

Univariate Analysis - ADIEWS

Notebook: 01_univariate_analysis.ipynb

Status: ☒ Complete

Visualizations: 8 PNG files

Overview

This phase conducts comprehensive single-variable analysis of Aadhaar demographic update patterns to understand distributions, central tendencies, and temporal trends. The analysis forms the statistical foundation for more complex multivariate explorations.

☒ Analysis Components

1. Child Updates Distribution

Histogram Analysis (01_child_updates_histogram.png) **Key Findings:** - **Mean:** 1.9 updates per record - **Median:** 0 (extreme right skew) - **Mode:** 0 (44% of records) - **Maximum:** 2,690 updates - **Standard Deviation:** High variability

Interpretation: - Massive concentration at zero indicates many areas with minimal child documentation activity - Long right tail suggests specific locations (schools, enrollment centers) drive high volumes - Policy implication: 44% zero rate signals geographic gaps in child Aadhaar coverage

Box Plot Analysis (02_child_updates_boxplot.png)

- **Q1 (25th percentile):** 0
 - **Q2 (50th percentile/Median):** 0
 - **Q3 (75th percentile):** 2
 - **IQR:** 2
 - **Outliers:** >1000 records with counts >50
-

2. Adult Updates Distribution

Histogram Analysis (03_adult_updates_histogram.png) **Key Findings:** - **Mean:** 19.1 updates per record - **Median:** 5 - **Mode:** 0-5

range (most common) - **Maximum:** 16,166 updates - **Distribution:** Right-skewed but less extreme than child updates

Interpretation: - Adult updates 10× more frequent than child updates - More consistent across geography (lower zero rate ~20%)
- Outliers likely represent urban service centers or corporate enrollment drives

Box Plot Analysis (04_adult_updates_boxplot.png)

- **Q1:** 0
 - **Q2:** 5
 - **Q3:** 16
 - **IQR:** 16
 - **Outliers:** >2000 records with counts >100
-

3. Child vs Adult Comparison

Comparative Box Plots (05_child_vs_adult_comparison.png) Key

Contrasts: | Metric | Child | Adult | Ratio | |----|----|----|----| |
Median | 0 | 5 | ∞ | | Mean | 1.9 | 19.1 | 10.05× | | Q3 | 2 | 16 | 8× | |
Max | 2,690 | 16,166 | 6× |

Insights: - Adult updates dominate across all statistical measures - Child documentation systematically lags behind adult activity - Geographic disparities more pronounced for children

4. Temporal Patterns

Monthly Trends Time Series (06_monthly_trends_timeseries.png)

Observed Patterns: - **Peak Month:** December 2025 (10.51M total updates) - **Lowest Month:** January 2026 (583K updates) - **Amplitude:** 18.02× variation - **Trend:** Strong seasonality with sharp December spike

Month-by-Month Breakdown:

2025-03: 5.2M updates
2025-04: 4.8M
2025-05: 5.1M
2025-06: 4.9M
2025-07: 5.0M
2025-09: 4.7M
2025-10: 5.3M

2025-11: 3.8M
 2025-12: 10.5M (PEAK)
 2026-01: 0.6M (DROP)

Interpretation: - December peak likely driven by: - Year-end administrative drives - School enrollment deadlines - Financial benefit claims (LPG, subsidies) - January drop suggests post-holiday slowdown - Missing data (Feb, Aug) indicates project collection period

Monthly Growth Rates (07_monthly_growth_rates.png) **Volatility Analysis:** - **Highest Growth:** Nov→Dec (+175%) - **Sharpest Decline:** Dec→Jan (-94.5%) - **Average MoM Change:** ±40% typical - **Stable Periods:** Mar-Jul (consistent 4.7-5.3M range)

5. Seasonal Decomposition

Seasonal Patterns (08_seasonal_patterns.png) **Components Identified:**

1. **Trend Component:** Relatively flat 4-5M baseline
2. **Seasonal Component:**
 - Strong Q4 effect (Oct-Dec surge)
 - Q1 slowdown (Jan-Mar)
3. **Residual Component:** Low noise, high explainability

Policy Implications: - Predictable December surge enables resource pre-positioning - Q1 lull presents opportunity for targeted outreach campaigns - Seasonal budgeting should account for 3× December capacity needs

□ Summary Statistics

Descriptive Statistics Table

	child_updates	adult_updates	total_updates	child_share_pct
count	2,375,882	2,375,882	2,375,882	2,375,882
mean	1.90	19.12	21.02	9.07%
std	14.83	130.45	133.21	18.32%
min	0	0	0	0.00%
25% (Q1)	0	0	0	0.00%
50% (Q2)	0	5	5	0.00%
75% (Q3)	2	16	18	14.29%
max	2,690	16,166	16,190	100.00%

□ Key Insights

Child Documentation Gaps

- 1. **44% Zero Records:** Nearly half of all geographic areas recorded zero child updates
- 2. **Low Median:** Even at the 75th percentile, only 2 child updates per record
- 3. **Urban Concentration:** High-volume outliers cluster in city districts

Adult Update Patterns

- 1. **10× Volume Advantage:** Adults consistently update 10 times more than children
- 2. **More Distributed:** Less geographic inequality compared to child updates
- 3. **Service Center Effect:** Outliers indicate concentrated enrollment locations

Temporal Dynamics

- 1. **18× Seasonal Variation:** Strongest swing between December peak and January trough
 - 2. **Predictable Cycles:** Year-end surge is consistent and plannable
 - 3. **Data Gaps:** Missing months (Feb, Aug) limit full annual analysis
-

□ Visualizations Generated

File	Type	Description
01_child_updates_histogram.png	Histogram	Child update frequency distribution
02_child_updates_boxplot.png	Box Plot	Child update quartile analysis
03_adult_updates_histogram.png	Histogram	Adult update frequency distribution
04_adult_updates_boxplot.png	Box Plot	Adult update quartile analysis
05_child_vs_adult_comparative_boxplots.png	Comparative Box Plots	Side-by-side age group comparison
06_monthly_trends_timeseries.png	Time Series	Monthly update trends over 10 months

File	Type	Description
07_monthly_growth_line_chart	Line Chart	Month-over-month percentage changes
08_seasonal_pattern_decomposition	Seasonal Decomposition	Trend, seasonal, residual components

□ Next Steps

Based on univariate findings:

1. **Bivariate Analysis** → Explore relationships between child/adult updates
2. **Geographic Exploration** → Map zero-update clusters
3. **Layer 2: Child Risk** → Quantify documentation gap severity
4. **Seasonal Modeling** → Build predictive capacity planning tools

□ Statistical Methods

Techniques Applied

- **Descriptive Statistics:** Mean, median, mode, quartiles, range
- **Distribution Analysis:** Skewness, kurtosis, normality tests
- **Time Series:** Trend analysis, seasonal decomposition (additive model)
- **Growth Rate Calculation:** $\text{Growth} = \frac{V_t - V_{t-1}}{V_{t-1}} \times 100$

Software Tools

- **Python Libraries:** pandas, numpy, matplotlib, seaborn, statsmodels
- **Visualization:** Matplotlib 3.5+, Seaborn 0.11+
- **Statistical Tests:** Shapiro-Wilk (normality), Levene (homogeneity)

Last Updated: January 2026

Maintainer: ADIEWS Project Team