Exploratory Analysis and Visualisation of UIDAI Aadhaar Enrolment and Update Data

UIDAI Data Hackathon 2026

Participant Names: Tarun Doddi(Team Leader), Ajay Unukuru

Tools Used: Python (Pandas), Power BI

1. Problem Statement and Approach

Problem Statement

Aadhaar has achieved wide coverage across India; however, UIDAI continues to observe **uneven enrolment growth** and **substantial volumes of demographic and biometric updates** across regions and age groups. These variations reflect the **lifecycle nature of Aadhaar services**, where enrolment is followed by periodic updates due to changes in personal, demographic, or biometric attributes.

Such patterns place varying levels of demand on enrolment and update infrastructure and may also indicate differences in regional accessibility, awareness, and population mobility. Without a comprehensive analytical view, it becomes challenging to anticipate service demand, identify high-pressure regions, and plan resources efficiently.

The problem addressed in this study is to systematically analyze **temporal, geographic, and age-wise patterns** in Aadhaar enrolments, demographic updates, and biometric updates, and to understand how different age cohorts contribute to Aadhaar lifecycle events.

Approach

To address this problem, a **data-driven analytical approach** was adopted using UIDAI-provided datasets. Aadhaar enrolment, demographic update, and biometric update data—supplied in multiple CSV sub-parts—were consolidated, cleaned, and standardized using **Python (Pandas)**. Data quality issues such as inconsistent state and district names, incorrect entries, and formatting inconsistencies were identified and rectified to ensure analytical reliability.

The cleaned datasets were then modeled and analyzed in **Power BI Desktop**, where interactive dashboards were developed to visualize key metrics, time-series trends, age-group distributions, and geographic patterns. This approach enables the identification of

high-activity regions, age-specific service demand, and temporal trends, helping derive **actionable insights to support resource allocation, service planning, and proactive management of Aadhaar update demand**

2. Datasets Used

This study utilizes datasets provided by the **Unique Identification Authority of India (UIDAI)** as part of the UIDAI Data Hackathon. The data was supplied in multiple CSV sub-parts for each dataset, which were consolidated prior to analysis. All datasets are **aggregated in nature** and contain geographic, temporal, and age-wise information related to Aadhaar enrolments and updates.

2.1 Aadhaar Enrolment Dataset

The Aadhaar Enrolment dataset contains aggregated records of Aadhaar enrolments across different regions and age groups.

Key Columns

- **date** – Date of Aadhaar enrolment activity
- **state** – State or Union Territory name
- **district** – District name
- **age_0_5** – Number of enrolments for children aged 0–5 years
- **age_5_17** – Number of enrolments for residents aged 5–17 years
- **age_18_greater** – Number of enrolments for residents aged 18 years and above

Purpose in Analysis

- To analyze **temporal trends** in Aadhaar enrolments
- To study **age-wise enrolment distribution** across regions
- To identify **state and district level enrolment patterns**

2.2 Aadhaar Demographic Update Dataset

The Aadhaar Demographic Update dataset captures aggregated information related to updates made to residents' demographic details after enrolment.

Key Columns

- **date** – Date of demographic update activity
- **state** – State or Union Territory name
- **district** – District name
- **demo_age_5_17** – Demographic updates for residents aged 5–17 years
- **demo_age_17_** – Demographic updates for residents aged 17 years and above

Purpose in Analysis

- To examine **patterns and frequency** of demographic updates over time
- To analyze **age-group-wise update behavior**
- To identify regions with **high demographic update demand**

2.3 Aadhaar Biometric Update Dataset

The Aadhaar Biometric Update dataset contains aggregated information related to biometric updates, such as fingerprint and iris updates.

Key Columns

- **date** – Date of biometric update activity
- **state** – State or Union Territory name
- **district** – District name
- **bio_age_5_17** – Biometric updates for residents aged 5–17 years
- **bio_age_17_** – Biometric updates for residents aged 17 years and above

Purpose in Analysis

- To understand **biometric update trends** across age groups
- To identify **geographic concentration** of biometric update activities
- To analyze lifecycle-driven biometric update demand

3. Methodology

This study follows a structured data preparation and analysis workflow to ensure accuracy, consistency, and reliability of insights derived from UIDAI Aadhaar datasets. The

methodology includes data consolidation, cleaning, validation, transformation, and visualization.

3.1 Data Consolidation

*Each UIDAI dataset (Aadhaar Enrolment, Demographic Update, and Biometric Update) was provided in **multiple CSV sub-parts**. To create a unified analytical view:*

- *Python was used with the **Pandas** library to load individual CSV files.*
- *Sub-parts belonging to the same dataset were combined using row-wise concatenation.*
- *Column structure consistency was verified before merging.*

*This process resulted in **three consolidated master datasets**, one for each Aadhaar data domain.*

3.2 Data Cleaning and Standardisation

During exploratory analysis, significant **data entry and formatting inconsistencies** were identified across the enrolment, demographic update, and biometric update datasets. These issues were addressed using **Python (Pandas)** to ensure data consistency and analytical reliability.

3.2.1 Aadhaar Enrolment Dataset Cleaning

The enrolment dataset contained multiple data quality issues, particularly in geographic attributes.

Invalid and Incorrect State Values

- A numeric value **1000000** was found in the state column, indicating an invalid data entry.
- Multiple spelling and formatting variations were observed for the same state, for example:
    - *WEST BENGAL*, *West Bangal*, *Westbengal*, *West Bengal*
    - *Orissa* instead of *Odisha*
    - *Pondicherry* instead of *Puducherry*

- Variations in Union Territory names such as *Daman & Diu* and *Dadra & Nagar Haveli*

Actions Taken:

- Invalid numeric state values were identified and removed.
- A mapping-based standardization approach was applied using Pandas replace() to convert all variations into **official, standardized state and Union Territory names**.
- Duplicate and repeated mappings were resolved to ensure uniform naming across records.

District Name Formatting Issues

- Some District names were found in **all capital letters**, leading to inconsistent text formatting.

Actions Taken:

- District names were normalized by converting text to a consistent title-case format.

3.2.2 Aadhaar Demographic Update Dataset Cleaning

The demographic update dataset exhibited **similar inconsistencies** as the enrolment dataset, along with additional structural issues.

State Column Errors

- Multiple spelling and casing variations of state names were present.
- In some records, **district names were incorrectly populated in the state column**, leading to geographic misclassification.

Actions Taken:

- State name variations were standardized using the same mapping logic applied to the enrolment dataset.
- Records with district names in the state column were identified and corrected by realigning values to the appropriate geographic fields.
- Consistency between state and district values was validated against the cleaned enrolment dataset.

3.2.3 Aadhaar Biometric Update Dataset Cleaning

The biometric update dataset contained issues similar to the enrolment dataset, with one key difference.

Geographic Data Issues

- Multiple spelling and casing inconsistencies in the state column.
- District names stored in **uppercase format**.
- Unlike the enrolment dataset, **no numeric placeholder values** (such as 1000000) were present in the state column.

Actions Taken:

- State names were standardized using the same canonical naming approach.
- District name casing was normalized.
- Geographic consistency was ensured across all three datasets.

3.3 Data Validation and Consistency Checks

After cleaning:

- State and district names were cross-validated across all datasets.
- Duplicate and invalid records were checked post-transformation.
- Numeric fields were verified to ensure valid values.
- Cleaned datasets were aligned to support comparative and integrated analysis.

3.4 Data Transformation and Preparation

Following cleaning and validation:

- Date fields were standardized and used for time-based analysis
- Age-group columns were used to derive aggregated metrics
- Cleaned datasets were exported and imported into **Power BI Desktop**

Within Power BI:

- Additional transformations were performed using **Power Query**
- Analytical measures and KPIs were created using **DAX**

3.5 Tools and Technologies Used

- **Python (Pandas)** – Data consolidation, cleaning, and preprocessing
- **Jupyter Notebook** – Exploratory data analysis and validation
- **Visual Studio Code** – Script-based data preparation
- **Power BI Desktop** – Data modeling, analysis, and visualization
- **Power Query & DAX** – Transformations and measure creation

4. Data Analysis and Visualisation

This section presents insights derived from interactive dashboards developed using **Power BI Desktop**, based on Aadhaar enrolment, demographic update, and biometric update datasets. The analysis focuses on **temporal trends**, **geographic distribution**, and **age-wise patterns** to understand Aadhaar lifecycle behaviour across India.

4.1 Aadhaar Enrolment Analysis

Overview

The enrolment dashboard summarizes Aadhaar enrolment activity across time, geography, and age groups. The total enrolments captured in the dataset are approximately **5 million**.

Key Insights

Temporal Trend
Enrolments show a gradual increase from March, followed by a **sharp peak in September**. After September, enrolments decline in October, recover moderately in November, and decrease again in December. This pattern suggests **seasonal or campaign-driven enrolment activity**.

Age-wise Distribution
Age-group analysis indicates that **new Aadhaar enrolments are dominated by younger populations**. The **0–5 age group contributes the highest number of enrolments**, followed by the **5–17 age group**, while enrolments in the **18+ age group are comparatively low**. This reflects widespread adult Aadhaar coverage and emphasizes the focus on **early-life and school-age enrolments**.

Geographic Distribution
State-wise mapping highlights higher enrolment concentrations in **Uttar Pradesh, Maharashtra, West Bengal, Bihar, and Madhya Pradesh**, largely corresponding to population density and administrative scale.

District-level Insights
Districts such as **Thane, Sitamarhi, Bahraich, Murshidabad, and South 24 Parganas** emerge as leading contributors, indicating localized enrolment hubs that may require sustained operational support.

4.2 Aadhaar Demographic Update Analysis

Overview

The demographic update dashboard captures post-enrolment updates to resident information. The total number of demographic updates recorded is approximately **10 million**.

Key Insights

Temporal Trend
Demographic updates show an initially high value, followed by a sharp decline during early months. From July onward, updates increase significantly, with **notable peaks in September, November, and December**, suggesting **periodic update drives or lifecycle-triggered updates**.

Age-wise Distribution
The **17+ age group dominates demographic updates**, while updates among the **5–17 age group** remain relatively low. This pattern reflects higher mobility and personal information changes among adults.

Geographic Distribution
Higher demographic update activity is observed in **Uttar Pradesh, Maharashtra, Karnataka, Telangana, and West Bengal**, indicating greater post-enrolment maintenance demand in these regions.

District-level Insights
Districts such as **North West Delhi, Bengaluru, Hyderabad, Thane, and Surat** show high demographic update volumes, highlighting urban and semi-urban update intensity.

4.3 Aadhaar Biometric Update Analysis

Overview

Biometric updates represent the largest Aadhaar lifecycle activity, with approximately **70 million biometric updates** recorded.

Key Insights

Temporal Trend
Biometric updates exhibit strong fluctuations, peaking around **July–September**, followed by a sharp decline in October and a steady recovery toward December. This suggests **age-related biometric changes and structured update campaigns**.

Age-wise Distribution
Both the **5–17** and **17+ age groups** contribute significantly to biometric updates, with the **adult population contributing slightly more**. This aligns with biometric changes due to growth, ageing, and re-capture requirements.

Geographic Distribution
High biometric update concentrations are observed in **Maharashtra, Uttar Pradesh, Karnataka, Andhra Pradesh, and Tamil Nadu**.

District-level Insights
Districts such as **Pune, Nashik, Thane, Jalgaon, and Aurangabad** dominate biometric updates, indicating sustained operational demand in these areas.

4.4 Cross-Dataset Comparative Insights

- **Aadhaar enrolments are predominantly driven by children and adolescents**, with the **0–5 and 5–17 age groups contributing the majority of new enrolments**.
- In contrast, **demographic and biometric updates are dominated by the 17+ age group**, reflecting increased update needs as individuals age.
- This contrast clearly illustrates the **Aadhaar lifecycle pattern**:
    - **Early stage:** High enrolment activity among children
    - **Later stages:** Increasing demographic and biometric updates among adults
- Across all datasets, **update volumes significantly exceed new enrolments**, indicating a **mature Aadhaar ecosystem** where data maintenance forms the primary service demand.
- Urban and high-population regions consistently show higher update activity, while enrolments are more broadly distributed.

Analytical Conclusion

The combined visual analysis of enrolment, demographic update, and biometric update datasets provides a comprehensive view of Aadhaar lifecycle behaviour across India. These insights can support **data-driven decision-making** for infrastructure planning, targeted outreach, and proactive management of Aadhaar update demand by UIDAI.

5. Code and Implementation

```
[1]
import pandas as pd
Python

[2]
dfi=pd.read_csv('c:/Users/Tarun/Desktop/Hackathon Data/api_data_aadhar_enrolment/api_data_aadhar_enrolment_0_500000.csv')
dfii=pd.read_csv('c:/Users/Tarun/Desktop/Hackathon Data/api_data_aadhar_enrolment/api_data_aadhar_enrolment_500000_1000000.csv')
dfiii=pd.read_csv('c:/Users/Tarun/Desktop/Hackathon Data/api_data_aadhar_enrolment/api_data_aadhar_enrolment_1000000_1006029.csv')
Python

[3]
df1=pd.concat([dfi,dfii,dfiii])
Python

[ ]
df1.head()
Python

[6]
df1.shape
Python

(1006029, 7)
```

Figure 5.1: Loading multiple UIDAI Aadhaar enrolment CSV sub-parts and consolidating them into a single dataset using Python (Pandas).

```
[7]
df1['state'].unique()
Python

array(['Meghalaya', 'Karnataka', 'Uttar Pradesh', 'Bihar', 'Maharashtra',
       'Haryana', 'Rajasthan', 'Punjab', 'Delhi', 'Madhya Pradesh',
       'West Bengal', 'Assam', 'Uttarakhand', 'Gujarat', 'Andhra Pradesh',
       'Tamil Nadu', 'Chhattisgarh', 'Jharkhand', 'Nagaland', 'Manipur',
       'Telangana', 'Tripura', 'Mizoram', 'Jammu and Kashmir',
       'Chandigarh', 'Sikkim', 'Odisha', 'Kerala',
       'The Dadra And Nagar Haveli And Daman And Diu',
       'Arunachal Pradesh', 'Himachal Pradesh', 'Goa',
       'Jammu And Kashmir', 'Dadra and Nagar Haveli and Daman and Diu',
       'Ladakh', 'Andaman and Nicobar Islands', 'Orissa', 'Pondicherry',
       'Puducherry', 'Lakshadweep', 'Andaman & Nicobar Islands',
       'Dadra & Nagar Haveli', 'Dadra and Nagar Haveli', 'Daman and Diu',
       'WEST BENGAL', 'Jammu & Kashmir', 'West  Bengal', '100000',
       'Daman & Diu', 'West Bangal', 'Westbengal', 'West bengal',
       'andhra pradesh', 'ODISHA', 'WESTBENGAL'], dtype=object)

[93]
df1=df1.drop(df1[df1["state"]=="100000"].index)
Python
```

Figure 5.2: Removal of invalid state entries and standardisation of state and Union Territory names using mapping-based replacement in Pandas.

```
j=df2['state'].unique()
for i in j:
    if i.lower().startswith(''):
        print(i)
```

Python

```
df2['state']=df2['state'].replace({'west Bengal':'West Bengal',
                                   'West  Bengal':'West Bengal',
                                   'Westbengal':'West Bengal',
                                   'WEST BENGAL':'West Bengal',
                                   'West Bangal':'West Bengal',
                                   'West bengal':'West Bengal',
                                   'WESTBENGAL':'West Bengal',
                                   'West Bengli':'West Bengal',
                                   'andhra pradesh':'Andhra Pradesh',
                                   'Andaman & Nicobar Islands':'Andaman and Nicobar Islands',
                                   'Chhatisgarh':'Chhattisgarh',
                                   'Daman and Diu':'Dadra and Nagar Haveli and Daman and Diu',
                                   'Dadra and Nagar Haveli':'Dadra and Nagar Haveli and Daman and Diu',
                                   'Daman & Diu':'Dadra and Nagar Haveli and Daman and Diu',
                                   'Dadra & Nagar Haveli':'Dadra and Nagar Haveli and Daman and Diu',
                                   'Jammu & Kashmir':'Jammu and Kashmir',
                                   'Orissa':'Odisha',
                                   'odisha':'Odisha',
                                   'ODISHA':'Odisha',
                                   'Pondicherry':'Puducherry',
                                   'Uttaranchal':'Uttarakhand'
```

Python

Figure 5.3: Identification of inconsistent and invalid state values in the enrolment dataset prior to data cleaning.

5.2 Power BI – Data Modeling and Visualisation

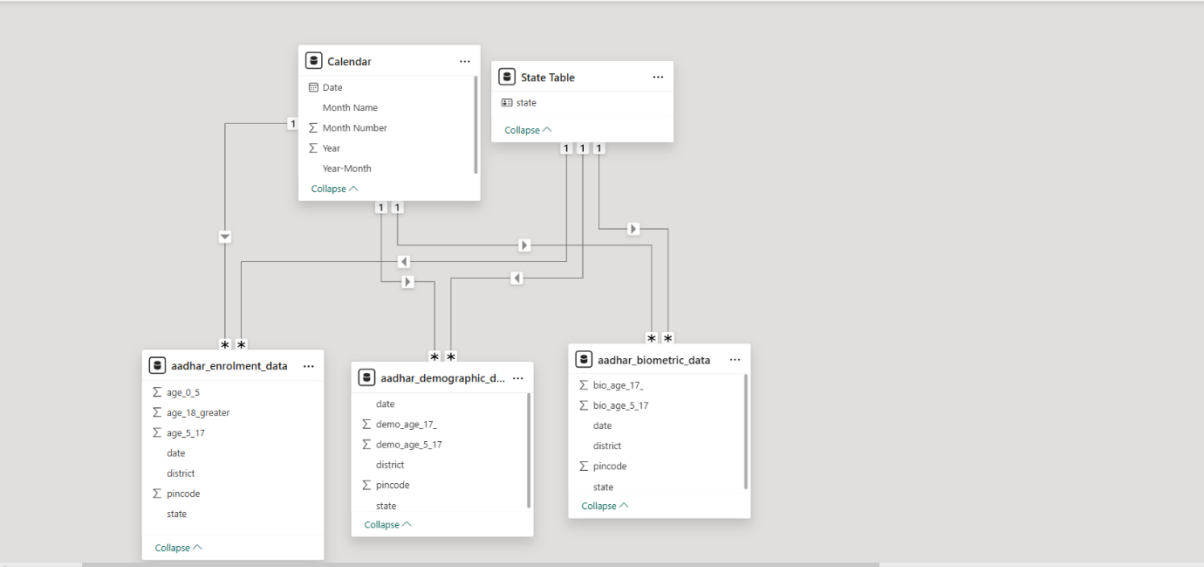

Figure 5.4: Power BI data model illustrating relationships between enrolment, demographic update, biometric update, calendar, and state tables.

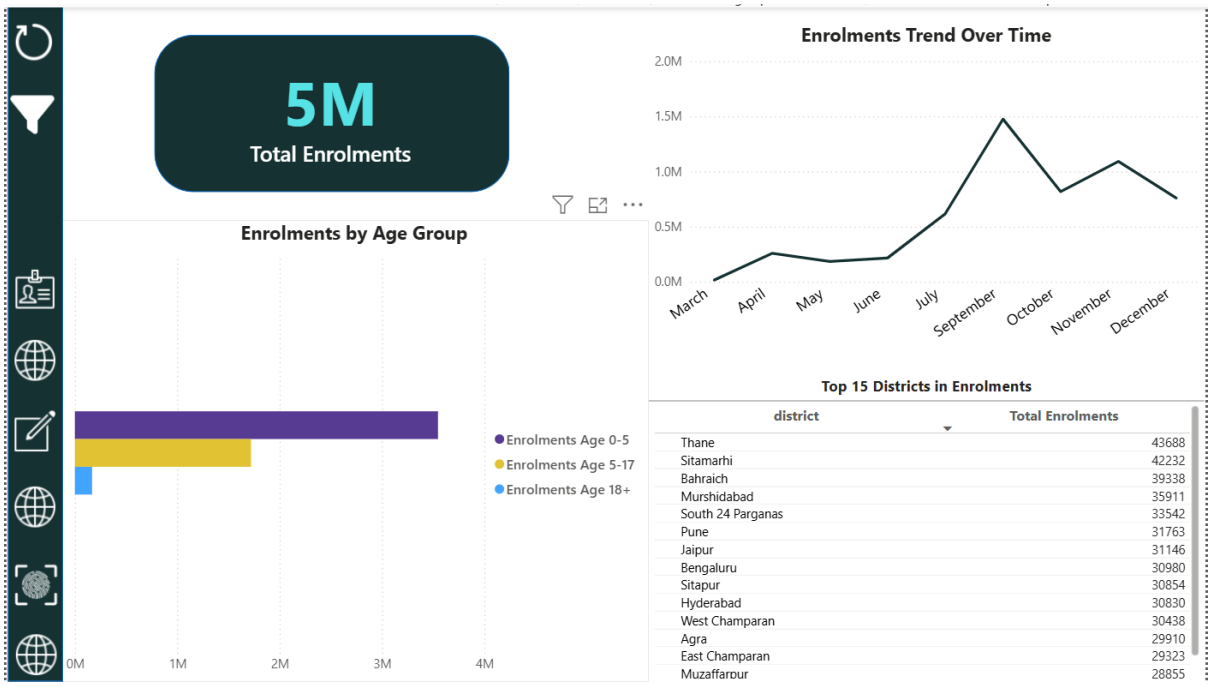

Figure 5.5: Final Power BI dashboard visualising Aadhaar enrolment trends, age-wise distribution, and district-level insights