# Project Report

## Unlocking Societal Trends in Aadhaar Enrolment and Updates

## CHAPTER 1- PROBLEM STATEMENT AND APPROACH

### Problem Statement

Aadhaar is India's largest digital identity system, covering more than a billion residents. While Aadhaar enrolment has reached maturity across most regions, continuous demographic and biometric updates reflect population mobility, lifecycle transitions, and administrative demand.

Despite the availability of aggregated Aadhaar data from UIDAI, there is limited structured analysis that connects enrolment trends with update behaviour across age groups, regions, and time. Understanding these patterns is crucial for improving service delivery, planning update drives, and optimising Aadhaar infrastructure.

**The problem addressed in this project is:**

*How can Aadhaar enrolment and update data be analysed to uncover meaningful societal trends, operational signals, and actionable insights that support informed decision-making and system improvements?*

### Analytical Approach

The project adopts a **data-driven exploratory and comparative analysis approach**, focusing on:

- Aggregated UIDAI enrolment and update datasets
- Age-wise, state-wise, and year-wise analysis
- Comparative analysis between enrolments and updates
- Simple anomaly detection to identify unusual patterns

The analysis is implemented entirely using **Python and Jupyter notebooks**, ensuring transparency, reproducibility, and interpretability.

# CHAPTER 2- DATASET USED

The analysis uses **official aggregated datasets published by UIDAI**, ensuring data authenticity and relevance.

### 2.1 Aadhaar Enrolment Dataset

This dataset provides aggregated information on Aadhaar enrolments across India.

**Key columns used:**

- date – Date of enrolment record
- state – State or Union Territory
- district – District name
- pincode – PIN code
- age_0_5 – Enrolments for age group 0–5 years
- age_5_17 – Enrolments for age group 5–17 years
- age_18_greater – Enrolments for age group 18+ years
- year – Extracted year for temporal analysis

**Derived column:**

- total_enrolment = age_0_5 + age_5_17 + age_18_greater

### 2.2 Aadhaar Demographic Update Dataset

This dataset captures aggregated demographic updates made to Aadhaar records.

**Key columns used:**

- state, district, pincode
- demo_age_5_17 – Demographic updates for age 5–17
- demo_age_17_ – Demographic updates for age 17+
- year

**Derived column:**

- total_updates = demo_age_5_17 + demo_age_17_

### 2.3 Aadhaar Biometric Update Dataset

This dataset reflects biometric updates, which are often mandatory during lifecycle transitions.

**Key columns used:**

- state, district, pincode
- bio_age_5_17 – Biometric updates for age 5–17
- bio_age_17_ – Biometric updates for age 17+
- year

**Derived column:**

- total_biometric_updates = bio_age_5_17 + bio_age_17_

## Note:

Raw and processed datasets are excluded from public sharing due to size constraints. All analysis is reproducible using the provided notebooks.

# CHAPTER 3-METHODOLOGY

The project follows a structured analytics pipeline:

## 3.1 Data Loading and Understanding

- Multiple CSV files per dataset were combined into unified DataFrames
- Column structure, data types, and completeness were examined
- Source file metadata was preserved for traceability

## 3.2 Data Cleaning and Preprocessing

- Standardised column naming conventions
- Removed inconsistencies and ensured numeric integrity
- Extracted year from date fields
- Created derived total columns from age-wise aggregates

This step ensured that all datasets were **analysis-ready** and consistent across notebooks.

## 3.3 Exploratory Data Analysis (EDA)

EDA was conducted separately for:

- Aadhaar enrolment
- Demographic updates
- Biometric updates

The analysis focused on:

- State-wise distributions
- Age-group contributions
- Year-wise trends

### 3.4 Comparative Analysis

To understand system maturity and maintenance demand:

- Enrolments were compared against demographic and biometric updates
- Update-to-enrolment ratios were calculated at the state level
- Year-wise trends across datasets were visualised

### 3.5 Anomaly Detection

Simple statistical techniques were used:

- Mean and standard deviation based spike detection
- Interquartile Range (IQR) for state-level outliers

This helped identify:

- Sudden surges in updates
- Regions with unusually high maintenance demand

# CHAPTER 4- DATA ANALYSIS AND VISUALISATION

This section summarises **key findings** along with visualisations developed during the analysis.

## 4.1 Aadhaar Enrolment Analysis

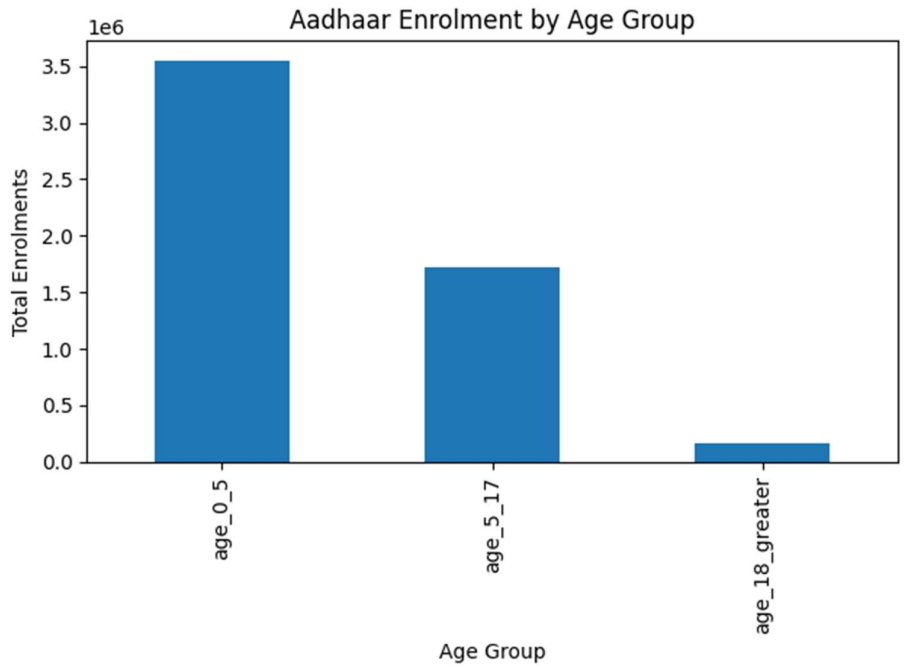


**FIG 1:Age-group enrolment distribution**

**Insight:**

- Adult (18+) enrolments form the majority
- Lower enrolment among children indicates saturation
- Highlights lifecycle enrolment patterns

**Code Snippet :**

```
enrol_df.groupby("state")["total_enrolment"].sum()
```

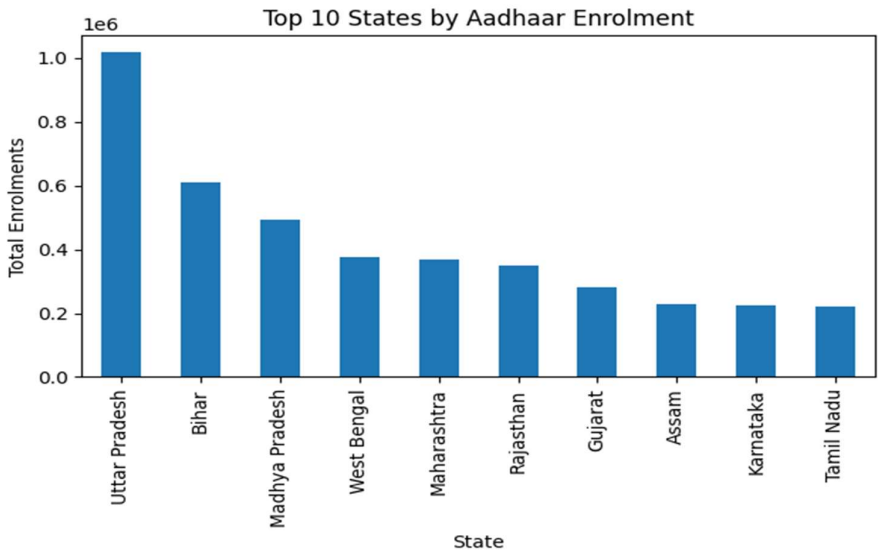


**FIG 2:State-wise enrolment bar charts**

## Insights:

- Enrolment varies significantly across states
- High-population states dominate total enrolments
- Reflects demographic size and early adoption

## Code Snippet :

```
enrol_df[["age_0_5","age_5_17","age_18_greater"]].sum()
```

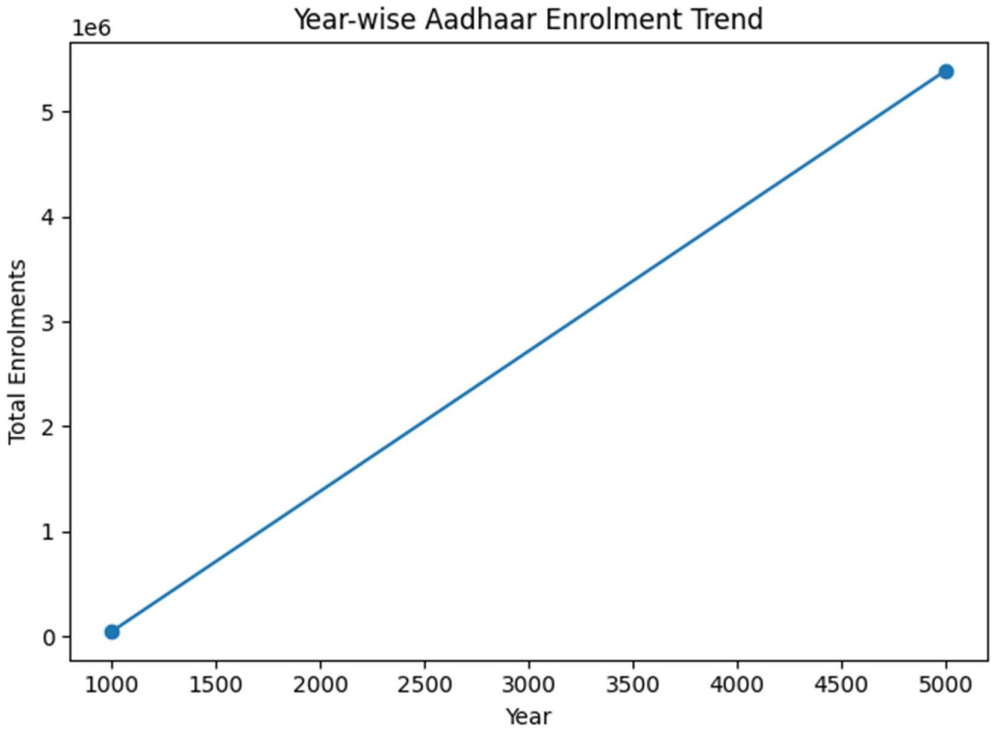


**FIG 3-Year-wise enrolment trends**

## Insight:

- Enrolments increase sharply during initial years
- Growth stabilises in later years
- Indicates Aadhaar system maturity

## Code Snippet :

```
enrol_df.groupby("year")["total_enrolment"].sum()
```

## 4.2 Demographic Update Analysis

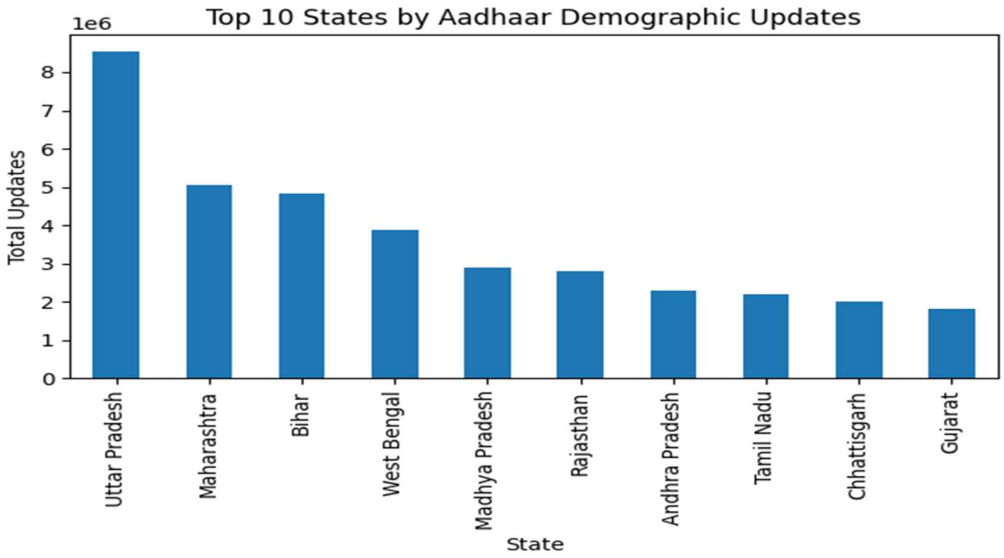


**FIG 4-State-wise demographic updates**

**Insight:**

- Some states show disproportionately high update volumes
- Reflects migration and address changes

**Code Snippet :**

```
demo_df.groupby("state")["total_updates"].sum()
```

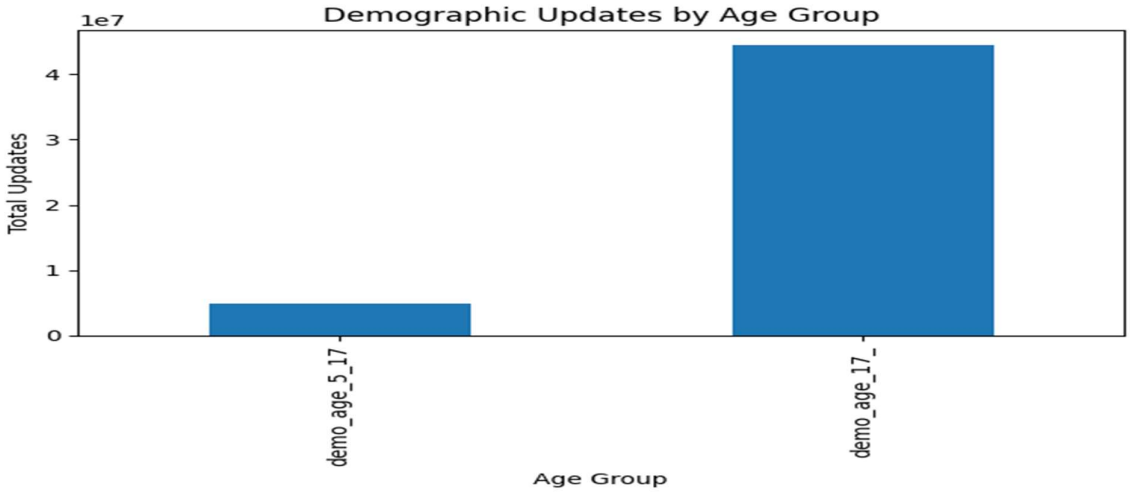


**FIG 5-Age-group update distribution**

## Insight:

- Majority of updates are driven by adults (17+)
- Indicates higher identity maintenance needs

## Code Snippet :

```
demo_df.groupby("state")["total_updates"].sum()
```

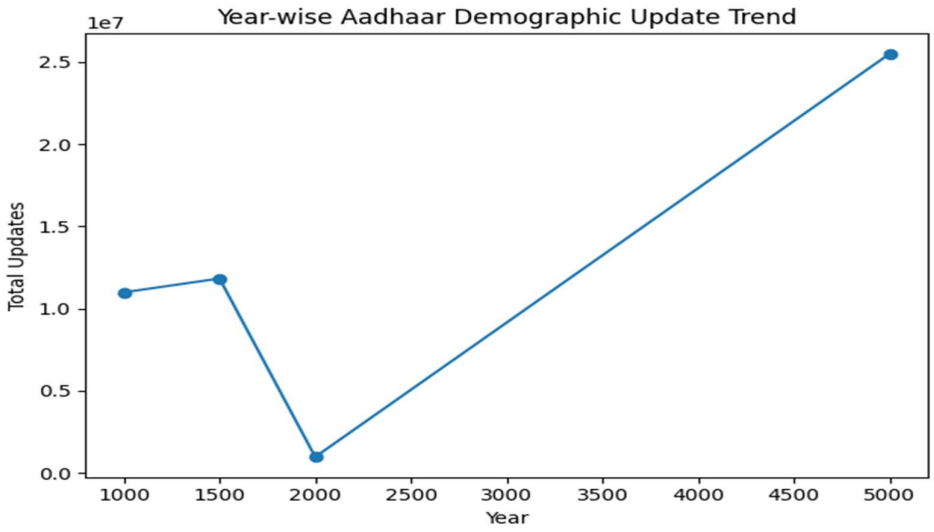


**FIG 6-Year-wise update trends**

## Insight:

- Demographic updates remain consistently high
- Spikes correspond to KYC and policy drives

## Code Snippet :

```
demo_df.groupby("year")["total_updates"].sum()
```

## 4.3 Biometric Update Analysis

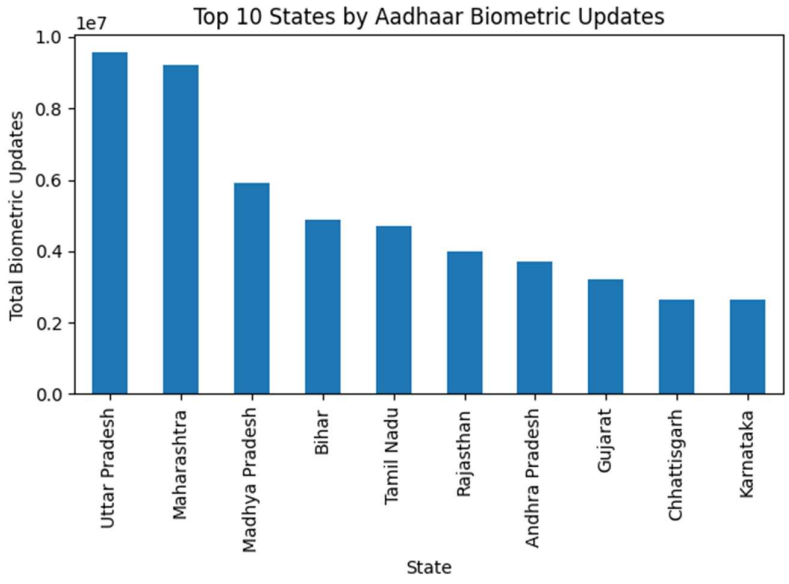


**FIG 7- State-wise biometric updates**

### Insight:

- Certain states show high biometric maintenance demand
- Reflects population density and compliance drives

### Code Snippet :

```
bio_df.groupby("state")["total_biometric_updates"].sum()
```

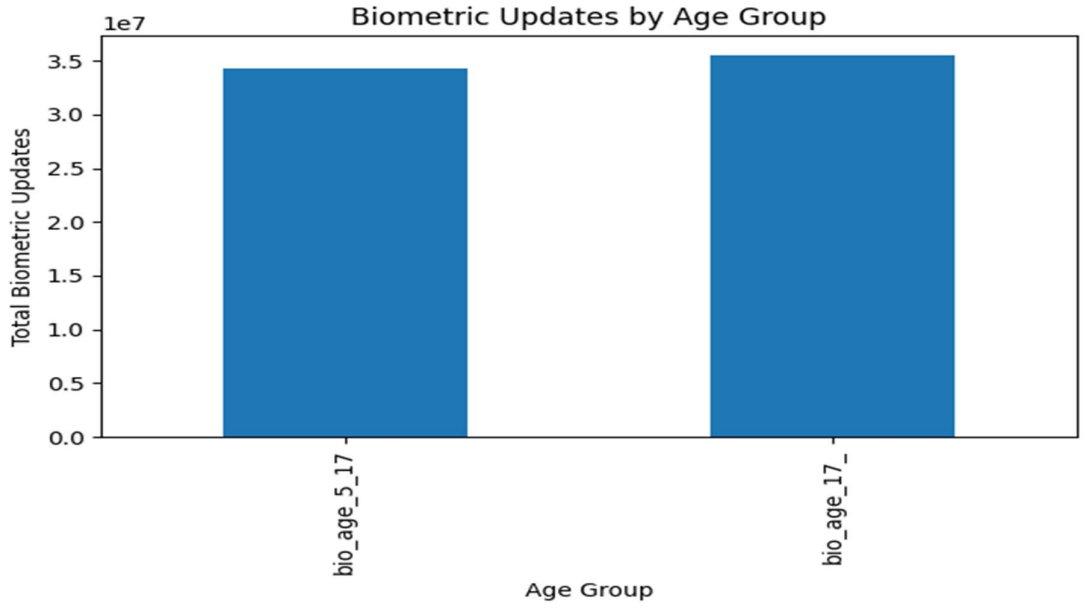


**FIG 8- Age-group biometric update comparison**

## Insight:

- Adult biometric updates dominate
- Supports mandatory biometric refresh during adulthood

## Code Snippet :

```
bio_df[["bio_age_5_17","bio_age_17_"]].sum()
```

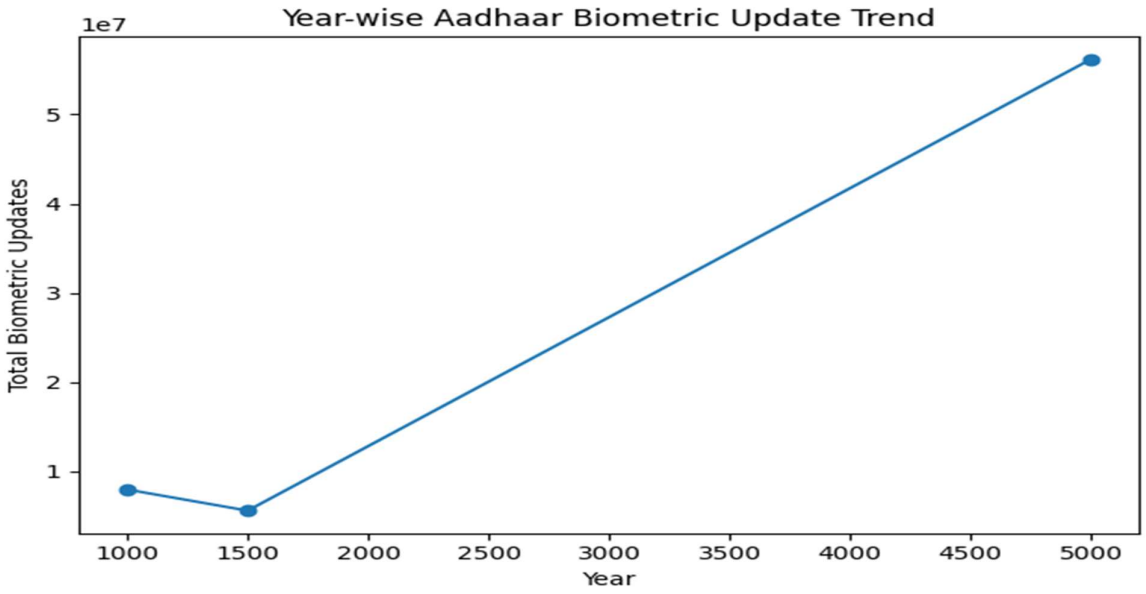


**FIG 9-Year-wise biometric update trends**

## Insight:

- Periodic spikes indicate large-scale biometric update campaigns
- Confirms lifecycle-driven biometric maintenance

## Code Snippet :

```
bio_df.groupby("year")["total_biometric_updates"].sum()
```

### 4.4 Code and Notebooks Used

The analysis was implemented using the following notebooks:

- 01_data_overview_and_loading.ipynb
- 02_data_cleaning_and_preprocessing.ipynb
- 03_enrolment_analysis.ipynb
- 04_demographic_update_analysis.ipynb
- 05_biometric_update_analysis.ipynb
- 06_comparative_and_trend_analysis.ipynb
- 07_anomaly_detection_and_insights.ipynb

Code snippets and outputs from these notebooks are embedded throughout the report.

# CHAPTER-5 CONCLUSION

This project demonstrates how Aadhaar enrolment and update data can be transformed into meaningful societal and administrative insights. By combining age-wise, geographic, and temporal analysis, the study highlights system maturity, lifecycle patterns, and operational signals that can support data-driven governance.

The methodology and findings showcase the potential of public administrative data to inform policy planning, service optimisation, and digital identity management.

# CHAPTER 6-FUTURE SCOPE

- Predictive modelling for update demand forecasting
- District-level hotspot analysis
- Integration with census or migration datasets
- Dashboard-based real-time monitoring