Technical Documentation: Data Engineering Pipeline (Notebook 1)

Project: Aadhaar Societal Trends Analysis

Notebook: Merging.ipynb

Objective: To ingest, consolidate, clean, and merge fragmented Aadhaar datasets (Enrollment, Biometric, Demographic) into a single, reliable analytical master dataset.

1. Data Ingestion & Concatenation

The raw data was provided in multiple split CSV files to handle file size limits.

Enrollment Data

- **Source:** 3 split CSV files (0_500000, 500000_1000000, 1000000_1006029).
- **Action:** Loaded individually and concatenated using pd.concat.
- **Raw Shape:** ~1.006 Million rows.
- **Key Columns:** date, state, district, pincode, age_0_5, age_5_17, age_18_greater.

Biometric Data

- **Source:** 4 split CSV files.
- **Action:** Concatenated into Biometric_df.
- **Raw Shape:** ~1.86 Million rows.
- **Key Columns:** bio_age_5_17, bio_age_17_ (Note: Infants 0-5 do not perform biometric updates).

Demographic Data

- **Source:** 5 split CSV files.
- **Action:** Concatenated into Demography_df.
- **Raw Shape:** ~2.07 Million rows.
- **Key Columns:** demo_age_5_17, demo_age_17_.

2. Initial Cleaning & Feature Engineering

Before merging, each individual dataframe underwent standard preprocessing to ensure schema consistency.

- **Duplicate Removal:** df.drop_duplicates(inplace=True) removed approximately 22k enrollment duplicates and 94k biometric duplicates.
- **Date Standardization:** Converted the string date column to Python datetime objects using format='%d-%m-%Y' to prevent parsing errors.
- **Total Calculation:** Created new aggregate columns to simplify downstream analysis:
    - total_enrollment = Sum of all age groups.
    - total_biometric = Sum of biometric updates.

- total_demography = Sum of demographic updates.

3. The Master Merge Strategy

This was the most critical step to align the three disparate datasets.

- **Merge Keys:** ['date', 'state', 'district', 'pincode']. We joined on *all* four grains to ensure specific location-time accuracy.
- **Join Type: Outer Join** (how='outer').
  - *Reasoning:* An area might have updates but zero new enrollments on a specific day (or vice versa). An inner join would have resulted in massive data loss.
- **Handling Nulls:** Post-merge NaN values were filled with 0.
  - *Logic:* If a district appears in the Enrollment data but not Biometric data for a specific date, it implies **0 biometric updates** happened, not missing data.

4. The "Gold Standard" Cleaning Logic

The raw data contained significant inconsistencies in State/District naming (e.g., "West Bangal" vs "West Bengal"). We implemented a custom function clean_aadhar_master_final to fix this.

A. Text Standardization

- Converted all string columns to **Title Case**.
- Removed extra whitespaces and asterisks (e.g., Gondiya * $\rightarrow$ Gondia).

B. Dictionary-Based Mapping (Hard Cleaning)

We manually mapped incorrect spellings to the official government census names.

- **States Fixed:** Odisha (from Orissa), West Bengal (from West Bangal/Westbengal), Uttarakhand (from Uttaranchal), Jammu & Kashmir, etc.
- **Districts Fixed:** Barabanki, Haora (Howrah), Hooghly (Hugli), South 24 Parganas, Sri Potti Sriramulu Nellore, Rangareddy.

C. Pincode Validation

- Forced pincode to numeric type.
- **Filter Logic:** Retained only valid 6-digit Indian pincodes within the range 110000 to 850000. This removed junk entries like 100000 or 999999.

D. Final Aggregation

- **Problem:** Renaming "Orissa" to "Odisha" creates duplicate rows for the same date/district (one from the original "Orissa" rows, one from "Odisha").
- **Solution:** Applied a final groupby(['date', 'state', 'district', 'pincode']).sum() to merge these rows back together.

5. Final Data Integrity Check

Before export, the data was validated to ensure mathematical consistency.

- **Consistency Check:** Verified that age_0_5 + age_5_17 + age_18_greater strictly equals total_enrollment.
- **Final Output Stats:**
    - **Total Rows:** 2,234,840
    - **Unique States:** 36 (Cleaned & Standardized)
    - **Unique Districts:** 938
    - **Missing Values:** 0 (All handled)

6. Output

- **File Name:** Aadhar_Data.csv
- **Description:** A fully cleaned, merged, and standardized dataset ready for Exploratory Data Analysis (EDA) and Feature Engineering.