Age-wise Aadhaar Enrolment Analysis at Pincode Level

1. Problem Statement and Approach

This report analyzes Aadhaar enrolment data at the pincode level to identify patterns and potential areas for targeted intervention in enrolment monitoring and outreach. The primary goal is to understand how age composition, enrolment volume, and state-level differences relate to enrolment patterns, with a particular focus on child (0–5) enrolments.

Analytical approach:

- **Unit of analysis:** State–district–pincode–date level Aadhaar enrolment records.
- **Broad to detailed:** Start with national and state-level aggregations, then drill down to pincode-level volume tiers.
- **Derived indicators:** Construct total_enrol, volume tiers, and a child_under_focus_flag to highlight pincodes where child-age enrolment share is unusually low compared with similar areas.
- **Segmentation:** Segment by enrolment volume tiers and focus on top 3 high-volume states (Uttar Pradesh, Bihar, Madhya Pradesh) for state-specific patterns.
- **Robustness:** Compare patterns across states and across volume tiers (low vs high) and check extremes (top/bottom percentiles).
- **Decision orientation:** Translate patterns into simple, interpretable indicators and suggested use-cases for programme teams.

---

2. Datasets Used

Dataset source: Aadhaar enrolment dataset shared for UIDAI Data Hackathon 2026, derived from the Aadhaar Enrolment and Update Data on the Open Government Data Platform.

Dataset granularity: Each row represents Aadhaar enrolments for a given state–district–pincode–date, split into three age bands (0–5, 5–17, 18+).

Time coverage: Data spans from early to late 2025 (specific date range: min date to max date as observed in the dataset).

Dataset size: Approximately 3.3 million enrolment records across multiple states and thousands of unique pincodes.

Columns used:

Column Name	Data Type	Purpose
date	DATE	Reference date for the enrolment period
state	VARCHAR	State identifier
district	VARCHAR	District identifier
pincode	VARCHAR	Pincode (postal code) as text to preserve leading zeros
age_0_5	BIGINT	Count of enrolments in 0–5 age group
age_5_17	BIGINT	Count of enrolments in 5–17 age group
age_18_greater	BIGINT	Count of enrolments in 18+ age group

Derived columns (created during analysis):

- total_enrol = age_0_5 + age_5_17 + age_18_greater
- Age share columns: share_0_5, share_5_17, share_18_plus (computed as percentage of total_enrol per row)
- volume_tier (low, medium, high) based on pincode-level enrolment volume percentiles
- child_under_focus_flag (0 or 1) for low-volume pincodes in top 3 states

---

3. Methodology

3.1 Data Cleaning and Preprocessing

- **Date type conversion:** The date column is explicitly cast to DATE type to ensure correct temporal ordering and grouping.
- **Pincode standardisation:** The pincode column is converted from numeric to VARCHAR to preserve leading zeros and treat it as a categorical identifier rather than a numerical value.

- **Null value checks:** Confirmed there are no missing values in critical columns (date, state, district, pincode, age groups).
- **Type conversions:** Age columns (age_0_5, age_5_17, age_18_greater) are converted to numeric; no negative values are present.

Code example (data cleaning):

```
CREATE OR REPLACE VIEW aadhaar_clean_step1 AS
SELECT CAST(date AS DATE) AS date, state, district, pincode, age_0_5, age_5_17,
age_18_greater
FROM read_csv_auto('api_data_aadhar_enrolment.csv');

CREATE OR REPLACE VIEW aadhaar_clean_step2 AS
SELECT date, state, district, CAST(pincode AS VARCHAR) AS pincode, age_0_5, age_5_17,
age_18_greater
FROM aadhaar_clean_step1;
```

3.2 Feature Engineering and Derived Variables

Total enrolment: Created by summing enrolments across all age groups:

```
CREATE OR REPLACE VIEW aadhaar_with_total_enrol AS
SELECT *, age_0_5 + age_5_17 + age_18_greater AS total_enrol
FROM aadhaar_clean_step2;
```

Volume tiers: Pincodes are categorized into 'low', 'medium', and 'high' volume tiers based on the 20th and 80th percentiles of total_enrol across the entire dataset.

- 20th percentile (p20): 1.0
- 80th percentile (p80): 6.0
- Classification: low if total_enrol ≤ 1.0; high if > 6.0; medium otherwise.

Note: These percentiles are based on observed enrolments in this dataset and are used as a practical segmentation tool for pincode-level analysis, not as policy thresholds.

Code example (volume tiers):

```sql
CREATE OR REPLACE VIEW aadhaar_with_volume_tiers AS
SELECT *,
       CASE WHEN total_enrol > 6.0 THEN 'high'
            WHEN total_enrol <= 1.0 THEN 'low'
            ELSE 'medium'
       END AS volume_tier
FROM aadhaar_with_total_enrol;
```

Child under-focus flag: For the top 3 states (Uttar Pradesh, Bihar, Madhya Pradesh), pincodes are further segmented using state-specific percentiles (p20 = 2.0, p80 = 10.0).

For low-volume pincodes in these states:

1. Calculate the median age_0_5 share (proportion of 0–5 enrolments to total enrolments) for all low-volume pincodes within that state.
2. Flag a pincode as child_under_focus_flag = 1 if its age_0_5 share falls **5 percentage points or more below** this state-specific median.
3. Otherwise, set child_under_focus_flag = 0.

This rule is intentionally simple and interpretable so that state teams can understand and adapt it; it identifies relative under-focus within a state and volume tier, not absolute under-coverage in the population.

Code example (child under-focus flag):

```sql
CREATE OR REPLACE TEMPORARY VIEW median_age_0_5_low_volume AS
SELECT state, MEDIAN(CAST(age_0_5 AS REAL) / total_enrol) AS median_share_0_5_low
FROM aadhaar_top_3_states_with_volume_tiers
WHERE volume_tier = 'low'
GROUP BY state;

CREATE OR REPLACE VIEW aadhaar_with_child_under_focus_flag AS
SELECT t.*,
       CASE WHEN (CAST(t.age_0_5 AS REAL) / t.total_enrol) < (m.median_share_0_5_low
0.05) THEN 1 ELSE 0 END AS child_under_focus_flag
FROM aadhaar_top_3_states_with_volume_tiers t
JOIN median_age_0_5_low_volume m ON t.state = m.state
WHERE t.volume_tier = 'low';
```

3.3 Analysis Design (Univariate, Bivariate, Trivariate)

- **Univariate analysis:** Distributions of total_enrol by state, and age shares overall and per state.
- **Bivariate analysis:** Volume tier vs age shares; state vs age shares; time trends.
- **Trivariate analysis:** State × volume tier × age share (comparing high vs low volume pincodes within each of the top 3 states); state × volume tier × time (if sufficient date granularity).

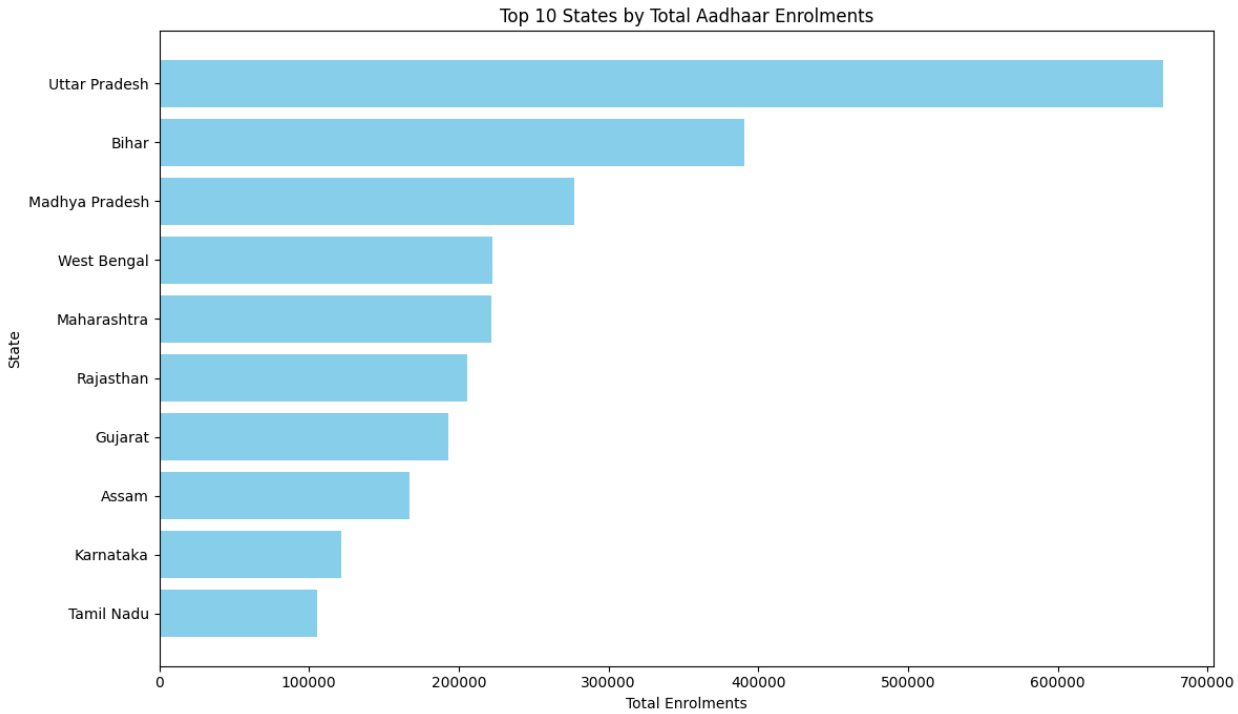
Tools used: DuckDB (SQL), Python (pandas, matplotlib, seaborn) for analysis, aggregation, and visualisation.

---

4. Data Analysis and Visualisation

This section presents key findings derived from the analysis through a series of charts and summarised data, illustrating enrolment patterns, age distributions, and geographical concentrations.

Chart 1: Top 10 States by Total Aadhaar Enrolments



What the chart shows:

- A horizontal bar chart displays total Aadhaar enrolments across the top 10 states, ranked by volume.

- Uttar Pradesh, Bihar, and Madhya Pradesh together account for approximately 40.54% of total enrolments.

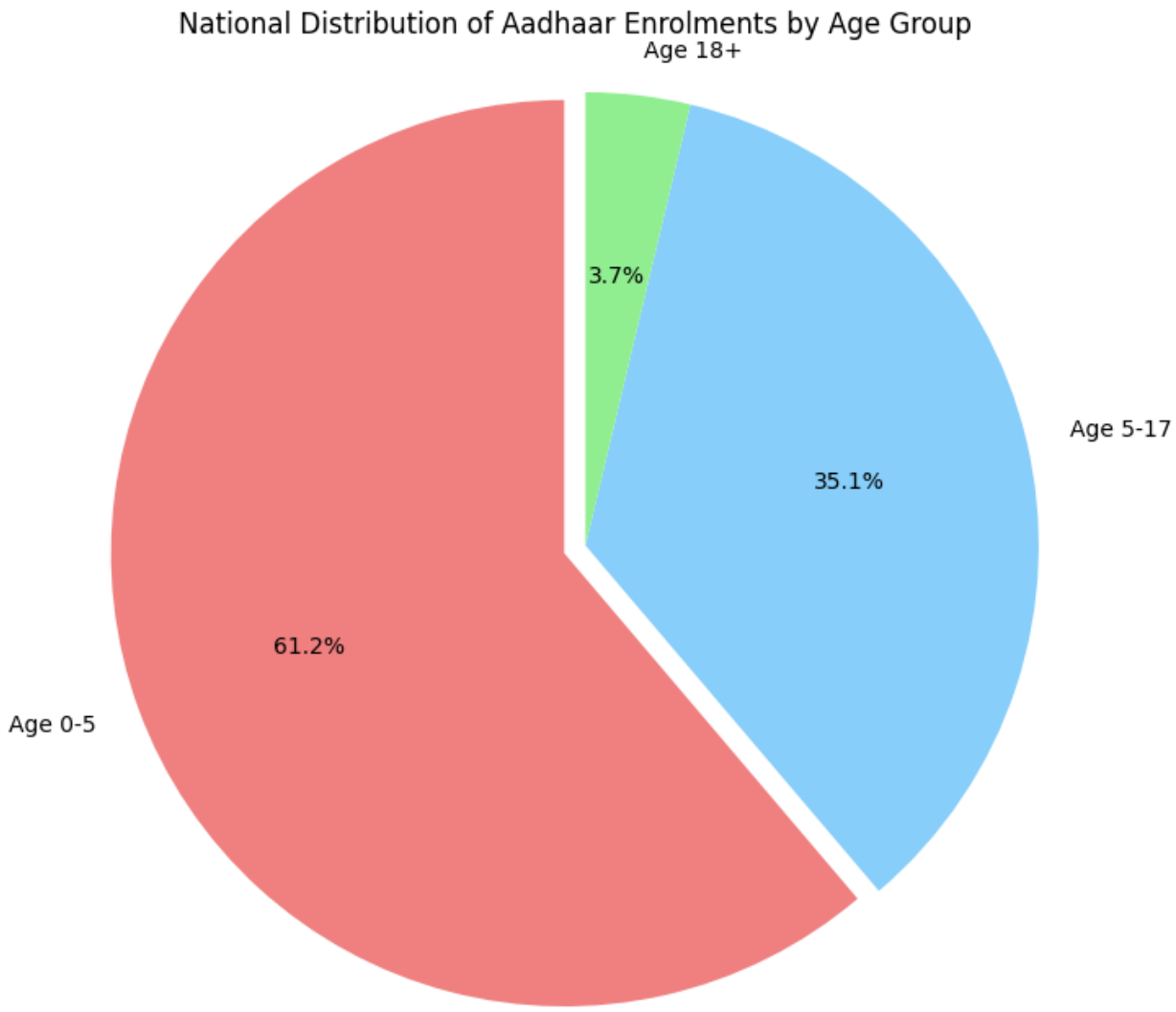
What it suggests:

- Aadhaar enrolment activity is geographically concentrated in a small number of high-volume states.
- This concentration is crucial for resource allocation decisions, as a large portion of overall enrolments originates from these regions.

What it does not prove:

- We cannot infer population coverage or per-capita enrolment rates without demographic denominators.

Chart 2: National Age Distribution



What the chart shows:

- A pie chart displays the national distribution of Aadhaar enrolments across the three age groups.
- 0–5 age group: approximately 61.21% of enrolments.
- 5–17 age group: approximately 35.08% of enrolments.
- 18+ age group: approximately 3.72% of enrolments.
- Together, 0–17 age groups account for over 96% of all enrolments.

What it suggests:

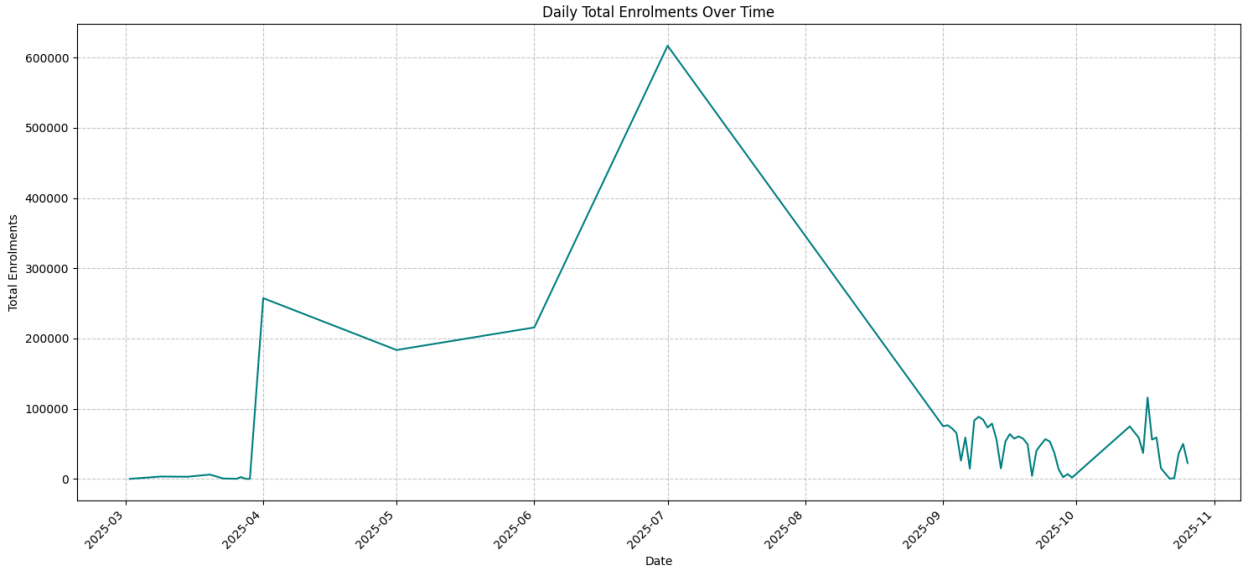
- Within this dataset, the Aadhaar system during the observed period is heavily oriented towards children and adolescents.

- This may reflect early-life identification and school-age use-cases.

What it does not prove:

- The data alone does not indicate population coverage levels or the underlying programme design intent, as no population denominators are available.

Chart 3: Daily Total Enrolments Over Time



What the chart shows:

- A line chart displays daily aggregate enrolments over the observed period (early to late 2025).
- The series exhibits noticeable spikes and dips, indicating volatility.

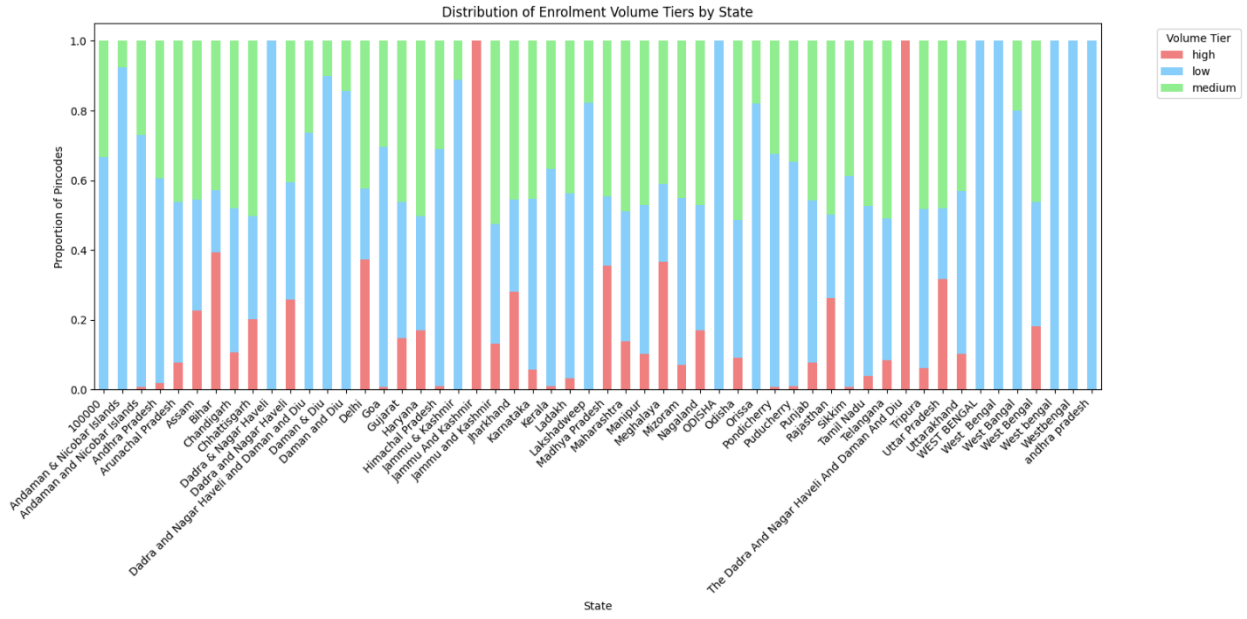
What it suggests:

- Enrolment activity during the observed period is not uniform; there are days with significantly higher and lower activity.
- This pattern may reflect operational capacity fluctuations, scheduled campaigns, or local events.

What it does not prove:

- Specific drivers of volatility (e.g., campaign timing, centre closures, seasonal demand) are not directly visible in the dataset and would require external programme or operational data.

Chart 4: Distribution of Enrolment Volume Tiers by State



What the chart shows:

- A stacked bar chart shows the proportion of low, medium, and high volume pincodes within each state.
- States vary significantly in their volume tier distributions; some are dominated by low-volume pincodes, while others have a more balanced or high-volume distribution.

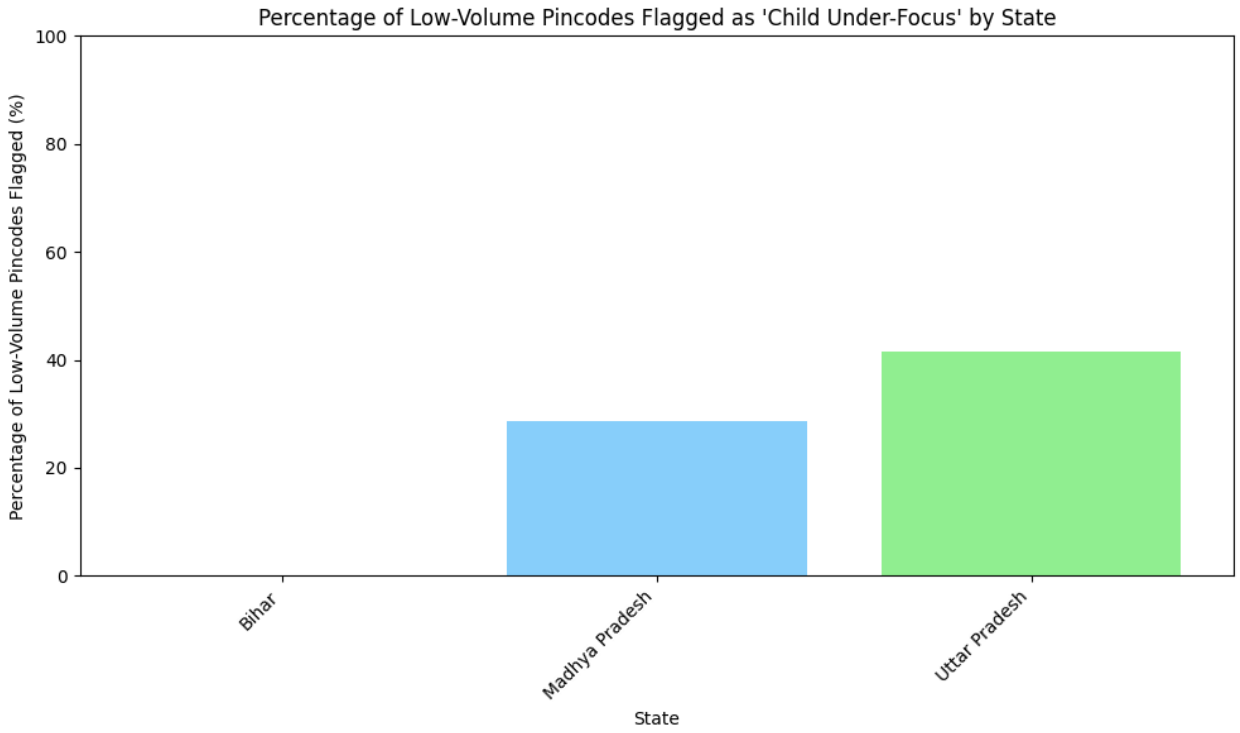
What it suggests:

- Enrolment penetration and activity levels differ materially across states, with implications for tailoring strategies to local contexts.

What it does not prove:

- We cannot conclude whether differences reflect programme design, operational capacity, or underlying population demand without additional data.

Chart 5: Child Under-Focus Flag Summary (Percentage of Flagged Low-Volume Pincodes)



What the chart shows:

- A bar chart displays the percentage of low-volume pincodes flagged as child_under_focus_flag = 1 for each of the top 3 states.
- Uttar Pradesh: approximately 41.50% of low-volume pincodes flagged.
- Madhya Pradesh: approximately 28.56% of low-volume pincodes flagged.
- Bihar: 0% of low-volume pincodes flagged.

What it suggests:

- In Uttar Pradesh and Madhya Pradesh, a substantial share of low-volume pincodes have 0–5 enrolment shares that are relatively low compared with other low-volume pincodes in the same state.
- These flagged pincodes are reasonable candidates for targeted review and potential child-focused outreach.
- Bihar's 0% flagging warrants investigation, as it may indicate data quality aspects, different enrolment strategies, or genuinely different age patterns.

What it does not prove:

- Flagging does not indicate absolute under-coverage of children; only relative under-focus within state and volume tier comparisons.

- Field validation and demographic data would be needed to assess actual coverage gaps.

---

5. Key Findings

Univariate Analysis

Finding 1 – Overall Enrolments:

- The dataset contains approximately 3.3 million Aadhaar enrolments across multiple states and thousands of pincodes.
- This represents a substantial volume of enrolment activity captured during the observed period.

Finding 2 – National Age Distribution:

- At the national level, Aadhaar enrolments are predominantly concentrated in the 0–17 age groups, which together account for over 96% of all enrolments.
- The 0–5 age group alone represents approximately 61.21% of enrolments.
- This indicates that, within this dataset, Aadhaar enrolments during the observed period are heavily oriented towards younger demographics.
- However, population denominators are not available, so coverage levels cannot be inferred.

Finding 3 – Enrolment Volatility:

- Daily Aadhaar enrolment activity exhibits significant fluctuations over the observed period, with noticeable spikes and dips.
- This pattern indicates that enrolment activity is volatile rather than uniform, suggesting varying daily demands, operational capacity, or campaign scheduling.

Bivariate Analysis

Finding 4 – Geographic Concentration:

- The top 3 states (Uttar Pradesh, Bihar, and Madhya Pradesh) collectively account for approximately 40.54% of total enrolments.
- This indicates that enrolment activity is geographically concentrated, with a large portion originating from these high-volume states.

Finding 5 – Age Shares by Volume Tier (State-Specific Patterns):

- Age distribution patterns in pincodes vary between high and low enrolment volume tiers.

- In Madhya Pradesh and Uttar Pradesh, low-volume pincodes have a higher 0–5 enrolment share (by approximately 3.2 and 1.2 percentage points, respectively) compared with high-volume pincodes.
- In Bihar, the pattern reverses: high-volume pincodes have a slightly higher 0–5 share than low-volume pincodes (by approximately 3.7 percentage points).
- These patterns are modest in magnitude but consistent within each state, suggesting that enrolment dynamics related to age groups differ based on both the state and the overall enrolment volume of a pincode.

Trivariate Analysis

Finding 6 – Child Under-Focus Areas (State-Specific Identification):

- Using the child_under_focus_flag, we identify low-volume pincodes in Uttar Pradesh and Madhya Pradesh where the 0–5 enrolment share is relatively low compared with the state's typical low-volume pincode pattern.
- In Uttar Pradesh, 41.50% of low-volume pincodes are flagged.
- In Madhya Pradesh, 28.56% of low-volume pincodes are flagged.
- In Bihar, 0% are flagged, which acts as a trigger for data and operational review rather than a conclusion about performance; it may indicate data quality characteristics, different enrolment strategies, or genuinely different age patterns.
- These flagged pincodes represent areas where targeted interventions (awareness, outreach, operational review) may be warranted to improve relative child enrolment activity.

---

6. Indicators and Decision Frameworks

6.1 Enrolment Volume Tiers (volume_tier)

Construction:
Pincodes are categorized into 'low', 'medium', and 'high' volume tiers based on the 20th and 80th percentiles of their total_enrol. For the overall dataset, p20 = 1.0 and p80 = 6.0. For the top 3 states, state-specific percentiles (p20 = 2.0, p80 = 10.0) are used.

Monitoring value:
This indicator provides a segmentation framework for understanding enrolment activity levels across different geographic units. By identifying areas with consistently low or high enrolment volumes, decision-makers can:

- Tailor resource allocation and operational strategies (e.g., permanent enrolment centres vs mobile camps).
- Adjust campaign intensity and outreach approach to local context.
- Assess saturation or growth potential in various regions.

6.2 Child Under-Focus Flag (child_under_focus_flag)

Construction:
This flag is applied to low-volume pincodes within the top 3 states. It is set to 1 if a pincode's age_0_5 share is 5 percentage points or more below its state's median age_0_5 share for other low-volume pincodes; otherwise, it is 0.

Monitoring value:
The child_under_focus_flag serves as a critical monitoring signal to identify specific low-volume areas where child enrolments are disproportionately low relative to their state's typical performance in similar areas. It allows for:

- Efficient allocation of limited resources to areas with the most significant need for improving child enrolment rates.
- Targeted outreach programmes, awareness campaigns, or removal of specific enrolment barriers in flagged pincodes.
- Prioritisation of field review and operational investigation.

The same construction can be extended to other states and time windows, enabling a consistent, rule-based monitoring framework for detecting relative child-enrolment under-focus across the Aadhaar ecosystem.

---

7. Impact and Decision Hooks

7.1 Targeted Child Enrolment Drives in Flagged Low-Volume Pincodes

Who acts: State-level Aadhaar authorities, local district administrators, and field teams.

What decision changes: Allocation of resources, deployment of mobile enrolment units, and design of localized outreach campaigns.

Metric enabling action: The child_under_focus_flag (specifically, the count and percentage of flagged low-volume pincodes per state) identifies areas requiring focused intervention for the 0–5 age group. In Uttar Pradesh (41.50% flagged) and Madhya Pradesh (28.56% flagged), these pincodes become candidates for priority outreach.

7.2 Investigation of Bihar's Low-Volume Child Enrolment Anomaly

Who acts: Data quality and analytics teams; state-level program managers.

What decision changes: Initiation of a data audit for Bihar's low-volume pincodes, potential recalibration of calculations, or investigation into on-ground enrolment practices.

Metric enabling action: Bihar's 0% flagging, combined with its observed median characteristics in low-volume pincodes, acts as a trigger for review rather than a conclusion about performance. This may indicate data quality aspects, distinct enrolment strategies, or different demographic patterns warranting clarification.

7.3 Age-Specific Campaign Adjustments in High-Volume Areas

Who acts: National and state-level marketing and programme strategists.

What decision changes: Shifting campaign focus in 'high' volume tier pincodes to include more messaging and facilities for 5–17 and 18+ age groups, where these groups are relatively more prominent.

Metric enabling action: The state-specific variations in age share distributions between volume tiers inform where broader age-group targeting would be more effective.

7.4 State-Specific Strategy Customisation

Who acts: National policy makers and state programme directors.

What decision changes: Development of tailored enrolment policies and resource distribution models that account for state-level variations in age mix and volume tier distributions.

Metric enabling action: The heterogeneous patterns observed across states (e.g., Uttar Pradesh vs Bihar vs Madhya Pradesh) in age_0_5 share gaps between volume tiers guide the customisation of strategies and resource allocation.

---

8. Limitations

This analysis is subject to certain data constraints and limitations, which preclude drawing broader conclusions:

Data scope: The analysis is based on Aadhaar enrolment data from early to late 2025. This limited time window and sample may not fully represent all-India trends or be generalizable to other periods.

No population denominators: The dataset contains only enrolment counts, not population baselines. Consequently, coverage rates, per-capita enrolment, or demographic penetration cannot be calculated. Findings describe enrolment patterns, not population coverage.

Exclusion of external factors: The dataset lacks crucial external factors such as socio-economic indicators (income, education, service accessibility), geographical features, or programme-specific campaign details. Consequently, the analysis cannot establish causal relationships between observed patterns and these unmeasured variables.

Correlational vs causal findings: While the analysis identifies correlations and patterns (e.g., varying age mixes in different volume tiers), it does not provide insights into underlying causes. For instance, it cannot explain why certain age groups are dominant in particular areas or why daily enrolment activity is volatile.

Snapshot nature: The data provides a snapshot of enrolment activities during a specific period. Dynamic changes in population demographics, policy interventions, or operational strategies over time are not explicitly captured, limiting long-term trend analysis or predictive capabilities.

Data quality considerations: The observation of 0% flagging in Bihar suggests potential data quality characteristics or specific enrolment dynamics that warrant further investigation, which cannot be fully addressed within the confines of the current dataset.

---

9. Conclusion

This analysis of Aadhaar enrolment data at the pincode level provides focused insights into age-wise patterns, enrolment volume tiers, and geographic concentration during the observed period. By identifying patterns such as the dominance of younger age enrolments nationally, state-specific variations in age-volume relationships, and the presence of low-volume pincodes with relatively low child enrolment shares in specific states, the work supports more targeted monitoring and intervention design.

Key takeaways for decision-makers:

1. Enrolment activity is geographically concentrated (top 3 states = 40.54%), making them critical focus areas for programme impact.
2. Age-volume relationships are state-specific (Uttar Pradesh and Madhya Pradesh show child-heavy low-volume pincodes; Bihar shows the opposite), necessitating customised strategies.

3. The child_under_focus_flag identifies 41.50% of low-volume pincodes in Uttar Pradesh and 28.56% in Madhya Pradesh as candidates for child-focused outreach.
4. Simple, interpretable indicators like the flag enable consistent monitoring and operational prioritisation across states.

Future work could integrate demographic baselines, additional time periods, qualitative inputs from field teams, and validation of the flag threshold to refine and operationalise these monitoring frameworks at scale.