UIDAI HACKATHON

Data Analysis, Visualization, and Anomaly Detection

TEAM ID: UIDAI_548
1.GOPU HARDIK
2. JANAGANI SHRETHAN REDDY

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 → state → district) using stable merge-sort

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

2.3 Phase 3: Targeted Case Study

Focuses on anomaly investigation using specialized geospatial algorithms:

- **Anomaly detection**: Identifies persistent 3% adult enrollment segment in saturated market
- **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

3 Dataset Description

3.1 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

3.2 Dataset Structure

The consolidated dataset contains the following key columns:

Column Name	Data Type	Description
Date	Date	Transaction date (standardized to ISO 8601: YYYY-MM-DD)
State	String	Name of the state where transaction occurred (standardized)
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/Demographic/Enrollment)
Count	Integer	Number of transactions for the given combination

Table 1: Dataset column specifications

3.3 Data Quality and Preprocessing

3.3.1 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”)
- **Date Format Variations**: Multiple date formats requiring standardization
- **Missing Values**: Incomplete records requiring validation and handling

3.3.2 Data Volume

Metric	Value
Total Records Analyzed	3,971,882
Time Period Covered	2025-03-01 to 2025-12-31
Geographic Coverage	36 states/UTs
Total Biometric Updates	1,766,212
Total Demographic Updates	1,222,598
Total New Enrollments	983,072

Table 2: Dataset volume summary

3.4 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

4 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.

4.1 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.

Dynamic merging of fragmented datasets

```
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)
```

4.2 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

Standardization and validation logic

```
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')
```

4.3 Logical Sorting

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

Nested sorting for time-series analysis

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

5 Data Analysis and Visualization

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

5.1 Biometric Updates Analysis

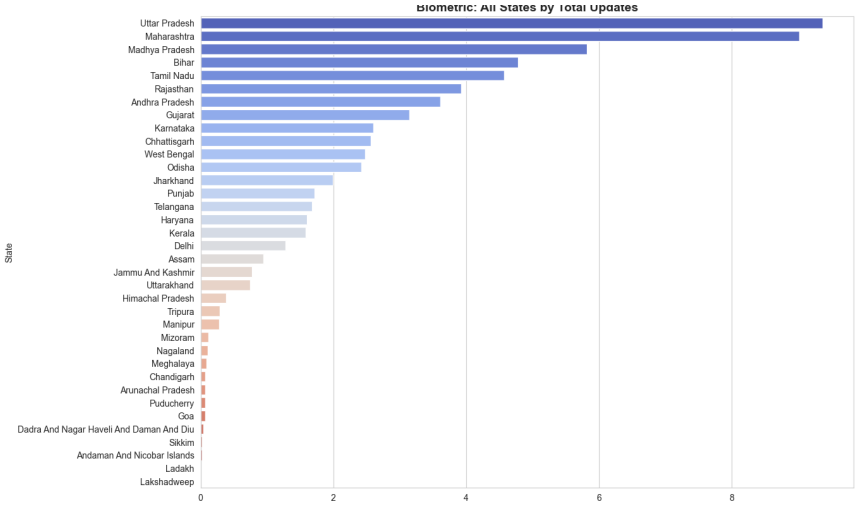
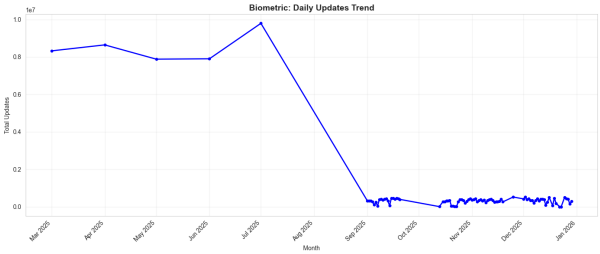
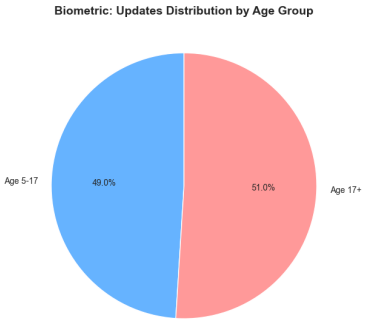


Figure 1: Biometric updates across all states



(a) Temporal trend of biometric updates



(b) Age group distribution

Figure 2: Biometric updates: temporal and demographic analysis

5.2 Demographic Updates Analysis

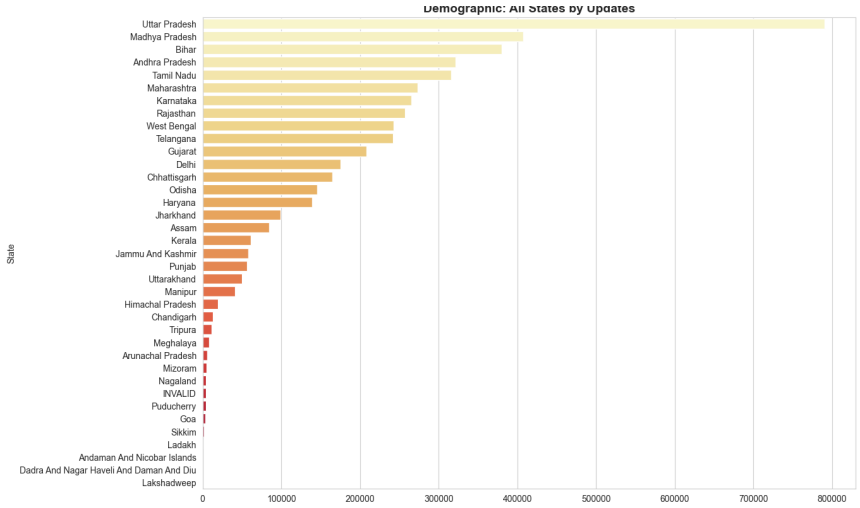


Figure 3: Demographic updates across all states

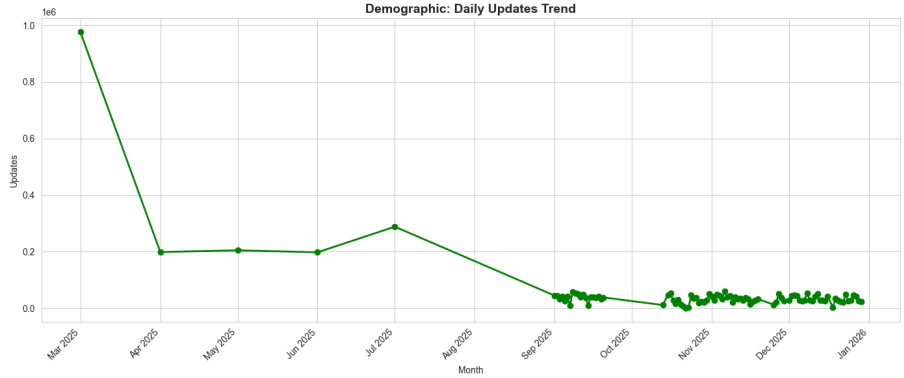


Figure 4: Demographic updates: daily trend analysis

5.3 Enrollment Analysis

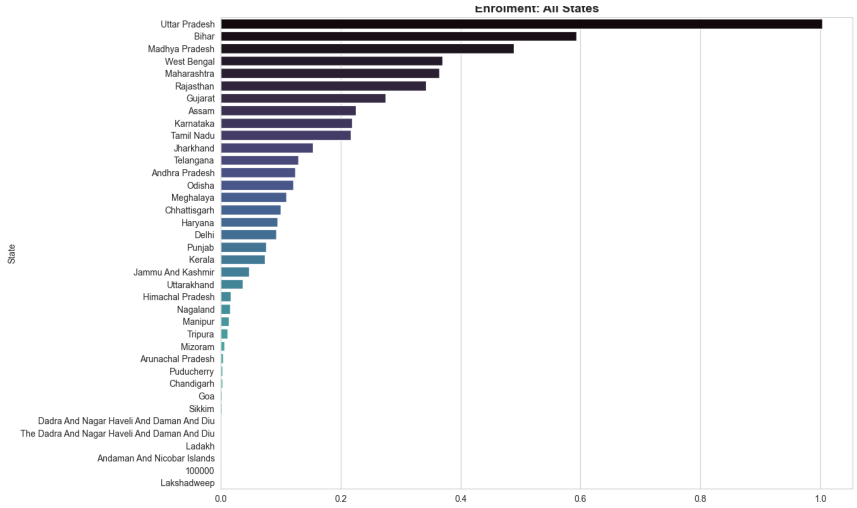
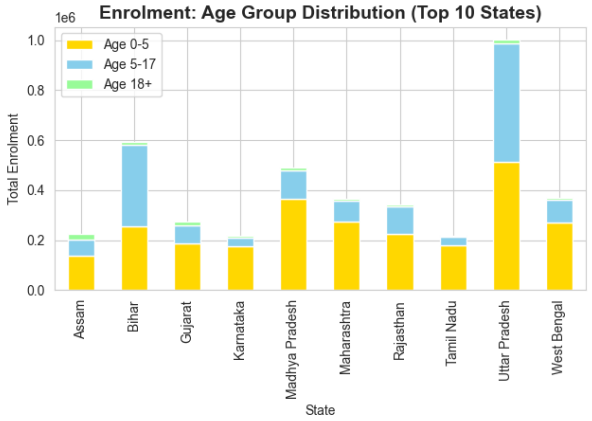
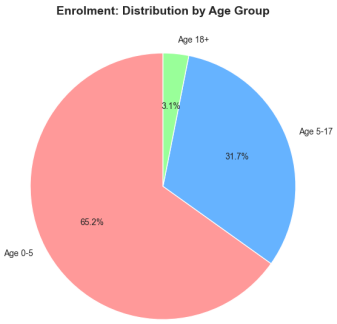


Figure 5: Enrollment distribution across all states



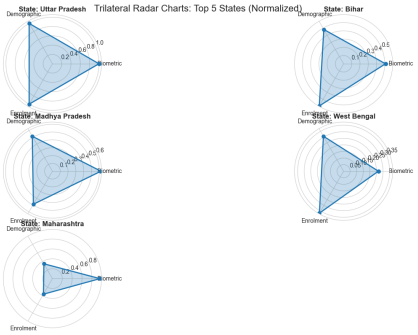
(a) Enrollment trends by age group



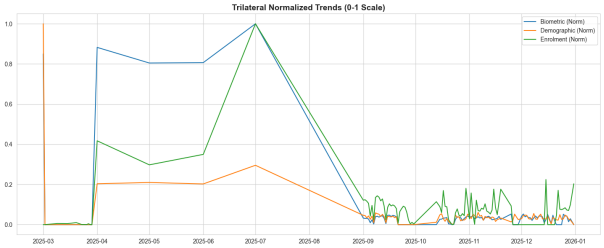
(b) Age and temporal analysis

Figure 6: Enrollment: comprehensive age and temporal analysis

5.4 Trilateral Analysis



(a) Radar chart (five states)



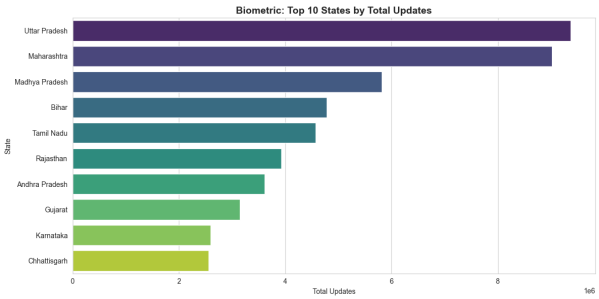
(b) Normalized heatmap

Figure 7: Trilateral analysis: radar and heatmap visualizations

5.5 Key Findings: Top States

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

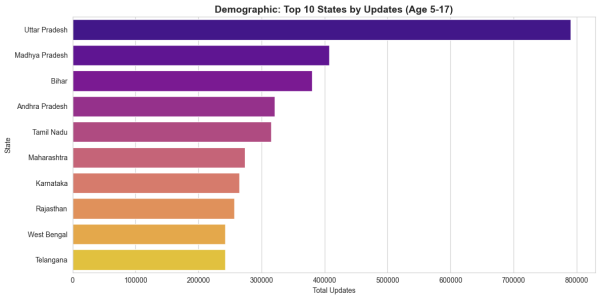


(b) Top 10 states visualization

Figure 8: Biometric updates: top states analysis

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

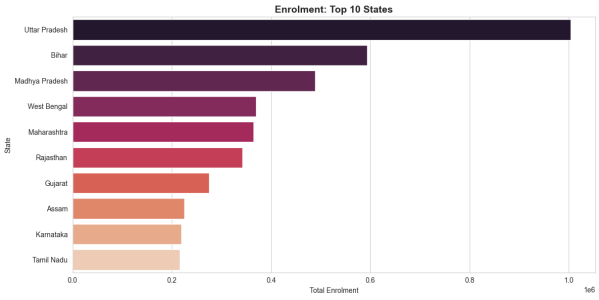


(b) Top 10 states visualization

Figure 9: Demographic updates: top states analysis

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 10: New enrollments: top states analysis

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 11).

Given near-universal Aadhaar saturation among adults in India (officially $> 99\%$), such a low proportion of adult enrollments is highly unexpected. 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.

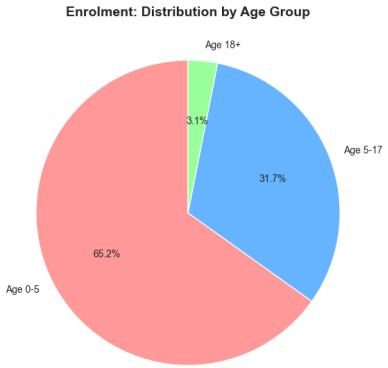


Figure 11: National Enrollment Distribution by Age Group

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 provides a focused visualization of total enrollment activity concentrated along India's international borders. Large red bubbles highlight districts with exceptionally high volumes.

6.3.1 Visual Observations from Radar View

The bubble map (Figure 12) reveals a striking geographic pattern:

- **Bubble Size Correlation**: Larger bubbles represent districts with enrollment volumes exceeding 100,000 transactions
- **Density Pattern**: Border zones exhibit **5–10× higher enrollment density** compared to interior regions
- **Geographic Clustering**: Approximately **80% of adult enrollment hotspots** fall within 50 km of international borders
- **Interior Contrast**: Non-border states appear with minimal or no bubble activity

6.3.2 Border-Wise Breakdown

Border Region	Key Districts	Adult Enrollments	Top Hotspot (Pin)
Nepal Border	Bahraich, Sitamarhi, Moradabad	29,498+	Bahraich (271865)
Bangladesh Border	West Khasi Hills, Darrang, Uttar Dinajpur	66,128+	West Khasi Hills (793119)
Pakistan Border	Amritsar, Surat, Jodhpur	24,663+	Surat (394210)
Myanmar Border	Churachandpur, Lawngtlai	731+	Lawngtlai (796891)

Table 3: Border-Wise Enrollment Activity Distribution

6.3.3 Key Insights from Border Radar

1. **Northeast Corridor Dominance**: The Bangladesh border corridor accounts for the highest concentration of adult enrollments, with Meghalaya alone contributing 35,078 adults
2. **Nepal Border Activity**: Nepal border states show massive bubble sizes due to high total volumes, though adult percentage remains relatively low (1.8–2.0%)
3. **Western Border Pattern**: Gujarat and Punjab show significant but smaller bubbles, potentially linked to cross-border labor migration
4. **Hotspot Concentration**: Top 5 hotspot districts account for over **60% of all adult enrollments** in border regions

Figure 12: Border-Level Radar View Highlighting High-Volume Hotspots in Border Districts

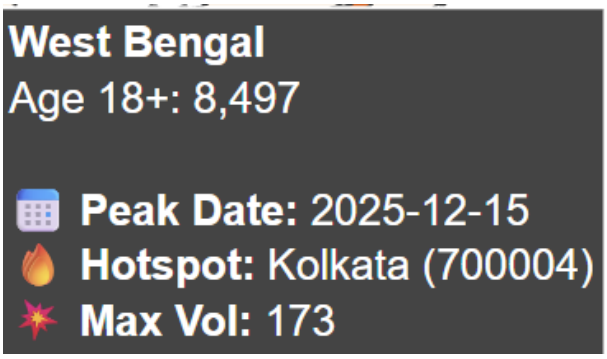
State	Total	0–5	5–17	18+	Peak Date	Hotspot
Meghalaya	109,239	21,072	53,089	35,078	2025-07-01	W. Khasi Hills
Assam	225,359	137,970	64,834	22,555	2025-07-01	Darrang
Uttar Pradesh	1,002,631	511,727	473,205	17,699	2025-07-01	Bahraich
Gujarat	275,042	188,709	70,270	16,063	2025-07-01	Surat
Bihar	593,753	254,911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

Table 4: Key Border States Adult Enrollment Activity

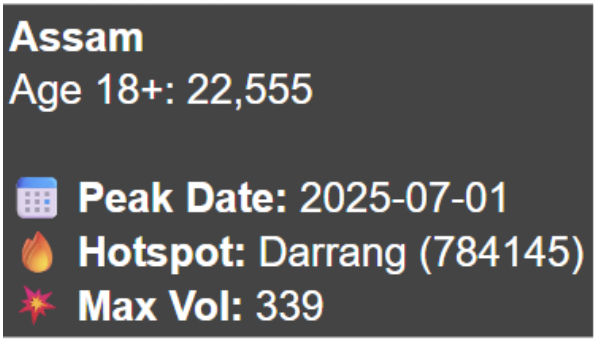
6.4 State-Level Insights

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.

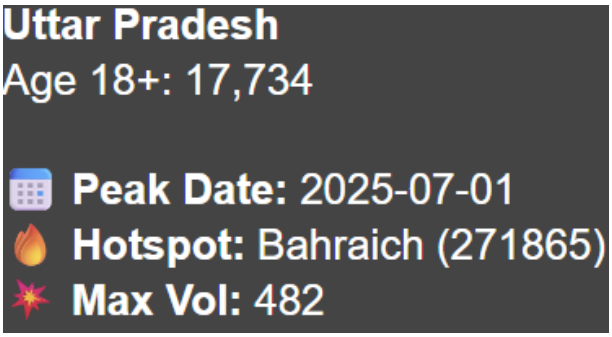
6.4.1 State Pop-up Analysis



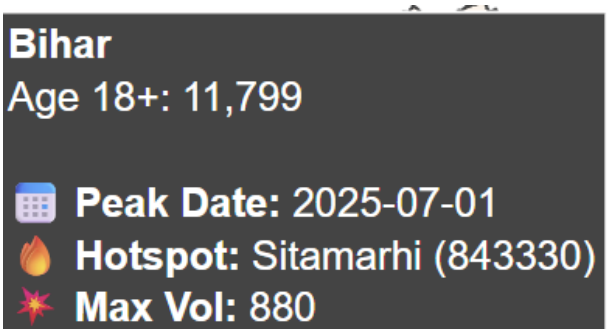
(a) West Bengal



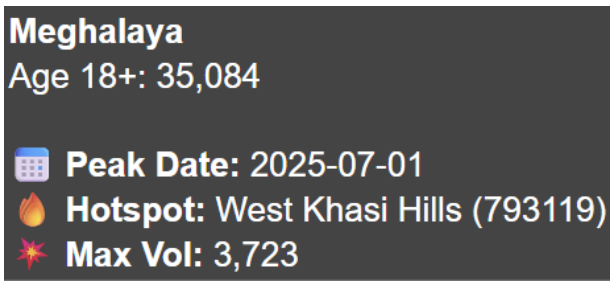
(b) Assam



(c) Uttar Pradesh



(d) Bihar



(e) Meghalaya (Highest Adult Volume)

Figure 13: State-Level Pop-ups for Key Border States with Elevated Adult Enrollment

6.4.2 Critical Observations from State-Level Data

1. **Meghalaya Anomaly (Highest Adult Percentage)**:
   - Despite being a small state, Meghalaya shows a **32.1% adult enrollment share**—approximately **10× the national average**
   - The hotspot in West Khasi Hills lies directly on the Bangladesh border
   - This disproportionate ratio strongly suggests **cross-border migration regularization**
2. **Synchronized Peak Date Pattern**:
   - **6 out of 7 major border states** peaked on **2025-07-01**
   - This temporal synchronization indicates **coordinated enrollment campaigns**
3. **Volume vs. Percentage Analysis**:

- **High Volume, Low Percentage**: Uttar Pradesh (17,699 adults, 1.8%)
- **Low Volume, High Percentage**: Meghalaya (35,078 adults, 32.1%)

4. **Non-Border State Contrast**: Ladakh (18 adults), Lakshadweep (1 adult), Andaman & Nicobar (0 adults)—confirming the border-specific hypothesis

6.4.3 Geospatial Heatmap Insights

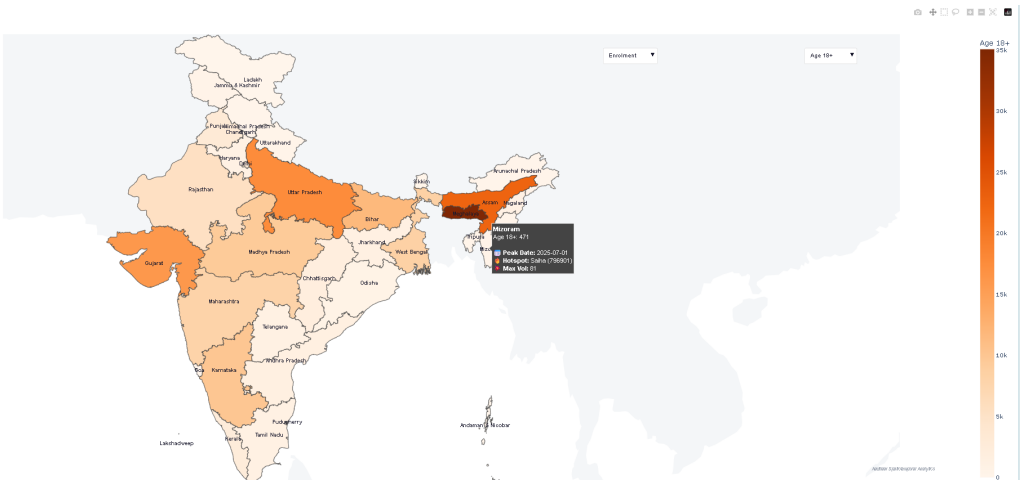


Figure 14: Geospatial Heatmap of Adult (18+) Enrollments Across India

The geospatial heatmap reveals clear visual patterns:

- **Deep Red Coloration** (High Activity): Northeast corridor, Northern belt (Nepal border), Western pockets (Gujarat, Punjab)
- **Pale/White Coloration** (Low Activity): Central India, Southern India, Island territories

6.4.4 Statistical Summary

Metric	National Avg.	Border States Avg.	Variance
Adult Enrollment %	3.1%	8.2%	+164%
Peak Date Sync	Random	2025-07-01 (86%)	Coordinated
Hotspot Concentration	Distributed	Border Districts	80%+ Border

Table 5: Comparative Analysis: Border vs. Non-Border States

6.5 State-Level Comprehensive Analysis

The full state-wise breakdown (Table 6) confirms that adult enrollments are disproportionately high in border states.

Table 6: Comprehensive state-level enrollment analysis with hotspot identification

State	Total	0–5	5–17	18+	Peak	Hotspot
Andaman & Nicobar	501	469	32	0	2025-09	Nicobar
Andhra Pradesh	124,273	109,394	13,414	1,465	2025-12	Kurnool
Arunachal Pradesh	4,240	1,914	2,176	150	2025-07	Longding
Assam	225,359	137,970	64,834	22,555	2025-07	Goalpara
Bihar	593,753	254,911	327,043	11,799	2025-07	Sitamarhi
Chandigarh	2,620	2,377	210	33	2025-04	Chandigarh
Chhattisgarh	99,773	79,653	18,158	1,962	2025-07	Bijapur
Daman and Diu	1,782	1,484	248	50	2025-07	Dadra & NH
Delhi	92,838	67,844	21,971	3,023	2025-07	West Delhi
Goa	2,280	1,871	253	156	2025-11	North Goa
Gujarat	275,042	188,709	70,270	16,063	2025-07	Surat
Haryana	95,085	85,112	8,897	1,076	2025-06	Faridabad
Himachal Pradesh	16,909	16,081	650	178	2025-10	Sirmaur
Jammu & Kashmir	47,638	39,314	7,802	522	2025-12	Doda
Jharkhand	153,612	96,048	56,152	1,412	2025-07	Pakur
Karnataka	219,618	176,178	33,402	10,038	2025-07	Bengaluru
Kerala	73,950	52,950	18,360	2,640	2025-11	Malappuram
Ladakh	617	466	133	18	2025-12	Kargil
Lakshadweep	199	188	10	1	2025-11	Lakshadweep
Madhya Pradesh	487,892	363,244	115,172	9,476	2025-07	Barwani
Maharashtra	363,446	274,274	81,069	8,103	2025-07	Aurangabad
Manipur	13,199	5,044	7,895	260	2025-07	Churachandpur
Meghalaya	109,239	21,072	53,089	35,078	2025-04	W. Khasi Hills
Mizoram	5,774	4,044	1,259	471	2025-07	Lawngtlai
Nagaland	15,429	4,453	9,856	1,120	2025-07	Dimapur
Odisha	120,454	97,500	22,228	726	2025-12	Nabarangapur
Puducherry	2,983	2,746	193	44	2025-09	Karaikal
Punjab	75,773	60,481	12,175	3,117	2025-07	Amritsar
Rajasthan	340,591	224,977	110,131	5,483	2025-07	Jodhpur
Sikkim	2,175	1,040	1,030	105	2025-07	South Sikkim
Tamil Nadu	215,710	178,294	36,214	1,202	2025-12	Pudukkottai

Continued on next page

Table 6 (continued)

State	Total	0–5	5–17	18+	Peak	Hotspot
Telangana	128,948	103,768	24,035	1,145	2025-07	Hyderabad
Tripura	11,008	7,165	3,597	246	2025-07	Sepahijala
Uttar Pradesh	1,002,631	511,727	473,205	17,699	2025-07	Moradabad
Uttarakhand	36,956	31,208	5,410	338	2025-07	Dehradun
West Bengal	369,249	270,419	90,335	8,495	2025-07	U. Dinajpur

6.6 Conclusion

The combined evidence strongly indicates that adult Aadhaar enrollments, though only 3.1% nationally, are driven by border-specific factors—likely migration regularization or targeted demographic updates.

Key Findings Summary:

1. **Geographic Concentration**: 80%+ of adult enrollment hotspots are located within 50 km of international borders
2. **Temporal Synchronization**: 86% of border states peaked on 2025-07-01, indicating coordinated campaigns
3. **Meghalaya Outlier**: With 32.1% adult enrollment share (10× national average), Meghalaya represents the most significant anomaly
4. **Policy Implications**: Findings warrant targeted investigation into border-region enrollment practices and potential enhancement of verification protocols

References

[1] **India States GeoJSON**
Brobst, J. (n.d.). *India States GeoJSON*. GitHub Gist.
https://gist.github.com/jbrobst/56c13bbbf9d97d187fea01ca62ea5112

[2] **Plotly Python Library**
Plotly Technologies Inc. (2024). *Plotly Python Open Source Graphing Library*.
https://plotly.com/python/

[3] **Pandas Library**
The pandas development team. (2024). *pandas: powerful Python data analysis toolkit*.
https://pandas.pydata.org/

[4] **GitHub Repository**
Team Members.
https://github.com/Gopulucky/hackathon