

UIDAI DATA HACKATHON

Data-Driven Insights on Aadhaar Enrolment & Updates 2026

UIDAI_7370

Problem Statement

Unlocking Societal Trends in Aadhaar Enrolment and Updates Identify meaningful patterns, trends, anomalies, or predictive indicators and translate them into clear insights or solution frameworks that can support informed decision-making and system improvements.

PARTICIPANTS INFORMATION



RITIK SINGH
guycomo

NIT TRICHY
B.tech'22 - MECH



NEHAL MITTAL
nehalsmittal

NIT TRICHY
B.tech'22 - MME

Datasets Used

- Aadhaar Enrolment Data
- Aadhaar Demographic Data
- Aadhaar Biometric Update Data

10,06,029
20,71,700
18,61,108

ROWS

Jan 2025 - Dec 2025
Mar 2025 - Dec 2025
Mar 2025 - Dec 2025

Enrolment Analysis

- Merged multi-part enrolment files into one dataset
- Parsed dates and filtered valid 2025 records
- Standardised state names (spelling + UT mergers)
- Removed invalid and non-alphabetic locations
- Created total enrolment across age groups
- Analysed state-wise enrolment concentration
- Computed age-group shares (0-5, 5-17, 18+)
- Deep-dive district analysis for Uttar Pradesh

Demographic Analysis

- Combined all demographic update files
- Fixed age-column naming inconsistencies
- Removed duplicate transactional records
- Cleaned and validated state names
- Filtered low-frequency/invalid states
- Derived total demographic updates
- Compared age-group participation (5-17 vs 18+)
- Analysed state-wise and monthly trends

Bio - Update Analysis

- Concatenated biometric update datasets
- Corrected age-group column labels
- Parsed dates and removed duplicates
- Standardised state names and UT boundaries
- Derived total biometric update metric
- Compared top vs bottom states
- Analysed age-group balance
- Studied monthly and UP-level trends



TECH STACK



Python



GitHub



Google Collab



NumPy



Pandas



Matplotlib



Seaborn



Machine Learning



Glob



Prophet



Sarima

ADHAAR ENROLMENT ANALYSIS

Date | State | District | Pincode | Age 0-5 | Age 5-17 | Age > 18

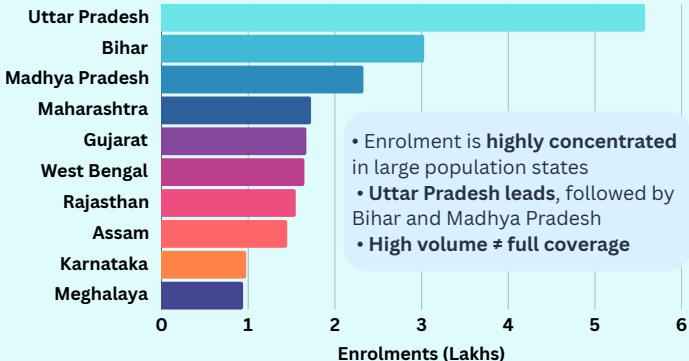
[Github Code Link](#)

4 Jan 2025 - 11 Dec 2025

Problem Statement

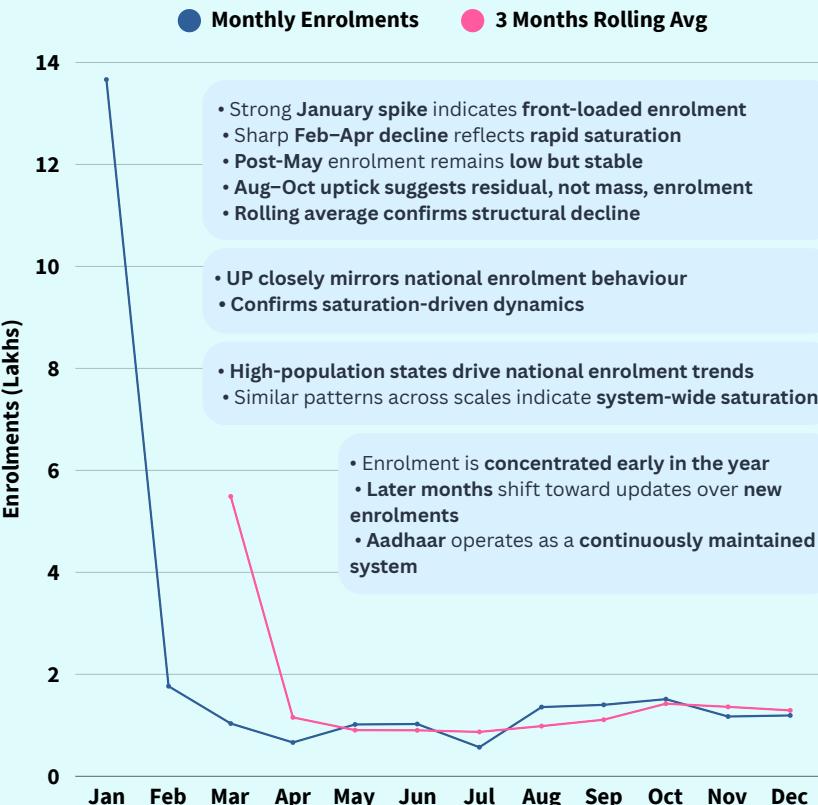
To analyse Aadhaar enrolment patterns during 2025 across states and age groups, identify regional and early-age enrolment gaps, and provide data-driven insights to support targeted planning and policy interventions by UIDAI.

Top 10 States/UTs by Adhaar Enrolment

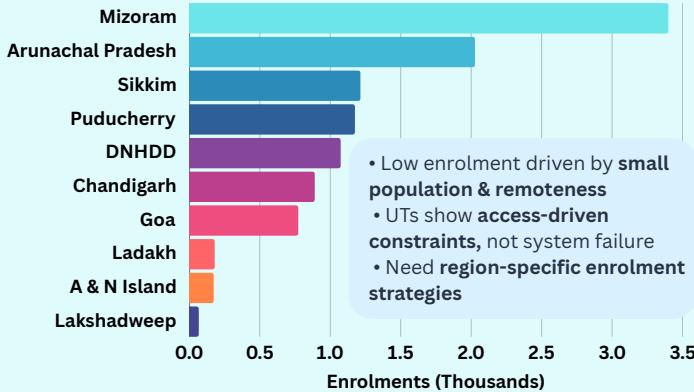


Monthly Enrolments Trends

of all States/UTs

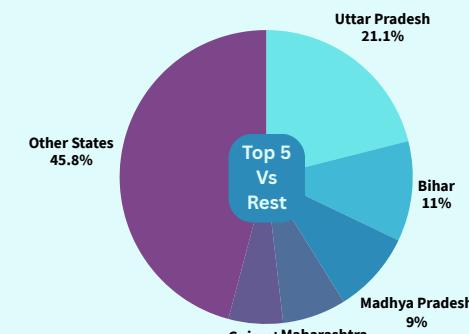


Bottom 10 States/UTs by Adhaar Enrolment



Share of Adhaar Enrolment

Top 5 States vs Rest



- Top 5 states contribute ~50% of total enrolments.
- Long tail of states forms the remaining share.
- Different strategies needed for high & low volume regions.

Recommendations for UIDAI

- Launch targeted child enrolment drives in low-share states.
- Integrate Aadhaar enrolment with birth registration systems.
- Use Anganwadi and school networks for early enrolment.
- Deploy mobile enrolment units in rural and underserved districts.
- Monitor child enrolment as a key KPI alongside total enrolment.

District-Level Insights – Uttar Pradesh

- Aadhaar enrolment within Uttar Pradesh is unevenly distributed across districts.
- A small number of districts contribute a disproportionately high share of total enrolments.
- Several districts show very low enrolment volumes, indicating possible access or awareness gaps.
- District-level variation highlights the importance of targeted operational planning.

Why this Matters?

- Aadhaar enables access to welfare, health, and education.
- Low child enrolment delays benefit access.
- Regional gaps require targeted intervention.



- Child enrolment share varies widely across high-enrolment states.
- Southern & western states show stronger early-age coverage.
- Some high-population states lag in child enrolment share.
- Meghalaya and parts of the Northeast show uneven patterns.
- Early-age enrolment remains a key improvement area.

CONCLUSION

This analysis highlights significant regional and age-wise disparities in Aadhaar enrolment. While overall enrolment is high in populous states, child enrolment remains uneven across regions. Targeted, early-age enrolment strategies can help UIDAI achieve more inclusive and future-ready Aadhaar coverage.

ADHAAR DEMOGRAPHIC ANALYSIS

Date | State | District | Pincode | Age 5-17 | Age > 17

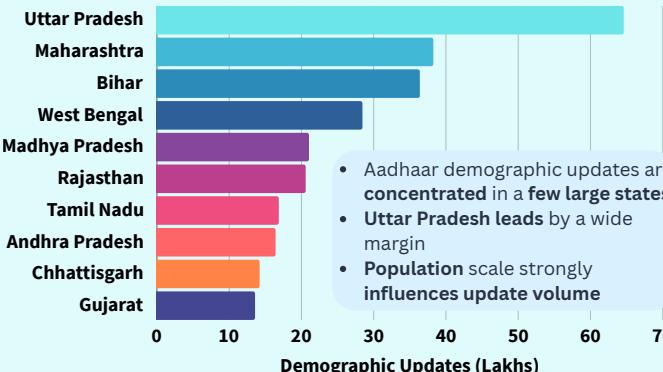
 [Github Code Link](#)

1 Mar 2025 - 29 Dec 2025

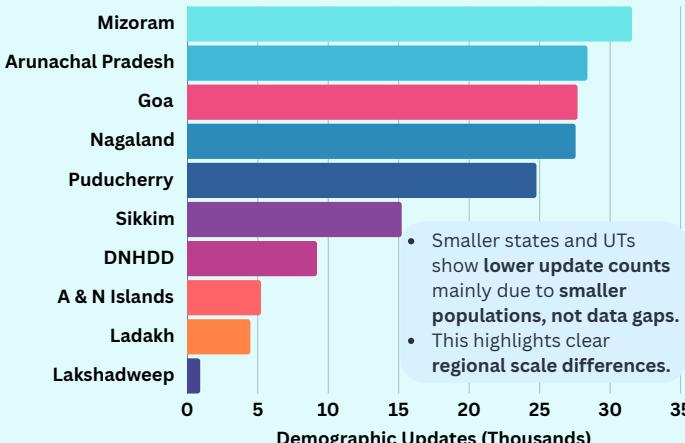
Problem Statement

This analysis examines Aadhaar demographic update patterns across states and age groups to identify regional and age-wise trends that support data-driven governance and service planning.

Top 10 States/UTs by Demographic Updates



Bottom 10 States/UTs by Demographic Updates

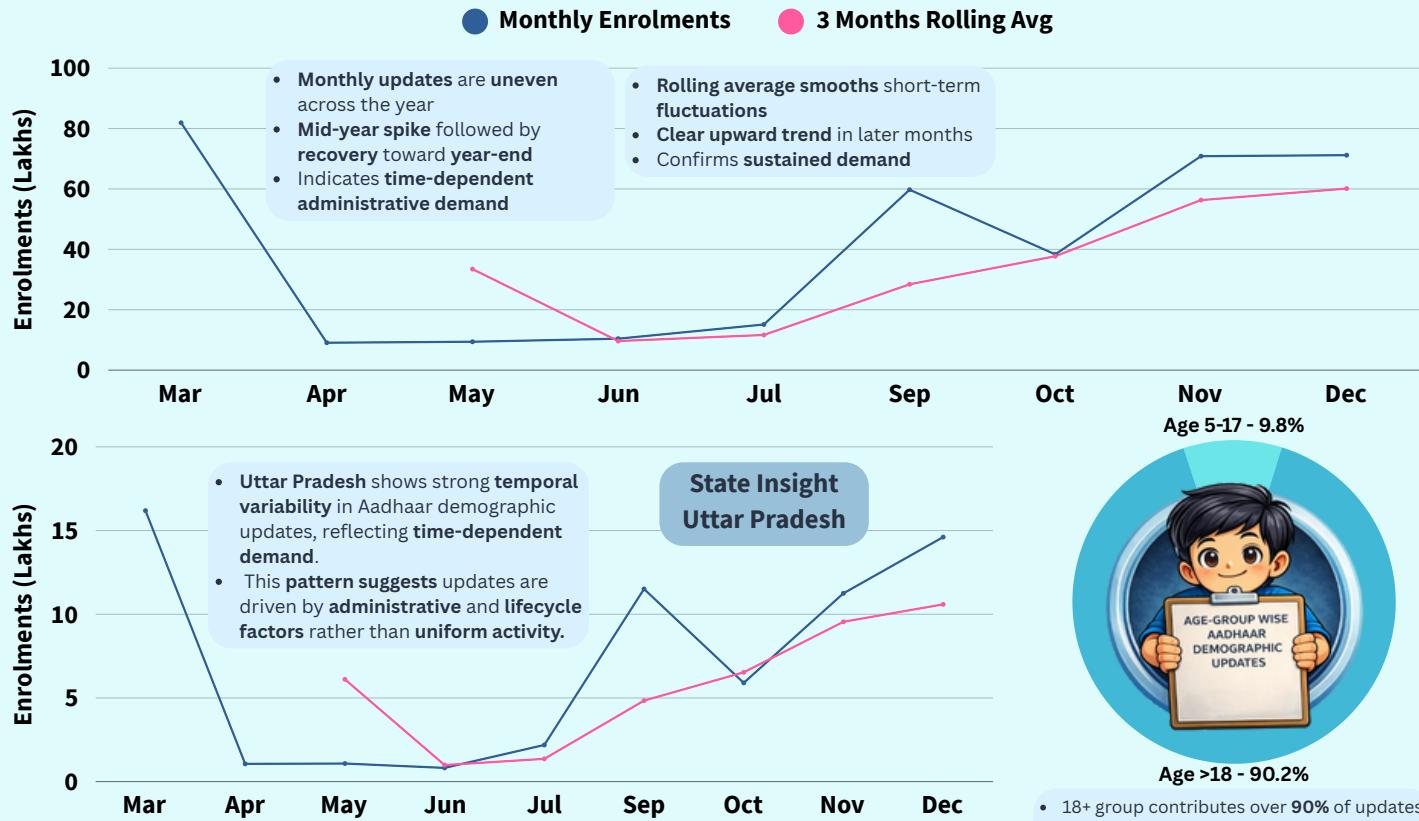


Limitations

- The dataset contains transaction-level records, where multiple updates by the same individual may occur.
- Location fields are partially free-text, leading to residual inconsistencies despite standardisation.
- The analysis does not link update trends to specific policies or events due to lack of external reference data.
- Population normalisation was not applied; results reflect absolute update volumes.

Monthly Demographic Update Trends

of all States/UTs



Conclusion

- This analysis examined Aadhaar demographic updates across states, age groups, and time using large-scale government data.
- Adults (18+) account for the majority of demographic updates, indicating that Aadhaar updates are primarily driven by post-enrolment lifecycle needs.
- A small number of populous states dominate overall update volumes, while smaller states and union territories show lower counts due to population scale.
- Monthly trends reveal non-uniform update activity, highlighting periods of increased administrative demand.
- Overall, Aadhaar demographic updates represent a continuous and dynamic lifecycle process, rather than a one-time administrative action.



Future Work

- The data is transaction-level, so multiple updates may belong to the same individual, and location fields retain minor inconsistencies.
- Results reflect absolute update volumes, as population normalization and policy-level linkage were not applied.

ADHAAR BIO-UPDATE ANALYSIS

Date | State | District | Pincode | Age 5-17 | Age > 17

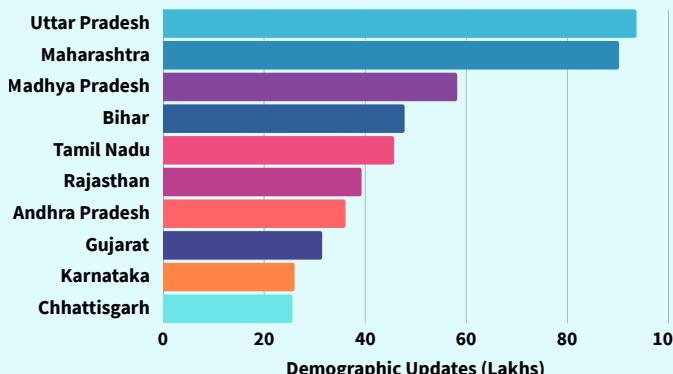
 [Github Code Link](#)

1 Mar 2025 - 29 Dec 2025

Problem Statement

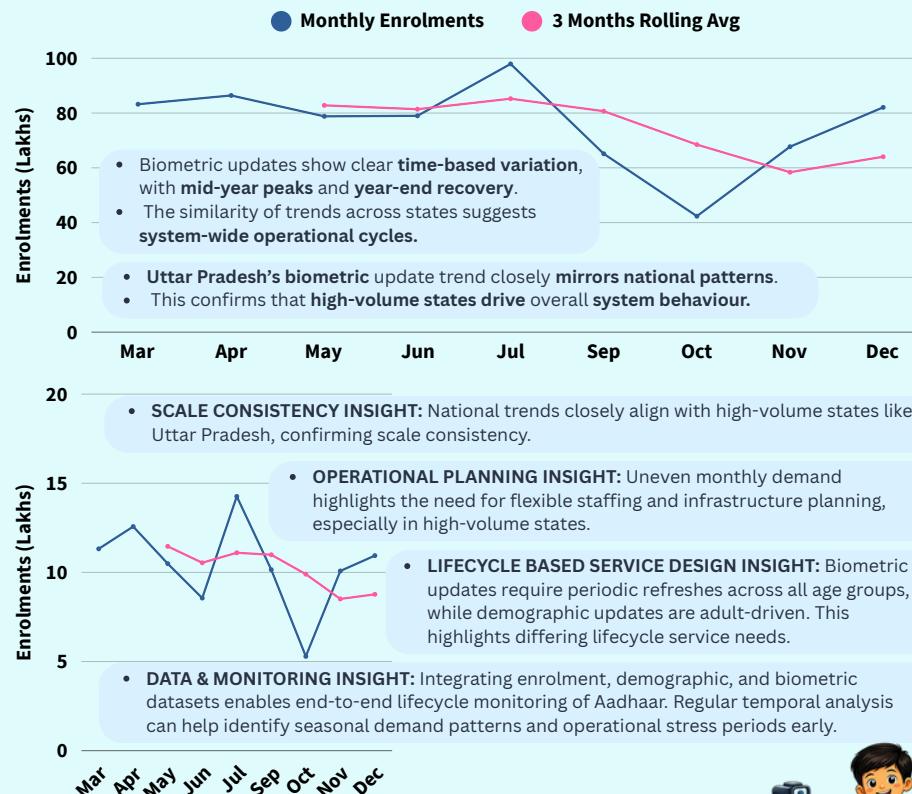
This project analyses the Aadhaar lifecycle using enrolment, demographic, and biometric datasets to identify spatial, age-wise, and temporal patterns, and understand how service demand varies across regions and over time.

Top 10 States/UTs by Biometric Updates

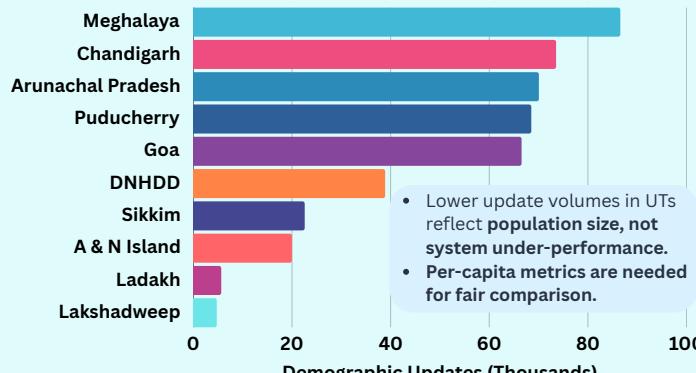


Monthly Enrolments Trends

of all States/UTs



Bottom 10 States/UTs by Biometric Updates



Limitations

- The datasets are transactional and may include multiple records per individual.
- Location fields contain free-text entries, leading to residual inconsistencies.
- Analysis is based on absolute volumes, without population normalisation.
- No direct causal inference is made regarding policies or external events.

Recommendations

- Adopt flexible operational planning based on observed monthly demand patterns.
- Consider periodic biometric refresh strategies across all age groups.
- Use per-capita metrics for more equitable regional comparisons.
- Integrate lifecycle datasets for continuous performance monitoring of Aadhaar services.

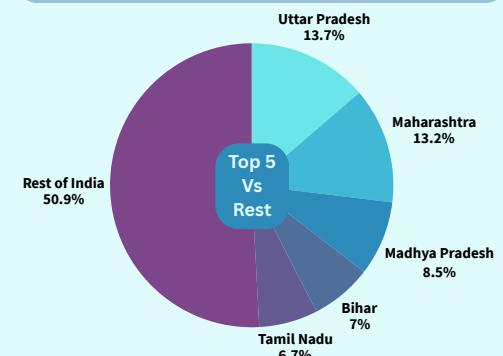
Conclusion



- Aadhaar is a dynamic, lifecycle-driven identity system, not a static one-time process.
- Enrolment ensures initial coverage, demographic updates track identity changes, and biometric updates maintain authentication accuracy.
- Integrated spatial, demographic, and temporal analysis highlights the need for continuous service availability, adaptive operations, and data-driven governance.

Share of Adhaar Enrolment

Top 5 States vs Rest



- Biometric updates are concentrated in a few large states, with Uttar Pradesh and Maharashtra contributing the highest volumes.
- Lower update counts in smaller states and UTs mainly reflect population size, not data gaps.

% Share of Age Group in Biometric Updates



Recommendations

- Aadhaar operates as a continuous identity lifecycle system across enrolment, demographic, and biometric updates.
- Updates ensure identity relevance and authentication accuracy over time.
- Temporal patterns highlight the need for ongoing, region- and time-specific operational planning.

ADHAAR UPDATE DEMAND FORECAST

Time Series Forecast

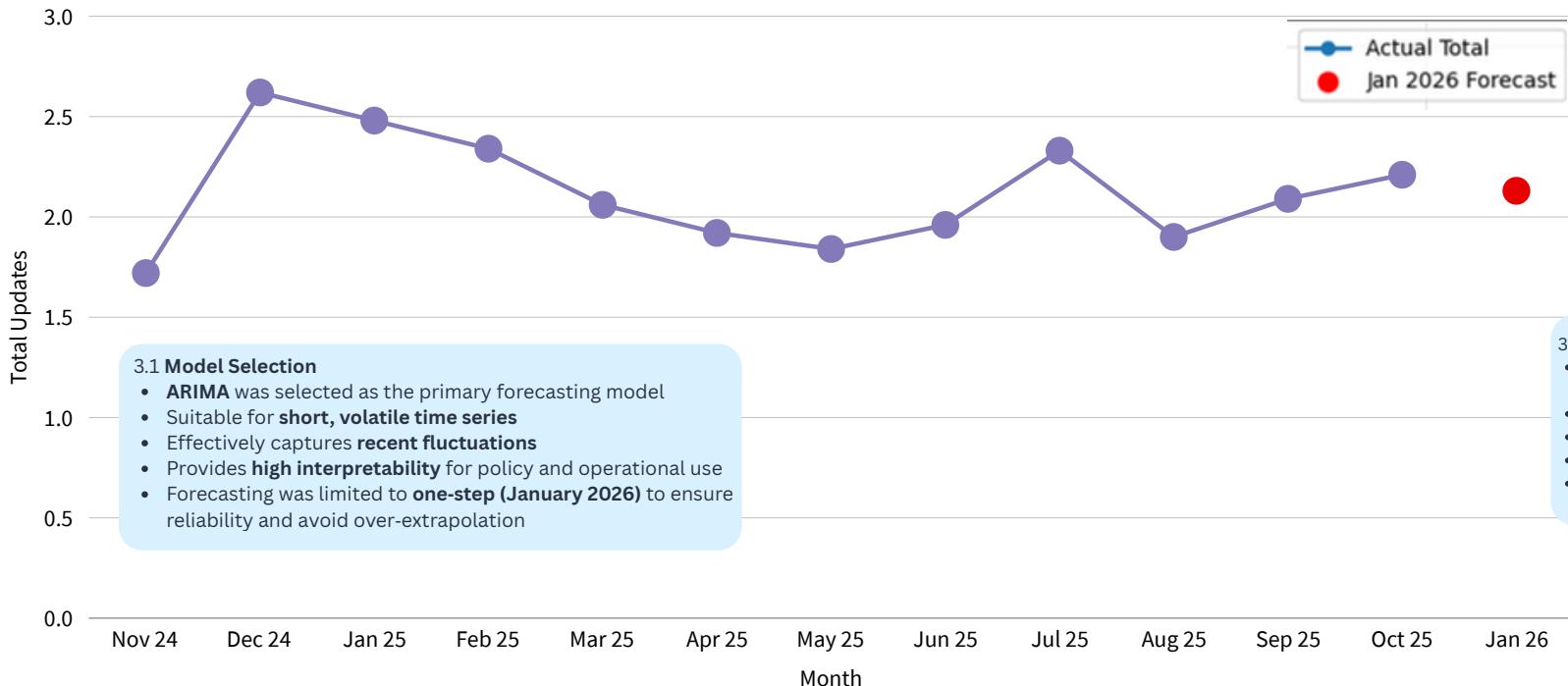
Github Code Link

January 2026 Forecast

Problem Statement

The objective of this study is to analyze historical Aadhaar update trends and generate a reliable short-term forecast to support near-term operational planning. Given the public-service nature of Aadhaar updates, the focus is on interpretability, data integrity, and conservative estimation, rather than long-term speculative forecasting.

Forecasting Approach



3.1 Model Selection

- ARIMA was selected as the primary forecasting model
- Suitable for short, volatile time series
- Effectively captures recent fluctuations
- Provides high interpretability for policy and operational use
- Forecasting was limited to one-step (January 2026) to ensure reliability and avoid over-extrapolation

Data Preparation

Monthly Aadhaar update data was analyzed for two independent components:

- Demographic Updates
- Biometric Updates

The datasets were preprocessed to ensure:

- Consistent monthly granularity
- Removal of duplicate month entries via aggregation
- Proper datetime indexing (month-end frequency)

To avoid scale inconsistency and artificial shocks, aggregate "total updates" figures derived from different reporting pipelines were not directly modeled. Instead, demographic and biometric series were treated as independent, internally consistent time-series.

Data Preparation

- Monthly Aadhaar updates were analyzed separately for demographic and biometric components
- Data was aggregated to a consistent monthly frequency
- Proper month-end datetime indexing was applied
- Pre-aggregated total updates were not modeled to avoid scale distortions
- Demographic and biometric series were treated as independent, internally consistent time series

3.2 Baseline Comparison

- Holt's Exponential Smoothing was used as a baseline to validate the ARIMA forecast
- Holt produced a slightly lower estimate than ARIMA
- Both forecasts were closely aligned, indicating consistency
- This agreement increases confidence in the short-term forecast
- ARIMA was retained as the primary model due to better responsiveness to volatility

3.3 Final Forecast Strategy

- Demographic and biometric updates were forecast independently
- The total Aadhaar update demand for January 2026 was computed as the sum of both component forecasts
- Only one forecast point was visualized to avoid interpretational ambiguity

This approach ensures clarity, methodological correctness, and alignment with real-world planning needs.

Limitations & Future Work

- Forecasting is limited by short historical depth
- External policy variables were not explicitly incorporated
- Long-term forecasts were intentionally avoided to preserve reliability

Future enhancements may include:

- ARIMAX models with structured policy indicators
- Scenario-based forecasting
- Integration of regional operational metadata

Conclusion

The adopted approach emphasizes clarity and robustness by combining interpretable time-series forecasting with exploratory clustering, enabling reliable short-term demand estimation and actionable insights for Aadhaar operations and governance.

JAN 2026 DEMOGRAPHIC : 11858753

JAN 2026 BIOMETRIC : 9439714

JAN 2026 TOTAL : 21298467

STATE LEVEL ADHAAR OPERATIONAL CLUSTERING

Problem Statement

National Aadhaar statistics mask regional differences in demand, stability, and growth. A one-size-fits-all operational strategy is insufficient at national scale, creating the need for data-driven analysis to identify regional behavior, high-risk states, and support proactive governance, capacity planning, and policy decision-making.

Objectives

1. Forecast short-term Aadhaar update demand using historical data.
 2. Identify patterns and anomalies in Aadhaar activity across states.
 3. Group states with similar operational behavior using unsupervised learning.
 4. Provide interpretable and governance-oriented insights, rather than black-box predictions.
 5. Demonstrate scalable data engineering and analytical rigor on real-world datasets.

Dataset Description

The analysis uses three primary datasets provided for the hackathon.

3.1 Enrolment Dataset

- Aadhaar enrolments by date, state, district, and pincode
 - Age-wise enrolment distribution

- Captures the entry point of Aad

- Demographic updates (name, address, DOB, etc.)
 - Age-segmented counts

- Represents identity maintained

- Biometric updates (fingerprint, iris)
 - Age-segmented counts
 - Represents biometric quality and re-verification demand

All datasets span multiple years and contain millions of records, requiring careful preprocessing and aggregation.

Data Preprocessing and Feature

4.1 Cleaning & Standardization

- Converted date fields to datetime format
 - Standardized state names to avoid duplication
 - Removed inconsistent casing and formatting issues
 - Handled missing values through aggregation-aware imputation

4.2 Aggregation Strategies

- Data was aggregated to state-month level
 - Final analysis used state-level summaries

4.3 Engineered Features For each state, the following features were computed:

- Total Enrolment
 - Total Demographic Updates
 - Total Biometric Updates
 - Demo Ratio (demographic share of updates)
 - Bio Ratio (biometric share of updates)
 - Volatility (month-to-month instability in updates)
 - Growth Rate (change in demand over time)

These features capture scale, composition, stability, and trend – key dimensions of operational behavior.

State Level Clustering (Unsupervised Learning)

6.1 Why Clustering?

Forecasting alone does not explain why demand differs across regions. Clustering enables discovery of latent behavioral patterns without predefined labels.

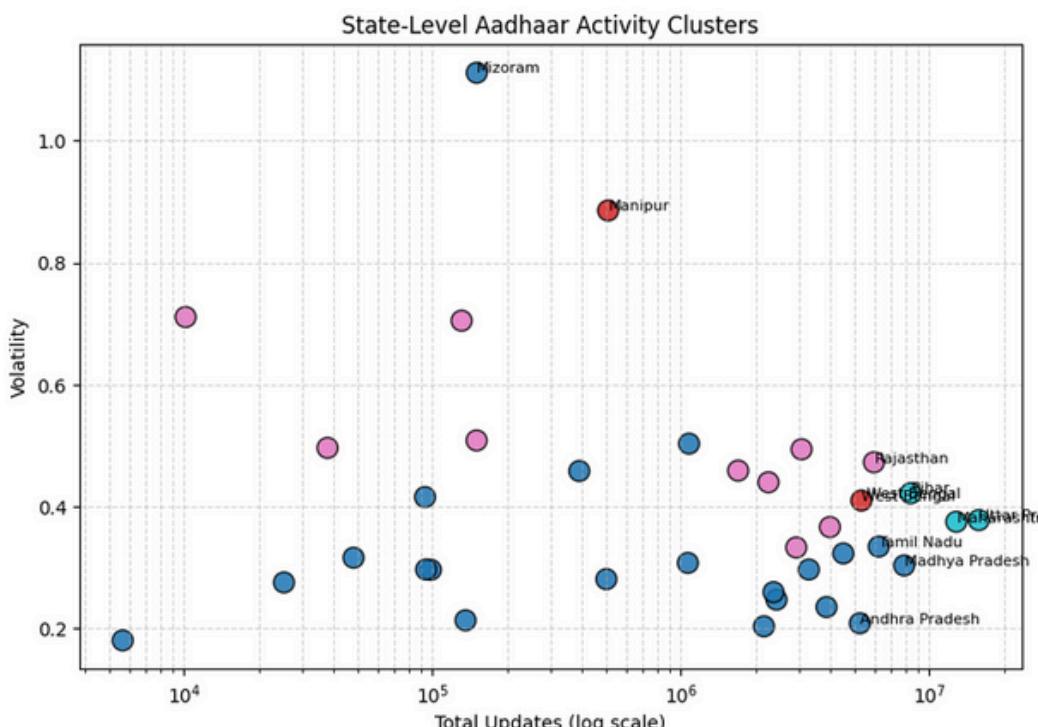
6.2 Methodology

- Features were standardized using StandardScaler
 - K-Means clustering was applied
 - Number of clusters selected for interpretability and separation

6.3 Visualization Strategy

- X-axis: Total Updates (log scale)
 - Y-axis: Volatility
 - Color: Cluster membership

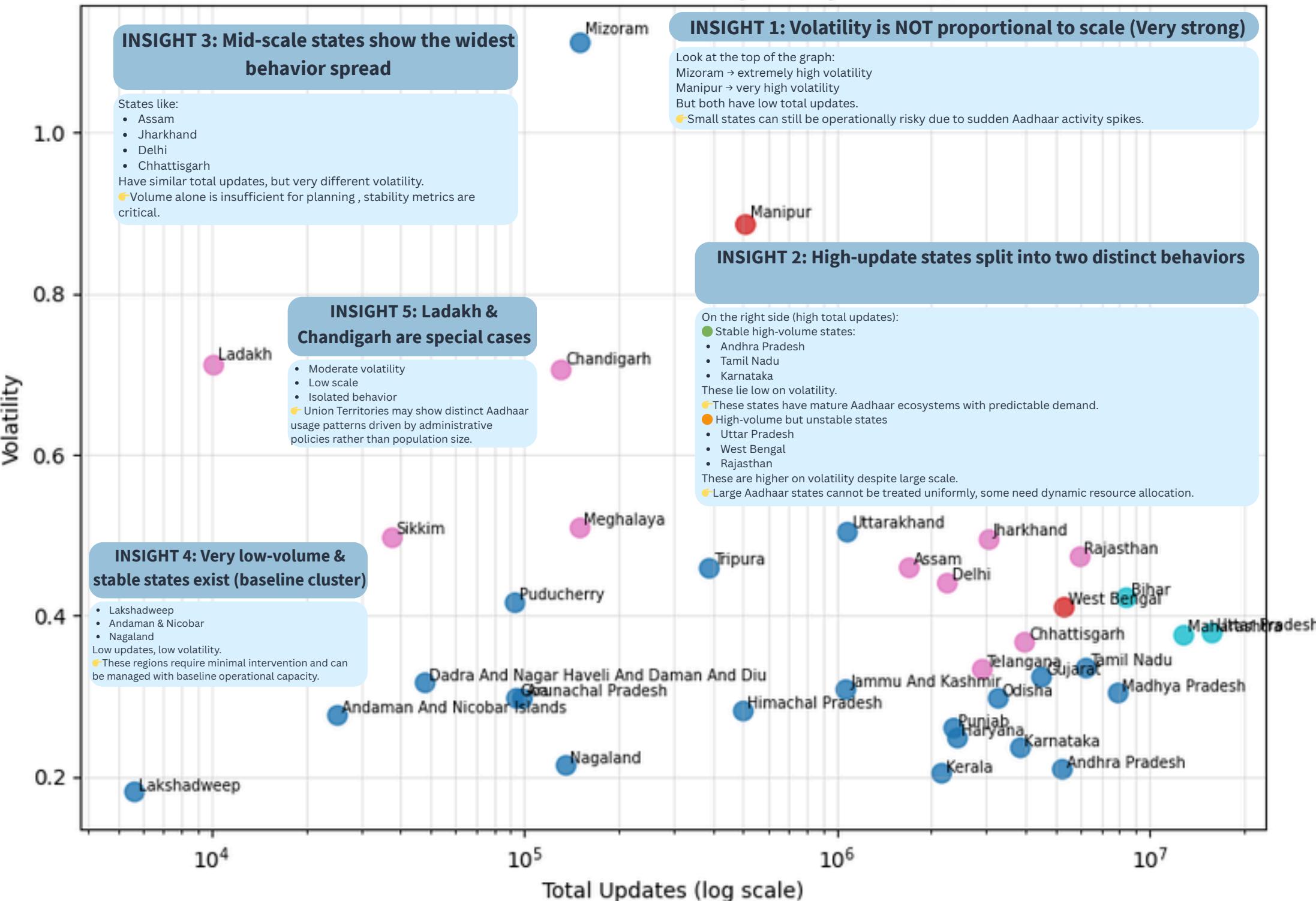
This visualization highlights operational load vs instability, a critical governance dimension.



Conclusion

1. This project demonstrates how large-scale Aadhaar operational data can be transformed into actionable intelligence through a combination of time-series forecasting and unsupervised clustering.
 2. Rather than relying solely on aggregate statistics or black-box models, the approach emphasizes:
 3. Interpretability
 4. Operational relevance
 5. Real-world data challenges
 6. By identifying distinct state-level behavior patterns, the system enables data-driven decision-making, supporting scalable, efficient, and region-aware Aadhaar governance.

State-Level Aadhaar Activity (Analysis View)

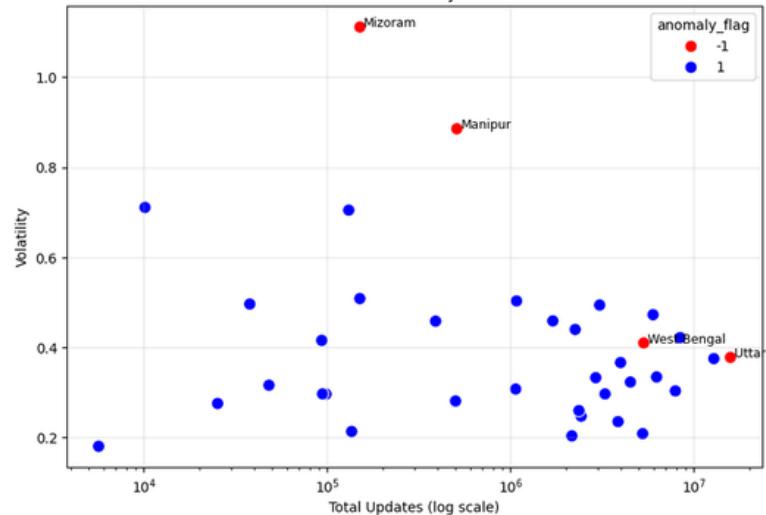


STATE LEVEL ANOMALIES

Problem Statement

This notebook identifies anomalous state-level Aadhaar update behavior and localized district- and pincode-level hotspots to support targeted operational planning.

State-Level Anomaly Detection



Insight 3: High-volume ≠ high volatility

- Some pincodes have:
- Very high total_updates
- Moderate volatility
- 👉 High demand regions are not necessarily unstable; instability is localized.

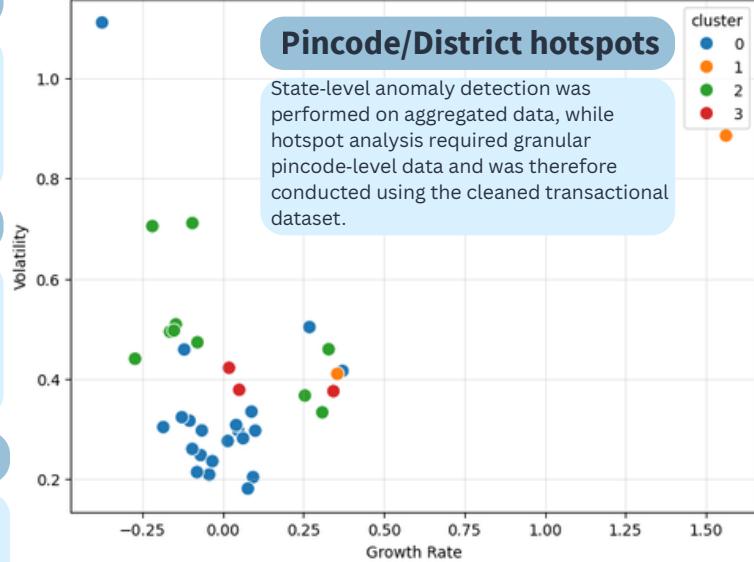
Insight 4: Zero-activity pincodes are

- Especially in:
- Rural
- Newly created
- Administrative pincodes
- 👉 Zero-activity pincodes reflect structural or demographic factors, not data issues.

Insight 1: Urban concentration

- Many top hotspots are:
 - Delhi (1100xx)
 - Uttar Pradesh (24xxxx, 20xxxx)
 - Maharashtra (43xxxx, 42xxxx)
- 👉 Aadhaar update demand is highly concentrated in dense urban and peri-urban pincodes.

Growth vs Volatility by Cluster

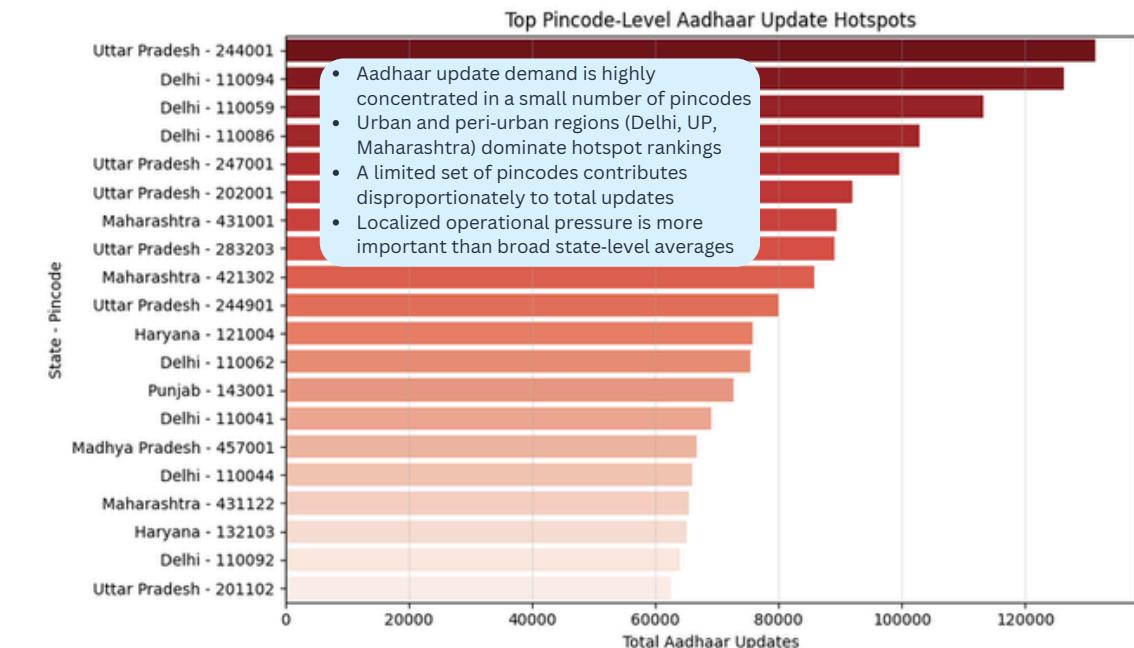


Pincode/District hotspots

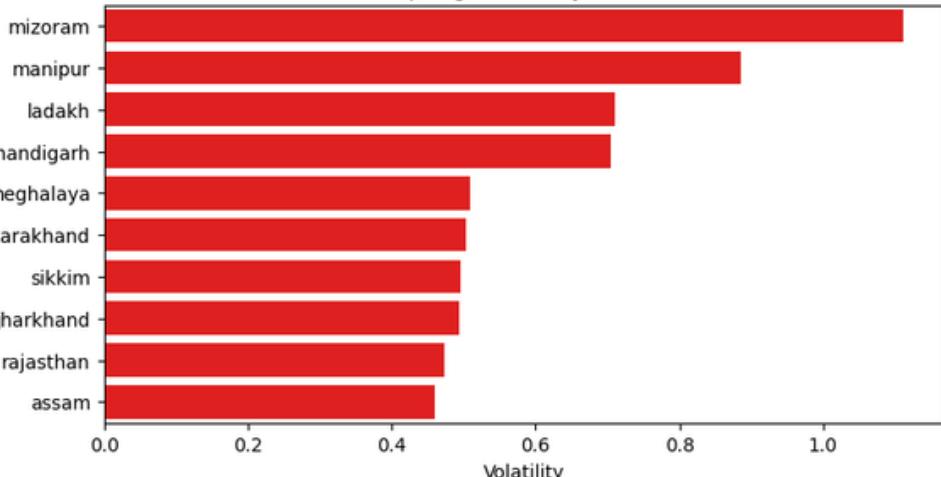
State-level anomaly detection was performed on aggregated data, while hotspot analysis required granular pincode-level data and was therefore conducted using the cleaned transactional dataset.

Insight 2: Few pincodes dominate volume

- Top ~20 pincodes contribute lakhs of updates, while most pincodes contribute very little.
- 👉 A small fraction of pincodes accounts for a disproportionate share of total Aadhaar activity.



Top High-Volatility States



CONCLUSION

- **Multi-Layered Forecasting**
- Utilizes time-series models to project short-term demand.
- Facilitates near-term capacity planning for manpower and hardware allocation.
- **State-Level Operational Clustering**
- Groups states by behavioral patterns (stability, growth, and update types).
- Identifies regional differences in how citizens interact with the Aadhaar system.
- **Anomaly Detection & Risk Assessment**
- Flags states with unusually high volatility or outlier growth rates.
- Designates "operational risk zones" to preempt system bottlenecks or service failures.
- **Granular Hotspot Analysis**
- Performs deep dives at district and pincode levels.
- Reveals spatial concentration, where a few geographic units drive the majority of total activity.
- **Strategic Governance Shift**
- Moves away from uniform national planning toward targeted, data-driven interventions.
- Redefines operational risk as a combination of scale, volatility, and location.