure 11: Demographic updates: top states analysis

New Enrollments:

Rank	State	Total Enrollments
1	Uttar Pradesh	1,003,760
2	Bihar	593,753
3	Madhya Pradesh	489,212
4	West Bengal	369,725
5	Maharashtra	364,496

(a) Top 5 states by new enrollments



(b) Top 10 states visualization

Figure 12: New enrollments: top states analysis

5 Case Study: Adult Enrollment Anomalies

6 Case Study: Adult Enrollment Anomalies

6.1 Observation and Hypothesis

A national-level pie chart of enrollment distribution by age group reveals a clear anomaly: only **3.1%** of new enrollments are adults (age 18+), while 65.2% are children aged 0–5 and 31.7% are aged 5–17 (Figure 13).

Given near-universal Aadhaar saturation among adults in India (officially $> 99\%$), such a low proportion of adult enrollments is highly unexpected under normal demographic patterns. We hypothesized that this small but significant adult segment is not randomly distributed but heavily concentrated in geographic hotspots, particularly along India's international borders.

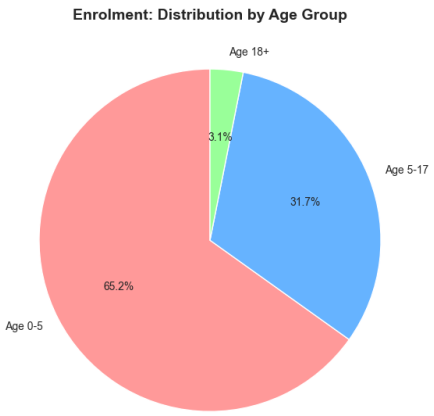


Figure 13: National Enrollment Distribution by Age Group

6.2 Methodology

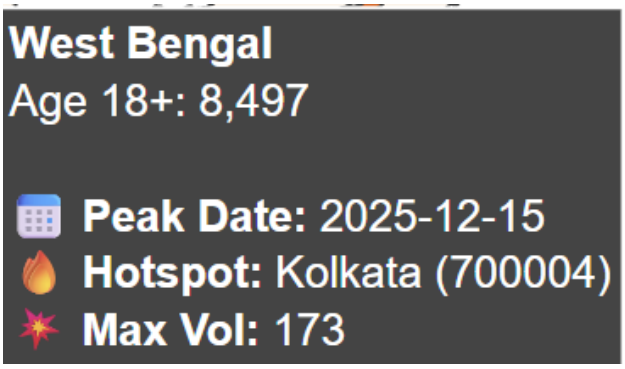
We conducted granular spatiotemporal analysis using custom visualization tools, isolating the `age_18_greater` metric to map enrollment volumes at state, district, and pin-code levels, identify peak dates, and detect hotspots through heatmaps, state pop-ups, and radar views.

6.3 Border-Level Radar Analysis

The border radar view (bubble map) provides a focused visualization of total enrollment activity concentrated along India's international borders. Large red bubbles highlight districts with exceptionally high volumes, overwhelmingly located in border states:

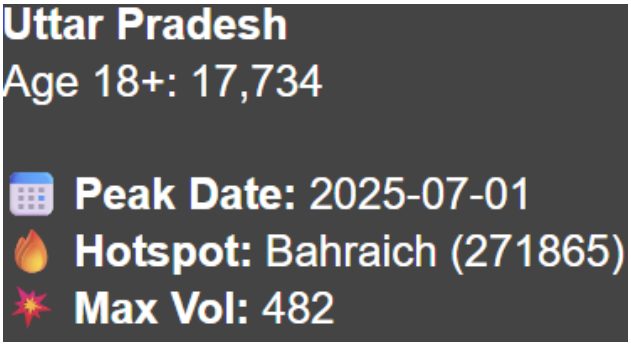
- **Uttar Pradesh**: Massive bubble over Bahraich district (Total Activity 101,003, pin-code 271865 highlighted)
- **Bihar**: Very large bubble in northern border districts (e.g., Sitamarhi region)
- **West Bengal**: Prominent activity in northern and eastern border areas
- Smaller but significant bubbles in northeastern states (Assam, Meghalaya) and western borders (Punjab, Gujarat)

This radar view confirms that the highest enrollment volumes — both total and disproportionately adult — are clustered in sensitive border zones (Bangladesh, Nepal, Pakistan, and Myanmar borders). The size and density of bubbles underscore that border regions drive the national anomaly, with interior states showing minimal or no activity.

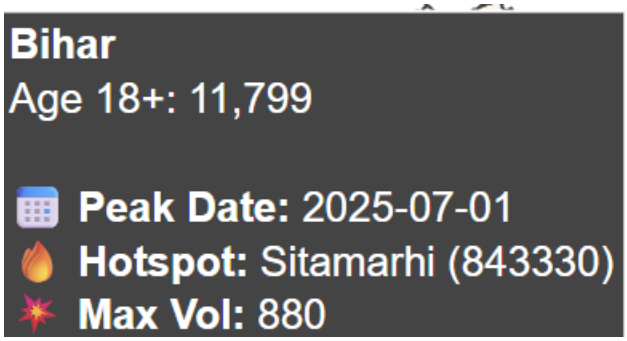


(a) West Bengal: Trends and Hotspot

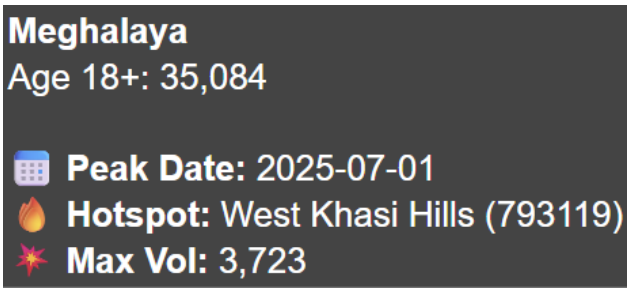
(b) Assam: Trends and Hotspot



(c) Uttar Pradesh: Trends and Hotspot



(d) Bihar: Trends and Hotspot



(e) Meghalaya: Trends and Hotspot (Highest Adult Volume)

Figure 14: State-Level Pop-ups for Key Border States with Elevated Adult Enrollment (Including Meghalaya as the Top Anomaly)

Figure 15: Border-Level Radar View Highlighting High-Volume Hotspots in Border Districts

Table 1: Key Border States Adult Enrollment Activity (Derived from Radar and Pop-up Data)

State	Total	Age 0–5	Age 5–17	Age 18+	Peak Date	Hotspot
Meghalaya	109,239	21,072	53,089	35,078	2025-07-01	West Khasi Hills (793119)
Assam	225,359	137,970	64,834	22,555	2025-07-01	Darrang (784145)
Uttar Pradesh	1,002,631	511,727	473,205	17,699	2025-07-01	Bahraich (271865)
Bihar	593,753	254,911	327,043	11,799	2025-07-01	Sitamarhi (843302)
West Bengal	369,249	270,419	90,335	8,495	2025-07-01	Uttar Dinajpur (733207)
Gujarat	275,042	188,709	70,270	16,063	2025-07-01	Surat (394210)
Punjab	75,773	60,481	12,175	3,117	2025-07-01	Amritsar (143001)

6.4 State-Level Analysis

Geospatial heatmaps and state pop-ups reveal pronounced adult enrollment spikes in specific border states, with synchronized peaks on **2025-07-01** across most hotspots — indicating coordinated activity rather than organic growth.

Key state-level observations:

- **Meghalaya**: Highest adult volume (35,078) despite smaller population; hotspot in West Khasi Hills near Bangladesh border
- **Assam**: 22,555 adults; hotspot in Darrang district
- **Uttar Pradesh**: 17,699 adults; concentrated in Nepal-border districts (Bahraich, Moradabad)
- **Bihar**: 11,799 adults; hotspot in Sitamarhi near Nepal border
- **West Bengal**: 8,495 adults; activity in northern border districts

Non-border states show pale colors and negligible adult enrollments, reinforcing the border-specific nature of the anomaly.

6.5 State-Level Comprehensive Analysis

The full state-wise breakdown (Table 2) confirms that adult enrollments are disproportionately high in border states, with Meghalaya, Assam, Uttar Pradesh, Gujarat, and Bihar leading. Non-border states show significantly lower adult shares.

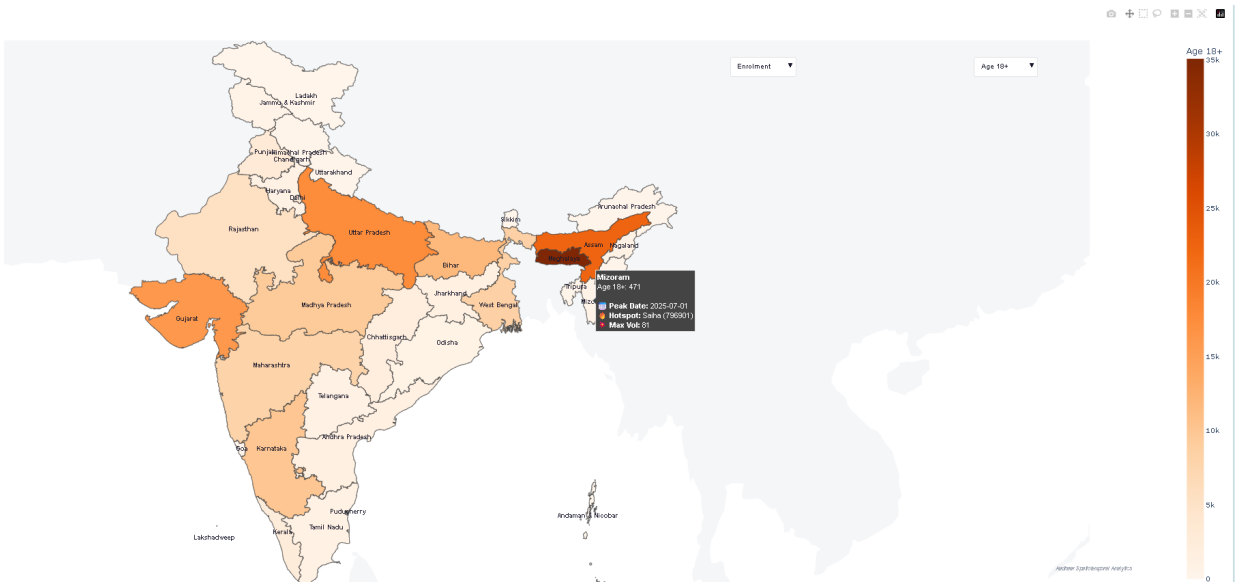


Figure 16: Geospatial Heatmap of Adult (18+) Enrollments Across India

Table 2: Comprehensive state-level enrollment analysis with hotspot identification

State	Total	Age 0–5	Age 5–17	Age 18+	Peak date	Hotspot
Andaman & Nicobar	501	469	32	0	2025-09-01	Nicobar (744301)
Andhra Pradesh	124,273	109,394	13,414	1,465	2025-12-22	Kurnool (518360)
Arunachal Pradesh	4,240	1,914	2,176	150	2025-07-01	Longding (792131)
Assam	225,359	137,970	64,834	22,555	2025-07-01	Goalpara (783129)
Bihar	593,753	254,911	327,043	11,799	2025-07-01	Sitamarhi (843302)
Chandigarh	2,620	2,377	210	33	2025-04-01	Chandigarh (160036)
Chhattisgarh	99,773	79,653	18,158	1,962	2025-07-01	Bijapur (494444)
Daman and Diu	1,782	1,484	248	50	2025-07-01	Dadra and Nagar Haveli
Delhi	92,838	67,844	21,971	3,023	2025-07-01	West Delhi (110059)
Goa	2,280	1,871	253	156	2025-11-14	North Goa (403401)
Gujarat	275,042	188,709	70,270	16,063	2025-07-01	Surat (394210)
Haryana	95,085	85,112	8,897	1,076	2025-06-01	Faridabad (121004)
Himachal Pradesh	16,909	16,081	650	178	2025-10-30	Sirmaur (173025)
Jammu & Kashmir	47,638	39,314	7,802	522	2025-12-15	Doda (182203)
Jharkhand	153,612	96,048	56,152	1,412	2025-07-01	Pakur (816107)
Karnataka	219,618	176,178	33,402	10,038	2025-07-01	Bengaluru (560068)
Kerala	73,950	52,950	18,360	2,640	2025-11-15	Malappuram (679325)
Ladakh	617	466	133	18	2025-12-15	Kargil (194301)
Lakshadweep	199	188	10	1	2025-11-02	Lakshadweep (682551)
Madhya Pradesh	487,892	363,244	115,172	9,476	2025-07-01	Barwani (451666)
Maharashtra	363,446	274,274	81,069	8,103	2025-07-01	Aurangabad (431001)
Manipur	13,199	5,044	7,895	260	2025-07-01	Churachandpur (795128)
Meghalaya	109,239	21,072	53,089	35,078	2025-04-01	West Khasi Hills (793119
Mizoram	5,774	4,044	1,259	471	2025-07-01	Lawngtlai (796891)
Nagaland	15,429	4,453	9,856	1,120	2025-07-01	Dimapur (797112)
Odisha	120,454	97,500	22,228	726	2025-12-15	Nabarangapur (764076)

Continued on n

Table 2 (continued)

State	Total	Age 0–5	Age 5–17	Age 18+	Peak date	Hotspot
Puducherry	2,983	2,746	193	44	2025-09-01	Karaikal (609602)
Punjab	75,773	60,481	12,175	3,117	2025-07-01	Amritsar (143001)
Rajasthan	340,591	224,977	110,131	5,483	2025-07-01	Jodhpur (342001)
Sikkim	2,175	1,040	1,030	105	2025-07-01	South Sikkim (737121)
Tamil Nadu	215,710	178,294	36,214	1,202	2025-12-15	Pudukkottai (614616)
Telangana	128,948	103,768	24,035	1,145	2025-07-01	Hyderabad (500008)
Tripura	11,008	7,165	3,597	246	2025-07-01	Sepahijala (799102)
Uttar Pradesh	1,002,631	511,727	473,205	17,699	2025-07-01	Moradabad (244001)
Uttarakhand	36,956	31,208	5,410	338	2025-07-01	Dehradun (248001)
West Bengal	369,249	270,419	90,335	8,495	2025-07-01	Uttar Dinajpur (733207)

6.6 Conclusion

The combined evidence from radar views and state-level analysis strongly indicates that adult Aadhaar enrollments, though only 3.1% nationally, are driven by border-specific factors — likely migration regularization or targeted demographic updates — rather than standard population growth.

This analysis strongly indicates that the observed adult enrollments are primarily driven by **migration regularization, border-area demographic updates, or coordinated enrollment campaigns**, rather than standard organic growth expected from natural population increase.