PROJECT REPORT

TEAM - UIDAI_2536

Github link:

https://github.com/khushi-infinity/Aadhar_Analysis

1.Problem Background & Motivation

Aadhaar is world's largest biometric identity system (~1.3 billion enrollees), clearly indicating that the system dominate workload.

The central questions addressed are:

How UIDAI can use 2025 Aadhaar data to plan capacity, infrastructure, and policy interventions for 2026 and beyond?

How can Aadhaar enrollment, demographic update, and biometric update data be analyzed to understand operational demand, regional disparities, behavioral patterns, and future infrastructure needs across India?

This creates three strategic needs:

- Anticipate **when** and **where** demand peaks will occur (temporal and spatial forecasting)
- Identify **which districts and states** are under stress (high-churn, high-ratio, or anomalous behavior).
- Design a **targeted infrastructure and policy response** (centers, mobile units, awareness, data quality) to ensure service quality and equity.

Framed in the “what–why–what next” lens:

- **What is happening**: Aadhaar usage is dominated by updates, with strong seasonality and regional inequality.
- **Why it is happening**: Demographic transition, urbanization, migration, and uneven infrastructure/awareness.
- **What should be done**: Cluster-based infrastructure planning, focus on high-churn districts, and early-childhood and awareness interventions.

2. Dataset Description

Three primary UIDAI datasets were used, all for calendar year 2025, covering around **763 districts** across **36 states/UTs** after cleaning and standardization.

2.1 Aadhaar Enrollment Dataset

Purpose: Tracks new Aadhaar enrollments across age groups.

Columns Used: date, state, district, pincode, age_0_5, age_5_17, age_18_greater

Scale :

- ~10 lakh records after merging monthly files
- Covers all states and union territories

Relevance:

- Measures enrollment saturation
- Helps identify regions with low Aadhaar coverage
- Enables age-based enrollment analysis

2.2 Aadhaar Demographic Update Dataset

Purpose: Captures updates related to name, address, date of birth, gender, etc.

Columns Used: date, state, district, pincode, demo_age_5_17, demo_age_17_plus

Scale:

- ~20 lakh records
- High-frequency update dataset

Relevance:

- Indicates migration, marriage, correction cycles
- Reflects demographic transitions and mobility

2.3 Aadhaar Biometric Update Dataset

Purpose: Records biometric re-enrollments (fingerprint and iris).

Columns Used: date, state, district, pincode, bio_age_5_17, bio_age_17_plus

Scale: ~18 lakh records

Relevance:

- Strongly linked to employment, banking, and compliance
- Indicates maturity and security maintenance of Aadhaar system

---

3. Methodology and Solution Approach

3.1 Data collection and challenges

Data collection:

- Multiple CSV files per theme (enrollment, demographic, biometric) for 2025 were concatenated to form full-year datasets.

Key challenges:

- **Duplicates**: All three datasets had significant duplicates, which were removed to avoid over-counting.
- **Inconsistent naming**: States/districts had spelling and naming variations (e.g., “Delhi” vs “New Delhi”); normalized via mapping.
- **Type mismatches**: Date columns stored as strings and numeric columns with formatting noise; converted to datetime and numeric types.
- **Missing values**: Handled by dropping or aggregating, depending on impact on district or month totals.

3.2 Integration and feature engineering

Merge logic:

- The three datasets were merged into a master DataFrame on **date, state, district, pincode**.
- Column prefixes: **enroll_** for enrollment, **demo_** for demographic updates, **bio_** for biometric updates.

Engineered features:

- **total_enrollment** = enroll_age_0_5 + enroll_age_5_17 + enroll_age_18_plus.
- **total_demo_updates** = demo_age_5_17 + demo_age_17_plus.
- **total_bio_updates** = bio_age_5_17 + bio_age_17_plus.
- **total_updates** = total_demo_updates + total_bio_updates.
- **update_to_enroll_ratio** = total_updates ÷ total_enrollment.
- **month** extracted from date for time-series analysis.

These features enabled unified analysis across functions and time, and the creation of derived KPIs for clustering and forecasting.

3.3 Aggregation Strategy

Data was aggregated at three levels:

- **District level (~763 districts)**: total_enrollment, total_demo_updates, total_bio_updates, update_to_enroll_ratio.
- **State/UT level (36 units)**: totals and ratios for ranking and inequality analysis.
- **Monthly national level (Mar–Dec 2025)**: time-series of enrollments and updates.

This multi-level aggregation supports both **macro** (national, state) and **micro** (district) insights.

Month 2025	Enrollment	Demo updates	Bio updates
March	16,582	8,797,695	8,578,228
April	273,737	977,613	8,641,679
July	676,317	1,619,721	9,792,552
September	1,839,644	7,452,260	7,795,565
November	1,309,561	8,301,109	7,778,956
December	843,874	7,946,674	8,969,038

Transformations:

- 3-month rolling averages to smooth volatility.
- Month-on-month growth rates to detect spikes and drops.
- Ratios and scaling to compare districts/states with very different sizes.

3.4 Analytical Methods

- **Univariate analysis**: distributions of enrollment, updates, and ratios across states/districts.
- **Bivariate analysis**: correlations between enrollment and demographic/biometric updates; between enrollment and update-to-enrollment ratio.
- **Multivariate state×month analysis**: seasonal patterns and cross-metric behaviors in top 3–5 states.
- **K-Means clustering (4 clusters)**: grouping districts by enrollment and update volumes and ratios.
- **Lifecycle classification**: labeling districts as update-heavy, legacy-only, balanced, or enrollment-heavy.

- **Time-series trend analysis**: month-on-month growth and rolling averages for each stream.
- **Forecasting**: seasonal decomposition of 2025 patterns and a 10–15% saturation factor to forecast 2026 enrollment load.

---

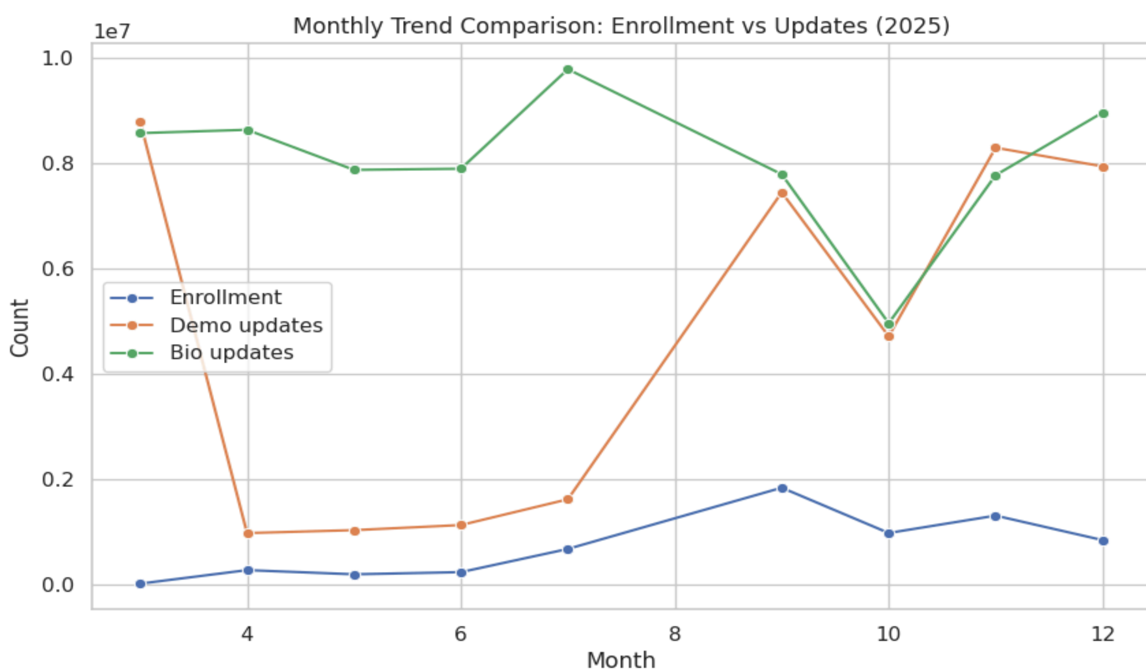
4. Experiments & Results

4.1 Monthly Trends

Using aggregated monthly totals, the following patterns emerged:

- **Monthly trend analysis (2025)**
  - **Peak enrollment**: September 2025 with **1,839,644** enrollments (≈11× March).
  - **Peak demographic updates**: March (8,797,695) and November (8,301,109).
  - **Peak biometric updates**: July (9,792,552) and December (8,969,038).

Result: Enrollment and updates show **strong seasonality**, with different peaks for demographic and biometric streams.



4.2 State-Level Ratios and Volumes

Top 8 states/UTs by update-to-enrollment ratio:

State/UT	Ratio

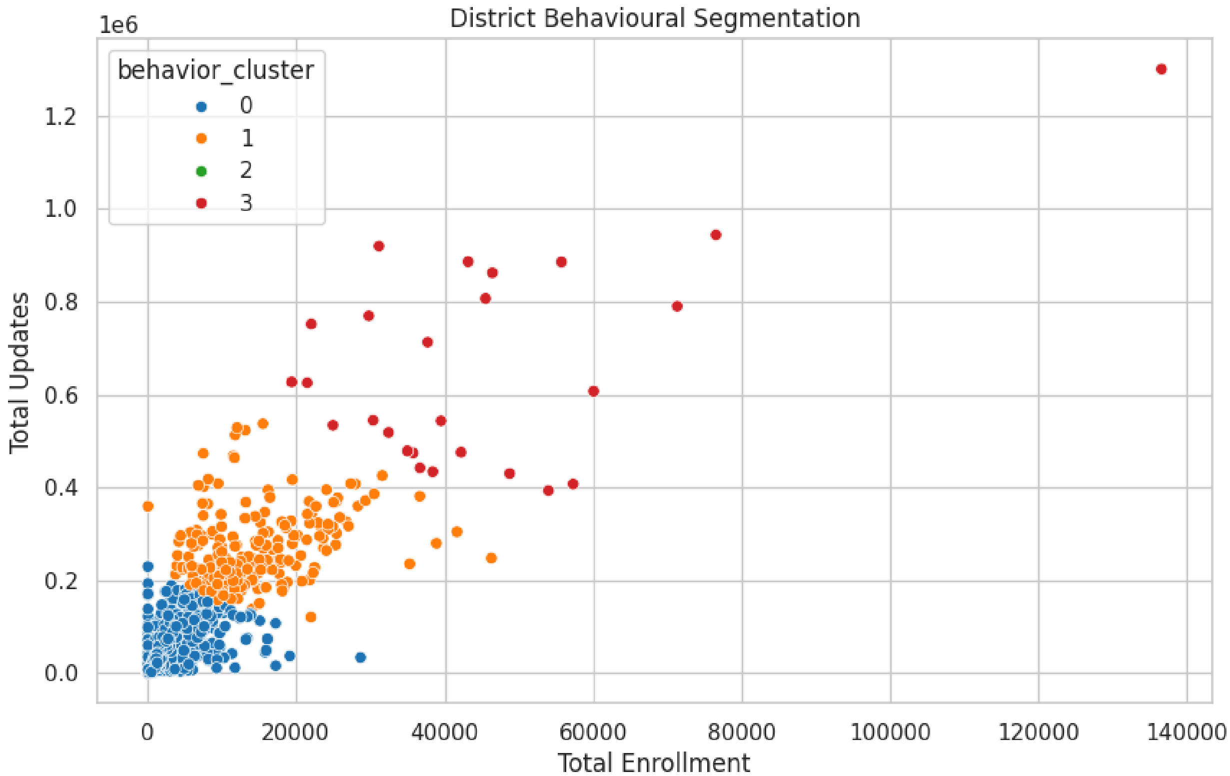
Chandigarh	49.98
Andaman & Nicobar	42.74
Goa	41.05
Manipur	38.58
Chhattisgarh	37.29
Tripura	34.73
Andhra Pradesh	34.24
Maharashtra	30.57

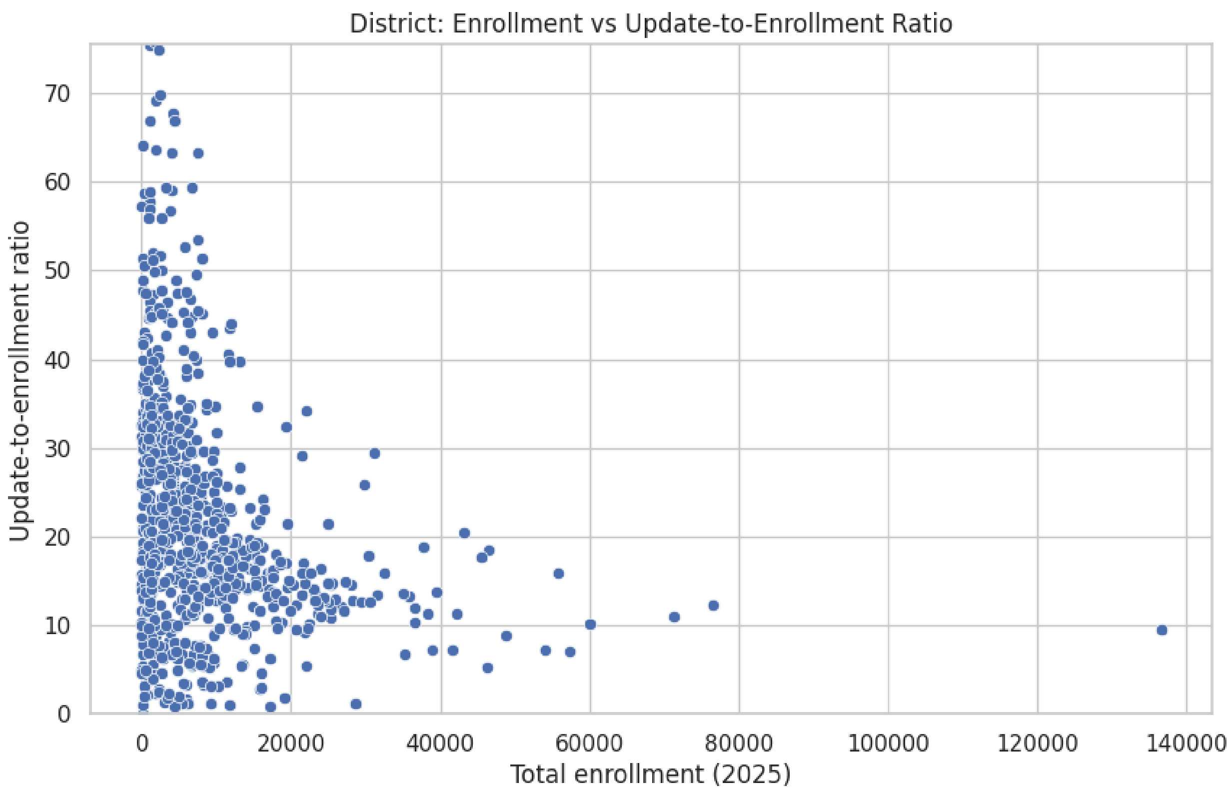
Combined with absolute volumes, Maharashtra and Andhra Pradesh are **workload giants**, while Small UTs show very high ratios, indicating **update-heavy usage** over fresh enrollments.

4.3 District Clustering and Lifecycle

K-Means cluster centroids:

Cluster	Avg enroll	Avg demo	Avg bio	Avg ratio	Interpretation
0	3,387	21,360	42,291	23.11	Small, infrastructure-constrained
1	14,022	93,946	167,129	22.73	Moderate-capacity growth zones
2	9	95	27,194	3,032.11	Anomaly / crisis districts
3	45,083	305,262	354,732	16.90	Metro/high-capacity hubs





Lifecycle classification:

Stage	Count	Interpretation

Update-heavy	765	Mature Aadhaar maintenance stage
Legacy-only	55	Only updates, no new enrollments
Balanced	1	Rare equilibrium
Enrollment-heavy	1	Early-stage adoption

Result: Almost all districts are **update-heavy/legacy-only (≈99.7%)**, confirming Aadhaar's mature, maintenance-focused lifecycle.

4.4 High-Churn Districts

High update-to-enrollment ratio (>50) identifies high-churn districts:

District	State	Enroll	Updates	Ratio	Driver
Mahabubnagar	Andhra Pradesh	95 or 9*	27,194	286–3,032:1	Data anomaly / crisis
Imphal East	Manipur	1,103	105,319	95:1	In-migration + re-certification
Serchhip	Mizoram	100	9,210	92:1	Remote, high errors
Wardha	Maharashtra	1,103	99,963	90:1	Migration and employment
Ratnagiri	Maharashtra	2,512	175,610	69:1	Coastal seasonal labour

*Different views show 9 or 95 enrollments for Mahabubnagar; in both cases, ratios are impossibly high.

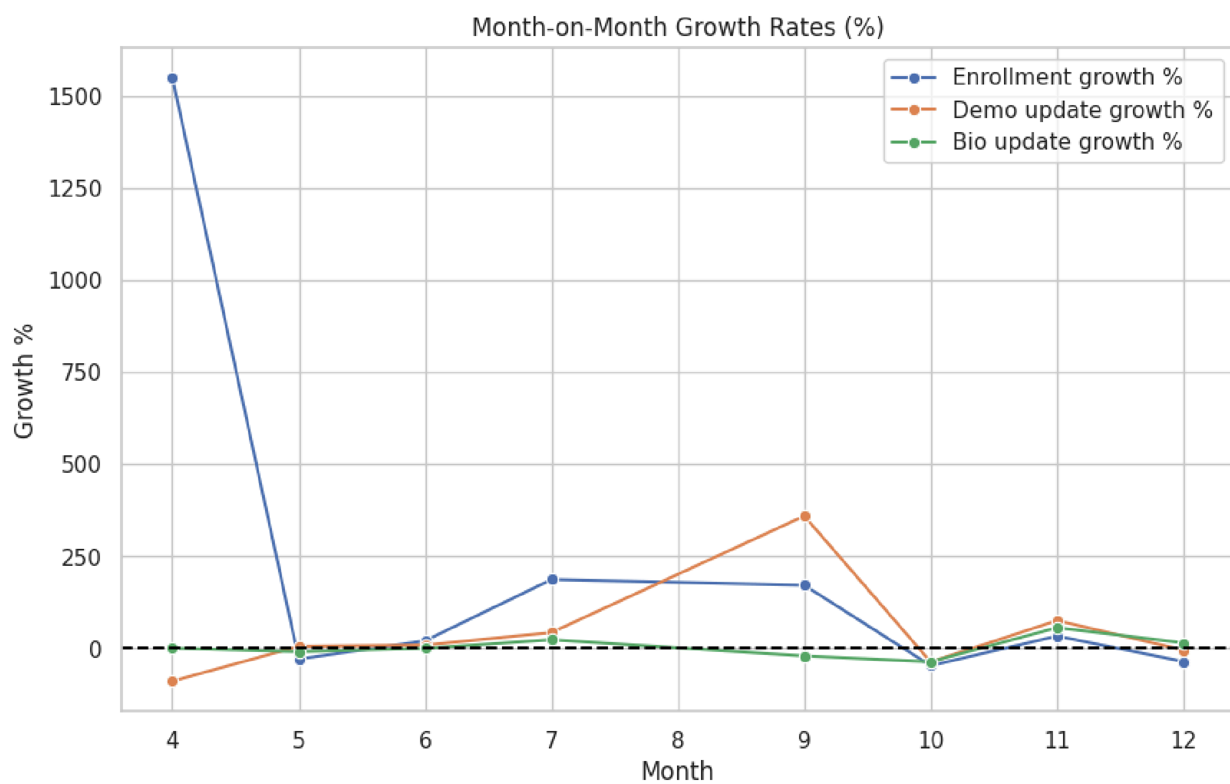
Result: High-churn districts cluster in **Manipur, Mizoram, Maharashtra, Andhra Pradesh, Punjab**, signaling both migration-driven churn and data-quality issues.

4.5 Forecasting Experiments

From 2025 trends and a saturation adjustment, forecasted **2026 enrollment**:

Month	2025 actual	2026 forecast	Change	Capacity load
Jan	~200K proj.	180K	-10%	Low

Apr	273,737	320K	+17%	Medium
June	235,286	280K	+19%	Med-high
July	1,619,721	1.6M	-1%	Peak
Aug	~1.4M proj.	1.35M	-4%	Peak
Sept	1,839,644	1.75M	-5%	Peak
Oct	979,011	900K	-8%	High
Nov	1,309,561	1.2M	-8%	High
Dec	843,874	750K	-11%	High



- **Annual 2026 forecast**: **~9.5–10M** enrollments, about 8% below 2025's ~10.3M.
- **January 2026** updates forecast:
  - Demographic: 8.0–8.5M.
  - Biometric: 8.2–8.8M.
  - Enrollments: 0.9–1.1M.

---

5. Insights & Business Interpretation

5.1 What is happening

- Aadhaar has become a **maintenance system**: update volumes vastly exceed new enrollments, and nearly all districts are in update-heavy lifecycle stages.
- Demand is **temporal and spatially concentrated**:
    - Temporal: enrollment peaks in **July–September**, demographic updates in **March/Nov**, biometric updates in **July/Dec**.
    - Spatial: absolute volumes are highest in **Maharashtra, Uttar Pradesh, Andhra Pradesh, West Bengal, Bihar**, while **Chandigarh, Goa, Andaman & Nicobar, Manipur** are extremely update-intense.

5.2 Why it is happening

- **Demographic transition & saturation**: Most adults and school-age children are already enrolled, so activity shifts to updating records, consistent with a post-expansion demographic system.
- **Urbanization & migration**: Metros and industrial/coastal districts show high update-to-enrollment ratios and heavy biometric activity, reflecting employment-driven and migration-driven churn.
- **Infrastructure and awareness gaps**: Remote and North-Eastern districts have low enrollments but high update ratios, indicating insufficient center coverage, poor initial documentation, and limited awareness.

5.3 What Should Be Done – Operational Recommendations

1. **Stabilize anomalies and high-churn districts**
    - Audit Mahabubnagar's 3,032:1 ratio and similar anomalies for data duplication or structural errors.
    - Create a **"Churn Reduction Task Force"** for top 10 churn districts (Imphal East, Serchhip, Wardha, Gadchiroli, Ratnagiri, Mansa, Srikakulam, etc.) with a target of **30% churn reduction** via digital updates, mobile camps, and employer partnerships.
2. **Implement 3-tier infrastructure model**
    - **Tier 1: Mega-centers** in metros (Mumbai, Bengaluru, Delhi NCR, Hyderabad, Chennai), each with **50+ biometric stations**, strong queue management, and appointment systems; ~61 mega-centers recommended.
    - **Tier 2: Regional hubs** in high-churn districts (Imphal, Serchhip, Wardha, Gadchiroli, Ratnagiri, Mansa, plus Pune, Ahmedabad, Jaipur, Lucknow, Kolkata), with 15–25 stations and one mobile unit per hub.
    - **Tier 3: Mobile vans** for low-volume rural districts (~33 vans for ~163 districts), running 2–3 village camps per month.
3. **Phase-wise rollout aligned to seasons**
    - **Phase 1 (0–3 months)**: Activate 15 mega-centers + 5 hubs; cost ≈₹50 Cr.
    - **Phase 2 (3–9 months)**: Add 46 mega-centers + 17 hubs; cost ≈₹120 Cr.

    - **Phase 3 (9–18 months)**: Deploy 33 mobile vans; cost ≈₹60 Cr. Total ≈₹230 Cr.
4. **Target early childhood enrollment and awareness**
    - Embed Aadhaar enrollment units into **50 high-birth government hospitals** (AIIMS Delhi, KEM Mumbai, CMC Vellore, Osmania Hyderabad, and selected rural hospitals) to register newborns at or soon after birth.
    - Run awareness campaigns in **Chandigarh, Goa, Manipur, Andaman & Nicobar** on correct documentation and one-time accurate enrollment, to reduce repeated corrections.
5. **Strengthen data quality and monitoring**
    - Introduce **real-time validation rules** for impossible update ratios and abnormal month-on-month spikes, with automated alerts to state offices.
    - Institutionalize this analysis as an **annual monitoring framework**, with dashboards tracking cluster movement, lifecycle stage shifts, and saturation levels.

---

6. Limitations & Future Scope

6.1 Limitations

- **Age-group detail**: While the framework supports age analysis (0–5, 5–17, 17+), some key age-wise outputs remain hidden in collapsed notebook cells, limiting precise quantification of early-childhood gaps.
- **Forecasting rigor**: Forecasts use seasonal decomposition plus a saturation assumption; they are not yet backed by dedicated models like SARIMA/Prophet with formal error metrics.
- **Data quality dependence**: Extreme anomalies (e.g., Mahabubnagar) show that mis-recorded counts or duplicates can distort ratios, so insight quality depends heavily on upstream data integrity.

6.2 Future Scope

- **Advanced time-series models**: Build state- and cluster-level forecasting models (SARIMA/Prophet/XGBoost) and validate against held-out monthly data to quantify forecast error.
- **Richer age-group analytics**: Fully expose and analyze 0–5 vs 5–17 vs 17+ trends to better design child, student, and worker-focused interventions.
- **Policy-linked covariates**: Integrate external signals (academic calendars, major schemes, elections, natural disasters) to explain peaks and improve predictive power.
- **Real-time monitoring system**: Turn this one-off analysis into a live system with monthly auto-refresh and alerts for high-churn or anomaly districts.
