Data Preparation - ADIEWS

Notebook: `00_data_preparation.ipynb`
Status: Complete
Output Dataset: `aadhaar_demographics_cleaned.csv`

Overview

This phase establishes the foundational dataset for the entire ADIEWS framework by cleaning, transforming, and enriching raw Aadhaar demographic update records. The preparation ensures data quality, consistency, and readiness for downstream analysis.

Input Data

Raw Dataset Characteristics

- **Source Files**: Multiple CSV files in `DemographicData/` folder
- **Initial Records**: 2,375,882 raw entries
- **Date Range**: March 2025 - January 2026 (10 months)
- **Geographic Scope**: 1,056 districts across India

Raw Schema

```
Columns:
- date: Update date (YYYY-MM-DD format)
- pincode: 6-digit area code
- district: District name
- state: State/UT name
- child_updates: Updates for ages 5-17
- adult_updates: Updates for ages 17+
```

Data Cleaning Steps

1. Missing Value Handling

- **Approach**: Complete case analysis (remove rows with any null values)
- **Rationale**: Geographic identifiers (pincode, district, state) are critical for spatial analysis
- **Result**: 100% complete records retained

2. Date Standardization

- **Conversion**: String dates → `datetime64[ns]` type
- **Validation**: Ensure all dates fall within expected range
- **Temporal Features**: Extract month, year for aggregation

3. Geographic Data Validation

- **Pincode Format**: Verify 6-digit numeric codes
- **District Names**: Standardize spelling and formatting
- **State Mapping**: Validate against known state list
- **Deduplication**: Remove duplicate pincode-district-date combinations

4. Numeric Data Quality

- **Non-Negative Constraint**: Ensure child_updates ≥ 0, adult_updates ≥ 0
- **Outlier Detection**: Flag extreme values (>99.9th percentile)
- **Data Type**: Convert to `int64` for update counts

Feature Engineering

Derived Metrics

1. Total Updates

```
total_updates = child_updates + adult_updates
```

Represents total demographic activity per record.

2. Child Share Percentage

```
child_share_pct = (child_updates / total_updates) * 100
```

Indicates proportion of updates involving children (ages 5-17).

3. Age Ratio

```
age_ratio = adult_updates / (child_updates + 1)  # +1 to avoid division by zero
```

Measures adult-to-child update intensity.

4. Temporal Features

```
month = date.month
year = date.year
month_name = date.strftime('%Y-%m')
```

Enables time-series and seasonal analysis.

Aggregation Pipelines

1. District-Level Monthly Aggregates (`df_monthly_district.csv`)

```
Grouping: [district, state, month]
Aggregations:
  - child_updates: sum, mean, std
  - adult_updates: sum, mean, std
  - total_updates: sum
  - child_share_pct: mean
  - pincode_count: unique count
```

Records: 10,560 (1,056 districts × 10 months)

2. State-Level Monthly Aggregates (`df_monthly_state.csv`)

```
Grouping: [state, month]
Aggregations:
  - child_updates: sum
  - adult_updates: sum
  - district_count: unique count
```

Records: 630 (63 states × 10 months)

3. Pincode-Level Monthly Aggregates (`df_monthly_pincode.csv`)

```
Grouping: [pincode, district, state, month]
Aggregations:
  - child_updates: sum
  - adult_updates: sum
  - child_share_pct: mean
```

Records: ~200,000 pincode-month combinations

Data Quality Metrics

Completeness

Metric	Value
Total Records	2,375,882
Complete Records	2,375,882 (100%)
Missing Values	0
Duplicate Records	0

Statistical Summary

```
Child Updates (ages 5-17):
  Mean: 1.9 per record
  Median: 0 (44% are zero)
  Max: 2,690
  Total: 4.5M updates

Adult Updates (ages 17+):
  Mean: 19.1 per record
  Median: 5
  Max: 16,166
  Total: 45.4M updates
```

Output Files

File	Format	Records	Description
`aadhaar_demographics_cleaned.csv`	CSV	2.4M	Cleaned daily pincode-level data
`df_monthly_district.csv`	CSV	10,560	District monthly aggregates
`df_monthly_state.csv`	CSV	630	State monthly aggregates
`df_monthly_pincode.csv`	CSV	~200K	Pincode monthly aggregates
`district_summary.csv`	CSV	1,056	District-level summary statistics
`pincode_summary.csv`	CSV	~20K	Pincode-level summary statistics

🔑 Key Insights

Geographic Coverage

- **1,056 districts** across 63 states/UTs
- **~20,000 unique pincodes**
- Comprehensive national coverage

Temporal Patterns

- **Peak Month**: December 2025 (10.51M updates)
- **Lowest Month**: January 2026 (583K updates)
- **18× variation** between peak and trough

Child vs Adult Updates

- **Child Updates**: 9.07% of total
- **Adult Updates**: 90.93% of total
- **Zero-Child Records**: 44% of all records

✅ Validation Checks

Data Integrity

- ✅ No negative update counts
- ✅ All dates within valid range
- ✅ All pincodes 6-digit numeric
- ✅ All districts mapped to valid states
- ✅ No orphaned geographic records

Statistical Consistency

- ✅ Total updates = child + adult (all records)
- ✅ Child share ≤ 100% (all records)
- ✅ Age ratio ≥ 0 (all records)
- ✅ Monthly aggregates sum to raw totals

🚀 Next Steps

1. **Univariate Analysis** → Explore distributions
2. **Bivariate Analysis** → Identify relationships
3. **Trivariate Analysis** → Uncover complex patterns

4. **Layer 1: Migration Radar** → Detect population movements
5. **Layer 2: Child Risk Map** → Identify documentation gaps
6. **Layer 3: System Intelligence** → Assess stability
7. **Layer 4: Early Warning** → Generate alerts

Technical Notes

Performance Optimizations

- **Data Types**: Optimized int32/int64 for memory efficiency
- **Indexing**: Date and district columns indexed for faster queries
- **Pickle Serialization**: `.pkl` files for faster loading in subsequent notebooks

Reproducibility

- **Random Seed**: Not applicable (deterministic processing)
- **Environment**: Python 3.8+, pandas 1.3+, numpy 1.21+
- **Execution Time**: ~5 minutes on standard hardware

Last Updated: January 2026
Maintainer: ADIEWS Project Team