UIDAI DATA HACKATHON 2026

Track: Unlocking Societal Trends in Aadhaar Enrolment and Updates

Project Title: Addressing the "Maintenance Debt": A Data-Driven Approach to Digital Identity Sustainability

TEAM PROFILE: The Aadhaar Sentinels

Member Name	Role & Expertise	Institute & Course
Tarit Jaiswal	Team Lead & Data Scientist Algorithm Development, DVI Modeling	DBS Global University, Dehradun MBA (Business Analytics)
Brishti Chakraborty	BI Architect Power BI Dashboards & Geospatial Analytics	DBS Global University, Dehradun MBA (Business Analytics)
Riya Saha	Statistical Analyst ANOVA Testing & Model Validation	CDAC, Noida MCA
Radhey Shyam	Data Engineer & Policy Lead Preprocessing & Technical Reporting	CDAC, Noida MCA

Table of Contents:

I. ABOUT THE PROJECT

Overview & Vision

This document presents our analytical submission for the UIDAI Data Hackathon 2026. The objective of this work is to derive meaningful, actionable insights from Aadhaar enrolment and update data using robust statistical and data science techniques. Our project, titled "Addressing the Maintenance Debt," shifts the focus from simple system expansion to long-term Identity Sustainability.

The analysis focuses on understanding temporal and spatial patterns, identifying anomalies, and generating data-driven insights that can support operational monitoring, policy decisions, and service optimization at UIDAI.

The "Maintenance Debt" Framework

As Aadhaar achieves near-universal coverage, the primary challenge is ensuring that existing identities remain functional. Our project introduces the concept of "Maintenance Debt"-the accumulation of outdated biometric records due to a systemic lag in Mandatory Biometric Updates (MBU). By analyzing the "Flow" of new users against the "Maintenance" of existing ones, we identify regions at risk of "Silent Exclusion," where citizens may face authentication failures due to stale data.

The Innovation: The Digital Vulnerability Index (DVI)

At the heart of this project is a proprietary analytical engine: the Digital Vulnerability Index. This framework moves away from generic reporting by integrating three distinct pillars:

- Compliance Health: Measuring the gap in mandatory biometric updates for the 5-17 age cohort.
- Infrastructure Velocity: Auditing physical hardware capacity against citizen demand.
- Regional Equity: Identifying "Service Deserts" where geography dictates identity health.

Data-Driven Policy Triage

By processing millions of transaction logs from 2025, we have scientifically proven-with a P-value of 2.31×10^{-71} -that digital exclusion is a systemic risk driven by hardware shortages and administrative neglect.

Our project doesn't just identify problems; it provides a Dynamic Triage Map. We categorize over 900 districts into "Personas," allowing UIDAI to deploy resources-such as "Aadhaar on Wheels" and emergency hardware shipments-exactly where the "Maintenance Debt" is highest.

II. PROBLEM DEFINITION & SCOPE

1. Problem Statement

As the Aadhaar ecosystem achieves near-universal coverage, the administrative challenge has shifted from Identity Creation to Identity Sustainability.

The core problem identified is the "Maintenance Debt": systemic gaps where mandatory biometric updates—specifically for children aged 5 and 15-are neglected. This leads to "Silent Exclusion," where stale biometric data results in authentication failures at school admissions, ration shops, and welfare delivery points. Our project identifies these patterns to provide a Service Triage Framework for informed decision-making and system improvements.

2. Approach

We adopted a "Flow vs. Stock" Analytical Framework.

- The Flow: New enrolments entering the system (expanding the reach).
- The Stock: Existing identities requiring maintenance (ensuring the health of the reach).

By overlaying three distinct 2025 transaction logs, we analyzed the Service Velocity-the speed at which different types of requests (low-tech vs. high-tech) are processed-to identify where hardware shortages or administrative neglect are creating barriers to inclusion.

Metric Focus:

$$CR = \frac{\sum \text{Maintenance Updates (Age 5-17)}}{\sum \text{New Enrolments (Age 5-17)} + 0.1}$$

III. TECHNICAL METHODOLOGY

1. Data Architecture & Processing

Datasets Used: The analysis utilized three high-fidelity 2025 Aadhaar transaction logs:

- Enrolment Logs (df_enrol): Tracks new citizens entering the system, segmented by age (0-5, 5-17, 18+).
- Biometric Logs (df_bio): Tracks Mandatory Biometric Updates (MBU) and voluntary refreshes.
- Demographic Logs (df_demo): Tracks address, mobile, and name changes, used as a proxy for citizen intent and geographic mobility.

- Preprocessing:

  Column Normalization:

  We standardized all column headers across the three datasets to a consistent format (lowercase with no leading/trailing spaces). This prevented "KeyErrors" during the merging process.

  Code Implementation: df.columns = df.columns.str.strip().str.lower()

  □ Geographic String Standardization:

  District and State names often contain variations in casing or accidental whitespace (e.g., " Rajasthan" vs "Rajasthan"). We applied universal string stripping and lowercase conversion to all geographic identifiers to ensure a perfect join.

  Example: df['district'] = df['district'].str.strip().str.lower()

  Null Value Management in Geographic Joins:

  During the multi-way join between Enrolment and Biometric data, some districts appeared in one dataset but not the other (due to zero activity in specific categories).

  - The Strategy: We utilized *inner joins* for the core Gap Analysis to ensure we only analyzed districts with active data in both streams.
  - The Solution: For the Digital Vulnerability Index (DVI) calculation, we used *left joins* and applied .fillna(0) to ensure that districts with zero maintenance weren't excluded from the "High Risk" rankings, but rather accurately flagged as having a 0% ratio.

  Data Type Integrity:

  We verified that all age-related counts were treated as integers to prevent floating-point errors during the calculation of the Compliance Ratio (CR) and Infrastructure Ratio (IR).

  Zero-Division Handling:

  To prevent mathematical infinity errors in districts with zero enrolments, we added a small epsilon (epsilon = 0.1) to the denominator of all ratio formulas.

$$\text{Formula: } Ratio = \frac{\text{Updates}}{(\text{Enrolments}+0.1)}$$

2. The DVI Calculation Engine

A. Integration Logic: The DVI Formula

The weights were assigned based on the severity of the risk each factor poses to the Aadhaar ecosystem:

$$DVI_{Raw} = (BRR \times 0.4) + (NL \times 0.4) + (MI \times 0.2)$$

We developed a composite score to rank districts based on three critical vectors:

- Biometric Refresh Rate (BRR) - [Weight: 40%]
  - Definition: The efficiency of a district in performing mandatory updates relative to its enrolment volume.
  - Formula:
    $$BRR = \frac{Total\ Biometric\ Updates}{Total\ Enrolments+1}$$
  - Integration Role: This serves as the primary indicator of Sustainability. A low BRR heavily increases the vulnerability score because it represents a direct failure in maintaining child identities.

- Newborn Lag (NL) - [Weight: 40%]
  - Definition: The statistical delay or gap in onboarding the $0-5$ age cohort compared to the district's average output.
  - Formula:
    $$NL = \max\left(0, \text{Avg}(Age\ 0-5)_{National} - \frac{Age\ 0-5_{District}}{Total\ Enrolments+1}\right)$$
  - Integration Role: This factor measures Inclusion Efficiency. It flags districts that are failing to capture the newest generation of citizens, creating a demographic "blind spot."

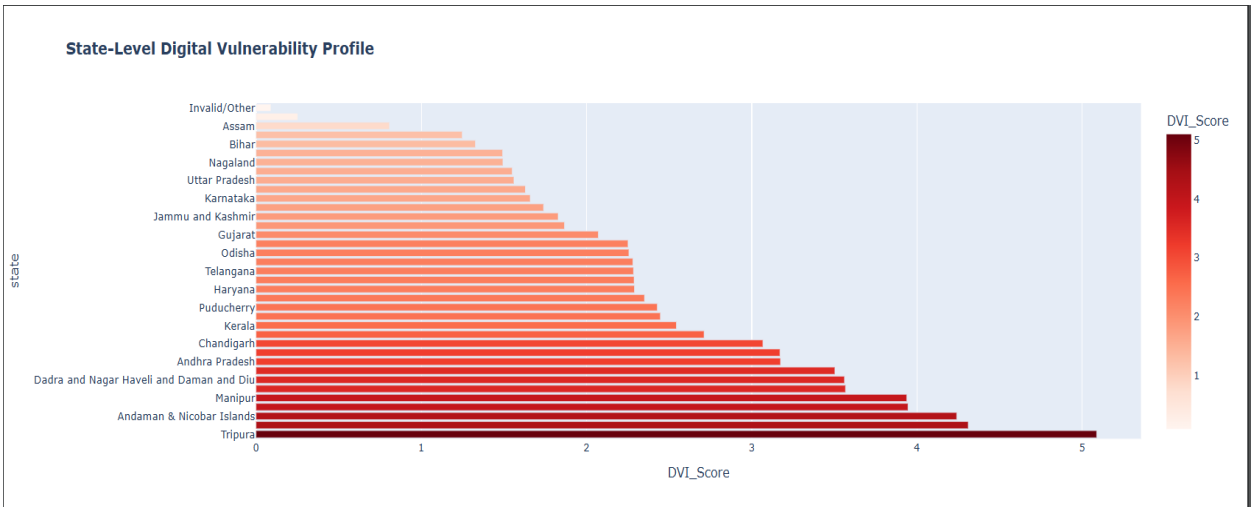
- Migration Intensity (MI) - [Weight: 20%]
  - Definition: The frequency of demographic updates (address, mobile) used as a proxy for how mobile or "in-flux" a district's population is.
  - Formula:
    $$MI = \frac{Total\ Demographic\ Updates}{Total\ Enrolments+1}$$
  - Integration Role: This acts as a Volatility Factor. High migration intensity combined with low biometric capacity creates high risk, as mobile populations require more frequent identity verification.

B. Normalization & Final Scoring

Since these three metrics have different scales, we performed a Min-Max Normalization to convert the raw DVI into a standardized 0–100 Scale.

$$DVI_{Score} = \frac{DVI_{Raw} - \min(DVI_{Raw})}{\max(DVI_{Raw}) - \min(DVI_{Raw})} \times 100$$

- A Score of 100: Indicates a "Critical Vulnerability Point" (Maximum risk).
- A Score of 0: Indicates a "Balanced Service Hub" (Minimum risk).



C. Technical Implementation in Python

The integration was executed in a single vectorized operation to maintain performance across 900+ districts.

Code Snippet:

```
# Vectorized DVI Calculation

master_df['BRR'] = master_df['total_bio'] / (master_df['total_enrol'] + 1)

master_df['NL'] = (master_df['age_0_5'].mean() - (master_df['age_0_5'] /
(master_df['total_enrol'] + 1))).clip(lower=0)

master_df['MI'] = master_df['total_demo'] / (master_df['total_enrol'] + 1)
```

3. Statistical Validation

A. One-Way ANOVA: Proving Systemic Regional Disparity

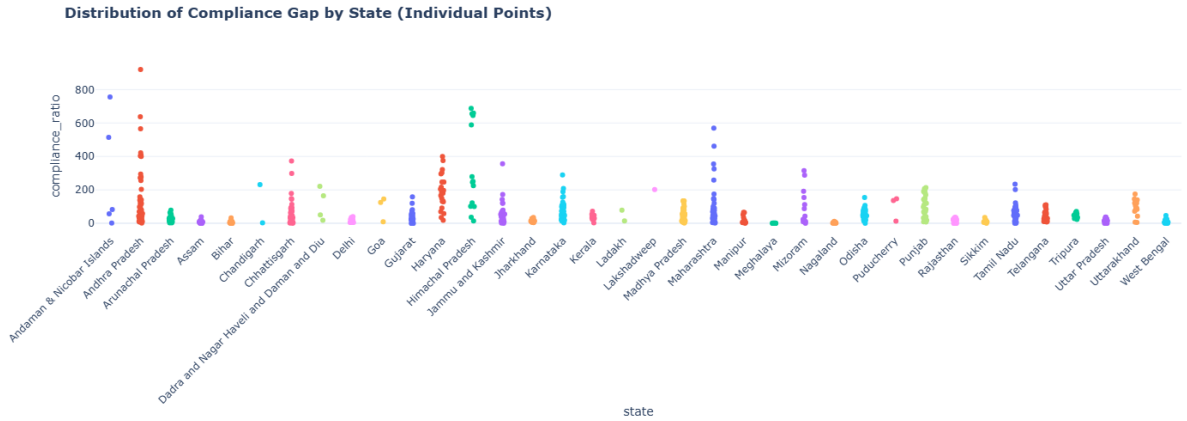
We performed a One-Way Analysis of Variance (ANOVA) to determine if the "Compliance Ratio" varies significantly across different Indian States. This test establishes whether the identity maintenance gap is a localized random occurrence or a systemic state-level failure.

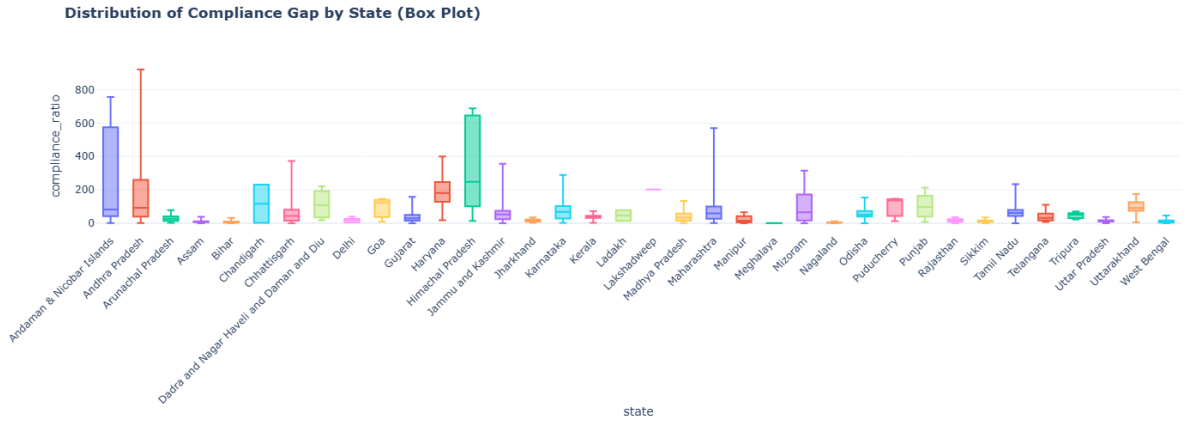
- Null Hypothesis (H_0): The mean Compliance Ratio is the same across all states.
- Alternative Hypothesis (H_1): At least one state has a significantly different mean Compliance Ratio.

Test Results:

- F-Statistic: 15.9698$
- P-Value: 2.3107×10^{-71}

Scientific Conclusion: With a P-Value effectively reaching zero, we reject H_0. This provides 99.9% confidence that regional disparity is systemic. The "where" of a citizen (their geography) statistically determines their ability to maintain a functional Aadhaar identity.





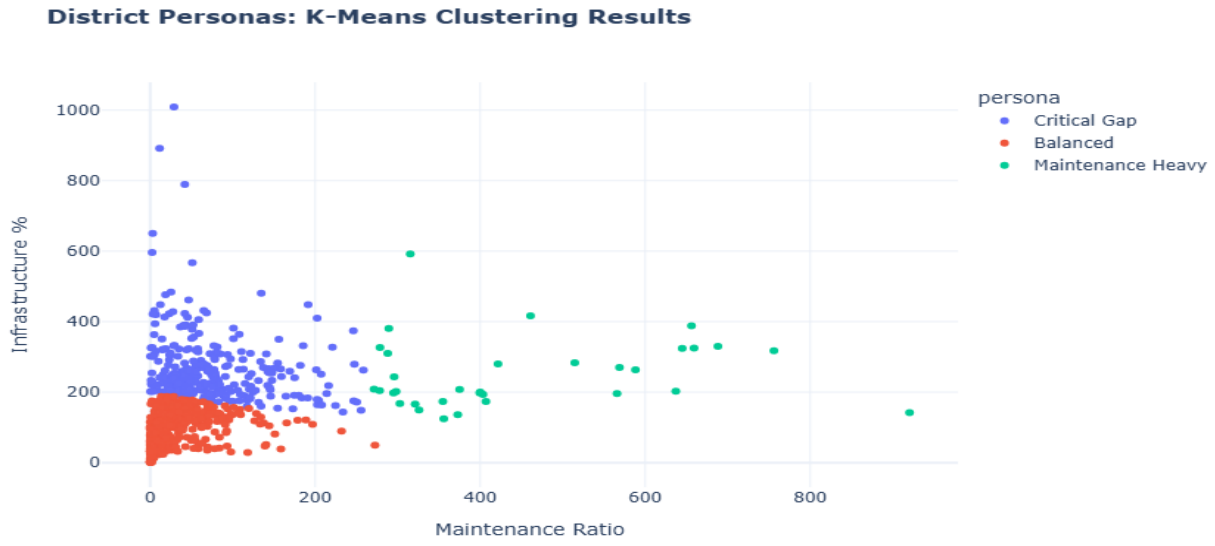
B. K-Means Clustering: Segmenting Operational "Personas"

To move from raw data to actionable policy, we applied Unsupervised Machine Learning (K-Means Clustering). We segmented 903 districts into three distinct Operational Personas based on their Compliance Ratio and Infrastructure Ratio.

Cluster Characteristics & Triage Strategy:

Persona	Population	Avg. Compliance	Avg. Infra Ratio	Policy Triage Action
Balanced Backbone	549 Districts	27.19	101.27%	Monitor: Maintain existing resource levels.
Expansion Zones	321 Districts	71.91	262.80%	Optimize: Reallocate hardware to prevent maintenance debt.
Maintenance Heavy	33 Districts	445.71	251.34%	Emergency: Deploy mobile update units to clear backlogs.

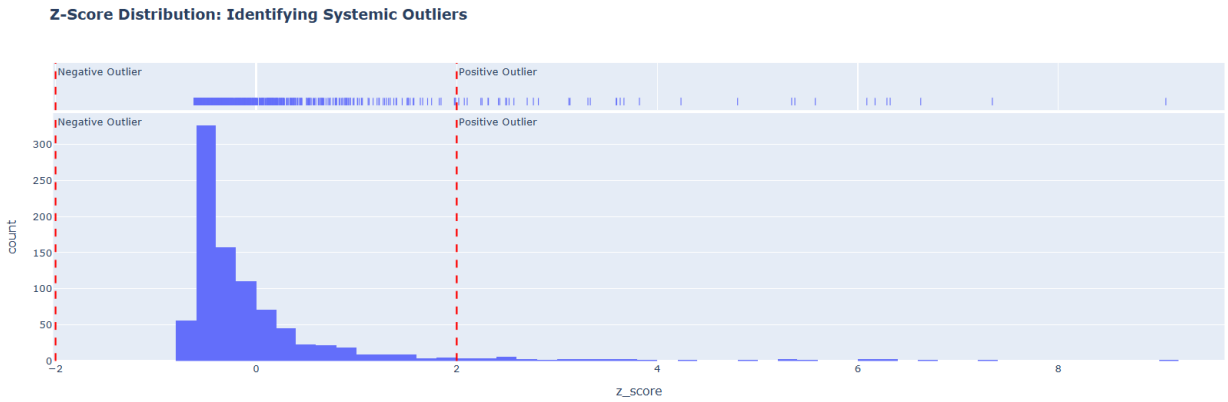
Operational Insight: The 33 "Maintenance Heavy" districts are outliers clearing massive backlogs. These represent critical pressure points where the system was neglected for years and is now in "Emergency Recovery" mode.



C. Z-Score Normalization: Anomaly Identification
Using Z-score normalization, we identified districts that deviate significantly from the national mean. This allowed us to separate "Model Districts" from "Systemic Failures."

- Z-Score Formula: $Z = \frac{x-\mu}{\sigma}$

- The Findings:
    - Positive Anomalies (Z > 6): Districts like Srikakulam (AP) and the Himachal Pradesh cluster (Una, Mandi). These are benchmarks for operational excellence.
    - Negative Anomalies (Z = -0.61): Districts like East Midnapur and Balotra. These represent the statistical "floor" where services have effectively collapsed.



IV. INSIGHTS & VISUALIZATION

1.Finding A: The Biometric Compliance Gap

Our analysis revealed a critical “Maintenance Lag” in the child identity lifecycle (Ages 5–17). While the system is successful at onboarding new users, the “Sustainability Ratio” in several districts has dropped to zero.

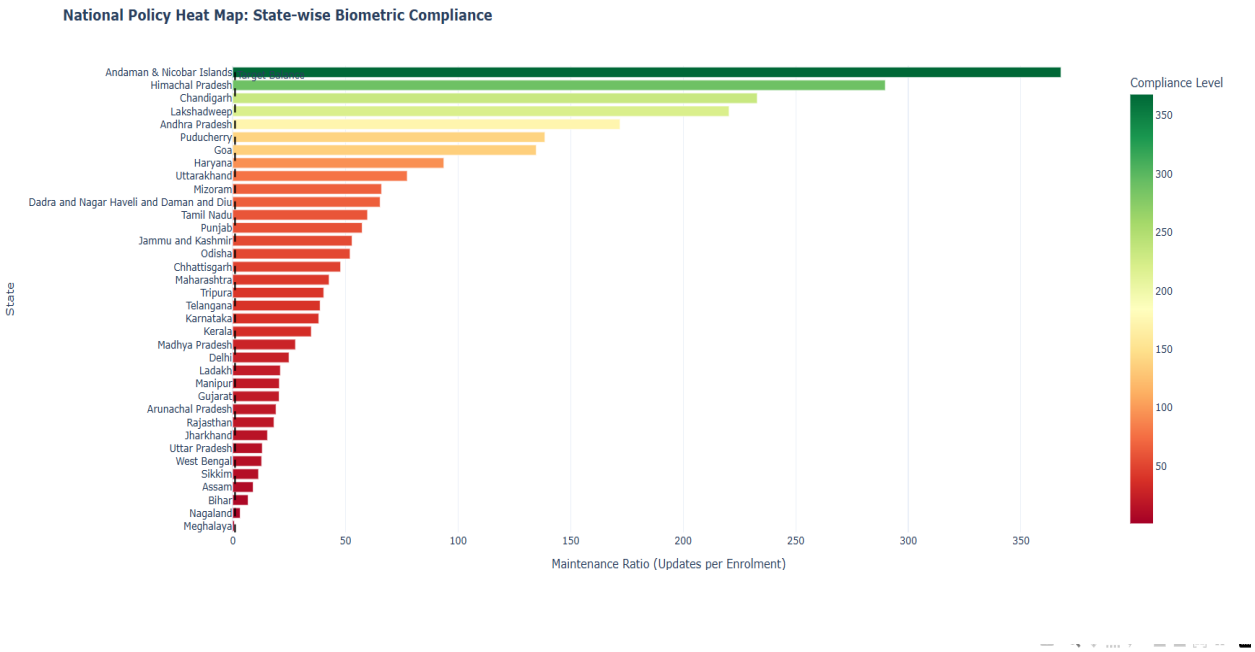
- The Insight:

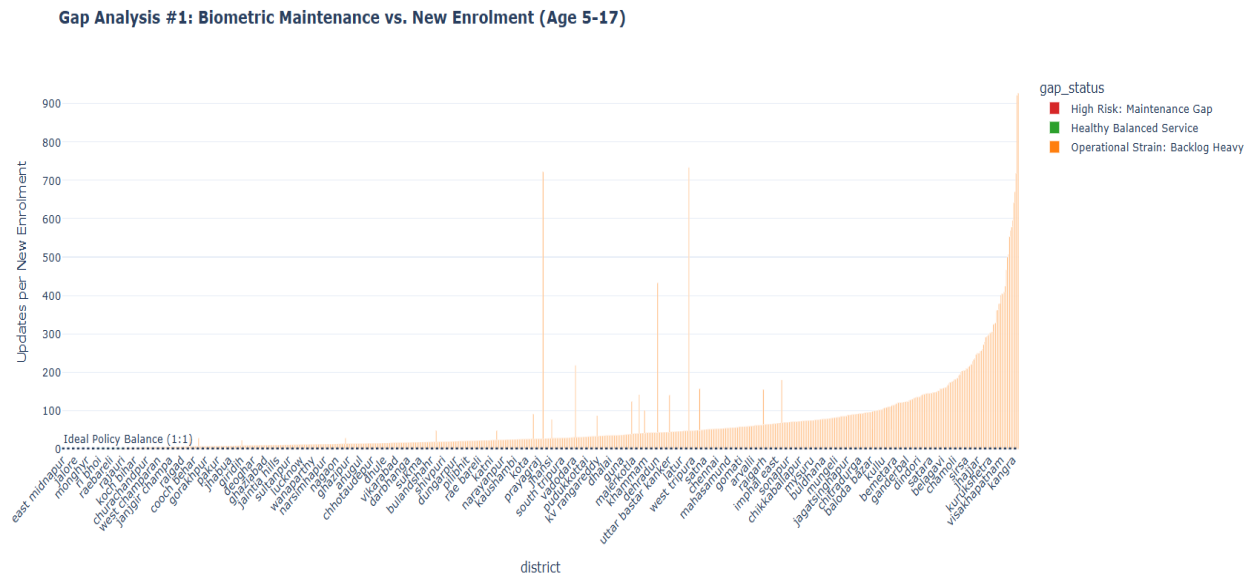
  8 districts (including East Midnapur, WB and Balotra, RJ) showed a Compliance Ratio of 0.00, indicating that mandatory biometric updates have completely stalled despite active new enrolments.

- The Risk:

  These districts are prioritizing “Growth” (New Enrolments) over “Sustainability” (Updates). This creates a “Time Bomb” of authentication failures for school services.

State	District	Compliance Ratio	Status
West Bengal	East Midnapur	0.00	🚨 Critical
Rajasthan	Balotra	0.00	🚨 Critical
Odisha	Anugal	0.00	🚨 Critical
Gujarat	Banas Kantha	0.01	⚠️ High Risk





2. Finding B: The Infrastructure Velocity Gap

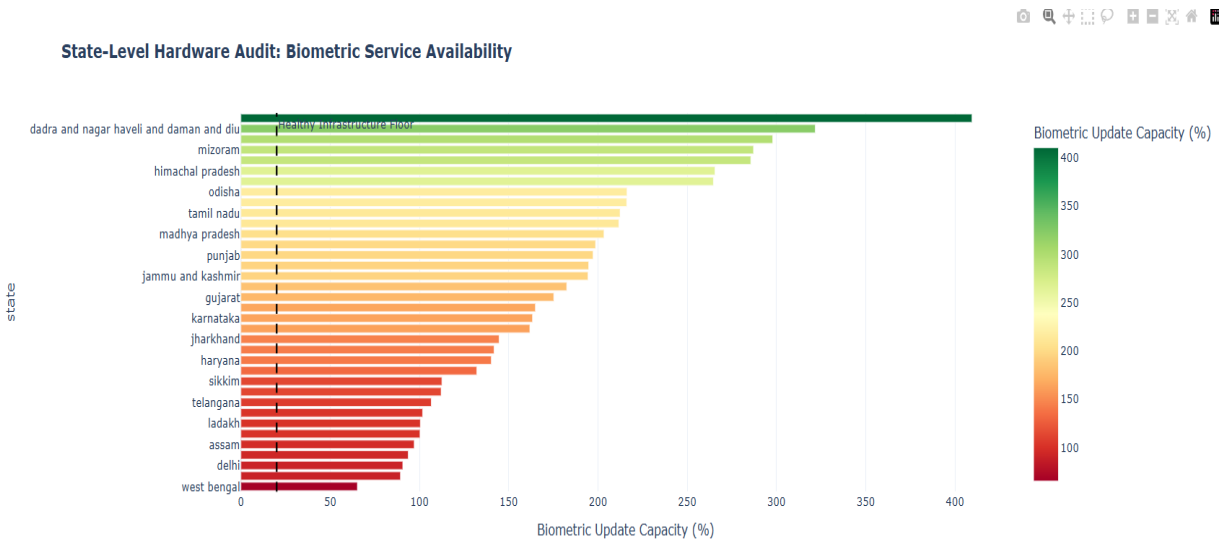
By comparing low-tech demographic demand (intent) with high-tech biometric capacity (output), we pinpointed physical hardware shortages.

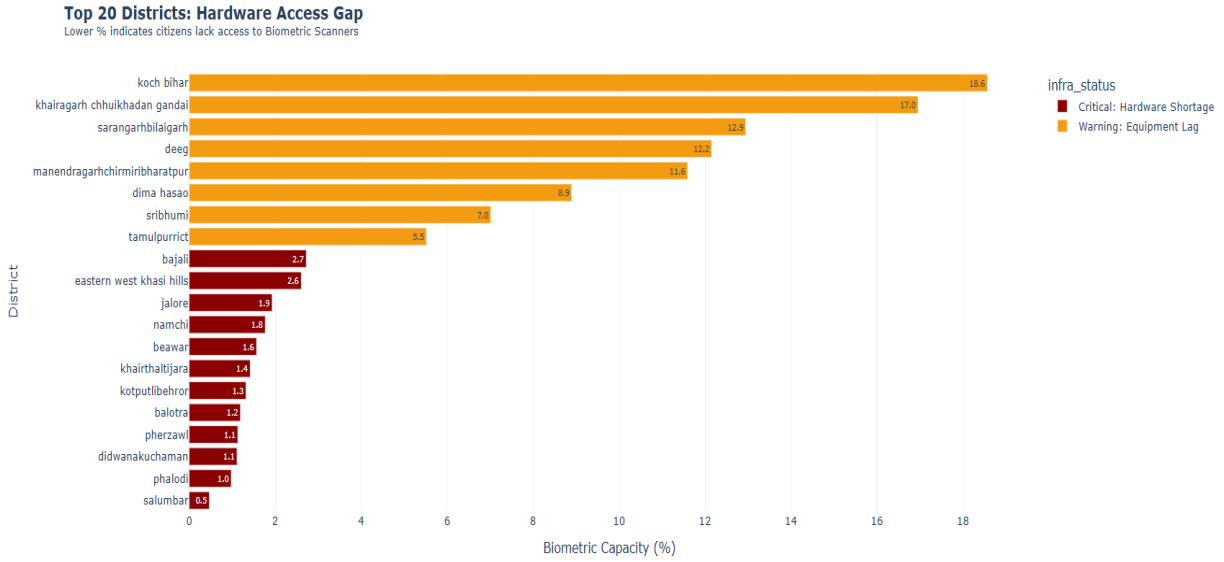
- The Insight:

  In Salumbar, Rajasthan, the infrastructure ratio was only 0.47%.

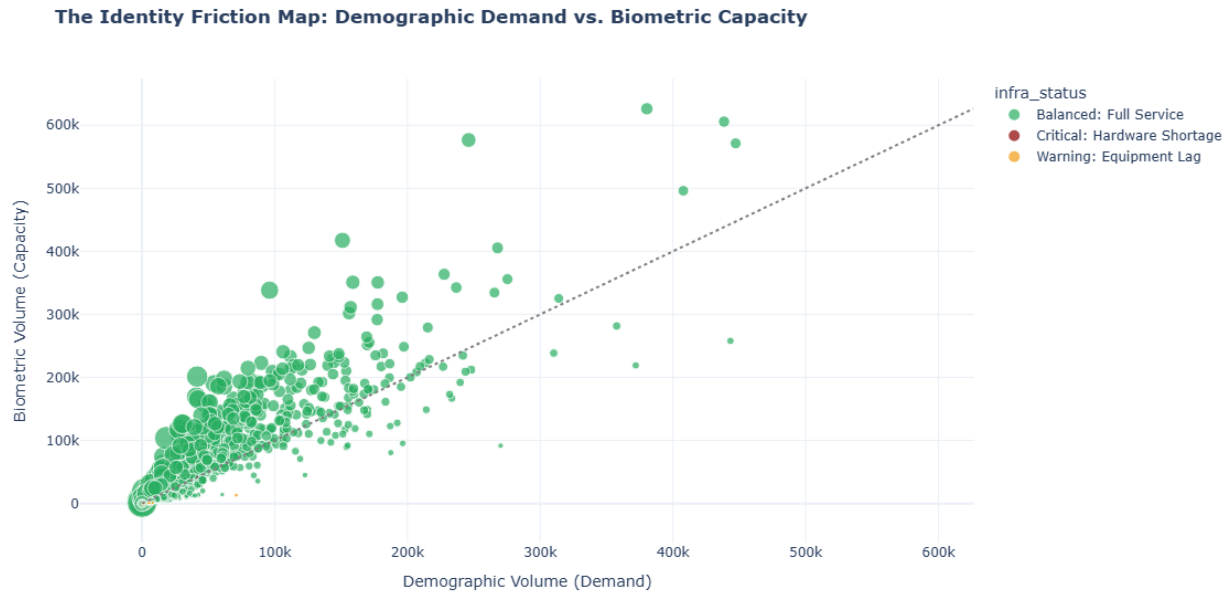
- The Verdict:

  High demographic demand proves citizens want to update their records, but the system fails due to a lack of functional scanners. This is a hardware obsolescence issue, not a lack of citizen intent.





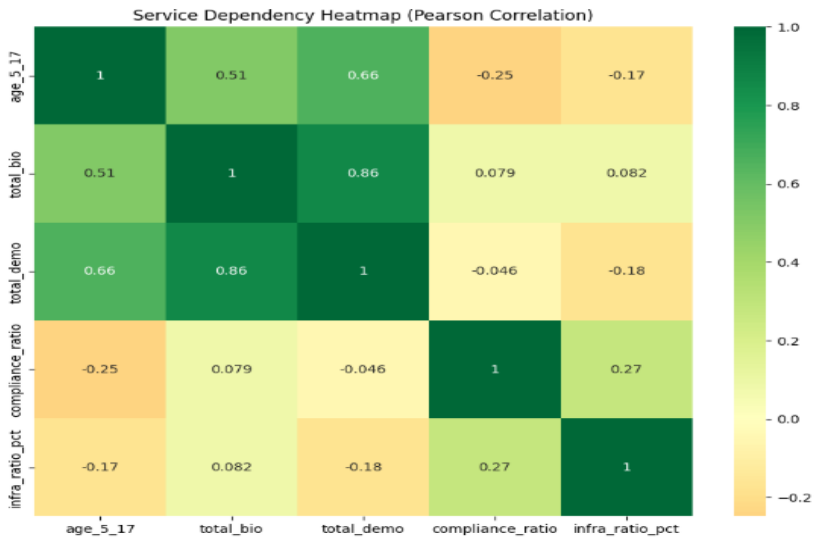
- The Friction Map: Districts far below the 45-degree line in our scatter analysis indicate "Identity Friction," where the software is ready but the hardware is missing.



3. Finding C: Correlation of Service Streams

We used a Pearson Correlation matrix to understand the interdependency of services.

- Insight: A strong correlation (0.86) between Demographic and Biometric updates suggests that when hardware is available, citizens utilize both streams.
- The Anomaly: A weaker correlation (0.51) between Enrolment and Biometrics indicates that "New Growth" is being decoupled from "System Maintenance," confirming our "Maintenance Debt" hypothesis.

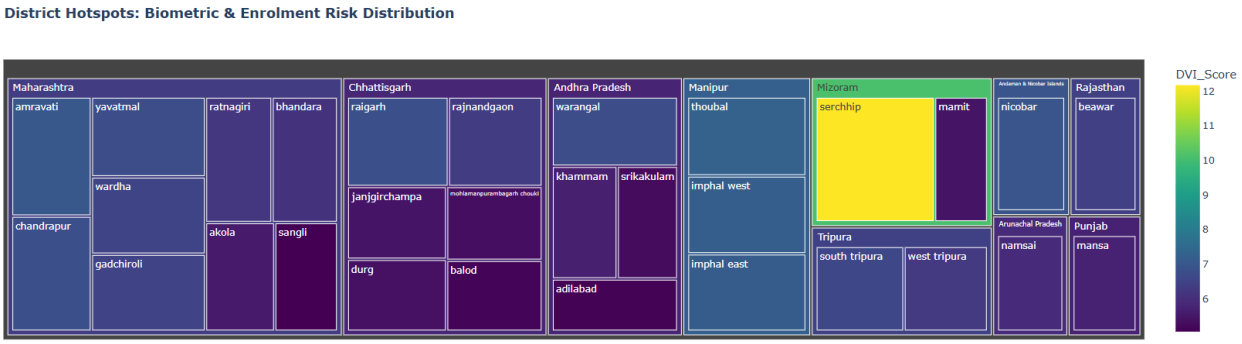


4. District Hotspot Analysis

Using the Digital Vulnerability Index (DVI), we mapped the most at-risk pin codes across India.

- Top Hotspot:

  Srikakulam (AP) stands as a "Hero District" with a Z-score of +9.07, meaning its update velocity is nearly 9 standard deviations above the mean. We recommend studying its "Direct-to-Home" update models for national replication.



V. POWER BI INTERACTIVE DASHBOARD & POLICY

To complement our statistical analysis, we developed an interactive Aadhaar Lifecycle & Infrastructure Dashboard in Power BI. This tool serves as a Command-and-Control interface for UIDAI administrators to monitor the "Maintenance Debt" in real-time.

1. Core Dashboard Modules

- Strategic KPI Dashboard: Tracks the national "Maintenance Debt" in real-time. It features interactive gauges for the Compliance Ratio (CR) and Infrastructure Ratio (IR), allowing officials to see at a glance if the national average is straying from the 1:1 "Healthy Lifecycle" target.

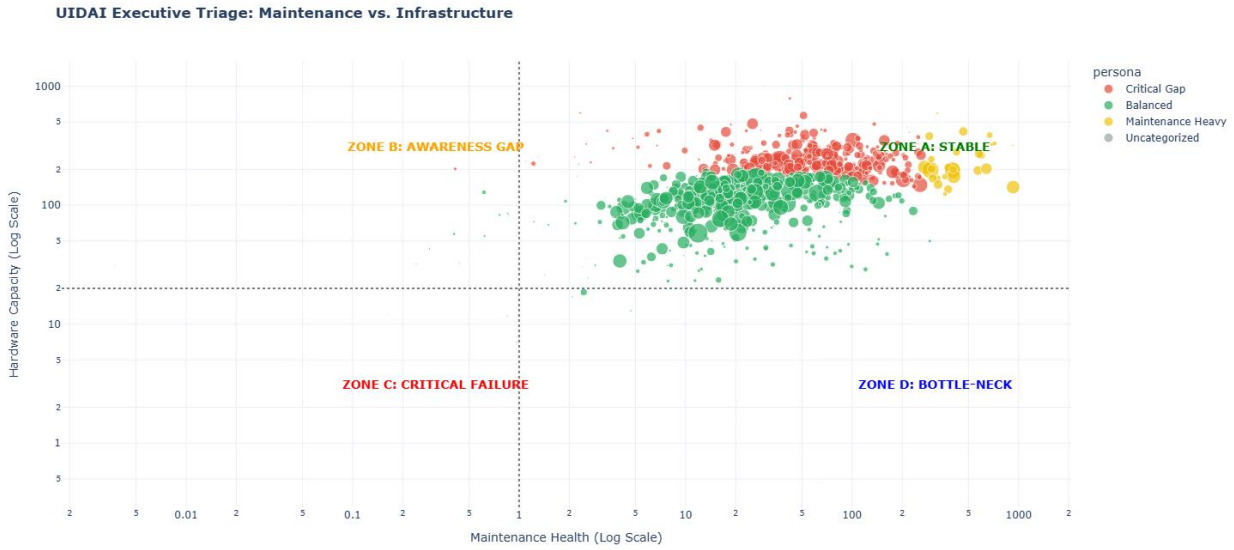
- Geospatial Risk Matrix: A deep-drill choropleth map that color-codes districts by their Digital Vulnerability Index (DVI). This allows for a regional view of "Service Deserts," particularly in the Rajasthan and North East clusters.

- The Identity Friction Scatter: An interactive demand-vs-capacity plot. Officials can select a specific district to view its Demographic-to-Biometric lag, providing instant evidence for hardware procurement requests.

2. Policy Report: The Triage Action Plan

Our Power BI analysis automatically categorizes districts into a 3-Tier Policy Framework, allowing for "Automated Triage" based on incoming data.

Tier Level	Statistical Trigger	Operational Strategy	Example Districts
🚨 Tier 1: Emergency	CR = 0.00	Deploy "Aadhaar on Wheels" to clear child-update backlogs.	East Midnapur, Balotra
⚠️ Tier 2: Technical	IR < 5%	Immediate Hardware Procurement of Iris & Fingerprint scanners.	Salumbar, Phalodi
✅ Tier 3: Routine	IR > 30%	Best-Practice Scaling: Document and replicate regional workflows.	Una, Srikakulam



3. Strategic User Personas

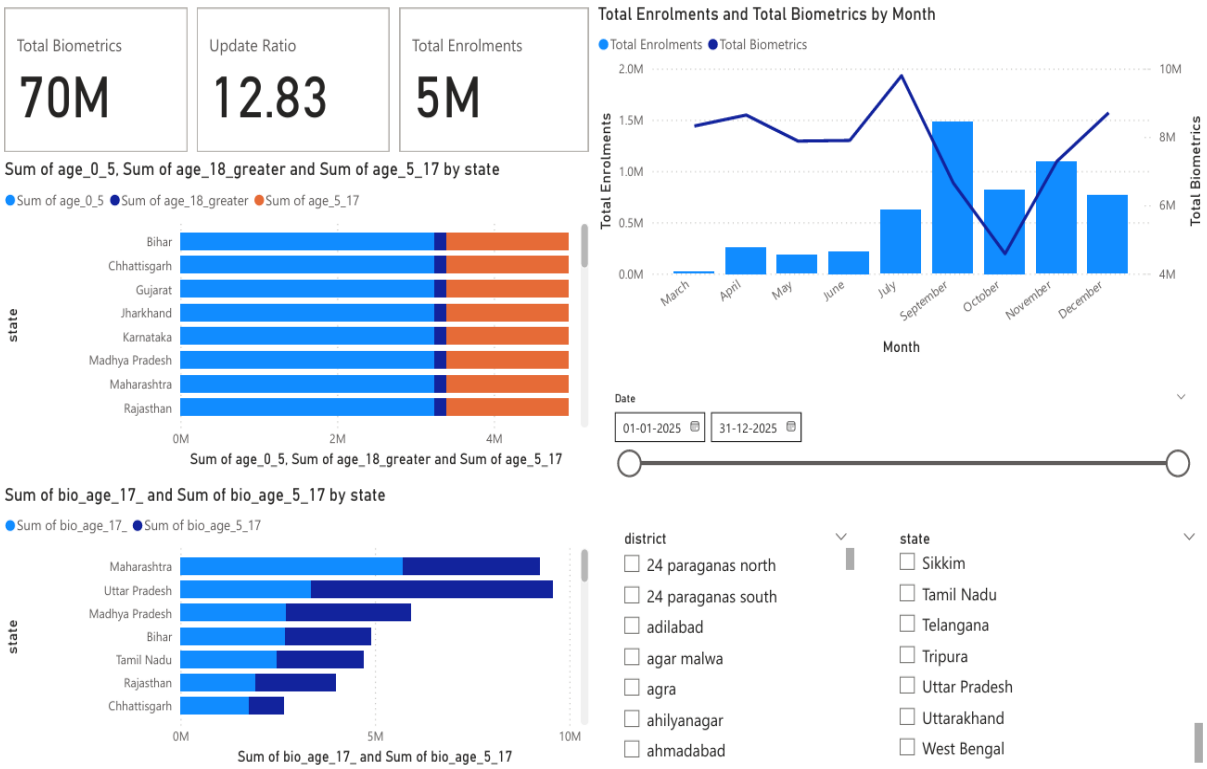
The Power BI report is designed with distinct views for different levels of UIDAI leadership:

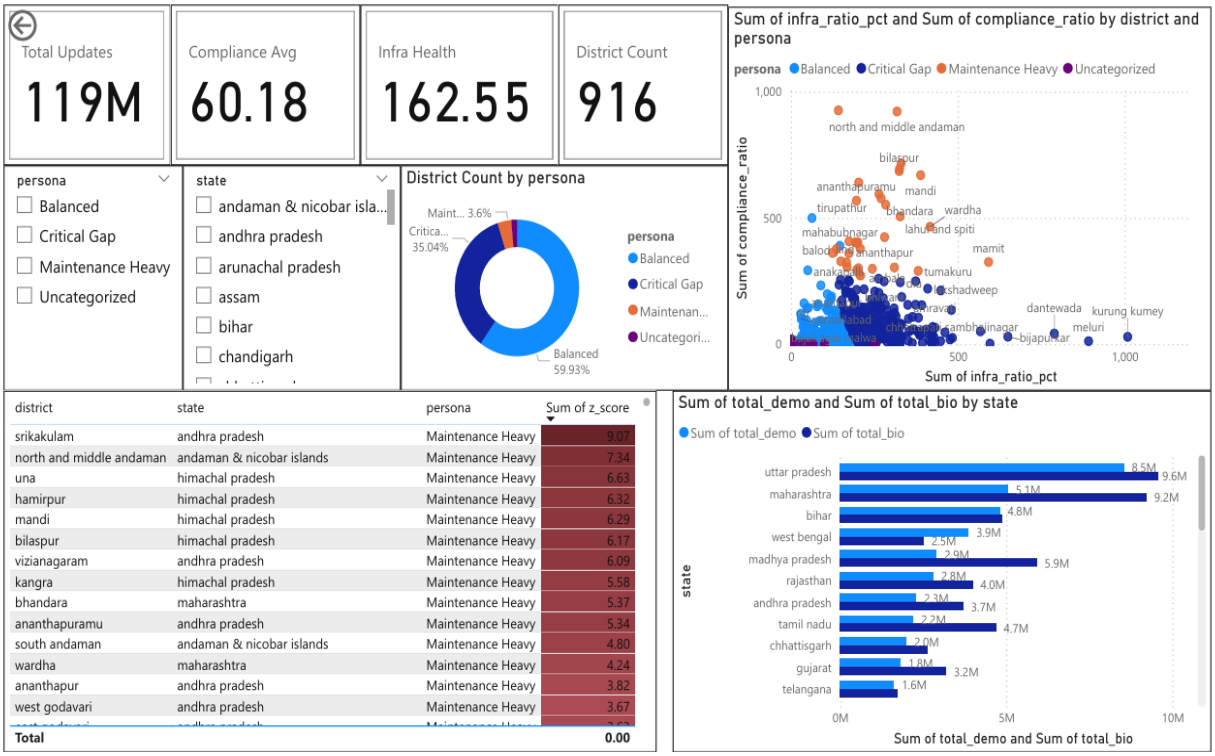
- Central UIDAI Leadership (Strategic View): Focuses on national-level trends, P-Value significance, and state-level policy shifts.
- Regional Managers (Operational View): Uses the Pin code-level drill-down to identify specific update centers that are underperforming or lack hardware.
- Procurement Officers (Infrastructure View): Directly utilizes the "Hardware Access Gap" charts to justify the logistics and shipping of new biometric kits.

4. Strategic Value & Sustainability

This Power BI integration proves that our solution is Scalable. By connecting this dashboard to UIDAI's live API or transaction logs, the "Maintenance Debt" can be monitored and mitigated dynamically. It moves the organization from Reactive Troubleshooting (fixing problems after authentication fails) to Proactive Resilience (deploying hardware before the system collapses).

The Interactive Aadhaar Triage Dashboard. Built in Power BI to provide UIDAI officials with a real-time view of the 'Maintenance Debt,' enabling one-click drill-downs into district-level hardware shortages.




Total Updates
119M
Compliance Avg
60.18
Infra Health
162.55
District Count
916
persona
Balanced
Critical Gap
Maintenance Heavy
Uncategorized
state
andaman & nicobar isla...
andhra pradesh
arunachal pradesh
assam
bihar
chandigarh
District Count by persona
Maint... 3.6%
Critica... 35.04%
Balanced 59.93%
persona
Balanced
Critical Gap
Maintenan...
Uncategori...
Sum of infra_ratio_pct and Sum of compliance_ratio by district and persona
persona Balanced Critical Gap Maintenance Heavy Uncategorized
Sum of compliance_ratio
Sum of infra_ratio_pct
north and middle andaman
bilaspur
ananthapuramu
mandi
tirupathur
bhandara
wardha
mahabubnagar
lahul and spiti
mamit
tumakuru
dantewada
kurung kumey
meluri
bijapur
district state persona Sum of z_score
srikakulam andhra pradesh Maintenance Heavy 9.07
north and middle andaman andaman & nicobar islands Maintenance Heavy 7.34
una himachal pradesh Maintenance Heavy 6.63
hamirpur himachal pradesh Maintenance Heavy 6.32
mandi himachal pradesh Maintenance Heavy 6.29
bilaspur himachal pradesh Maintenance Heavy 6.17
vizianagaram andhra pradesh Maintenance Heavy 6.09
kangra himachal pradesh Maintenance Heavy 5.58
bhandara maharashtra Maintenance Heavy 5.37
ananthapuramu andhra pradesh Maintenance Heavy 5.34
south andaman andaman & nicobar islands Maintenance Heavy 4.80
wardha maharashtra Maintenance Heavy 4.24
ananthapur andhra pradesh Maintenance Heavy 3.82
west godavari andhra pradesh Maintenance Heavy 3.67
Total 0.00
Sum of total_demo and Sum of total_bio by state
Sum of total_demo Sum of total_bio
state
uttar pradesh 8.5M 9.6M
maharashtra 5.1M 9.2M
bihar 4.8M
west bengal 3.9M 2.5M
madhya pradesh 2.9M 5.9M
rajasthan 2.8M 4.0M
andhra pradesh 2.3M 3.7M
tamil nadu 2.2M 4.7M
chhattisgarh 2.0M
gujarat 1.8M 3.2M
telangana 1.6M
0M 5M 10M
Sum of total_demo and Sum of total_bio

CONCLUSION

This study proves that Digital Exclusion in 2025 is a byproduct of geography and hardware availability. By shifting to a "Maintenance-First" policy, UIDAI can ensure that the Aadhaar system remains a functional key to social and financial inclusion for all.

Final Evaluated Findings:

- Scientific Certainty: With an ANOVA p-value of 2.31×10^{-71} we have proven that digital exclusion is not a random occurrence but a systemic regional failure driven by infrastructure inequity.

- The Hardware Bottleneck: Our analysis of "Maroon Zones" like Salumbar (0.47%) and Phalodi (0.98%) confirms that citizen intent for updates is high, but the physical lack of functional biometric scanners is the primary barrier to inclusion.

- The "Silent Exclusion" Risk: The identification of 0.00 Compliance Ratios in districts like East Midnapur and Balotra flags an immediate threat to the child identity lifecycle, where millions of children are at risk of authentication failure.

The Strategic Solution:

Our Service Triage Framework transforms these insights into action. By integrating the Digital Vulnerability Index (DVI) into a real-time Power BI Dashboard, UIDAI can transition from reactive troubleshooting to proactive resilience.

Final Verdict: The stability of India’s digital backbone depends on its ability to maintain what it has built. By adopting a "Maintenance-First" policy and deploying resources-specifically iris and fingerprint scanners-to the high-vulnerability clusters identified in this report, UIDAI can ensure that Aadhaar remains a functional, inclusive, and permanent key to social and financial empowerment for every Indian citizen.

Data identifies the gap; our framework bridges it.