UIDAI DATA HACKATHON 2026

Report

INDEX:

1. Problem Statement

Aadhaar transaction data reveals a system that is predominantly maintenance-driven rather than enrollment-driven. Nationally, biometric updates account for over 70% of total Aadhaar activity, while new enrollments contribute less than 3%, indicating that the Aadhaar ecosystem has largely matured in terms of coverage but requires continuous upkeep.

However, deeper analysis uncovers regional and age-group-level anomalies. Certain states exhibit disproportionately high enrollment activity among older age groups (0–5, 5–17, and 18+), rather than early childhood alone, suggesting delayed enrollment or regional accessibility gaps. At the same time, biometric and demographic updates are heavily concentrated in high-population states, placing operational pressure on existing infrastructure.

Additionally, monthly trends show system-wide dips in activity, notably during October, pointing toward external influences such as festive periods that affect service utilization.

This study aims to:

- Identify operational patterns in enrollments and updates,
- Highlight regional deviations from national trends,
- Assess whether Aadhaar infrastructure and policies are aligned with actual usage patterns,
- And provide data-backed recommendations to improve efficiency, accessibility, and system resilience.

2. Dataset Used

Three Aadhaar-related datasets were analyzed, each compiled from multiple CSV files:

Dataset	**Description**
Enrollment Dataset	Aadhaar enrollment counts by age group (0–5, 5–17, 18+)
Biometric Updates Dataset	Biometric updates Counts across age groups (5-17 and 18+)
Demographic Updates Dataset	Demographics update counts across age groups(5-17 and 18+)

Common columns Used

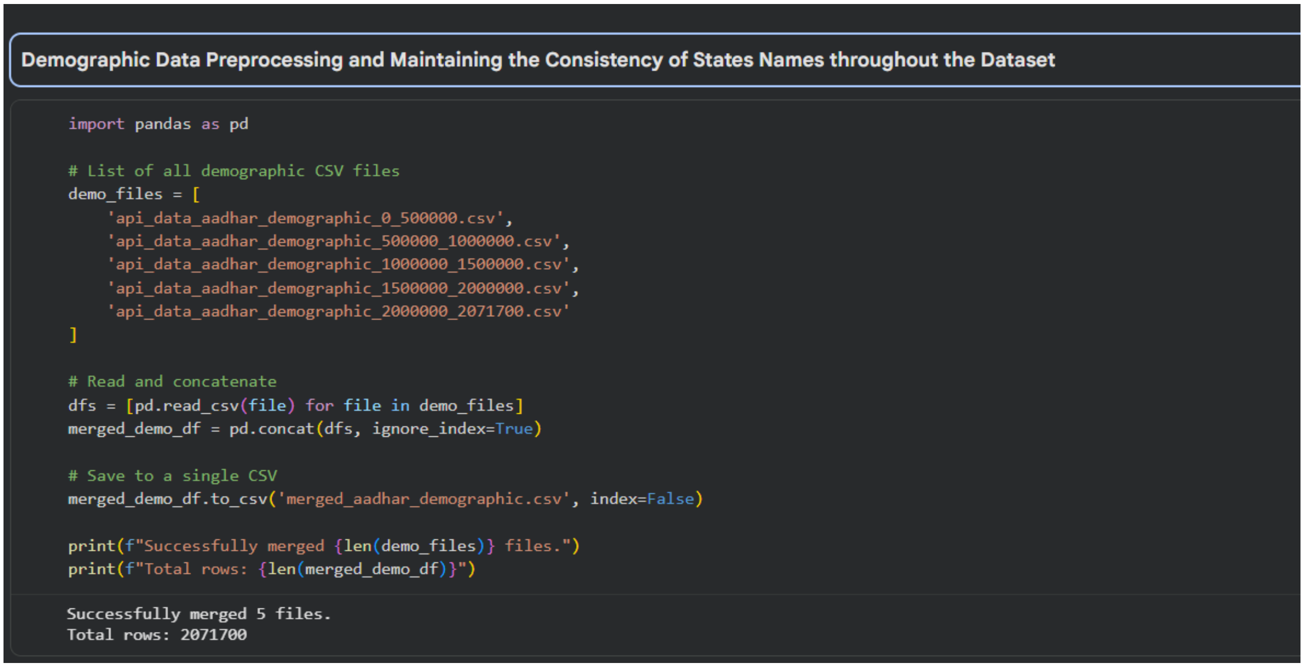
- State
- District
- Date(Month)
- Pincode

These fields enabled temporal, geographic, and demographic analysis at both macro and micro levels.

3. Methodology

Data Preparation and Cleaning in Jupyter Notebook / Google Colab

- Initially, multiple CSV files from different sources were imported into Jupyter Notebook/Google Colab for preprocessing.
- State-level cleaning was performed here, which included:
  - Merging multiple CSV files into unified datasets for enrollment, biometric updates, and demographic updates.
  - Standardizing state names by correcting spelling inconsistencies and removing duplicates.
  - Handling missing values and ensuring consistent data types.
  - Preliminary aggregations and sanity checks to validate data quality before importing into Power BI.
- These preprocessing steps ensured the base data was clean and structured for further refinement.



Demographic Data Preprocessing and Maintaining the Consistency of States Names throughout the Dataset

```
import pandas as pd

# List of all demographic CSV files
demo_files = [
    'api_data_aadhar_demographic_0_500000.csv',
    'api_data_aadhar_demographic_500000_1000000.csv',
    'api_data_aadhar_demographic_1000000_1500000.csv',
    'api_data_aadhar_demographic_1500000_2000000.csv',
    'api_data_aadhar_demographic_2000000_2071700.csv'
]

# Read and concatenate
dfs = [pd.read_csv(file) for file in demo_files]
merged_demo_df = pd.concat(dfs, ignore_index=True)

# Save to a single CSV
merged_demo_df.to_csv('merged_aadhar_demographic.csv', index=False)

print(f"Successfully merged {len(demo_files)} files.")
print(f"Total rows: {len(merged_demo_df)}")
```

```
Successfully merged 5 files.
Total rows: 2071700
```

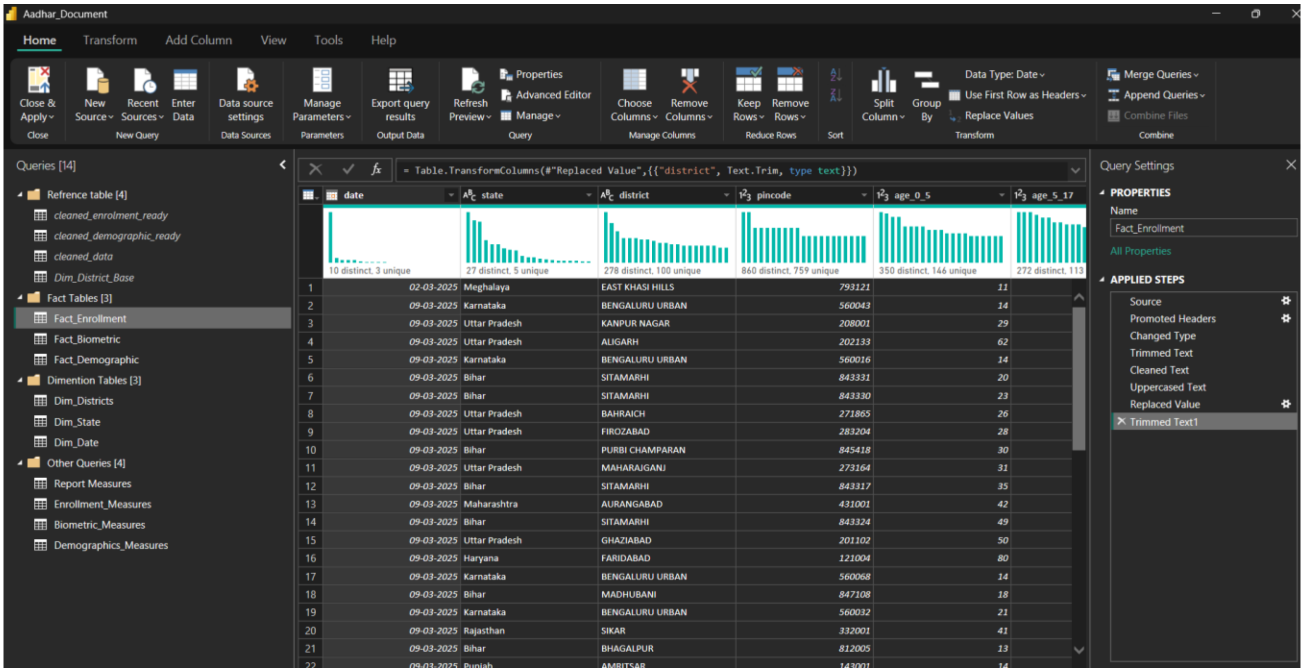
Power Query Transformations

Data cleaning and transformation were performed extensively in Power BI's Power Query editor to ensure consistency and accuracy in the district-level data. Key transformation steps included:

- **Trimming Text:** Removed leading and trailing whitespace from district names to avoid mismatches during analysis.
- **Text Cleaning:** Corrected spelling inconsistencies and unified text formats.
- **Uppercasing:** Converted district names to uppercase to standardize naming conventions and facilitate reliable joins and filtering.
- **Value Replacement:** Fixed inconsistent district names by replacing variants with a single standardized form.

- **Type Conversion:** Ensured all columns had appropriate data types for accurate processing (e.g., date, text, numeric).

Approximately 50% of district records required these cleaning steps to resolve discrepancies across datasets, thereby enabling robust aggregation and analysis.



Queries [14]

- Refrence table [4]
  - cleaned_enrolment_ready
  - cleaned_demographic_ready
  - cleaned_data
  - Dim_District_Base
- Fact Tables [3]
  - Fact_Enrollment
  - Fact_Biometric
  - Fact_Demographic
- Dimention Tables [3]
  - Dim_Districts
  - Dim_State
  - Dim_Date
- Other Queries [4]
  - Report Measures
  - Enrollment_Measures
  - Biometric_Measures
  - Demographics_Measures

```
= Table.TransformColumns(#"Replaced Value",{{"district", Text.Trim, type text}})
```

	date	state	district	pincode	age_0_5	age_5_17
	10 distinct, 3 unique	27 distinct, 5 unique	278 distinct, 100 unique	860 distinct, 759 unique	350 distinct, 146 unique	272 distinct, 113
1	02-03-2025	Meghalaya	EAST KHASI HILLS	793121	11	
2	09-03-2025	Karnataka	BENGALURU URBAN	560043	14	
3	09-03-2025	Uttar Pradesh	KANPUR NAGAR	208001	29	
4	09-03-2025	Uttar Pradesh	ALIGARH	202133	62	
5	09-03-2025	Karnataka	BENGALURU URBAN	560016	14	
6	09-03-2025	Bihar	SITAMARHI	843331	20	
7	09-03-2025	Bihar	SITAMARHI	843330	23	
8	09-03-2025	Uttar Pradesh	BAHRAICH	271865	26	
9	09-03-2025	Uttar Pradesh	FIROZABAD	283204	28	
10	09-03-2025	Bihar	PURBI CHAMPARAN	845418	30	
11	09-03-2025	Uttar Pradesh	MAHARAJGANJ	273164	31	
12	09-03-2025	Bihar	SITAMARHI	843317	35	
13	09-03-2025	Maharashtra	AURANGABAD	431001	42	
14	09-03-2025	Bihar	SITAMARHI	843324	49	
15	09-03-2025	Uttar Pradesh	GHAZIABAD	201102	50	
16	09-03-2025	Haryana	FARIDABAD	121004	80	
17	09-03-2025	Karnataka	BENGALURU URBAN	560068	14	
18	09-03-2025	Bihar	MADHUBANI	847108	18	
19	09-03-2025	Karnataka	BENGALURU URBAN	560032	21	
20	09-03-2025	Rajasthan	SIKAR	332001	41	
21	09-03-2025	Bihar	BHAGALPUR	812005	13	
22	09-03-2025	Punjab	AMRITSAR	143001	14	

Query Settings

PROPERTIES

Name: Fact_Enrollment

All Properties

APPLIED STEPS

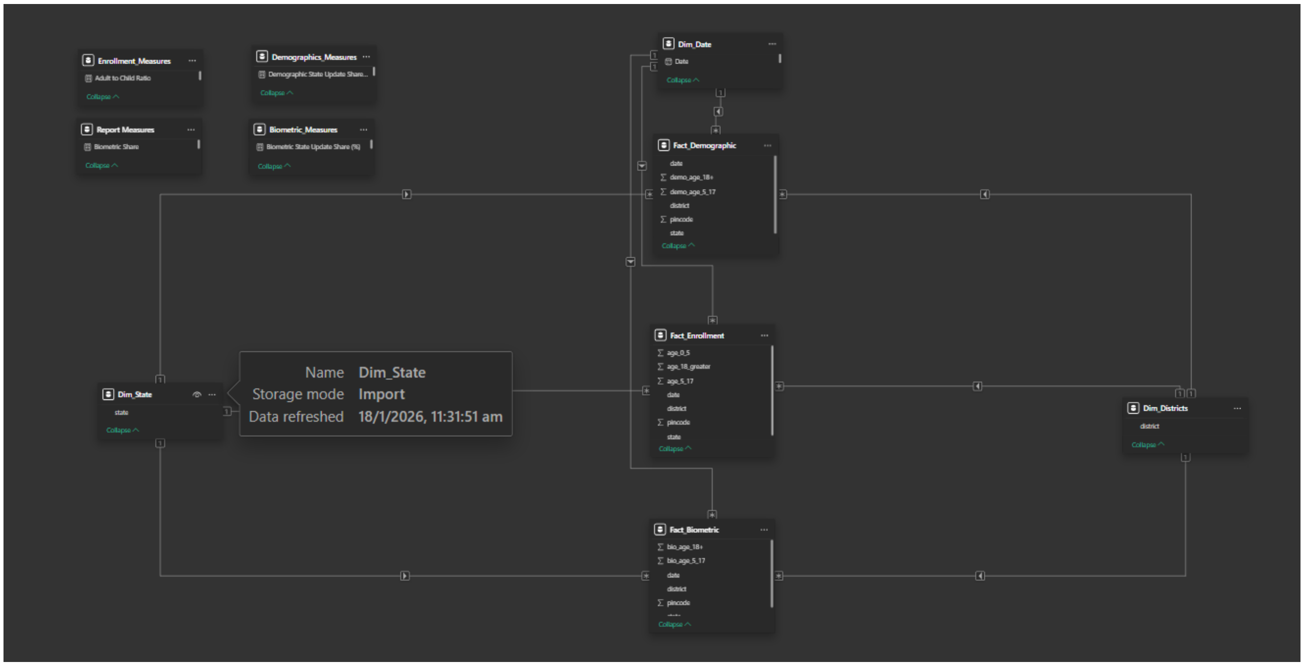
- Source
- Promoted Headers
- Changed Type
- Trimmed Text
- Cleaned Text
- Uppercased Text
- Replaced Value
- Trimmed Text1

Data Modeling

In Power BI's Model View, a star schema was created to organize and relate the data efficiently:

- **Fact Tables:**
  - Fact_Enrollment — containing enrollment counts segmented by age groups and geography.
  - Fact_Biometric — containing biometric update records by age and region.
  - Fact_Demographic — containing demographic update records by district and state.
- **Dimension Tables:**
  - Dim_Date — a comprehensive date table to enable time-based filtering and calculations.
  - Dim_State — holding state-level metadata.
  - Dim_Districts — standardized district names for consistent geographic analysis.

Relationships were established between fact tables and corresponding dimension tables using primary keys like state, district, and date, enabling efficient slicing and dicing across multiple dimensions.



Name Dim_State
Storage mode Import
Data refreshed 18/1/2026, 11:31:51 am

DAX Measures and Visualizations

To extract meaningful insights, various DAX (Data Analysis Expressions) measures were created. Some examples include:

- **Biometric Share:** Calculated the percentage share of biometric updates relative to total Aadhaar transactions, helping identify states with higher maintenance activity.
- **Age Group Ratios:** Measures comparing enrollments or updates across age groups (e.g., adult to child ratio) to detect anomalies or demographic trends.
- **Deviation Metrics:** Calculated deviations from national averages at the state or district level to highlight regions with atypical patterns.

These measures powered interactive visuals on the Power BI dashboard, allowing dynamic filtering by geography, time, and age groups for comprehensive analysis.

4. Data Analysis and Key Insights (Power BI)

Enrollment Insights

- Pan-India enrollments remain low relative to updates, confirming Aadhaar's mature coverage stage.
- Meghalaya shows higher adult (18+) enrollments, but this is not uniform across all districts.
- East Khasi Hills emerges as the dominant contributor, indicating localized drivers rather than a state-wide pattern.
- This suggests that district-targeted interventions will be more effective than blanket state-level policies in addressing enrollment gaps.

Biometric Updates Insights

- Biometric updates dominate Aadhaar transactions and are largely driven by high-population states such as Uttar Pradesh, Maharashtra, and Madhya Pradesh.
- At the national level, updates are balanced between all age groups (0–5, 5–17, and 18+), indicating continuous lifecycle maintenance.
- Deviation analysis highlights certain states updating biometrics more frequently than the national average, pointing to infrastructure load concentration.
- A clear drop in October is observed across most states, plausibly linked to Diwali and festive season effects, when citizens deprioritize administrative activities.

Demographic Updates Insights

- Demographic updates are concentrated in high-population states, particularly Uttar Pradesh and Maharashtra.
- These updates can be performed both at physical centers and via online portals, reflecting digitization efforts in Aadhaar update processes.
- The pattern reflects population scale and volume effects, rather than abnormal behavior.
- Consistent monthly drops mirror biometric trends, reinforcing the presence of system-wide temporal factors.
- Government campaigns, including school and community outreach programs, may influence spikes in enrollments and updates during certain months (e.g., July).

5. Recommendations and Actionable Outcomes

Pan-India Recommendations

- Shift policy focus from enrollment expansion to maintenance optimization, as updates dominate system usage.
- Introduce seasonal capacity planning, especially around festive months like October, to manage predictable drops and post-festival surges.
- Use deviation metrics as an early warning system to identify overburdened regions.

Enrollment-Specific Recommendations

- Implement district-level monitoring instead of state-level assumptions, as seen in Meghalaya (East Khasi Hills dominance).
- Strengthen early-age (0–5) enrollment awareness programs to reduce delayed enrollments at later ages.
- Deploy mobile enrollment units in districts showing adult-heavy enrollment patterns.

Biometric Updates Recommendations

- Expand and upgrade biometric infrastructure in high-population, high-update states to reduce processing delays.
- Introduce update bundling or reminders, allowing citizens to complete multiple updates in a single visit.
- Use age-wise parity insights to design uniform biometric refresh cycles.

Demographic Updates Recommendations

- Prioritize capacity in large states where volume-driven updates are unavoidable.
- Improve data validation at enrollment to reduce future correction load.
- Consider enhancing online update portals and promoting digital channels to reduce physical center load.

6. Expected Impact

By aligning Aadhaar operations with actual usage patterns identified through this dashboard:

- Authorities can optimize resource allocation,
- Reduce operational strain in high-impact states,
- Improve citizen experience through targeted interventions,
- And transition Aadhaar management from reactive maintenance to data-driven governance.