Unlocking Societal Trends in Aadhaar Enrolment and Updates

Data Analysis, Visualization, and Anomalies Detection

Team Analysis

January 20, 2026

1 Introduction

The Aadhaar system represents one of the largest biometric identification databases in the world. As the ecosystem matures, the nature of data transactions shifts from primarily new enrolments to a mix of demographic updates, biometric refreshes, and residual enrolments.

The objective of this report is to identify meaningful patterns, trends, and anomalies within the provided Aadhaar datasets. By analyzing variables such as age distribution, geographic concentration, and temporal spikes, we aim to translate raw data into clear insights that can support informed decision-making.

This document details our technical approach, visualizes key trends across Biometric, Demographic, and Enrolment categories, and concludes with a deep-dive case study into specific anomalies detected in adult enrolment patterns in border regions.

2 Methodology and Analytical Framework

Our project execution followed a linear analytical path designed to uncover hidden trends within the Aadhaar ecosystem. The process evolved through three distinct phases:

1. **Data Engineering (Clean):** We first established a robust pipeline to **Merge, Clean, and Sort** the fragmented raw datasets. This addressed data inconsistencies and ensured the datasets were strictly time-ordered and geographically standardized.

2. **Exploratory Visualization (Visualize):** We generated comprehensive unilateral, bilateral, and trilateral visualizations to establish baseline trends. During this phase, the *Enrolment Age Distribution Pie Chart* revealed a critical anomaly: a persistent **3% segment of adult enrolments (Age 18+)** in an otherwise saturated market.

3. **Targeted Case Study (Deep Dive):** This visual observation triggered a specific investigation. We developed specialized geospatial algorithms to trace these adult enrolments, ultimately identifying specific border-region hotspots as the primary drivers of this trend.

The following subsections detail the technical implementation of this framework.

2.1 1. Data Aggregation (Merging)

The raw data was provided in multiple fragmented CSV files. We utilized a dynamic file detection approach to identify all relevant datasets in the directory and consolidated them into unified master files (Demographic, Biometric, and Enrolment) using the Pandas library.

Key Technique: We utilized ‘glob‘ for pattern matching to automatically detect input files and ‘pd.concat‘ for batch merging, ensuring no manual selection errors occurred.

```
import glob
import pandas as pd
import os
from datetime import datetime
# 1. Dynamic File Detection
```

```
csv_files = glob.glob(os.path.join(BASE_DIR, "api_data_aadhar_enrolment*.csv"))

if not csv_files:
    raise RuntimeError("No CSV files found!")

# 2. Batch Loading
df_list = []
for file in csv_files:
    # low_memory=False used to handle mixed types
    df_list.append(pd.read_csv(file, low_memory=False))

# 3. Concatenation
final_df = pd.concat(df_list, ignore_index=True)
```

Listing 1: Dynamic Merging of Fragmented Datasets

2.2 2. Data Cleaning and Standardization

The merged data contained duplicate records, inconsistent state names (e.g., "Orissa" vs. "Odisha"), and varying date formats. We implemented a chunk-based processing method to clean the data efficiently without exhausting memory resources.

Key Transformations:

- **Deduplication:** Removed exact duplicate rows to prevent skewed counts.
- **Entity Resolution:** Mapped variations of state names to their official titles using a correction dictionary.
- **Date Standardization:** Converted all date columns to ISO 8601 (`YYYY-MM-DD`).

```
state_corrections = {
    'Orissa': 'Odisha',
    'Uttaranchal': 'Uttarakhand',
    'Dadra & Nagar Haveli': 'Dadra And Nagar Haveli And Daman And Diu'
}

for chunk in pd.read_csv(INPUT_FILE, chunksize=200000):
    # 1. Text Standardization
    chunk["state"] = (
        chunk["state"].str.strip().str.title().replace(state_corrections)
    )

    # 2. Hard Validation
    invalid_mask = (~chunk["state"].isin(VALID_STATES))
    chunk.loc[invalid_mask, "state"] = "INVALID"

    # 3. Date Formatting
    chunk["date"] = pd.to_datetime(
        chunk["date"], errors="coerce", dayfirst=True
    ).dt.strftime("%Y-%m-%d")
```

Listing 2: Standardization and Validation Logic

2.3 3. Logical Sorting (Time-Series Preparation)

To facilitate time-series forecasting, we applied a nested sorting strategy: *Date* → *State* → *District*. We utilized the `mergesort` algorithm (stable sort) to preserve the relative order of records.

```
# Primary: Date (Chronological), Secondary: State, Tertiary: District
sort_cols = ["date", "state", "district"]

df.sort_values(
    by=sort_cols,
    ascending=[True, True, True],
    inplace=True,
    kind="mergesort"  # Stable sort algorithm
)
```

Listing 3: Nested Sorting for Time-Series Analysis

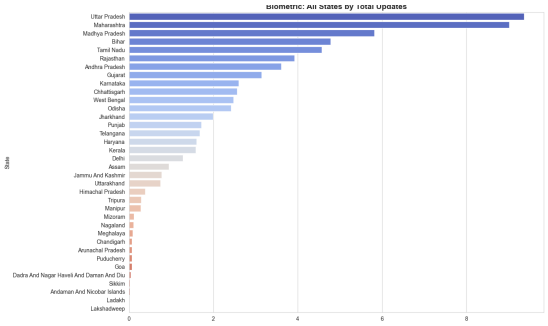
3 Data Analysis and Visualization

Our analysis was divided into three distinct modules to explore individual trends, pairwise relationships, and holistic patterns across the datasets.

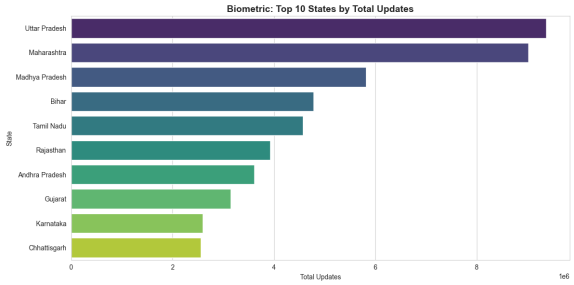
3.1 1. Visualization Summary

A. Unilateral Analysis (Individual Trends)

This module focused on understanding the distribution and trends of single variables (Biometric, Demographic, and Enrolment) independently.

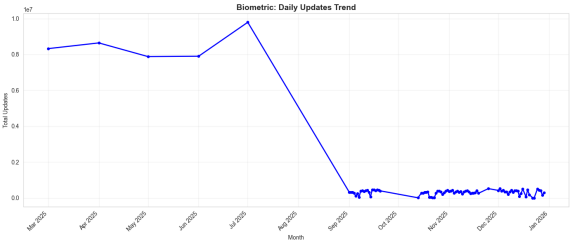


(a) All States by Total Updates

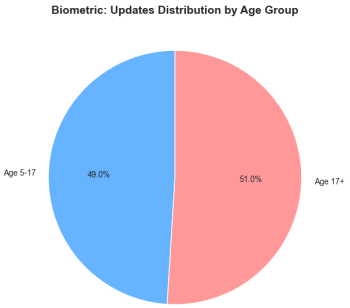


(b) Top 10 States by Total Updates

Figure 1: Biometric Updates Geographic Distribution



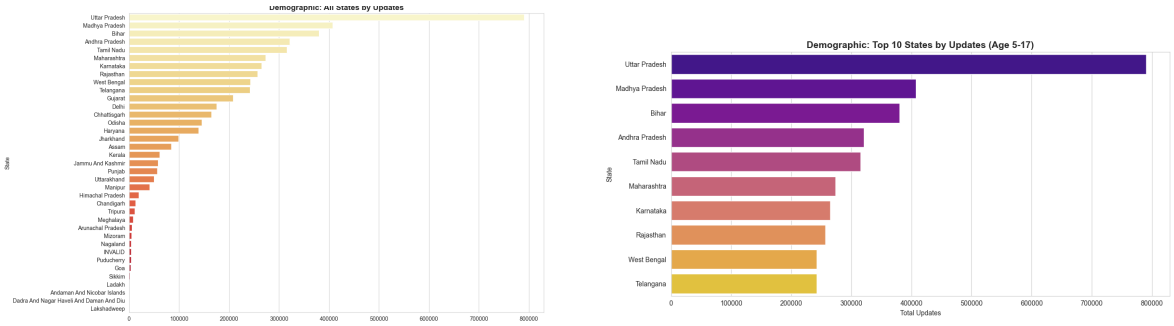
(a) Daily Updates Trend



(b) Distribution by Age Group

Figure 2: Biometric Updates Temporal and Demographical Breakdown

Biometric Updates Analysis



(a) All States by Updates

(b) Top 10 States (Age 5-17)

Figure 3: Demographic Updates Geographic Distribution

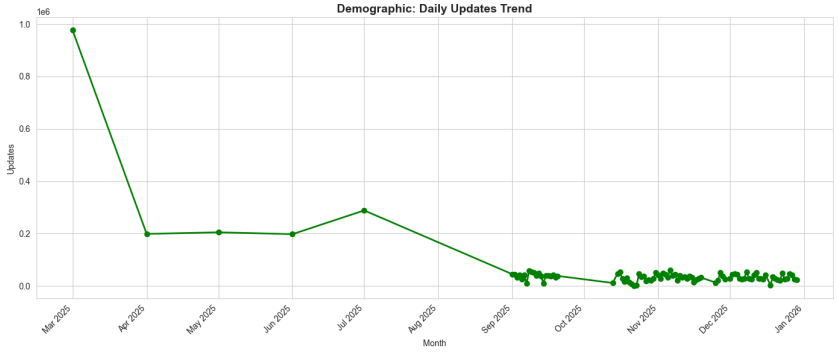
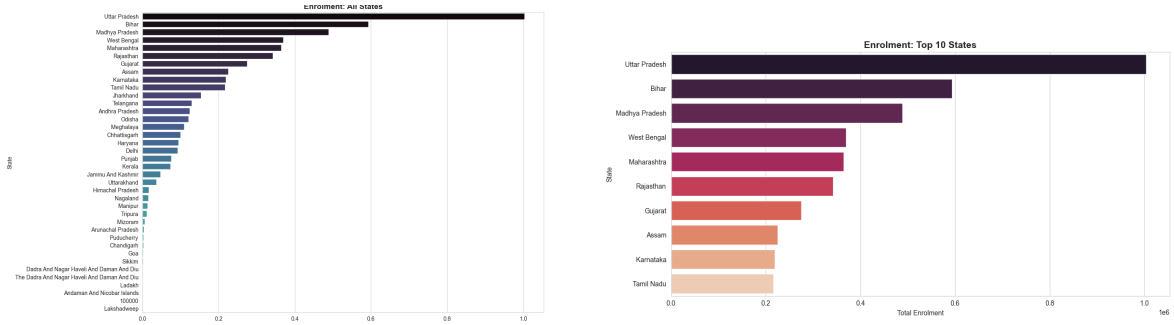


Figure 4: Demographic: Daily Updates Trend

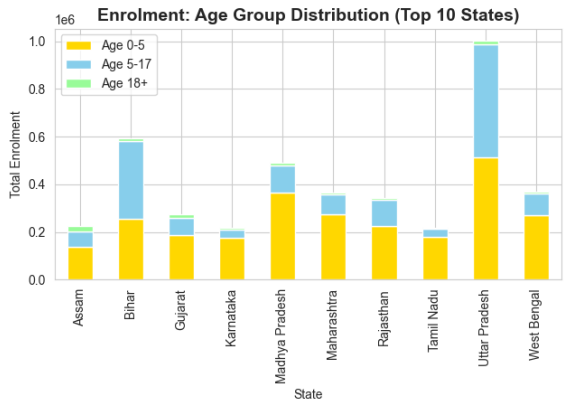
Demographic Updates Analysis



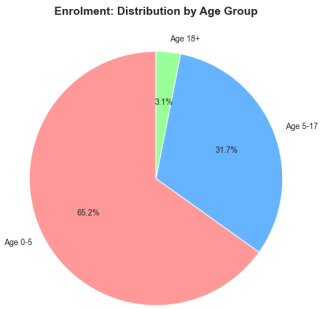
(a) Enrolment: All States

(b) Enrolment: Top 10 States

Figure 5: Enrolment Geographic Distribution



(a) Daily Trends by Age Group



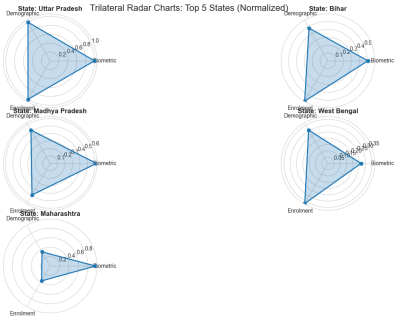
(b) Distribution by Age Group

Figure 6: Enrolment Age and Temporal Analysis

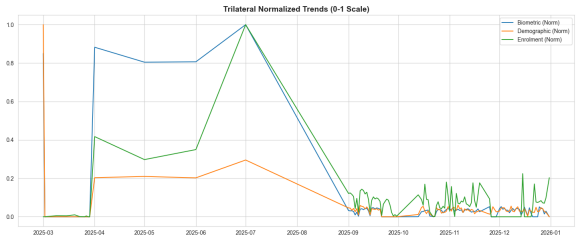
Enrolment Analysis

B. Bilateral and Trilateral Analysis

The Trilateral Analysis integrated all three datasets to provide a comprehensive view of the ecosystem. This included unified trends and normalized heatmaps to compare performance across states regardless of absolute population size.



(a) Trilateral Radar Chart (5 States)



(b) All States Normalized Heatmap

Figure 7: Holistic Trilateral Analysis

3.2 2. Key Findings: Top States by Category

Table 1: Top States for Biometric Updates (Includes fresh captures and mandatory updates)

Rank	State	Total Biometric 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

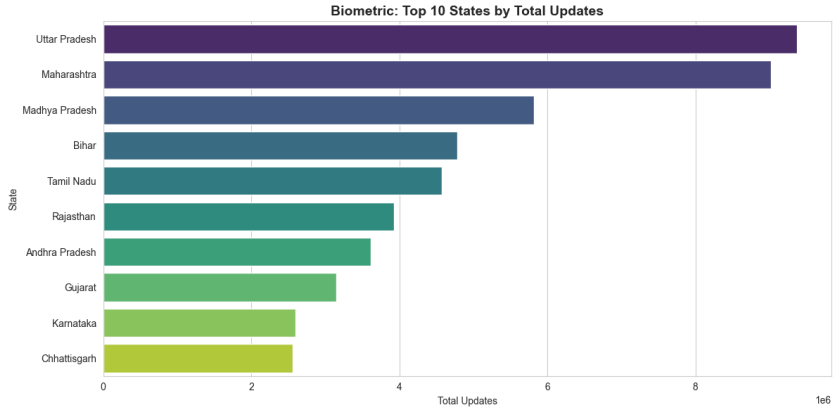


Figure 8: Biometric Top 10

Table 2: Top States for Demographic Updates (Includes updates to name, address, DOB)

Rank	State	Total Demographic 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

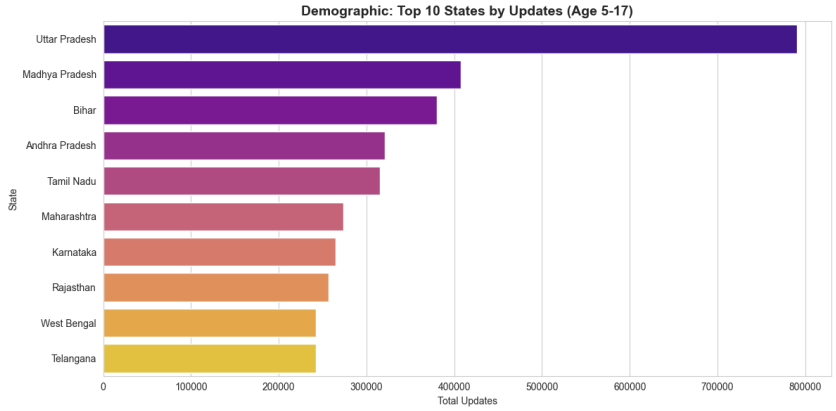


Figure 9: Demographic Top 10

Table 3: Top States for New Enrolments (New Aadhaar generations)

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

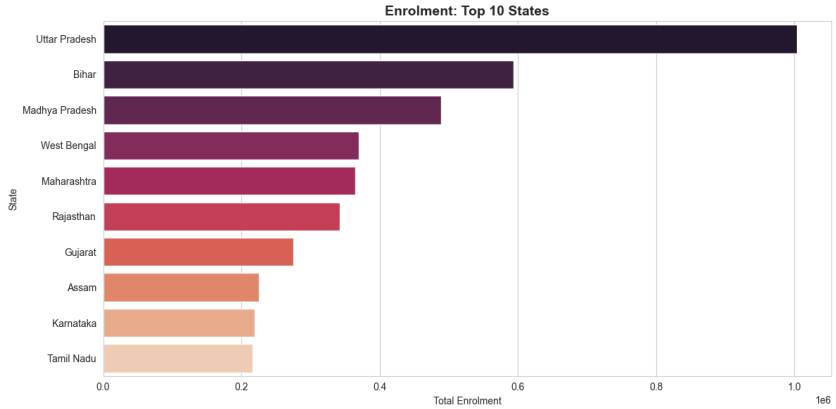


Figure 10: Enrollment Top 10

4 Case Study: Investigating Adult Enrolment Anomalies

4.1 Observation and Hypothesis

During the visualization phase, we identified an anomaly in the age distribution of new enrolments. While child enrolments (ages 0–5 and 5–17) followed expected growth curves, a persistent segment of approximately **3%** represented **Adult Enrolments (Age 18+)**. Given the high saturation of Aadhaar among the adult population, we hypothesized that this activity was not random but concentrated in specific "hotspot" regions.

4.2 Methodology: Geospatial Drill-Down

To investigate this, we developed a specialized Python module (`indiaenroll.py`) to perform granular geospatial analysis. Unlike standard state-level aggregation, this script isolates the `age_18_greater` metric and identifies hyper-local hotspots at the district and pincode levels.

4.3 Findings: The Border State Correlation

The analysis revealed a strong correlation between adult enrolment volumes and international border regions. The top contributing districts are primarily located in states sharing borders with Bangladesh, Nepal, and Pakistan.

Table 4: Top Districts for Adult Enrolment (Age > 18) indicating Border Region Activity

State	District	Geographic Context	Count (18+)
Meghalaya	East Khasi Hills	Bangladesh Border	9,841
Meghalaya	West Khasi Hills	Bangladesh Border	5,259
Meghalaya	West Garo Hills	Bangladesh Border	4,532
Karnataka	Bengaluru Urban	Urban Migration Hub	3,796
Bihar	Sitamarhi	Nepal Border	2,694
Gujarat	Banaskantha	Pakistan Border	1,629

4.4 Comprehensive State-Level Analysis

The complete state-level breakdown reveals the geographic distribution of adult enrolments and their relationship to border regions and migration hubs.

Table 5: Comprehensive State-Level Enrolment 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	DadraAndNagarHave (396230)
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)
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)

This data suggests that adult enrolments are likely driven by migration regularization or border-area demographic updates rather than standard organic growth.