SHIVAM
2211CS010536
SECTION-2

An Analytical Deep Dive into Aadhaar Monthly Enrolment Data

A Big Data Approach to Demographic Enrolment Trend Analysis

Project Report

Table of Contents

---

1. Executive Summary & Introduction

1.1 Project Overview

This report presents an in-depth analysis of Aadhaar monthly enrolment figures across various states, districts, and pin codes. The study focuses on quantifying total enrolments and identifying key demographic and geographic hotspots based on enrolment volume between **March and July 2025**.

By leveraging a **Big Data Analytics** approach using **PySpark**, the project successfully processes large volumes of enrolment records to generate actionable insights into demographic penetration and regional performance.

1.2 Core Focus: Demographic and Geographic Trends

The primary goal of this analysis is to:

- Determine the total number of Aadhaar enrolments within the period.
- Identify the age group distribution across various states (Age 0-5, 5-17, 18+ years).

- Pinpoint specific pin codes and states exhibiting the highest and lowest enrolment activities.

---

2. Methodology and Technical Stack

2.1 Data Source and Schema

- **Data Source:** Enrolment_data_March-July (2).csv
- **Time Period:** March to July 2025 (as indicated by file name)
- **Key Columns:** Date, State, District, Pincode, Age_0_5, Age_5_17, and Age_18_greater.
- **Derived Metric:** A new column, **Total_Enrolments**, was calculated as the sum of all three age groups (Age_0_5 + Age_5_17 + Age_18_greater) to represent the total monthly enrolment activity at the record level.

2.2 Technology Stack (PySpark & Visualization)

Tool	Purpose
PySpark	Core framework used for distributed processing, data manipulation, aggregation, and calculation of metrics (Big Data Approach).
Python/Pandas	Used for data transfer and structuring of results for visualization.
Matplotlib/Seaborn	Used to generate insightful data visualizations (bar plots).
Jupyter Notebook	Interactive environment for development and documentation of the analysis.

---

3. Key Findings and Detailed Analysis

3.1 Overall Enrolment Volume

The **Overall Total Aadhaar Enrolments** recorded across all districts and states in the dataset is **1,291,042**. This figure represents the total volume of enrolment activity during the period under review.

3.2 Geographic Hotspots (Top/Minimum Enrolments)

This section pinpoints the extreme values in enrolment activity, highlighting areas of high focus and areas that may require additional support.

- **Top 5 Pincodes by Aadhaar Enrolments:** The analysis reveals specific micro-regions driving the highest enrolment volumes, likely due to local campaigns or high population density.

Rank	Pincode	Total Enrolments
1	**244001**	**10,380**
2	**793119**	**9,690**
3	**202001**	**8,574**
4	**110059**	**7,424**
5	**244901**	**6,594**

- **State with Minimum Aadhaar Enrolments:** The state of **Goa** recorded the lowest enrolment volume at **42**. This may suggest near-saturation of the target population or limited reporting during the analyzed period.

3.3 Demographic Distribution by Age (State-Wise)

The notebook generates a data frame (`state_age`) to visualize the enrolment breakdown by age group for the top 10 states. This is crucial for understanding whether enrolment drives are effectively targeting children (**Age 0-5** and **Age 5-17**) versus adults (**Age 18+**).

Key Columns Calculated in `state_age` Data Frame:

- `State`
- `Age 0-5` (Total enrolments)
- `Age 5-17` (Total enrolments)
- `Age 18+` (Total enrolments)

4. Data Visualization (Conceptual)

Age Group Distribution Across States

Insight: The visualization, generated using PySpark and Matplotlib, clearly shows the contribution of each age group to the overall enrolment figures for the top 10 states. While specific distribution varies by state, the pattern helps identify if enrolment drives are successfully targeting the critical **Age 0-5** cohort (new enrolments) or the larger **Age 18+** cohort (updates/corrections/initial adult enrolment).

---

5. Conclusion & Future Scope

5.1 Project Conclusion

This analysis, powered by **PySpark's distributed computing capabilities**, provides a clear, quantitative snapshot of Aadhaar enrolment trends between March and July 2025. Key insights include the identification of high-performing pincodes and the significant regional disparities in enrolment volume. The overall enrolment figure of **1,291,042** serves as a vital benchmark for performance evaluation. The methodology employed is scalable and highly efficient for handling national-level, large-scale datasets, consistent with a Big Data approach.

5.2 Future Scope

To further enhance this study, future work could include:

- **Time Series Analysis:** Analyzing enrolment data over multiple years to identify long-term trends and seasonality.
- **Geospatial Integration:** Mapping the enrolment hotspots (pincode data) to visualize geographic coverage and identify areas needing enhanced outreach.
- **Target Population Comparison:** Comparing the total enrolments to estimated state and district-wise target populations to calculate true coverage or penetration rates.