UIDAI Data Hackathon 2026

Early Warning & Decision Support System for Aadhaar Enrolment Operations

Operational risk analytics, early alerts, and capacity planning for UIDAI field operations.

Participants

Abhay Gupta
IIT Gandhinagar | abhay.gupta@iitgn.ac.in

Aayushi Vyas
IIT Gandhinagar | aayushi.vyas@iitgn.ac.in

Prepared for evaluation and operational decision support

Contents

Executive Summary

This project develops an Early Warning and Decision Support System (EWDSS) for Aadhaar enrolment and update operations by transforming large-scale administrative data into actionable operational intelligence. Instead of relying on raw monthly volumes, the system analyzes how enrolment activity behaves over time and across geography, enabling proactive intervention rather than reactive firefighting.

The framework monitors enrolment patterns at district and pincode levels, decomposing operational stress into four complementary signals: volatility (chronic month-to-month unpredictability), spike frequency (recurring abnormal surges), peak stress (capacity shocks in extreme months), and severity (operational impact of abnormal events). These signals are integrated into a single Instability Score, providing a clear, comparable measure of operational risk across regions. The outputs are also showcased on a locally hosted Dashboard.

To support early warning, the system employs robust anomaly detection to flag unusual activity before it escalates into service congestion. Severity scoring further prioritizes anomalies by real-world impact, ensuring attention is directed to events that matter operationally. Concentration analysis (Top-X contribution and inequality measures) highlights whether instability is localized in a small number of regions or systemic, allowing targeted allocation of resources.

Finally, unsupervised clustering groups districts and pincodes into interpretable instability types (e.g., stable, spiky, peaky, mixed), which are mapped to reason codes and corresponding operational responses. This converts complex analytics into decision-ready insights—such as where to deploy surge staffing, activate early alerts, adjust seasonal capacity, or conduct root-cause investigations.

Overall, the proposed EWDSS enables UIDAI to anticipate enrolment stress, prioritize high-risk regions, and optimize operational planning, strengthening service reliability while improving efficiency and responsiveness of Aadhaar enrolment operations.

Key Decision Outcomes

- Proactive capacity planning using month-ahead forecasts for both district and pincodes.
- Risk-based monitoring and audits using anomaly rankings, instability scores, and reason codes.
- Cluster-based governance: common playbooks for regions with similar behaviour (scalable policy design).
- Stable benchmarks via consistent-performer identification.

Problem Statement

Build a system that enables UIDAI to proactively monitor Aadhaar enrolment activity, detect abnormal patterns early, explain what is happening, and forecast future workload to support evidence-based operational decisions.

Innovation & Novelty

1. Explainability-first scoring: transparent calculations and reason codes for administrators.
2. Composite instability index combining coefficient of variation, spike frequency, peak behaviour and severity into a single operational risk score.
3. Cluster-based governance using K-Means to segment regions by behaviour for scalable policy design.
4. Lightweight EWMA forecasting for actionable month-ahead planning with uncertainty bands.

5. A dashboard hosted locally for a quick glimpse of outputs.

Datasets

Primary focus: Aadhaar Enrolment dataset. Supporting datasets (Demographic and Biometric updates) demonstrate extensibility of the same framework.

- Geographic keys: state, district, optionally pincode.
- Time key: date (converted to month).
- Activity columns: age-segmented enrolment counts plus total.

The evaluations are done for districts and pincodes separately.

Pipeline Overview

1. Parse date and derive month.
2. Identify Age columns and converting it to numeric.
3. Monthly aggregation at district and pincode levels.
4. Spike metrics and severity scoring (abnormal month-to-month changes).
5. Top anomalies by month (ranked alerts).
6. Instability scoring and reason codes (explainability layer).
7. Top X contribution in instability
8. Clustering of regions by behavioral similarity using K-Means.
9. Preparing reason code (Volatile, Spiky, Peaky, Mixed).
10. EWMA forecasting for next-month workload.
11. Identification of consistent performers (stable benchmarks).
12. Locally hosted Dashboard through Streamlit.

Technical Deep Dive - Function-by-Function

Spike Detection/ Severity Scoring for both districts and pincodes

The severity score is constructed as a multiplicative composite of three interpretable components:

- relative magnitude (absolute month-over-month percent change),
- statistical unusualness (a robust z-score of the month-to-month change using median and MAD), and
- impact/scale (a log-transformed current-month volume).

This design reflects a standard "risk = likelihood × impact" philosophy: the robust z-score captures how unlikely the change is relative to the location's own historical behavior (reducing sensitivity to outliers and non-normality), the percent change captures how large the shift is in relative terms, and the log-volume term ensures the score reflects real-world operational impact without letting very large regions dominate purely by size.

Severity score can be as a prioritization and response tool. The score helps identify which events demand immediate attention because they combine large change, rarity, and high operational impact. High-severity spikes can trigger actions such as surge staffing, temporary capacity expansion, targeted audits, or advance planning for expected drives. Aggregated over time and locations, severity scores also reveal districts that experience intense stress during spike months, informing where contingency plans, buffers, and preventive interventions will deliver the greatest benefit.

Severity Score = (|%Δ|+∈) × ($Z_Δ$+∈) × (log(1+X))

%Δ = $\frac{Absolute\ Change}{previous\ month\ total}$ where, Absolute Change = current month total - previous month total

The value obtained by %Δ is clipped between [0,20] that means changes beyond 2000% (20) clips down to 20. Thus, the values remain between [0,20].

Z Score = $\frac{x - median}{1.4826 * MAD}$

MAD = Mean Absolute Deviation

For a standard normal variable $Z \sim \mathcal{N}(0,1)$:

median(|Z|) ≈ 0.67449

MAD ≈ 0.67449 × Standard Deviation

Standard Deviation = 1.4826 × MAD

$Z_Δ$ = evaluating Z Score on value obtained on Absolute Change

∈ = slack variable 0.1. If either %Δ or $Z_Δ$ is 0, the product would become 0 and you'd lose information. 0.1 makes severity still non-zero for mild anomalies and prevents hard zeros.

log(1+x) = How big is the real-world impact of this anomaly without letting large regions overpower the analysis?

Categorizing the Severity Score in Spike or Not a Spike:

The Spike condition is based on either of two conditions:

- Top 1% by severity globally --> Severity Score is in the top 1%, OR
- $Z_Δ \geq 3$ AND %Δ > 1 that is 100%

If either of the Severity Scores passes through these criteria then the Flag is_spike = TRUE.

Top Anomalies by Month for both districts and pincodes.

Anomalies typically represent unusual surges, drops, or disruptions that cannot be explained by normal variability. In this case, the Severity Scores are ranked and top-N districts or pincodes are where is_spike Flag is True are shown. Here Total = total enrolments

District anomalies: Interpretation: Large negative percent change (near -100%) indicates abrupt collapse vs previous month.

Anomalies on District level							
month	**state**	**district**	**total**	**prev**	**pct change**	**severity**	**rank in month**
Feb 2025	Bihar	Pashchim Champaran	3	16,149	-100.0%	8,302.210	1
Feb 2025	Gujarat	Banas Kantha	8	12,522	-99.9%	5,097.387	2
Feb 2025	Gujarat	Ahmadabad	35	5,733	-99.4%	2,152.230	3
Feb 2025	Gujarat	Dohad	97	11,239	-99.1%	1,880.588	4
Feb 2025	Madhya Pradesh	Morena	518	11,215	-95.4%	1,633.360	5
Feb 2025	Delhi	New Delhi	10	1,458	-99.3%	1,280.295	6
Feb 2025	Gujarat	Botad	45	1,805	-97.5%	1,219.875	7
Feb 2025	Maharashtra	Hingoli	134	4,851	-97.2%	1,191.902	8
Feb 2025	Meghalaya	East Khasi Hills	272	22,688	-98.8%	1,168.819	9

Feb 2025	Madhya Pradesh	Barwani	481	13,677	-96.5%	1,147.696	10

Anomalies on Pincode level								
month	**state**	**district**	**pincode**	**total**	**prev**	**pct change**	**severity**	**rank in month**
Feb 2025	Bihar	Pashchim Champaran	845438	3	3,114	-99.9%	3,197.160	1
Feb 2025	Tamil Nadu	Namakkal	637406	75	6	1,150.0%	2,343.024	2
Feb 2025	Uttar Pradesh	Barabanki	225305	2	1,069	-99.8%	1,736.595	3
Feb 2025	Uttar Pradesh	Shahjahanpur	242221	40	2,842	-98.6%	1,524.144	4
Feb 2025	Gujarat	Dohad	389151	27	2,203	-98.8%	1,519.250	5
Feb 2025	Uttar Pradesh	Etah	207001	64	4,489	-98.6%	1,503.819	6
Feb 2025	Gujarat	Dohad	389170	2	1,699	-99.9%	1,381.058	7
Feb 2025	Gujarat	Dohad	389380	3	638	-99.5%	1,298.774	8
Feb 2025	Uttar Pradesh	Sitapur	261001	54	1,389	-96.1%	1,273.870	9
Feb 2025	Gujarat	Dohad	389180	15	1,854	-99.2%	1,251.346	10

Pincode anomalies enable micro-targeted intervention (specific centres/camps).

Instability Score for both districts and pincodes.

Instability Score is a composite indicator that summarizes how operationally unpredictable a district's Aadhaar activity is over time. Rather than relying on raw monthly volumes, it integrates four complementary signals:

- volatility (how much monthly activity fluctuates relative to its average),
- spike frequency (how often abnormal surges occur),
- peak stress (how extreme the highest-load month is compared to normal), and
- spike intensity (how severe those abnormal surges are when they occur).

Together, these dimensions capture both chronic instability and event-driven risk, an approach widely used in operations and risk management literature where composite indices are favored for their robustness and interpretability over single metrics.

For a decision maker, the Instability Score functions as a prioritization and planning tool. A higher score indicates districts or pins that are harder to manage operationally and more prone to service disruptions without proactive intervention.

The score translates complex time-series behavior into a single, actionable measure that supports informed resource allocation, risk mitigation, and policy decisions.

Core components (per region):

instability_score_with_severity(aggregated monthly df, severity df, level, min_months, severity_agg, severity_stat, peak_mode)

- level = either district or pin based on if the user is evaluating instability on district or pincode.
- Min_month = 4 (Can be changed, it it is kept as a threshold so that districts/pins if having months with count >= min_month will be considered).
- Coefficient of Variation (CoV): std(total) / mean(total).

- Spike frequency: (# spike months) / (# observed months) - how often abnormal behaviour occurs.
- Peak factor: max(total) / mean(total) - how extreme the highest month is vs normal.
  - Peak_mode: There is a provision to choose either the normal ratio or log ratio for peak factor. It can be mentioned at the time of calling function.
- Severity metric: typically the 95th percentile of monthly spike severity - how bad spikes tend to get.
- Drivers that drives Severity Score in the calculation of Instability
  - **Severity_Agg:** Choose either of the one while calling the function-
    - sum: total severity across spiky PINs/Districts in a district-month (captures distributed stress)
    - max: worst spike in that district-month (captures extreme local pocket)
    - mean: average severity across spiky PINs/Districts (captures typical spike intensity), how intense is a typical spike pocket?
  - **Severity_Stat:** Choose either of the while calling the function-
    - Mean:
      - What it emphasizes
        - Chronic, everyday stress
        - Repeated moderate problems
        - Long-term operational load
      - What it ignores
        - Rare but very dangerous months
        - Extreme one-off stress events
    - P95:
      - What it emphasizes
        - Chronic, everyday stress
        - Repeated moderate problems
        - Long-term operational load
      - What it ignores
        - Rare but very dangerous months
        - Extreme one-off stress events

How components are combined: Each component is robust-normalized (0-1) and combined into a 0-100 index with fixed weights (variance, spike frequency, peak behaviour, severity).

The instability score is a weighted sum of Coefficient of Variation (COV), Spike Frequency, Peak Factor and Instability score.

Instability score = (0.30 * COV) + (0.25 * Spike Frequency) + (0.20 * Peak Factor) + (0.25 * Peak Factor)

Instability Score of top 10 districts of Aadhar Enrollment:

state	district	months	mean	std	cov	max_total	peak_factor	spike_months	spike_freq	severity_metric	cov_norm	spike_norm	peak_norm	severity_norm	instability_score
Gujarat	Ahmadabad	12	503.5	1646.902	3.270907	5733	11.38629593	2	0.166667	2048.219	1	1	1	1	100
Gujarat	Botad	12	181.4167	511.6169	2.820121	1805	9.94947175	2	0.166667	1169.141	1	1	1	1	100
Madhya Pradesh	Barwani	12	1375.667	3875.243	2.816993	13677	9.942088684	2	0.166667	1094.297	1	1	1	1	100
Gujarat	Gir Somnath	12	342.0833	951.2091	2.780636	3361	9.825091352	2	0.166667	593.2323	1	1	1	1	100

state	district	months	mean	std	cov	max_total	peak_factor	spike_months	spike_freq	severity_metric	cov_norm	spike_norm	peak_norm	severity_norm	instability_score
Bihar	Sitamarhi	12	2589.417	7423.521	2.86687	26151	10.09918579	2	0.166667	757.2784	1	1	1	1	100
Maharashtra	Hingoli	12	490.1667	1373.787	2.802693	4851	9.896633798	2	0.166667	1135.318	1	1	1	1	100
Madhya Pradesh	Morena	12	1214.75	3151.257	2.594161	11215	9.232352336	2	0.166667	1554.971	0.944997	1	0.949405	1	97.33801
Assam	Kamrup	12	784.8333	2029.015	2.585282	7225	9.205776173	2	0.166667	873.5231	0.941344	1	0.946173	1	97.16377
Chhattisgarh	Sukma	12	105.8	286.1	2.703	1014	9.5811023	2	0.166667	398.9384	0.99004	1	0.991	0.82455	95.15189
Uttar Pradesh	Varanasi	12	1218.75	2974.631	2.440723	10658	8.745025641	2	0.166667	734.6371	0.881876	1	0.890127	1	94.25882

Instability score of top 10 pincodes of Aadhar Enrollment:

state	district	pincode	months	mean	std	cov	max_total	peak_factor	spike_months	spike_freq	severity_metric	cov_norm	spike_norm	peak_norm	severity_norm	instability_score
Madhya Pradesh	Sehore	466331	12	96.667	279.297	2.889	983	10.169	2	0.167	1000.98	1	1	1	1	100
Gujarat	Chhotaudepur	391152	9	17.444	47.463	2.721	144	8.255	1	0.111	227.88	1	1	1	1	100
West Bengal	Uttar Dinajpur	733129	12	90.917	260.254	2.863	917	10.086	2	0.167	705.08	1	1	1	1	100
Uttar Pradesh	Pilibhit	262001	12	350.000	1019.303	2.912	3586	10.246	2	0.167	1022.39	1	1	1	1	100
Uttar Pradesh	Siddharthnagar	272153	12	177.833	443.758	2.495	1586	8.918	2	0.167	304.30	1	1	1	1	100
Uttar Pradesh	Sitapur	261204	12	127.167	366.285	2.880	1290	10.144	2	0.167	616.77	1	1	1	1	100
Bihar	Gaya	824232	12	143.833	404.888	2.815	1429	9.935	2	0.167	348.99	1	1	1	1	100
Bihar	Madhubani	847227	12	45.667	108.738	2.381	390	8.540	3	0.250	152.44	1	1	1	1	100
Bihar	Madhubani	847108	12	223.750	635.778	2.841	2237	9.998	2	0.167	147.20	1	1	1	1	100
Uttar Pradesh	Mathura	281121	12	44.917	122.585	2.729	434	9.662	2	0.167	762.37	1	1	1	1	100

Top X contribution in instability of districts and pincodes

Top-X contribution analysis is a concentration metric used to quantify how much of a system's total instability is driven by a small subset of units (e.g., districts or PIN codes). Rather than only identifying the most unstable entities, it measures the cumulative share of instability accounted for by the top X% of units, providing insight into the distributional structure of risk across the system.

The primary objective of Top-X contribution analysis is to answer the question:

"Is instability widespread across the system, or is it concentrated in a small number of hotspots?"

By expressing instability in cumulative terms (e.g., "Top 10% of districts contribute 55% of total instability"), this metric enables:

- Prioritization of interventions, by identifying whether targeting a small subset of districts can significantly reduce system-wide risk.
- Efficient resource allocation, avoiding uniform responses when problems are structurally concentrated.
- Strategic planning, by distinguishing between systemic instability and localized operational stress.

This makes Top-X contribution especially valuable for decision makers who must choose where limited monitoring, staffing, or corrective capacity should be deployed.

Mathematical Logic of Instability Concentration Analysis:

Each district or PIN code is assigned a non-negative **Instability Score**, representing its operational risk. Let these scores be $I_1, I_2, \dots, I_N$.

To assess **risk concentration**, scores are first ranked in descending order. For a chosen proportion x(e.g., top 10%), the top $k = \lceil xN \rceil$units are selected. The **Top-X Contribution** is computed as the ratio of instability contributed by these units to total system instability:

$$\text{Top-X Contribution} = \frac{\sum_{i=1}^{k} I(i)}{\sum_{i=1}^{N} I(i)} \times 100$$

To provide a single summary measure of inequality, the Gini coefficient is computed over instability scores:

$$G = \frac{n + 1 - 2 * \frac{\sum_{i=1}^{n} C(i)}{C(n)}}{n}$$ for which:

- C_i is the cumulative sum of sorted instability scores
- C_n is the total instability
- n is the number of regions

To summarize inequality in a single measure, the **Gini coefficient** is computed over the instability scores.
Interpretation

- $G = 0$: instability evenly distributed across all regions
- $G \to 1$: instability highly concentrated in a small subset

Thresholds are used to classify concentration as:

- **Fairly distributed** ($G < 0.35$)
- **Moderately concentrated** ($0.35 \leq G < 0.50$)
- **Highly concentrated** ($G \geq 0.50$)

Together, Top-X Contribution and Gini analysis convert region-level instability scores into **system-level insights**, enabling decision makers to determine whether instability is localized or widespread and to prioritize targeted interventions efficiently. This makes Top-X contribution especially valuable for decision makers who must choose where limited monitoring, staffing, or corrective capacity should be deployed.

Top-X Contribution on district level

Enrolments Updates (districts): Instability is highly concentrated (Gini=0.63). Concentration: top 5% contribute 22.19%; top 10% contribute 41.36%; top 20% contribute 64.26% of total instability. Highest-risk districts: district=Ahmadabad, state=Gujarat → 100.0; district=Botad, state=Gujarat → 100.0; district=Gir Somnath, state=Gujarat → 100.0; district=Barwani, state=Madhya Pradesh → 100.0; district=Hingoli, state=Maharashtra → 100.0.

Top-X Contribution on pincode level

Enrolments Updates (PIN codes): Instability is highly concentrated (Gini=0.64). Concentration: top 5% contribute 30.82%; top 10% contribute 49.51%; top 20% contribute 72.64% of total instability. Highest-risk PIN codes: pincode=382440, district=Ahmedabad, state=Gujarat → 100.0; pincode=271503, district=Gonda, state=Uttar Pradesh → 100.0; pincode=466331, district=Sehore, state=Madhya Pradesh → 100.0; pincode=391152, district=Chhotaudepur, state=Gujarat → 100.0; pincode=261001, district=Sitapur, state=Uttar Pradesh → 100.0.

K Means and Cluster Plotting

We use K-means clustering here to discover natural groups of districts/PINs that behave similarly, rather than evaluating each location in isolation.

What K-means adds beyond ranking or scoring?

- Instability score ranks severity, but it does not explain why a district is unstable.
- K-means clusters explain structure, grouping districts with similar instability drivers even if their total scores differ.

Two districts/pin-codes may have the same instability score:

- one driven by frequent small spikes,
- another driven by a single extreme peak.

K-means separates them into different clusters, enabling tailored interventions. In this analysis, K-Means is applied to normalized instability components (volatility, spike frequency, peak stress, and severity) to identify natural groupings of districts or PIN codes based on how instability manifests, rather than how large it is.

Cluster profiling then interprets each cluster by examining the average values of these components, translating abstract clusters into operational archetypes (e.g., stable, spiky, peaky, mixed). For decision makers, this enables differentiated responses: regions in the same cluster can be managed using similar strategies, while different clusters require distinct interventions. Together, K-Means and cluster profiling transform complex multi-dimensional instability metrics into interpretable categories that support targeted, cause-specific operational planning.

The number of clusters k is selected using **silhouette scores**, which balance cluster compactness and separation.

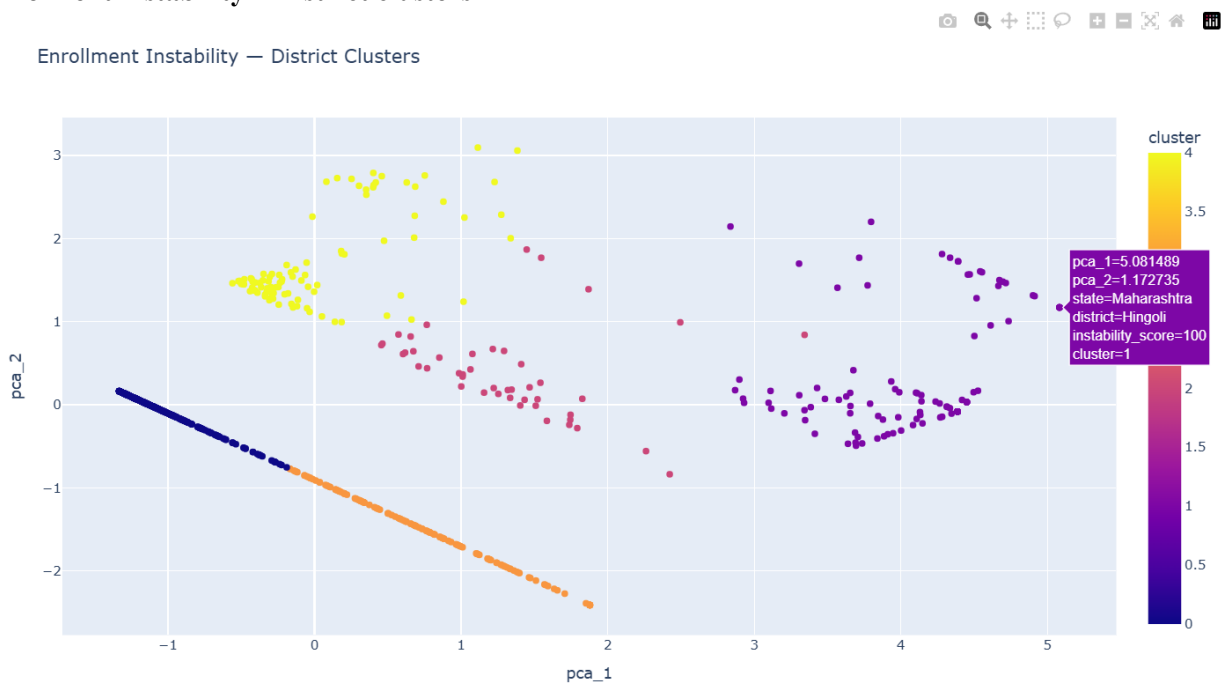
To support visualization and interpretability, **Principal Component Analysis (PCA)** is applied as a linear transformation that projects the original feature space into a lower-dimensional orthogonal basis capturing maximum variance. PCA does not affect clustering results; it is used solely to visualize cluster separation and structure in two dimensions.

Cluster profiling is performed by computing summary statistics (typically means or medians) of each instability component within clusters:

$$X_j = \frac{1}{|C(j)|} * \sum_{x(i) \in C(j)} x(i)$$

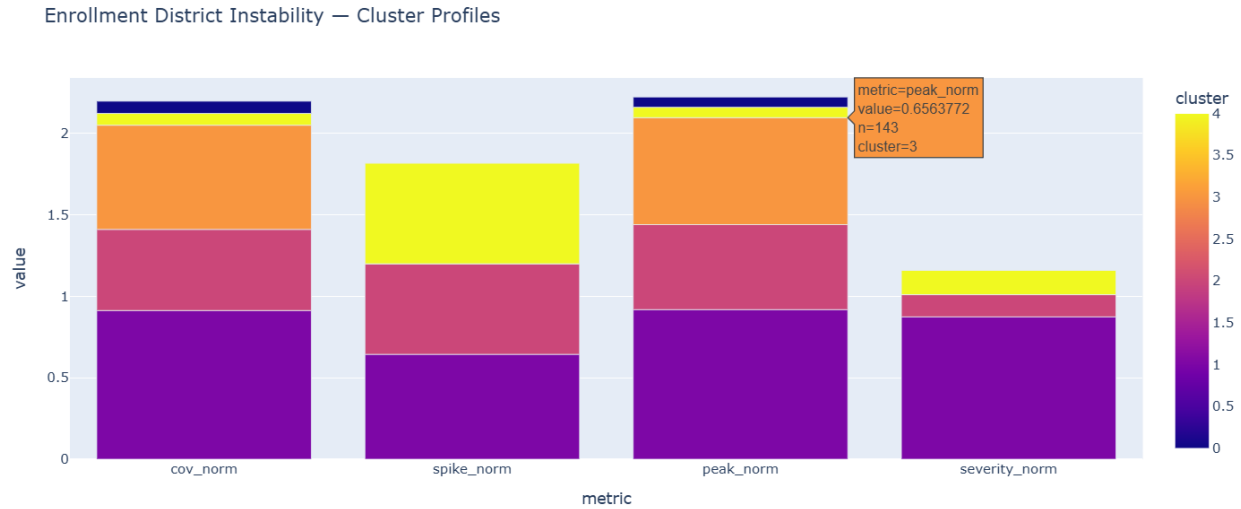
These profiles translate geometric clusters into interpretable operational patterns (e.g., high volatility vs. high spike frequency). Together, K-Means, PCA, and cluster profiling convert multi-dimensional instability measures into distinct, interpretable region types that enable targeted and cause-specific decision making.

Enrolment instability – District clusters



The above Cluster plot is showing 5 clusters that is Cluster 0 = Blue, Cluster 1 = Purple, Cluster 2 = Pink, Cluster 3 = Orange, Cluster 4 = Yellow.

Cluster Profile – District Clusters



Decoding cluster profiles – Aadhar Enrolment – District wise instability

🔵 Cluster 0 (Blue) — Stable / Low-Risk

Profile: cov_norm ≈ 0.075, spike_norm=0, peak_norm ≈ 0.063, severity_norm=0

This cluster represents districts with highly stable enrollment behavior. Low normalized volatility and peak stress indicate that demand remains close to its typical level, while the absence of spikes and severity suggests no abnormal surges over the observed months. In operational terms, these districts behave predictably and require only routine monitoring and standard staffing. Cluster 0 can serve as a benchmark baseline for expected system behavior.

Operational implications:

- Maintain standard staffing and processes
- Use as a benchmark baseline for expected performance
- No intervention required beyond routine monitoring

Probable Reason Code: STABLE

🟣 Cluster 1 (Purple) — Extreme Mixed Instability / High-Risk Hotspots

Profile: cov_norm≈0.913, spike_norm≈0.644, peak_norm≈0.918, severity_norm≈0.874

This is the highest-risk cluster, characterized by simultaneous elevation across all four instability dimensions. High CoV indicates chronic month-to-month turbulence; high spike frequency shows that abnormal months occur repeatedly; high peak stress implies extreme capacity-shock months; and high severity confirms that spikes are operationally intense, not merely statistical noise. In practice, these are the districts most likely to trigger service congestion or SLA risk and should be prioritized for comprehensive intervention—buffer capacity, proactive forecasting, special monitoring, and root-cause investigation.

Operational implications:

- Highest priority for intervention
- Deploy buffer capacity and surge staffing
- Enable proactive monitoring and early-warning systems
- Conduct root-cause analysis (infrastructure, demand shocks, policy drives)

Probable Reason Code: MIXED

🌸 Cluster 2 (Pink) — Moderate Mixed Variability / Managed Instability

Profile: cov_norm≈0.498, spike_norm≈0.556, peak_norm≈0.522, severity_norm≈0.136

Cluster 2 reflects districts with moderate instability driven by regular fluctuations and recurring spikes, but with low spike intensity. The combination of mid-level CoV, spike frequency, and peak factor suggests these areas experience operational variability (likely due to routine administrative cycles or predictable drives), yet the low severity indicates that spike events are generally contained and manageable. Operationally, this cluster benefits from improved scheduling and early-warning triggers rather than major capacity expansion.

Operational implications:

- Improve scheduling discipline and demand smoothing
- Use rolling forecasts and adaptive staffing
- Monitor trends rather than respond to incidents

Probable Reason Code: VOLATILE

🟠 Cluster 3 (Orange) — Peaky-Volatile Without Spikes / Structural Swings

Profile: cov_norm≈0.639, spike_norm=0, peak_norm≈0.656, severity_norm=0

This cluster shows structural variation and peak pressure without being classified as "spiky." Higher volatility and peak factor indicate enrollment demand swings and elevated maximum months relative to baseline, but the absence of spike frequency/severity implies these changes are either gradual, seasonal, or do not meet the anomaly threshold used for spike classification. Operationally, this cluster represents districts where capacity planning should focus on seasonality management and forecasting, rather than incident response.

Operational implications:

- Plan seasonal or policy-driven capacity surges
- Temporary staffing during known peak windows
- Focus on capacity planning, not anomaly response

Probable Reason Code: PEAKY

🟡 Cluster 4 (Yellow) — Spiky-Low-Volume / Event-Driven Surges

Profile: cov_norm≈0.072, spike_norm≈0.619, peak_norm≈0.064, severity_norm≈0.150

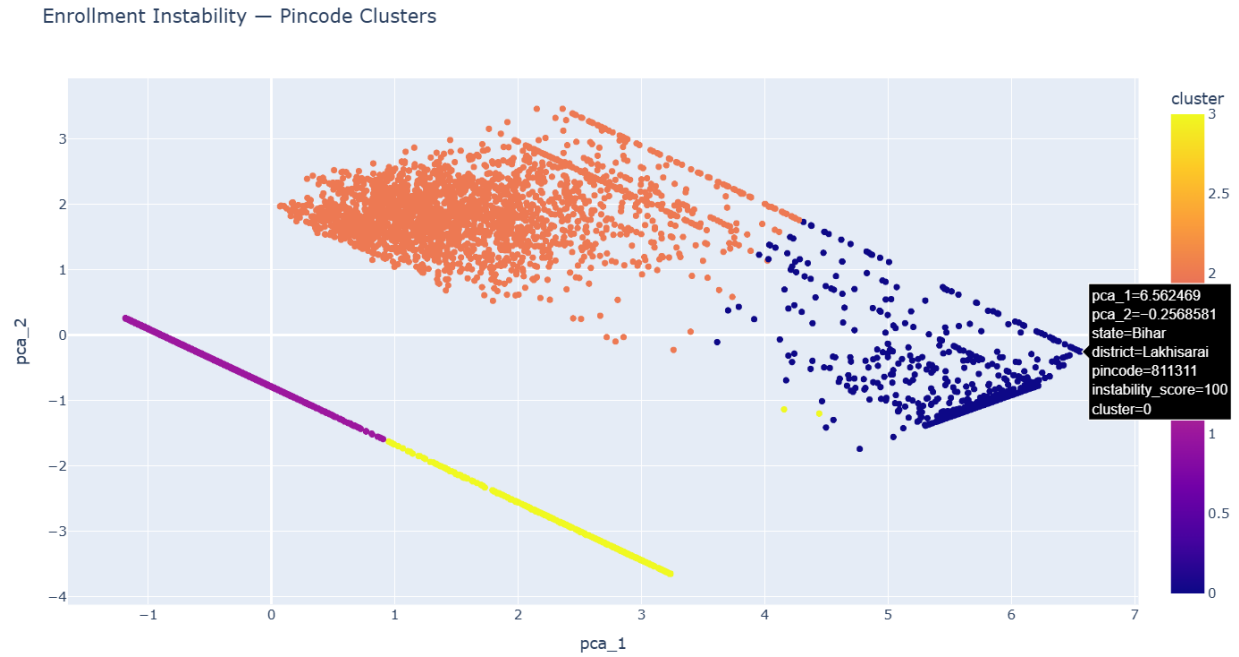
Cluster 4 is defined by frequent spikes occurring on an otherwise stable baseline. Low CoV and peak factor suggest typical volumes are steady, but spike_norm is high—meaning spikes occur repeatedly. Severity is low-to-moderate, implying these spikes are not extremely intense but are persistent. This pattern is consistent with event-driven behavior (scheduled drives, periodic backlogs, outreach campaigns) that causes recurring surges without fundamentally destabilizing the baseline. Operationally, this cluster is best managed through event calendars, pre-emptive staffing for known spike windows, and queue management, rather than permanent capacity increase.

Operational implications:

- Plan seasonal or policy-driven capacity surges
- Temporary staffing during known peak windows
- Focus on capacity planning, not anomaly response

Probable Reason Code: SPIKY

Enrolment instability – Pincode clusters



The above cluster plot is showing 4 clusters that is Cluster 0 = Blue, Cluster 1 = Pink, Cluster 2 = Orange, Cluster 3 = Yellow.

Cluster Profile – Pincode Clusters

Decoding cluster profiles – Aadhar Enrolment – Pincode wise instability

🔵 Cluster 0 (Blue) — Extreme Mixed Instability Hotspots

Profile: cov_norm ≈ 0.96, spike_norm ≈ 0.81, peak_norm ≈ 0.96, severity_norm ≈ 0.91

This cluster represents high-risk pincode hotspots where all dimensions of instability are simultaneously elevated. Enrollment activity in these pincodes is highly volatile month-to-month, experiences frequent spikes, reaches extreme peak loads, and those spikes are operationally severe. This pattern indicates

localized breakdowns or intense operational stress, often driven by concentrated population demand, repeated special drives, or infrastructure constraints. These pincodes should be the top priority for intervention, requiring granular monitoring, surge capacity, and possibly structural fixes at the enrolment-center level.

Operational implication: Requires deep root-cause analysis and multi-pronged intervention.

Probable Reason Code: MIXED

🌸 Cluster 1 (Pink) — Stable Low-Activity Pincodes

Profile: cov_norm ≈ 0.11, spike_norm = 0, peak_norm ≈ 0.09, severity_norm = 0

Cluster 1 captures pincodes with stable and predictable enrollment behavior. Low volatility and peak stress indicate consistent demand, and the complete absence of spikes and severity suggests no abnormal surges across the observation window. These pincodes form the operational baseline of the system. They require minimal intervention beyond routine operations and can be used as reference benchmarks when evaluating abnormal behavior elsewhere.

Operational implication: Routine operations; no special action required.

Probable Reason Code: STABLE

🟠 Cluster 2 (Orange) — Spiky Event-Driven Pincodes

Profile: cov_norm ≈ 0.19, spike_norm ≈ 0.86, peak_norm ≈ 0.17, severity_norm ≈ 0.41

This cluster is characterized by frequent spike occurrences on an otherwise stable baseline. Volatility and peak stress remain low, but spike frequency is high, and severity is moderate—indicating repeated, localized surges that are noticeable but not catastrophic. These patterns are consistent with event-driven behavior, such as enrollment camps, periodic backlog clearances, or targeted outreach programs. Operationally, these pincodes benefit most from anticipatory planning (event calendars, temporary staffing) rather than permanent capacity expansion.

Operational implication: Routine operations; no special action required.

Probable Reason Code: SPIKY

🟡 Cluster 3 (Yellow) — Peaky but Non-Spiky Structural Load

Profile: cov_norm ≈ 0.85, spike_norm = 0, peak_norm ≈ 0.84, severity_norm = 0

Cluster 3 represents pincodes with large structural swings and high peak months, but without being flagged as anomalous spikes. High volatility and peak factor suggest strong seasonal or cyclical demand patterns, yet the absence of spike frequency and severity implies these changes are gradual or expected, rather than sudden disruptions. These pincodes require forecast-driven capacity planning and seasonal staffing adjustments, not anomaly response mechanisms.

Operational implication: Event-driven planning, early warnings, and short-term staffing buffers.

Probable Reason Code: PEAKY

Reason Code

Reason codes are categorical labels assigned to regions to explain why instability occurs, not just how much instability exists. While numerical scores rank regions by risk, reason codes translate analytical results into interpretable causes—such as chronic volatility, repeated abnormal spikes, extreme peak months, or a combination of factors—making the outputs actionable for operations and policy teams.

The goal of the reason-code logic is to identify the dominant driver of instability for each region (district or PIN), based on the relative strength of normalized instability components. Rather than relying on absolute scores, the method compares components within each region, ensuring interpretability and fairness across geographies.

For each region i, four normalized instability components are available:

$$\mathbf{x}_i = [c_i \quad s_i \quad p_i \quad v_i]$$

where:

- c_i= normalized volatility (CoV)
- s_i= normalized spike frequency
- p_i= normalized peak stress
- v_i= normalized severity

All components are scaled to $[0, 1]$, making them **comparable in magnitude**.

Step 1: Dominance Identification

For each region, the components are ranked internally:

$$x_{i,(1)} \leq x_{i,(2)} \leq x_{i,(3)} \leq x_{i,(4)}$$

Let:

- $x_{i,\max} = x_{i,(4)}$(largest component)
- $x_{i,2\text{nd}} = x_{i,(3)}$(second-largest component)

The index of the maximum component is computed as:$j = \arg\max \{c_i, s_i, p_i, v_i\}$

This identifies which instability dimension is *numerically strongest* for the region.

Step 2: Mixed vs Dominant Classification

To avoid overconfident labeling when multiple drivers are similarly strong, a **margin-based dominance rule** is applied.

A region is classified as **MIXED** if:

$$x_{i,\max} - x_{i,2\text{nd}} \leq \delta$$

where $\delta$is the user-defined margin (default $\delta = 0.10$).

This means that if the strongest and second-strongest components are too close in value, no single driver is considered dominant.

Step 3: Reason Code Assignment

If the dominance condition is satisfied $\left(x_{i,\max} - x_{i,2\text{nd}} > \delta\right)$, the region is assigned a reason code based on the dominant component:

$$\text{Reason}_i = \begin{cases} \text{VOLATILE}, & j = 0 \\ \text{SPIKY}, & j = 1 \\ \text{PEAKY}, & j = 2 \\ \text{INTENSE}, & j = 3 \end{cases}$$

Otherwise:

$$\text{Reason}_i = \text{MIXED}$$

Interpretability

VOLATILE

-Chronic month-to-month instability

-CoV high relative to other districts

-Interpretation: The district's workload is structurally unstable (planning + staffing mismatch, inconsistent throughput, or irregular operations).

SPIKY

-Repeated abnormal events (drives, outages, backlog clearances).

-spike months occur often across the district timeline

-Interpretation: The district is usually stable, but repeatedly gets disrupted by episodic events.

PEAKY

-One or two extreme months causing capacity shock.

-A single "mountain" month. Flat before and after. That one month explains most of the perceived instability.

-Capacity planning problem, not chronic instability. You should plan temporary infrastructure / staff surge for known seasonal or campaign-like peak windows.

MIXED

-Multiple instability drivers → needs deeper investigation.

-Both structural instability and event shocks exist; needs deeper investigation and multi-pronged intervention.

Top 10 Districts by Instability Score (Enrolment) and Reason Code

state	district	months	mean	cov	spike freq	peak factor	severity metric	instability score	reason code
Gujarat	Ahmadabad	12	503.500	3.271	16.7%	11.386	2,048.219	100.000	MIXED
Gujarat	Botad	12	181.417	2.820	16.7%	9.949	1,169.141	100.000	MIXED
Madhya Pradesh	Barwani	12	1,375.667	2.817	16.7%	9.942	1,094.297	100.000	MIXED

state	district	months	mean	cov	spike freq	peak factor	severity metric	instability score	reason code
Gujarat	Gir Somnath	12	342.083	2.781	16.7%	9.825	593.232	100.000	MIXED
Bihar	Sitamarhi	12	2,589.417	2.867	16.7%	10.099	757.278	100.000	MIXED
Maharashtra	Hingoli	12	490.167	2.803	16.7%	9.897	1,135.318	100.000	MIXED
Madhya Pradesh	Morena	12	1,214.750	2.594	16.7%	9.232	1,554.971	97.338	MIXED
Assam	Kamrup	12	784.833	2.585	16.7%	9.206	873.523	97.164	MIXED
Chhattisgarh	Sukma	12	105.833	2.704	16.7%	9.581	398.938	95.152	MIXED
Uttar Pradesh	Varanasi	12	1,218.750	2.441	16.7%	8.745	734.637	94.259	MIXED

Top 10 Pincodes by Instability Score (Enrolment) and Reason Code

state	district	pincode	months	mean	std	cov	peak_factor	spike_freq	severity_metric	Instability score	reason_code
Madhya Pradesh	Sehore	466331	12	96.666	279.29	2.889	10.168	0.1666	1000.978	100	MIXED
Gujarat	Chhotaudepur	391152	9	17.444	47.463	2.720	8.2547	0.1111	227.8812	100	MIXED
West Bengal	Uttar Dinajpur	733129	12	90.916	260.253	2.862	10.0861	0.1666	705.0817	100	MIXED
Uttar Pradesh	Pilibhit	262001	12	350	1019.30	2.912	10.2457	0.1666	1022.391	100	MIXED
Uttar Pradesh	Siddharthnagar	272153	12	177.83	443.758	2.495	8.91846	0.1666	304.2989	100	MIXED
Uttar Pradesh	Sitapur	261204	12	127.16	366.284	2.880	10.1441	0.1666	616.7704	100	MIXED
Bihar	Gaya	824232	12	143.83	404.888	2.814	9.93511	0.1666	348.991	100	MIXED
Bihar	Madhubani	847227	12	45.666	108.737	2.381	8.54014	0.25	152.4376	100	MIXED
Bihar	Madhubani	847108	12	223.75	635.7782	2.841	9.99776	0.1666	147.1982	100	MIXED
Uttar Pradesh	Mathura	281121	12	44.9	122.5846	2.72	9.66233	0.166	762.3673	100	MIXED

Interpretation: High instability implies higher monitoring priority and often need for operational redesign or verification.

Exponential Smoothing for Forecasting of Aadhar Enrolment

What it does: Forecasts next-month enrolment per region using Exponential Weighted Moving Average (EWMA) and produces prediction intervals.

This function implements **Simple Exponential Smoothing (SES)**, also known as an **Exponentially Weighted Moving Average (EWMA)**, to generate a **one-step-ahead forecast** for monthly Aadhaar activity.

For each region, the observed time series $y_t$is smoothed recursively as:

$$\hat{y}_t = \alpha y_t + (1-\alpha)\hat{y}_{t-1}$$

where:

- $\hat{y}_t$is the smoothed level (EWMA),
- $y_t$is the observed value at month t,
- $\alpha \in (0,1]$is the smoothing parameter controlling how much weight is given to recent observations.

The **forecast for the next month** $t+1$is simply the last smoothed level:

$$\hat{y}_{t+1} = \hat{y}_t$$

To quantify uncertainty, the function computes an **EWMA-based standard deviation**, which captures recent variability in the series. A heuristic **95% prediction interval** is then constructed as:

$$[\hat{y}_{t+1} \pm 1.96 \cdot \sigma_{\text{EWMA}}]$$

This approach assumes short-term continuity in enrollment behavior and prioritizes recent trends, making it well-suited for **operational planning and early warning**, where responsiveness is more important than long-horizon accuracy.

Top 10 District Forecasts for Jan 2026 (EWMA)

state	district	forecast month	forecast	pi low	pi high	last value
West Bengal	Murshidabad	Jan 2026	1,018.925	592.401	1,445.450	906.000
West Bengal	South 24 Parganas	Jan 2026	890.802	465.546	1,316.058	790.000
Rajasthan	Jaipur	Jan 2026	798.976	308.366	1,289.586	831.000
West Bengal	North 24 Parganas	Jan 2026	777.983	599.212	956.755	724.000
Maharashtra	Thane	Jan 2026	770.159	0.000	1,716.954	732.000
Maharashtra	Pune	Jan 2026	765.034	395.843	1,134.224	768.000
Bihar	East Champaran	Jan 2026	609.132	105.286	1,112.977	603.000
West Bengal	Uttar Dinajpur	Jan 2026	604.618	217.043	992.192	545.000
Uttar Pradesh	Bahraich	Jan 2026	593.328	0.000	1,767.230	615.000
Maharashtra	Nashik	Jan 2026	577.492	287.343	867.640	575.000

Top 10 Pincode Forecasts for Jan 2026 (EWMA)

state	district	pincode	Forecast month	forecast	Pi low	Pi high	Last value
Gujarat	Banas Kantha	385310	Jan-26	879.5	0	3311.805906	2
Uttar Pradesh	Moradabad	244001	Jan-26	152.0019	0	702.746689	141
Maharashtra	Aurangabad	431001	Jan-26	125.91796	0	378.6520145	118
Meghalaya	West Khasi Hills	793119	Jan-26	120.40869	0	606.3266664	181
West Bengal	Uttar Dinajpur	733207	Jan-26	118.80664	35.02718	202.5860962	114
West Bengal	Murshidabad	742202	Jan-26	112.8691	48.56873	177.1695487	109
Uttar Pradesh	Saharanpur	247001	Jan-26	110.28955	0	440.1559559	98

state	district	pincode	Forecast month	forecast	Pi low	Pi high	Last value
Delhi	West Delhi	110059	Jan-26	110.078125	0	503.3608096	108
Uttar Pradesh	Ghaziabad	201102	Jan-26	107.1088	0	331.3401906	85
Uttar Pradesh	Aligarh	202001	Jan-26	105.9707	0	535.7188674	114

Forecasting is used by decision makers to **anticipate near-term operational load** and prepare resources in advance. By providing a one-month-ahead estimate along with an uncertainty range, the forecast helps planners decide **how much staffing, infrastructure, and monitoring capacity** will be required in the upcoming period. When combined with instability and anomaly indicators, forecasting enables proactive actions—such as activating surge capacity in high-risk regions or reallocating resources from stable areas—thereby reducing the likelihood of service congestion and reactive crisis management.

Consistent performers

What it does: Identifies stable regions with low CoV, low median month-to-month percent change, and meaningful volume.

What the calculation evaluates: Measures operational predictability and stability.

Why it matters for UIDAI decisions: Stable regions become benchmarks and can be monitored with lower intensity.

Dashboard

The **Aadhar Enrollment Operations Dashboard** is an interactive decision-support application designed to provide a consolidated, data-driven view of Aadhaar enrollment and update activities across India.

The dashboard enables to:

- Monitor **overall system workload**
- Distinguish between **new enrollments and maintenance operations**
- Identify **geographic concentration and operational imbalance**
- Detect **temporal instability and abnormal spikes**
- Segment districts based on **behavioral patterns**

The application is built using **Streamlit** and integrates multiple Aadhaar operational datasets into a unified analytics and visualization layer, enabling both **high-level executive insight** and **deep operational diagnostics**.

High-Level Architecture

At a high level, the application follows a layered architecture:

1. **Data Ingestion Layer**
    - CSV-based data loading
    - Schema normalization and date parsing

2. **Processing & Metrics Layer**
    - Transaction unification
    - KPI computation
    - Aggregation and statistical transformations
3. **Analytics Layer**
    - Monthly aggregation
    - Spike detection
    - Instability scoring
    - Clustering and dimensionality reduction
4. **Visualization Layer**
    - Executive KPI ribbon
    - Operational charts
    - Geographic maps
    - Advanced analytical visualizations
5. **Presentation Layer (Streamlit UI)**
    - Interactive filters
    - Modular charts
    - Optional advanced analytics section

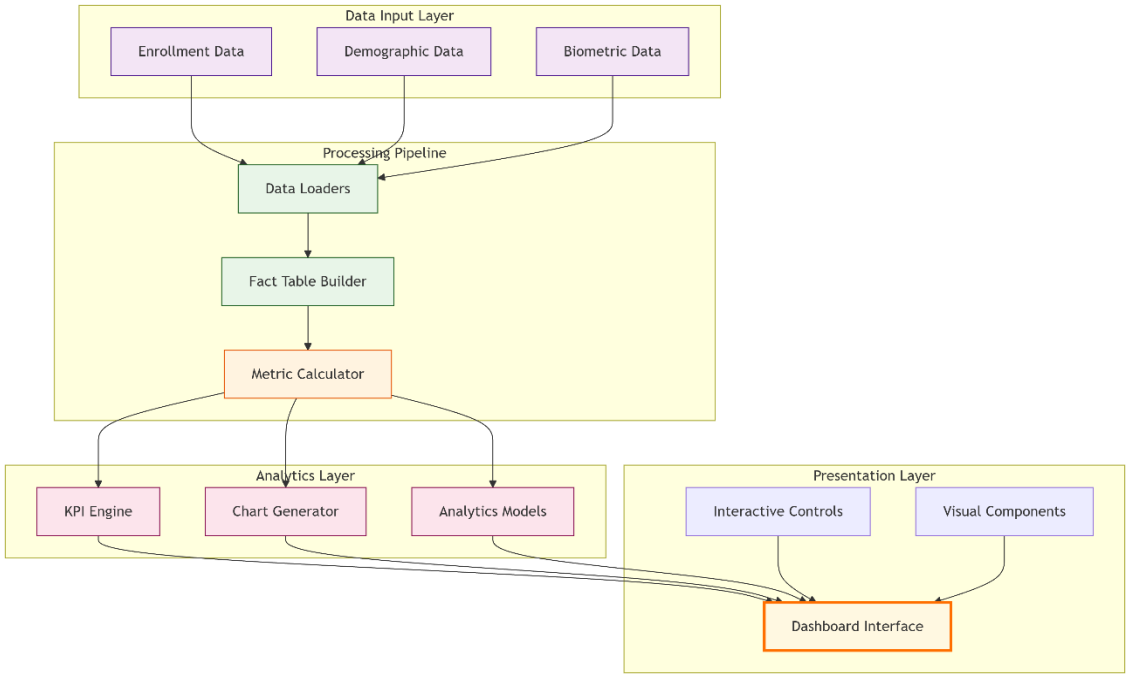


Figure 1

Backend File Execution Flow

The dashboard backend follows a **controller-driven execution model**, with app.py acting as the single-entry point.

1. app.py — Entry Point & Orchestrator

- This is the **only file executed directly** by Streamlit.
- On every run or user interaction, Streamlit executes app.py **top to bottom**.
- app.py does **not** contain business logic; it coordinates other modules.

Responsibilities:

- Load data
- Apply filters
- Call metrics, analytics, and visualization functions
- Render outputs to the UI

2. processing/loaders.py — Data Preparation

- Called by app.py during data ingestion.
- Standardizes schemas across datasets.
- Builds the unified **transaction fact table**.

Executed when:

- Data is first loaded or rerun occurs.

Output:

- Cleaned DataFrames
- fact_df used for KPIs

3. processing/metrics.py — KPI Computation

- Called by app.py after filters are applied.
- Computes executive metrics using the fact table.
- Contains only aggregation logic (no visualization).

Executed when:

- executive_kpis() is called.

Output:

- Dictionary of KPI values

4. analytics/model.py — Advanced Analytics Engine

- Invoked indirectly through chart functions.
- Performs:
    - Monthly aggregation

    - Spike detection
    - Instability scoring
    - Clustering and PCA
- Heavy computations are cached.

Executed when:

- Advanced analytics charts are requested.

Output:

- Analytical DataFrames
- Intermediate results for plotting

5. visuals/charts.py — Chart Construction

- Called by app.py to generate charts.
- Uses processed or analytical data.
- Returns **Plotly figure objects** only.

Executed when:

- Corresponding chart function is rendered.

Output:

- Plotly figures

6. visuals/kpis.py — KPI Rendering

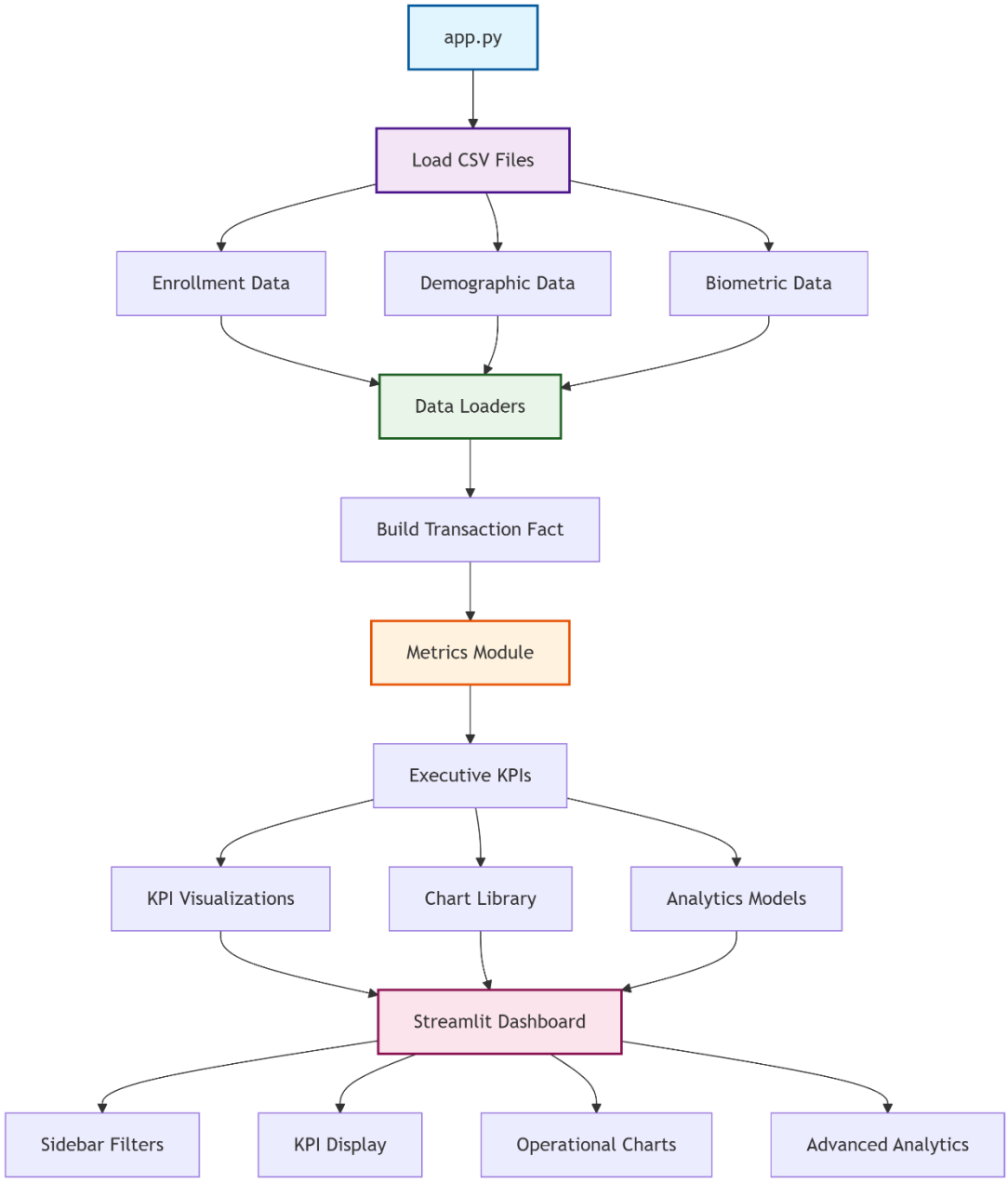
- Called by app.py after KPI computation.
- Formats KPI values (K/M notation).
- Renders metrics in Streamlit.

Output:

- KPI ribbon in UI

7. visuals/maps.py — Geographic Visuals

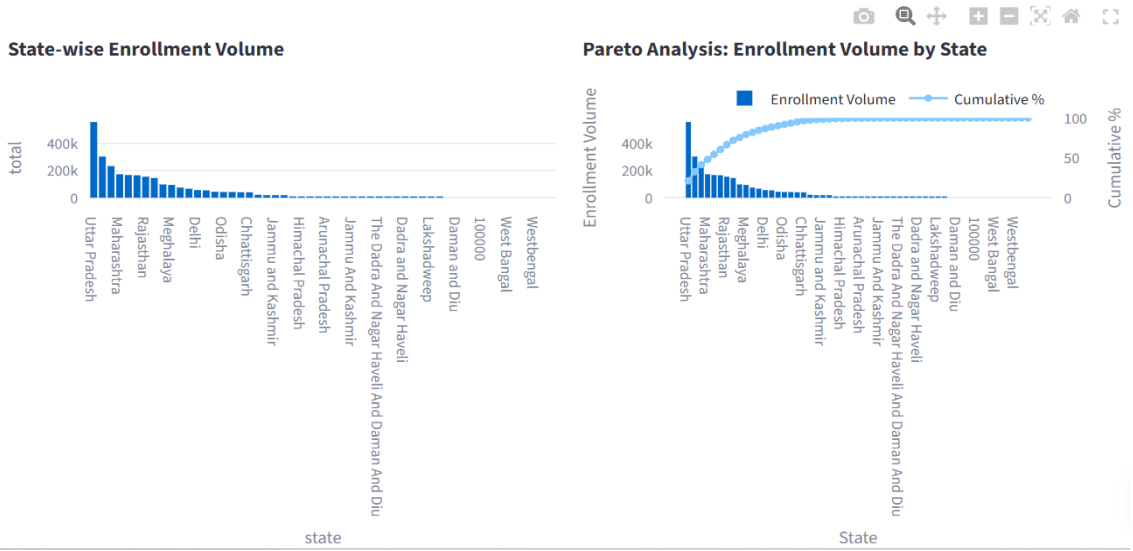
- Called by app.py for spatial analysis.
- Converts aggregated data into map-based Plotly visuals.

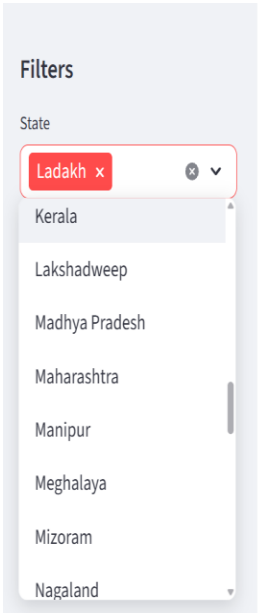

app.py
Load CSV Files
Enrollment Data
Demographic Data
Biometric Data
Data Loaders
Build Transaction Fact
Metrics Module
Executive KPIs
KPI Visualizations
Chart Library
Analytics Models
Streamlit Dashboard
Sidebar Filters
KPI Display
Operational Charts
Advanced Analytics

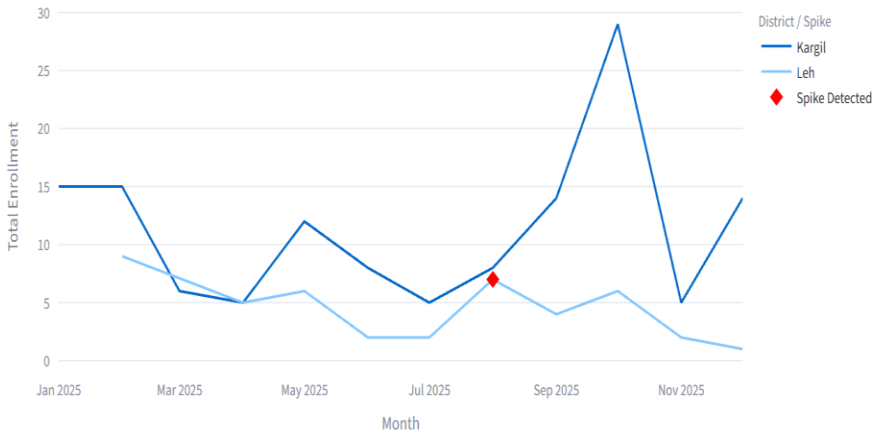
Aadhaar Enrollment & Operations Dashboard

Total Transactions	Enrollments	Updates	Update %	Top State Share	Concentration Index
121.70M	2.64M	119.06M	97.8%	15.3%	0.0715

Operational Overview (Enrollment Dataset Only)



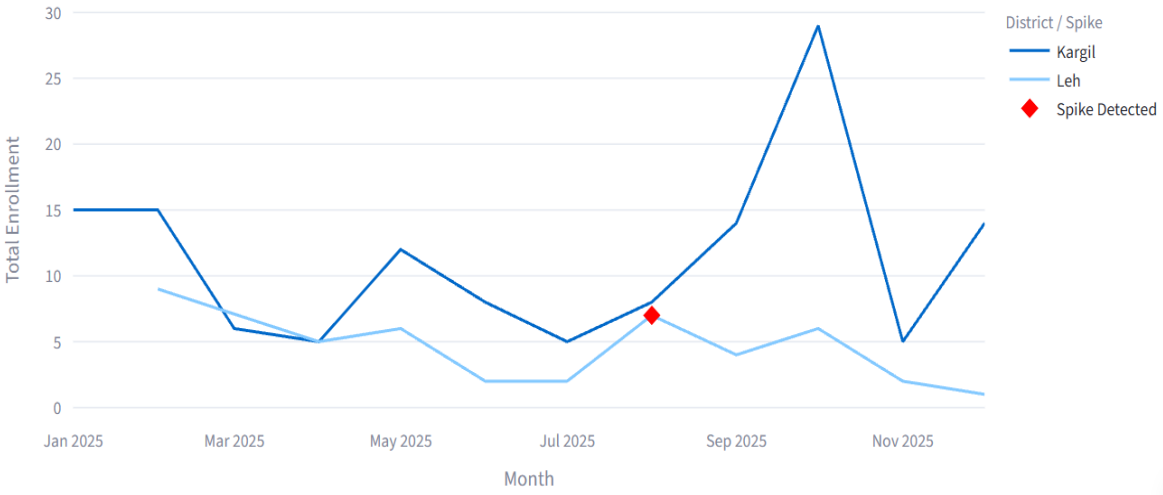




D

📈 Monthly Enrollment Spikes

Monthly Enrollment Spikes

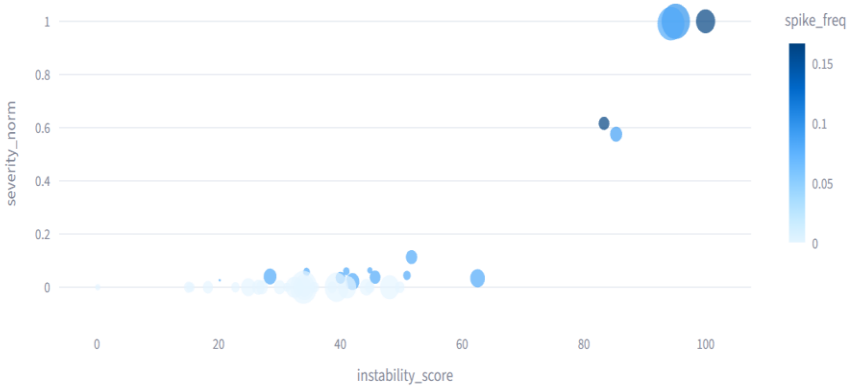


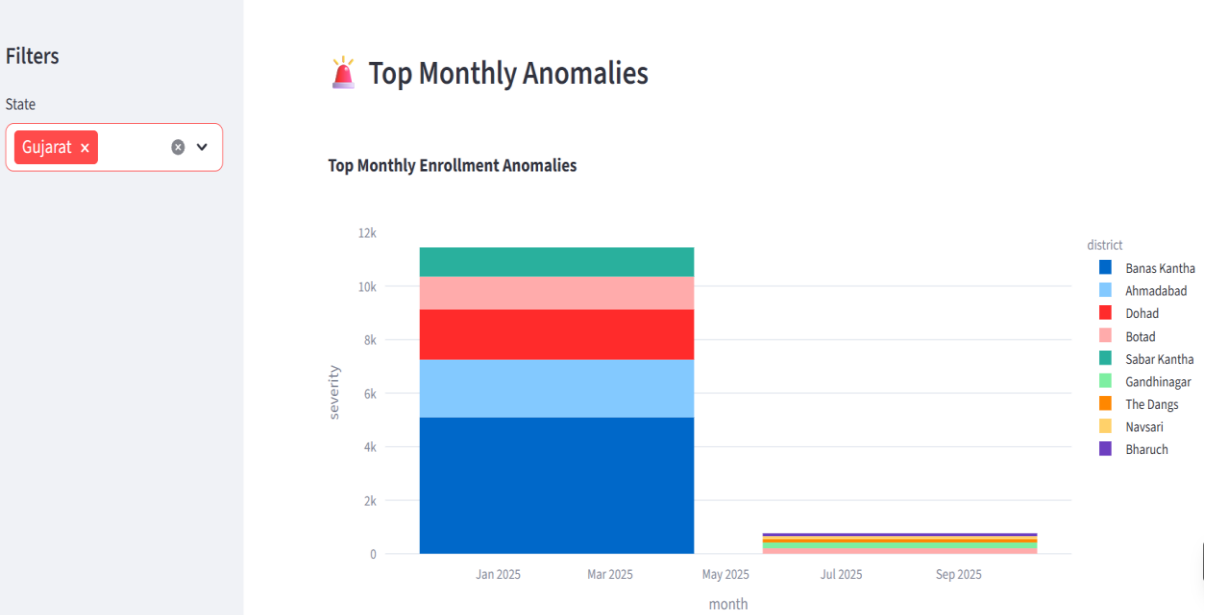
Filters

State

Gujarat ×

District Instability vs Severity





Note: These charts are generated for different states as per filter. For the working of the dashboard please refer to the video link given below.

Webpage Preview:

Video Link: https://drive.google.com/file/d/1m22SNSsf2C6selg0rqzFpe9Id_8jScWa/view?usp=sharing

Github link to access the code and output files.

https://github.com/Aayushivyas/UIDAI-Hackthon