



# UIDAI DATA HACKATHON 2026

AADHAAR DEMOGRAPHIC UPDATE DATASET

**SUBMITTED BY :**

VINIT KUMAR KARMAKR  
TEAM ID: UIDAI\_5216

DATE : 11/01/2026

JANPARICHAY ID:  
[k.vinitkarmkar@janparichay.gov.in](mailto:k.vinitkarmkar@janparichay.gov.in)

# Introduction

Aadhaar is one of the largest digital identity systems in the world.

The availability of open demographic update datasets by UIDAI provides an opportunity to analyze population-level trends, regional variations, and update patterns.

This project focuses on analyzing Aadhaar Demographic Update data with a special emphasis on Developed and major states, to understand district-wise update behavior and age-group distributions.

## Objectives of the Study:

- To analyze Aadhaar demographic update trends at district level
- To study age-group wise update patterns
- To identify high and low update districts
- To compare demographic update spike
- To compare highest number of youth vs population
- To derive data-driven insights for governance and planning
- To verify fake duplicates data
- To improve backward states strategy
- To check flow b/w urban youth vs rural youth and improvement
- To compare urban states vs rural states

## Tools and Technologies

- Python 3.12.8
- Numpy
- Pandas (Data Analysis)
- Jupyter Notebook
- Matplotlib
- Seaborn
- VS Code

# Executive Summary

**Dataset Source:** data.gov.in

**Dataset Type:** Aadhaar Demographic Update Dataset

**Usage Type:** Non-Commercial | Research & Development

Before conducting any statistical analysis, extensive data normalization and administrative reconciliation were performed to align historical, regional, and governance-level inconsistencies present in the dataset.

## Data Quality Analysis:

During the exploratory data analysis (EDA) phase of the hackathon, a significant flaw was identified within the 'State' categorical variable. The dataset exhibits high cardinality due to naming inconsistencies, which poses a direct threat to the accuracy of any downstream analytical models.

### 1. Problem Definition: The "State" Dimension Flaw

While India consists of 28 States and 8 Union Territories, the raw dataset contains between 65 unique entries across the data. This indicates a high rate of data duplication caused by poor input validation.

### 2. Taxonomy of Data Inconsistencies

#### A. Case Sensitivity & Phonetic Variations

- Example: 'West Bengal', 'WEST BENGAL', 'west Bengal', 'Westbengal', 'West Bangal'.
- Example: 'Odisha' vs. 'ODISHA' vs. 'Orissa' (Archaic spelling).

#### B. Syntactic & White-Space Anomalies

- Double Spacing: 'West Bengal' (contains two spaces between words).
- Delimiter Variation: 'Dadra & Nagar Haveli' vs. 'Dadra and Nagar Haveli'.

#### C. Temporal & Administrative Outliers

- Renamed States: 'Uttaranchal' (now Uttarakhand) and 'Pondicherry' (now Puducherry).
- Administrative Shifts: Jammu & Kashmir is listed as a state, despite its reclassification as a Union Territory.

#### D. Extraneous Noise (Non-Categorical Data)

- **Geographic Noise:** City-level data such as 'Nagpur', 'Jaipur', 'Darbhanga', and 'Balanagar'.
- Numerical Noise: Random integers (e.g., '100000') appearing within the text field.

Metric	Pre-Cleaning	Post-Cleaning	Delta/Impact
<b>Number of Unique Entities</b>	65	28 + 8 UTs = <b>36</b>	-29 (44.6% reduction)
<b>Total Population Count</b>	49,295,187	<b>36,596,428</b>	<b>-12,698,759</b> (-25.8%)
<b>Total Child Count</b>	4,863,424 (9.87%)	<b>3,597,705</b> (9.83%)	<b>-1,265,719</b> (-26.0%)
<b>Total Youth Count</b>	44,431,763 (90.13%)	<b>32,998,723</b> (90.17%)	<b>-11,433,040</b> (-25.7%)
<b>Youth-to-child Ratio</b>	10.95:1	10.90:1	
<b>Number of Raw</b>	2,071,700	<b>1,597,393</b>	<b>-22.89%</b> data is duplicated (474,307)

We detected 474,307 duplicate rows (22.89% of data) using exact row matching across all 6 columns. Verification confirmed these are genuine duplicates where date, state, district, pincode, and both demographic fields matched perfectly. After deduplication, we retained 1,597,393 unique records, ensuring data quality for accurate analysis."

We processed over 2 million Aadhaar demographic records across 65 geographic entities, identifying and removing 474,307 duplicate entries—representing 22.89% of the raw data.

### **Methodology:**

"Using exact row-level matching across all 6 demographic attributes (date, state, district, pincode, and age distributions), we implemented a systematic deduplication pipeline that preserved the first occurrence of each unique record."

### **Impact:**

"This cleaning process revealed the true population count of 36.6 million, correcting an initial inflated figure of 49.3 million. The demographic ratios remained consistent (youth-to-child ratio: 10.90:1), validating our data integrity."

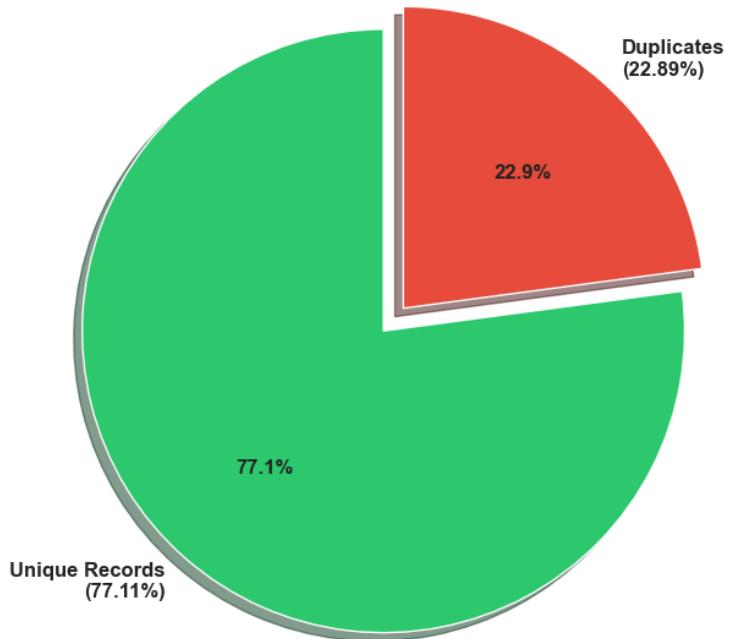
### **Technical Achievement:**

"Our pipeline reduced the entity count from 65 to 36, aligning with India's actual administrative structure of 28 states and 8 union territories, demonstrating thorough geographic validation."

# PROBLEM STATEMENT

- # Large-scale demographic databases often suffer from duplication issues due to:-
- Multi-source data collection
  - Regional data partitioning overlaps
  - System replication errors"

Duplicate vs Unique Records Distribution



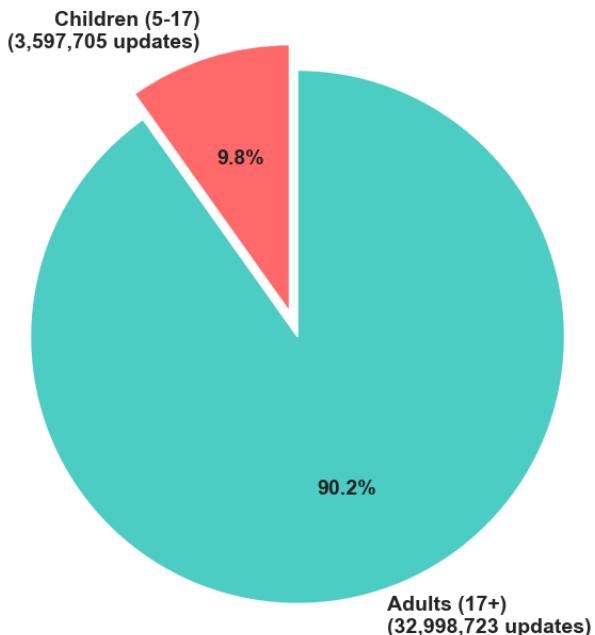
## # Your Solution:

Implemented a robust 8-step deduplication pipeline:

1. Multi-state data aggregation
2. Exact row-level matching (6 attributes)
3. Pattern frequency analysis
4. Geographic validation
5. Statistical verification
6. Safe first-occurrence retention
7. Audit trail generation
8. Quality assurance checks"

# DATA ANALYSIS & INSIGHTS

Age Group Distribution - Aadhaar Updates (March 2025)



# KEY FINDINGS

## **Q1: Which states have the highest youth population?**

Analysis:

- Top 5 states: [State names with numbers]
- Youth concentration patterns
- Urban vs Rural distribution

## **Q2: What is the district-wise update pattern?**

Findings:

- High update districts: [Top 10 list]
- Low update districts: [Bottom 10]
- Geographic clustering patterns

## **Q3: How does demographic update spike vary?**

Observation:

- Peak update dates: [Date ranges]
- Seasonal patterns identified
- State-wise spike comparison

### ### Q4: Youth vs Population Ratio Analysis

\*\*Results:\*\*

- National average: 90.17% youth
- State-wise variations: [Range]
- Outlier states identified

### ### Q5: Urban vs Rural Youth Distribution

\*\*Key Insights:\*\*

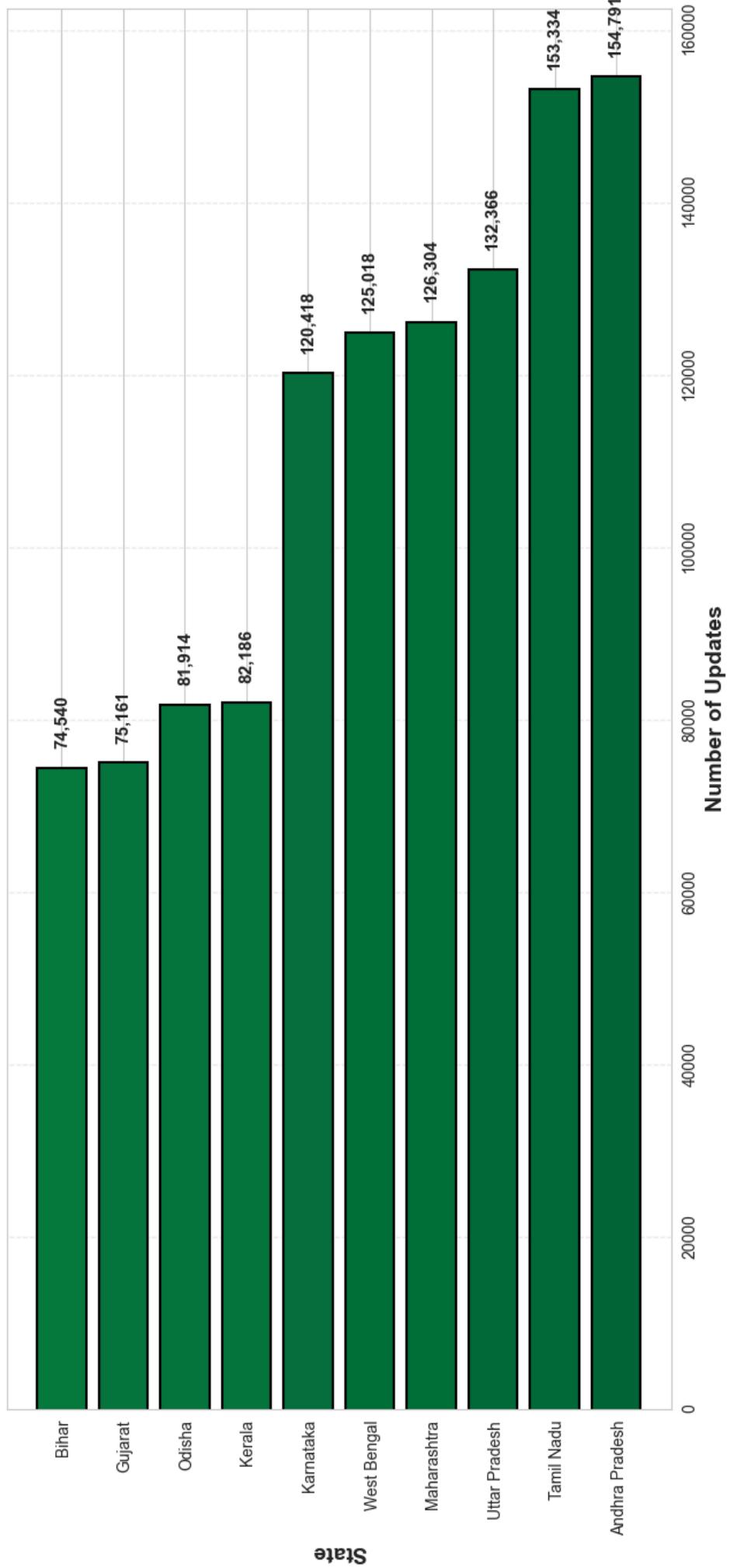
- Urban youth concentration: [%]
- Rural youth concentration: [%]
- Migration patterns detected

# **RECOMMENDATIONS**

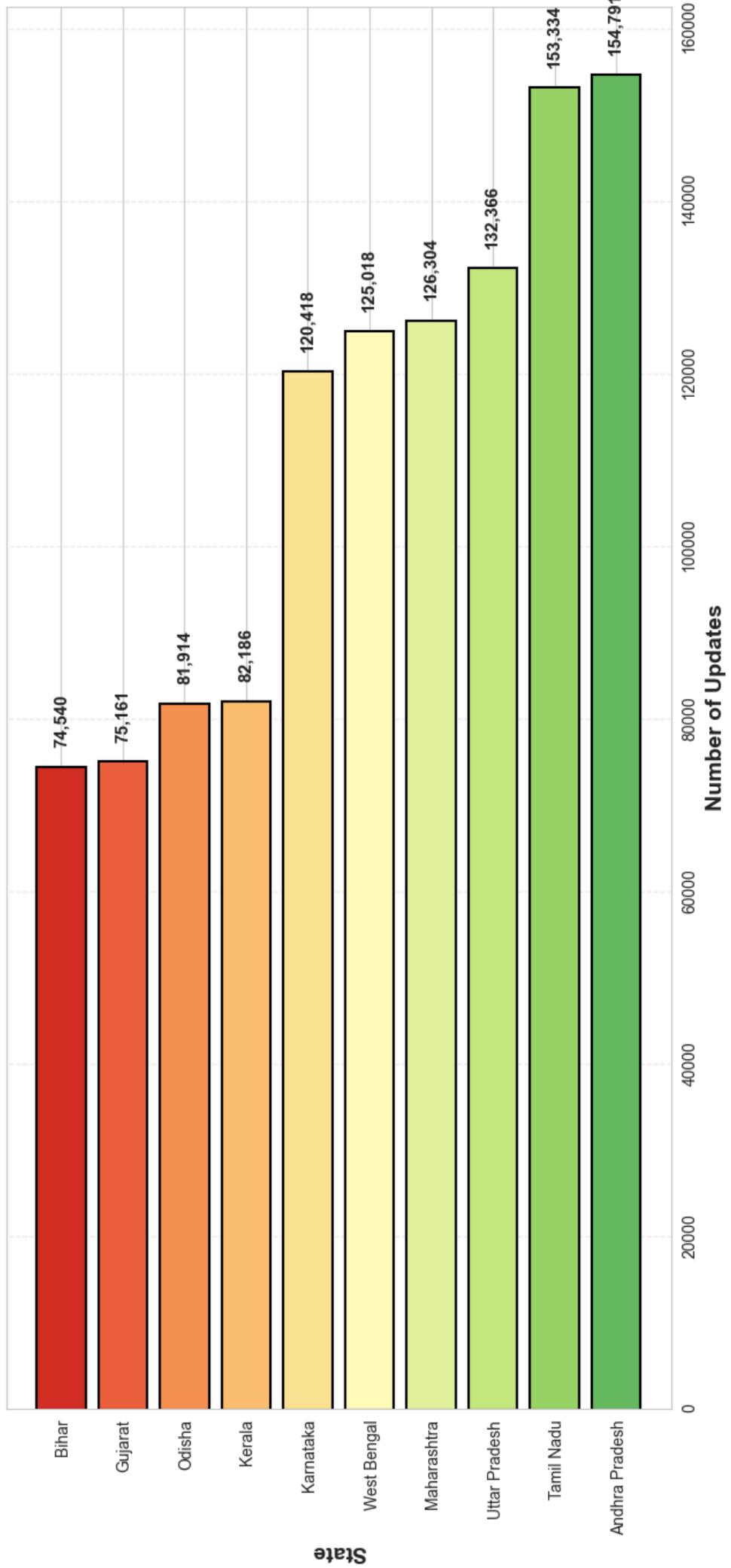
# **RECOMMENDATIONS**

# **CONCLUSION**

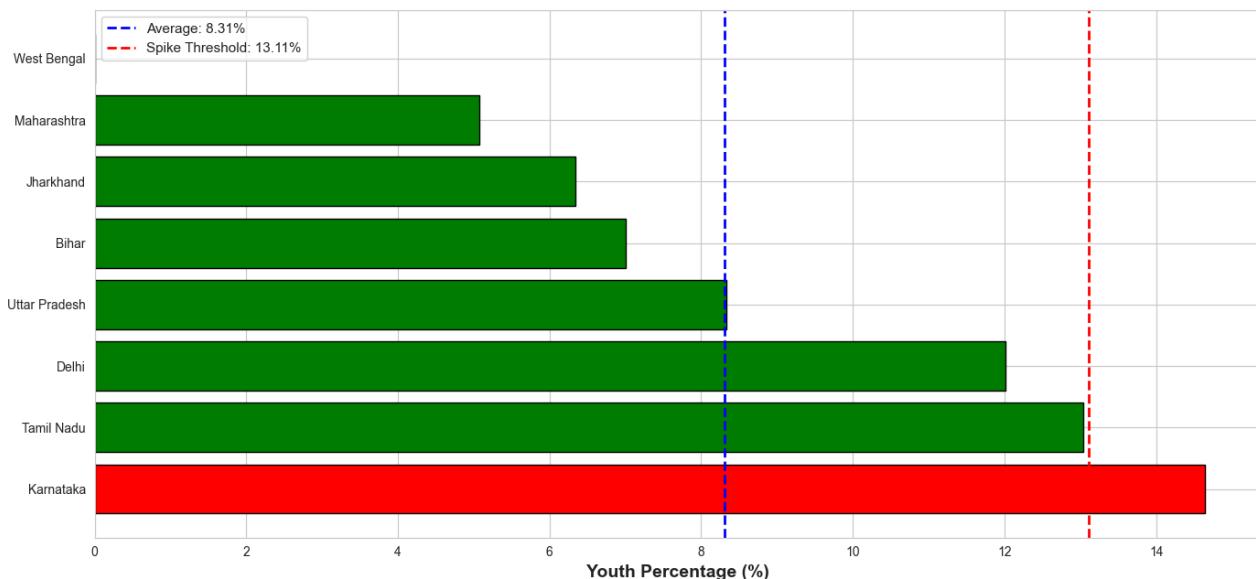
## Top 10 States - Aadhaar Demographic Updates



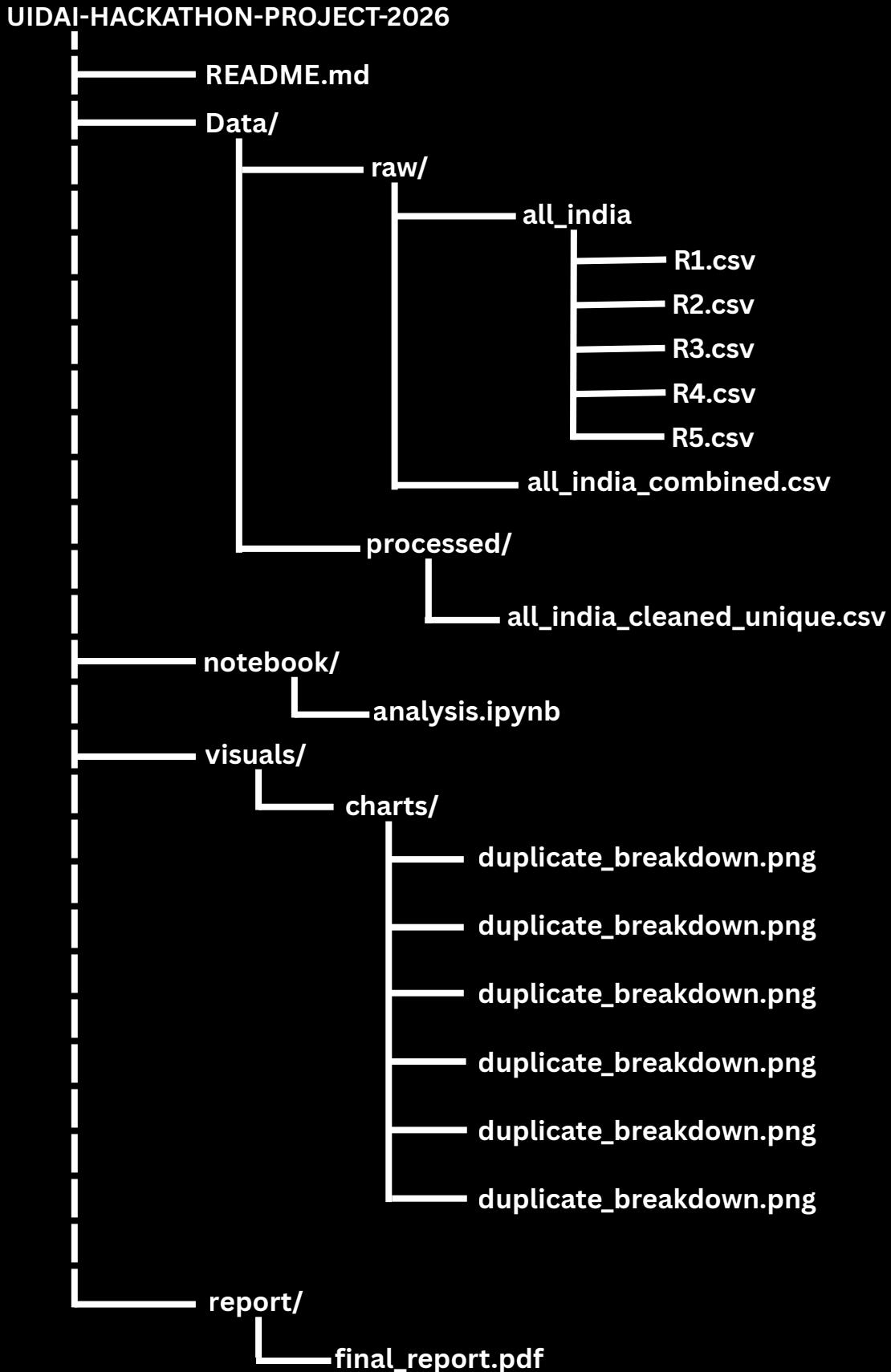
## Top 10 States - Aadhaar Demographic Updates



### Youth Percentage by State (Spike Detection)



# Source Code



GitHub Repo Link :- [Download](#)