UIDAI DATA HACKATHON 2026

Project Submission

Vaibhav R. Bagde

Team leader

Prerna P. Tekale

Team Meader

Adnath N. Nage

Team Member

Project Dataset

Aadhaar Enrolment Dataset

Problem Statement and Approach

Context

Regarding the Aadhaar programme, India's Unique Identification Authority (UIDAI) has laid the foundations for the country's digital identity system.

The programme, which enables residents to access a multitude of public services, welfare schemes and financial systems, has over the years achieved a fairly extensive coverage in the country.

However, the pattern of Aadhaar enrolment is by no means uniform and is often found to differ across age, geographical and time segments.

These deviations are an effect of societal factors, demographic disparities and even the distribution of services, with early-life inclusion, population distributions, and the efficacy of drive drives all adding to these dynamics.

Problem Statement

Despite the availability of large-scale Aadhaar enrolment data, it is often used primarily for **operational reporting**, with limited analytical exploration of underlying patterns and trends. As a result:

- Temporal fluctuations in enrolment activity are not fully understood
- Regional disparities across states and districts remain under-analysed
- Age-wise enrolment behaviour and late adoption patterns are not systematically quantified
- Administrative insights that could support capacity planning and service optimisation are underutilised

Objectives of the Study

This study aims to address the above gap by pursuing the following objectives:

1. To analyse **age-wise enrolment patterns** and assess the balance between early-life inclusion and adult enrolment
2. To examine **geographic variations** in enrolment across states and districts
3. To identify **temporal trends and seasonality** in enrolment activity
4. To detect **anomalies and volatility** in enrolment behaviour across regions
5. To derive **actionable insights** that can support policy planning and administrative decision-making

Systematic Analytical Approach

Step 1: Data Understanding
Step 2: Data Cleaning and Preprocessing
Step 3: Univariate Analysis
Step 4: Bivariate Analysis
Step 5: Trivariate Analysis
Step 6: Indicator Development
Step 7: Insight Translation

Expected Outcomes

- Reveal how Aadhaar enrolment behaviour differs across age groups and regions
- Highlight seasonal and campaign-driven enrolment dynamics
- Identify regions with unstable or irregular enrolment activity
- Support data-driven improvements in enrolment planning and resource allocation

Datasets Used

Dataset : Aadhaar Enrolment Dataset

Source : Unique Identification Authority of India (UIDAI)

For this project, UIDAI provided multiple Aadhaar-related datasets as part of the challenge requirements. While three datasets were made available, this study **intentionally focuses on a single dataset-the Aadhaar Enrolment Dataset**-to enable a deeper, more structured analysis of enrolment behaviour across India.

The Aadhaar Enrolment Dataset contains aggregated information on Aadhaar registrations across different geographic and demographic dimensions. The data does not include any personally identifiable information and is suitable for large-scale analytical and policy-oriented evaluation.

- **Date of Enrolment:**
  Used to analyse temporal trends, seasonality, and fluctuations in enrolment activity over time.
- **State and District:**
  Enable geographic comparison and help identify regional disparities in enrolment patterns.
- **PIN Code:**
  Provides fine-grained spatial context and supports micro-level analysis of enrolment distribution.
- **Age-wise Enrolments:**
  - 0–5 years
  - 5–17 years
  - 18 years and above

  These categories allow the study of early-life inclusion, school-age enrolment, and adult or late Aadhaar adoption.

Methodology

1. Dataset Selection

- Three Aadhaar-related datasets were provided as part of the challenge.
- For this project, **only the Aadhaar Enrolment Dataset** was selected.
- The dataset was chosen because it provides comprehensive coverage across **time, geography, and age groups**, enabling deeper and more focused analysis.

2. Data Collection and Integration

- The Aadhaar Enrolment data was provided in multiple CSV files.
- All files were merged into a single consolidated dataset to ensure uniform processing and analysis.
- Initial inspection was carried out to understand the structure, variables, and scale of the data.

3. Data Cleaning

- Date fields were converted into a standard datetime format.
- State and district names were cleaned by:
  - Removing extra spaces
  - Correcting spelling variations
  - Standardising text case
- Known inconsistencies in state names were resolved using a predefined mapping.
- Duplicate records were identified and removed to avoid double counting.
- Records with invalid or zero enrolment values across all age groups were excluded.
- PIN codes were treated as categorical values to preserve formatting.

4. Data Preprocessing

- A new variable, **total enrolment**, was created by summing enrolments across all age groups.
- Time-based features such as **year** and **month** were extracted from the enrolment date.
- The dataset was aggregated at different levels, including:
  - State–month level
  - National age-group level

5. Data Transformation

- Aggregation operations were applied to summarise enrolment activity across geographic and temporal dimensions.
- Ratio-based indicators were created, such as:
  - Adult enrolment share as an indicator of late Aadhaar adoption.
- Statistical measures (e.g., standard deviation) were calculated to assess enrolment volatility across states.

6. Exploratory Data Analysis

- **Univariate analysis** was performed to examine individual variables such as age-wise enrolment distribution.
- **Bivariate analysis** was used to study relationships between:
  - Age group and geography
  - Time and enrolment volume
- **Trivariate analysis** combined state, time, and age group to identify seasonal patterns and enrolment surges.

7. Visualisation

- Appropriate charts were created to represent trends and patterns, including:
  - Bar charts
  - Line charts
  - Stacked bar charts
- All visualisations were clearly labelled and supported by written interpretations.

8. Insight Generation

- Analytical results were interpreted to identify:
  - Societal trends such as early-life inclusion
  - Administrative patterns such as seasonal enrolment drives
- Observed patterns were translated into meaningful insights rather than raw statistics.

9. Tools and Reproducibility

- The analysis was carried out using Python and Jupyter Notebook.
- Libraries such as Pandas and Matplotlib were used for data processing and visualisation.
- The entire workflow was designed to be reproducible and transparent.

GitHub Link: https://github.com/Vaibhavbagde551/UIDAI-Data-Hackethon-2026-Project

Data Analysis and Visualisation

This section presents the analytical findings derived from the Aadhaar Enrolment dataset using systematic exploratory data analysis techniques. The analysis is structured into **univariate, bivariate, and trivariate analysis**, supported by visualisations created using Python.

1. Univariate Analysis

Univariate analysis was performed to understand the **overall distribution and magnitude** of Aadhaar enrolments across individual variables.

1.1 Age-wise Enrolment Distribution (National Level)

Objective:

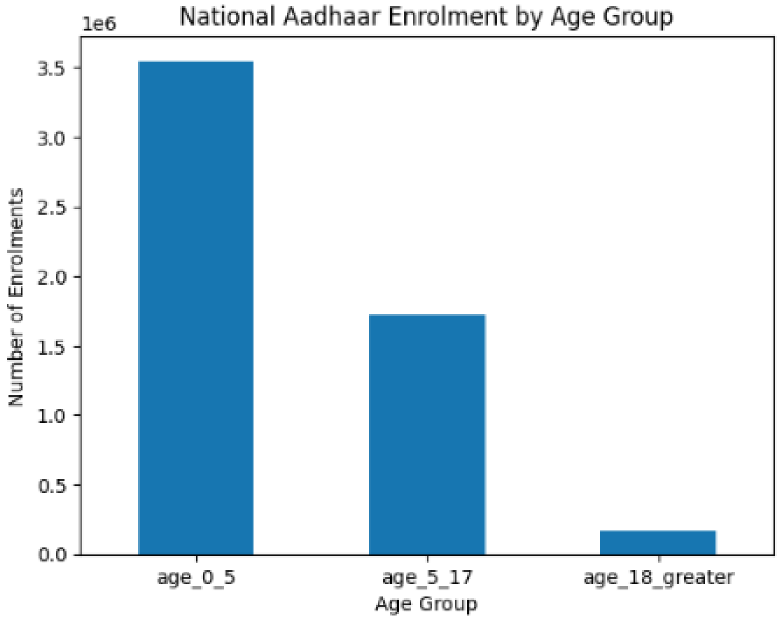
To understand how Aadhaar enrolment is distributed across different age groups at the national level.

Code Used:

```
age_totals = df[['age_0_5', 'age_5_17', 'age_18_greater']].sum()

plt.figure()
age_totals.plot(kind='bar')
plt.title("National Aadhaar Enrolment by Age Group")
plt.xlabel("Age Group")
plt.ylabel("Number of Enrolments")
plt.xticks(rotation=0)
plt.show()
```

Visualisations:



Key Observations:

- Enrolment in the **0–5 years age group** is significantly higher than other age groups.
- Enrolment decreases progressively with increasing age.
- Adult enrolment (18+) forms a very small proportion of total enrolments.

Interpretation:

- Aadhaar enrolment in India is largely driven by **early-life registration**, indicating strong institutional mechanisms such as birth registration and child welfare programmes.
- The low adult enrolment suggests that Aadhaar coverage among adults is largely saturated.

- Adult enrolment (18+) forms a very small proportion of total enrolments.

Interpretation:

- Aadhaar enrolment in India is largely driven by **early-life registration**, indicating strong institutional mechanisms such as birth registration and child welfare programmes.
- The low adult enrolment suggests that Aadhaar coverage among adults is largely saturated.

1.2 Top States by Total Aadhaar Enrolment

Objective:

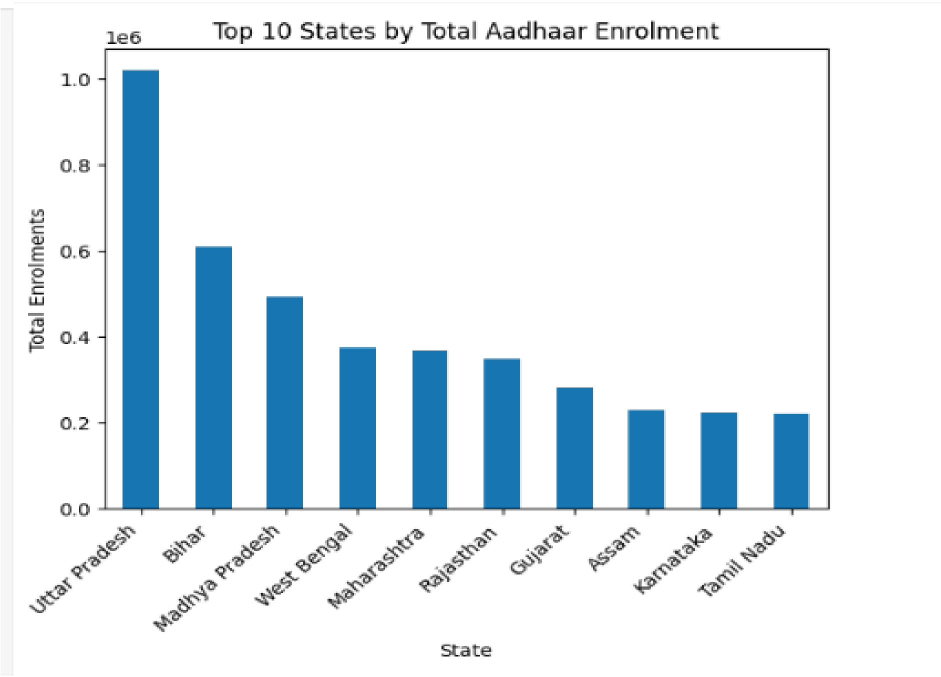
To identify states with the highest overall Aadhaar enrolment volumes.

Code Used:

```
state_totals = (df.groupby('state')['total_enrolment'].sum().sort_values(ascending=False).head(10))

plt.figure()
state_totals.plot(kind='bar')
plt.title("Top 10 States by Total Aadhaar Enrolment")
plt.xlabel("State")
plt.ylabel("Total Enrolments")
plt.xticks(rotation=45, ha='right')
plt.show()
```

Visualisations:



Key Observations:

- High-population states dominate total enrolment counts.
- Smaller states and Union Territories show lower absolute enrolment numbers.

Interpretation:

- Enrolment volume is influenced by population size as well as administrative capacity.
- Absolute counts alone may not reflect enrolment efficiency, highlighting the need for deeper analysis.

1.3 Daily Aadhaar Enrolment Trend

Objective:

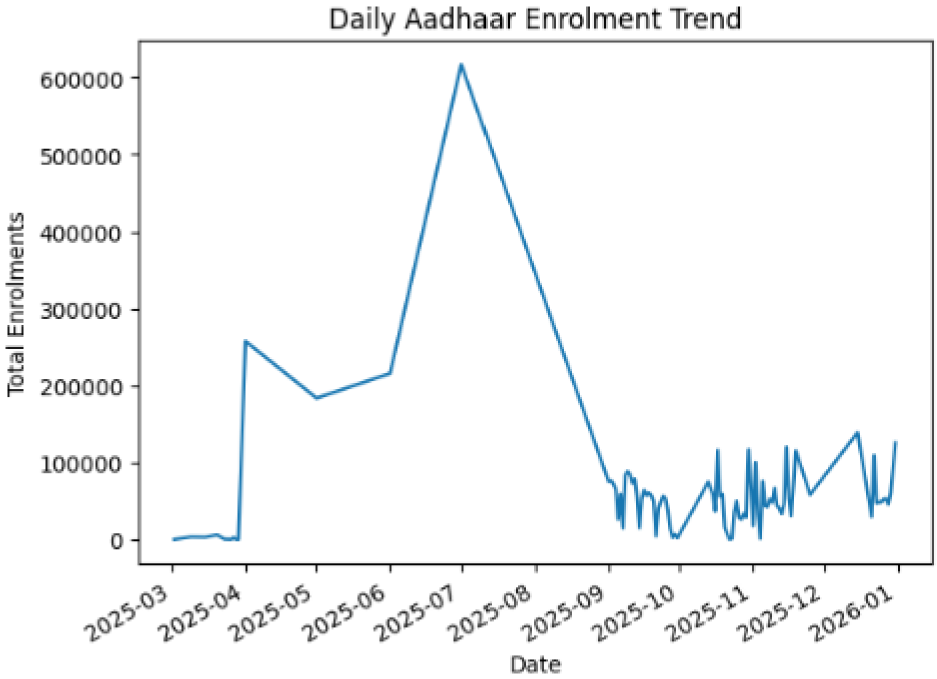
To observe fluctuations in enrolment activity over time.

Code Used:

```
daily_totals = df.groupby('date')['total_enrolment'].sum()

plt.figure()
daily_totals.plot()
plt.title("Daily Aadhaar Enrolment Trend")
plt.xlabel("Date")
plt.ylabel("Total Enrolments")
plt.show()
```

Visualisations:



Key Observations:

- Enrolment activity shows sharp spikes and dips.
- The trend is not uniform across days.

Interpretation:

- These fluctuations indicate **campaign-driven or time-bound enrolment activities**, rather than steady daily enrolment.

2. Bivariate Analysis

Bivariate analysis was conducted to examine **relationships between two variables**, such as age group and geography, or time and enrolment volume.

2.1 Age-wise Enrolment Distribution Across Top States

Objective:

To compare age-wise enrolment patterns across high-enrolment states.

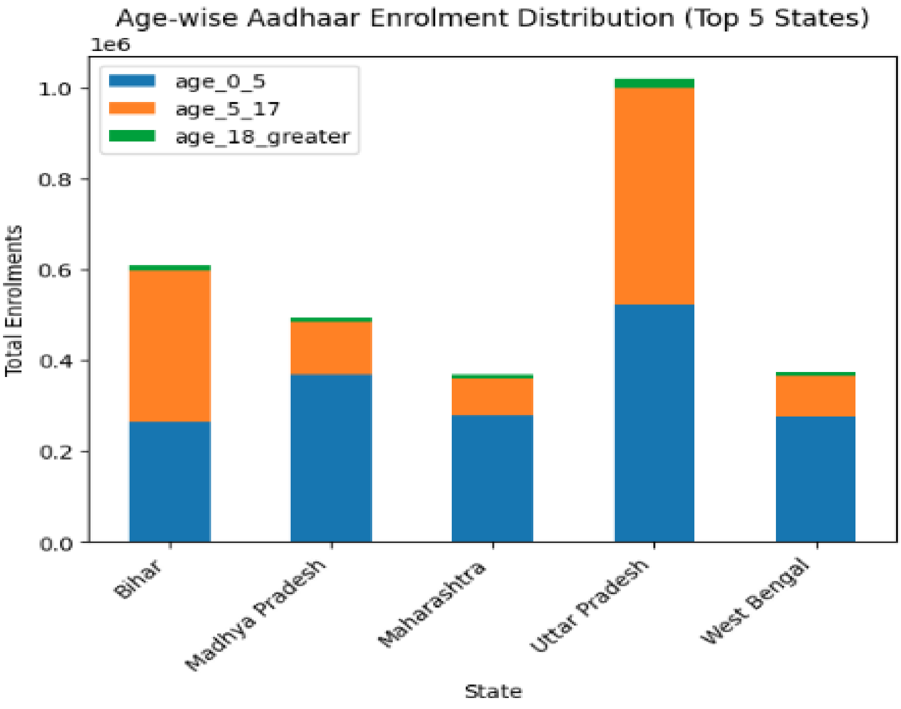
Code Used:

```
top_states =
(df.groupby('state')['total_enrolment'].sum().sort_values(ascending=False).head(5).index)

state_age = (df[df['state'].isin(top_states)].groupby('state')[['age_0_5', 'age_5_17',
'age_18_greater']].sum())

plt.figure()
state_age.plot(kind='bar', stacked=True)
plt.title("Age-wise Aadhaar Enrolment Distribution (Top 5 States)")
plt.xlabel("State")
plt.ylabel("Total Enrolments")
plt.xticks(rotation=45, ha='right')
plt.show()
```

Visualisations:



Key Observations:

- Child enrolment dominates across all top states.
- Adult enrolment remains consistently low in every state.

Interpretation:

- The dominance of child enrolment across regions suggests a **uniform national pattern of early-life inclusion**.

2.2 Monthly Enrolment Trends by Age Group

Objective:

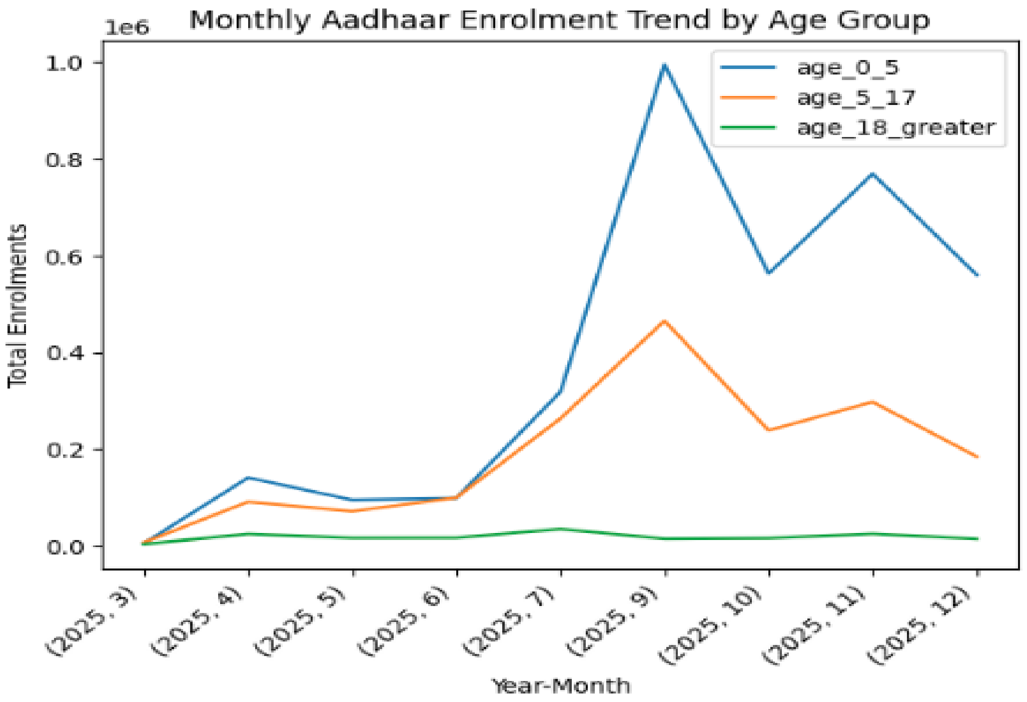
To analyse how enrolment trends vary over time for different age groups.

Code Used:

```
monthly_age = (df.groupby(['year', 'month'])[['age_0_5', 'age_5_17', 'age_18_greater']].sum())

plt.figure()
monthly_age.plot()
plt.title("Monthly Aadhaar Enrolment Trend by Age Group")
plt.xlabel("Year-Month")
plt.xticks(rotation=45, ha='right')
plt.ylabel("Total Enrolments")
plt.show()
```

Visualisations:



Key Observations:

- Clear seasonal peaks are visible, especially for child enrolments.
- Adult enrolment remains relatively flat across months.

Interpretation:

- Seasonal enrolment patterns indicate alignment with **institutional schedules and enrolment drives**.

2.3 Adult Enrolment Share by State

Objective:

To identify states with relatively higher adult enrolment proportions.

Code Used:

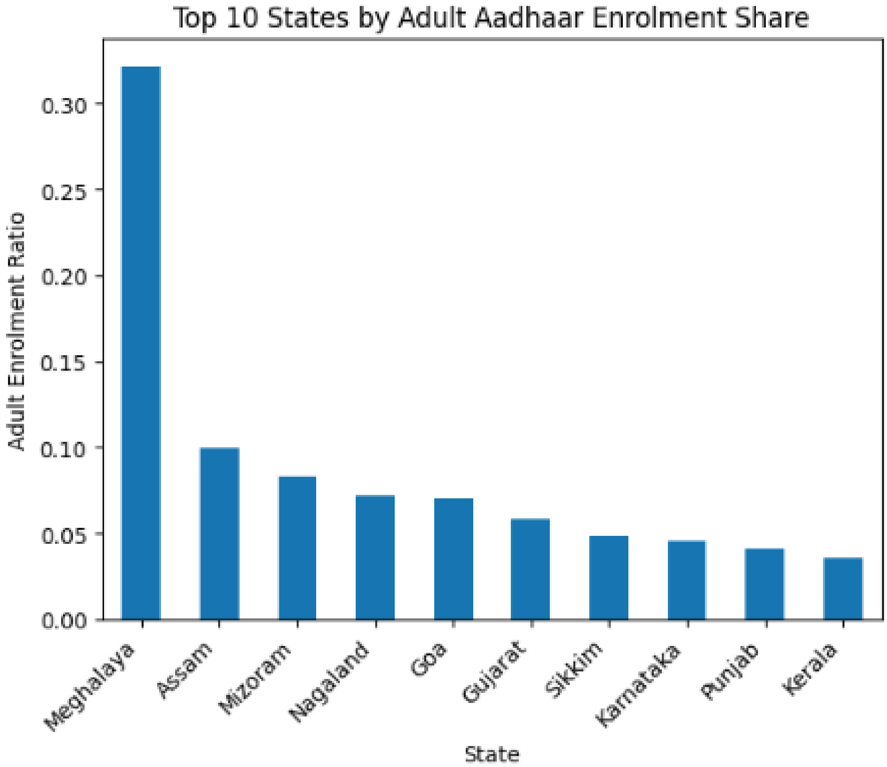
```
adult_share = (df.groupby('state')[['age_18_greater', 'total_enrolment']].sum())

adult_share['adult_ratio'] = adult_share['age_18_greater'] / adult_share['total_enrolment']

adult_share_top = adult_share['adult_ratio'].sort_values(ascending=False).head(10)

plt.figure()
adult_share_top.plot(kind='bar')
plt.title("Top 10 States by Adult Aadhaar Enrolment Share")
plt.xlabel("State")
plt.ylabel("Adult Enrolment Ratio")
plt.xticks(rotation=45, ha='right')
plt.show()
```

Visualisations:



Key Observations:

- Smaller states and UTs show relatively higher adult enrolment ratios.

Interpretation:

- Higher adult enrolment ratios may indicate **late adoption, migration, or targeted outreach efforts**.

3. Trivariate Analysis

Trivariate analysis was used to examine interactions between **state, time, and age group**.

3.1 Monthly Age-wise Trends for High-Enrolment States

Objective:

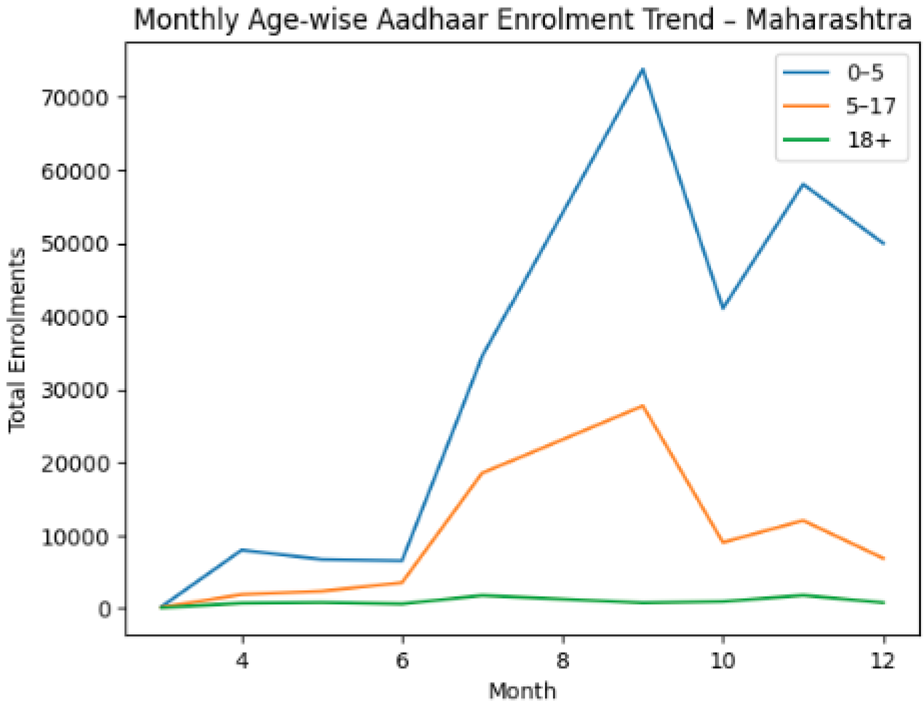
To analyse how enrolment patterns vary across states and months for different age groups.

Code Used:

```
for state in top_states:
    subset = (
        df[df['state'] == state].groupby('month')[['age_0_5', 'age_5_17',
'age_18_greater']].sum().reset_index())

plt.figure()
plt.plot(subset['month'], subset['age_0_5'], label='0–5')
plt.plot(subset['month'], subset['age_5_17'], label='5–17')
plt.plot(subset['month'], subset['age_18_greater'], label='18+')
plt.title(f"Monthly Age-wise Aadhaar Enrolment Trend – {state}")
plt.xlabel("Month")
plt.ylabel("Total Enrolments")
plt.legend()
plt.show()
```

Visualisations:



Key Observations:

- Enrolment peaks are concentrated in specific months.
- Child enrolment drives these peaks.

Interpretation:

- Aadhaar enrolment is largely **campaign-based rather than continuous**, especially for children.

3.2 Enrolment Volatility Across States

Objective:

To identify regions with unstable enrolment activity.

Code Used:

```
state_volatility =
(df.groupby('state')['total_enrolment'].std().sort_values(ascending=False).head(10))
```

Key Observations:

- Smaller states and UTs exhibit higher enrolment volatility.

Interpretation:

- High volatility suggests **episodic enrolment drives and limited infrastructure availability**.

3.3 Delayed Aadhaar Adoption Indicator

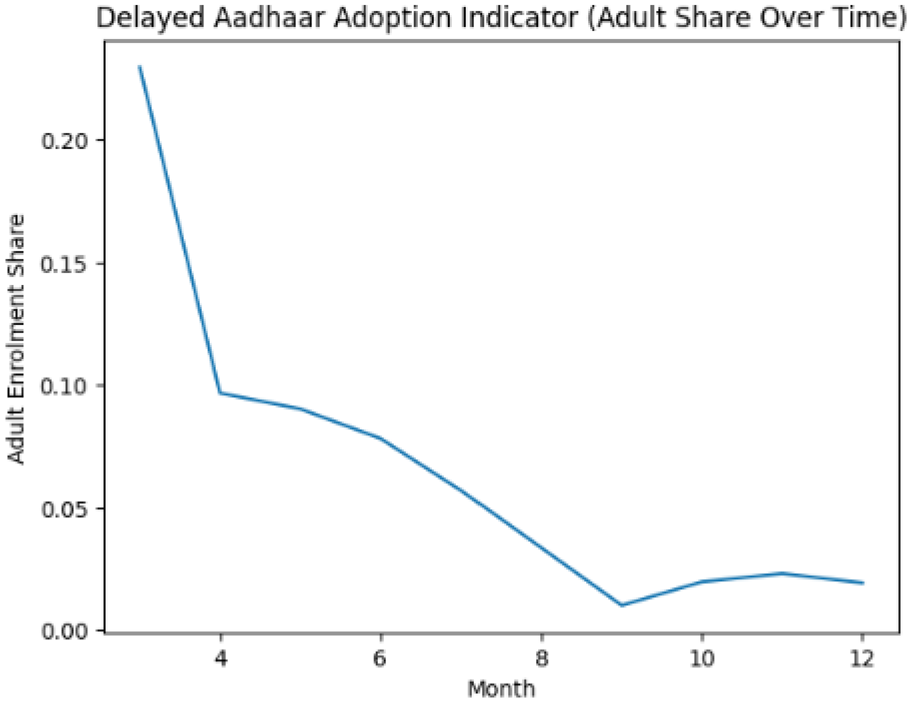
Objective:

To examine adult enrolment trends over time.

Code Used:

```
adult_monthly = (df.groupby(['year', 'month'])[['age_18_greater',
'total_enrolment']].sum().reset_index())
adult_monthly['adult_share'] = adult_monthly['age_18_greater'] / adult_monthly['total_enrolment']
plt.figure()
plt.plot(adult_monthly['month'], adult_monthly['adult_share'])
plt.title("Delayed Aadhaar Adoption Indicator (Adult Share Over Time)")
plt.xlabel("Month")
plt.ylabel("Adult Enrolment Share")
plt.show()
```

Visualisations:



Key Observations:

- Adult enrolment share declines steadily over time.

Interpretation:

- This confirms **progressive saturation of Aadhaar among adults**, with remaining enrolments limited to migrants or previously excluded populations.

Significance of the Study

This analysis demonstrates how administrative datasets, when analysed systematically, can move beyond descriptive statistics to provide **decision-support insights**. The findings have practical relevance for:

- Government agencies
- Programme administrators
- Policy planners
- Digital governance initiatives

By leveraging Aadhaar enrolment data effectively, the study contributes to more **efficient, inclusive, and evidence-based governance**

Impact & Applicability

Administrative Impact

- **Capacity Planning:**
  Seasonal enrolment surges indicate the need for flexible staffing and infrastructure during peak months.
- **Targeted Outreach:**
  Adult enrolment patterns can help identify migrant or previously excluded populations.
- **Operational Efficiency:**
  Volatility analysis helps identify regions where enrolment services are irregular and require stabilisation.

Societal Impact

- Confirms **early-life digital inclusion** as the dominant enrolment pathway.
- Highlights the transition of Aadhaar from an enrolment-focused system to a **maintenance and update-oriented ecosystem**.

Policy Recommendations

1. **Seasonal Resource Allocation**
   Deploy additional enrolment centres and mobile units during historically high-demand months.
2. **Focus on Migrant Populations**
   Use adult enrolment as a proxy to design targeted outreach for migrant and mobile populations.
3. **Stabilise High-Volatility Regions**
   Replace episodic enrolment drives with continuous service availability in smaller states and UTs.
4. **Data-Driven Planning**
   Institutionalise enrolment analytics to guide annual operational planning and budgeting.